Sistema de diagnóstico de enfermedades de vías urinarias

Norma Natalia Rubín Ramírez

Instituto Tecnológico de Tepic

nrubin@ittepic.edu.mx

Daniel Martín Preciado Ibarra

Instituto Tecnológico de Tepic

damapreciadoib@ittepic.edu.mx

Ángel Ríos Chávez

Instituto Tecnológico de Tepic

arios@ittepic.edu.mx

Irving Aldahyr Marín Bautista

Instituto Tecnológico de Tepic

a12400663@gmail.com

Resumen

Este artículo se centra en el desarrollo de esquemas de diagnóstico automático y flexible. Para ello se exploran diferentes alternativas capaces de utilizar eficientemente la información de un grupo de casos "etiquetados" para el diagnóstico de enfermedades de vías urinarias que se distinguen en dos tipos de infecciones: Inflamación aguda de la vejiga y Nefritis aguda. Este tipo de herramientas de diagnóstico sirven para la detección a tiempo de dichas enfermedades. En este artículo para realizar la separación de este tipo de infecciones se han explorado diferentes tipos de algoritmos de clasificación como M5P, RepTree, KStar, MultiplayerPerceptron y M5Rules, además del algoritmo K-Medias que el autor original utilizó en su artículo "Application of Rough Sets in the Presumptive Diagnosis of Urinary System Diseases" (J. Czerniak, 2003). La eficiencia de los algoritmos mencionados se puede determinar a través de la evaluación de la calidad de clasificación mediante la tasa de error, la rapidez en clasificar, la interpretabilidad y la simplicidad del algoritmo.

Abstract

This article focuses on the development of schemes and flexible automatic diagnosis. Acute inflammation of the bladder and acute nephritis: To do various alternatives capable of efficiently use information from a group of cases " labelled " for the diagnosis of urinary tract diseases that are distinguished in two types of infections are explored. Such diagnostic tools used for early detection of such diseases. In this paper for the separation of these infections have explored different types of classification algorithms as M5P, reptree, KStar, MultiplayerPerceptron and M5Rules, besides the K -means algorithm as the original author used in his article "Application of Rough Sets in the Presumptive Diagnosis of Urinary System Diseases" (J. Czemik, 2003). The efficiency of the above algorithms can be determined through the evaluation of the quality of classification by the error rate, speed sorting, interpretability and the simplicity of the algorithm.

Palabras Clave: algoritmo de decisión, reglas de decisión, modelos de árboles de decisión.

Introducción

Hoy en día con las grandes cantidades de información que se extraen en todo tipo de disciplinas, han llevado a emplear diferentes herramientas para el tratado de los mismos. En el área de Inteligencia Artificial se cuenta con diferentes tipos de técnicas para el procesado de información con la finalidad de extraer todo el conocimiento significativo que ayuden a resolver problemas relacionados con la medicina. De la técnica a la que se hace referencia es la Minería de Datos que es la que ayuda a extraer información de una manera eficiente, para después emplear otra área de la Inteligencia Artificial como el aprendizaje automático. La clasificación es la atribución de una clase específica a un objeto, esta atribución necesita un cierto grado de abstracción para poder extraer generalidades a partir de los ejemplos disponibles. Para una computadora la clasificación de rostros, de datos médicos o de formas son tareas

bastantes difíciles, mientras que para un ser humano son cuestiones cotidianas. Por ejemplo, en el caso de reconocimiento de caracteres manuscritos, es difícil enunciar una descripción general que tenga en cuenta todas las variaciones particulares de cada carácter. Una técnica que puede ser utilizada para resolver este problema es el aprendizaje, así, el criterio de decidir si una imagen corresponde a una letra "A" consiste en comparar si esta imagen es similar a otras "A" que se hayan introducido a la máquina previamente; con este enfoque, uno no solamente realiza la clasificación de las letras, sino que ayuda a que el algoritmo aprenda a partir de ejemplos etiquetados (S., 2012).

