Técnicas de Minería de Datos para la Predicción de Enfermedades Cardiovasculares

Autores: Antonio Martínez Casado y José María Sandoval Cerezo
Tutores: Dr. D. José Tomás Palma Méndez y M.ª Trinidad Cámara Meseguer

Resumen

En este trabajo tratamos de crear y validar un modelo que permita determinar si un nuevo paciente padece o no enfermedad coronaria a partir de la información dada por un número pequeño de atributos utilizando técnicas de Minería de Datos.
Partimos de una base de datos formada por 303 pacientes y 70 atributos y tras realizar los experimentos oportunos obtuvimos un modelo de tipo Perceptrón Multicapa formado por 3 neuronas en la capa oculta y con un factor de aprendizaje de 0.1 y que determina si el paciente pertenece a la clase sano o enfermo a partir de 5 atributos con un error del 23,3% para los resultados enfermo y del 13,1% para los resultados sanos.

Introducción

Las enfermedades coronarias se producen cuando las arterias que suministran la sangre al músculo cardíaco se endurecen y se estrechan, reduciendo el flujo que llega al corazón.
La Minería de Datos se centra en la extracción no-trivial de información no conocida previamente y potencialmente útil a partir de grandes cantidades de datos. Su principal objetivo es resolver el problema analizando los datos que se encuentran en las bases de datos. El proceso de descubrimiento de conocimiento en el que se engloba la Minería de Datos se puede ver en la Figura 1.
Las técnicas utilizadas en Minería de Datos para generar un modelo pueden ser: predictivas (clasificación, modelo de regresión,…) o descriptivas (agrupar los individuos en clases, descubrir patrones secuenciales…).
En este trabajo utilizaremos tres clasificadores para crear modelos predictivos. Un clasificador es una función definida sobre los objetos y que asocia a cada vector de características la clase a la que pertenece.
Tras la creación de modelos es preciso evaluarlos, es decir, estimar su calidad.
Como paso previo a la creación de modelos se realiza una selección de atributos. Ésta se puede hacer por Filtrado de características o usando técnicas de búsquedas. El filtrado de características consiste en evaluar los atributos con alguna medida, ordenarlos y descartar los que caen por debajo de un determinado umbral. Las técnicas de búsqueda son aquellas en las que se realiza una búsqueda en el espacio de subconjuntos de atributos y se evalúa cada uno de ellos.

Objetivos

- Conocer qué es la minería de datos.
- Aplicar técnicas de minería de datos a una base de datos, para extraer información relacionada con enfermedades coronarias.
- Crear y validar un modelo que nos informe si un nuevo paciente puede padecer o no enfermedades coronarias.

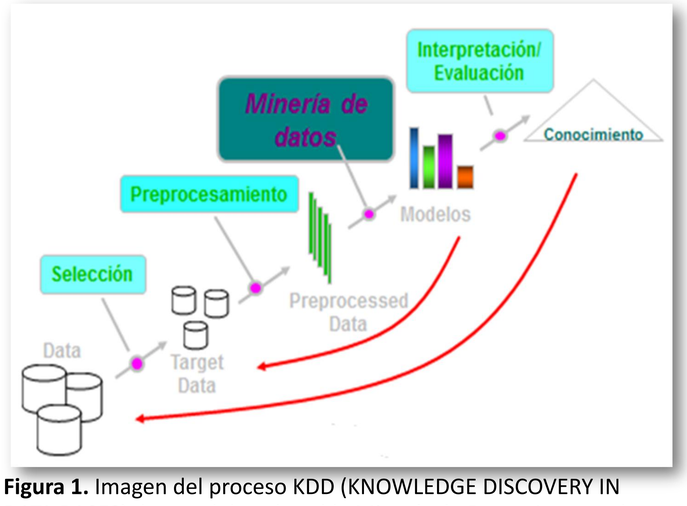

Figura 1. Imagen del proceso KDD (KNOWLEDGE DISCOVERY IN DATA BASES) dentro del cual está la Minería de Datos. Imagen de los apuntes de Palma, J. (2015) que, a su vez es una adaptación de la de U. Fayyad, et al. (1995), "From Knowledge Discovery to Data Mining: An Overview," Advances in Knowledge Discovery and Data Mining, U. Fayyad et al. (Eds.), AAAI/MIT Press.

