Diseño de Sistemas Neurocomputacionales en el Ámbito de la Biomedicina

TESIS DOCTORAL

D. Daniel Urda Muñoz

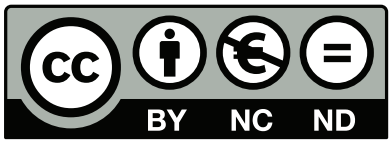
Departamento de Lenguajes y Ciencias de la Computación
Escuela Técnica Superior de Ingeniería Informática
Universidad de Málaga

Diciembre 2014

Publicaciones y
Divulgación Científica

AUTOR: Daniel Urda Muñoz

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga



Diseño de Sistemas Neurocomputacionales en el Ámbito de la Biomedicina

Memoria que presenta para optar al título de Doctor en Informática por la Universidad de Málaga

D. Daniel Urda Muñoz

Dirigida por los Doctores

Dr. D. José Manuel Jerez Aragonés
Dr. D. Leonardo Franco

Departamento de Lenguajes y Ciencias de la Computación
Escuela Técnica Superior de Ingeniería Informática
Universidad de Málaga

Diciembre 2014

UNIVERSIDAD DE MÁLAGA

Departamento de Lenguajes y Ciencias de la Computación
Escuela Técnica Superior de Ingeniería Informática
Universidad de Málaga

El Dr. D. José Manuel Jerez Aragonés, Profesor Titular de Universidad perteneciente al área de Lenguajes y Sistemas Informáticos, y el Dr. D. Leonardo Franco, Profesor Titular de Universidad perteneciente al área de Ciencia de la Computación e Inteligencia Artificial, ambos de la E.T.S. Ingeniería Informática de la Universidad de Málaga,

Certifican que,

D. Daniel Urda Muñoz, Ingeniero en Informática, ha realizado en el Departamento de Lenguajes y Ciencias de la Computación de la Universidad de Málaga, bajo su dirección, el trabajo de investigación correspondiente a su Tesis Doctoral titulada:

Diseño de Sistemas Neurocomputacionales en el Ámbito de la Biomedicina

Revisado el presente trabajo, estimamos que puede ser presentado al tribunal que ha de juzgarlo. Y para que conste a efectos de lo establecido en la legislación vigente, autorizamos la presentación de este trabajo en la Universidad de Málaga.

Málaga, Diciembre de 2014

Fdo.: Dr. José M. Jerez Aragonés

Fdo.: Dr. Leonardo Franco

A mi mujer, Esperanza, por sus consejos y apoyo incondicional, por estar siempre a mi lado en los buenos y malos momentos. Ojalá pueda devolverte tan solo la mitad del cariño que tú me has dado.

-

A mis padres, Jesús y Rosario, y a mi hermano, Jesús, por haberme formado como persona y animarme siempre a lograr mis objetivos, sin importar la dificultad del camino.

"La intersección de la genómica y la bioinformática en la práctica clínica es una necesidad que debe incorporarse en la investigación médica si se quiere que la revolución genómica incida efectivamente en la medicina."

-

Chris Gunter

Agradecimientos

"La gratitud da sentido a nuestro pasado, trae paz al presente y crea una visión para el mañana."

En primer lugar, me gustaría expresar mi más sincero agradecimiento tanto a José Manuel Jerez Aragonés como a Leonardo Franco, los directores de mi tesis doctoral. Ambos han creído en mi capacidad y potencial, y me dieron la oportunidad de descubrir este mundo que ni en sueños hubiera imaginado al finalizar la carrera. Sin las enseñanzas de ambos, su constante apoyo y paciencia durante muchos momentos de este largo viaje, estas líneas no estarían escritas. Ha sido también todo un lujo y un privilegio haber formado parte del grupo de investigación de Inteligencia Computacional en Biomedicina (ICB), donde he podido desarrollarme tanto personal como profesionalmente.

He de expresar también mi gratitud, de forma muy especial, a muchos compañeros y amigos con los que he compartido momentos muy importantes en este caminar. A Jose Luis, Rafa, Paco y Julio, con los que he tenido la oportunidad y el placer de poder trabajar codo con codo en algún momento puntual de mi carrera investigadora. Tampoco puedo olvidarme Esteban, Héctor, Cristóbal, Patxi, Antonio, Roberto, Mónica, María, Pilar y Jose, todos ellos compañeros del laboratorio que siguen o han estado en algún momento de estos años al lado, viviendo y conociendo las horas y el empeño que he dedicado en que este trabajo saliera adelante.

Tampoco me puedo olvidar del grupo de investigación que dirige el profesor Paulo Lisboa. Tanto Paulo, Ian, Simon, Sandra e Iván me acogieron como uno más desde el primer minuto en su equipo de trabajo en Liverpool, Reino Unido. Durante mi estancia en Liverpool me ofrecieron su ayuda y experiencia para profundizar en la línea de investigación de esta tesis doctoral y por todo ello les estoy profundamente agradecido.

Por último, no quiero acabar estas líneas sin dar las gracias, de todo corazón, a mi mujer, Esperanza, que ha vivido de primera mano todas mis frustraciones y agobios, se ha alegrado por mis satisfacciones y logros y siempre ha estado ahí en todo momento y lugar con su apoyo incondicional. Tampoco

puedo olvidar a mis maravillosos padres, Jesús y Rosario, y a mi hermano Jesús: gracias a ellos soy lo que soy, formándome como persona, apoyándome y guiándome desde que tengo uso de razón a alcanzar aquellos objetivos que me he planteado, enseñándome que con trabajo, voluntad y esfuerzo todos ellos pueden conseguirse.

De nuevo, gracias, pues este trabajo también es en parte al apoyo de todos vosotros.

Resumen

El área de la biomedicina es un área extensa en el que las entidades públicas de cada país han invertido y continúan invirtiendo en investigación una gran cantidad de financiación a través de proyectos nacionales, europeos e internacionales. Los avances científicos y tecnológicos registrados en los últimos quince años han permitido profundizar en las bases genéticas y moleculares de enfermedades como el cáncer, y analizar la variabilidad de respuesta de pacientes individuales a diferentes tratamientos oncológicos, estableciendo las bases de lo que hoy se conoce como medicina personalizada. Ésta puede definirse como el diseño y aplicación de estrategias de prevención, diagnóstico y tratamiento adaptadas a un escenario que integra la información del perfil genético, clínico, histopatológico e inmuhistoquímico de cada paciente y patología. Dada la incidencia de la enfermedad de cáncer en la sociedad, y a pesar de que la investigación se ha centrado tradicionalmente en el aspecto de diagnóstico, es relativamente reciente el interés de los investigadores por el estudio del pronóstico de la enfermedad, aspecto integrado en la tendencia creciente de los sistemas nacionales de salud pública hacia un modelo de medicina personalizada y predictiva.

El pronóstico puede ser definido como conocimiento previo de un evento antes de su posible aparición, y puede enfocarse a la susceptibilidad, supervivencia y recidiva de la enfermedad. En la literatura, existen trabajos que utilizan modelos neurocomputacionales para la predicción de casuísticas muy particulares como, por ejemplo, la recidiva en cáncer de mama operable, basándose en factores pronóstico de naturaleza clínico-histopatológica. En ellos se demuestra que estos modelos superan en rendimiento a las herramientas estadísticas tradicionalmente utilizadas en análisis de supervivencia por el personal clínico experto. Sin embargo, estos modelos pierden eficacia cuando procesan información de tumores atípicos o subtipos morfológicamente indistinguibles, para los que los factores clínicos e histopatológicos no proporcionan suficiente información discriminatoria. El motivo es la heterogeneidad del cáncer como enfermedad, para la que no existe una causa individual caracterizada, y cuya evolución se ha demostrado que está determinada por factores no sólo clínicos sino también genéticos. Por ello, la integración de los datos clínico-histopatológicos y proteómico-genómica pro-

porcionan una mayor precisión en la predicción en comparación con aquellos modelos que utilizan sólo un tipo de datos, permitiendo llevar a la práctica clínica diaria una medicina personalizada.

En este sentido, los datos de perfiles de expresión provenientes de experimentos con plataformas de microarrays de ADN, los datos de microarrays de miRNA, o más recientemente secuenciadores de última generación como RNA-Seq, proporcionan el nivel de detalle y complejidad necesarios para clasificar tumores atípicos estableciendo diferentes pronósticos para pacientes dentro de un mismo grupo protocolizado. El análisis de datos de esta naturaleza representa un verdadero reto para clínicos, biólogos y el resto de la comunidad científica en general dado el gran volumen de información producido por estas plataformas. Por lo general, las muestras resultantes de los experimentos en estas plataformas vienen representadas por un número muy elevado de genes, del orden de miles de ellos. La identificación de los genes más significativos que incorporen suficiente información discriminatoria y que permita el diseño de modelos predictivos sería prácticamente imposible de llevar a cabo sin ayuda de la informática. Es aquí donde surge la Bioinformática, término que hace referencia a cómo se aplica la ciencia de la información en el área de la biomedicina.

El objetivo global que se intenta alcanzar en esta tesis consiste, por tanto, en llevar a la práctica clínica diaria una medicina personalizada. Para ello, se utilizarán datos de perfiles de expresión de alguna de las plataformas de microarrays más relevantes con objeto de desarrollar modelos predictivos que permitan obtener una mejora en la capacidad de generalización de los sistemas pronóstico actuales en el ámbito clínico. Del objetivo global de la tesis pueden derivarse tres objetivos parciales: el primero buscaría (i) pre-procesar cualquier conjunto de datos en general y, datos de carácter biomédico en particular, para un posterior análisis. Al acudir a un repositorio de datos público para buscar y obtener un conjunto de datos, es común encontrarse ante un conjunto de datos completo, válido y listo para realizar cualquier tipo de análisis. Sin embargo, al trabajar con datos reales lo más normal es encontrarse ante un conjunto de datos que presente muestras incompletas y que requiere un pre-procesamiento. El desarrollo de un sistema que integre diferentes métodos estándar de imputación de datos perdidos facilitará a cualquier investigador la tarea de pre-procesar un conjunto de datos antes de realizar el análisis. Un segundo objetivo parcial buscaría (ii) analizar las principales deficiencias existentes en los sistemas de información actuales de un servicio de oncología para así desarrollar un sistema de información oncológico que cubra todas sus necesidades. A día de hoy, este tipo de sistemas suelen ser muy generalistas originando bastantes reticencias a su uso dentro del propio personal clínico. Además, no ofrecen la posibilidad de utilizar la información clínico-histopatológica almacenada para predecir la evolución clínica de cada paciente. El diseño y desarrollo de un sistema oncológico

específico que integre modelos predictivos basados en la información almacenada en el sistema facilitará a los clínicos el diagnóstico y la determinación del tratamiento a aplicar a un paciente. Dada la relevancia de la información proteómica-genómica en la casuística del cáncer, el tercer objetivo parcial de la tesis buscaría (iii) desarrollar nuevos modelos predictivos basados en perfiles de expresión obtenidos a partir de alguna plataforma de secuenciación. El desarrollo de estos modelos lleva implícito una selección de características o genes significativos (comúnmente conocido como firma genética). Las firmas genéticas encontradas, además de poseer altas capacidades predictivas, deben poseer características tales como robustez y relevancia biológica en la enfermedad. El desarrollo satisfactorio de estos modelos permitiría integrarlos en un sistema oncológico de información junto a los modelos de predicción que usan información clínico-histopatológica de cada paciente. Esto ofrecería a los clínicos la posibilidad de llevar a cabo en la práctica clínica diaria una medicina personalizada de cada paciente, proporcionando así una mayor precisión en el diagnóstico, evolución y tratamiento del paciente.

El pre-procesado de un conjunto de datos es un paso previo necesario a cualquier tipo de análisis que se desee realizar sobre él. En un marco de trabajo ideal, el conjunto de datos que se tiene para analizar estará completo y perfectamente válido para trabajar con él. Esto suele ser bastante habitual cuando se acude a repositorios públicos que contienen numerosos conjuntos de datos de diferentes áreas y que ya han sido pre-procesados con anterioridad. No obstante, la situación de partida inicial cuando se trabaja con conjuntos de datos reales es bien diferente. Cuando los datos con los que se trabaja no son datos de repositorios públicos sino datos provenientes de mediciones reales o ensayos experimentales, resulta muy habitual encontrarse con conjuntos de datos incompletos que requieren un pre-procesamiento. En clínica, por ejemplo, es muy común disponer de información de un conjunto de pacientes en los que determinadas variables no contengan información específica para una determinada muestra. Esto problema es conocido en la literatura como *missing data* o valores perdidos en los datos. Cuando el objetivo del análisis, además, pasa por hacer cualquier tipo de clasificación de los datos, entonces ocurre que los algoritmos estándares de clasificación no se comportan todo lo bien que de ellos se espera.

Por lo general, una medida sencilla y directa para solucionar este problema sería descartar aquellas muestras del conjunto de datos que presenten valores perdidos. De esta forma, el subconjunto resultante que quedaría sí estaría completo y sería válido para analizar con cualquier algoritmo de clasificación. Sin embargo, en conjuntos de datos biomédicos esta opción no es contemplable. Por un lado, este tipo de conjuntos suele presentar un número pequeño de muestras de manera que si, además, se eliminan aquellas muestras que estén incompletas, la eficacia del clasificador se vería trastocada negativamente. Además, los conjuntos de datos de perfiles de expresión

proteómica-genómica se caracterizan por tener muy pocas muestras (del orden de decenas) en comparación con el elevadísimo número de características que describen cada muestra (del orden de miles), por lo que la eficacia del clasificador disminuiría aún más si cabe. Por otro lado, en conjuntos de datos de estas características tampoco existe la opción de volver a repetir el experimento, bien por el costo que tiene asociado o bien debido a que el tejido que se utilizó para realizar el experimento no se encuentre ya disponible. Por tanto, surge la necesidad de desarrollar métodos efectivos para procesar este tipo de conjuntos de datos en particular, y cualquiera en general, e imputar o estimar los valores perdidos en base a los valores contenidos en el resto de muestras del conjunto.

En líneas generales, los diferentes métodos de imputación ya existen y están publicados en la literatura. Aun así, existe una serie de problemas comunes a un estudio de este tipo que surgen cuando se requiere aplicar un método de imputación. Primero, se debe conocer en profundidad cómo funcionan cada uno de los diferentes métodos de imputación que existen para, a continuación, poder deducir cuál de ellos es más apropiado para aplicar al conjunto de datos en cuestión. Una vez que el investigador se decide por el método, el siguiente problema que probablemente se tope es dónde encontrar la implementación del método o algún software, muy posiblemente de pago, que ya lo incorpore. En ocasiones, si el método no es muy complicado la solución pasa directamente por hacer una implementación propia del mismo, con el coste extra de validar la implementación, pero esto no siempre es posible y acaba llevando al investigador a pagar la licencia de un software comercial para poder realizar la imputación en el conjunto de datos. Finalmente, si suponemos resueltos los anteriores problemas, podría llegar a darse el caso en grupos de investigación pequeños de no disponer de recursos computacionales suficientes para poder ejecutar los métodos elegidos sobre un conjunto de datos y así poder obtener el resultado en un tiempo razonable.

Hasta la fecha, no se conoce ningún sistema público y gratuito y con una interfaz amigable que ofrezca la posibilidad a sus usuarios de realizar la imputación de datos perdidos, dando a elegir varios métodos de imputación disponibles, en un conjunto de datos cualquiera y en un tiempo razonable. Por ello, en el Capítulo 2 de esta tesis se describe el diseño y desarrollo de un sistema web denominado WIMP (Web IMPutation). Este sistema web es una herramienta web pública y gratuita para poder ser utilizada por los usuarios registrados en ella. WIMP ofrece a la comunidad investigadora la posibilidad de imputar valores perdidos en los conjuntos de datos que cada usuario suba al sistema y eligiendo entre varios métodos de imputación que se incorporan en él en su versión inicial. Estos métodos ya están implementados y previamente testados, es decir, el usuario no debe preocuparse de buscar ningún método de imputación o bien de pagar por un software comercial que les ofrezca esta posibilidad. De entre los diferentes métodos de

imputación de datos existentes, WIMP incorpora algunos basados en técnicas estadísticas (imputación por la media, imputación hot-deck e imputación múltiple) y otros basados en algoritmos de aprendizaje (imputación con mapas auto-organizados e imputación con un perceptrón multicapa). El número de métodos que incorpora el sistema en su versión inicial no es algo cerrado, sino que puede ampliarse en un futuro tan solo implementando otros métodos existentes, validarlos e incorporarlos en el núcleo del sistema.

Por otro lado, otra gran ventaja que ofrece WIMP a sus usuarios es la de evitarles la necesidad de disponer de una gran capacidad de cálculo computacional. Las tareas de imputación que los usuarios soliciten a través del sistema se procesan en el lado del servidor. Es más, el servidor está conectado a un clúster de computación, que es donde en realidad se ejecutan las tareas solicitadas por los usuarios del sistema. Esto posibilita que un simple usuario, desde su ordenador de casa, pueda ejecutar un método de imputación complejo y obtener así en poco tiempo el conjunto de datos completo tras la imputación realizada. Una vez que una tarea de imputación lanzada por un usuario finaliza, el servidor avisa al usuario a través del e-mail con un enlace de descarga que le permita recoger el conjunto de datos resultante de la ejecución. Este conjunto de datos está compuesto, ahora sí, por muestras completas y está disponible para realizar cualquier tipo de análisis sobre él, alcanzado así el primer objetivo parcial de esta tesis.

A continuación, en el Capítulo 3 se aborda el segundo objetivo parcial de esta tesis analizando las principales deficiencias existentes en los sistemas de información actuales de un servicio de oncología para así desarrollar un sistema de información oncológico que cubra todas sus necesidades y permita llevar a la práctica clínica diaria los modelos predictivos basados en la información clínico-histopatológica de cada paciente. En concreto, el grupo de investigación Inteligencia Computacional en Biomedicina (ICB) de la universidad de Málaga mantiene una estrecha colaboración con el equipo médico del servicio de oncología del hospital universitario Virgen de la Victoria (HUVV) de Málaga, España. Esta estrecha relación permite analizar con detalle las necesidades existentes en un servicio de oncología clínica de un hospital con objeto de poder implantar con éxito un sistema de estas características que incorpore, junto a la información clínico-histopatológica de cada paciente, modelos predictivos que ofrezcan al clínico mayor información para decidir acerca del diagnóstico, evolución y tratamiento a aplicar a un paciente.

En este capítulo de la tesis se muestra en detalle toda la experiencia adquirida en el diseño e implementación de un sistema de información oncológica para el servicio de oncología del HUVV. Tras un exhaustivo análisis, se analiza, detalla y aporta solución a los aspectos más críticos que faciliten alcanzar el éxito en la implantación del sistema. El primer aspecto a tener en cuenta se trata de la usabilidad del sistema. Un diseño centrado en el usuario

facilitará la participación activa de los usuarios finales del sistema a lo largo de su desarrollo, haciendo desde el principio que la experiencia del usuario sea agradable y su interacción con el sistema sencilla, evitando posibles futuros rechazos al uso del sistema. Otro aspecto a tener en cuenta es el uso de una tecnología adecuada que facilite todo el desarrollo así como el mantenimiento y futuras ampliaciones del sistema, pues de nada serviría usar la última tecnología disponible en el mercado si con ella no se logra desarrollar un sistema que facilite alcanzar una mejor medicina. También es importante que este tipo de sistemas integre en su totalidad todas las rutinas clínicas diarias del servicio de oncología y, para ello, nada mejor que particularizar el diseño y desarrollo en función del servicio en el que se pretende implantar el sistema. Soluciones generalistas que pretendan abarcar la implantación del sistema en numerosos hospitales hará que no se capturen todos los requisitos y necesidades propios de cada servicio de oncología provocando, en mayor o menor medida, reticencias a usar el sistema desarrollado. La seguridad del sistema, garantizando la privacidad de los datos contenidos en él, y su integración con otros sistemas de información son otros aspectos a tener en cuenta. Un sistema aislado con mucha información no sirve de mucho debido a la alta especialización del sistema. Para complementar la información contenida en un sistema de información oncológica hace falta interactuar con otros sistemas tales como radiología, hematología, farmacia, etc. Por último, un aspecto clave que marca la diferencia en este tipo de sistemas está en habilitar la posibilidad, dentro del propio sistema, de utilizar la información almacenada para realizar estudios de diferentes características sin necesidad de tener que recurrir a otro software ajeno al sistema.

El sistema final desarrollado e implantado con éxito en el HUVV consiste en una aplicación web con una arquitectura en capas que incorpora toda la información clínico-histopatológica de cada paciente. La arquitectura en capas del sistema busca una mayor usabilidad, la facilidad de mantenimiento y de futuras ampliaciones del sistema. Además, su diseño contempla la distribución de la información en los siguientes módulos funcionales: gestión de pacientes, hospital de día, ensayos clínicos, consejo genético y análisis estadístico. La característica más importante que hace que este sistema destaque sobre otros es que, además de almacenar toda la información relativa a los pacientes del servicio, contiene un módulo específico para realizar diferentes estudios a partir de esa información. Este módulo integra varios modelos de pronóstico clínico fruto de la investigación del grupo ICB en trabajos anteriores, ofreciendo a los clínicos la posibilidad de realizar análisis de supervivencia implementado bajo el algoritmo de Kaplan-Meier, análisis de regresión de Cox, calcular funciones de riesgo y obtener tablas de contingencia en base a la información clínico-histopatológica de un subconjunto de pacientes previamente seleccionado. La evaluación del éxito en la implantación del sistema se llevó a cabo a partir de encuestas realizadas a los usuarios finales

del mismo a los 3 y 15 meses tras su implantación en el hospital, mostrando una satisfacción generalizada del personal.

A pesar de lograr una exitosa implantación del sistema en el HUVV, en su versión actual sólo se dispone de información clínico-histopatológica de cada paciente. En sí, este sistema ha supuesto un gran paso hacia adelante en dirección a llevar a la práctica clínica diaria una medicina personalizada. Sin embargo, tal como se comentó anteriormente la heterogeneidad del cáncer como enfermedad hace que su evolución esté determinada por factores no sólo clínicos sino también genéticos, lo que lleva a plantearse la necesidad de integrar en un futuro la información proteómico-genómica de cada paciente dentro del sistema. A su vez, surge la necesidad de desarrollar nuevos modelos predictivos basados en datos de perfiles de expresión para incorporarlos a los ya existentes en el sistema. De esta forma, se daría la posibilidad a los clínicos de poder practicar una medicina personalizada a sus pacientes atendiendo a la información de ambos tipos y a los modelos predictivos incorporados en el modulo de análisis estadístico.

En este sentido, se abrió paralelamente una línea de investigación para analizar conjuntos de datos de perfiles de expresión génica. Este tipo de datos da lugar a diferentes tipos de análisis, cada uno de ellos lo suficientemente amplios para investigar. De forma global, se podrían diferenciar tres tipos de análisis principales: (i) descubrir la clase asociada a cada muestra, (ii) predecir la clase asociada a una nueva muestra y (iii) comparar muestras de la misma clase. El primero de los análisis tiene sentido cuando se desconoce la clase o etiqueta asociada a cada una de las muestras del conjunto de datos. Este tipo de análisis trata de dividir las muestras en clases o grupos a partir de patrones ocultos en los datos. En el segundo tipo de análisis, por contra, sí se conoce a priori la clase o etiqueta asociada a cada una de las muestras del conjunto de datos. En este caso, el reto consiste en intentar predecir la clase de nuevas muestras a partir del conocimiento o patrones ocultos en los datos iniciales y es, justamente, la línea de investigación que se avanzará en los Capítulos 4-6 para alcanzar el tercer objetivo parcial que se plantea en esta tesis. Por último, en el tercer tipo de análisis también se conoce a priori las clases de cada muestra del conjunto y lo que se pretende es descubrir los genes y los patrones de expresión que permiten diferenciar mejor ambos grupos por medio de la comparación de muestras de distinta clase.

En conjuntos de datos de otras áreas más cercanas a la Inteligencia Artificial (IA), el desarrollo de modelos predictivos es una tarea que está ampliamente estudiada y resuelta en la literatura, comúnmente conocido como *machine learning*. Por tanto, cabría esperar la posibilidad de que el desarrollo de modelos predictivos en base a datos de perfiles de expresión resulte un objetivo no demasiado difícil de alcanzar. Sin embargo, el proceso de clasificación o predicción en conjuntos de datos proteómico-genómicos no es una tarea trivial debido a la propia naturaleza de los datos. Estos conjuntos

de datos tienen una dimensionalidad muy alta, del orden de miles a decenas de miles de genes por muestras y tan solo unas decenas o un centenar de muestras. Esto hace que el uso de cualquier algoritmo de clasificación conlleve un sobreajuste o sobreentrenamiento del modelo, intentando clasificar todas y cada una de las muestras (incluso posibles muestras ruidosas) y, por tanto, produciendo una baja generalización en datos a futuro. Este problema se conoce ampliamente en la literatura como *curse of dimensionality*. Como consecuencia, este tipo de análisis lleva implícito un proceso de selección de características que identifica los genes más significativos y con mayor información discriminatoria entre las clases. Por tanto, el desarrollo de un modelo predictivo con datos de perfiles de expresión parte del conjunto original y realiza una búsqueda de un subconjunto de genes, conocido como firma genética, cuyos patrones de expresión sirven para diferenciar los grupos existentes en el análisis. Este proceso de selección de características se organiza en tres categorías dependiendo de cómo se combinan la búsqueda de selección de características con la construcción del modelo de clasificación: métodos de filtrado (filter) , métodos de envoltura (wrapper) y método embebidos (embedded).

El Apéndice A plantea el punto inicial de partida en esta línea de investigación. En él, se plantea el uso de una red neuronal constructiva (CNN) como clasificador alternativo a las redes neuronales artificiales clásicas (ANN). A su vez, como estrategia de selección de características se escoge un método wrapper ampliamente utilizado en la literatura como es el método Stepwise Forward Selection (SFS). El funcionamiento de este método es simple. En su primera iteración, evalúa la predicción de cada modelo utilizando un solo gen, escoge el modelo de mayor predicción y avanza a la iteración siguiente, que estudiará los modelos formados por dos genes (el que se acaba de seleccionar combinado con uno más de los restantes genes del conjunto). El método seguirá avanzando en iteraciones hasta que el modelo formado por un número de genes X no mejore la predicción del mejor modelo hallado en la iteración anterior (de $X-1$ número de genes). Los resultados obtenidos por la estrategia SFS+CNN planteado son competitivos en comparación con los obtenidos por la estrategia SFS+ANN tradicional utilizada en la literatura para identificar firmas genéticas con alta capacidad de predicción. Además, la ventaja añadida de utilizar una CNN reside en que la arquitectura de este tipo de red neuronal se ajusta completamente al problema que se intenta aprender en cada iteración del método SFS, mientras que si se utiliza una ANN se debería prefijar una arquitectura inicial "óptima" para todas las iteraciones. Pero, ¿cómo asegurar que esa arquitectura es la óptima para problemas de distinta dimensión (uno, dos, tres genes, etc.) que se plantea en el método SFS en función de la iteración en que se encuentre?

El siguiente paso llevó a pensar en un planteamiento que intentara aprovechar por un lado las ventajas de un método wrapper como es el caso del

SFS+CNN (resultados con muy buena predicción) y por otro las de los métodos de filtrado (menor complejidad y mucho más rápidos en su ejecución). Esta necesidad derivó en el planteamiento de un modelo híbrido que reunía ambas ventajas, descrito en detalle en el Apéndice B. Este método se basa en la capacidad de predicción individual de cada gen presente en los datos utilizando un modelo de red neuronal constructivo, C-Mantec, junto al uso del coeficiente de correlación lineal entre dos variables. El método desarrollado plantea tres estrategias de selección de variables: la primera, *ROnly*, se basa únicamente en la generalización escogiendo directamente los 10 genes con mayor capacidad individual de predicción; la segunda, *RelevanceF*, combina generalización y redundancia, de forma que del 10 % de genes con mayor capacidad individual de predicción, finalmente se escogen los 10 genes que presenten menor redundancia entre ellos; y la tercera de ellas, *RedundancyF*, es justo lo contrario a la anterior, es decir, del 10 % de genes con menor redundancia entre ellos se seleccionan los 10 con mayor capacidad individual de predicción. Comparando estas tres estrategias con los resultados del método SFS+CNN, se puede deducir como la estrategia *RedundancyF* obtiene unos resultados muy similares a SFS+CNN en cuanto a predicción. No obstante, los requisitos en cuanto a complejidad y necesidad de cómputo son mucho menores en la estrategia *RedundancyF* y, por tanto, el tiempo necesario para estimar la firma genética es considerablemente menor en este modelo híbrido.

Tras estos trabajos preliminares orientados a desarrollar nuevos modelos predictivos usando datos de perfiles de expresión, se llevó a cabo una revisión extensa de la literatura para identificar métodos de selección de características utilizados en este tipo de conjuntos de datos. En esta línea, otra familia de algoritmos muy utilizada para identificar firmas genéticas con altas capacidades predictivas son los algoritmos genéticos (GA). Los GA son, también, un método wrapper que en cada generación del algoritmo se evalúan un número determinado de individuos (o subconjunto de genes) utilizando un algoritmo de clasificación. El objetivo de un GA es minimizar una función, siendo la clave de este tipo de algoritmos lo que se conoce como *función de fitness*, una función matemática que se encarga de evaluar cómo de bueno es cada individuo que compone la población de la generación de estudio. Por lo general, los trabajos publicados que utilizan un GA para estimar modelos predictivos definen esta función buscando minimizar el error en clasificación (lo que equivale a maximizar la tasa de predicción) al mismo tiempo que se priorizan subconjuntos de genes compuestos por un menor número de genes. En base a esta evaluación de la población, los mejores individuos de la población pasan directamente a la siguiente generación mientras que los restantes individuos evolucionan a nuevos individuos a partir de operandos de selección y cruce, simulando el proceso reproducción biológica.

