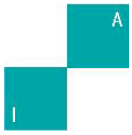
Inteligencia Artificial 45(2010), 1-4
doi:10.4114/ia.v14i45.1050



INTELIGENCIA ARTIFICIAL

http://erevista.aepia.org/

Resumen de Tesis: Modelo para la integración de conocimiento biológico explícito en técnicas de clasificación aplicadas a datos procedentes de microarrays de ADN

Daniel Glez-Peña
Departamento de Informática. Escuela Superior de Ingeniería Informática. Universidad de Vigo.
dgpena@uvigo.es

Resumen En esta Tesis se desarrolla un modelo para la integración de conocimiento biológico explícito en técnicas de clasificación aplicadas a datos procedentes de microarrays de ADN, en el contexto del diagnóstico de enfermedades complejas como el cáncer. El sistema desarrollado ha sido validado de forma exhaustiva concluyéndose que la integración de conocimiento biológico permite aumentar de forma significativa la coherencia de las técnicas estándar de selección de genes, mantener la precisión en las clasificaciones, y aumentar la robustez al combinar datos experimentales procedentes de distintos laboratorios.

Palabras clave: Integración de conocimiento, Clasificación, Microarrays de ADN.

1 Introducción

En este trabajo se presenta geneReasoner, un modelo de integración de conocimiento biológico explícito en técnicas de clasificación (aprendizaje supervisado) con el fin de mejorar sus resultados cuando se aplican al diagnóstico de enfermedades a partir de muestras tomadas con microarrays de ADN [1]. Se ha constatado que la clasificación en este contexto presenta problemas de sobreentrenamiento debido a la alta dimensionalidad de los datos procedentes de microarrays de ADN (miles de genes) [2], frente a la escasez de muestras (decenas de pacientes) que obliga a una selección de genes marcador que, sin embargo, difiere entre los distintos estudios llevados a cabo sobre las mismas enfermedades [3], además de carecer de significado biológico [4]. Por todo ello, y siguiendo las recomendaciones de algunos autores [5], se propone la inclusión de conocimiento biológico explícito para superar dichos problemas durante la creación de clasificadores orientados al diagnóstico eficaz de enfermedades complejas y relacionadas con la genética, como es el caso del cáncer.

El modelo propuesto en la presente investigación incluye una representación formal del conocimiento biológico basada en *conjuntos de genes*, junto con una arquitectura que incorpora técnicas de aprendizaje automático estándar para llevar a cabo diferentes tareas. En este sentido, el usuario del sistema propuesto no sólo proporciona los datos experimentales obtenidos mediante ensayos de microarrays, sino que aporta además una serie de grupos de genes con base en distintos criterios biológicos que guardan relación con el experimento (p.ej: pertenencia a una misma ruta metabólica, relación conocida con la patología, experiencia previa del experto, etc). Dicho conocimiento es automáticamente ampliado por el sistema a través de la adición de un nuevo grupo de genes, derivado de la aplicación de una técnica de selección de características sobre los datos de entrenamiento. A continuación, se utiliza un algoritmo genético para explorar el espacio de hipótesis, que tiene en cuenta todas las posibles intersecciones y uniones entre los grupos aportados, en busca de lo que se denomina una interpretación adecuada para los datos, consistente en una lista de genes con alto poder discriminante y capacidad de explicación simple. Posteriormente, se filtran los datos de forma que se tienen en cuenta únicamente los genes implícitos a la interpretación encontrada, y se entrena un clasificador estándar dado. De forma paralela, el modelo genera una explicación para la interpretación final utilizando un lenguaje cercano al usuario, basado en los conceptos biológicos a los que se vinculan los grupos de entrada. La arquitectura del sistema propuesto contempla la

ISSN: 1988-3064(on-line)

posibilidad de llevar a cabo una sustitución de las técnicas empleadas para realizar los procesos de selección de genes, clasificación y búsqueda, por lo que se consigue una alta flexibilidad posibilitando su configuración y adaptación a diferentes contextos.

La justificación de la hipótesis defendida en este trabajo se lleva a cabo de forma experimental, empleando distintos conjuntos de datos reales accesibles a través de Internet. Los resultados obtenidos a partir de los experimentos realizados con el modelo propuesto se comparan con los generados mediante la utilización de distintas técnicas de clasificación y selección de genes, lo que permite la realización de un análisis cuantitativo y cualitativo de la eficacia del sistema desarrollado.