Materiales

- Base de datos con 303 individuos y 70 atributos por individuo relacionada con enfermedades cardiovasculares.
- Programa informático WEKA.

Metodología

- Definición, características y elementos de la minería de datos, utilizando los apuntes de D. José Palma.
- Definición y características de Weka y de las enfermedades coronarias, usando Internet.
- Realización de experimentos para crear y validar un modelo que indique si un paciente padece o no enfermedades coronarias, empleando el programa Weka, a partir una base de datos con 303 individuos y 14 atributos cada individuo:
  1. Aplicación de tres técnicas para la selección de atributos: CFSSubSet, Evolutionarysearch y BestFirst.
  2. Aplicación de tres técnicas para construir un modelo predictivo: J48, RPART y MultilayerPerceptro.
  3. Validación del modelo: Validación cruzada y test T de Student.

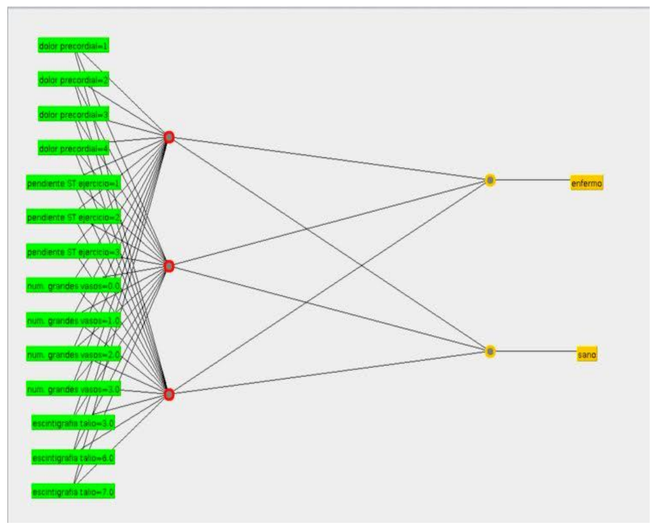

Figura 2. Imagen del Perceptrón Multicapa obtenido.

Experimentos realizados

1. Selección de atributos usando distintos métodos de búsqueda y distintas técnicas de clasificación.
2. Determinación del mejor clasificador para cada una de las bases de datos obtenidas en el experimento anterior.
3. Determinar la mejor base de datos de las obtenidas y el mejor clasificador.
4. Ajustar los parámetros del clasificador seleccionado anteriormente con los datos que tenemos, para que sea capaz de diagnosticar futuros casos.

Resultados obtenidos

1.

		Evaluador de Atributos		
		Wrapper+J48	Wrapper+PART	CFSSubsetEval
Método de Búsqueda	EvolutionarySearch	Reduced1	Reduced3	
	BestFirst	Reduced2	Reduced4	
	Greedy			Reduced5

2.

Base de datos	Reduced1	Reduced2	Reduced3	Reduced 4	Reduced 5
Mejor Clasificador	PART	PART	PART	PART	MP

3. reduced1.arff y MultilayerPerceptrón.

4. Perceptrón Multicapa con tres neuronas en la capa oculta y con un factor de aprendizaje de 0.1. Margen de error 23.3% para los pacientes sanos y 13.1% para los enfermos (ver Figura 2).

Conclusiones

- Este proyecto nos ha permitido conocer qué es la minería de datos y algunos de los métodos de trabajo que utiliza.
- Hemos comprobado la utilidad de esta rama de la informática para obtener información de bases de datos y, en nuestro caso, para colaborar en campos como el de la medicina.
- Con las técnicas aprendidas hemos construido un clasificador, concretamente un perceptrón multicapa, que nos permite determinar si un paciente padece una enfermedad coronaria a partir de cinco atributos y con un margen de error del 23.3% para los pacientes sanos y 13.1% para los enfermos.

Trabajo realizado dentro del proyecto IDIES 2016

Entidades colaboradoras: