Selección de Variables: Teoría de Testores *versus* Redes Bayesianas

Jorge Ochoa Somuano

Resumen

La selección de variables permite reducir la cantidad de información que se necesita para llevar a cabo procesos de clasificación, siendo ésta una de las principales ventajas que se puede obtener, ya que impacta de manera directa en ahorro de tiempo al tratar la información, y por ende, hay una reducción en los costos computacionales. *En este documento se muestran los resultados comparativos de técnicas que permiten aplicar la selección de variables, la clasificación se llevó a cabo con los mismos conjuntos de datos para cada herramienta utilizando el algoritmo k-means simple. Se utilizó un software propio denominado ReduceTT en el cual se implementó una técnica denominada teoría de testores y los resultados de la selección de variables con Redes Bayesianas. Una característica de la primera técnica, es la garantía de mantener la calidad en la clasificación para cualquier subconjunto de variables que se obtenga como resultado en la selección de variables, las Redes Bayesianas no ofrecen la misma garantía.*

Palabras clave: testores típicos, selección de variables, reducción de dimensiones, espacio de representación, atributos, objetos, datos, clasificación, redes bayesianas.

Abstract

Feature selection allows for a reduction in the amount of information that is needed to carry out classification processes, which is one of the main advantages that can be obtained from saving time in processing information; therefore, there is a reduction in the computational costs. *In this paper, the comparative results of techniques that allow feature selection to be carried out are shown. The classification was performed with the same data sets for each tool using the simple k-means algorithm. Software called ReduceTT was used, through which a technique called Testors Theory was implemented and the results of the selection of variables were obtained with Bayesian Networks. One characteristic of the Testors Theory that makes it different from others is its guarantee to maintain the quality of classification for any subset of features that is obtained as the result in the selection of variables; Bayesian Networks do not offer the same guarantee.*

Key words: Typical Testors, feature selection, dimension reduction, representation space, attributes, objects, data, classification, Bayesian Networks.

Recibido: 10 de enero de 2019

Aceptado: 15 de marzo de 2019

[1] Instituto de Industrias, Universidad del Mar campus Puerto Escondido. km 2.5 Carretera Puerto Escondido-Sola de Vega, Puerto Escondido 71980, San Pedro Mixtepec, Oaxaca, México.
* *Autor de correspondencia: ochoa@zicatela.umar.mx*

Introducción

Cuando se trabaja con información que está representada por muchas variables, se pueden encontrar subconjuntos que permiten mantener la calidad en la clasificación de los datos. Para obtener conjuntos más pequeños, se tiene que aplicar alguna técnica de selección de variables con el fin de obtener aquellas que mejor representen la información original.

El objetivo de la investigación es reducir el número de variables necesarias para representar un conjunto de datos, con la característica de mantener la calidad en la clasificación de la información. Precisamente es ahí donde radica el interés de este trabajo, hacer un comparativo de la teoría de testores con otra técnica con la finalidad de verificar cuál de ellas es la que mejor mantiene la calidad en la clasificación.

Para cumplir con el objetivo, se han estructurado los temas en el siguiente orden: se da información relevante de las herramientas utilizadas para la selección de variables, posteriormente se describen las bases de datos con las cuales se realizaron las pruebas de selección de variables, se indican las características de los experimentos, así como los resultados que se obtuvieron con cada herramienta, también se dan algunas conclusiones con base en los resultados y se exponen los trabajos futuros. Finalmente, se tiene una sección de referencias para quien desee profundizar en el tema.

Materiales y métodos

Herramientas utilizadas

Básicamente se está realizando la comparación entre los resultados de una investigación (Castro & Von-Zuben 2009) en la cual se utilizan las Redes Bayesianas para la selección de variables y los resultados obtenidos con la Teoría de Testores al llevar a cabo la misma tarea, por ello es que la comparación se realiza únicamente con dos herramientas:

En una aplicación implementada con un *Sistema Inmune-Inspirado* (Castro & Von-Zuben 2009), en la que se comparan los resultados de la investigación se reporta un sistema desarrollado con una metodología denominada Sistema Artificial Inmune, que es un mecanismo de búsqueda capaz de proponer múltiples redes bayesianas. Para lograr su objetivo utiliza las cadenas de Markov.