El aprendizaje consiste en la adaptación de los parámetros de un sistema, ya sea artificial o natural, en donde se busca obtener una respuesta frente a un estímulo externo. La definición de aprendizaje puede ser formalizada con el paradigma de aprendizaje supervisado que consiste en realizar técnicas iterativas de minimización de un costo (cuantificación de los errores en las respuestas), es decir, dispone de datos en forma de pares de entrada-salida a los que se denominan objetos. Si se posee un cierto número de estos objetos, entonces se tiene un conjunto de aprendizaje. Por lo tanto un clasificador se construye a partir de un conjunto de aprendizaje. La clasificación supervisada se caracteriza por tener clases ya determinadas y objetos caracterizados por atributos continuos o discretos que pertenecen a dichas clases. Por otro lado tenemos el aprendizaje no supervisado, que a diferencia del supervisado, no es necesario contar con pares de entrada y salida. En la clasificación no supervisada, no se tienen las clases determinadas sino que se van creando de acuerdo a las características de los objetos, es decir, los objetos que más se parezcan pertenecerán a una misma clase. En este artículo se realizaron pruebas con los datos donados por J.Czerniak (J. Czerniak, 2003), alojados en el repository Uci (UCI Repository Data Sets , s.f.), con los algoritmos M5P, RepTree, KStar, MultiplayerPerceptron y M5Rules, además del algoritmo K-Medias. Dentro de los métodos de validación o evaluación de la eficiencia de los algoritmos, en el marco de una aplicación de diagnóstico se consideró importante no solo la clasificación, sino también la validación de los

algoritmos utilizados por medio del estadístico de correlación lineal múltiplec (W. Mendelhall, 2010) (Navidi, 2006).

Métodos

Conceptos preliminares.

La importancia del proceso de selección de características en cualquier problema de clasificación, pone de manifiesto el permitir eliminar las características que puedan inducir el error, es decir, características ruidosas que no aporten información o aquellas que incluyen la misma información que otras (Pajares, 2010). Este proceso tiene varias ventajas como la disminución de los tiempos de procesamiento de los datos, un menor requerimiento de espacios para almacenar información y un bajo costo en la obtención de los datos. A continuación se mencionan brevemente conceptos importantes que se estudiaron para la realización de este proyecto.

Tareas de clasificación

Este tema tiene un lugar en un extenso campo de acción en diferentes áreas, en su concepto más amplio, el término de clasificación podría incluir cualquier contexto en el cual algún diagnóstico es hecho sobre la base de una información disponible, por lo que un procedimiento de clasificación es entonces un método formal que permita repetidamente realizar tales valoraciones ante nuevas situaciones. En (Pajares, 2010), se hace una detallada formulación para resolver un problema de clasificación, que es la técnica utilizada en este trabajo. Esta formulación la describe de la siguiente manera: si hay *J* clases de objetos de interés, que utilizan el subíndice j con j=1….J, para cada estado respectivamente. La información que se posee sobre los objetos es resumida en un número patrones, es decir, medidas de valor real denominadas características. Ya todas juntas forman un vector de características $x \in R^2$ es decir $x = (x_1, x_2, \dots., x_p)$ (Pajares, 2010). Se puede asumir que el problema tiene que ver con la construcción de un procedimiento que sea aplicado a una secuencia de casos, en el cual cada nuevo caso debe ser asignado a una clase de un conjunto predefinido de ellas sobre la base

de un grupo de rasgos observados. Para modelar la relación entre el vector de características se asume que un objeto de la clase $y \in \{1,2,\ldots.\}$, es un vector condicional a la clase $F_y(x)$.

En la actualidad muchos de los problemas que han surgido de la ciencia, la industria e incluso del comercio, requieren el uso de datos complejos que a menudo son muy extensos, pueden ser considerados como problemas de clasificación o decisión. En su forma más simple, la clasificación se presenta como un proceso de reconocimiento de determinados patrones. En este sentido contribuyen mucho en diferentes disciplinas científicas tales como medicina, bioinformática, ciencias biológicas etc., como ya se mencionó anteriormente para la clasificación podemos trabajar con aprendizaje supervisado y no supervisado, para la cual se describe a continuación algunos clasificadores que trabajan con este tipo de aprendizajes.

Tipos de clasificadores

El algoritmo MP5 es un algoritmo de regresión (Basilio, 2006), es un método de aprendizaje mediante arboles de decisión utiliza el criterio estándar de poda M5. Y constan con las siguientes características:

- Construcción de árbol mediante algoritmo inductivo de árbol de decisión.
- Decisiones de en rutado en nodos tomadas a partir de valores de los atributos.
- Cada hoja tiene asociada una clase que permite calcular el valor estimado de la instancia mediante una regresión lineal.

El Algoritmo K-Medias es un algoritmo clasificado como método de particionado y recolocación el cual ha sido utilizado en aplicaciones científicas e industriales (Basilio, 2006). El nombre le viene porque representa cada uno de los clústeres por la media (o media ponderada) de sus puntos, es decir, por su centroide. Este método únicamente se puede aplicar a atributos numéricos, y los *outliers* le pueden afectar muy negativamente. Sin embargo, la representación mediante centroides tiene la ventaja de que tiene un significado gráfico y estadístico inmediato. La suma de las discrepancias entre un punto y su centroide, expresado a través de la distancia apropiada, se usa

como función objetivo. La función objetivo, suma de los cuadrados de los errores entre los puntos y sus centroides respectivos, es igual a la varianza total dentro del propio clúster. La suma de los cuadrados de los errores se puede racionalizar, como el negativo del log-likelihood, para modelos mixtos que utilicen distribuciones normales.

El Clasificador REPTree está basado en arboles de decisión, que construye arboles de decisión utilizando información de la varianza y realiza la poda, usando un criterio de la reducción de errores.

KStar determina cuales son las instancias más parecidas, puede utilizar la entropía, o contenido de información de las instancias. Como medida de distancia entre ellas, son destacables las siguientes características: admite atributos numéricos y simbólicos, así como pesos por cada instancia y permite que la clase sea simbólica o numérica.

MultilayerPerceptron es una red neuronal artificial (RNA) (S., 2012), formada por múltiples capas, esto permite que sea capaz de resolver problemas que no son linealmente separables. Las capas del perceptron se pueden clasificar en 2:

- Capa de entrada: constituida por aquellas neuronas cuyas entradas provienen de unas capas interiores (proceso).
- Capa de salida: constituida por las neuronas que reciben el valor de salida de la capa oculta.

M5Rules: Genera una lista de decisiones para problemas de regresión lineal y utiliza la idea de “divide y vencerás”. En cada iteración se construye un modelo de árbol utilizando M5 y hace que el “mejor” de la hoja se convierta en una regla.

En la actualidad existen diversos clasificadores que nos ayudan a hacer una separación entre clases. Para tener acceso a estos algoritmos, nos auxiliamos del software WEKA el cual está considerado como el más utilizado para Minería de datos área de la Inteligencia Artificial.

El Software Weka, fue desarrollado por la Universidad de Waikato de Nueva Zelanda (Weka Software GNU , s.f.), y cuenta con una amplia gama de algoritmos de clasificación, además de algoritmos para el tratado de datos y señales.

Metodología del proyecto

A continuación se detalla los métodos empleados para la realización de este proyecto, la metodología consistió en 4 fases, que se muestran en el siguiente diagrama:

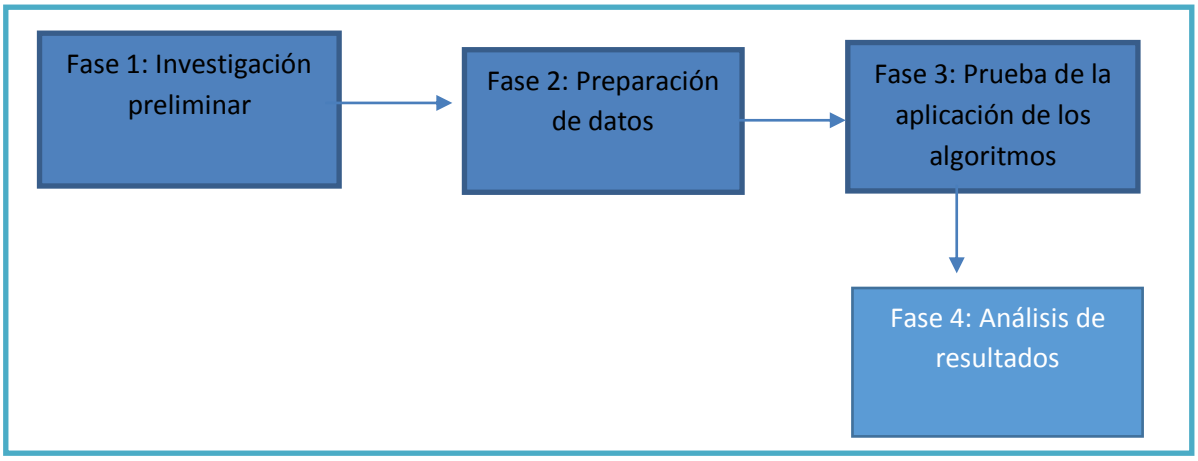

Diagrama 1. Fases del proyecto

Fase 1: Investigación preliminar

Para la realización de este proyecto, se realizó un análisis de cada uno de los algoritmos que se aplicaron en la base de datos "Infecciones de Vías Urinarias", alojada en el Repository UCI (UCI Repository Data Sets , s.f.), por el donador de la información (J. Czerniak, 2003), con la finalidad de ver los resultados obtenidos con el clasificador que utilizo, así mismo se dio a la tarea de ver el funcionamiento de los algoritmos que se utilizaron para la clasificación y así poder realizar las comparaciones de la tasa de error, la rapidez en clasificar, la interpretabilidad y la simplicidad de los algoritmos. Dicha base de datos consiste en realizar una clasificación de infecciones de vías urinarias las cuales cuenta con dos tipos de infecciones: inflamación aguda de la vejiga y nefritis aguda.

Fase 2: Preparación de datos

Una vez hecha la investigación, se exploró la base de datos, para ver la aplicabilidad de los algoritmos, con la finalidad de ver la posibilidad de saber si se sería necesario realizar alguna normalización de los datos. Por las características de los algoritmos,

pueden tener diferentes formas para introducir la información, como ya lo se mencionó en la fase 1. la base de datos con la que se trabajó, contiene la siguiente información: datos de 120 pacientes, recabando características o atributos que determinan las clases de infecciones de vías urinarias que a continuación se detallan:

1. Temperatura del paciente en un rango de 35 a 42 C.
2. Presencia de nausea {si, no}
3. Dolor lumbar{si, no}
4. Necesidad frecuente de orinar{si no}
5. Dolor al orinar
6. Ardor de uretra, comezón e inflamación en la salida de la uretra {si, no}
7. Diagnóstico de inflamación aguda de la vejiga{si, no}
8. Diagnóstico de nefritis aguda en las vías renales{si, no}

Fase 3: Prueba de la aplicación de los algoritmos

Para la aplicación de los algoritmos: M5P, RepTree, KStar, MultiplayerPerceptron y M5Rules y Kmedias, utilizamos el software Weka, desarrollado en 1997, en la Universidad de Waikato de NUEVA ZELANDA y utiliza los lenguaje TCL/TK y C, sin embargo en el año de 1997 emigro al lenguaje de programación JAVA y actualmente cuenta con una nutrida colección de algoritmos de clasificación además de muchos algoritmos para el tratado y filtrado de datos (Weka Software GNU , s.f.).

Fase 4. Análisis de resultados.

Una vez aplicados los algoritmos, se realizaron varias pruebas con la finalidad de probar cada uno de los algoritmos y ver el comportamiento de los algoritmos y determinar cuál de ellos era el más eficaz, u óptimo para llevar a cabo la clasificación, para realizar la validación de los resultados se aplicó el estadístico de correlación que establece una medida del grado de asociación lineal entre la variable respuesta y la variable predictora, concretamente entre la variable respuesta y la recta de regresión estimada.

Resultados

Una vez realizado la aplicación de los algoritmos con el software Weka, se realizaron diferentes pruebas con la finalidad que los algoritmos aprendieran a clasificar, aunque es conveniente aclarar que se utilizó el aprendizaje supervisado y no supervisado para los diversos algoritmos que se utilizaron, en la figura 1 se muestra la clasificación obtenida.