Por último, y a la vista de los resultados obtenidos tras la experimentación realizada, se concluye que la integración de conocimiento biológico permite aumentar de forma significativa la coherencia de las técnicas estándar de selección de genes, mantener la precisión en las clasificaciones, y aumentar la robustez al combinar datos experimentales procedentes de distintos laboratorios.

2 Contribuciones

La principal aportación del trabajo desarrollado ha sido la definición de un modelo de clasificación completo que, a diferencia de las técnicas de clasificación estándar, incorpora conocimiento biológico previo en forma de conjuntos de genes que representan alguna relación biológica relevante y de interés para el problema concreto a resolver. A partir de dicho conocimiento y de un conjunto de datos de entrenamiento, geneReasoner encuentra un conjunto de genes que son, por un lado, relevantes biológicamente en el sentido de que aportan una explicación en términos del conocimiento previo disponible, y a su vez tienen poder discriminante entre las muestras de las diferentes clases del conjunto de datos de entrenamiento. El conjunto de genes resultante se emplea para generar una explicación al usuario y para entrenar un clasificador estándar e intercambiable, encargado de asesorar al especialista en decisiones futuras.

Además son relevantes dos aportaciones adicionales derivadas del propio desarrollo del modelo. Por un lado, geneReasoner aporta un nuevo modelo formal de representación del conocimiento que posee varias propiedades, entre las que se encuentran: la habilidad para definir nuevas hipótesis en forma de listas de genes estructuradas y biológicamente relevantes, la posibilidad de generar explicaciones simples a partir de estas listas y la capacidad para evaluar la complejidad de las mismas. Por otro lado, geneReasoner define una arquitectura flexible y genérica basada en distintas técnicas estándar e intercambiables con el objetivo de la creación de un modelo de clasificación completo. Se contempla el uso de técnicas de selección de características, como mRMR, capaces de complementar el conocimiento previo disponible e identificar aquellos genes que presentan un patrón altamente discriminante entre los aportados por el experto, el uso de algoritmos genéticos para la exploración del espacio de hipótesis definido a partir del conocimiento existente, de medidas multivariantes, como CFS, para el cálculo del poder discriminante de genes, y de clasificadores estándar, como SVM, Random Forest y GCS entre otros, para llevar a cabo la tarea de clasificación de muestras empleando únicamente los genes biológicamente relevantes.

Acompañando a geneReasoner, se han realizado trabajos de carácter traslacional con el fin último de acercar este tipo de modelos al ámbito clínico. Entre ellos destacan geneCBR [6], una herramienta de apoyo a la toma de decisiones en diagnósticos basados en datos procedentes de microarrays de ADN y WhichGenes [7], una herramienta que permite la creación sencilla e intuitiva de grupos de genes a partir de diversas fuentes de información disponible. Estas herramientas hacen posible la integración de distinto conocimiento y representan la base para la efectiva explotación del modelo desarrollado en la presente investigación.

3 Estructura

El trabajo realizado se presenta organizado en 7 capítulos más un apéndice, de los cuales el primero de ellos sirve para definir, en líneas generales, el problema que se pretende solucionar, establecer la hipótesis de partida comentando las fases de la investigación llevada a cabo y presentar la estructura organizativa de la memoria.

En el capítulo dos se introducen los fundamentos teóricos de la tecnología de los microarrays de ADN, haciendo un breve recorrido histórico desde su aparición en los años 80. Se presentan las principales áreas de aplicación médica que abarcan desde el análisis de expresión en diferentes tejidos y durante el desarrollo, hasta el análisis del genotipo y exploración de mutaciones, pasando por el estudio de patrones de expresión génica en sistemas modelo y en patógenos, el análisis de expresión diferencial en enfermedades y el estudio de la respuesta a tratamientos farmacológicos. Se realiza una explicación sobre los pasos a ejecutar durante la realización de un experimento y se documentan las técnicas habituales de pre-procesamiento de datos resultantes. Se identifican los tipos de análisis comúnmente realizados a partir de la información proporcionada y se procede a su clasificación en función del objetivo que persiguen. El capítulo finaliza realizando un resumen del mismo y presentado las conclusiones más relevantes.