ReduceTT (Reduce Typical Testors) es una adaptación del sistema ReduClass desarrollado en Ochoa-Somuano (2005), es una herramienta que se ha implementado para la selección de variables, basada en el enfoque lógico combinatorio y la cual permite obtener todos los testores típicos (Ruiz *et al.* 1999, Ochoa-Somuano 2007, Alba *et al.* 2000, Díaz *et al.* 2011, Santiesteban & Pons 2003, Santos *et al.* 2004, Ochoa-Somuano 2013) de una base de datos.

Bases de datos

Para la experimentación que se realizó en esta investigación, se emplearon 5 bases de datos (BD) de las 10 utilizadas en (Castro & Von-Zuben 2009) para su experimentación. Las otras 5 bases de datos no se consideraron porque no son compatibles con la teoría de testores y no habría forma de establecer un comparativo. A continuación se da una breve descripción de cada una de las BD utilizadas para la experimentación:

Bupa (Bache & Lichman 2013). Base de datos de trastornos hepáticos por consumo excesivo de alcohol, propiedad de Richard S. Forsyth. Contiene 345 casos de los cuales 145 son positivos y 200 son negativos. Los casos están representados por 6 variables.

Ionosphere (Sigillito *et al.* 1989). Esta base de datos contiene información de datos obtenidos con un radar, propiedad de Vince Sigillito. Contiene 351 instancias de las cuales 126 son buenas y 225 son malas. Las instancias están representadas por 34 atributos.

Pima (Sigillito 1990). Base de datos de diabetes de los indios pima, propiedad del Instituto Nacional de Diabetes y Enfermedades Digestivas y del Riñón. Contiene 768 instancias de las cuales 500 son casos negativos y 268 positivos. La información está representada por 8 atributos.

Sonar (Aeberhard 1992). Esta base de datos contiene información de señales de sonar, propiedad de Terry Sejnowski. Contiene 208 casos de los cuales 111 se obtuvieron por rebote de señales en un cilindro de metal y 97 se obtuvieron a partir de rocas. Los casos están representados por 60 variables.

Wine (Gorman & Sejnowski 1988). Base de datos de resultados de análisis químicos a vinos que se cultivan en una misma región de Italia, propiedad del Instituto de Productos Farmacéuticos de Análisis y Tecnologías de Alimentos. Se aplicó a tres diferentes cultivos que representan el número de clases, el análisis determinó 13 tipos de constituyentes que equivalen a las variables y la dispersión por clases es de 59, 71 y 48 patrones.

En la Tabla I se puede ver un compilado de la información de las bases de datos usadas en la experimentación, que a su vez es una sub tabla de la mostrada en (Castro & Von-Zuben 2009). Es importante señalar que la teoría de testores tiene ciertas limitaciones en el tamaño de las bases de datos, principalmente con la cantidad de variables a utilizar, debido a su propia naturaleza para buscar los testores típicos, ya que lo hace de forma secuencial. Con base en lo anterior, se recomienda que las bases de datos no excedan de 200 variables. La limitante de la Teoría de Testores respecto a las bases de datos que no se utilizaron de (Castro & Von-Zuben 2009) es que no se pueden separar linealmente en un espacio de representación.

Desarrollo de experimentos

Lo que se realiza como parte de los experimentos es tomar las bases de datos de la Tabla I y se le aplican tres procesos, primero se lleva a cabo la clasificación de la información utilizando todas las variables, en segundo lugar se hace el proceso de selección de variables con el sistema ReduceTT, de forma consecutiva y como tercer paso se clasifica la información sólo con las variables seleccionadas. Posteriormente se realiza el comparativo de los resultados obtenidos en la fase experimental con los datos reportados en (Castro & Von-Zuben 2009). Finalmente se indican las diferencias entre cada uno de los resultados para poder establecer las conclusiones del comparativo.

Los experimentos realizados con ReduceTT se desarrollaron en un equipo con CPU Intel Core i3-2310M a una velocidad de 2.1GHz y 3.0GB de memoria RAM.

