Comparación de técnicas y herramientas para la selección de variables

Jorge Ochoa-Somuano[1]*, José Francisco Delgado-Orta[1], Ángel Salvador López-Vásquez[1], Ángel Antonio Ayala-Zúñiga[1], Omar Antonio Cruz-Maldonado, María Alejandra Menéndez-Ortiz[1] & Omar de Jesús Reyes-Pérez [2]

Resumen

La selección de variables permite reducir la cantidad de información necesaria para realizar procesos como la 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 presentan resultados comparativos de técnicas que permiten aplicar la selección de variables, se llevó a cabo la clasificación con los mismos conjuntos de datos para cada herramienta utilizando el algoritmo k-medias simple incluido en Weka; con los resultados se realizó la comparación. Se usaron los algoritmos de las técnicas que incluyen R, Weka, Tanagra y un software propio denominado ReduceTT en el cual se implementó la técnica teoría de testores, ésta forma parte del enfoque lógico combinatorio. Una característica de la última aplicación, que la hace diferente a las demás, 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.*

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

Abstract

Feature selection can reduce the amount of information required to perform the classification process, which is one of the major benefits that can be obtained as a direct impact in saving time to process information, and thus, there a reduction in the computational costs. *This paper presents comparative results of applying the techniques for feature selection, the classification was performed with the same data sets for each tool using the k-means algorithm simply included in Weka, with the results we compared. We used the algorithms techniques including R, Weka, Tanagra and an own software called ReduceTT in which the testors theory technique was implemented, this approach is part of combinational logic. A feature of the last application that makes it different from the others, is guaranteed to maintain quality in the classification for any subset of variables is obtained as a result in the feature selection.*

Key words: typical testors, feature selection, dimension reduction, representation space, attributes, objects, data, classification.

Recibido: 10 de diciembre de 2022.

Aceptado: 27 de marzo de 2023.

[1] *Instituto de Industrias, Universidad del Mar campus Puerto Escondido. Ciudad Universitaria, Vía Sola de Vega km 1.5 Carretera Puerto Escondido- Oaxaca. San Pedro Mixtepec-Juquila, Oaxaca, México, 71980.*

[2] *Instituto de Estudios Internacionales. Universidad del Mar campus Huatulco. Ciudad Universitaria, La Crucecita, Huatulco 70989, Oaxaca, México.*

* **Autor de correspondencia:** *ochoa@zicatela.umar.mx (JOS)*

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 diferentes técnicas 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 describe la secuencia de las actividades realizadas para la elaboración de este documento, 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. Finalmente, se tiene una sección de referencias para quien desee profundizar en el tema.

Herramientas utilizadas

Las herramientas que se utilizaron para hacer las comparaciones de las técnicas de selección de variables son las siguientes:

R (2022). Es un sistema para análisis estadísticos y gráficos creado por Ross Ihaka y Robert Gentleman. Se distribuye gratuitamente bajo los términos de la GNU (2022) *General Public Licence*. Para la experimentación se utilizó un paquete denominado *FSelector* (CRAN 2022). Una ventaja de este software es que cuenta con una interfaz que permite al usuario realizar sus propios programas, integrando las diferentes funciones que tiene implementadas, para realizar análisis más complejos (Contreras *et al*. 2010).

Weka (2020) (*Waikato Environment for Knowledge Analysis*, *Waikato* entorno para el Análisis del Conocimiento) es una extensa colección de algoritmos de Máquinas de conocimiento implementados en *Java* y desarrollados por la universidad de Waikato en Nueva Zelanda. Además, *Weka* contiene las herramientas necesarias para realizar transformaciones sobre los datos, tareas de clasificación, regresión, agrupamiento, asociación y visualización (García 2006).

Tanagra. Se define por el autor Rakotomalala (2005), profesor de la Universidad de Lione, Francia, como un software de minería de datos libre para la investigación y la educación. Tanagra sólo está disponible para Windows y se puede descargar libremente.

ReduceTT (Ochoa 2013). (*Reduce Typical Testors*) es una adaptación del sistema *ReduClass* desarrollado en Ochoa (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, Alba *et al*. 2000, Santos *et al*. 2004, Ochoa *et al*. 2007, Díaz *et al*. 2011) de una base de datos.

Bases de datos

El término base de datos (BD) se usará para referirse a un archivo de texto formado por filas y columnas. Cada fila representa una instancia y cada columna constituye un atributo de la instancia. Los atributos son datos numéricos separados entre sí por una coma.

Las bases de datos utilizadas para los experimentos se tomaron de diferentes fuentes. A continuación, se da una breve descripción de cada una de ellas:

Diabetes indios pima (BD-1). Base de datos de diabetes de los indios pima (Sigillito 1990), propiedad del Instituto Nacional de Diabetes y Enfermedades Digestivas y del Riñón. Contiene 768 instancias (objetos) de las cuales 500 son casos negativos y 268 positivos.

Imágenes de pastos y veredas (BD-2). Base de datos de imágenes segmentadas (Brodley 1990), se tomaron dos de las siete clases originales, tiene información de 400 imágenes de las cuales 200 son de pastos y 200 de veredas.

Imágenes de escenas (BD-3). Es una base de datos binaria, generada en (Ochoa 2005), con información de tres diferentes tipos de imágenes (cada una con 21 variables): flores de primavera, montañas suizas y parque de Yellowstone. Formada por 48, 30 y 45 instancias respectivamente, lo cual da un total de 123 instancias.