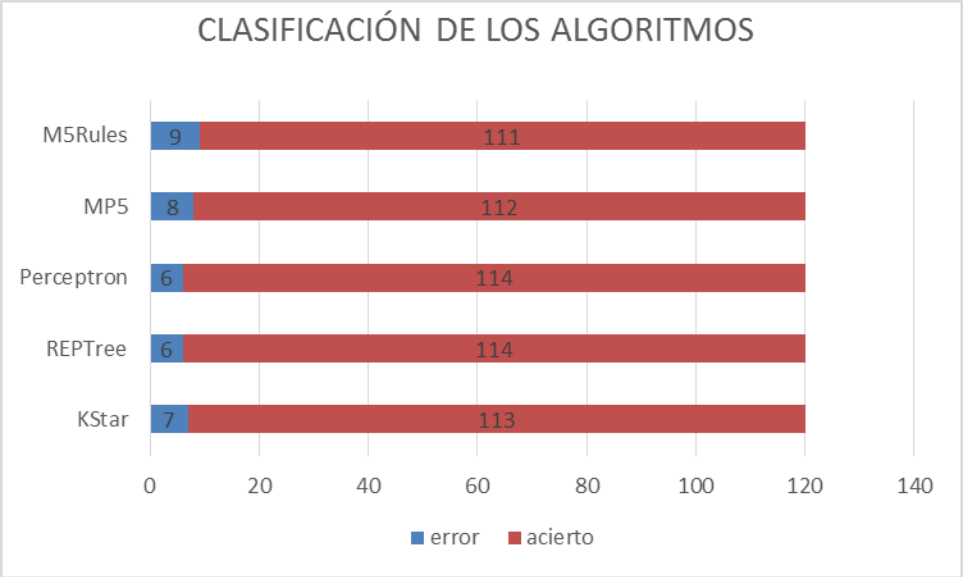

Figura 1. Clasificación de los algoritmos

Como se puede apreciar en la figura 1, se ve claramente que la clasificación ha sido muy parecida en todos los algoritmos puesto que de las 120 instancias el 90% fueron clasificadas de una manera correcta, por lo que su porcentaje de mal clasificadas fue de un 1 0%, como no se puede afirmar que la predicción sea la correcta. Se validaron estadísticamente los resultados de la clasificación.

En la tabla 1, se muestra el comportamiento del coeficiente de correlación, que nos ayuda a validar el algoritmo REPTree y Perceptron que son los que aparecen con el coeficiente de correlación más alto(W. Mendelhall, 2010) (Navidi, 2006), en el entendido que el coeficiente de correlación entre más tiende a 1, significa que las variables independientes tienen mucha relación con la variable dependiente, por lo que se puede decir que REPTree es el que tienen el mejor tiempo de ejecución tiene más elevado el

coeficiente de correlación, pero tiene el error absoluto medio más bajo que el perceptron que es el que tiene casi las mismos resultados que el de REPTree.

Tabla 1. Resultados estadísticos de los algoritmos

Algoritmos	Coeficiente de correlación	Error Absoluto Medio	Raíz de error absoluto	Error absoluto relativo	Raíz de error cuadrático relativo	Tiempo (s)
KStar	0.7476	163.7440	209.1165	61.67%	68.31%	0.02
REPTree	0.7709	165.8835	185.6498	54.67%	58.76%	0.02
Perceptron	0.7727	189.6330	217.3540	67.59%	72.07%	1.40
MP5	0.7613	140.2117	145.7819	47.12%	48.24%	0.30
M5Rules	0.7597	141.6854	153.0973	48.37%	48.98%	0.41
Kmeans	0.7627	182.6320	105.3876	67.59%	72.07%	0.99

En la fig. 2, se puede observar el comportamiento del estadístico de correlación, donde se puede ver claramente que efectivamente que los dos algoritmos que tuvieron más éxito en clasificar son REPTree y Perceptron.