Resultados

Esta sección se divide en tres etapas. En la primera se muestran los resultados de la selección de variables, obtenidos con la herramienta ReduceTT indicando el número de variables seleccionadas. En la segunda, se muestran los resultados del proceso de clasificación con todas las variables y con las

Tabla I. Bases de datos para las pruebas de selección de variables.

Base de datos	Clases	Variables	Objetos	Tipo de datos	Objetos por clase
Bupa	2	6	345	Numéricos	145, 200
Ionosphere	2	34	351	Numéricos	126, 225
Pima	2	8	768	Numéricos	500, 268
Sonar	2	60	208	Numéricos	111, 97
Wine	3	13	178	Numéricos	59, 71, 48

variables seleccionadas. Y en la tercera etapa se indican los comparativos de los resultados obtenidos en (Castro & Von-Zuben 2009) con los resultados obtenidos en la experimentación con ReduceTT.

Selección de variables

Como parte del proceso de selección de variables se ingresaron las bases de datos de la Tabla I al sistema ReduceTT. En la Tabla II se listan los resultados obtenidos con la herramienta ReduceTT en el proceso de selección de variables.

Tabla II. Resultados de la selección de variables con la herramienta ReduceTT.

BD	Número de variables	Lista de variables
Bupa	3	1, 2, 3
Ionosphere	Sin reducción	Sin reducción
Pima	4	2, 3, 5, 8
Sonar	Sin reducción	Sin reducción
Wine	Ver en Tabla III	Ver en Tabla III

Como se puede ver en la Tabla II con las bases de datos *Bupa* y *Pima* se pudo reducir el número de variables a 3 y 4 respectivamente, con respecto al número de variables originales mostradas en la Tabla I, se logró una reducción del 50% en las bases de datos *Bupa* y *Pima*. En las bases de datos *Ionosphere* y *Sonar* no se logró reducir el número de variables, esto se debe a que la Teoría de Testores no encontró ningún subconjunto de variables (Testores Típicos) que permitiera diferenciar entre las clases, con lo anterior se puede argumentar que para las bases de datos *Ionosphere* y *Sonar* no existe una solución a la partición. Los resultados de la base de datos *Wine* se muestran en la Tabla III.

Una de las características de la Teoría de Testores es que sí existe más de un subconjunto de variables representativas indica todos los subconjuntos posibles. Por lo anterior es que los resultados de la selección de variables para la base de datos *Wine* se presentan en la Tabla III ya que para esa base de datos se encontraron 19 Testores Típicos, los cuales están formados por 2 ó 3 variables cada uno. Comparando con el número de variables que contiene la base de datos Wine originalmente (Tabla I) se estaría reduciendo el número de variables entre un 76.92% y un 84.61%.

En la Tabla IV se muestran los subconjuntos de variables obtenidos con Redes Bayesianas y reportados en (Castro & Von-Zuben 2009) con la intención de tener un referente visual para la comparación de resultados.

Tabla III. Resultados de la selección de variables con la herramienta ReduceTT.

Número de variables	Lista de variables
2	12, 13
2	11, 12
2	10, 13
2	9, 13
2	9, 11
3	8, 11, 13
2	7, 13
2	6, 13
2	6, 12
3	5, 11, 13
3	5, 8, 13
3	4, 8, 13
3	4, 5, 13
2	3, 13
2	3, 12
2	2, 13
2	2, 12
2	1, 13
2	1, 12

Tabla IV. Resultados de la selección de variables obtenidos con Redes Bayesianas (Castro & Von-Zuben 2009).

BD	Número de variables	Lista de variables
Bupa	3	1, 3, 4
Pima	3	2, 6, 7
Wine	6	1, 3, 8, 9, 12, 13

Clasificación

Una vez obtenidos los Testores Típicos se procedió a realizar la clasificación de los datos únicamente con las variables seleccionadas en el proceso anterior y que se muestran en las tablas 2 y 3, para la clasificación de los datos se utilizó el algoritmo *k-means* implementado en Weka (2018). En la Tabla V se muestran los porcentajes del proceso de clasificación de las bases de datos *Bupa* y *Pima* sólo con las variables seleccionadas por ReduceTT.