El capítulo tres realiza un estudio de las técnicas y retos que presenta la clasificación de pacientes a partir de datos de microarrays en el contexto del problema planteado en la presente investigación. Se lleva a cabo una revisión de las técnicas comúnmente empleadas para la identificación de genes relevantes. En concreto, se analizan técnicas para el cálculo de la relevancia de genes, se presentan diversos esquemas de búsqueda y selección de genes, y se detallan los distintos marcos de integración con clasificadores. A continuación se realiza una revisión de los distintos modelos y técnicas de clasificación agrupándolos en aprendizaje basado en instancias, árboles de decisión, redes neuronales, clasificadores lineales y otro tipo de enfoques. Se identifican y explican los principales retos y dificultades existentes hoy en día en el campo, para finalmente, realizar un resumen del capítulo y esbozo de las conclusiones más relevantes.

El capítulo cuatro lleva a cabo una revisión sobre las distintas alternativas existentes para la integración de datos experimentales e información de tipo biológico, centrando el interés en aquellas propuestas relacionadas con problemas de clasificación. Se presentan las principales fuentes de conocimiento biológico disponibles, clasificándolas en genómicas, ontológicas, experimentales, metabólicas y patológicas. Se identifican y agrupan las técnicas más relevantes para la realización de análisis de tipo funcional a partir de la expresión diferencial. Se presentan las alternativas más relevantes para la mejora de técnicas mediante la incorporación de conocimiento previo y se revisan en detalle los aspectos relacionados con la clasificación de muestras. Finalmente, el capítulo concluye realizando un resumen y exponiendo las conclusiones más relevantes que motivan el presente trabajo de investigación.

En el capítulo cinco se presenta, geneReasoner, el modelo propuesto en esta investigación para la integración de conocimiento biológico explícito en técnicas de clasificación que utilizan datos provenientes de microarrays de ADN. Se describe en detalle la representación utilizada para modelar el conocimiento en forma de conjuntos de genes y cómo estos se integran en la propuesta global. Se presenta el esquema general de funcionamiento del modelo detallando, para cada etapa, los distintos elementos que lo componen y cómo éstos trabajan de forma conjunta para la búsqueda de una solución óptima. Se definen las relaciones existentes entre datos y conocimiento y se establecen los parámetros que permiten modelar el comportamiento global del sistema. El capítulo termina con un breve resumen y las conclusiones más destacables del modelo propuesto.

El capítulo seis presenta los resultados obtenidos por el modelo final propuesto, analizando las contribuciones del mismo desde dos puntos de vista diferentes: cuantitativa y cualitativamente. En primer lugar, se procede a la evaluación cuantitativa de los resultados obtenidos centrando el interés en el estudio de la precisión lograda por las distintas alternativas empleadas. En segundo lugar, se analiza a nivel cualitativo el grado de bondad de la solución, estudiando la coherencia, robustez y capacidad para generar explicaciones de tipo biológico. Como último punto, se presentan las conclusiones que se derivan del análisis de los resultados obtenidos.

Por último, el capítulo siete presenta una serie de comentarios relativos a las conclusiones extraídas de la aplicación del modelo desarrollado al problema planteado. Se detallan los objetivos alcanzados, la aplicabilidad del modelo presentado y las repercusiones directas de la investigación desarrollada. Finalmente se detallan las líneas de trabajo futuro a desarrollar, utilizando como punto de partida la investigación presentada en esta tesis.

4 Resumen y Conclusiones

El modelo geneReasoner, desarrollado para llevar a cabo la integración de conocimiento biológico, ha implicado la formalización del conocimiento a partir de conjuntos de genes de interés y el diseño de una arquitectura de integración de dicha formalización con algoritmos de clasificación estándar.

La formalización del conocimiento propuesta permite (*i*) la definición de un espacio de búsqueda de posibles explicaciones biológicamente relevantes para los datos experimentales, (*ii*) la cuantificación de la complejidad de las explicaciones generadas posibilitando primar aquellas más simples y (*iii*) facilita la construcción de explicaciones biológicas a partir de listas de genes con capacidad discriminante entre muestras de varias condiciones.

Por su parte, la arquitectura de integración de conocimiento permite la inclusión de la representación propuesta, en colaboración con otros algoritmos de minería de datos, para la mejora de técnicas estándar de clasificación, que se traduce en una selección de genes anterior al entrenamiento de cualquier técnica de clasificación específica, que además tiene en cuenta el conocimiento previo.

La validación se ha llevado a cabo utilizando varios conjuntos de datos públicos compuestos por pacientes que presentan distintos tipos de leucemia, así como con conocimiento procedente de distintas bases de datos que contienen información acerca de rutas metabólicas y genes asociados a diferentes tipos de patología.

Los resultados obtenidos han demostrado que geneReasoner se comporta de una forma mucho más coherente que otras técnicas de selección de genes, manteniendo la precisión y aumentando ligeramente la robustez sobre todo cuando se llevan a cabo clasificaciones que combinan varios conjuntos de datos procedentes de distintos laboratorios. Por otro lado, la integración de conocimiento biológico previo se ha demostrado eficaz, ya que geneReasoner es capaz de obtener buenos resultados de clasificación empleando únicamente los genes aportados

por el experto, que además recibe como salida una explicación simplificada de la selección realizada, que mantiene un alto grado de similitud, independientemente del conjunto de datos empleado.

Los trabajos llevados a cabo en la presente tesis se enmarcan dentro de varios proyectos nacionales e internacionales de investigación entre los que se encuentran la acción integrada *Development of computational tools for cancer diagnosis using gene expression data* (HP2006-0125) financiada por el Ministerio de Educación y Ciencia para el bienio 2007-2008 en colaboración con miembros de Departamento de Informática de la Universidade do Minho (Portugal), el proyecto *BioTools: Integración de conocimiento biológico en técnicas de IA aplicadas al agrupamiento y clasificación de datos de expresión génica* (2008-INOU-2) financiado por la Universidad de Vigo durante el año 2008, el proyecto *Desarrollo de herramientas computacionales de clasificación y clustering para el descubrimiento de información biológica relevante en el diagnóstico del cáncer a partir de datos de expresión genética* (VA100A08) concedido por la Junta de Castilla y León durante el año 2008 y el proyecto *Investigación en Bioinformática Traslacional* (08VIB6) concedido por la Universidad de Vigo durante el período 2008-2009.

Referencias

[1] Quackenbush,J. (2006). Microarray analysis and tumor classification. *The New England Journal of Medicine* 354:2463-2472. doi:10.1056/NEJMra042342

[2] Allison,D.B., Cui,X., Page,G.P., y Sabripour,M. (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics* 2006 7(1):55-65. doi:10.1056/NEJMra042342

[3] Fan,C., Oh,D.S., Wessels,L., Weigelt,B., Nuyten,D.S.A., Nobel,A.B., van't Veer,L.J., y Perou,C.M. (2006). Concordance among gene-expression-based predictors for breast cancer. *The New England Journal of Medicine* 355(6):560-569. doi:10.1056/NEJMoa052933

[4] Lottaz,C., y Spang,R. (2005). Molecular decomposition of complex clinical phenotypes using biologically structured analysis of microarray data. *Bioinformatics* 21(9):1971-1978. doi:10.1093/bioinformatics/bti292

[5] Bellazzi,R., y Zupan, B. (2007). Towards knowledge-based gene expression data mining. Journal of *Biomedical Informatics* 40(6):787-802. doi:10.1016/j.jbi.2007.06.005

[6] Glez-Peña,D., Díaz,F., Hernández,J.M., Corchado,J.M., y Fdez-Riverola,F. (2009). geneCBR: a translational tool for multiple-microarray analysis and integrative information retrieval for aiding diagnosis in cancer research. BMC *Bioinformatics* 10:187. doi:10.1186/1471-2105-10-187

[7] Glez-Peña,D., Gómez-López,G., Pisano,D.G., y Fdez-Riverola,F. (2009). WhichGenes: a web-based tool for gathering, building, storing and exporting gene sets with application in gene set enrichment analysis. *Nucleic Acids Research*. doi:10.1093/nar/gkp263