Un Sistema para la Extracción de Conocimiento en Bioinformática

Vicent Estruch, Cèsar Ferri, José Hernández, María José Ramírez

DSIC, Universitat Politècnica de València, Camí de Vera s/n, 46020 Valencia, Spain
{vestruch, cferri, jorallo, mramirez}@dsic.upv.es

Resumen

El área del aprendizaje automático, en especial el aprendizaje automático inductivo, ha perseguido durante décadas la generación automática de modelos generales a partir de conjuntos de datos particulares. Así, el uso de técnicas de aprendizaje automático para la extracción de conocimiento se ha convertido en una práctica frecuente y ha servido de impulso y de característica diferenciadora de los nuevos sistemas de minería de datos frente a los tradicionales sistemas de análisis de datos.

En particular, las técnicas de aprendizaje automático se han mostrado especialmente adecuadas en áreas caracterizadas por grandes volúmenes de datos pero con pocos modelos teóricos, como es el caso del diagnóstico médico o la biología molecular. Otra de las ventajas esgrimidas a favor de esta aproximación es que los modelos generados pueden fácilmente ser adaptados ante cambios de entorno (las hipótesis pueden revisarse ante la presencia de nuevos datos dando lugar a nuevas hipótesis), situación habitual en este tipo de aplicaciones. Una variedad de técnicas de aprendizaje automático han sido aplicadas en muchos problemas de bioinformática: redes neuronales, métodos bayesianos, programación lógica inductiva, árboles de decisión, máquinas de soporte vectorial, algoritmos genéticos, modelos de Markov, etc.

Sin embargo, aunque estos métodos son capaces de obtener modelos mucho más precisos que otros métodos clásicos, muchos de ellos han sacrificado algunos de los requerimientos de las aplicaciones en medicina y bioinformática: la comprensibilidad, la expresividad y la adaptación al contexto. Así, una de las principales críticas a algunas de las técnicas de aprendizaje automático es su poca (o nula) inteligibilidad. Este es el caso de las redes neuronales o de los modelos de Markov ocultos en los que el resultado del proceso es una caja negra que sirve para predecir o clasificar nuevos casos, pero no se sabe cómo, por lo que finalmente no se ha obtenido conocimiento útil. Otras aproximaciones, como los árboles de decisión o la programación lógica inductiva, generan modelos comprensibles pero con una menor precisión o eficiencia. Los métodos combinados mejoran la precisión pero tienen el inconveniente de que se pierde la comprensibilidad del modelo y, además, se incrementa enormemente el coste en recursos computacionales necesarios para obtener y almacenar el conjunto de hipótesis. Aun cuando el modelo generado sea preciso con respecto a los casos más frecuentes, esto no significa que pueda ser usado con seguridad en general: un error de clasificación con respecto a un caso infrecuente puede tener un coste mucho mayor que en la situación contraria.

Otra cuestión que apenas ha sido considerada en los sistemas de aprendizaje automático y que es fundamental en el análisis de datos en ámbitos científicos es el coste de los tests, es decir, el coste asociado con evaluar una determinada condición, como p.ej. “presión sanguínea > 12” o “colesterol > 200”. Muchos métodos de aprendizaje automático derivan un modelo que es función de todos los atributos, lo que obliga a aplicar todos los tests. Esto es inaceptable en muchos casos, especialmente en el diagnóstico médico.

En este trabajo presentamos un sistema de aprendizaje de modelos declarativos a partir de conjuntos de datos de entrenamiento (evidencia). Dicho sistema puede usarse para la extracción de conocimiento a partir de grandes volúmenes de datos relacionales, minimizando *conjuntamente* los costes de transformación, los costes de inteligibilidad y difusión de los modelos, los costes de clasificación errónea, los costes de aplicación de tests y los costes computacionales, muchos de ellos cruciales en aplicaciones científicas. Se trata, por lo tanto, de una aproximación innovadora en la visión de los costes de la minería de datos y en el desarrollo de herramientas que tengan en cuenta todos ellos y que sean parametrizables según la importancia que el usuario quiera dar a cada uno de los costes asociados al proceso de aprendizaje. Asimismo, el sistema integra nuevas técnicas en aprendizaje automático que hasta ahora se habían desarrollado por separado: análisis y toma de decisiones ROC, métodos de combinación de hipótesis, multi-árboles de decisión, etc. Se analiza en qué problemas de biología molecular y medicina dicho sistema puede ser más apropiado y se describe brevemente su uso, descarga, instalación y ejemplos que se pueden encontrar en su sitio web: http://www.dsic.upv.es/~flip/smiles/.