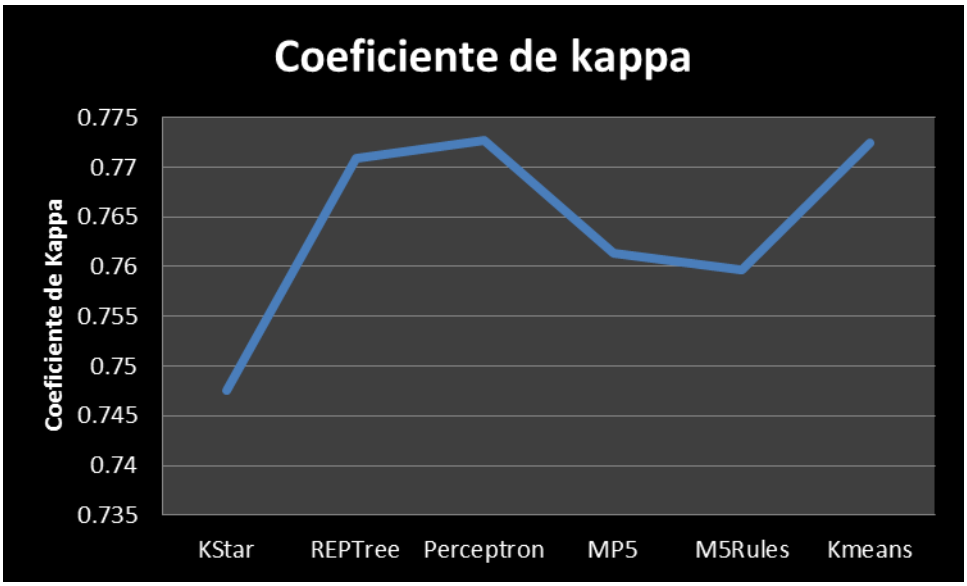

Figura 2. Comportamiento del coeficiente de correlación

Discusión

De acuerdo los resultados obtenidos de cada uno de los algoritmos, se puede concluir que los algoritmos de clasificación que utilizamos en general dieron resultados bastante buenos, para hacer la clasificación de los dos tipos de enfermedades de infecciones de vías urinarias: inflamación aguda de la vejiga y nefritis aguda, también podemos decir que el autor original al probar el algoritmo KMeans, con el que trabajaron en el artículo "*Application of Rough Sets in the Presumptive Diagnosis of Urinary System Diseases*" (J. Czerniak, 2003), sus resultados fueron bastante similares con la clasificación obtenida con los algoritmos analizados.

Podemos mencionar que el objetivo de este artículo se cumplió debido a que se obtuvieron resultados bastante buenos, lo que nos llevó a conocer el comportamiento de los algoritmos, no solo en la tasa de error, la rapidez en clasificar, la interpretabilidad y la simplicidad del algoritmo, si no que se logró obtener resultados positivos en la validación de los mismos, porque se puede afirmar que los datos independientes fueron correlacionados fuertemente con los datos dependientes, esto hace posible afirmar que se obtuvo una clasificación de calidad de las clases.

Referencias

(s.f.). Obtenido de UCI Repository Data Sets : http://archive.ics.uci.edu/ml/index.html

Basilio, S. A. (2006). *Aprendizaje Automático: conceptos básicos y avanzados, aspectos prácticos utilizando software WEKA.* Pearson.

J. Czemik, H. Z. (2003). Application of rough sets in the presumptive diagnosis of Urinary System Diseases Artificial Inteligence and Security in Computing Systems. *ACS 2002 9th International Conference Proceedings* (pág. 41). Kluwer Academic Publishers 2003.

Navidi, W. (2006). *Estadística para Ingenieros.* MC Graw Hill.

Pajares, M. G. (2010). *Aprendizaje Automático un Enfoque Práctico.* RAMA.

S., A. G. (2012). *Inteligencia Artificial.* RC LIBROS.

W. Mendelhall, R. J. (2010). *Introducción a la Probabilidad y Estadística.* Cengage Learning .

Weka Software GNU . (s.f.). Obtenido de http://www.cs.waikato.ac.nz/ml/weka/