En este este sentido, el Capítulo 4 propone el uso de un GA utilizando

C-Mantec como algoritmo de clasificación para la estimación de modelos predictivos con datos de perfiles de expresión. Aparte de la utilización de una red neuronal constructiva como clasificador dentro del método de selección de características, el GA desarrollado incluye principalmente dos novedades no contempladas anteriormente en la literatura: (i) la realización de un pre-filtrado de genes para su posterior análisis con el método GA+CMantec, y (ii) la definición de una función de fitness del GA que incluya información sobre la correlación existente en el subconjunto de genes que se evalúa. Debido a que los GA se tratan de métodos wrapper y buscando disminuir la alta necesidad de cómputo que estos métodos requieren, se incluye una primera novedad en la estrategia GA+CMantec propuesta que busca pre-filtrar, antes de comenzar el análisis, el conjunto total de genes aplicando el test estadístico Welch's t-test. Este test estadístico asigna un valor a cada gen, conocido como *p-value*, que indica cómo de importante es ese gen para discriminar ambas clases o grupos del conjunto de datos. Como resultado de este test, sólo se retiene el 5 % total de los genes, siendo estos los que pasan a ser analizados con el método GA+CMantec. Por otro lado, la segunda novedad del GA propuesto se encuentra en la definición de la función de fitness. Ésta se compone de tres términos. El primero de ellos busca minimizar el error de clasificación del algoritmo C-Mantec (se prefieren subconjuntos de genes con mejor tasa de predicción). El segundo término persigue minimizar el número de genes que compongan la firma genética encontrada (se prefieren soluciones con el menor número de genes posible). Como novedad respecto a los trabajos publicados en la literatura, se incluye un tercer término que busca minimizar la correlación entre variables (se penaliza a aquellos subconjuntos de genes que presenten información redundante). Además, los tres términos que forman la función de fitness se encuentran ponderados por unas tasas constantes que permiten dar mayor o menor importancia a uno u otro término de cara a guiar la búsqueda hacia soluciones que cumplan unas características determinadas.

Para validar los modelos predictivos obtenidos a partir de la estrategia GA+CMantec, se comparó los resultados obtenidos con los de la estrategia SFS+CNN planteada en el trabajo preliminar incluyéndole el pre-filtrado de genes acorde al resultado de aplicar el Welch's t-test. Además, aparte del algoritmo C-Mantec se utilizaron otros algoritmos de clasificación ampliamente conocidos tales como Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), Redes Bayesianas (NB), k-Nearest Neighbors (kNN) y el Perceptrón Multicapa (MLP). Todas estas estrategias se probaron utilizando conjuntos de datos de microarrays de ADN de distintos tipos de cáncer. Los resultados obtenidos muestran en general mejores resultados de predicción de la estrategia GA respecto a SFS. Por contra, la estrategia GA obtiene como resultado subconjuntos de genes más grandes y es computacionalmente más exigente que SFS al evaluar muchos más subconjuntos de genes. Respecto a

la influencia del algoritmo de clasificación en la estrategia de selección, con MLP, LDA y SVM se obtienen los mejores resultados mientras que C-Mantec y kNN le siguen de cerca aunque con una predicción ligeramente menor. Además, C-Mantec, MLP y LDA dan lugar a firmas genéticas con menor número de genes en comparación con SVM, NB y kNN, y en particular C-Mantec es el clasificador más robusto en cuánto al ajuste de los parámetros del modelo.

Llegados a este punto, se encontró la necesidad de analizar con mayor detenimiento las tres tareas más complejas que hay que abordar al estimar modelos predictivos con datos de perfiles de expresión, que son: la selección de características, el modelo de clasificación y la estrategia de validación. Revisando la literatura, se puede encontrar trabajos que centran su esfuerzo en solventar, de forma individual, los problemas que surgen en alguna de esas tareas. Por ejemplo, algunos se centran en ver que método de selección de características es mejor sin prestar atención a la estrategia de validación utilizada. Otros autores centran su atención únicamente en las ventajas de utilizar un algoritmo de clasificación u otro sin tener en cuenta el método de selección de características. Sin embargo, hasta la fecha no se encontró ningún trabajo que englobe todas estas tareas a la vez ofreciendo una visión global del problema.

El Capítulo 5 presenta una extensa comparativa entre diversos métodos de selección de características existentes utilizando varios algoritmos estándar de clasificación. Los métodos de selección de características considerados en este estudio son los métodos SFS y GA descritos con anterioridad (métodos wrapper), los métodos Correlation-based Feature Selection (CFS), Consistency-based Filter (Cons), Information Gain (IG), Minimun-Redundancy Maximum-Relevance (mRMR) y ReliefF (todos ellos métodos de filtrado), y SVM-Recursive Feature Elimination (método embebido). Elegido uno de estos métodos de selección, se han utilizado varios algoritmos de clasificación como son LDA, SVM, kNN, NB, C-Mantec y MLP. Además, para evaluar las diferentes estrategias que surgen combinando los métodos de selección con los algoritmos de clasificación, se aplicó un esquema de validación honesto. En este sentido y dado que el número de muestras disponibles es muy bajo, es muy recomendable el uso de metodologías que utilicen técnicas de *resampling*. Por ello, la estrategia que se plantea utilizar para esta comparativa es una estrategia Bootstrap-CrossValidation (BCV) que permita obtener la tasa de predicción de una validación externa y al set también muy recomendada cuando hay muy pocas muestras en el análisis, tal y como es nuestro caso. Los resultados obtenidos tras ejecutar, bajo el mismo esquema de validación descrito, las diferentes estrategias surgidas de combinar los métodos de selección con los algoritmos de clasificación, muestran que los métodos de filtrado o embebidos proporcionan mayor predicción siendo menos complejos y más rápidos de ejecutar cque los métodos wrapper. Además, se analizó la robustez de las firmas genéticas obtenidas en función del

método de selección, y de nuevo los métodos de filtrado aparecen, en general, como los métodos más robustos.

La robustez y relevancia biológica de las firmas genéticas, junto a su capacidad predictiva, son otras características muy deseadas. El principal inconveniente que presentan estos los métodos selección de características, independientemente de cuál se aplique, es la escasa robustez y relevancia biológica que ofrecen las soluciones finales, es decir, cada ejecución del método produce una firma genética diferente. Este tipo de soluciones resultan poco útiles de cara a utilizar estas firmas genéticas en estudios clínicos o biomédicos. Lo ideal sería poder, de alguna forma, dotar a los métodos de selección de características de algún mecanismo que aporte información biológica de la enfermedad de estudio al propio proceso. Si esto se consiguiera, el método de selección podría promocionar aquellos genes más relevantes desde el punto de vista clínico/médico para que formen parte de la solución final y, al contrario, descartar aquellos genes que poco o nada influyen en la enfermedad de estudio. De esta forma, además de tener más garantías acerca de la relevancia biológica de las firmas genéticas que se encuentren, sería de esperar que la robustez de las mismas también aumente ya que la búsqueda se restringe y se guía hacia soluciones cuyas características son más restrictivas que antes.

En esta el Capítulo 6 se propone un nuevo método de selección de características en dos etapas con el objetivo de obtener firmas genéticas robustas, con alta capacidad predictiva y de relevancia biológica. Este nuevo método utiliza información biológica asociada a la enfermedad de estudio. En este sentido, se utiliza la información disponible en la base de datos Kyoto Enciclopedia of Genes and Genomes (KEGG) al ser una base de datos de referencia ampliamente utilizada en los últimos años como fuente importante para la construcción de modelos de pathways (vías metabólicas) para descifrar el genoma y vincularlo a los sistemas biológicos. Esta base de datos se basa en la asignación manual de un código KEGG a una determinada secuencia que haya implicado una evidencia en un ensayo experimental. Si una proteína codificada por una secuencia no produce actividad enzimática o no es parte de un pathway, nunca tendrá asociado un código KEGG. Afortunadamente, la mayoría de los genes implicados en el cáncer tienen actividad enzimática y pertenecen a algún pathway.

Las dos etapas del nuevo modelo que se propone en esta tesis consisten en lo siguiente: (i) la primera etapa hace un filtrado inicial, basada en la base de datos KEGG, del número total de genes reteniendo únicamente aquellos que representan una enzima. Además, establece un ranking ordenado de todos los pathways asociados a la enfermedad de estudio acorde a la capacidad predictiva de los genes que forman parte de cada pathway junto al número de palabras clave (previamente identificadas por un experto en el área) encontradas en un proceso de *data mining* sobre la descripción tex-

tual de cada pathway ofrecida en la web de KEGG. A continuación, (ii) la segunda etapa ejecuta un GA sobre los mejores pathways para estimar un modelo predictivo. La principal novedad de este GA respecto al explicado con anterioridad se encuentra en la función de fitness. De nuevo, esta función está compuesta por tres términos. Los dos primeros son idénticos al anterior GA (minimización del error en clasificación y minimización del número de genes que componen la firma genética). Sin embargo, el tercer término, en esta ocasión, trata de maximizar una función que puntúa la "calidad" de la firma en base al número de genes que ésta contiene y que además forman parte del pathway de estudio, o bien, de cualquiera de los otros pathways seleccionados en la primera etapa para analizar. De igual forma que en el anterior GA propuesto, cada uno de estos términos está ponderado por una tasa constante que permite guiar la búsqueda hacia soluciones que cumplan ciertas características. Tras evaluar este nuevo método en tres conjuntos de datos de microarrays de ADN de distintos tipos de cáncer, los resultados alcanzados muestran que esta nueva estrategia mejora la consistencia, robustez y predicción de las firmas genéticas obtenidas. Este nuevo enfoque ofrece la posibilidad de definir firmas genéticas que puedan ser utilizadas para predecir el diagnóstico y la evolución clínica de pacientes con cáncer.

Finalmente, se puede concluir que los resultados obtenidos en esta tesis doctoral permitirían ofrecer, en un futuro cercano, una medicina personalizada en la práctica clínica diaria. Los modelos predictivos basados en datos de perfiles de expresión que se han desarrollado en la tesis podrían integrarse en el propio sistema de información oncológico implantado en el HUVV. Además, se podría incorporar la información proteómico-genómica de cada paciente para poder aprovechar al máximo las ventajas añadidas mencionadas a lo largo de esta tesis. Por otro lado, gracias a todo el trabajo realizado en esta tesis, el doctorando ha podido profundizar y adquirir una extensa formación investigadora en un área tan amplia como es la Bioinformática.

Summary

The biomedicine is a vast area in which public entities of each country have invested and still continue invest in research a lot of funding through national, European and international projects. Scientific and technological progress registered in the last fifteen years have allowed to go further in the genetic and molecular basis of diseases like cancer, as well as analyze the variability of response of individual patients to different cancer treatments, establishing the basis for what is currently known as personalized medicine. This can be defined as the design and implementation of strategies for prevention, diagnosis and treatment tailored to a scenario that integrates information from genetic profile, clinical, histopathological and immunohistochemical for each patient and pathology. Given the incidence of cancer disease in society, and although research has traditionally focused on the aspect of diagnosis, it is relatively recent the interest of researchers on the study of prognosis, an integrated aspect into the growing trend of national public health systems towards a model of personalized and predictive medicine.

The prognosis can be defined as prior knowledge of an event before its possible appearance, and can focus on the susceptibility, survival and disease recurrence. In the literature, there are several studies that use neurocomputational models for the prediction of a very specific caseload such as relapse in operable breast cancer, based on prognosis factors of clinical histopathological nature. These studies demonstrated that these models outperform traditional statistical tools used in survival analysis by expert clinicians. However, these models are less efficient when processing information of atypical tumors or subtypes morphologically indistinguishable, for which clinical and histopathological factors do not provide sufficient discriminatory information. The reason is the heterogeneity of cancer as a disease, for which there is no single cause characterized, and whose evolution has been shown to be determined by not only clinical but also genetic factors. Therefore, the integration of clinical and histopathological and genetic profile data provide a more accurate prediction compared to those models that use only one type of data, allowing to carry daily clinical practice a personalized medicine.

In this sense, the expression profiling data from experiments with DNA microarray platforms, data from miRNA microarrays, or more recently the

last generation sequencers such as RNA-Seq, provide the level of detail and complexity necessary to classify atypical tumors setting different prognoses for patients within a notarized group. The analysis of data of this nature represents a real challenge for clinicians, biologists and the rest of the scientific community in general due to the large volume of information produced by these platforms. Usually, the samples resulting from experiments on these platforms are represented by a large number of genes, in the order of thousands. The identification of significant genes that incorporate sufficient discriminatory information and which allows the design of predictive models would be virtually impossible to accomplish without the aid of computers. This is where bioinformatics arises, a term that refers to how information science is applied in the area of biomedicine.

The overall goal that is trying to be achieved within this thesis is, therefore, to offer a personalized medicine in the daily clinical practice. To this end, expression profiling data of some of the most relevant microarray platforms will be used in order to develop predictive models that allow to obtain an improvement in the generalization ability of existing prognosis systems in the clinical area. The overall objective of this thesis can be derived into three partial objectives: the first one will try to (i) pre-process any data in general, and biomedical data in particular, for further analysis. When you go to a public data repository to seek and obtain a data set, it is common to find a complete and valid data set ready for any kind of analysis. However, when working with real data sets it is normal to face with a dataset that contains incomplete samples and thus requires some pre-processing. The development of a system that integrates different standard methods of missing data imputation will make the task of pre-processing a data set easier to any researcher before doing the analysis. A second partial objective will try to (ii) analyze the main gaps in current information systems of an oncology service in order to develop a oncologic information system (OIS) that covers all the needs. Nowadays, these systems are usually very general causing quite reluctant to use it within their own clinical staff. In addition, they do not offer the possibility of using clinical and histopathological information stored in the system to predict the clinical course of each patient. The design and development of a OIS that integrates predictive models based on the information stored in it will help clinicians the diagnosis and determination of the treatment to be applied to a patient. Given the importance of genetic profiles data in the casuistry of cancer, the third partial objective of this thesis will try to (iii) develop new predictive models based on expression profiles obtained from any sequencing platform. The development of these models implies a selection of features or significant genes (commonly known as genetic signature). The genetic signatures found, besides having high predictive capabilities, should possess characteristics such as robustness and biological relevance in the studied disease. The successful development of these models will allow to

integrate them into an OIS together with predictive models based on clinical and histopathological information of every patient. This would give clinicians the ability to perform in daily clinical practice a personalized medicine for each patient, thus providing greater accuracy in diagnosis, prognosis and treatment of the patient.

The pre-processing of a dataset is a necessary precursor to any type of analysis you want to perform on it. In an ideal framework, the dataset that is going to be analyzed will be complete and perfectly valid to work with it. This is usually quite common when datasets are obtained from public repositories that contain numerous data sets from different areas and have already been pre-processed before. However, the initial starting position when working with real data sets is quite different. When the data with which we work are not from public data repositories but data from actual measurements or experimental testing, it is very common to find incomplete data sets that require pre-processing. In clinic, for example, is very common to have information of a group of patients in which certain variables do not contain specific information for a given sample. This problem is well-known in the literature as *missing data* or lost data values. When the analysis also tries to make any type of data classification, then it happens that standard classification algorithms do not behave as well as they are expected to do it.

Usually, a simple and straightforward way of addressing this problem would be to discard those samples in the dataset that contain missing data. Thus, the resulting subset would be complete itself and would be valid to analyze with any classification algorithm. However, when working with biomedical data sets this option is not contemplable. On one hand, this type of data usually has a small number of samples so that if, in addition, those samples that are incomplete are removed, the effectiveness of the classifier will be adversely disrupted. Moreover, data sets based on genetic profiles are characterized by presenting very few samples (on the order of tens) compared to the very high number of characteristics that describe each sample (on the order of thousands), so that the efficacy of the classifier decrease even further. On the other hand , on datasets of this type there is no option to repeat the experiment, either the cost associated with it or because the tissue that was used for the experiment is not already available. Therefore, it arises the need to develop effective methods for processing such data sets in particular, and anyone in general, and imputing or estimating missing data based on the values contained in the remaining samples of the dataset.

In general, different imputation methods already exist and are published in the literature. There still are a number of common problems related to a study of this type that arise when the application of an imputation method is required. First, you must know in depth how each existing imputation method works to, then, be able to deduce which of them is more suitable to apply to the dataset of the study. Once the researcher decides on one

method, the following problem that a researcher may find is where to look for the implementation of the method or some software, quite possibly by paying for a license, that does already incorporate it. Sometimes, if the method is not very complicated the solution passes directly to make our own implementation, with the extra cost of validating the implementation, but this is not always possible and eventually leads the researcher to pay the license of a commercial software to impute missing data in the dataset. Finally, assuming solved the above problems, it could be the case of a small research group that might not have sufficient computational resources to run the chosen imputation method on a dataset and thereby obtain the result within a reasonable time.

To date, there is no free public system known with a friendly user interface that offers the possibility to its users to impute missing data, giving the option of choosing among several imputation methods available, in any dataset and in a reasonable time. Therefore, in Chapter 2 of this thesis the design and development of a web system called WIMP (Web imputation) is described. This system is a free public web tool available to be used by users registered in the system. WIMP provides the research community the ability to impute missing data in the datasets that each user upload into the system by choosing among multiple imputation methods that are incorporated into it in its original version. These methods are already implemented and have been previously tested, thus the user should not worry about finding the implementation of any imputation method or pay for a commercial software that offers this possibility. Among the existing imputation methods, WIMP incorporates some based on statistical techniques (mean imputation, hot-deck imputation and multiple imputation) and others based on machine learning algorithms (imputation using Self-Organizing Maps and using a Multilayer Perceptron). The number of methods incorporated into the system in its initial version is not closed, but may be expanded in the future just implementing other existing methods, validating and incorporating them into the kernel of the system.

On the other hand, another great advantage offered in WIMP to its users is to spare them the need for a large capacity of computational resources. The imputation tasks that users request through the system are processed on the server side. Moreover, the server is connected to a computing cluster where tasks requested by users of the system are executed. This architecture allows a single user, from a personal computer, executing a complex imputation method and quickly obtain the entire data set after executing the imputation process. Once a task launched by a user is finished, the server notifies the user via e-mail and provides a download link that allows them to collect the resulting data set from the execution. Now, this dataset is composed by complete samples and is totally available for any kind of analysis, thus reaching the first partial objective of this thesis.

Then, in Chapter 3 the second partial objective of this thesis is addressed by analyzing the major gaps in current information systems of an oncology service in order to develop a OIS that covers all the needs and allows using predictive models based on clinical and histopathological information of each patient. Specifically, the research group "Inteligencia Computacional en Biomedicina (ICB)" of the University of Málaga works closely with clinicians from the oncology service at the "Hospital Universitario Virgen de la Victoria (HUVV)" in Málaga, Spain. This close relationship allows detailed analysis of the needs in the oncology service of a hospital in order to successfully implement a system that incorporates both the clinical and histopathological information of each patient and predictive models that provide more information to clinicians in order to decide about the diagnosis, prognosis and treatment to be applied to a patient.

This chapter of the thesis shows in detail all the experience gained in the design and implementation of an OIS for the oncology service of the HUVV. After a thorough analysis, common problems are identified and solutions to the most critical aspects that facilitate success in implementing the system are provided. The first aspect to consider is the usability of the system . A user-centered design will facilitate the active involvement of end users of the system throughout its development, ensuring from the beginning an enjoyable user experience and a simple interaction with the system, avoiding possible future declines to use the system. Another aspect to consider is the use of appropriate technology to make the development, maintenance and future system upgrades easier, because there is no worth in using the latest technology available in the community if it does not guarantee the development of a system that allows achieving a better medicine. It is also important that this type of system fully integrates all daily clinical oncology service routines and for that, nothing better to particularize the design and development depending on the service where the system is going to be deployed. General solutions that seek to cover the deployment of the system in many hospitals will make all the requirements and specific needs of oncology service not to be caught causing reluctance to use the developed system. System security (ensuring the privacy of the data contained on it) and its interaction with other information systems of the hospital are other aspects to be considered. An isolated system with lots of information is not valuable due to the high specialization of the system. To complement the information contained in an oncology information system, the interaction with other systems such as radiology, hematology, pharmacy, etc, is a must. Finally, a key aspect that makes the difference in these systems is the ability to enable, within the system, the use the information stored to do some kind of study with no need of going to another third party software.

The final system developed and successfully deployed in the HUVV is a web application with a layered architecture that incorporates all the clinical

and histopathological information of each patient. The layered architecture of the system seeks greater usability, ease of maintenance and future system expansion. Its design includes the distribution of information in the following functional modules: patient management, treatment outpatient unit, clinical research, genetic counseling and statistical analysis. The most important feature that makes this system stand out from others is that, in addition to storing all information relating to patients of the oncology service, it contains a specific module for doing several studies based on that information. This module integrates several models of clinical prognosis as a result of previous works done by the ICB research group, thus giving clinicians the possibility of estimating survival analysis implemented under the Kaplan-Meier algorithm, Cox regression analysis, calculate risk functions and obtain contingency tables based on the clinical an histopathological information of a subset of patients previously selected. The evaluation of the successful deployment of the system was conducted from surveys done by end users at 3 and 15 months after deploying the system in the hospital, showing widespread satisfaction.

Despite a successful development and deployment of the OIS in the HUVV, in its current version it is only available clinical and histopathological data of each patient. In this sense, this system has been a major step forward towards in order to lead in the daily clinical practice a personalized medicine. However, it was previously noted that the heterogeneity of cancer as a disease makes its evolution to be determined by not only clinical but also genetic factors, which leads to consider the need to integrate in a near future the genetic profiles of each patient within system. Furthermore, it arises the need to develop new predictive models based on expression profiling data in order to incorporate them into the existing system. Thus, this would allow clinicians to practice a personalized medicine to their patients according to both types of data and to predictive models integrated in the statistical analysis module.

In this sense, a new and parallel research line was opened in order to analyze datasets of gene expression profiles. This kind of data results in different types of analysis, each large enough to investigate. Overall,three main types of analysis could be distinguished: (i) identify the class associated to each sample, (ii) predict the class associated to a new sample and (iii) compare samples of the same class. The first type of analysis has sense when the class or label associated to each sample of the dataset is unknown. This type of analysis tries to split samples into classes or groups of patterns that are hidden within the data. In the second type of analysis, the class or label associated to each sample of the dataset is a priori known. In this case, the challenge is to try to predict the class of new samples from knowledge or hidden patterns from the initial dataset, being this one precisely the research line that will follow in Chapters 4-6 to reach the third partial objective set

in this thesis. Finally, in the third type of analysis classes are also known a priori, although now the aim is to discover genes and expression patterns that allow better differentiation between groups by comparing samples of different class.

In other data sets closer to areas of Artificial Intelligence (AI), the development of predictive models is a task that is widely studied and solved in the literature, commonly known as *machine learning*. Therefore, one would expect that the development of predictive models based on expression profiling data should not be very difficult to achieve. However, the process of classification or prediction based on genetic profiles is not a trivial task due to the nature of the data. These data sets have a very high dimensionality, on the order of thousands to tens of thousands of genes per sample and only a few dozen or a hundred samples. This causes an overfitting problem when using any classification algorithm, since they will try to classify every sample in the dataset (even possible noisy samples) and thus producing a low generalization in future data. This problem is well known in the literature as *curse of dimensionality*. Consequently, this type of analysis comprises an implicit feature selection process which identifies the most significant features with most discriminatory information between classes. Therefore, the development of a predictive model based on expression profile data starts with the original dataset and performs a search of a subset of genes, known as genetic signature, whose expression patterns are used to differentiate existing groups in the analysis. This feature selection process is organized into three categories depending on how is is combined the search of feature selection and the construction of the classification model: filter methods, wrapper methods and embedded methods.

Appendix A shows the work that raises as the initial starting point in this research line. It describes the use of a constructive neural network (CNN) as an alternative classifier to traditional artificial neural networks (ANN). At the same time, a wrapper feature selection method widely used in the literature, called Stepwise Forward Selection (SFS) procedure, is chosen as feature selection strategy. The workflow of this method is simple. In its first iteration, the method evaluates each prediction model built by using a single gene, choose the model that provides higher prediction and advances to the next iteration, which will then study models built by two genes (the one already selected combined with one more gene of the remaining genes from the dataset). The method will continue going further on iterations until the model that is being evaluated and is built using a number of genes X does not improve the prediction of the best model found in the previous iteration (built using $X-1$ genes). The results obtained by the SFS+CNN strategy are competitive in comparison to those obtained by the traditional SFS+ANN strategy used in the literature to identify genetic signatures with high predictive capabilities. In addition, the added advantage of using a CNN is that

the architecture of this kind of neural network is built dynamically during the learning phase when they are trying to learn on each iteration of SFS method, whereas an ANN model should prefix an initial architecture called "optimal" that will be used on all the iterations. But how can we ensure that this architecture is the optimal one for problems of different sizes (one, two, three genes, etc.) that are going to be evaluated in the SFS method depending on the iteration that is executing the algorithm?

The next step led to believe in an approach that tries to match on one hand the advantages of wrapper methods such as SFS+CNN (very good prediction results) and on the other hand the main advantages of filter methods (less complex and much faster in execution). This need led to propose an approach of a hybrid model which brought both benefits and is described in detail in Appendix B. This method is based on the prediction cability of each individual gene in the data model using a constructive neural network, C- named Mantec, together with the use of the linear correlation coefficient between two variables. The developed method proposes three feature selection strategies: first, *ROnly* , is solely based on the generalization by directly choosing 10 genes with greater individual predictability; the second one, *RelevanceF*, combines generalization and redundancy, so that from the 10 % of the genes with higher individual predictability, finally the 10 genes presenting a lower redundancy are chosen; and the third one, *RedundancyF*, is just the opposite to the previous one, thus from the 10 % of genes with less redundancy, the 10 genes presenting greater individual predictability are selected. A comparison of these three strategies was conducted against the results obtained using the SFS+CNN method where it can be deducted that the *RedundancyF* strategy obtains very similar results to SFS+CNN in terms of prediction. However, the requirements in terms of computation complexity and resources are much lower in the proposed strategy *RedundancyF* and therefore, the time required to estimate the genetic signature is considerably lower in this hybrid model.

After this preliminary work mainly oriented on the development of new predictive models based on expression profiling data, an extensive review of literature was conducted in order to identify feature selection methods used in these types of datasets. In this sense, another family of algorithms widely used to identify genetic signatures with high predictive capabilities are genetic algorithms (GA). This algorithm is also a wrapper method that in each generation of the algorithm a number of individuals (or subset of genes) are assessed using a classification algorithm. The goal of a GA is to minimize a function, where the key to this kind of algorithms is known as *fitness function* that is a mathematical function that is responsible for evaluating the quality of each individual within the population of the study. Generally, previous published works that use a GA to estimate a predictive model tend to define this function seeking to minimize the classification error (which is equiva-

lent to maximizing the rate prediction) while compounds subsets of genes are prioritized by fewer genes. Based on this assessment of the population, the best individuals of the population directly passed to the next generation while the remaining individuals evolve to new individuals using selection and crossover operands, simulating the biological reproduction process.

In this this sense, Chapter 4 proposes the use of a GA using C-Mantec as classification algorithm for estimating predictive models with data from expression profiles. Apart from the use of a constructive neural network classifier within the feature selection method, the developed GA mainly includes two new features not covered previously in the literature : (i) it performs a pre-filtering of genes for further analysis with the GA+CMantec method, and (ii) it defines a fitness function of the GA by including information of the correlation between genes in the subset that is being evaluated. Since GAs are a wrapper method and looking for a lower need of required computing resources by these methods, this proposal includes a first novelty in the GA+CMantec strategy doing a pre-filter on the total set of genes using the Welch's statistical t-test before starting the rest of the analysis. This statistical test assigns a value to each gene, known as *p-value*, which indicates how important is that gene to discriminate two classes or groups in the dataset. As a result of this test, only the 5 % of the total amount of genes are retained, thus being this subset of genes the one that will be analyzed with the GA+CMantec strategy proposed. On the other hand, the second novelty of the proposed GA is localized in the definition of the fitness function that is composed by three terms. The first one seeks to minimize the classification error of the C-Mantec algorithm (subsets of genes with better prediction rate are preferred). The second term aims to minimize the number of genes that compose the genetic signature found by the algorithm (solutions with the lowest possible number of genes are preferred). Finally, as a novelty with respect to previous published works in the literature, a third term that seeks to minimize the correlation between variables (those subsets of genes that present redundant information are penalized) is included . Moreover, the three terms that form the basis of the fitness function are weighted by a constant rate that gives more or less importance to each term in order to guide the search towards solutions that meet certain characteristics.

In order to validate the predictive models obtained from the GA+CMantec strategy, the results where compared to the ones obtained using the SFS+CNN strategy proposed in the preliminary work but including in this previous strategy the pre-filtering of genes by applying the Welch's statistical t-test. Furthermore, apart from the C-Mantec algorithm other widely known classification algorithms were used such as Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), Bayesian Networks (NB), k -Nearest Neighbors (kNN) and a Multilayer Perceptron (MLP). All these strategies were tested using DNA microarray datasets of different types of cancer. The results show

in general better prediction results using the GA strategy than using the SFS method. In contrast, the GA strategy obtains as a result subsets of larger genes and is computationally more demanding than SFS as it evaluates much more subsets of genes. Regarding the influence of the classification algorithm on the feature selection strategy, MLP, LDA and SVM lead to better prediction results while C-Mantec and kNN are close behind them with a slightly lower prediction. Furthermore, C-Mantec, MLP and LDA find genetic signatures with fewer genes in comparison to SVM, NB and kNN, and particularly C-Mantec is the most robust classifier in terms of adjusting the parameters of the model.

At this point, it was necessary to analyze in more detail the three most complex tasks that need to be addressed when estimating predictive models using data from expression profiles: feature selection, classification model and validation strategy. Reviewing the literature, one can find works that focus their efforts on solving, individually, problems that arise in any of these tasks. For example, some focus on identifying the best feature selection method without paying attention to the validation strategy used. Other authors focus their attention solely on the advantages of using a certain classification algorithm regardless of feature selection method used. However, to date no work that encompasses all these tasks was found thus offering an overview of the whole problem on a single work.

Chapter 5 presents an extensive comparison between different existing feature selection methods using several standard classification algorithms. The feature selection methods considered in this study are the SFS and GA methods previously described (both are wrapper methods), the Correlation-based Feature Selection (CFS), Consistency-based Filter (Cons), Information Gain (IG), Minimun-Redundancy Maximum-Relevance (mRMR) and ReliefF (all of them are filter methods) and SVM-Recursive Feature Elimination (as an embedded method). Once it is chosen one of these feature selection methods, several standard classification algorithms were used such as LDA, SVM, kNN, NB, C-Mantec and MLP. In addition, to evaluate the different strategies that arise by combining the feature selection methods and the classification algorithms, a honest validation scheme was applied. In this sense and since the number of samples available is very low, it is recommended to use methodologies that use *resampling* techniques. Therefore, the strategy that is proposed to be used for this comparison is a Bootstrap-CrossValidation (BCV) strategy to obtain the rate prediction of external validation as well as it is also highly recommended when there are few samples in the analysis, as occurs in our case. The results obtained after executing each strategy under the same validation scheme described show that filter methods or embedded provide greater predictability being less complex and faster running than wrapper methods. Furthermore, the robustness of the obtained genetic signatures depending on the feature selection method

was analyzed, and again filter methods appear generally as the most robust methods.

Robustness and biological relevance of genetic signatures, along with its predictive cability, are other very desirable characteristics. The main drawback of these feature selection methods, regardless of the applied strategy, is the limited biological relevance and robustness offered by the final solutions, ie, each execution of the method produces a different genetic signature. These solutions are useless if the final goal would try to use these gene signatures in clinical or biomedical studies. Then, the idea would try to provide these feature selection methods with a mechanism that incorporates biological information of the studied disease within the process. If this is reached, the feature selection method may promote the most relevant genes from a clinical and biological point of view to become part of the final solution and, in contrast, discard those genes that have little or no influence on the studied disease. Thus, besides having more guarantees about the biological relevance of the genetic signatures, one would expect that the robustness of these signatures do also increase because the search is restricted and guided to solutions whose characteristics are more restrictive than in the traditional approaches.

Chapter 6 proposes a new two-stage feature selection method in order to obtain robust genetic signatures with high predictive capabilities and biological relevance in the studied disorder. This new method uses biological information associated with the disease under study. In this sense, the information available in the database Kyoto Encyclopedia of Genes and Genomes (KEGG) is used, since it is considered a database of reference widely used in recent years as an important source for the modeling of pathways to decode the genome and link it to biological systems. This database is based on the definition of a manual KEGG code to a particular sequence to has been involved in experimental testing. If a protein encoded by a sequence produces no enzymatic activity or is not part of a pathway, then it will never have an associated KEGG code. Fortunately, most of the genes involved in cancer do have enzymatic activity and belong to a pathway.

The two stages of the new model proposed in this thesis consist of the: (i) the first stage makes an initial filtering of the total number of genes based on the KEGG database by retaining only those genes that represent an enzyme. It also establishes an ordered ranking of all pathways associated with studied disease according to the predictive ability of the genes that are part of each pathway and to the number of keywords (previously identified by an expert in the area) found in a data mining process on the textual description of each pathway offered on the KEGG website. Then, (ii) the second stage runs a GA on the best pathways to estimate a predictive model. The main novelty of this GA respect to one previously explained is located in the fitness function. Once again, this function is composed of three terms. The first two are

identical to the previous GA (minimizing the classification error and minimizing the number of genes that comprise the genetic signature). However, now the third term tries to maximize a function that scores the "quality" of the genetic signature based on the number of genes that it contains and are also part of the pathway of study or in any of the other selected pathways in the first stage for further analyze. Similarly as the previous proposed GA, each of these terms is weighted by a constant rate that allows to guide the search solutions to meet certain characteristics. After evaluating this new method in three sets of DNA microarray data from different types of cancer, the obtained results show that this new strategy improves consistency, robustness and prediction of genetic signatures. This new approach offers the possibility to define genetic signatures that could be used in a near future to predict the diagnosis and clinical outcome of patients with cancer.

Finally, it can be concluded that the results obtained in this thesis represents an advance to achieve in a near future the ability of offering a personalized medicine in the daily clinical practice. Predictive models based on expression profiling data that have been developed in this thesis could be integrated into the OIS deployed in the HUVV. Furthermore, it could incorporate genetic profiles of each patient to maximize the added benefits mentioned throughout this thesis. Moreover, thanks to all the work done in this thesis, the PhD student was able to deepen and acquire an extensive research background in such a wide area such as Bioinformatics.

Índice

Capítulo 1

Introducción

1.1. Motivación

La predicción de la evolución clínica de un paciente con cáncer de mama constituye un aspecto fundamental en la medicina oncológica personalizada. Por ejemplo, en el año 2010, la Sociedad Americana del Cáncer (ACS) preveía que para ese año habría 207.090 nuevos casos de cáncer de mama en las mujeres y 1.970 en los hombres. De éstos, se estimó que más de 40.000 fallecerían, lo cual supone más del 19 % de los pacientes con cáncer de mama diagnosticados. En España, se diagnostican cada año 15.000 nuevos casos de cáncer de mama, y actualmente ya se manejan cifras que indican que una de cada 16-18 españolas tendrá un cáncer de mama a lo largo de su vida. Hoy día, el cáncer de mama en particular y cualquier tipo de cáncer en general sigue siendo una enfermedad con una respuesta a los tratamientos y unos resultados a menudo inciertos.

En los últimos años, el avance en la tecnología y en la ciencia ha permitido profundizar en las bases genéticas y moleculares de enfermedades como el cáncer, y analizar la variabilidad de respuesta de pacientes individuales a diferentes tratamientos oncológicos, estableciendo las bases de lo que hoy se conoce como medicina personalizada. El objetivo principal de ésta persigue diseñar y aplicar estrategias de prevención, diagnóstico y tratamiento adaptadas a un escenario que integra la información del perfil genético, clínico, histopatológico e inmuhistoquímico de cada paciente y patología. En este sentido, existen trabajos que utilizan modelos predictivos basados en factores pronóstico de naturaleza clínico-histopatológica con muy buenas tasas de acierto. No obstante, estos modelos pierden eficacia cuando procesan información de tumores atípicos o subtipos morfológicamente indistinguibles, para los que los factores clínicos e histopatológicos no proporcionan suficiente información discriminatoria. Esto se debe, fundamentalmente, a la heterogeneidad del cáncer como enfermedad, para la que no existe una causa individual caracterizada, y cuya evolución se ha demostrado que está

determinada por factores no sólo clínicos sino también genéticos.

Aquí surge la necesidad de integrar los datos clínico-histopatológicos junto a la información proteómico-genómica en los modelos predictivos, siendo de esperar que el uso de ambos tipos de datos proporcionen una mayor precisión en comparación con aquellos modelos predictivos que sólo utilicen uno de ellos. Debido a que las células tumorales son genéticamente inestables, es probable que un subconjunto reducido de genes sea capaz de satisfacer los requisitos necesarios para la invasión y metástasis. De esta forma, la detección o identificación de perfiles genéticos indicativos de agresividad tumoral se considera de vital importancia para el diagnóstico temprano de la enfermedad, el pronóstico y la predicción de la respuesta al tratamiento que se le aplique. Para ello, existen diversas plataformas que aportan un gran volumen de datos de perfiles de expresión. En concreto, la tecnología de microarrays de ADN ha revolucionado la forma de analizar las expresiones genéticas de los pacientes, ofreciendo la posibilidad de proyectar y analizar simultáneamente cientos de miles de genes que se encuentran representados en base a la intensidad de luz que se expresa en una matriz.

Los repositorios públicos de bases de datos de perfiles de expresión son una gran fuente de datos para poder realizar cualquier tipo de análisis en general y, en particular, para estimar modelos predictivos basados en información proteómico-genómica. Estos datos están, en su mayoría, pre-procesados y listos para analizar. Sin embargo, al trabajar con datos reales es común encontrar que la matriz de datos presenta algunos valores que faltan en una o varias muestras, lo que se conoce como valores perdidos. Los valores perdidos se producen por diversas razones, como pueden ser una resolución insuficiente, la corrupción de la imagen, o simplemente debido al polvo o arañazos en la diapositiva. En otras áreas, una posible solución podría pasar por descartar aquellas muestras que presenten valores perdidos. Sin embargo, dado el número tan escaso de muestras del que se dispone en este tipo de problemas, esta opción resulta ser inviable. Por otro lado, también se podría repetir el experimento para solución el problema, aunque obviamente esto aumentaría los costos inevitablemente y no siempre será posible obtener de nuevo la muestra para llevar a cabo el experimento. Por tanto, surge la necesidad de pre-procesar los datos antes de cualquier otro tipo de análisis y aplicar técnicas conocidas de imputación de datos perdidos. De esta forma, a los valores perdidos se le asignarían los valores resultantes de aplicar estas técnicas y estaríamos paliando los efectos negativos de los anteriores casos.

El análisis de datos de perfiles de expresión para la estimación de modelos predictivos es, dado el gran volumen de información, una tarea muy compleja que presenta diversos problemas. Por lo general, los modelos de Inteligencia Artificial (IA) funcionan muy bien cuando se tiene un número suficientemente grande de muestras N y cada muestra se encuentra descrita por un número no demasiado grande de atributos M (esto es $N \gg M$). Sin

embargo, en los conjuntos de datos de perfiles de expresión ocurre justamente lo contrario. Este motivo hace que los algoritmos de predicción tradicionales de la IA no se comporten todo lo bien que de ellos se espera, siendo necesario realizar una búsqueda e identificar un subconjunto de genes significativo, conocido como firma genética, que maximicen la precisión alcanzada por el modelo predictivo. A pesar de toda la variedad de técnicas de selección de características existentes en la literatura, la selección de una firma genética con altas capacidades predictivas, robusta e informativa desde el punto de vista clínico y biológico sigue siendo un problema abierto en la literatura. A menudo, los trabajos publicados se centran en estimar firmas genéticas con altas capacidades predictivas, obviando la importancia que tiene la robustez y la relevancia biológica de la firma de cara a implantarla en un chip y utilizarla en la práctica clínica diaria. Por ello, es necesario prestar atención también a estas características para lograr desarrollar modelos predictivos que puedan ser utilizados por los clínicos y facilitándoles el diagnóstico, la evolución y la determinación del tratamiento a aplicar a un paciente.

1.2. La naturaleza de la información biomédica

Los avances científicos y tecnológicos registrados en los últimos quince años han permitido profundizar en las bases genéticas y moleculares de enfermedades como el cáncer y analizar la variabilidad de respuesta de pacientes individuales a diferentes tratamientos oncológicos, estableciendo las bases de lo que hoy se conoce como medicina personalizada. Los datos de perfiles de expresión provenientes de experimentos con plataformas de microarrays de ADN, los datos de microarrays de miRNA, o más recientemente secuenciadores de última generación como RNA-Seq, proporcionan el nivel de detalle y complejidad necesarios para clasificar tumores atípicos estableciendo diferentes pronósticos para pacientes dentro de un mismo grupo protocolizado. Esto se debe a la heterogeneidad del cáncer como enfermedad, para la que no existe una causa individual caracterizada, la cual hace que la evolución de un paciente esté determinada por factores no sólo clínicos sino también genéticos.

La aplicación de modelos predictivos basados en información clínico-histopatológica de cada paciente es algo que ya se logró alcanzar con muy buenos resultados en algunos trabajos existentes en la literatura [Gómez-Ruiz et al. (2004); Jerez et al. (2005); Jerez-Aragonés et al. (2003); Hadjianastassiou et al. (2006)]. No obstante, si el objetivo es tratar de llevar a la práctica clínica diaria una medicina personalizada, es necesario desarrollar modelos predictivos basados en datos de perfiles de expresión. Por lo general, las muestras resultantes de los experimentos en plataformas de secuenciación están representadas por un número muy elevado de genes, del orden de miles o decenas de miles de ellos. El análisis de un volumen tan

grande de información supone un reto casi insalvable para clínicos y biólogos. La Bioinformática, término que hace referencia a cómo se aplica la ciencia de la información en el área de la biomedicina, viene a paliar las dificultades que clínicos y biólogos encuentran al tratar de analizar datos de perfiles de expresión.

En la última década, la plataforma predominante en la investigación y en clínica para el análisis de datos de perfiles de expresión ha sido la tecnología de microarrays de ADN, aunque en los últimos años aumentó también el número de trabajos de investigación utilizando microarrays de miRNA o la plataforma RNA-Seq. Además, los microarrays de ADN son, hoy día, bastante utilizados en el ámbito clínico pues a pesar de que la tecnología de ultrasecuenciación es claramente el futuro, los microarrays de ADN siguen teniendo un coste mucho más bajo en comparación con esta nueva tecnología. La tecnología de microarrays de ADN surge durante la segunda mitad de la década de 1990, fruto de los informes de unos experimentos iniciales publicados a mediados de la década de 1970 que indicaban que los ácidos nucleicos marcados se podrían utilizar para controlar la expresión de moléculas de ácidos nucleicos unidas a un soporte sólido. Sin embargo, no fue hasta 1995 cuando aparece el primer artículo describiendo la aplicación de la tecnología de microarrays de ADN para el análisis de expresión, publicada por Patrick Brown y sus colegas en la Universidad de Stanford [Shena et al. (1995)].

1.2.1. Microarrays de ADN

Un microarray de ADN, o comúnmente conocido como chip de ADN, es una herramienta analítica que consta de muchos fragmentos de ADN densamente colocados sobre un sustrato de resina o vidrio para detectar los cambios de expresión en los genes que se encuentran en las muestras. La tecnología de microarrays es una plataforma potente que ha sido utilizada con éxito en una variedad de estudios experimentales en diversas áreas de investigación [Hero et al. (2004); Friedland et al. (2006); Stangegaard (2009); Wang y Simon (2011); Houseman et al. (2012); Elhamifar y Vidal (2013)], ya que permiten un análisis exhaustivo de varios cientos a varios cientos de miles de genes con una pequeña cantidad de muestra y en un tiempo relativamente corto. En los últimos años, esta tecnología ha permitido a la comunidad científica entender los aspectos fundamentales que subrayan el crecimiento y desarrollo de la vida, así como explorar las causas genéticas de las anomalías que se producen en el funcionamiento del cuerpo humano.

Un chip de ADN consiste en una estructura ordenada de miles de genes secuenciados, identificados e impresos sobre un soporte sólido impermeable, generalmente de vidrio, chips de silicio o membrana de nylon (ver Figura 1.1). Cada uno de los genes secuenciados corresponde a un fragmento de ADN genómico y representa un único gen. Por lo general, un microarray de

ADN contiene miles de puntos, donde cada uno de ellos de manera individual representa a un solo gen y, colectivamente, a todo el genoma de un organismo.

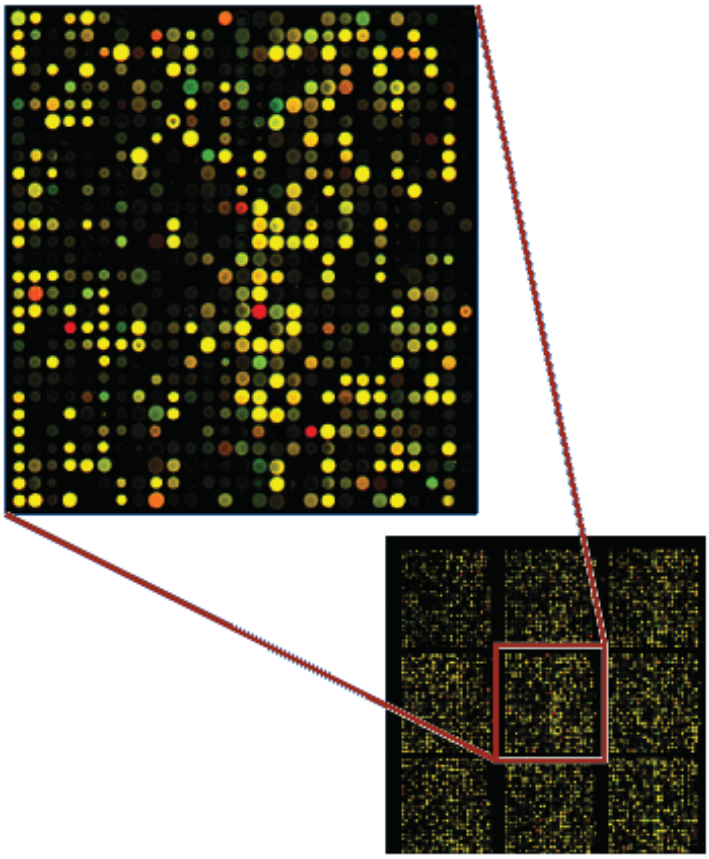


Figura 1.1: Imagen de un microarray de ADN con el nivel de expresión de los genes reflejado por la intensidad luminosa reflejada.

La gama de aplicaciones de la tecnología de microarrays es enorme. De entre todas ellas cabría destacar las siguientes:

- *Descubrimiento de genes.* La tecnología de microarrays de ADN ayuda a identificar nuevos genes, saber acerca de su funcionamiento y conocer sus niveles de expresión en diferentes condiciones.
- *Diagnóstico de enfermedades.* Esta tecnología ayuda a los investigadores a aprender más acerca de las diferentes enfermedades, tales como las enfermedades del corazón, enfermedades mentales, enfermedades infecciosas y especialmente a estudiar del cáncer. Hasta hace poco, los diferentes tipos de cáncer se han clasificado a partir de los órganos en los que se desarrollan los tumores. Ahora, con la evolución de la tecnología de microarrays, se abre la posibilidad para que los investigadores

puedan clasificar aún más los tipos de cáncer en base a los patrones de actividad génica en las células tumorales. Esto ayudará enormemente a la comunidad farmacéutica a desarrollar medicamentos más eficaces al mismo tiempo que permitirá definir mejores estrategias de tratamiento dirigidas directamente al tipo específico de cáncer que padezca el paciente.

- *Descubrimiento de medicamentos.* La tecnología de microarrays tiene una amplia aplicación en la farmacogenómica. La farmacogenómica es el estudio de las correlaciones entre las respuestas terapéuticas a los medicamentos y los perfiles genéticos de los pacientes. Los análisis comparativos de los genes de una célula enferma y una célula normal ayudará a la identificación de la constitución bioquímica de las proteínas sintetizadas por los genes enfermos. De esta forma, los investigadores pueden utilizar esta información para sintetizar medicamentos que combatan con estas proteínas y reducir así su efecto.
- *Investigación toxicológica.* Esta tecnología proporciona una plataforma sólida para la investigación de los efectos de las toxinas en las células y su transmisión a los descendientes. La toxicogenómica establece la correlación entre las respuestas a las sustancias tóxicas y los cambios en los perfiles genéticos de las células expuestas a dichas sustancias tóxicas.

La Figura 1.2 muestras las típicas fases presentes en la tecnología de microarrays. El punto de partida es el cuestionamiento de algún proceso biológico/biomédico que guarde relación con algún hecho observable. A partir de aquí, se diseña y se lleva a cabo un experimento de microarray dando lugar a una imagen de características similares a la mostrada en la Figura 1.1. El análisis y pre-procesado de esta imagen es un área amplia de investigación cuyo objetivo principal es lograr una matriz de datos limpia de ruido y completa para su posterior análisis. A continuación, se pretende aplicar diferentes técnicas de análisis de datos para extraer conocimiento que permita verificar las cuestiones planteadas como hipótesis de partida en los diferentes estudios.

1.3. Estado del arte

La tecnología de microarrays de ADN ha aumentado, en gran medida, la velocidad a la que podemos generar datos sobre los sistemas biológicos, lo cual permite por primera vez poder observar a nivel molecular la respuesta global de un organismo a un estímulo particular. Debido a ello, se incrementaron sustancialmente los desafíos asociados con la recolección, gestión y análisis de los datos de cada experimento. Existen diversos problemas asociados al análisis de datos de microarrays según el área de aplicación en el

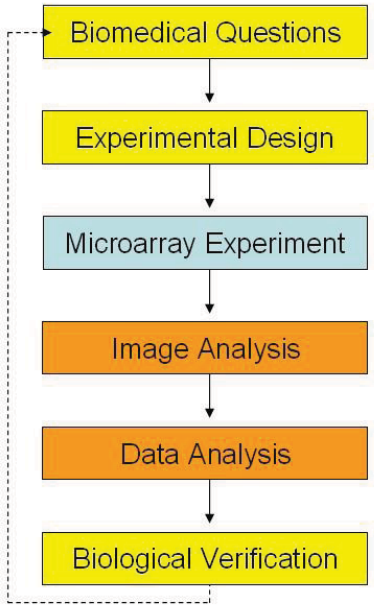


Figura 1.2: Etapas presentes en la tecnología de microarrays (fuente en http://www.dmi.unict.it)

que nos movamos, siendo los más destacados: (i) compresión de imágenes, (ii) recogida, transformación y representación de datos, (iii) descubrimiento de la clase asociada a la muestra, (iv) predicción de la clase de nuevas muestras y, (v) comparación de muestras de la misma clase.

1.3.1. Compresión de imágenes

Los experimentos de microarrays de ADN dan como resultado unas imágenes de las que se extrae información acerca de la intensidad de luz o nivel de expresión del gen. Estas imágenes se analizan con diferentes técnicas para posteriormente tener expresada toda la información genética en un ordenador. Sin embargo, mantener sólo esta información genética y desechar las imágenes de microarrays no es algo deseable debido a que, desafortunadamente, las técnicas de análisis de estas imágenes no están universalmente aceptadas [Allison et al. (2006)]. Si las técnicas de análisis cambian con el tiempo, será muy deseable volver a analizar las imágenes para obtener datos genéticos más precisos. Sin embargo, en tales casos, la repetición de todo el experimento podrá no ser una opción dado que las muestras biológicas necesarias para el mismo ya no estén disponibles. Por ello, es importante almacenar las imágenes de microarrays de ADN junto con los datos genéticos extraídos. En la tarea de almacenamiento de estas imágenes, resulta esencial

ejecutar un proceso de compresión de imágenes que típicamente comprende hasta 5 etapas: preprocesamiento, transformación, cuantificación, codificación y posprocesamiento.

La etapa de preprocesamiento comprende cualquier cálculo realizado en una imagen para prepararla para los procesos de compresión o de análisis, como es la eliminación de ruido [Adjeroh et al. (2006); Lukac et al. (2005); Smolka y Plataniotis (2005); Chen y Duan (2007); Zifan et al. (2010)] y la segmentación [Faramarzpour y Shirani (2004); Hua et al. (2003); Chen et al. (1997); Bierman et al. (2006); Battiato et al. (2008); Karimi et al. (2010); Uslan y Bucak (2010); Li y Weng (2011)]. La etapa de transformación consiste en cambiar el dominio de la imagen a otro donde puede ser procesado o codificado de manera más eficiente, como por ejemplo al dominio de la frecuencia. Hay pocos artículos que aporten un trabajo novedoso en esta área [Peters et al. (2007); Zifan et al. (2010); Avanaki et al. (2011)]. La cuantificación trata de dividir en grupos los valores representados en la imagen, reduciendo el número total de símbolos necesarios para representarla imagen y aumentando, por tanto, la capacidad de compresión [Jornsten et al. (2003); Jornsten y Yu (2000); Peters et al. (2007)]. En la etapa de codificación de la imagen, el objetivo principal es generar un flujo de bits más compactos a partir de los datos obtenidos en las etapas anteriores, bien por segmentación [Faramarzpour y Shirani (2004); Bierman et al. (2006); Battiato y Rundo (2009)] o bien tratando de predecir la intensidad de los píxeles más cercanos basándose en los píxeles vecinos [Zhang et al. (2005); Zhang y Adjeroh (2008); Neves y Pinho (2009)]. Por último, las imágenes resultantes de las etapas anteriores deberían procesarse para mejorar su calidad visual o proporcionar nuevas características aplicando indicadores de calidad específicos para las imágenes de microarrays [Wang et al. (2001); Sauer et al. (2005); Kim et al. (2005)].

1.3.2. Recogida, transformación y representación de datos

Dada una imagen de microarray de ADN, e independientemente de la plataforma utilizada para realizar los experimentos (Affymetrix, cDNA, Agilen arrays, etc.), los datos que se utilizarán en posteriores análisis son los niveles de expresión de cada gen en cada experimento. Estos datos suelen representarse como una matriz en la que cada fila representa un gen particular y cada columna representa una muestra biológica específica. Es decir, cada fila puede verse como un vector de expresión génica y cada columna como un vector de expresión de la muestra que registra la expresión de todos los genes en esa muestra.

Los datos que componen la matriz suelen normalizarse para facilitar la comparación entre ensayos compensando posibles diferencias en el etiquetado, hibridación y detección de eficiencias. Existe un buen número de enfoques propuestos en la literatura orientados a la normalización de datos de micro-

arrays, aunque estos dependen de la plataforma utilizada para los ensayos y de las hipótesis formuladas en base al sesgo que pueda darse en los datos [Irizarry et al. (2003); Quackenbush (2002); Schadt et al. (2001); Yang et al. (2002a,b)]. Además, también es común aplicar técnicas de filtrado basadas en métodos estadísticos que, por ejemplo, elimine aquellos los genes que tienen una varianza mínima o aquellos que no presentan información en la mayoría de los experimentos. El valor que añaden la aplicación de estas técnicas de filtrado reside en la reducción de la complejidad del conjunto de datos.

Aunque esta etapa resulta imprescindible de llevar a cabo como paso previo al análisis de los datos, cabe destacar que tanto la normalización como el filtrado pueden tener un efecto considerable en la matriz de datos resultante [Hoffmann et al. (2002)]. Esto se debe a que la normalización ajusta la intensidad de luz en cada ensayo pudiendo, por tanto, cambiar la diferencia relativa observada entre las diferentes muestras. Y de igual forma, las técnicas de filtrado de los datos puede producir resultados muy diferentes ya que todas las técnicas estadísticas que se aplican se basan en suposiciones acerca de la variación existente en las diferentes mediciones.

1.3.3. Descubrimiento de la clase asociada a una muestra

El análisis y descubrimiento de clases asociadas a un experimento, de forma que se identifiquen grupos presentes en los datos, es otra área principal asociada al estudio de microarrays de ADN. Por ejemplo, se podría examinar un grupo de pacientes de cáncer para ver si en base a sus perfiles de expresión, éstos podrían ser colocados en grupos distintos sin necesidad de disponer de ningún conocimiento previo como la progresión de la enfermedad, el estado sano o tumoral de la muestra, o la respuesta al tratamiento. Una vez identificados los grupos, el objetivo pasa a ser la búsqueda de alguna conexión a un factor clínico o biológico que explique la diferencia.

Este tipo de análisis que buscan separar las diferentes muestras en grupos se basan en métodos de análisis de datos no supervisados o clustering, permitiendo ambos métodos explorar patrones de expresión existentes en los datos. La pregunta que se lanza en este tipo de experimentos es: "¿hay algunos patrones ocultos en los datos que resulten interesante desde el punto de vista biológico?". Los métodos no supervisados no utilizan la etiqueta o clase asociada a la muestra como entrada (por ejemplo, no tienen en cuenta si las muestras son de un paciente sano o de un paciente con algún tumor). La idea principal de estos métodos consiste en agrupar las muestras acorde a alguna medida de similitud. Dos de los métodos no supervisados más utilizados son el clustering jerárquico [Eisen et al. (1998); Weinstein et al. (1997); Wen et al. (1998)] y k-means [Soukas et al. (2000)].

Existen numerosos métodos que se puede aplicar en el análisis no supervisado, como pueden ser, entre otros, los mapas auto-organizados (SOM) [Chavez-Alvarez et al. (2014); Tamayo et al. (1999); Törönen et al. (1999);

Wang et al. (2002)], los árboles auto-organizados (SOTA) [Herrero et al. (2001)], las redes de relevancia [Butte y Kohane (1999)], o el análisis de componentes principales [Raychaudhuri et al. (2000)]. Cada uno de estos métodos utiliza, principalmente, alguna característica de los datos y una regla para determinar relaciones que agrupen los genes (o muestras) que compartan patrones similares de expresión. Si se traslada este mismo razonamiento al contexto del análisis de la enfermedad, la aplicación de estos métodos puede ser muy útil para identificar nuevas subclases en los datos. Estos algoritmos, por tanto, dividirán los datos en grupos o clusters, pero para poder determinar si realmente son significativos o no esos grupos desde el punto de vista clínico/biológico se necesitará la participación de expertos en el área en cuestión. La valoración crítica de los resultados por parte de estos expertos resulta esencial debido a la sensibilidad de los experimentos de microarrays al ruido técnico y biológico que puede producirse en las muestra. En este sentido, hay informes recientes que confirman las buenas prácticas adquiridas en los laboratorios al llevar cabo un análisis cuidadoso de los datos, produciendo resultados reproducibles de alta calidad en los que se observa la relación existente entre el sistema biológico del estudio y los perfiles de expresión recogidos tras el experimento [Bammler et al. (2005); Dobbin et al. (2005); Irizarry et al. (2005); Larkin et al. (2005)].

1.3.4. Predicción de la clase de nuevas muestras

La tarea de predicción de la clase de una muestra intenta ir más allá de los enfoques de clustering simples vistos en la sección anterior en los que el objetivo era descubrir esa clase. Los enfoques que buscan predecir la clase utilizan perfiles de expresión ya catalogados o clasificados como medio para predecir el grupo o clase al que pertenece una nueva muestra. La pregunta que se hace en este tipo de experimento es: "¿Se puede encontrar un patrón particular y una regla matemática que permita predecir el grupo al que pertenece una muestra?" Por lo general, en este tipo de problemas se parte de un conjunto de muestras ya caracterizado o clasificado y, a continuación, se trata de descubrir aquellos genes cuyos patrones de expresión sirven para diferenciar los grupos existentes en el análisis comparando los perfiles de expresión de los genes. En este sentido, estos métodos intentan utilizar un conjunto de genes importantes o significativos que permitan desarrollar una regla matemática (o algoritmo computacional) que usen datos de perfiles de expresión de manera que dada la muestra de un individuo cualquiera, a ésta se le pueda asignar el grupo al que corresponde. El objetivo, por tanto, no es simplemente separar las muestras en grupos, sino crear una regla (o algoritmo) que permita predecir la clase o grupo en base al perfil de expresión genético de un individuo.

En los algoritmos de clasificación, las reglas matemáticas que se usan en el análisis de nuevas muestras están codificadas en el propio algoritmo

de clasificación. En la literatura existe una amplia gama de algoritmos que se han utilizado para este propósito, tales como algoritmos de voto ponderado [Golub et al. (1999)], redes neuronales artificiales (ANN) [Chou et al. (2013); Tong y Schierz (2011); Bloom et al. (2004); Ellis et al. (2002)], análisis lineal discriminante (LDA) [Antoniadis et al. (2003); Le et al. (2003); Nguyen y Rocke (2002); Orr y Scherf (2002)], árboles de decisión y regresión (CART) [Chou et al. (2013); Boulesteix et al. (2003)], support vector machines (SVM) [Xu et al. (2010); Brown et al. (2000); Ramaswamy et al. (2001)], los k-vecinos más cercanos (kNN) [Theilhaber et al. (2002)] o redes bayesianas [Piao (2011)] entre otros muchos. El principio de cualquiera de estos métodos se basa en utilizar un conjunto original de muestras, también llamado conjunto de entrenamiento, para desarrollar una regla que, a partir de una nueva muestra de otro conjunto de pruebas o conjunto de test, permita colocar esta nueva prueba dentro del contexto del conjunto de datos original, identificando así su clase. Para ello, este proceso requiere la utilización de un subconjunto de genes significativos y reducido respecto al número total de genes del conjunto de datos original, también conocido en la literatura como firma genética, y que se obtienen tras la ejecución de un método de selección de características.

1.3.4.1. Métodos de selección de características

La tarea de clasificar o predecir el cáncer utilizando datos de microarray de ADN no es una tarea trivial debido a la propia naturaleza de los datos. Los conjuntos de datos de expresión genética tienen una dimensionalidad muy alta, del orden de miles a decenas de miles de genes por muestras y tan solo unas decenas o un centenar de muestras. Esto conlleva a que al usar cualquier algoritmo de clasificación se caiga un sobreajuste o sobreentrenamiento del modelo. Este problema se conoce ampliamente en la literatura como *curse of dimensionality*.

Por lo general, las técnicas que buscan reducir la dimensionalidad de un conjunto de datos se pueden dividir en dos categorías: reducción por transformación o reducción por selección. La principal diferencia entre una u otra reside en si la técnica de reducción transforma o preserva la semántica del conjunto de datos en el proceso de reducción. Es decir, una técnica de reducción como puede ser el Análisis de Componentes Principales (PCA) transforma las características originales del conjunto de datos a un número reducido de ellas. Por contra, las técnicas de reducción por selección tratan de hallar un subconjunto mínimo de características dentro del conjunto global de ellas, pero siempre conservando el significado del conjunto de datos original. Debido a esto, en muchas aplicaciones de la bioinformática se prefiere utilizar técnicas de reducción por selección ya que ofrecen la ventaja de poder interpretar los subconjuntos resultantes por un experto en el área de estudio.

La selección de características es, por tanto, un proceso por el cual se busca reducir sistemáticamente la dimensionalidad de un conjunto de datos hacia un subconjunto óptimo de atributos a efectos de clasificación. No hay que perder de vista que el objetivo final es clasificar o predecir la clase de nuevas muestras. Además, se ha demostrado que un proceso de selección de características mejora la capacidad de predicción de un clasificador en muchas aplicaciones [Guyon (2003)]. En este contexto, las técnicas de selección de características se organizan en tres categorías dependiendo de cómo se combinan la búsqueda de selección de características con la construcción del modelo de clasificación: método de filtrado (filter) , métodos de envoltura (wrapper) y método embebidos (embedded).

Los métodos de filtrado establecen un ranking de características en base a alguna métrica univariada, siendo las características que estén situadas más arriba en el ranking las que se utilizarán en el algoritmo de clasificación. El resto de características simplemente se descartan del estudio. Estos tipos de métodos se basan, por tanto, en las características generales de los datos de entrenamiento para realizar la selección, sin intervención alguna de un algoritmo de aprendizaje. Los resultados de estos métodos no afectan de forma alguna al algoritmo de clasificación y tienen la ventaja añadida de ser muy fáciles de calcular y aplicar a conjuntos de datos más grandes, pues requieren de un tiempo de ejecución mínimo. Las redes bayesianas (NB) [Giallourakis et al. (2005)], Information Gain (IG) y Signal-to-Ratio (SNR) [Wang et al. (2006); Golub et al. (1999)], y la distancia Euclídea [Cho y Won (2003); Hu et al. (2006)] son algunos de los métodos de filtrado univariados que se han utilizado ampliamente sobre conjuntos de datos de microarrays. Otros métodos de filtrado se basan en técnicas no paramétricas como el coeficiente de correlación de Pearson [Cho y Won (2003); Hu et al. (2006)] y el análisis de significancia de microarrays (SAM) [Fung y Ng (2003)]. El principal inconveniente que presentan estos métodos es las más que probable redundancia en los genes seleccionados. Es decir, puede que ocurra que los genes mejor posicionados en el ranking presente una información similar de cara a discriminar las clases y que a pesar de que eliminemos uno de ellos, la precisión del algoritmo de aprendizaje no se vea alterada. En [Koller y Sahami (1995)] se desarrolló un método de selección llamado Markov Blanket que puede descartar genes redundantes eliminando así este problema. Basándose en este método, [Yu y Liu (2004)] propuso el método Redundancy Based Filter (RBF) para afrontar el problema de la redundancia en los genes, proporcionando unos resultados bastante prometedores.

Los métodos de envoltorio, también llamados métodos wrapper, utilizan un algoritmo de clasificación en el propio proceso de selección de características. Los métodos wrapper [Kohavi y John (1997)] realizan una búsqueda en todo el espacio de características, de manera que se evalúa la aptitud de diferentes subconjuntos de características en base al porcentaje de clasificación

obtenido de entrenar un algoritmo de aprendizaje sólo con ese subconjunto de características. Por lo general, los métodos wrapper obtienen mejores resultado de clasificación que los métodos de filtrado [Zhang et al. (2006)] aunque es importante considerar el mayor coste computacional que requieren. Dentro de los métodos wrapper se distingue entre algoritmos de búsqueda determinista y aleatorios. Los algoritmo genéticos (GA), por ejemplo, son un tipo algoritmo de búsqueda aleatoria y optimización que se fundamentan en la imitación de la evolución y la genética natural. Los GA se han empleado tanto para clasificación binaria como multiclase en el ámbito de la enfermedad de cáncer [Tong y Schierz (2011); Liu et al. (2005); Li et al. (2001)]. Por contra, un inconveniente común que presentan los métodos wrapper es que tienen un mayor riesgo de caer en el sobreajuste o sobreentrenamiento, en comparación con los métodos de filtrado, y requieren un cómputo intensivo para su ejecución.

Los métodos embebidos surgen como una alternativa que incorpore la principal ventaja de los métodos de filtrado (poca necesidad de cómputo) junto a la de los métodos wrapper (interacción del algoritmo de clasificación en el propio proceso de selección de características). Este tipo de métodos incorporan dentro del clasificador un criterio que establezca un ranking de características. El método embebido que probablemente sea el más conocido y utilizado en la literatura son las Support Vector Machine basadas en la eliminación recursiva de características (SVM-RFE) propuesto por [Guyon et al. (2002)] y utilizado, además, para la selección de genes en datos de microarrays de ADN. SVM-RFE realiza la selección de manera iterativa, entrenando un clasificador SVM con el conjunto de características de la iteración y eliminando la peor característica indicada por SVM.

1.3.4.2. Robustez de la firma genética

En la literatura existen un gran número de trabajos en los que se publica la identificación de una firma genética que proporciona buenos resultados, en términos de predicción, tras el análisis de un conjunto de muestras de microarrays de ADN. Sin embargo, es muy habitual dejar de lado una característica muy deseada y, generalmente, poco presente en estos resultados: la robustez de las firmas genéticas encontradas. Uno de los principales problemas que se puede observar si se repiten sucesivos análisis sobre un conjunto de muestras de microarrays es que al comparar las diferentes firmas genéticas obtenidas, aparecen muy pocos genes comunes a todas ellas. Es decir, en cada ejecución o análisis se obtiene como resultado una firma genética prácticamente distinta a la anterior aunque, eso sí, con muy buena capacidad de predicción.

La raíz de este problema reside, fundamentalmente, en el escaso número de muestras disponibles sobre las que se realiza la selección de genes [Ein-Dor et al. (2005)]. Para demostrar este problema, en [Ein-Dor et al. (2005)]

se concentraron en un solo conjunto de datos y repitieron muchas veces el análisis realizado en [van de Vijver et al. (2002)]. Al generar muchos subconjuntos de muestras de entrenamiento diferentes, mostraron como en [van de Vijver et al. (2002)] se podría haber obtenido distintas firmas genéticas con similar capacidad predictiva pero en las que apenas se compartiría algún gen. Esta misma hipótesis de trabajo fue apoyada por [Michiels et al. (2005)] que extendió el estudio a otros conjuntos de datos de microarrays, más allá de centrase exclusivamente en el cáncer de mama. En él muestran como, efectivamente, las firmas genéticas resultantes dependen mucho del conjunto de muestras utilizadas en la etapa de entrenamiento y en el proceso de selección de características. Por todo ello, y a pesar de que se han publicado numerosas firmas genéticas en diferentes trabajos realizados, [Ioannidis (2005); Lønning et al. (2005); Ahmed y Brenton (2005); Brenton et al. (2005)] ponen en duda la madurez de estas firmas de cara a implantar su uso en la rutina clínica diaria. En definitiva, concluyen que parece existir problemas en la metodología de selección de características por lo que, antes de aplicar y utilizar las firmas genéticas en clínica, habría que profundizar mucho más en la investigación sobre la robustez de las mismas.

1.3.5. Comparación de muestras de la misma clase

Los experimentos de comparación de clases se centran en la comparación de diferentes grupos: grupos de muestras a las que se aplicó cierto tratamiento frente a grupos de control, muestras de tejido con enfermedad y muestra de tejido sano, etc. De esta forma, lo que se persigue es descubrir los genes y los patrones de expresión que permiten diferenciar mejor ambos grupos. En este tipo de enfoque, se parte del supuesto de que se conocen las clases que están representadas en los datos. Se empieza, por tanto, asignando determinadas clases biológicas, basándose en algún criterio objetivo, a las diferentes muestras. Por ejemplo, se podría dar el caso de que los datos representen muestras tratadas con dos fármacos diferentes que producen distintas respuestas en tejidos tumorales y en tejidos sanos. En este momento, la primera pregunta que debe hacerse es: " ¿Qué genes permiten diferenciar mejor las distintas clases?" El objetivo en esta etapa es encontrar, basándose en las clases presentes en los datos, aquellos genes que son más informativos de cara a diferenciar ambas clases.

Afortunadamente, existe una amplia variedad de herramientas estadísticas que se pueden utilizar para afrontar este tipo de problemas, como pueden ser el t-test (para dos clases) [Trevino et al. (2007); Sreekumar y Jose (2008)] o el análisis de la varianza ANOVA (para tres o más clases) [Kerr y Churchill (2001); Lönnstedt et al. (2005)]. Estas técnicas estadísticas asignan un valor a cada gen, conocido como p-value, en base a la capacidad de cada gen para distinguir los grupos presentes en los datos. Uno de los problemas que presentan estos métodos se deriva de las comparaciones múltiples. Esto quiere

decir que, por ejemplo, en una matriz de datos con 10000 genes, si se aplica límite de confianza del 95 % (p=0, 05) se esperaría encontrar como resultado un total de 500 genes significativos. Resulta evidente que siguen siendo un número considerablemente alto de genes de manera que si se quiere evitar problemas a posteriori, habría que ser aún más estrictos en esta selección. Sin embargo, no hay que perder de vista que lo que realmente ofrecen estos métodos no es más que un medio para priorizar la selección de un subconjunto de genes significativos para un posterior análisis. De igual forma, existen otros enfoques como el análisis de significancia de microarrays (SAM) [Tusher et al. (2001)] que utiliza un t-estadístico ajustado para estimar la tasa de descubrimientos falsos (False Discovery Rate, FDR) en cualquier conjunto de genes significativos seleccionado.

El resultado de este tipo de análisis es, por tanto, una colección de genes que se consideran importantes para diferenciar los grupos biológicos que se comparan en el estudio. Por lo general, este subconjunto pasa a ser el punto de partida de un análisis posterior (por ejemplo, la predicción de nuevas muestras aplicando algún método que se englobe en los presentados en el Apartado 1.3.4). Pero el auténtico reto en este enfoque es, generalmente, situar estos genes en un contexto biológico para así centrarse en los elementos clave que casualmente están implicados en cualquier proceso y utilizarlos para determinar si un compuesto o un tratamiento específico puede producir una respuesta concreta.

1.4. Objetivos

Esta tesis doctoral tiene como objetivo avanzar en el conocimiento científico y tecnológico necesario para implantar, en un futuro cercano, el concepto de medicina personalizada. Para ello, se utilizarán datos de perfiles de expresión de microarrays de ADN para desarrollar modelos predictivos que permitan obtener una mejora en la capacidad de generalización de los sistemas pronóstico actuales en el ámbito clínico. De forma más detallada, esta tesis busca alcanzar los siguientes tres objetivos parciales:

- Desarrollar una herramienta pública y gratuita que incorpore diversos métodos estándar de imputación de datos perdidos. Este sistema facilitará a sus usuarios la posibilidad de adjuntar un conjunto de datos con muestras incompletas y, tras elegir el método que desea aplicar, obtener como resultado un nuevo conjunto de datos completo, sin valores perdidos. Esta tarea es fundamental como paso previo al análisis de datos de perfiles de expresión debido a que la presencia de valores perdidos es algo habitual al trabajar con conjuntos de datos reales.
- Analizar las principales deficiencias existentes en los sistemas de información actuales de un servicio de oncología y desarrollar un sistema

de información oncológico que cubra todas las necesidades para implantarlo en el servicio de oncología del hospital universitario Virgen de la Victoria (HUVV) de Málaga, España. El sistema debe integrar modelos predictivos basados en la información clínico-histopatológica para facilitar a los clínicos el diagnóstico y tratamiento de un paciente.

- Desarrollar nuevos modelos predictivos basados en datos de perfiles de expresión obtenidos a partir de alguna plataforma de secuenciación. Dado que el proceso de estimación del modelo predictivo lleva asociado un proceso de selección de características, se pretende revisar los distintos métodos de selección y clasificación más utilizados en la literatura con objeto de comparar el rendimiento obtenido. Además, las firmas genéticas encontradas y utilizadas para predecir la clase de nuevas muestras deben tener una alta capacidad predictiva así como ser robustas y relevantes desde el punto de vista clínico y biológico.

1.5. Estructura de la tesis

Tras este capítulo introductorio, la memoria de esta tesis se encuentra estructurada en los siguientes capítulos. El Capítulo 2 contempla el problema de los conjuntos de datos con valores perdidos, que es muy importante en conjuntos de datos biomédicos y, en particular, en datos de perfiles de expresión debido al alto costo de los experimentos y al escaso número de muestras disponibles. En él, se analiza y desarrolla un sistema público y gratuito que incorpora distintos métodos estándar de imputación de datos perdidos. En el Capítulo 3 se describe la experiencia vivida en el desarrollo de un sistema de información oncológico implantado en el HUVV de Málaga que permite almacenar la información clínico-histopatológica de los pacientes e integra modelos predictivos basados en este tipo de datos que facilite a los clínicos diagnosticar y predecir la evolución de un paciente. El Capítulo 4, basado en los trabajos preliminares recogidos en los Apéndices A y B, muestra una novedosa estrategia de selección de características basada en un algoritmo genético (GA) modificado al incluir dentro de la función de fitness del GA un término correspondiente a la información mutua. Además, se utiliza un modelo de red neuronal constructivo, C-Mantec, como algoritmo de clasificación. A continuación, el Capítulo 5 muestra una comparativa de los distintos métodos estándar de selección de características existentes en la literatura (wrappers, de filtrado y embebidos) utilizando para ellos diversos algoritmos de clasificación comúnmente conocidos (LDA, SVM, NB, C-Mantec, kNN y MLP). Posteriormente, el Capítulo 6 muestra una novedosa estrategia en dos etapas para buscar firmas genéticas con altas capacidades predictivas, robustas y con relevancia biológica en la enfermedad de estudio. El método desarrollado es un GA que incorpora en el propio de proceso de selección de características información biológica relativa a la enfermedad basándose en

la información contenida en la base de datos pública KEGG. Finalmente, el Capítulo 7 muestra las conclusiones derivadas de esta tesis doctoral junto a posibles líneas futuras de trabajo en este campo.

Capítulo 2

Imputación de datos perdidos en conjuntos de datos biomédicos

Resumen: La imputación de datos perdidos es una tarea crucial en el análisis de conjuntos de datos biomédicos. Al trabajar con datos reales, es muy común que se produzcan situaciones en las que es necesario clasificar muestras descritas por vectores que presenten valores perdidos, dando lugar a que algoritmos estándares de clasificación/predicción no se comporten todo lo bien que de ellos se espera. Por ello, es necesario desarrollar métodos efectivos para imputar o estimar valores perdidos en base a los valores contenidos en el resto de muestras del conjunto de datos. Estos métodos, por lo general, ya existen pero tienen el inconveniente de saber cuál de ellos es el apropiado para aplicar a los datos en cuestión. Además, la aplicación de estos métodos no suele ser fácil ni directa (en la mayoría de los casos), siendo indispensable tener un conocimiento técnico sobre el método de imputación de datos y, particularmente en el caso de trabajar con datos de perfiles de expresión, disponer de una gran potencia de cálculo para poder ejecutar estos métodos. Por lo que sabemos, a día de hoy no existe ninguna aplicación software pública y gratuita que ofrezca la posibilidad de realizar la imputación de datos perdidos en un conjunto de datos en un tiempo razonable. En este capítulo, se presenta una nueva herramienta pública que incorpora varios métodos estándares de imputación de datos perdidos, estando además ligada a un cluster de computación que ofrece la posibilidad de ejecutar métodos complejos y con grandes requisitos de computación. WIMP (Web IMPutation) es una aplicación web pública y gratuita donde los usuarios que se registren pueden crear, ejecutar, analizar y almacenar tareas relacionadas con la imputación de valores perdidos.

Título: WIMP: Web server tool for missing data imputation
Autores: Urda, D., Subirats, J.L., García-Laencina, P.J., Franco, L., Sancho-Gómez, J.L. and Jerez, J.M.
Revista: Computer Methods and Programs in Biomedicine
Volumen: 108 (3)
Páginas: 1247-1254
Año: 2012
DOI: 10.1016/j.cmpb.2012.08.006
Abstract: The imputation of unknown or missing data is a crucial task on the analysis of biomedical datasets. There are several situations where it is necessary to classify or identify instances given incomplete vectors, and the existence of missing values can much degrade the performance of the algorithms used for the classification/recognition. The task of learning accurately from incomplete data raises a number of issues some of which have not been completely solved in machine learning applications. In this sense, effective missing value estimation methods are required. Different methods for missing data imputations exist but most of the times the selection of the appropriate technique involves testing several methods, comparing them and choosing the right one. Furthermore, applying these methods, in most cases, is not straightforward, as they involve several technical details, and in particular in cases such as when dealing with microarray datasets, the application of the methods requires huge computational resources. As far as we know, there is not a public software application that can provide the computing capabilities required for carrying the task of data imputation. This paper presents a new public tool for missing data imputation that is attached to a computer cluster in order to execute high computational tasks. The software WIMP (Web IMPutation) is a public available web site where registered users can create, execute, analyze and store their simulations related to missing data imputation.

Capítulo 3

Sistema de información oncológico para la explotación de la información clínico-histopatológica de pacientes

Resumen: En este capítulo se muestra la experiencia adquirida en el diseño e implementación de un sistema de información oncológica para el servicio de oncología del hospital universitario Virgen de la Victoria de Málaga (España). El objetivo de este trabajo busca lograr el éxito en la implantación de un sistema de información oncológico atendiendo los siguientes aspectos más críticos: usabilidad, utilización de la tecnología adecuada, integración de las rutinas clínicas diarias del servicio, posibilitar la explotación estadística de los datos, seguridad e interconexión del sistema con el resto de sistemas del hospital. El sistema desarrollado consiste en una aplicación web con una arquitectura en capas buscando la usabilidad, facilidad de mantenimiento y futuras ampliaciones del sistema, incorporando la información clínico-histopatológica de los pacientes en diferentes módulos (gestión de pacientes, hospital de día, ensayos clínicos, consejo genético y análisis estadístico). Este último módulo es el que hace que este sistema destaque sobre otros similares, ya que permite ejecutar modelos predictivos basados en la información clínica (análisis de supervivencia, regresión de Cox, funciones de riesgo y tablas de contingencias). La evaluación del éxito en la implantación del sistema se llevó a cabo a partir de encuestas realizadas a los usuarios finales del mismo a los 3 y 15 meses tras su implantación en el hospital, mostrando una satisfacción generalizada del personal.

Título: Addressing critical issues in the development of an Oncology Information System
Autores: Urda, D., Ribelles, N., Subirats, J.L., Franco, L., Alba, E. and Jerez J.M.
Revista: International Journal of Medical Informatics
Volumen: 82 (5)
Páginas: 398-407
Año: 2013
DOI: 10.1016/j.ijmedinf.2012.08.001
Abstract: Purposes: This paper presents the experience on the design and implementation of a user-centered Oncology Information System developed for the Medical Oncology Department at the "Hospital Universitario Virgen de la Victoria", in Málaga, Spain. The project focused on the aspects considered in the literature as critical factors for a successful deployment and usage of a health information system.
Methods: System usability, adequate technology, integration of clinical routines, real-time statistical analysis of data, information confidentiality and standard protocol-based external interconnection were the key aspects considered.
Results: The developed system is based on a web application with a modular and layered architecture accounting for usability, ease of maintenance and further system development. Evaluation of system usability was carried at three and fifteen months after system deployment to analyze the advantages and disadvantages experienced by the end-users.
Conclusions: A thorough prior analysis of clinical activities and workflows, the use of the adequate technology, and the availability of data analysis tools will almost guarantee success in the deployment of an Oncology Information System.

Capítulo 4

Estimación de modelos predictivos usando algoritmos genéticos y C-Mantec

Resumen: La selección de características es un aspecto muy importante en el análisis de los datos de cara a identificar los genes más relevantes que maximicen la capacidad de predicción. En este capítulo, se lleva a cabo un estudio comparativo entre un algoritmo de selección de características por pasos hacia adelante (SFS) y un algoritmo genético (GA) con el objetivo de identificar subconjuntos de genes que tengan altas capacidades predictivas y relevancia biológica. Para evaluar los diferentes subconjuntos de genes testados, se han utilizado seis modelos estándares de clasificación: "Linear Discriminant Analysis" (LDA), "Support Vector Machines" (SVM), Redes Bayesianas (NB), C-MANTEC, "K-Nearest Neighbors" (kNN) y el Perceptrón Multicapa (MLP). Ambos marcos de selección de características se ejecutaron sobre seis bases de datos públicas y gratuitas de microarrays de ADN. Los resultados obtenidos muestran mejores resultados, en términos de predicción, de la estrategia GA aunque cabe mencionar que, en comparación con SFS, esta estrategia da como resultados subconjuntos de genes más grandes y es computacionalmente más exigente, ya que evalúa muchos más subconjuntos de genes. En cuanto a los modelos de clasificación utilizados, MLP, LDA y SVM obtienes los mejores resultados mientras que C-MANTEC y kNN le siguen de cerca aunque con una predicción ligeramente menor. Además, C-MANTEC, MLP y LDA dan lugar a resultados con menor número de genes en comparación con SVM, NB y kNN, y en particular C-MANTEC es el clasificador más robusto en cuánto al ajuste de los parámetros del modelo.

Título: Application of genetic algorithms and constructive neural networks for the analysis of microarray cancer data
Autores: Luque-Baena, R.M., Urda, D., Subirats, J.L., Franco, L. and Jerez J.M.
Revista: Theoretical Biology and Medical Modelling
Volumen: 11 (Suppl. 1)
Páginas: S1-S7
Año: 2014
DOI: 10.1186/1742-4682-11-S1-S7
Abstract: Background: Extracting relevant information from microarray data is a very complex task due to the characteristics of the data sets, as they comprise a large number of features while few samples are generally available. In this sense, feature selection is a very important aspect of the analysis helping in the tasks of identifying relevant genes and also for maximizing predictive information.
Methods: Due to its simplicity and speed, Stepwise Forward Selection (SFS) is a widely used feature selection technique. In this work, we carry a comparative study of SFS and Genetic Algorithms (GA) as general frameworks for the analysis of microarray data with the aim of identifying group of genes with high predictive capability and biological relevance. Six standard and machine learning-based techniques (Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), Naive Bayes (NB), C-MANTEC Constructive Neural Network, K-Nearest Neighbors (kNN) and Multilayer perceptron (MLP)) are used within both frameworks using six free-public datasets for the task of predicting cancer outcome.
Results: Better cancer outcome prediction results were obtained using the GA framework noting that this approach, in comparison to the SFS one, leads to a larger selection set, uses a large number of comparison between genetic profiles and thus it is computationally more intensive. Also the GA framework permitted to obtain a set of genes that can be considered to be more biologically relevant. Regarding the different classifiers used standard feedforward neural networks (MLP), LDA and SVM lead to similar and best results, while C-MANTEC and k-NN followed closely but with a lower accuracy. Further, C-MANTEC, MLP and LDA permitted to obtain a more limited set of genes in comparison to SVM, NB and kNN, and in particular C-MANTEC resulted in the most robust classifier in terms of changes in the parameter settings.
Conclusions: This study shows that if prediction accuracy is the objective, the GA-based approach lead to better results respect to the SFS approach, independently of the classifier used. Regarding classifiers, even if C-MANTEC did not achieve the best overall results, the performance was competitive with a very robust behaviour in terms of the parameters of the algorithm, and thus it can be considered as a candidate technique for future studies.

Capítulo 5

Métodos de selección de características y algoritmos de clasificación

Resumen: En este capítulo se realiza un estudio exhaustivo de los diferentes métodos de selección de características y algoritmos estándar de clasificación (LDA, SVM, kNN, Naive-Bayes, C-MANTEC, y MLP), utilizando un esquema de validación honesto basado en la estrategia "bootstrap cross-validation" (BCV). Hasta la fecha, no conocemos de la existencia de un trabajo similar que englobe en su totalidad un estudio similar de estas características. El estudio realizado se probó en seis conjuntos de datos de microarrays relacionados con el cáncer. Los resultados obtenidos muestran que los métodos de selección de características de filtrado o embebidos son mejor, en general, que los métodos "wrapper", ya que los resultados son mejores en términos de predicción y, además, son menos exigentes en cuanto al coste computacional que necesitan para poder ejecutarlos, obteniendo lo resultados en un tiempo menor. Respecto a los algoritmos de clasificación, kNN y MLP son los algoritmos que pueden considerarse más robustos independientemente del método de selección de características utilizado. Sin embargo, otros clasificadores como SVM o C-Mantec podrían alcanzar resultados similares con un ajuste óptimo de los respectivos parámetros de ambos modelos.

An insight into feature-selection procedures and prognostic models for microarray data

Daniel Urda[a,*], Rafael M. Luque-Baena[b], Noelia Sánchez-Maroño[c], Leonardo Franco[a], Jose M. Jerez[a]

[a] *Department of Computer Science, University of Málaga, Spain.*
[b] *Department of Computer Systems and Telematics Engineering. University of Extremadura, Mérida, Spain*
[c] *Department of Computing, University of La Coruña, Spain*

Abstract

The analysis of DNA microarray datasets usually involves the use of a feature-selection process to find the best subset of genes that produces a better prediction assessment. In this paper, an exhaustive study - including different types of feature selection methods and classifiers (LDA, SVM, kNN, Naive-Bayes, C-MANTEC, and Artificial Neural Networks) - using a bootstrap cross-validation strategy as an honest validation scheme is presented, since no similar works have previously included this kind of analysis in a single paper. The results of this study for six benchmark datasets show that filter or embedded methods are preferred, in general, to wrapper methods according to their better statistical significant results, in terms of accuracy, and lower demand for computational resources.

Keywords: Microarray data, Cancer disease, Feature selection,

*Corresponding author
Email addresses: durda@lcc.uma.es (Daniel Urda), rmluque@unex.es (Rafael M. Luque-Baena), nsanchez@udc.es (Noelia Sánchez-Maroño), lfranco@lcc.uma.es (Leonardo Franco), jja@lcc.uma.es (Jose M. Jerez)

Classification model, Genetic profile, Evolutionary algorithm

1. Introduction

DNA microarray technology is a powerful platform for the analysis of gene expression data and it has been been widely used for the prediction of cancer disease outcome [1, 2, 3, 4]. Although NGS clearly seems to be the predominant technology for the near future of biomedical research [5], DNA microarray still remains as a low-cost sequencing technology easily available in clinical and biological experimentation [6, 7, 8, 9, 10, 11, 12, 13]. One of its present application is the estimation of gene expression signatures in cancer studies [14, 15, 16, 17], which is a complex task that involves three different steps that should be carried out rigorously: feature selection, classification-model selection, and prediction assessment [18].

Feature selection refers to which variables will be included in the prognosis model, and it is a crucial step in developing a class predictor. Filter methods such as Partial Least Squares (PLS) regression [19], Information Gain (IG) [20], Minimum-Redundancy Maximum-Relevance (mRMR), and ReliefF [21] are among the statistical techniques proposed to address this problem. On the other hand, wrapper methods such as Stepwise Forward Selection (SFS) [22, 23], Ant Colony optimization [24], and evolutionary models [25, 26, 27, 28] have been applied as heuristic methods from the computational intelligence perspective.

With regard to classification model selection, different algorithms have been studied for the identification of differentially expressed genes in microarray data. Classification methods such as Neural Net-Multilayer Perceptron

(NN) [29, 22], Support Vector Machines (SVM) [30, 29], Naive Bayes (NB) [31, 32], k-Nearest Neighbour (kNN) [33, 34], Decision Trees (DT) [35, 36], and RF (Random Forest) [34, 37] have been used in recent studies.

With regard to prediction assessment, the obtained performance of the predictive models is also relevant. As few samples are typically available in microarray data, resampling techniques are a suitable methodology. In this sense, the feature selection procedure is performed within each resampling step in order to estimate prediction errors. This process is known as honest performance assessment [38], a necessary process in the analysis of microarray data that has been overlooked in several works [22, 39, 40, 41]. On the other hand, honest validation strategies are presented in [18], where the .632+ bootstrap method is highlighted for high-dimensional genomic studies and a number of existing bootstrap methods are compared (out-of-bag estimation and a bootstrap cross-validation (BCV) method [42]).

Several works have focused on the analysis of model performance within different feature selection strategies [22, 40, 29]. Others centred their attention on prediction assessment [18], while other authors focused on the comparison of different feature selection procedures [43]. Additionally, in [44] an evolutionary method with a fixed feature size is combined with filter techniques and classifiers, but one specific dataset is analysed, and there is no estimation of the parameters of the classification method. Thus, conclusions according to the robustness and accuracy of the filter and classification techniques in prognosis estimation would not be suitably inferred, due to the low number of datasets analysed. But up to date, these three steps involved in the estimation of prediction models have not been analysed in a single

work.

The contributions of this paper are an exhaustive computational work combining the use of eight feature-selection (FS) procedures and six classification algorithms (including several machine-learning algorithms introduced in recent years) to analyse six public cancer microarray datasets; a comparison between the robustness obtained by different feature-selection techniques with regard to the variability of the subsets of genes; an analysis of the influence of the validation scheme in the gene selection profile; and an analysis of the frequency of selected genes for each feature-selection method and the relationship between these selected genes and the studied disorder according to biological information. The rest of the paper is structured as follows. Section 2 shows the databases used within this study. Section 3 describes the feature-selection techniques tested in this work to obtain a subset of features and estimate prediction errors. Section 4 presents the methodology of our approach, and Section 5 shows the experimental results for different databases and feature-selection techniques. Finally, Section 6 provides the final conclusions of this paper.

2. Datasets

Six free public high-dimensional biomedical datasets[1,2] have been used within this work. The information of each dataset is shown in Table 1, and each is related to the study of a specific cancer: breast, leukaemia, lung, colon, and prostate cancer diseases.

[1]http://datam.i2r.a-star.edu.sg/datasets/krbd/

[2]http://cilab.ujn.edu.cn/datasets.htm

Dataset	*#Genes*	*#Samples*	*Class 0 (**"normal"**)*	*Class 1 (**"cancer"**)*	*Proportion of class "normal"*
West_ER	7129	49	25	24	0.51
Breast	24481	78	34	44	0.436
Leukaemia	7129	72	25	47	0.347
Lung	12533	181	150	31	0.829
Colon	2000	62	22	40	0.355
Prostate	12600	102	50	52	0.49

Table 1: Information about the six databases analysed.

The West_ER dataset uses microarray technology to analyse primary breast tumours in relation to oestrogen receptor (ER) status. The Breast dataset was reported for patients' outcome prediction related to breast cancer disorder. The Leukaemia dataset contains measures that correspond to Acute Lymphoblastic Leukaemia (ALL) and Acute Myeloid Leukaemia (AML) samples and is a dataset that often serves as a benchmark for microarray analysis methods. The Lung dataset presents a classification between malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung. The Colon dataset collects samples from colon-cancer patients. Finally, the Prostate dataset contains different neoplastic samples as these tumours are among the most heterogeneous of cancers both histologically and with respect to highly divergent clinical outcomes.

3. Feature-Selection Framework

Filter, wrapper, and embedded methods are the three main categories into which feature selection techniques can be divided [43]. In filter methods, some statistical procedures are applied to remove irrelevant features, as it is a method that is completely independent of the classifier. Wrapper methods evaluate different subsets of features within a classification algorithm comparing their accuracy, thus requiring more computational resources in contrast to filter methods. Finally, embedded methods could be seen as a mix of filter and wrapper methods where the search space is composed of the feature-selection procedure and the classification algorithm as a whole, thus also being a classifier-dependent method.

This work aims to select the most significant subset of features, in terms of prediction, that are also associated with the studied disorder. Several feature-selection procedures are analysed, looking for good generalization rates in the prediction stage.

3.1. Wrapper methods

3.1.1. Stepwise Forward Selection (SFS) procedure

This procedure analyses the inclusion of one or several features in order to improve the performance of the classification task. Thus, sequential forward selection [45] chooses the best variable in each iteration by minimizing the misclassification rate, and includes it in the final subset of features, starting with an empty set. The algorithm will continue to add variables until the resulting subset does not improve, in terms of a specific criterion. In our implementation, an improvement on the stopping criteria was introduced.

In order to decide the incorporation of a new variable in the final subset of features, Lilliefors' test is applied to test the normality condition of both distributions and, depending on the verification or rejection of this hypothesis, a T-test or Wilcoxon rank sum test is applied to check that the independent samples come from distributions with equal means or medians. In others words, we check that the results obtained by adding a new variable to the final subset of features are statistically significant.

3.1.2. Evolutionary strategy (GA)

GAs are a class of optimization procedure inspired by the biological mechanisms of reproduction. In this kind of optimization problem, a fitness function $f(\mathbf{x})$ should be maximized or minimized over a given space X of arbitrary dimension. In this work, a previously published GA [46] has been used to perform the analysis.

This GA uses a simple encoding scheme where genes are associated to features and are represented using a string of bits whose length is determined by the total number of genes. If the i^{th} bit of the string is active (value 1), then the i^{th} gene is selected as part of the feature subset. Otherwise, a value of 0 indicates that the corresponding feature is ignored. In all the experiments, a population size of 100 random individuals and an elite count value of 10 (the number of chromosomes which are retained in the next generation) were used. Furthermore, a selection strategy based on a roulette wheel and uniform sampling, a scattered crossover with a crossover rate set to 0.8 and a mutation operator with a probability rate of 0.2 were applied.

Fitness function. The fitness function assesses each chromosome in the population so that it can be ranked against all the other chromosomes. The main goal of feature subset selection is to use fewer features to achieve the same or better performance. Additionally, it has been found that the combination of features with low redundancy among them, that is, that provide different information about the target class, and with a certain resemblance to the target class can improve the performance rates [47]. Therefore, the fitness function should contain three terms: the misclassification error, the number of features selected, and a redundancy measure among them. Datasets are split into training and testing sets in order to evaluate the generalization ability of the proposed chromosome.

Statistical techniques such as mutual information [48] give us an idea of the correlation between a pair of features. The mutual information between two continuous random variables y and z is given by

$$I(y,z) = \int\int p(y,z) \log\left(\frac{p(y,z)}{p(y)\,p(z)}\right) dy\,dz \tag{1}$$

where $p(y,z)$ is the joint probability density function of y and z, and $p(y)$ and $p(z)$ are the marginal probability density functions of y and z respectively. The mutual information is symmetric.

Moreover, it is non-negative, with a zero value indicating that the variables are independent. The more correlated two variables are, the greater their mutual information. Advantages of mutual information are that the dependency between variables is no longer restricted to being linear and it can handle nominal or discrete features. Although it is hard to compute for continuous data, the probability densities can be discretized using histograms, which are considered as good approximations [49]. A measure which incor-

porates the correlation of the features with the target class and penalizes the redundancy among the selected features is described as follows [47]:

$$corr(\mathbf{x}) = \frac{1}{t}\sum_{i=1}^{k}\sum_{j=i+1}^{k} I(x_j, x_i) - \frac{1}{k}\sum_{j=1}^{k} I(x_j, C) \tag{2}$$

where k is the number of features selected, C is the target class, and t is the number of combinations among the pairs of chromosome x analysed. Finally, the function to be minimized is represented as follows:

$$fitness(\mathbf{x}) = (1 - ACC(\mathbf{x})) + \lambda\frac{k}{\mathcal{N}} + \beta corr(\mathbf{x}) \tag{3}$$

where $fitness(\mathbf{x})$ is the fitness value of the feature subset represented by $\mathbf{x}$; $ACC(\mathbf{x})$ is the accuracy rate obtained by the classifier using the test set; $\mathcal{N}$ is the total number of extracted features; finally, $corr(\mathbf{x})$ defines the correlation among the features and the target class, with the aim of avoiding the redundancy in the feature vector (Equation 2). The parameters λ and β can take values in the interval $(0, 1)$ and were empirically chosen to be 0.4 and 0.25, respectively.

Therefore, if two subsets achieve the same performance while containing different numbers of features, the subset with fewer features is preferred. We also prefer the mixture of features that are less redundant among them, which is considered a good quality for classification tasks. Nevertheless, among the three terms - error, feature subset size, and correlation - the first one is our major concern.

3.2. Filter methods

With regard to the relationship between a feature selection algorithm and the inductive learning method used to infer a model, filter methods rely on the general characteristics of training data and carry out the feature selection process as a pre-processing step with independence of the induction algorithm.

3.2.1. Correlation-based Feature Selection (CFS)

This is a simple multivariate filter algorithm that ranks feature subsets according to a correlation-based heuristic evaluation function [50]. The bias of the evaluation function is toward subsets that contain features that are highly correlated with the class and uncorrelated with each other. Irrelevant features should be ignored because they will have low correlation with the class. Redundant features should be screened out as they will be highly correlated with one or more of the remaining features. The acceptance of a feature will depend on the extent to which it predicts classes in areas of the instance space not already predicted by other features.

3.2.2. Consistency-based filter (Cons)

The consistency-based method [51] uses an inconsistency rate over the dataset for a given feature. This rate is based on counting, in a specific way, inconsistent patterns, where a pattern is considered inconsistent if there exist at least two samples such that their features have the same values but the output class is different.

3.2.3. Information Gain (IG)

The IG [52] filter is one of the most common univariate methods of evaluating attributes. It is a univariate filter that evaluates the features according to their information gain and considers a single feature at a time. It provides an orderly classification of all the features, and then a threshold is required to select a certain number of them according to the order obtained.

3.2.4. Minimum-Redundancy Maximum-Relevance (mRMR)

The mRMR method [47] selects features that have the highest relevance with the target class and are also minimally redundant; that is, it selects features that are maximally dissimilar to each other. Both optimization criteria (maximum relevance and minimum redundancy) are based on mutual information.

3.2.5. ReliefF

This algorithm [53] is an extension of the original Relief algorithm [54]. The original Relief algorithm works by randomly sampling an instance from the data and then locating its nearest neighbour from the same and opposite class . The values of the attributes of the nearest neighbours are compared to the sampled instance and used to update relevance scores for each attribute. The rationale is that a useful attribute should differentiate between instances from different classes and have the same value for instances from the same class. ReliefF adds the ability to deal with multiclass problems and is also more robust and capable of dealing with incomplete and noisy data. This method may be applied in all situations, has low bias, includes interaction among features, and may capture local dependencies which other methods

miss.

3.3. Embedded methods

Embedded methods perform feature selection in the process of training and are usually specific to given learning machines. One of the most well-known embedded methods is SVM-RFE (Recursive Feature Elimination for Support Vector Machines), which was introduced by Guyon et al. in [55]. They defined an iterative procedure called Recursive Feature Elimination (RFE) consisting of three steps: 1) training the classifier, 2) ranking features according to a given criterion, and 3) removing features with the smallest criterion. SVM-RFE is an application of RFE using the weight magnitude as a ranking criterion.

4. Validation Scheme

In this paper, an honest validation strategy is applied with the aim of obtaining a final subset of features with high prediction capabilities. In this sense, it is important to highlight the choice of a BCV strategy to obtain an accuracy measure of an external validation since its good behaviour in estimating misclassification error with small samples, as is the particular case of DNA microarray datasets, has been previously demonstrated in [42, 56]. A high-level description of our methodological approach is shown in Figure 1 as well as a brief pseudocode of the algorithm is described in Algorithm 1. In concrete, the developed procedure executes a 50-bootstrap resampling as external validation and 5-k-fold for internal validation techniques. Thus, this scheme will lead us to find subsets of features with high generalization rates

in the prediction stage, as this is essential for determining the probability of suffering from a specific condition.

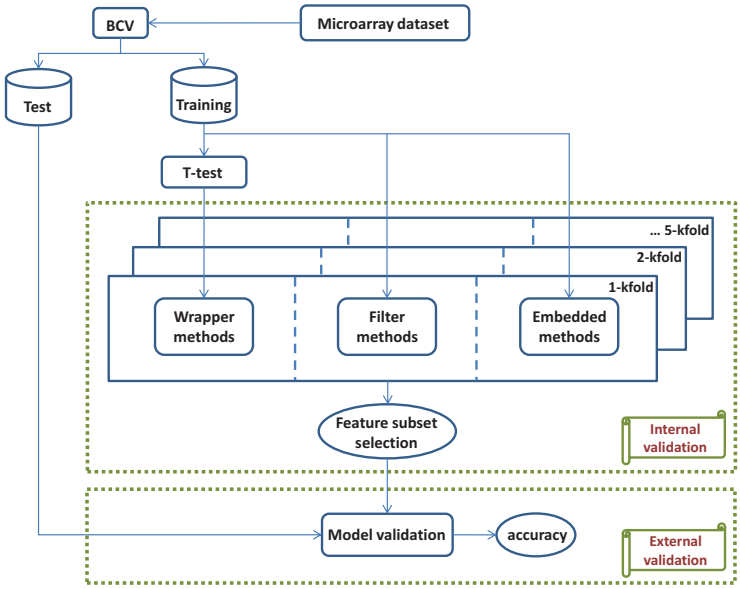


Figure 1: Honest validation scheme used in the feature-selection procedure and prediction error estimation.

A particularity of this approach is that the applied BCV strategy uses a stratified bootstrap resampling method. Therefore, each resampling keeps the proportion of positive and negative instances of the "class" attribute reducing the variance in the simulations. Moreover, in the case of wrapper methods, Welchś t-test [57] is applied assuming that the two classes (the patient does or does not have cancer) have unknown and unequal variances, because it is not advisable to use the basic form if we are unsure whether the requirements of the test are satisfied [43]. The top 200 of the total number of genes are retained according to the p-value descending sort, which will be

Algorithm 1 Brief pseudocode of our methodological approach

{initialization}
$[Train, Test]\{1..50\} \Leftarrow BCV(dataset, 50)$
$FeatureSelectionMethods$ =[SFS,GA,CFS, ... ,SVM-RFE]

for all method m in $FeatureSelectionMethods$ **do**
  {first-step: feature selection process}
  for $i = 1 \rightarrow 50$ **do**
    $TR_i \Leftarrow Train[i]$
    if (IsWrapperMethod(m)) **then**
      $TR\_Reduced_i \Leftarrow Ttest(TR_i)$
      $[IntVal_i, Features_i] \Leftarrow ExecWrapper(m, TR\_Reduced_i)$ //involves execution of a classification method
    else if (IsFilterMethod(m)) **then**
      $[Features_i] \Leftarrow ExecFilter(m, TR_i)$
      $[IntVal_i] \Leftarrow ExecClassificationMethod(TR_i, Features_i)$
    else
      $[Features_i] \Leftarrow ExecEmbedded(m, TR_i)$
      $[IntVal_i] \Leftarrow ExecClassificationMethod(TR_i, Features_i)$
    end if
  end for
  $InternalValidation \Leftarrow mean(IntVal_i)$

  {second-step: model validation}
  for $i = 1 \rightarrow 50$ **do**
    $T_i \Leftarrow Test[i]$
    $ExtVal_i \Leftarrow Accuracy(T_i, Features_i)$
  end for
  $ExternalValidation \Leftarrow mean(ExtVal_i)$
end for

the input of a wrapper method feature-selection procedure.

Several standard and well-known classification models have also been tested in this paper: Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), k-Nearest Neighbours (kNN), Naive-Bayes (NB), C-MANTEC (CM)[58] as a constructive neural network model, and a standard

Multilayer Perceptron (NN). The aim of the LDA classifier is to find a linear combination of features which separates two or more classes of patterns. SVM is a more sophisticated and widely used method that finds the optimal separation margin between two classes. The standard kNN algorithm tries to classify an object according to the majority vote of its neighbours, with the object being assigned to the most common class amongst its k nearest neighbours. NB is a probabilistic model classifier that assumes that the presence or absence of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. CM is a novel neural network constructive algorithm that utilizes competition between neurons and a modified perceptron learning rule to build compact architectures with good prediction capabilities. Finally, NN is a standard feedforward artificial neural network model that maps sets of input data onto a set of appropriate outputs, consisting of multiple layers of neurons in a directed graph, with each layer fully connected to the next one, and using a supervised learning technique called backpropagation for training the network.

Table 2 shows the set of parameters needed for each classifier and the different values that have been tested in this paper. The combination of these values generates a set of configurations that are evaluated for the 50 resamplings since the train dataset is different in each one. In this sense, the model selection implies a huge effort in terms of computational and time resources, since in the case of wrapper methods many different subsets of features are evaluated in each iteration. Therefore, the authors propose to adjust the parameters of each classification method by using the top 200 variables after sorting the p-values obtained by the application of Welchś

t-test. As a result of this parameter estimation phase, every configuration is labelled by an accuracy measure in a reasonable time, keeping the parameter configuration with the highest result as the one to be used in the rest of the procedure (feature selection and external validation). The accuracy measure is obtained through the .632+ bootstrap method [59], as it is highlighted for high-dimensional genomic studies.

Algorithm	**Test Parameters**
LDA	No parameters
SVM	Kernel, t= {linear, polynomial, radial base, sigmoid} Cost, C= {1, 3, 5, 7, 9, 10, 12, 15} Degree, d= {1, 2, 3, 4, 5} Gamma, g= {0.001, 0.005, 0.1, 0.15, 0.2, 0.4, 0.6, 0.8, 1, 2, 3, 5} Coef0, r= {0, 1, 2}
kNN	Neighbours, k= {1, 2, 3, 4, 5, 6, 7, 8, 9, 10} Distance type, d= {euclidean, chi-squared, cosine-similarity}
NB	Kernel density, K= {0, 1} Supervised discretization, D= {0, 1}
CM	Max. Iterations, I_{max}= {1000, 10000, 100000} GFac, g_{fac} = {0.01, 0.05, 0.1, 0.2, 0.25, 0.3} Phi, ϕ = {1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6}
NN	Hidden neurons, $NHidden$= {2, 3, 4, 5, 6} Alpha, α= {0.05, 0.1, 0.2, 0.3, 0.5} Number of cycles, $NCycles$= {10, 25, 50}

Table 2: Parameter settings tested during evaluation of the classification algorithms. The combination of all the values of the parameters generates a set of configurations for each method.

5. Results

Figure 2 presents a summary of the raw data results shown in Tables A.1 and A.2. The bar diagram over each cancer dataset represents the performance of eight feature-selection procedures analysed in this paper: two different wrapper methods (SFS and GA), five different filter methods (CFS,

Cons, IG, mRMR, and ReliefF), and one embedded method (SVM-RFE). This performance is computed after averaging the accuracy obtained with six machine-learning classifiers (LDA, SVM, kNN, NB, CM, and NN). In general, filter and embedded methods are distinguished by the most accurate results in contrast to wrapper methods, independently of the cancer microarray dataset analysed. In concrete, mRMR emerges as the one with the best performance in three out of six analysed cancer datasets (Leukaemia, Lung, and Colon). Therefore, the results suggest the use of filter or embedded methods instead of wrapper methods, since the latter are more highly computationally demanding, leading to lower performance results on average.

Regarding the six cancer microarray datasets analysed, for three of them (the Leukaemia, Lung, and Prostate datasets), a very good classification result is obtained independently of the feature-selection procedure (over a 90%). On the other hand, the West_ER and Colon datasets present good classification results (over 80%) while the Breast cancer dataset appears to be the most difficult problem as a success rate of only 65% is achieved, which could lead us to think that more patient samples are needed to estimate gene-expression-based predictors.

5.1. Classification models' performance

Regarding the classifier models' performance, a deeper analysis has been carried out and is represented in Figures 3 and 4, which summarize the raw data results shown in Tables A.3 and A.4. In particular, Figure 3 shows a pairwise analysis of the number of times that one classifier is statistically significant better than another one according to the multiple comparison test. The thickness of the lines connecting the different classifiers indicates

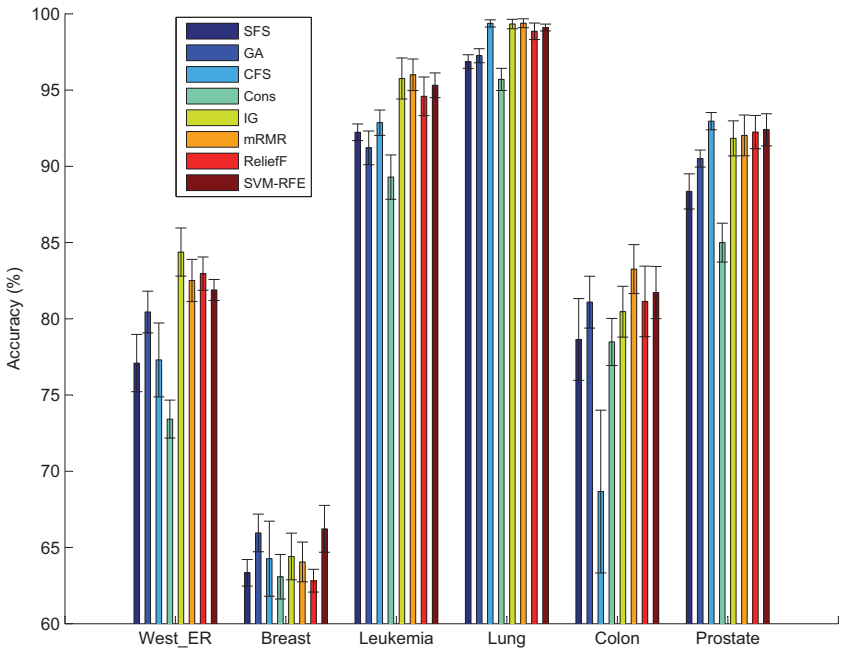


Figure 2: Performance comparison (after averaging the accuracy of six machine learning classifiers) for eight feature selection procedures over six different cancer microarray datasets.

four categories in which the comparison was done: classifiers that are better than the other less than four times are discarded and not represented on the graph; those that are four to six times better than the other are represented by the thinnest line on the graph; those that are seven to nine times better than the other are indicated with a middle thickness line; finally, the last category represents classifiers that are better than the other more than nine times (shown in the graph with the thickest line). The results suggest that the kNN, NN, or NB classifier should be used rather than LDA and the rest. LDA is the only classifier that does not outperform any other classifier (does

not reach the category that is seven to nine times better), so in principle it could be discarded as a classifier for the DNA microarray analysis.

Regarding the classifier models performance, a deeper analysis has been carried out and represented in Figures 3-4, which are a summary of the raw data result shown in Tables A.3-A.4. In particular, Figure 3 shows a pairwise analysis of the number of times that one classifier is more statistically significant than another one according to the multiple comparison test. The thickness of the lines connecting the different classifiers indicates four categories in which the comparison was done: classifiers that are better than other less than 4 times are discarded and not represented on the graph; those that are 4-6 times better than other are represented with the thinnest line on the graph; then, classifiers that are 7-9 times better than others are indicated with a middle thickness line; and finally, the last category represent classifiers that are better than the other classifiers more than 9 times (shown in the graph with the thickest line). The results suggest the use of kNN, NN or NB classifiers in comparison to LDA and the rest. LDA is the only classifier that does not overcome any other classifier (does not reach the 7-9 category), so in principle it could be discarded as a classifier for the DNA microarray analysis.

Finally, Figures 4a) and 4b) summarize the findings of Tables A.4 and A.3 respectively using bar diagrams. Figure 4a) shows the percentage value of the number of times (occurrences) that a given classifier leads to statistically significant different results in comparison to a control group (the lowest in performance) computed among all analysed cases (different datasets and FS procedures), while Fig. 4b) shows a similar analysis but for different FS

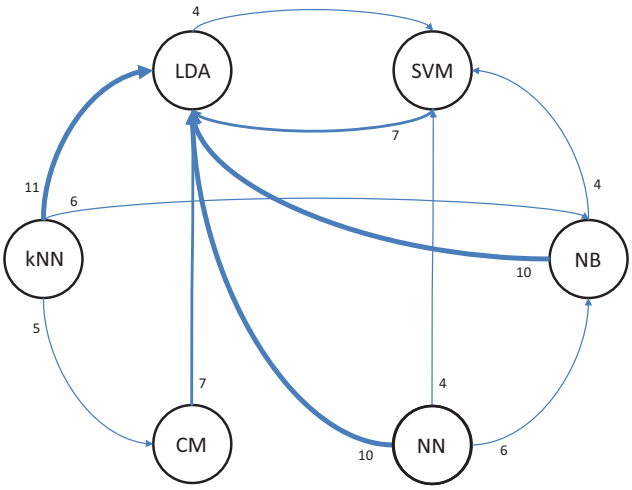
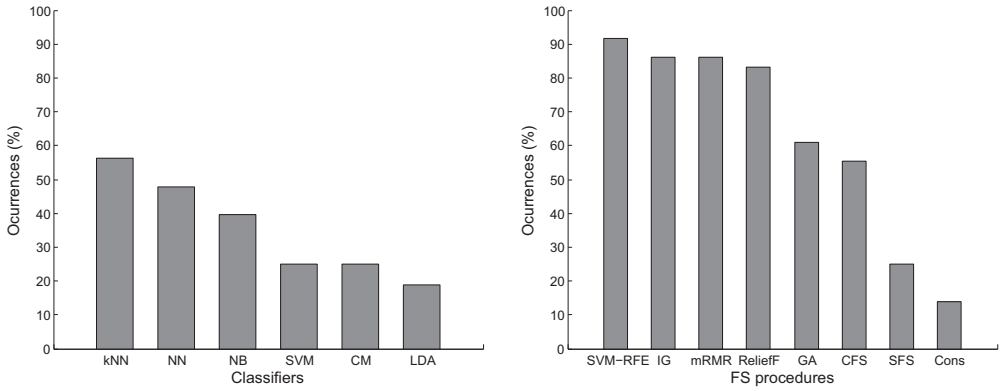


Figure 3: Pairwise graph showing the number of times that the performance of a given classifier is statistically significant in comparison to other classifiers.

procedures among all datasets and classifiers. The histogram shown in 4a) indicates that kNN is the preferred classifier as it does outperform other classifiers almost 60% of the time, while LDA behaves quite poorly as it achieves the best results less than 20% of the time.

5.2. Honest validation scheme

The use of an honest validation scheme is relevant, as performance results could be very optimistic otherwise. In this sense, and according to the results shown in Figure 2, we selected three a priori more difficult datasets (West_ER, Breast, and Colon) in order to perform a detailed analysis using SFS and GA as FS procedures. Table 3 shows the performance results of the LDA and SVM classifier models for each dataset with and without using an honest validation scheme. As expected, the behaviour of FS procedures is very optimistic if no honest validation is applied, independently of the classifier



(a) (b)

Figure 4: Summary of results. (a) Percentage value of the number of times (occurrences) that a given classifier leads to statistically significant different results in comparison to a control group computed for all analysed cases (different datasets and FS procedures). (b) Analysis similar to before as before but for different FS procedures performed over all datasets and classifiers.

used. The final accuracy measures vary from honest validation schemes to non-honest ones approximately a 20% (West_ER), $16-20\%$ (Breast), and $5-18\%$ (Colon). Regarding the overfitting problem in the feature selection, an overfitting index (OI) was computed to analyse how much this effect affects the FS procedures, classifiers, and datasets. It was computed as $OI = 1-(with\_hv/without\_hv)$, and averaged across datasets and classifiers. The results of the overfitting index were $OI = 0.1683$ for the SFS procedure and $OI = 0.2123$ for the GA. Therefore, previous articles that did not use an honest validation scheme presented over-optimistic results as no test set was kept apart from the FS procedure.

5.3. Robustness of the FS procedures

An important aspect of the FS procedure is the variability observed in the set of selected subsets of genes in different executions of a given algo-

		SFS		GA	
	Classifier	*with honest validation*	*without honest validation*	*with honest validation*	*without honest validation*
West_ER	LDA	73.29±9.23	95.14±1.38	81.16±9.34	99.52±0.37
	SVM	77.78±7.10	96.63±1.05	79.58±8.83	99.32±0.47
Breast	LDA	63.35±7.00	79.45±4.34	66.23±7.74	95.10±1.38
	SVM	64.27±6.87	82.36±2.26	66.72±8.54	97.78±0.95
Colon	LDA	81.69±5.72	86.04±2.01	83.70±7.09	92.45±1.22
	SVM	78.39±7.83	88.27±1.72	78.61±8.08	95.28±1.11

Table 3: Performance comparison among two different wrapper methods (SFS and GA) and two classifiers (LDA and SVM) using three datasets (West_ER, Breast and Colon). The results shown correspond to the accuracy of each classification method using the honest validation scheme proposed in this work and without using it, both shown in the format of *mean±standard deviation*.

rithm. In order to quantify this, we compute a robustness index for each FS procedure used, taking into consideration the subset of genes obtained for every resampling of the dataset. First, the absolute frequency for each gene is computed in order to retain those genes selected at least 5% of the time . Then the set of selected genes is sorted in descending order according to the relative frequency, discarding those genes for which the cumulative frequency is greater than 80%. Finally, the robustness measure is calculated as the average of the relative frequencies of the resulting genes.

Figure 5 shows the robustness value obtained for each dataset, depending on the FS procedure used. ReliefF could be considered the most robust FS procedure according to our analysis, since it leads to the highest robustness values for three datasets with competitive values for the other three. Moreover, IG and mRMR are a step backward in comparison to ReliefF but they also have competitive robustness values. On the other hand, the remainder of

the FS procedures have values of less than 0.5 for almost all of the datasets, and thus it can be derived that on several executions of the algorithm, a different subset of genes will be obtained. It should be noted that there is no clear correlation between the robustness and the accuracy measure, since the most robust method (ReliefF) is not the same as the most accurate technique (SVM-RFE). Between the wrapper methods, GA overcomes SFS in both robustness and accuracy. To further confirm the results shown in the figure, permitting a more direct comparison of the robustness of the FS procedures, we compute a weighted average of the results shown by averaging the observed values re-scaled in relationship to the maximum value obtained within each dataset (an average value of 1 would indicate that the FS method obtained the best robustness index for all datasets), obtaining the following values: ReliefF: 0.87179; IG: 0.78078; mRMR: 0.64941; GA: 0.61396; SFS: 0.48686; CFS: 0.45904; SVM-RFE: 0.43212; Cons: 0.30266.

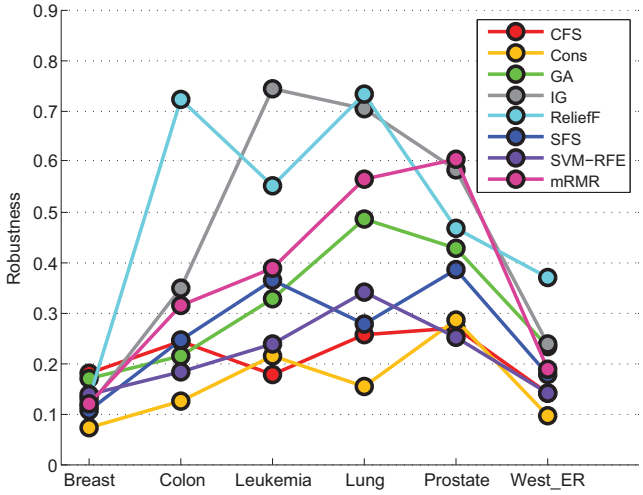


Figure 5: Robustness measure for each FS procedure among the different resamplings of each dataset, computed taking into account the variability of the selected genes.

5.4. Number of selected genes

The number of genes obtained by the different FS procedures studied varies depending on several factors. According to the raw data results shown in Tables A.1 and A.2, it can be appreciated that the SFS, GA, and Cons procedures are more aggressive in the gene-selection procedure, as few genes are kept in the final solutions. In the case of the SFS procedure, this could be explained by the nature of the algorithm, as it begins from solutions with only one gene and then iteratively adds new genes, while the performance is statistically significantly better than in the previous iteration. In a similar way, the GA procedure includes in its fitness function the desired characteristics of the solutions, maximizing the accuracy result while at the same time keeping those configurations with a smaller number of genes.

In comparison to wrapper methods (SFS and GA), filter and embedded methods are independent from the classification model. As these methods usually retain many more genes in their solutions, we established the following cut-off criteria: if the solution has more genes than the number of samples available in the dataset, then only the first $\#samples/8$ genes are kept (genes are sorted according to their suitability). Thus, there are some cases of FS procedures (i.e., SVM-RFE, mRMR, ReliefF, ...) where the number of selected genes is constant for all resamplings (the standard deviation is equal to zero in these cases). This criterion was set, firstly in order to reduce the number of selected genes per resampling and secondly in order to apply similar criteria over all filter and embedded methods so that a fair comparison could be made.