En la Tabla VI se presentan los resultados del proceso de clasificación para cada Testor Típico de los listados en la Tabla III. Como se puede apreciar, no importa con cual Testor Típico se realice la clasificación, en todos los casos se obtiene el mismo porcentaje de clasificación (70.22 %). Lo anterior se debe a que cualquier subconjunto de variables considerado como Testor Típico, siempre mantendrá la calidad en el proceso de clasificación, tal como si se realizara con las variables originales (70.22 %).

Tabla V. Resultados de la clasificación para las bases de datos Bupa y Pima.

BD	Lista de variables	Porcentaje de clasificación
Bupa	1, 2, 3	55.36%
Pima	2, 3, 5, 8	66.01%
Wine	Ver en Tabla 5	Ver en Tabla 5

Tabla VI. Resultados de la clasificación para la base de datos Wine.

Lista de variables	Porcentaje de clasificación
12, 13	70.22%
11, 12	70.22%
10, 13	70.22%
9, 13	70.22%
9, 11	70.22%
8, 11, 13	70.22%
7, 13	70.22%
6, 13	70.22%

continuación de la tabla VI...

Lista de variables	Porcentaje de clasificación
6, 12	70.22%
5, 11, 13	70.22%
5, 8, 13	70.22%
4, 8, 13	70.22%
4, 5, 13	70.22%
3, 13	70.22%
3, 12	70.22%
2, 13	70.22%
2, 12	70.22%
1, 13	70.22%
1, 12	70.22%

Comparativo

En esta última etapa se colocan en la Tabla VII los porcentajes del proceso de clasificación usando todas las variables, usando las variables de los testores típicos y usando las variables reportadas en Castro & Von-Zuben (2009) obtenidas con Redes Bayesianas. Para la base de datos *Wine* sólo se reportan los resultados en una única línea ya que cualquier combinación de variables obtenida con los Testores Típicos da el mismo porcentaje.

Tabla VII. Resultados comparativos del proceso de clasificación.

BD	%Todas las Variables	%ReduceTT	%Redes Bayesianas
Bupa	55.36%	55.36%	44.98%
Pima	66.01%	66.01%	73.56%
Wine	70.22%	70.22%	70.22%

Como se puede apreciar en la Tabla VII, para la base de datos *Bupa* en el proceso de clasificación se obtiene el mismo porcentaje con todas las variables que al utilizar las recomendadas por los Testores típicos (1, 2, 3), sin embargo, con las variables obtenidas con Redes Bayesianas (1, 3, 4), el porcentaje de clasificación se reduce en un 10.38%, se puede observar que hay dos variables que son comunes en ambos procesos de selección de variables; la variable 1 y la variable 3. Para el caso de la base de datos *Pima* el porcentaje de clasificación obtenido con todas las variables y con las variables del Testor Típico (2, 3, 5, 8)

es el mismo, pero con las variables seleccionadas con Redes Bayesianas (2, 6, 7) el porcentaje de clasificación aumenta en un 7.55%, en este comparativo hay sólo una variable que es común en ambos procesos de selección de variables; la variable 2. Finalmente para la base de datos *Wine* se obtiene el mismo porcentaje con todas las variables, con las variables obtenidas con Redes Bayesianas (1, 3, 8, 9, 12, 13) y con todos los Testores Típicos, ver la combinación de variables en la Tabla VI.

Discusión

Como se puede observar en la sección de resultados, una de las principales características que tiene la Teoría de Testores, es que garantiza que no se pierde la calidad en la clasificación de la información al utilizar las variables resultantes del proceso de selección de variables. Lo anterior se puede sustentar con los resultados reportados en la Tabla VII en la cual se puede ver que los porcentajes de clasificación se mantienen aún con menos variables. La teoría de testores no busca reducir o aumentar el porcentaje de certeza en procesos de clasificación, su objetivo es encontrar, si existe, un conjunto con menos variables que permita mantener exactamente el mismo porcentaje de clasificación, es por ello que, en caso de existir un conjunto de variables que pudiera mejorar o empeorar la calidad de la clasificación la teoría de testores no lo encontraría. En cambio las Redes Bayesianas reportadas en (Castro & Von-Zuben 2009) no garantizan dicha situación, como se puede ver en la misma tabla 7, para esta última se presentan las tres situaciones posibles, en el primer caso el porcentaje de clasificación es menor que al usar todas las variables, en el segundo caso el porcentaje aumenta y en el tercer caso el porcentaje sí se mantiene.