5.5. Biological analysis

We are interested in getting good results not only in prognosis prediction but also in examining whether the selected genes provide biological information related to the diseases studied. Therefore, if the proposed models provide this consistency between the computational and biological fields, we could have more confidence in the results and the selected genes would be more reliable from a clinical perspective. In this section the analysis of the selected genes for the Prostate dataset is carried out.

Table 4 presents the 15 most frequently selected genes among all the feature-selection techniques analysed. The top bar graph splits the frequency of selection (the fifth column of the table) of each gene according to the CFS, Cons, IG, mRMR, ReliefF, GA, SFS, and SVM-RFE strategies. The first four columns show information about the gene, such as the rank, the internal index (ID), the gene symbol (name of the gene), and the probe set ID, which is related to the chip from which the dataset has been extracted (e.g., Affymetrix). Most of the gene symbols have been found from their probe set ID by using tools such as IPA[3] or NCBI[4].

A higher frequency of selection might imply a higher relevance of the gene in the prognosis of the disease. Those genes that are selected with similar frequencies for all classifiers are considered independent with respect to the classification method. For instance, the HPN gene is more significant than the MAF gene, since it has been selected more times and all the classifiers

[3]Data were analysed through the use of IPA (Ingenuity® Systems, `www.ingenuity.com`).

[4]http://www.ncbi.nlm.nih.gov/gene/

select it with the same frequency. Thus, feature selection methods such as CFS, Cons, and IG barely take into account the MAF gene, whereas for the SVM-RFE technique it is one of the main genes.

According to the biological behaviour of the genes, it is possible to find references in the literature for several of the most selected genes after applying these computational techniques. For example, the gene **HPN**, officially named hepsin, which encodes a type II transmembrane serine protease (most frequent, 76.33%) [60], **PTGDS** (fourth more frequent, 39.04%), which encodes the protein glutathione-independent prostaglandin D synthase, which catalyses the conversion of prostaglandin H2 (PGH2) to prostaglandin D2 (PGD2) [61], **MAF** (sixth most frequent, 27.75%), which encodes a protein related to DNA-binding [62], and **GSTP1**, belonging to the family of Glutathione S-transferases (GSTs) enzymes (twelfth most frequent, 19.04%) [63], are biologically related to the absence or presence of prostate cancer. This supports the idea that our computational approach is robust and consistent with the results obtained in biological studies.

6. Conclusion

In Figure 2 we presented a summary of the average accuracy results shown in Tables A.1-A.2 for all of the analysed datasets. First, the results of this study clearly indicate the presence of three less complex datasets (Leukaemia, Lung, and Prostate) for which, independently of the FS and classification method used, the accuracy is always larger than 84%. Moreover, the use of honest validation schemes leads to less overfitting in the feature selection (an overfitting index was computed by dividing the accuracy with the proposed

Rank	*ID*	*Symbol*	*Probe ID*	*Set*	*Freq.(%)*
1	6185	HPN	37639_at		76.33
2	10494	NELL2	32598_at		48.63
3	8965	HSPD1	37720_at		43.75
4	9172	PTGDS	38406_f_at		39.04
5	9850	CFD	40282_s_at		38.21
6	10234	MAF	41504_s_at		27.75
7	4365	Unknown	41468_at		26.08
8	8850	PDLIM5	37366_at		25.29
9	9034	LMO3	38028_at		22.25
10	9593	CDKN1C	39545_at		20.96
11	5890	PEX3	36864_at		20.13
12	7756	GSTP1	33396_at		19.04
13	6462	RBP1	38634_at		18.33
14	12153	Unknown	556_s_at		17.25
15	10956	TGFB3	1767_s_at		16.13

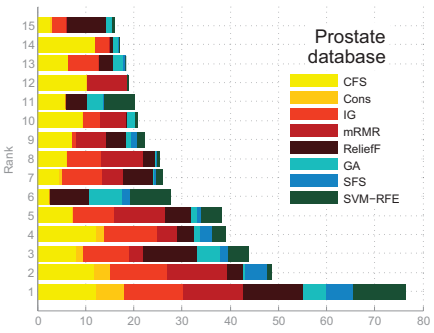


Table 4: The fifteen most selected genes for the Prostate dataset. Frequency selection is represented by a horizontal bar, divided according to the eight feature selection methods used in the analysis; CFS, Cons, IG, mRMR and ReliefF (filter methods); GA and SFS (wrapper methods); and SVM-RFE (embedded methods). The rank (also shown in the top figure as the y coordinate), index, gene symbol and probe set ID of each gene are shown in columns one to four.

honest validation scheme and without using an honest validation scheme and then averaged across the three selected datasets and the two classification algorithms analysed).

The results shown in Figures 3 and 4 suggest that the kNN and NN classifiers could be considered more robust methods independently of the FS method used and the dataset. Nevertheless, other classification techniques such as SVM or CM, which require the adjustment of several parameters, could lead to the achievement of similar results after a fine-tuning in the parameter estimation stage. Regarding the FS methods, the embedded

SVM-RFE and three other filter methods (IG, mRMR, and ReliefF) behave qualitatively better than the rest of the methods, indicating a superior performance in comparison to wrapper methods. Further, taking into account that wrapper methods tend to be more computationally intensive, the previous results clearly suggest an advantage of filtering (or embedded) FS schemes. In relation to the number of selected genes, SFS and Cons lead to more restricted sets, but with the disadvantage of worse performance, indicating that except when the size of the final set is a very important factor, these two FS procedures should not be the preferred option. Finally, by analysing references in the literature for the case of the prostate cancer dataset (see Section 5.5), the biological relevance of our analysis can be confirmed, as most of the selected genes have been mentioned as relevant regarding the absence or presence of the disease.

Finally, the overall conclusion of the present study is that, in general, filter methods are preferred over wrapper ones, as they lead to more robust results in terms of both percentages of better statistical significant results and overfitting effects, and are also less computationally intensive.

Acknowledgments

The authors acknowledge support through grants TIN2010-16556 from MICINN-SPAIN and P08-TIC-4026 (Junta de Andalucía), all of which include FEDER funds.

References

[1] J. S. Wei, B. T. Greer, F. Westermann, S. M. Steinberg, C.-G. Son, Prediction of Clinical Outcome Using Gene Expression Profiling and Artificial Neural Networks for Patients with Neuroblastom, Cancer Research 64 (2004) 6883 – 6891.

[2] J. Phan, A. Young, M. Wang, Robust microarray meta-analysis identifies differentially expressed genes for clinical prediction, The Scientific World Journal 2012.

[3] X. Wang, R. Simon, Microarray-based cancer prediction using single genes, BMC Bioinformatics 12 (2011) 391.

[4] A. Kulkarni, B. Naveen Kumar, V. Ravi, U. Murthy, Colon cancer prediction with genetics profiles using evolutionary techniques, ESWA 38 (3) (2011) 2752–2757.

[5] A. K. Gupta, U. Gupta, Chapter 19 - next generation sequencing and its applications, in: A. S. V. Singh (Ed.), Animal Biotechnology, Academic Press, San Diego, 2014, pp. 345 – 367.

[6] D. Slonim, I. Yanai, Getting started in gene expression microarray analysis, PLoS Computational Biology 5 (10) (2009) 0.

[7] S. Karimi, M. Farrokhnia, Leukemia and small round blue-cell tumor cancer detection using microarray gene expression data set: Combining data dimension reduction and variable selection technique, Chemometrics and Intelligent Laboratory Systems 139 (0) (2014) 6 – 14.

[8] S. Kar, K. D. Sharma, M. Maitra, Gene selection from microarray gene expression data for classification of cancer subgroups employing {PSO} and adaptive k-nearest neighborhood technique, Expert Systems with Applications 42 (1) (2015) 612 – 627.

[9] J. Nahar, T. Imam, K. S. Tickle, A. S. Ali, Y.-P. P. Chen, Computational intelligence for microarray data and biomedical image analysis for the early diagnosis of breast cancer, Expert Systems with Applications 39 (16) (2012) 12371 – 12377.

[10] C.-K. Chen, The classification of cancer stage microarray data, Computer Methods and Programs in Biomedicine 108 (3) (2012) 1070 – 1077.

[11] C. Orsenigo, C. Vercellis, A comparative study of nonlinear manifold learning methods for cancer microarray data classification, Expert Systems with Applications 40 (6) (2013) 2189 – 2197.

[12] M. Fabregue, S. Bringay, P. Poncelet, M. Teisseire, B. Orsetti, Mining microarray data to predict the histological grade of a breast cancer, Journal of Biomedical Informatics 44, Supplement 1 (0) (2011) S12 – S16.

[13] E. Lotfi, A. Keshavarz, Gene expression microarray classification using pca bel, Computers in Biology and Medicine 54 (0) (2014) 180 – 187.

[14] V. Lopes, A. Sims, Z. Sadiq, A. Saeed, Integration of multiple microarray datasets generating an eight gene signature as a validated biomarker for the detection of oral squamous cell carcinoma, British Journal of Oral and Maxillofacial Surgery 52 (8) (2014) e61 –.

[15] T. P. Stricker, A. M. L. Madrid, A. Chlenski, L. Guerrero, H. R. Salwen, Y. Gosiengfiao, E. J. Perlman, W. Furman, A. Bahrami, J. M. Shohet, P. E. Zage, M. J. Hicks, H. Shimada, R. Suganuma, J. R. Park, S. So, W. B. London, P. Pytel, K. H. Maclean, S. L. Cohn, Validation of a prognostic multi-gene signature in high-risk neuroblastoma using the high throughput digital nanostring ncounter system, Molecular Oncology 8 (3) (2014) 669 – 678.

[16] M. Shi, M.-S. Chen, K. Sekar, C.-K. Tan, L. L. Ooi, K. M. Hui, A blood-based three-gene signature for the non-invasive detection of early human hepatocellular carcinoma, European Journal of Cancer 50 (5) (2014) 928 – 936.

[17] H. Mirghani, N. Ugolin, C. Ory, M. Lefèvre, S. Baulande, P. Hofman, J. L. S. Guily, S. Chevillard, R. Lacave, A predictive transcriptomic signature of oropharyngeal cancer according to {HPV16} status exclusively, Oral Oncology 50 (11) (2014) 1025 – 1034.

[18] A. M. Molinaro, R. Simon, R. M. Pfeiffer, Prediction error estimation: a comparison of resampling methods, Bioinformatics 21 (2005) 3301–3307(7).

[19] S. Student, K. Fujarewicz, Stable feature selection and classification algorithms for multiclass microarray data, Biol Direct 7 (1) (2012) 33.

[20] D. Dittman, T. Khoshgoftaar, R. Wald, A. Napolitano, Similarity analysis of feature ranking techniques on imbalanced DNA microarray

datasets, Proceedings - 2012 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2012 (2012) 398–402.

[21] Y. Zhang, C. Ding, T. Li, Gene selection algorithm by combining relieff and mrmr, BMC Genomics 9 (SUPPL. 2).

[22] L. J. Lancashire, R. C. Rees, G. R. Ball, Identification of gene transcript signatures predictive for estrogen receptor and lymph node status using a stepwise forward selection artificial neural network modelling approach, Artif. Intell. Med. 43 (2) (2008) 99–111.

[23] H. Peng, Y. Fu, J. Liu, X. Fang, C. Jiang, Optimal gene subset selection using the modified SFFS algorithm for tumor classification, Neural Comput Appl (2012) 1–8.

[24] H. Yu, J. Ni, J. Zhao, ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data, Neurocomputing 101 (2013) 309–318.

[25] J. A. Castellanos-Garzón, F. Díaz, An evolutionary computational model applied to cluster analysis of DNA microarray data, ESWA 40 (7) (2013) 2575 – 2591.

[26] A. Sungheetha, J. Suganthi, An efficient clustering-classification method in an information gain NRGA-KNN algorithm for feature election of micro array data, Life Science Journal 10 (SUPPL. 7) (2013) 691–700.

[27] E. Keedwell, A. Narayanan, Gene expression rule discovery and multi-objective ROC analysis using a neural-genetic hybrid, Int J Data Min Bioin 7 (4) (2013) 376–396.

[28] S. Gupta, S. Garg, Multiobjective optimization using genetic algorithm, Advances in Chemical Engineering 43 (2013) 206–245.

[29] M. Pirooznia, J. Yang, M. Qu, Y. Deng, A comparative study of different machine learning methods on microarray gene expression data, BMC Genomics 9 (SUPPL. 1) (2008) S1–S13.

[30] M. Zervakis, M. Blazadonakis, G. Tsiliki, V. Danilatou, M. Tsiknakis, D. Kafetzopoulos, Outcome prediction based on microarray analysis: A critical perspective on methods, BMC Bioinformatics 10 (2009) 53.

[31] M. Wu, D. Dai, Y. Shi, H. Yan, X. Zhang, Biomarker identification and cancer classification based on microarray data using laplace naive bayes model with mean shrinkage., IEEE/ACM T Comput Bi 9 (6) (2012) 1649–1662.

[32] L. Wang, X. Wei, C. Cao, X. Li, Microarray data classification under the active learning framework of naive bayes, Advanced Science Letters 7 (2012) 496–500.

[33] S. Li, E. Harner, D. Adjeroh, Random kNN feature selection - a fast and stable alternative to random forests, BMC Bioinformatics 12 (1) (2011) 450.

[34] M. Dhawan, S. Selvaraja, Z.-H. Duan, Application of committee kNN classifiers for gene expression profile classification, Int J Bioinformatics Res Appl 6 (4) (2010) 344–352.

[35] Q. Han, G. Dong, Using attribute behavior diversity to build accurate decision tree committees for microarray data, J. Bioinformatics and Computational Biology 10 (4) (2012) 0.

[36] Z. Yang, Z. Yang, Y. Wang, A novel gene selection algorithm based on binary decision treea, International Journal of Digital Content Technology and its Applications 6 (11) (2012) 237–246.

[37] M. Burton, M. Thomassen, Q. Tan, T. Kruse, Gene expression profiles for predicting metastasis in breast cancer: A cross-study comparison of classification methods, The Scientific World Journal 2012.

[38] S. Dudoit, J. Fridlyand, Statistical Analysis of Gene Expression Microarray Data, Chapman and Hall, 2003.

[39] C. Ambroise, G. J. McLachlan, Selection bias in gene extraction on the basis of microarray gene-expression data, Proceedings of the National Academy of Sciences 99 (10) (2002) 6562–6566.

[40] A. R. Statnikov, C. F. Aliferis, I. Tsamardinos, D. P. Hardin, S. Levy, A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis., Bioinformatics 21 (5) (2005) 631–643.

[41] R. L. Somorjai, B. Dolenko, R. Baumgartner, Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions, Bioinformatics 19(12) (2003) 1484–1491.

[42] W. J. Fu, R. J. Carroll, S. Wang, Estimating misclassification error

with small samples via bootstrap cross-validation, Bioinformatics 21 (9) (2005) 1979–1986.

[43] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (19) (2007) 2507–2517.

[44] T. Jirapech-Umpai, S. Aitken, Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes, BMC Bioinformatics 6 (1) (2005) 148.

[45] A. R. Webb, Statistical Pattern Recognition, 2nd Edition, 3rd Edition, John Wiley & Sons, 2011.

[46] R. M. Luque-Baena, D. Urda, J. Subirats, L. Franco, J. M. Jerez, Application of genetic algorithms and constructive neural networks for the analysis of microarray cancer data, Theoretical Biology and Medical Modelling 2014 11(Suppl 1):S7.

[47] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.

[48] B. Guo, M. Nixon, Gait feature subset selection by mutual information, IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 39 (1) (2009) 36 –46.

[49] R. Moddemeijer, On estimation of entropy and mutual information of continuous distributions, Signal Processing 16 (3) (1989) 233 – 246.

[50] M. A. Hall, Correlation-based feature selection for machine learning, Tech. rep., University of Waikato, Hamilton, New Zealand (1998).

[51] M. Dash, H. Liu, Consistency-based search in feature selection, Artif. Intell. 151 (1-2) (2003) 155–176.

[52] J. R. Quinlan, Induction of decision trees, Machine learning 1 (1) (1986) 81–106.

[53] I. Kononenko, Estimating attributes: Analysis and extensions of relief, 1994, pp. 171–182.

[54] K. Kira, L. A. Rendell, A practical approach to feature selection, in: Proceedings of the ninth international workshop on Machine learning, ML92, 1992, pp. 249–256.

[55] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (1-3) (2002) 389–422.

[56] W. Pan, Bootstrapping likelihood for model selection with small samples (1998).

[57] B. L. Welch, The generalization of student's problem when several different population variances are involved, Biometrika 34 (1-2) (1947) 28–35.

[58] J. L. Subirats, L. Franco, J. M. Jerez, C-mantec: A novel constructive neural network algorithm incorporating competition between neurons, Neural Networks 26 (2012) 130 – 140.

[59] B. Efron, R. Tibshirani, Improvements on cross-validation: The .632+ bootstrap method, Journal of the American Statistical Association 92 (438) (1997) pp. 548–560.

[60] V. Srikantan, M. Valladares, J. S. Rhim, J. W. Moul, S. Srivastava, HEPSIN Inhibits Cell Growth/Invasion in Prostate Cancer Cells, Cancer Research 62 (23) (2002) 6812 – 6816.

[61] S. Ashida, H. Nakagawa, T. Katagiri, M. Furihata, M. Iiizumi, Y. Anazawa, T. Tsunoda, R. Takata, K. Kasahara, T. Miki, T. Fujioka, T. Shuin, Y. Nakamura, Molecular Features of the Transition from Prostatic Intraepithelial Neoplasia (PIN) to Prostate Cancer: Genome-wide Gene-expression Profiles of Prostate Cancers and PINs., Cancer Res 64 (17) (2004) 5963–5972.

[62] V. E. Steele, J. T. Arnold, H. Le, G. Izmirlian, M. R. Blackman, Comparative Effects of DHEA and DHT on Gene Expression in Human LNCaP Prostate Cancer Cells, Anticancer Res 26 (5A) (2006) 3205–3215.

[63] X. Lin, M. Tascilar, W.-H. Lee, W. J. Vles, B. H. Lee, R. Veeraswamy, K. Asgari, D. Freije, B. van Rees, W. R. Gage, G. S. Bova, W. B. Isaacs, J. D. Brooks, T. L. DeWeese, A. M. D. Marzo, W. G. Nelson, GSTP1 cpg island hypermethylation is responsible for the absence of GSTP1 expression in human prostate cancer cells, The American Journal of Pathology 159 (5) (2001) 1815 – 1826.

Appendix A. Supplementary tables

		West_ER		Breast		Leukaemia	
	Classifier	*accuracy*	*#genes*	*accuracy*	*#genes*	*accuracy*	*#genes*
SFS	LDA	73.29±9.23	2.96±0.83	63.35±7.00	3.64±1.55	**92.90±4.42**	2.56±0.64
	SVM	77.78±7.10	2.52±0.61	**64.27±6.87**	3.38±0.90	91.50±6.86	2.18±0.72
	kNN	77.86±10.56	2.40±0.81	62.56±7.57	2.22±1.20	91.95±5.67	2.16±0.79
	NB	77.51±9.85	2.70±0.81	62.10±5.85	3.68±1.52	92.85±4.09	2.32±0.59
	CM	78.01±7.72	2.08±0.70	64.19±6.85	2.46±1.11	92.01±5.10	1.96±0.45
	NN	**78.12±9.05**	2.48±0.71	63.57±7.67	3.62±1.07	92.18±5.69	2.36±0.66
GA	LDA	81.16±9.34	4.34±1.08	66.23±7.74	9.92±1.75	92.05±4.58	3.88±1.02
	SVM	79.58±8.83	4.14±1.01	66.72±8.54	9.44±3.00	89.44±7.25	3.38±0.97
	kNN	79.15±9.93	4.44±1.25	66.62±10.36	12.62±3.53	**92.47±6.00**	4.04±1.18
	NB	**82.60±8.84**	4.42±0.97	**67.10±7.29**	10.72±2.73	91.30±7.17	3.60±1.09
	CM	80.96±8.67	3.96±1.24	65.30±8.58	10.44±2.28	90.45±7.43	3.38±0.85
	NN	79.21±10.27	3.96±0.95	63.77±8.05	7.76±1.29	91.55±5.52	3.52±0.81
CFS	LDA	77.68±9.10	8.88±10.96	62.60±7.47	10.28±9.05	91.65±6.33	30.02±23.33
	SVM	78.42±10.12		63.96±6.11		92.22±5.83	
	kNN	77.05±11.49		63.69±6.36		93.08±5.18	
	NB	72.61±10.94		**69.14±7.82 ***		**93.90±6.38**	
	CM	**79.06±8.06**		63.60±5.76		92.78±4.31	
	NN	78.99±8.53		62.59±6.09		93.52±5.13	
Cons	LDA	72.54±8.46	2.12±0.44	63.33±6.17	3.36±0.56	86.44±7.58	1.84±0.51
	SVM	72.94±9.49		63.81±7.18		89.81±5.09	
	kNN	**74.83±9.30**		**65.27±8.39**		**90.62±4.78**	
	NB	71.69±8.38		61.13±7.72		89.43±5.08	
	CM	73.86±8.61		61.86±6.13		89.58±5.59	
	NN	74.67±8.50		63.08±6.21		89.87±5.12	
IG	LDA	84.91±6.41	6.00±0.00	65.01±7.03	9.00±0.00	93.38±3.14	9.00±0.00
	SVM	81.74±8.86		62.46±7.86		95.44±3.13	
	kNN	85.83±5.75		64.45±7.40		95.74±3.44	
	NB	**85.98±6.76 ***		**66.93±7.27**		**97.36±2.10**	
	CM	83.68±5.86		63.32±6.81		96.13±2.88	
	NN	84.11±6.92		64.28±7.21		96.51±2.71	
mRMR	LDA	82.54±7.56	6.00±0.00	64.83±7.33	9.00±0.00	94.88±3.29	9.00±0.00
	SVM	82.67±8.24		63.33±6.93		94.84±2.98	
	kNN	83.10±7.62		66.28±7.62		96.12±2.99	
	NB	79.82±8.86		63.78±6.97		**97.45±2.77 ***	
	CM	83.19±6.62		**63.40±5.42**		95.91±3.15	
	NN	**83.74±7.86**		62.68±6.03		96.81±2.44	
ReliefF	LDA	82.93±7.11	6.00±0.00	62.94±7.46	9.00±0.00	93.78±3.20	9.00±0.00
	SVM	81.42±7.86		62.73±8.14		93.27±3.92	
	kNN	83.95±8.35		**61.53±7.01**		94.34±3.51	
	NB	**84.40±6.30**		62.65±7.37		**96.95±2.91**	
	CM	82.26±7.37		63.68±6.41		94.50±3.27	
	NN	82.82±8.27		63.39±7.57		94.71±3.18	
SVM-RFE	LDA	82.13±7.36	6.00±0.00	66.72±9.49	9.00±0.00	95.22±3.89	9.00±0.00
	SVM	81.73±8.44		66.66±9.71		94.62±3.97	
	kNN	**83.10±7.81**		67.16±7.31		96.14±3.33	
	NB	81.39±6.54		63.23±6.87		**96.22±2.97**	
	CM	81.13±6.52		66.10±7.00		94.19±3.28	
	NN	81.86±8.00		**67.47±6.89**		95.49±3.44	

Table A.1: Performance comparison obtained for the eight analysed feature selection procedures (SFS, GA, CFS, Cons, IG, mRMR, ReliefF and SVM-RFE) and six classifiers (LDA, SVM, kNN, NB, CM and NN) for three cancer microarray datasets (West_ER, Breast and Leukaemia). The results shown correspond to the accuracy of each classification method and to the number of selected genes (*mean±standard deviation*).

		Lung		Colon		Prostate	
	Classifier	*accuracy*	*#genes*	*accuracy*	*#genes*	*accuracy*	*#genes*
SFS	LDA	96.45±2.11	3.00±0.83	**81.69±5.72**	2.42±0.97	88.96±4.85	3.54±1.05
	SVM	96.70±2.04	2.34±0.56	78.39±7.83	2.60±0.95	88.11±4.51	3.20±0.70
	kNN	97.30±1.60	2.20±0.57	74.68±9.24	2.06±0.98	87.48±5.60	2.38±0.75
	NB	96.92±2.21	2.30±0.81	80.45±6.18	2.04±0.92	89.12±4.13	3.08±0.85
	CM	96.38±2.16	2.14±0.45	76.39±8.30	1.86±0.78	86.67±6.78	2.84±0.82
	NN	**97.47±1.67**	2.26±0.53	80.25±8.91	2.52±0.97	**89.78±4.32**	3.30±0.91
GA	LDA	97.11±2.11	3.62±1.03	**83.70±7.09**	7.28±2.55	89.98±4.77	5.90±1.63
	SVM	96.52±2.32	2.52±0.65	78.61±8.08	8.60±3.14	89.92±5.12	6.04±1.75
	kNN	97.70±1.55	2.44±0.54	81.27±7.21	9.12±4.15	90.48±4.96	6.64±2.10
	NB	97.37±1.73	2.88±0.80	81.40±8.37	7.98±2.64	90.48±4.39	5.36±1.63
	CM	97.07±1.90	2.44±0.61	**81.56±7.36**	7.70±2.62	**91.44±4.44**	5.48±1.55
	NN	**97.75±1.86**	2.48±0.61	80.02±8.50	6.64±2.20	90.75±5.24	5.86±1.43
CFS	LDA	99.40±0.82	96.66±22.97	64.03±5.74	8.08±7.64	92.50±3.65	76.34±8.87
	SVM	99.50±0.63		66.04±8.86		93.36±3.09	
	kNN	**99.61±0.56**		**76.23±8.56**		93.27±3.37	
	NB	99.55±0.56		62.70±5.36		93.21±3.24	
	CM	99.16±0.50		69.87±8.26		**93.40±3.36**	
	NN	99.02±1.00		73.12±7.40		92.02±3.69	
Cons	LDA	94.37±2.68	1.84±0.42	79.63±7.39	3.28±0.76	82.99±6.03	3.20±0.67
	SVM	95.61±1.95		76.07±8.18		85.33±5.64	
	kNN	96.29±1.54		77.97±9.17		**86.77±5.71**	
	NB	95.66±1.83		77.57±6.66		84.24±7.07	
	CM	95.88±1.62		79.60±6.98		85.14±5.58	
	NN	**96.39±1.34**		**80.04±7.11**		85.48±4.94	
IG	LDA	98.74±1.07	22.00±0.00	81.03±8.62	9.58±9.94	92.42±2.88	12.00±0.00
	SVM	99.26±0.95		77.30±8.89		89.87±3.95	
	kNN	99.53±0.60		80.59±7.78		**92.88±3.27**	
	NB	99.41±0.59		**82.27±9.49**		92.34±3.63	
	CM	99.49±0.52		80.64±7.27		91.01±3.24	
	NN	**99.57±0.57**		80.96±6.44		92.46±2.85	
mRMR	LDA	98.86±0.91	22.00±0.00	84.07±5.24	7.00±0.00	92.05±2.58	12.00±0.00
	SVM	99.40±0.82		80.96±8.04		90.08±4.29	
	kNN	**99.63±0.55** *		84.28±7.54		**93.69±3.30**	
	NB	99.61±0.55		81.45±6.73		93.16±3.30	
	CM	99.24±0.60		**84.58±5.34** *		91.00±2.95	
	NN	99.56±0.54		84.21±5.62		92.20±2.91	
ReliefF	LDA	97.91±1.27	22.00±0.00	83.11±5.14	7.00±0.00	91.20±3.43	12.00±0.00
	SVM	98.99±0.96		77.30±8.40		91.90±4.04	
	kNN	99.13±1.07		81.86±9.16		93.34±3.18	
	NB	99.24±0.86		**83.70±6.51**		90.90±3.67	
	CM	**99.33±0.53**		80.26±5.96		92.63±4.86	
	NN	98.53±0.79		80.60±6.62		**93.51±3.78**	
SVM-RFE	LDA	99.28±0.80	22.00±0.00	82.45±6.97	7.00±0.00	92.85±3.56	12.00±0.00
	SVM	98.75±0.95		79.69±8.11		91.65±4.12	
	kNN	**99.35±0.64**		82.47±6.80		**93.81±3.29** *	
	NB	99.23±0.78		79.45±6.65		90.88±3.52	
	CM	98.97±0.62		82.76±6.69		92.12±3.65	
	NN	99.05±0.77		**83.51±7.09**		93.05±3.60	

Table A.2: Performance comparison obtained for the eight analysed feature selection procedures (SFS, GA, CFS, Cons, IG, mRMR, ReliefF and SVM-RFE) and six classifiers (LDA, SVM, kNN, NB, CM and NN) for three cancer microarray datasets (Lung, Colon and Prostate). The results shown correspond to the accuracy of each classification method and to the number of selected genes (*mean±standard deviation*).

Classifier	Database	p-value	Control	Statistically different FS procedures
LDA	West_ER	0	Cons	GA, SVM-RFE, ReliefF, mRMR, IG
	Breast	0.0058	ReliefF	SVM-RFE
	Leukaemia	$6.6613e^{-16}$	Cons	GA, CFS, SFS, IG, ReliefF, SVM-RFE, mRMR
	Lung	0	Cons	GA, ReliefF, IG, mRMR, SVM-RFE, CFS
	Colon	0	CFS	Cons, SFS, IG, SVM-RFE, ReliefF, GA, mRMR
	Prostate	0	Cons	SFS, GA, ReliefF, mRMR, IG, SVM-RFE, CFS
SVM	West_ER	$5.1310e^{-9}$	Cons	CFS, GA, IG, ReliefF, SVM-RFE, mRMR
	Breast	0.0722	-	-
	Leukaemia	$1.2568e^{-13}$	Cons	CFS, ReliefF, SVM-RFE, mRMR, IG
	Lung	0	Cons	SVM-RFE, ReliefF, IG, CFS, mRMR
	Colon	$1.9592e^{-12}$	CFS	Cons, ReliefF, SFS, IG, GA, SVM-RFE, mRMR
	Prostate	0	Cons	mRMR, GA, IG, ReliefF, SVM-RFE, CFS
kNN	West_ER	$9.1707e^{-12}$	Cons	SVM-RFE, mRMR, ReliefF, IG
	Breast	$4.5337e^{-4}$	ReliefF	GA, mRMR, SVM-RFE
	Leukaemia	$1.4988e^{-14}$	Cons	CFS, ReliefF, SVM-RFE, mRMR, IG
	Lung	0	Cons	ReliefF, SVM-RFE, CFS, IG, mRMR
	Colon	$9.1788e^{-8}$	SFS	IG, ReliefF, GA, SVM-RFE, mRMR
	Prostate	0	Cons	GA, IG, ReliefF, CFS, SVM-RFE, mRMR
NB	West_ER	0	Cons	mRMR, SVM-RFE, GA, ReliefF, IG
	Breast	$5.0040e^{-10}$	Cons	GA, IG, CFS
	Leukaemia	0	Cons	CFS, SVM-RFE, IG, ReliefF, mRMR
	Lung	0	Cons	SVM-RFE, ReliefF, IG, CFS, mRMR
	Colon	0	CFS	Cons, SVM-RFE, SFS, mRMR, ReliefF, GA, IG
	Prostate	0	Cons	SFS, GA, SVM-RFE, ReliefF, IG, CFS, mRMR
CM	West_ER	$6.2372e^{-11}$	Cons	GA, SVM-RFE, ReliefF, IG, mRMR
	Breast	0.2749	-	-
	Leukaemia	0	Cons	CFS, ReliefF, SVM-RFE, mRMR, IG
	Lung	0	Cons	SVM-RFE, CFS, mRMR, ReliefF, IG
	Colon	0	CFS	SFS, Cons, ReliefF, IG, GA, SVM-RFE, mRMR
	Prostate	0	Cons	mRMR, IG, GA, SVM-RFE, CFS, ReliefF
NN	West_ER	$1.0246e^{-9}$	Cons	GA, SVM-RFE, ReliefF, mRMR, IG
	Breast	$1.5604e^{-4}$	mRMR	SVM-RFE
	Leukaemia	0	Cons	ReliefF, CFS, SVM-RFE, IG, mRMR
	Lung	0	Cons	GA, ReliefF, CFS, SVM-RFE, IG, mRMR
	Colon	$1.3967e^{-13}$	CFS	ReliefF, GA, Cons, SFS, IG, SVM-RFE, mRMR
	Prostate	0	Cons	SFS, GA, mRMR, IG, CFS, SVM-RFE, ReliefF

Table A.3: Differences between feature selection algorithms for the six different classification methods used (first column). The lowest performant FS procedure is taken as control group (fourth column) while the last column of the table lists the procedures that lead to statistically significant results (corresponding p-value indicated in the third column)

FS procedure	Database	p-value	Control	Statistically different classifiers
SFS	West_ER	0.0144	LDA	kNN, NN
	Breast	0.3595	-	-
	Leukaemia	0.9650	-	-
	Lung	$6.3446e^{-4}$	LDA	NN, kNN
	Colon	$1.1155e^{-4}$	kNN	LDA, NN
	Prostate	0.0056	CM	NN
GA	West_ER	0.4007	-	-
	Breast	0.0924	-	-
	Leukaemia	0.1125	-	-
	Lung	0.0046	SVM	NN, kNN
	Colon	0.0068	SVM	LDA
	Prostate	0.0897	-	-
CFS	West_ER	$7.0441e^{-4}$	NB	kNN, CM, NN, SVM
	Breast	$1.3860e^{-5}$	LDA	NB
	Leukaemia	$3.4700e^{-5}$	LDA	NN, NB
	Lung	$2.2204e^{-16}$	NN	LDA, SVM, NB, kNN
	Colon	0	NB	CM, NN, kNN
	Prostate	$7.0186e^{-5}$	NN	NB, CM, SVM, kNN
Cons	West_ER	0.0037	NB	kNN
	Breast	$1.8876e^{-4}$	NB	kNN
	Leukaemia	$1.6352e^{-7}$	LDA	SVM, CM, NN, NB, kNN
	Lung	$9.5437e^{-12}$	LDA	SVM, CM, NB, NN, kNN
	Colon	0.0015	SVM	CM, NN
	Prostate	$1.9109e^{-5}$	LDA	CM, NN, SVM, kNN
IG	West_ER	0.0236	CM	kNN
	Breast	0.0073	SVM	NB
	Leukaemia	$9.2215e^{-13}$	LDA	SVM, kNN, CM, NN, NB
	Lung	0	LDA	NB, CM, NN, SVM, kNN
	Colon	$7.4769e^{-4}$	SVM	LDA, NN, NB
	Prostate	$1.4814e^{-4}$	SVM	LDA, NB, NN, kNN
mRMR	West_ER	0.0105	NB	SVM, NN
	Breast	0.2202	-	-
	Leukaemia	$1.2065e^{-11}$	LDA	CM, kNN, NN, NB
	Lung	0	LDA	NN, NB, SVM, kNN
	Colon	0.0117	SVM	CM
	Prostate	$8.4826e^{-9}$	CM	NB, kNN
ReliefF	West_ER	0.0678	-	-
	Breast	0.7720	-	-
	Leukaemia	$1.4892e^{-8}$	LDA	NB
	Lung	0	LDA	SVM, NB, CM, kNN
	Colon	$1.2489e^{-8}$	SVM	LDA, kNN, NB
	Prostate	$2.4679e^{-7}$	LDA	kNN, NN
SVM-RFE	West_ER	0.0103	CM	kNN
	Breast	0.0012	NB	SVM, LDA, kNN, NN
	Leukaemia	$4.9617e^{-8}$	CM	NN, NB, kNN
	Lung	$1.1538e^{-9}$	CM	LDA, NB, kNN
	Colon	$5.1422e^{-4}$	NB	NN
	Prostate	$1.9036e^{-6}$	NB	LDA, NN, kNN

Table A.4: Differences between classifiers for the eight different feature selection (FS) procedures used (first column). Best techniques are indicated as in Table A.3.

.

Capítulo 6

Robustez y relevancia biológica de las firmas genéticas

Resumen: Los algoritmos evolutivos son una familia de algoritmos muy utilizados para estimar firmas genéticas a partir de datos de perfiles de expresión. El principal inconveniente que presentan estos algoritmos es la escasa robustez que ofrecen las soluciones finales, de manera que cada ejecución produce una firma genética diferente. Este tipo de soluciones resultan poco útiles de cara a utilizar estas firmas genéticas en estudios clínicos o biomédicos. En este capítulo se propone una nueva estrategia en dos etapas para seleccionar características utilizando un algoritmo genético que incluye en su función de fitness información biológica extraída de la base de datos KEGG. Para evaluar el rendimiento de esta estrategia, se realizó un estudio comparativo sobre tres bases de datos de distintos tipos de cáncer (leucemia, pulmón y próstata), comparando los resultados obtenidos con los que se producen tras la ejecución de un algoritmo genético tradicional. Los resultados alcanzados muestran que esta nueva estrategia mejora la consistencia, robustez y predicción de las firmas genéticas obtenidas. Por ello, este nuevo enfoque ofrece la facilidad de definir firmas genéticas que puedan ser utilizadas en el futuro para diagnosticar y predecir la evolución clínica de pacientes con cáncer.

Título: Robust gene signatures from microarray data using genetic algorithms enriched with biological pathway keywords
Autores: Luque-Baena, R.M., Urda, D., Gonzalo-Claros, M., Franco, L. and Jerez J.M.
Revista: Journal of Biomedical Informatics
Volumen: 49
Páginas: 32-44
Año: 2014
DOI: 10.1016/j.jbi.2014.01.006
Abstract: Genetic algorithms are widely used in the estimation of expression profiles from microarrays data. However, these techniques are unable to produce stable and robust solutions suitable to use in clinical and biomedical studies. This paper presents a novel two-stage evolutionary strategy for gene feature selection combining the genetic algorithm with biological information extracted from the KEGG database. A comparative study is carried out over public data from three different types of cancer (leukemia, lung cancer and prostate cancer). Even though the analyses only use features having KEGG information, the results demonstrate that this two-stage evolutionary strategy increased the consistency, robustness and accuracy of a blind discrimination among relapsed and healthy individuals. Therefore, this approach could facilitate the definition of gene signatures for the clinical prognosis and diagnostic of cancer diseases in a near future. Additionally, it could also be used for biological knowledge discovery about the studied disease.

Capítulo 7

Conclusiones y trabajo futuro

Este capítulo final presenta las conclusiones derivadas del trabajo realizado en esta tesis doctoral, así como las posibles líneas futuras de investigación a seguir. Para ello, se expone primeramente un breve resumen de los contenidos más relevantes de cada capítulo introducido a lo largo de la memoria para luego sacar las conclusiones pertinentes finales.

El Capítulo 2 mostró un análisis y desarrollo de un sistema público y gratuito llamado WIMP que incorpora distintos métodos estándar de imputación de datos perdidos. A lo largo del capítulo, se contempló el problema de los conjuntos de datos con valores perdidos, que es muy importante en conjuntos de datos biomédicos y, en particular, en datos de perfiles de expresión debido al alto costo de los experimentos y al escaso número de muestras disponibles.

En el Capítulo 3 se describió la experiencia vivida en el desarrollo de un sistema de información oncológico implantado en el HUVV de Málaga que permite almacenar la información clínico-histopatológica de los pacientes e integra modelos predictivos basados en este tipo de datos que facilite a los clínicos diagnosticar y predecir la evolución de un paciente. El desarrollo del capítulo contempló los principales motivos que hacen que estos sistemas no tengan éxito en cuanto a uso y satisfacción de los usuarios y abordó los aspectos más críticos que hay que llevar a cabo para lograr implantar un sistema de estas características en un servicio de oncología de un hospital.

Con objeto de acercar a la práctica clínica diaria una medicina personalizada, los Capítulos 4, 5 y 6 mostraron los problemas que surgen en la estimación de modelos predictivos basados en información proteómico-genómica. El Capítulo 4 mostró una novedosa estrategia de selección de características basada en un algoritmo genético (GA) modificado al incluir dentro de la función de fitness del GA un término correspondiente a la información mutua. Además, se utilizó un modelo de red neuronal constructivo, C-Mantec, como algoritmo de clasificación. A continuación, el Capítulo 5 mostró una comparativa de los distintos métodos estándar de selección de característi-

cas existentes en la literatura (wrappers, de filtrado y embebidos) utilizando para ellos diversos algoritmos de clasificación comúnmente conocidos (LDA, SVM, NB, C-Mantec, kNN y MLP). Finalmente, el Capítulo 6 mostró una novedosa estrategia en dos etapas para buscar firmas genéticas con altas capacidades predictivas, robustas y con relevancia biológica en la enfermedad de estudio. El método desarrollado es un GA que incorpora en el propio de proceso de selección de características información biológica relativa a la enfermedad basándose en la información contenida en la base de datos pública KEGG.

WIMP es una herramienta web pública y gratuita desarrollada para poder ser utilizada por los usuarios registrados en ella. Esta herramienta se desarrolló dado que el pre-procesado de un conjunto de datos es un paso previo necesario a cualquier tipo de análisis que se desee realizar sobre él. Al trabajar con conjuntos de datos biomédicos reales, es muy frecuente encontrarse con muestras incompletas que impiden comenzar el análisis sin realizar antes un pre-procesamiento de los datos. En este sentido, WIMP ofrece a la comunidad investigadora la posibilidad de imputar valores perdidos en los conjuntos de datos que cada usuario suba al sistema y elegir entre varios métodos de imputación que se incorporan en él en su versión inicial. Estos métodos ya están implementados y previamente testados, es decir, el usuario no debe preocuparse de buscar ningún método de imputación o bien de pagar por un software comercial que les ofrezca esta posibilidad. De entre los diferentes métodos de imputación de datos existentes, WIMP incorpora algunos basados en técnicas estadísticas (imputación por la media, imputación hot-deck e imputación múltiple) y otros basados en algoritmos de aprendizaje (imputación con mapas auto-organizados e imputación con un perceptrón multicapa). El número de métodos que incorpora el sistema en su versión inicial no es algo cerrado, sino que puede ampliarse en un futuro tan solo implementando otros métodos existentes, validarlos e incorporarlos en el núcleo del sistema.

Por otro lado, otra gran ventaja que ofrece WIMP a sus usuarios es la de evitarles la necesidad de disponer de una gran capacidad de cálculo computacional. Las tareas de imputación que los usuarios soliciten a través del sistema se procesan en el lado del servidor. Es más, el servidor está conectado a un clúster de computación, que es donde en realidad se ejecutan las tareas solicitadas por los usuarios del sistema. Esto posibilita que un simple usuario, desde su ordenador de casa, pueda ejecutar un método de imputación complejo y obtener así en poco tiempo el conjunto de datos completo tras la imputación realizada. Una vez que una tarea de imputación lanzada por un usuario finaliza, el servidor avisa al usuario a través del e-mail con un enlace de descarga que le permita recoger el conjunto de datos resultante de la ejecución. Este conjunto de datos está compuesto, ahora sí, por muestras completas y está disponible para realizar cualquier tipo de

análisis sobre él, alcanzado así el primer objetivo parcial de esta tesis.

En el Capítulo 3 se analizó las principales deficiencias existentes en los sistemas de información actuales de un servicio de oncología y se desarrolló un sistema de información oncológico que cubre todas sus necesidades y permite llevar a la práctica clínica diaria los modelos predictivos basados en la información clínico-histopatológica de cada paciente. En concreto, gracias a la estrecha colaboración existente entre el grupo de investigación ICB de la universidad de Málaga y el equipo médico del servicio de oncología del HUVV de Málaga, se realizó un análisis detallado de las necesidades existentes en el servicio de oncología clínica del hospital con objeto de implantar con éxito un sistema de estas características que incorpore, junto a la información clínico-histopatológica de cada paciente, modelos predictivos que ofrezcan al clínico mayor información para decidir acerca del diagnóstico, evolución y tratamiento a aplicar a un paciente. En este capítulo se analizó y aportó soluciones a los aspectos más críticos que facilitan alcanzar el éxito en la implantación del sistema como son: (i) utilizar un diseño centrado en el usuario para aumentar la usabilidad y facilidad de uso del sistema; (ii) usar la tecnología adecuada que facilite todo el desarrollo así como el mantenimiento y futuras ampliaciones del sistema; (iii) integrar en el sistema todas las rutinas clínicas diarias del servicio de oncología; (iv) seguridad e integración del sistema con otros sistemas de información del hospital.

El sistema final desarrollado e implantado con éxito en el HUVV incorpora toda la información clínico-histopatológica de cada paciente del servicio. El diseño del sistema contempla la distribución de la información en los siguientes módulos funcionales: gestión de pacientes, hospital de día, ensayos clínicos, consejo genético y análisis estadístico. La característica más importante que hace que este sistema destaque sobre otros es que, además de almacenar toda la información relativa a los pacientes del servicio, contiene un módulo específico para realizar diferentes estudios a partir de esa información. Este módulo integra varios modelos de pronóstico clínico fruto de la investigación del grupo ICB en trabajos anteriores, ofreciendo a los clínicos la posibilidad de realizar análisis de supervivencia implementado bajo el algoritmo de Kaplan-Meier, análisis de regresión de Cox, calcular funciones de riesgo y obtener tablas de contingencia en base a la información clínico-histopatológica de un subconjunto de pacientes previamente seleccionado. La evaluación del éxito en la implantación del sistema se llevó a cabo a partir de encuestas realizadas a los usuarios finales del mismo a los 3 y 15 meses tras su implantación en el hospital, mostrando una satisfacción generalizada del personal y logrando alcanzar así el segundo objetivo parcial de esta tesis doctoral.

Este sistema ha supuesto un gran paso hacia adelante en dirección a llevar a la práctica clínica diaria una medicina personalizada en un futuro no muy lejano. Sin embargo, la heterogeneidad del cáncer como enfermedad hace

que su evolución esté determinada por factores no sólo clínicos sino también genéticos, lo que lleva a plantearse la necesidad de integrar en un futuro la información proteómico-genómica de cada paciente dentro del sistema. Por ello, surge la necesidad de desarrollar nuevos modelos predictivos basados en datos de perfiles de expresión para incorporarlos a los ya existentes en el sistema. De esta forma, se daría la posibilidad a los clínicos de poder practicar una medicina personalizada a sus pacientes atendiendo a la información de ambos tipos y a los modelos predictivos incorporados en el módulo de análisis estadístico. En este sentido, el Capítulo 4 describió el uso de un GA utilizando C-Mantec como algoritmo de clasificación para la estimación de modelos predictivos con datos de perfiles de expresión. Las principales novedades que presentó el GA son el uso de un pre-filtrado de genes aplicando el test estadístico Welch's t-test y la inclusión en la función de fitness del GA de un término que busca minimizar la correlación entre variables.

La validación de esta nueva estrategia planteada se realizó comparando los resultados obtenidos con los de la estrategia SFS+CNN planteada en el trabajo preliminar presentado en el Apéndice A, incluyéndole además el pre-filtrado de genes acorde al resultado de aplicar el Welch's t-test (para realizar una comparación justa). Los resultados obtenidos son en general mejores en cuanto a predicción de la estrategia GA respecto a SFS. Por contra, la estrategia GA obtiene como resultado subconjuntos de genes más grandes y es computacionalmente más exigente que SFS al evaluar muchos más subconjuntos de genes. Respecto a la influencia del algoritmo de clasificación en la estrategia de selección, con MLP, LDA y SVM se obtuvieron los mejores resultados mientras que C-Mantec y kNN le siguen de cerca aunque con una predicción ligeramente menor. Además, C-Mantec, MLP y LDA dan lugar a firmas genéticas con menor número de genes en comparación con SVM, NB y kNN, y en particular C-Mantec es el clasificador más robusto en cuánto al ajuste de los parámetros del modelo.

El Capítulo 5 presentó una extensa comparativa entre diversos métodos de selección de características existentes utilizando varios algoritmos estándar de clasificación, todo bajo un esquema de validación honesto. En este sentido y dado que el número de muestras disponibles en los conjuntos de datos de perfiles de expresión es bastante limitado, se utilizó una técnica de *resampling* conocida como es BCV. Los resultados obtenidos tras ejecutar, bajo el mismo esquema de validación descrito, las diferentes estrategias surgidas de combinar los métodos de selección con los algoritmos de clasificación, muestran que los métodos de filtrado o embebidos proporcionan, en general, mayor predicción siendo a su vez menos complejos y más rápidos de ejecutar que los métodos wrapper. También los métodos de filtrado aparecen como los métodos más robustos en cuanto a las firmas genéticas obtenidas. Respecto a los algoritmos de clasificación, kNN y MLP son los algoritmos que pueden considerarse más robustos independientemente del método de

selección de características utilizado. Sin embargo, otros clasificadores como SVM o C-Mantec podrían alcanzar resultados similares con un ajuste óptimo de los respectivos parámetros de ambos modelos, una tarea nada fácil dada la variabilidad en el tamaño de las firmas genéticas que se evalúan en los métodos de selección de características y, en este sentido, C-Mantec se presentó en el Capítulo 4 como un clasificador muy robusto en el ajuste de parámetros.

El nuevo método de selección de características planteado en el Capítulo 6 trata de obtener firmas genéticas robustas, con alta capacidad predictiva y de relevancia biológica. Para ello, este nuevo método utilizó información biológica disponible en la base de datos KEGG asociada a la enfermedad de estudio. Las dos etapas del nuevo modelo propuesto consisten en que: (i) la primera etapa hace un filtrado inicial del número total de genes basado en la base de datos KEGG y establece un ranking de pathways acorde a la capacidad predictiva de los genes involucrados en él y a un proceso de *data mining* realizado sobre la descripción textual de cada pathway ofrecida en la web de KEGG. A continuación, (ii) la segunda etapa ejecuta un GA sobre los mejores pathways encontrados en la primera etapa para estimar un modelo predictivo. La principal novedad de este GA está en la inclusión de un tercer término en la función de fitness que trata de maximizar una función que puntúa la "calidad" de la firma en base al número de genes que ésta contiene y que además forman parte del pathway de estudio, o bien, de cualquiera de los otros pathways seleccionados en la primera etapa para analizar. La evaluación de este nuevo método se realizó utilizando tres conjuntos de datos de microarrays de ADN de distintos tipos de cáncer, donde los resultados alcanzados muestran que esta nueva estrategia mejora la consistencia, robustez y predicción de las firmas genéticas obtenidas. Esta novedosa estrategia mejoró los resultados obtenidos con un GA clásico que no incorpore información biológica en el proceso de selección, independientemente del clasificador utilizado. Por tanto, este nuevo enfoque, junto a los modelos presentados en los Capítulos 4 y 5, logra satisfacer el tercer objetivo parcial de esta tesis y ofrece la posibilidad de definir en un futuro cercano firmas genéticas que puedan ser utilizadas para predecir el diagnóstico y la evolución clínica de pacientes con cáncer.

Finalmente, como conclusión global de esta tesis doctoral se puede decir que los resultados obtenidos en ella permiten acercarse en un futuro a una medicina personalizada en la práctica clínica diaria. Los modelos predictivos basados en datos de perfiles de expresión que se han desarrollado en la tesis podrían integrarse en el propio sistema de información oncológico implantado en el HUVV. Además, se podría incorporar la información proteómico-genómica de cada paciente para poder aprovechar al máximo las ventajas añadidas mencionadas a lo largo de esta tesis. En la faceta personal, todo el trabajo realizado en esta tesis me ha permitido profundizar y

adquirir una extensa formación investigadora en un área tan amplia como es la Bioinformática.

7.1. Trabajo futuro

Dado que los resultados obtenidos en esta tesis doctoral permiten acercarse en un futuro a una medicina personaliza en la práctica clínica diaria, se abren nuevas vías de trabajo futuro y líneas de investigación dentro del propio grupo de investigación ICB como posible extensión a todo el trabajo realizado en esta tesis. Entre las diversas ideas que han surgido se podrían destacar las siguientes:

- Integración de la información proteómico-genómica de cada paciente y de los modelos predictivos basados en perfiles de expresión en el sistema desarrollado e implantado en el HUVV de Málaga (ver Capítulo 3), cuya versión 2.0 se prevé implantar en el Hospital Regional Virgen de la Victoria, que unifica varios centros hospitalarios de la provincia de Málaga. Además, analizar si la estimación de un modelo predictor basado tanto en información clínico-histopatológica como en información proteómico-genómica mejora la predicción de los modelos predictivos existentes que se basan en el uso de un solo tipo de datos. Esto sería un gran paso adelante para poder ofrecer una medicina personalizada ya que los clínicos tendrían a su disposición, en una misma herramienta, toda la información clínico-histopatológica y proteómico-genómica junto a una serie de modelos predictivos previamente estimados en base a toda la población de pacientes oncológicos contenidos en el sistema de información.

- Aplicar técnicas estadísticas para pre-filtrar el conjunto de características presente en datos de perfiles de expresión antes de estimar un modelo predictivo. En este sentido, sería interesante aplicar un t-test modificado y utilizar los *q-valores* para identificar los genes más significativos a la hora de diferenciar el estado o clase de las muestras. Además, se podría incluir esta información en la estrategia planteada en el Capítulo 6 de manera que este nuevo enfoque debería producir resultados más rigurosos, desde el punto de vista estadístico, y de mayor robustez pues se buscarían preferiblemente firmas genéticas compuestas por alguno de los genes identificados como significativos. En este sentido, el trabajo preliminar desarrollado en esta línea arroja unos resultados prometedores que incitan a profundizar más en esta dirección de cara a obtener firmas genéticas robustas y de altas capacidades predictivas.

- Validar los nuevos modelos propuestos y cualquier otro nuevo que se

pueda plantear en tecnologías más recientes como RNA-Seq, Next Generation Sequencing (NGS), y utilizar datos de microarrays de miRNA. Dada la estrecha colaboración que existe entre el grupo ICB con el departamento de Biología Molecular de la Universidad de Málaga y la Unidad de Oncología del HUVV, sería interesante plantear la validación de las firmas genéticas encontradas en estos u otros trabajos del grupo utilizando repositorios públicos de información de perfiles de expresión.

Capítulo 8

Conclusions and further work

This final chapter presents the conclusions from the work done in this thesis as well as possible future work based on this research line. To do this, a brief summary of the relevant contents of each chapter presented along this memory is introduced in order to draw the final conclusions.

Chapter 2 showed an analysis and development of a free public web tool called WIMP incorporating various standard missing data imputation methods. Throughout the chapter, the problem of data sets with missing values is exposed, which is very important in biomedical data sets contemplated and, in particular, using expression profiling data due to the high cost of experiments and the small number of samples available.

Chapter 3 described the experience in developing a OIS deployed in the HUVV in Malaga to store clinical and histopathologic patient information and that integrates predictive models based on these type of data, thus providing clinicians facilities to diagnose and predict the evolution of a patient. This chapter analyzed the main reasons that make these systems not to be successful in terms of use and user satisfaction and it addressed the most critical aspects that need to be taken into account in order to deploy a system of this characteristics in a oncology service of a hospital.

In order to practice a personalized medicine in the daily clinical practice in a near future, Chapter 4, 5 and 6 showed the problems that arise in the estimation of predictive models based in expression profiling data. Chapter 4 showed a novel feature selection strategy based on a modified genetic algorithm (GA) that includes within the fitness function of the GA a mutual information term. In addition, a constructive neural network model called C- Mantec was used as classification algorithm. Then, Chapter 5 showed a comparative of different standard feature selection methods in the literature (wrappers, filter and embedded) when using various commonly well known classification algorithms (LDA, SVM, NB, C-Mantec, kNN and MLP). Finally, Chapter 6 showed a novel two-stage strategy to find genetic signatures with high predictive capabilities, more robust and with biological relevance

in the studied disease. The developed GA is a method that incorporated into feature selection process some kind of biological information related to the disease based on the information contained in the public KEGG database.

WIMP is a free public web tool developed to be used by registered users. This tool was developed due to the need of pre-processing a biomedical dataset as a necessary step before any type of analysis you want to perform on it. When working with real biomedical data sets, it is very common to find incomplete samples that have to be preprocessed to avoid problems during the analysis. In this sense, WIMP provides to the research community the ability to impute missing data values in datasets that each user load onto the system by choosing among multiple imputation methods that are incorporated within the system in its original version. These methods are already implemented and tested, ie , the user should not worry about finding any imputation method or pay for a commercial software that offers this possibility. Among the different imputation methods available, WIMP incorporates some based on statistical techniques (mean imputation, hot-deck imputation and multiple imputation) and others based on machine learning algorithms (Self-Organizing Maps and MLP imputation). The number of methods incorporated in the system in its initial version is not closed, so it could be expanded in the future by implementing other existing methods, validate and incorporate them into the kernel of the system.

On the other hand, another great advantage that WIMP offers to the users is to spare them the need for high computational resources. The imputation tasks that users request through the system are processed on the server side. Moreover, the server is connected to a computing cluster where the tasks requested by users of the system are finally executed. This allows a single user from their personal computer to execute a complex imputation method and obtain the entire data set after performing the imputation method in a reasonable time. Once a launched task by a user is completed, the server notifies the user the result via e-mail providing a download link that allows them to collect the data set resulting from the execution of the imputation method. This dataset is now composed by complete samples and is ready for any kind of analysis, thus reaching the first partial objective of this thesis.

Chapter 3 analyzed major shortcomings in the current information systems of an oncology service and presented the design and development of an OIS that covers all the needs and allows on the daily practice the use of predictive models based on clinical and histopathological information of each patient. In particular, thanks to the close cooperation between the ICB research group at the University of Málaga and clinicians from the oncology service at the HUVV in Malaga, a detailed analysis of the needs of this service was performed in order to successfully implement a system that incorporates together with clinical and histopathological information of each

patient, the developed predictive models thus offering more information to clinicians to decide about the diagnosis, prognosis and treatment to be applied to a patient. In this chapter, solutions to the most critical aspects that would facilitate the success in the deployment of the system were analyzed and provided, such as: (i) use a user-centered design to increase the usability and ease of maintainance of the system; (ii) use appropriate technology to facilitate the development, maintenance and future upgrades of the system; (iii) integration into the system of all the daily clinical routines of the oncology service; (iv) security and interaction of the OIS with other hospital information systems (HIS).

The final system developed and successfully deployed in the HUVV incorporated all the clinical and histopathologic information of each patient of the service. The system design included the distribution of the information in the following functional modules: patients manager, treatment outpatient unit, clinical research, genetic counseling and statistical analysis. The most important feature that makes this system stand out from others is that, in addition to store all the information related to patients of the service, it contains a specific module that allows the posibility of doing different type of studies based on that information. This module integrates several models for clinical prognosis as a result of previous studies performed by the ICB research group, giving clinicians the possibility of doing a survival analysis implemented using the Kaplan-Meier algorithm, Cox regression analysis, calculate risk functions and obtain contingency tables based on histopathological and clinical information of a subset of patients previously selected. The evaluation of the success on the deployment of the system was conducted from surveys filled by users at 3 and 15 months after deployment of the system at the hospital, showing widespread satisfaction and therefore reaching the second partial objective of this thesis.

This system has been a step forward in order to offer, in a near future, a personalized medicine in the daily clinical practice. However , the heterogeneity of cancer as a disease makes its evolution to be determined by not only clinical but also genetic factors, which leads to consider in the future the need to integrate the genetic profiles of patients within the system. Therefore, it arises the need to develop new predictive models based on expression profiling data and incorporate them into the OIS. Thus, clinicians would have the ability to practice a personalized medicine to their patients attending both types of data by using the predictive models incorporated in the statistical analysis module. In this regard, Chapter 4 proposed the use of a GA using C-Mantec as classification algorithm for estimating predictive models with data from expression profiles. The main novelties introduced in this GA was the application of a pre-filter of genes using the Welch's statistical t-test and the inclusion in the fitness function of the GA of a term that seeks to minimize the correlation between variables.

The validation of this new proposed strategy was conducted by comparing the results with the ones obtained with the SFS+CNN strategy proposed in the preliminary work presented in Appendix A, but also adding the prefiltering of genes according to the result of the Welch's statistical t-test (for a fair comparison). This comparison generally showed better results, in terms of prediction, of the GA strategy than SFS. In contrast, the GA strategy produced larger subsets of genes and is computationally more demanding than SFS since it evaluates much more subsets of genes. Regarding the influence of the classification algorithm on the feature selection process, MLP, LDA and SVM provided the best results while C-Mantec and kNN were close behind with a slightly lower prediction. Furthermore, C- Mantec, MLP and LDA produced genetic signatures with fewer genes in comparison to SVM, NB and kNN, and particularly C-Mantec is the most robust classifier when adjusting the parameters of the model.

Chapter 5 provided an extensive comparison between different feature selection methods using several standard classification algorithms, all under an honest schema validation. In this regard, and given the low number of samples available in expression profiling datasets, a *resampling* technique known as BCV was used. The results obtained using the same validation scheme described and executing the different strategies that arise by combining the feature selection methods and the classification algorithms showed that filter or embedded methods generally provide higher predictive capabilities, are less complex and much faster to execute than wrapper methods. Filter methods also appeared as the most robust methods regarding the obtained genetic signatures. Regarding classification algorithms, kNN and MLP are algorithms that can be considered more robust indepently of the feature selection method used. However, other classifiers such as SVM or C-Mantec could achieve similar results with an optimal adjustment of the respective parameters of both models, since it is not an easy task due to the variability of the size of the genetic signatures that are evaluated in the feature selection methods and, in this sense, C-Mantec was presented in Chapter 4 as a robust classifier in terms of parameter setting.

The novel feature selection method presented in Chapter 6 tries to obtain robust genetic signatures with high prediction capabilities and biological relevance on the studied disorder. In this sense, the new method used biological information available on the KEGG database associated to the disease. The two stages of the new proposed model are: (i) the first one filters the total number of genes based on the KEGG database and establishes a ranking of pathways according to the prediction ability of the genes involved in each pathway and a *data mining* process performed on the textual description of each pathway offered on the KEGG website. Then, (ii) the second stage runs a GA on the best pathways found in the first step to estimate a predictive model. The main novelty of this GA was the inclusion of a third term in the

fitness function that tries to maximize a function that scores the "quality" of the genetic signature based on the number of genes that it contains and are also part of the analyzed pathway, or of any of the other selected pathways in the first stage. The evaluation of this new method was performed using three sets of DNA microarray data from different types of cancer and the results showed that this new strategy improves consistency, robustness and prediction of the obtained genetic signatures. This novel strategy improved the results in comparison to a classic GA since it incorporates biological information in the selection process, regardless of the classifier used. Therefore, this new approach, together with the models presented in Chapers 4 and 5, reaches the third partial objective of this thesis and offers the possibility to define in a near future genetic signatures that can be used to predict the diagnosis and clinical outcome of patients with cancer.

Finally, as an overall conclusion of this thesis it could be said that the results obtained throughout the work presented in each Chapter could lead, in a near future, to a personalized medicine in the daily clinical practice. The predictive models based on expression profiling data that have been developed in this thesis could be integrated into the OIS deployed in the HUVV. Furthermore, it could incorporate the genetic profiles of each patient to maximize the added benefits mentioned throughout this thesis. On the personal side, all the work done in this thesis has allowed me to deepen and acquire an extensive background in such a wide area such as Bioinformatics.

8.1. Further work

Since the results obtained in this thesis open a new possibility of supplying in a near future a personalized medicine in the daily clinical practice, new avenues arise for future work and research lines within the ICB research group as a possible extension of all the work done in this thesis. Among the various ideas that have emerged, it could be included the next ones:

- Integration of genetic profiles of each patient and predictive models based on expression profiles in the OIS deployed in the HUVV in Malaga (see Chapter 3), where its new version 2.0 is expected to be deployed in the "Hospital Regional Virgen de la Victoria" that unifies several hospitals in the area of Málaga. Moreover, analyze whether the estimation of a prediction model based on both clinical and histopathological and genetic information improves existing predictive models based on the use of a single data type. This would be a great step forward to offer a personalized medicine and thus clinicians would have in the same tool all the clinical, histopathological and genetic information together with several predictive models previously estimated based on all the oncology patient population contained in the information system.

- Apply statistical techniques to pre-filter the set of features present in expression profiling data before estimating a predictive model. In this sense, it would be interesting to apply a modified t-test and use the *q-values* to identify the most significant genes that best differentiate the status or class of the samples. In addition, this information may be included in the strategy proposed in Chapter 6 so that this new approach should produce more rigorous results from the statistical point of view and therefore more robust genetic signatures, since the method would seek signatures consisting of any of the genes identified as significant. In this sense, the preliminary work done in this research line shows promising results that encourage to go deeper in this direction in order to obtain robust genetic signatures with high prediction capabilities.

- Validate the new models proposed and any new one that may arise using more recent technologies such as RNA-Seq, Next Generation Sequencing (NGS), or miRNA microarray data. Due to the close cooperation between the ICB research group with the department of Molecular Biology at the University of Malaga and the HUVV Oncology Unit, it would be interesting to consider the validation of the genetic signatures found in these or other related works by using expression profiling information available in public repositories.

Apéndice A

Selección de características utilizando redes neuronales constructivas

Resumen: El análisis de perfiles genéticos de microarrays, en los que se ven envueltos miles de genes, ha atraído el interés de la comunidad científica. En concreto, existen diversos métodos que utilizan redes neuronales artificiales para identificar un subconjunto óptimo de genes con altas capacidades predictivas en la enfermedad. Existen diversos métodos de selección de características que utilizan redes neuronales artificiales (RNA) para identificar un subconjunto óptimo de genes de un conjunto de datos de microarray. El problema de utilizar un algoritmo de selección de características por pasos hacia adelante (SFS) que, además, use RNAs para evaluar la capacidad de predicción de los subconjuntos de genes estudiados, reside en la óptima elección de los parámetros de una red neuronal y en elegir la arquitectura óptima de la red. En este sentido, nuestro primer trabajo trata de aplicar un algoritmo de red neuronal constructivo (C-Mantec) para predecir el estado positivo o negativo de muestras de microarrays asociadas a pacientes con cáncer de mama. Los resultados obtenidos muestran que C-Mantec mejora los resultados de una RNA clásica tanto en predicción como en tiempo de ejecución del modelo.

Título: Constructive Neural Networks to Predict Breast Cancer Outcome by Using Gene Expression Profiles
Autores: Daniel Urda, José Luis Subirats, Leo Franco, José Manuel Jerez
Revista: Lecture Notes in Computer Science
Volumen: 6096
Páginas: 317-326
Año: 2010
DOI: 10.1007/978-3-642-13022-9_32
Abstract: Gene expression profiling strategies have attracted considerable interest from biologist due to the potential for high throughput analysis of hundreds of thousands of gene transcripts. Methods using artifical neural networks (ANNs) were developed to identify an optimal subset of predictive gene transcripts from highly dimensional microarray data. The problematic of using a stepwise forward selection ANN method is that it needs many different parameters depending on the complexity of the problem and choosing the proper neural network architecture for a given classification problem is not a trivial problem. A novel constructive neural networks algorithm (CMantec) is applied in order to predict estrogen receptor status by using data from microarrays experiments. The obtained results show that CMantec model clearly outperforms the ANN model both in process execution time as in the final prognosis accuracy. Therefore, CMantec appears as a powerful tool to identify gene signatures that predict the ER status for a given patient.

Apéndice B

Método híbrido basado en la generalización-correlación para selección de características

Resumen: Este trabajo toma como base el trabajo presentado en el Apéndice A, ampliando la propuesta inicial en la que se comparan los dos algoritmos "wrapper" de selección de características (SFS+RNA y SFS+C-MANTEC) con una nueva propuesta de un modelo híbrido que combine las ventajas tanto de los métodos "wrapper" y los de filtrado. Para ello, se utiliza la medida estadística de la correlación entre variables junto con un algoritmo de predicción. El método desarrollado plantea tres estrategias de selección de variables: la primera, *ROnly*, se basa únicamente en la generalización escogiendo directamente los 10 genes con mayor capacidad individual de predicción; la segunda, *RelevanceF*, combina generalización y redundancia, de forma que del 10 % de genes con mayor capacidad individual de predicción, finalmente se escogen los 10 genes que presenten menor redundancia entre ellos; y la tercera de ellas, *RedundancyF*, es justo lo contrario a la anterior, es decir, del 10 % de genes con menor redundancia entre ellos se seleccionan los 10 con mayor capacidad individual de predicción. Los resultados obtenidos con la estrategia *RedundancyF* son muy similares a SFS+CNN en cuanto a predicción. Sin embargo, los requisitos en cuanto a complejidad y necesidad de cómputo son mucho menores en la estrategia *RedundancyF* y, por tanto, el tiempo necesario para estimar la firma genética es considerablemente menor en este modelo híbrido.

Título: Hybrid (Generalization-Correlation) Method for Feature Selection in High Dimensional DNA Microarray Prediction Problems
Autores: Yasel Couce, Leonardo Franco, Daniel Urda, José L. Subirats, José M. Jerez
Revista: Lecture Notes in Computer Science
Volumen: 6692
Páginas: 202-209
Año: 2011
DOI: 10.1007/978-3-642-21498-1_26
Abstract: Microarray data analysis is attracting increasing attention in computer science because of the many applications of machine learning methods in prediction problems. The process typically involves a feature selection step, important in order to increase the accuracy and speed of the classifiers. This work analyzes the characteristics of the features selected by two wrapper methods, the first one based on artificial neural networks (ANN) and the second in a novel constructive neural network (CNN) algorithm, to later propose a hybrid model that combines the advantages of wrapper and filter methods. The results obtained in terms of the computational costs involved and the prediction accuracy reached show the feasibility of the hybrid model proposed here and indicate an interesting research line for the near future.

Bibliografía

Adjeroh, D. A., Zhang, Y. y Parthe, R. On Denoising and Compression of DNA Microarray Images. *Pattern Recognition*, vol. 39(12), páginas 2478–2493, 2006.

Ahmed, A. A. y Brenton, J. D. Microarrays and breast cancer clinical studies: forgetting what we have not yet learnt. *Breast Cancer Res*, vol. 7(3), páginas 96–9, 2005.

Allison, D. B., Cui, X. Q., Page, G. P. y Sabripour, M. Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*, vol. 7, páginas 55–65+, 2006.

Antoniadis, A., Lambert-Lacroix, S. y Leblanc, F. Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics*, vol. 19(5), páginas 563–570, 2003.

Avanaki, M. R. N., Aber, A. y Ebrahimpour, R. Compression of cDNA Microarray Images based on Pure-Fractal and Wavelet-Fractal Techniques. *Graphics, Vision and Image Processing GVIP*, vol. 11(1), páginas 43–52, 2011.

Bammler, T., Beyer, R. P., Bhattacharya, S., Boorman, G. A., Boyles, A., Bradford, B. U. y others. Standardizing global gene expression analysis between laboratories and across platforms. *Nature methods*, vol. 2(5), páginas 351–356, 2005.

Battiato, S., Farinella, G., Gallo, G. y Guarnera, G. Neurofuzzy segmentation of microarray images. En *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, páginas 1–4. 2008.

Battiato, S. y Rundo, F. A bio-inspired CNN with re-indexing engine for lossless DNA microarray compression and segmentation. En *ICIP*, páginas 1737–1740. IEEE, 2009.

Bierman, R., Maniyar, N., Parsons, C. y Singh, R. MACE: Lossless Compression and Analysis of Microarray Images. En *Proceedings of the*

2006 ACM Symposium on Applied Computing, SAC '06, páginas 167–172. ACM, New York, NY, USA, 2006.

BLOOM, G., YANG, I., BOULWAR, D., KWONG, K., COPPOLA, D., ESCHRICH, S. ET AL. Multi-platform, multi-site, microarray-based human tumor classification. *American Journal of Pathology*, vol. 164(1), páginas 9–16, 2004.

BOULESTEIX, A.-L., TUTZ, G. y STRIMMER, K. A CART-based approach to discover emerging patterns in microarray data. *Bioinformatics*, vol. 19(18), páginas 2465–2472, 2003.

BRENTON, J. D., CAREY, L. A., AHMED, A. A. y CALDAS, C. Molecular classification and molecular forecasting of breast cancer: Ready for clinical application? *Journal of Clinical Oncology*, vol. 23(29), páginas 7350–7360, 2005.

BROWN, M. P. S., GRUNDY, W. N., LIN, D., CRISTIANINI, N., SUGNET, C. W., FUREY, T. S., ARES, M. y HAUSSLER, D. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, vol. 97(1), páginas 262–267, 2000.

BUTTE, A. J. y KOHANE, I. S. Unsupervised knowledge discovery in medical databases using relevance networks. *Proceedings/AMIA, Annual Symposium. AMIA Symposium*, páginas 711–715, 1999.

CHAVEZ-ALVAREZ, R., CHAVOYA, A. y MENDEZ-VAZQUEZ, A. Discovery of possible gene relationships through the application of self-organizing maps to dna microarray databases. *PLoS ONE*, vol. 9(4), 2014.

CHEN, X. y DUAN, H. A Vector-based Filtering Algorithm for Microarray Image. En *Complex Medical Engineering, 2007. CME 2007. IEEE/ICME International Conference on*, páginas 794–797. 2007.

CHEN, Y., DOUGHERTY, E. R. y BITTNER, M. L. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics*, vol. 2(4), páginas 364–374, 1997.

CHO, S.-B. y WON, H.-H. Machine learning in dna microarray analysis for cancer classification. En *Proceedings of the First Asia-Pacific Bioinformatics Conference on Bioinformatics 2003 - Volume 19*, APBC '03, páginas 189–198. Australian Computer Society, Inc., Darlinghurst, Australia, Australia, 2003.

CHOU, H.-L., YAO, C.-T., SU, S.-L., LEE, C.-Y., HU, K.-Y., TERNG, H.-J., SHIH, Y.-W., CHANG, Y.-T., LU, Y.-F., CHANG, C.-W., WAHLQVIST, M., WETTER, T. y CHU, C.-M. Gene expression profiling of breast

cancer survivability by pooled cDNA microarray analysis using logistic regression, artificial neural networks and decision trees. *BMC Bioinformatics*, vol. 14, 2013.

DOBBIN, K. K., BEER, D. G., MEYERSON, M., YEATMAN, T. J., GERALD, W. L., JACOBSON, J. W. y OTHERS. Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays. *Clinical Cancer Research*, vol. 11(2), páginas 565–572, 2005.

EIN-DOR, L., KELA, I., GETZ, G., GIVOL, D. y DOMANY, E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, vol. 21(2), páginas 171–178, 2005.

EISEN, M. B., SPELLMAN, P. T., BROWN, P. O. y BOTSTEIN, D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, vol. 95(25), páginas 14863–14868, 1998.

ELHAMIFAR, E. y VIDAL, R. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35(11), páginas 2765–2781, 2013.

ELLIS, M., DAVIS, N., COOP, A., LIU, M., SCHUMAKER, L., LEE, R. Y. ET AL. Development and Validation of a Method for Using Breast Core Needle Biopsies for Gene Expression Microarray Analyses. *Clinical Cancer Research*, vol. 8(5), páginas 1155–1166, 2002.

FARAMARZPOUR, N. y SHIRANI, S. Lossless and Lossy Compression of DNA Microarray Images. En *2004 Data Compression Conference (DCC 2004), 23-25 March 2004, Snowbird, UT, USA*, página 538. 2004.

FRIEDLAND, S., NIKNEJAD, A. y CHIHARA, L. A simultaneous reconstruction of missing data in {DNA} microarrays. *Linear Algebra and its Applications*, vol. 416(1), páginas 8 – 28, 2006. Special Issue devoted to the Haifa 2005 conference on matrix theory.

FUNG, B. Y. M. y NG, V. T. Y. Classification of heterogeneous gene expression data. *SIGKDD Explor. Newsl.*, vol. 5(2), páginas 69–78, 2003.

GIALLOURAKIS, C., HENSON, C., REICH, M., XIE, X. y MOOTHA, V. K. Disease gene discovery through integrative genomics. *Annual Review of Genomics and Human Genetics*, vol. 6(1), páginas 381–406, 2005.

GÓMEZ-RUIZ, J. A., JEREZ-ARAGONÉS, J. M., MUÑOZ-PÉREZ, J. y ALBA-CONEJO, E. A neural network based model for prognosis of early breast cancer. *Appl. Intell.*, vol. 20(3), páginas 231–238, 2004.

GOLUB, T. R., SLONIM, D. K., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J. P., COLLER, H., LOH, M. L., DOWNING, J. R., CALIGIURI, M. A., BLOOMFIELD, C. D. y LANDER, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, vol. 286(5439), páginas 531–537, 1999.

GUYON, I. An introduction to variable and feature selection. *Journal of Machine Learning Research*, vol. 3, páginas 1157–1182, 2003.

GUYON, I., WESTON, J., BARNHILL, S. y VAPNIK, V. Gene selection for cancer classification using support vector machines. *Machine Learning*, vol. 46(1-3), páginas 389–422, 2002.

HADJIANASTASSIOU, V., FRANCO, L., JEREZ, J., EVANGELOU, I., GOLDHILL, D., TEKKIS, P. y HANDS, L. Optimal prediction of mortality after abdominal aortic aneurysm repair with statistical models. *Journal of Vascular Surgery*, vol. 43, páginas 877–877, 2006.

HERO, A. O., FLEURY, G., MEARS, A. J., MESURES, S. D., OPHTHALMOLOGY, D. O., SCIENCES, V. y GENETICS, H. Multicriteria gene screening for analysis of differential expression with DNA microarrays. *EURASIP Journal on Applied Signal Processing*, vol. 1, páginas 43–52, 2004.

HERRERO, J., VALENCIA, A. y DOPAZO, J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, vol. 17(2), páginas 126–36, 2001.

HOFFMANN, R., SEIDL, T. y DUGAS, M. Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. *Genome Biology*, vol. 3(7), páginas 0033.1–0033.11, 2002.

HOUSEMAN, E., ACCOMANDO, W., KOESTLER, D., CHRISTENSEN, B., MARSIT, C., NELSON, H., WIENCKE, J. y KELSEY, K. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, vol. 13(1), 2012.

HU, H., LI, J., WANG, H. y DAGGARD, G. Combined Gene Selection Methods for Microarray Data Analysis. En *Knowledge-Based Intelligent Information and Engineering Systems* (editado por B. Gabrys, R. Howlett y L. Jain), vol. 4251 de *Lecture Notes in Computer Science*, páginas 976–983. Springer Berlin Heidelberg, 2006.

HUA, J., LIU, Z., XIONG, Z., WU, Q. y CASTLEMAN, K. R. Microarray BASICA: background adjustment, segmentation, image compression and analysis of microarray images. En *ICIP (1)*, páginas 585–588. 2003.

IOANNIDIS, J. Microarrays and molecular research: noise discovery? *The Lancet*, vol. 365(9458), páginas 454–455, 2005.

IRIZARRY, R. A., BOLSTAD, B. M., COLLIN, F., COPE, L. M., HOBBS, B. y SPEED, T. P. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, vol. 31(4), página e15, 2003.

IRIZARRY, R. A., WARREN, D., SPENCER, F., KIM, I. F., BISWAL, S., FRANK, B. C. y OTHERS. Multiple-laboratory comparison of microarray platforms. *Nature methods*, vol. 2(5), páginas 345–350, 2005.

JEREZ, J., FRANCO, L., ALBA, E., LLOMBART-CUSSAC, A., LLUCH, A., RIBELLES, N., MUNÁRRIZ, B. y MARTÍN, M. Improvement of breast cancer relapse prediction in high risk intervals using artificial neural networks. *Breast Cancer Research and Treatment*, vol. 94(3), páginas 265–272, 2005.

JEREZ-ARAGONÉS, J. M., GÓMEZ-RUIZ, J. A., RAMOS-JIMÉNEZ, G., MUÑOZ-PÉREZ, J. y ALBA-CONEJO, E. A combined neural network and decision trees model for prognosis of breast cancer relapse. *Artificial Intelligence in Medicine*, vol. 27(1), páginas 45–63, 2003.

JORNSTEN, R., WANG, W., YU, B. y RAMCHANDRAN, K. Microarray Image Compression: SLOCO and the Effect of Information Loss. *Signal Processing*, vol. 83(4), páginas 859–869, 2003.

JORNSTEN, R. y YU, B. "Comprestimation": Microarray Images in Abundance. En *Proceedings of the Conference on Information Sciences and Systems*. 2000.

KARIMI, N., SAMAVI, S., SHIRANI, S. y BEHNAMFAR, P. Segmentation of DNA microarray images using an adaptive graph-based method. *IET Image Processing*, vol. 4(1), páginas 19–27, 2010.

KERR, M. y CHURCHILL, G. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, vol. 7, páginas 819–837, 2001.

KIM, P., PARK, K. y CHO, H. A quality measure model for microarray images. *International Journal of Information Technology*, vol. 11, páginas 117–124, 2005.

KOHAVI, R. y JOHN, G. H. Wrappers for Feature Subset Selection. *Artificial Intelligence*, vol. 97(1), páginas 273–324, 1997.

KOLLER, D. y SAHAMI, M. Toward optimal feature selection. En *In 13th International Conference on Machine Learning*, páginas 284–292. 1995.

LARKIN, J. E., FRANK, B. C., GAVRAS, H., SULTANA, R. y QUACKENBUSH, J. Independence and reproducibility across microarray platforms. *Nat Methods*, vol. 2(5), páginas 337–344, 2005.

LE, Q.-T., SUTPHIN, P. D., RAYCHAUDHURI, S., YU, S. C. T., TERRIS, D. J., LIN, H. S., LUM, B., PINTO, H. A., KOONG, A. C. y GIACCIA, A. J. Identification of osteopontin as a prognostic plasma marker for head and neck squamous cell carcinomas. *Clinical Cancer Research*, vol. 9(1), páginas 59–67, 2003.

LI, L., DARDEN, T., WEINGBERG, C., LEVINE, A. y PEDERSEN, L. Gene Assessment and Sample Classification for Gene Expression Data Using a Genetic Algorithm / k-nearest Neighbor Method. *Combinatorial Chemistry & High Throughput Screening*, vol. 4, páginas 727–739(15), 2001.

LI, Z. y WENG, G. Segmentation of cDNA Microarray Image using Fuzzy c-mean Algorithm and Mathematical Morphology. *Key Engineering Materials*, vol. 464, páginas 159–162, 2011.

LIU, J. J., CUTLER, G., LI, W., PAN, Z., PENG, S., HOEY, T., CHEN, L. y LING, X. B. Multiclass cancer classification and biomarker discovery using ga-based algorithms. *Bioinformatics*, vol. 21(11), páginas 2691–2697, 2005.

LÖNNSTEDT, I., RIMINI, R. y NILSSON, P. Empirical bayes microarray ANOVA and grouping cell lines by equal expression levels. *Statistical Applications in Genetics and Molecular Biology*, vol. 7(1), página a7, 2005.

LØNNING, P. E., SØRLIE, T. y BØRRESEN-DALE, A.-L. Genomics in breast cancer-therapeutic implications. *Nature Clinical Practice Oncology*, vol. 2(1), páginas 26–33, 2005.

LUKAC, R., PLATANIOTIS, K. N., SMOLKA, B. y VENETSANOPOULOS, A. N. A Data-Adaptive Approach to cDNA Microarray Image Enhancement. En *International Conference on Computational Science (2)* (editado por V. S. Sunderam, G. D. van Albada, P. M. A. Sloot y J. Dongarra), vol. 3515 de *Lecture Notes in Computer Science*, páginas 886–893. Springer, 2005.

MICHIELS, S., KOSCIELNY, S. y HILL, C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *The Lancet*, vol. 365(9458), páginas 488–492, 2005.

NEVES, A. J. R. y PINHO, A. J. Lossless Compression of Microarray Images Using Image-Dependent Finite-Context Models. *IEEE Transactions on Medical Imaging*, vol. 28, páginas 194–201, 2009.

NGUYEN, D. V. y ROCKE, D. M. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, vol. 18(1), páginas 39–50, 2002.

ORR, M. y SCHERF, U. Large-scale gene expression analysis in molecular target discovery. *Leukemia*, vol. 16(4), páginas 473–7, 2002.

PETERS, T., SMOLIKOVA-WACHOWIAK, R. y WACHOWIAK, M. Microarray Image Compression Using a Variation of Singular Value Decomposition. En *Proceedings of the Annual International Conference of the IEE Engineering in Medicine and Biology Society*, páginas 1176–1179. 2007.

PIAO, H. DNA microarray data analysis using a correlational bayesian network. *Journal of Medical Imaging and Health Informatics*, vol. 1(4), páginas 366–370, 2011.

QUACKENBUSH, J. Microarray data normalization and transformation. *Nat Genet*, vol. 32 Suppl, páginas 496–501, 2002.

RAMASWAMY, S., TAMAYO, P., RIFKIN, R., MUKHERJEE, S., YEANG, C. H., ANGELO, M., LADD, C., REICH, M., LATULIPPE, E., MESIROV, J. P., POGGIO, T., GERALD, W., LODA, M., LANDER, E. S. y GOLUB, T. R. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences*, vol. 98(26), páginas 15149–15154, 2001.

RAYCHAUDHURI, S., STUART, J. M. y ALTMAN, R. B. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac Symp Biocomput*, páginas 455–466, 2000.

SAUER, U., PREININGER, C. y SCHMATZBERGER, H. Quick and simple: quality control of microarray data. *Bioinformatics*, vol. 21(8), páginas 1572–8, 2005.

SCHADT, E., LI, C., ELLIS, B. y WONG, W. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J Cell Biochem*, vol. Suppl 37, páginas 120–5, 2001.

SHENA, M., SHALON, D., DAVIS, R. y BROWN, P. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, vol. 270(5235), páginas 467–470, 1995.

SMOLKA, B. y PLATANIOTIS, K. N. Ultrafast Technique of Impulsive Noise Removal with Application to Microarray Image Denoising. En *Image Analysis and Recognition, Second International Conference, ICIAR 2005, Toronto, Canada, September 28-30, 2005, Proceedings*, páginas 990–997. 2005.

SOUKAS, A., COHEN, P., SOCCI, N. y FRIEDMAN, J. Leptin-specific patterns of gene expression in white adipose tissue. *Genes Dev*, vol. 8(8), páginas 963–80, 2000.

SREEKUMAR, J. y JOSE, K. Statistical tests for identification of differentially expressed genes in cDNA microarray experiments. *Indian Journal of Biotechnology*, vol. 7, páginas 423–426, 2008.

STANGEGAARD, M. *Gene Expression Analysis Using Agilent DNA Microarrays*, vol. 529 de *Methods in Molecular Biology*, páginas 133–145. Springer Science+Business Media B.V., 1 edición, 2009.

TAMAYO, P., SLONIM, D., MESIROV, J., ZHU, Q., KITAREEWAN, S., DMITROVSKY, E., LANDER, E. S. y GOLUB, T. R. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences, USA*, vol. 96, páginas 2907–2912, 1999.

THEILHABER, J., CONNOLLY, T., ROMAN-ROMAN, S., BUSHNELL, S., JACKSON, A., CALL, K., GARCIA, T. y BARON, R. Finding genes in the C2C12 osteogenic pathway by k-nearest-neighbor classification of expression data. *Genome research*, vol. 12(1), páginas 165–176, 2002.

TONG, D. y SCHIERZ, A. Hybrid genetic algorithm-neural network: Feature extraction for unpreprocessed microarray data. *Artificial Intelligence in Medicine*, vol. 53(1), páginas 47–56, 2011.

TREVINO, V., FALCIANI, F. y BARRERA-SALDAÑA, H. DNA Microarrays: a Powerful Genomic Tool for Biomedical and Clinical Research. *Molecular Medicine*, vol. 13(10), páginas 527–541, 2007.

TÖRÖNEN, P., KOLEHMAINEN, M., WONG, G. y CASTRÉN, E. Analysis of gene expression data using self-organizing maps. *{FEBS} Letters*, vol. 451(2), páginas 142 – 146, 1999.

TUSHER, V. G., TIBSHIRANI, R. y CHU, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, vol. 98(9), páginas 5116–5121, 2001.

USLAN, V. y BUCAK, I. Clustering-Based Spot Segmentation of cDNA Microarray Images. En *Proceedings of the 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, páginas 1828–1831. 2010.

VAN DE VIJVER, M. J., HE, Y. D., VAN 'T VEER, L. J., DAI, H., HART, A. A. ET AL. A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. *New England Journal of Medicine*, vol. 347(25), páginas 1999–2009, 2002.

WANG, J., DELABIE, J., AASHEIM, H., SMELAND, E. y MYKLEBOST, O. Clustering of the som easily reveals distinct gene expression patterns: re-

sults of a reanalysis of lymphoma study. *BMC Bioinformatics*, vol. 3(1), página 36, 2002.

WANG, X., GHOSH, S. y GUO, S. Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Research*, vol. 29, página e75, 2001.

WANG, X. y SIMON, R. Microarray-based cancer prediction using single genes. *BMC Bioinformatics*, vol. 12, 2011.

WANG, Z., PALADE, V. y XU, Y. Neuro-Fuzzy Ensemble Approach for Microarray Cancer Gene Expression Data Analysis. En *Proceeding of the Second International Symposium on Evolving Fuzzy System (EFS'06)*, páginas 241–246. Lancaster, UK, 2006.

WEINSTEIN, J., MYERS, T., O'CONNOR, P., FRIEND, S., FORNACE, A., KOHN, K., FOJO, T., BATES, S., RUBINSTEIN, L., ANDERSON, N., BUOLAMWINI, J., VAN OSDOL, W., MONKS, A., SCUDIERO, D., SAUSVILLE, E., ZAHAREVITZ, D., BUNOW, B., VISWANADHAN, V., JOHNSON, G., WITTES, R. y PAULL, K. An information-intensive approach to the molecular pharmacology of cancer. *Science*, vol. 275(5298), páginas 343–349, 1997.

WEN, X., FUHRMAN, S., MICHAELS, G. S., CARR, D. B., SMITH, S., BARKER, J. L. y SOMOGYI, R. Large-scale Temporal Gene Expression Mapping of Central Nervous System Development. *Proceedings of National Academy of Sciences*, vol. 95(1), páginas 334–339, 1998.

XU, H., LEMISCHKA, I. y MAÁYAN, A. SVM classifier to predict genes important for self-renewal and pluripotency of mouse embryonic stem cells. *BMC Systems Biology*, vol. 4, 2010.

YANG, I., CHEN, E., HASSEMAN, J., LIANG, W., FRANK, B., WANG, S., SHAROV, V., SAEED, A., WHITE, J., LI, J., LEE, N., YEATMAN, T. y QUACKENBUSH, J. Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biology*, vol. 3(11), páginas research00621–research006212, 2002a.

YANG, Y. H., DUDOIT, S., LUU, P., AVID M. LIN, D., PENG, V., NGAI, J. y SPEED, T. P. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, vol. 30(4), página e15, 2002b.

YU, L. y LIU, H. Redundancy based feature selection for microarray data. En *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining*, páginas 737–742. ACM Press, 2004.

ZHANG, H., HO, T. y KAWASAKI, S. Wrapper feature extraction for time series classification using singular value decomposition. *International Journal of Knowledge and Systems Science*, vol. 3(1), páginas 53–60, 2006.

ZHANG, Y. y ADJEROH, D. A. Prediction by partial approximate matching for lossless image compression. *IEEE Trans. Image Process*, vol. 17, páginas 924–935, 2008.

ZHANG, Y., PARTHE, R. y ADJEROH, D. A. Lossless Compression of DNA Microarray Images. En *CSB Workshops*, páginas 128–132. IEEE Computer Society, 2005.

ZIFAN, A., MORADI, M. H. y GHARIBZADEH, S. Microarray image enhancement by denoising using decimated and undecimated multiwavelet transforms. *Signal, Image and Video Processing*, vol. 4(2), páginas 177–185, 2010.