Otra situación que vale la pena destacar es, si existe por lo menos una solución posible al problema, la teoría de testores la va a encontrar por su naturaleza de ser una técnica de enfoque combinatorio, ya que realiza una búsqueda completa. Y en algunos casos como los presentados en la Tabla II si no existe un conjunto de variables que permita diferenciar entre las clases no se encontrarán Testores Típicos.

Uno de los objetivos centrales de esta investigación es tener elementos para elegir una de las dos técnicas comparadas en el presente trabajo de investigación para continuar con otra fase de investigación que es hacer selección de variables en datos temporales para mantener la calidad en procesos de clasificación.

Trabajos futuros

Uno de los trabajos principales que se pretende realizar como extensión a esta investigación es determinar la posibilidad de adecuar las bases de la minería de secuencias a la teoría de testores, con la finalidad de abordar problemas de tipo temporal, como las series de tiempo.

Agradecimientos

Agradezco las observaciones y recomendaciones realizadas por los revisores para mejorar el presente documento.

Referencias

Aeberhard, S., Coomans D. & O. de Vel. 1992. Comparison of classifiers in high dimensional settings. Tech. Rep. no. 92-02, Dept. of Computer Science and Dept. of Mathematics and Statistics, James Cook University of North Queensland.

Alba, E., Santana, R., Ochoa A. & M. Lazo. 2000. Finding typical testors by using an evolutionary strategy, Workshop on Pattern Recognition, Portugal.

Bache, K. & M. Lichman. 2013. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Castro, P. & F.J. Von-Zuben. 2009. Learning bayesian networks to perform feature selection, International Joint Conference on Neural Networks, pp. 467-473. DOI: 10.1109/IJCNN.2009.5178817.

Díaz, G., Piza, I., Sánchez, G., Mora, M., Reyes, O., Cardenas, A. & C. Aguirre. 2011. Typical testors generation based on an evolutionary algorithm. Pp. 58-65, *In:* Proceedings of the 12th international conference on Intelligent data engineering and automated learning (IDEAL'11), Berlin, Heidelberg, Springer-Verlag.

Gorman, R.P. & Sejnowski T.J. 1988. Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets, in Neural Networks, Vol. 1: 75-89.

Ochoa-Somuano, J. 2005. Técnicas de Selección de Atributos para la Categorización Automática de Escenas Visuales, Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), Tesis de Maestría, Cuernavaca, Morelos.

Ochoa-Somuano, J. 2013. ReduceTT - Reduce typical testors, Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), Laboratorio de Inteligencia Artificial, Cuernavaca, Morelos.

Ochoa-Somuano, J., Valdés-Marrero, M.A., Moctezuma, I. & C.A. Esquivel. 2007. Dimension reduction in images databases using the logical combinatorial approach. Pp. 260-265, *In:* Innovations Advanced Techniques in Computer and Information Sciences and Engineering, Springer, ISBN: 9781402062674.

Ruiz, J., Guzmán, A. & J.F. Martínez. 1999. Enfoque lógico combinatorio al reconocimiento de patrones, Avance en Reconocimiento de Patrones, Centro de Investigación en Computación, Instituto Politécnico Nacional, México.

Santiesteban, Y. & A. Pons. 2003. LEX: A new algorithm for the calculus of all typical testors. Revista Ciencias Matemáticas, 21(1).

Santos, J. & J. Carrasco. 2004. Feature selection using typical testors applied to estimation of stellar parameters. Computación y Sistemas, CIC-IPN 8(1): 15-23.

Sigillito, V.G. 1990. Pima indians diabetes database. National Institute of Diabetes and Digestive and Kidney Diseases, The Johns Hopkins University, RMI Group Leader Applied Physics Laboratory.

Sigillito, V.G., Wing S.P. Hutton, L.V. & K.B Baker. 1989. Classification of radar returns from the ionosphere using neural networks. APL Technical Digest 10: 262-266.

Weka. 2018. Waikato Environment for Knowledge Analysis, versión 3.7, The University of Waikato, Hamilton, New Zealand, http://www.cs.waikato.ac.nz/ml/weka/downloading.html.