Imágenes de escenas reducidas (BD-4). Es la misma base de datos binaria generada en Ochoa (2005), con la diferencia que se eligieron sólo nueve de las 21 variables de forma aleatoria.

Imágenes segmentadas (BD-5). Base de datos tomada de Brodley (1990), en la cual se tienen 810 instancias divididas en siete clases de la siguiente forma: 125 de fachadas de ladrillos, 110 de avenidas de cemento, 122 de follaje, 123 de pasto, 94 de veredas, 110 de cielo y 126 de ventanas.

En la tabla I se muestran, a manera de resumen, las características de las bases de datos para las pruebas de selección de variables.

Desarrollo de experimentos

Los experimentos se desarrollaron en un equipo con CPU Intel Core i3-2310M a una velocidad de 2.1GHz y 3.0GB de memoria RAM. Cada prueba se realizó con técnicas implementadas en R, Weka, Tanagra y ReduceTT.

Los experimentos consisten en tomar cada base de datos (BD) y en cada una de las herramientas anteriores realizar la selección de atributos con las técnicas (tcas) que tienen implementadas, los resultados que se obtienen son:

Tabla I. Características de las bases de datos utilizadas para las pruebas de selección de variables.

Base de datos	Clases	Variables	Objetos	Tipos de datos	Objetos por clase
BD-1	2	8	768	Numéricos	500, 268
BD-2	2	19	400	Numéricos	200, 200
BD-3	3	21	123	Binarios	48, 30, 45
BD-4	3	9	123	Binarios	48, 30, 45
BD-5	7	19	810	Numéricos	125, 110, 122, 123, 94, 110, 126

las variables seleccionadas (no reportadas aquí por limitaciones de espacio), el número de variables seleccionadas (nvs), el porcentaje de clasificación con todas las variables (%tv) y el porcentaje de clasificación con las variables seleccionadas (%vs). El símbolo (---) se utiliza para indicar que no se pudo obtener resultado en la técnica correspondiente.

Resultados

Esta sección se divide en dos etapas. En la primera, se muestran los resultados de la selección de variables, obtenidos en cada herramienta y en particular con cada una de las técnicas que tienen implementadas, indicando el número de variables seleccionadas. En la segunda, se muestran los comparativos de los resultados obtenidos con la clasificación utilizando todas las variables y con los resultados de la clasificación con las variables seleccionadas. En ambas etapas se agrupan los resultados por herramienta.

Selección de variables

En la tabla II se pueden ver los resultados obtenidos con la herramienta R.

En la tabla III se pueden observar los resultados obtenidos con Weka y en la En la tabla IV se muestran los resultados obtenidos en la selección de variables con la herramienta Tanagra.

En la tabla V se enlistan los resultados obtenidos con la herramienta ReduceTT, en esta tabla cambia un poco la representación, debido a que con la herramienta ReduceTT, se puede obtener más de un subconjunto reducido de variables, a estos subconjuntos se les denomina Testores Típicos (TT) [10].

Clasificación

En las siguientes tablas (VI-IX), específicamente en la última fila, se incorporan los resultados de clasificación obtenidos con todas las variables, así como los resultados que se obtuvieron al utilizar los subconjuntos de variables obtenidos por cada una de las técnicas implementadas en las herramientas antes mencionadas. El

Tabla II. Resultados de la selección de variables (nvs) con la herramienta R.

Tcas/ BD	BD-1	BD-2	BD-3	BD-4	BD-5
CFS filter	3	1	5	3	6
Chi-squared filter	5	5	5	5	5
OneR algorithm	5	5	5	5	5
Rrelief filter	2	2	2	2	2
Consistensy-based filter	6	1	4	5	8
RandomForest filter	5	5	5	5	5
Best-first search	2	1	5	3	5
Greedy search	2	1	2	3	4
Hill climbing search	2	11	10	6	12
Cutoff	1	9	2	2	11
Entropy-basedfilters	1	9	4	2	11

Tabla III. Resultados de la selección de variables (nvs) con la herramienta Weka.

Tcas/ BD	BD-1	BD-2	BD-3	BD-4	BD-5
CfsSubsetEval	4	11	10	7	8
ConsistencySubsetEval	---	1	4	5	8
FilteredSubsetEval	3	10	5	3	6
LatentSemanticAnalysis	---	2	14	7	19

Tabla IV. Resultados de la selección de variables (nvs) con la herramienta Tanagra.

Tcas/ BD	BD-1	BD-2	BD-3	BD-4	BD-5
Backward-logit	4	0	0	0	0
Fisher filtering	6	15	17	9	17
Forward-logit	4	1	0	0	0
ReliefF	3	8	9	4	11
Runs filtering	4	14	12	7	19
Stepdisc	5	12	13	7	12

Tabla V. Resultados de la selección de variables (nvs) con la herramienta ReduceTT.

Tcas/ BD	BD-1	BD-2	BD-3	BD-4	BD-5
TT-1	4	1	5	9	19
TT-2		1	5		
TT-3		1			
TT-4		1			
TT-5		1			
TT-6		1			
TT-7		1			

clasificador que se utilizó es el k-medias simple implementado en Weka.

En la tabla VI se muestran los resultados de la clasificación con las variables seleccionadas por los algoritmos que ya tiene implementados la herramienta R.

Como se puede observar, se realizó la clasificación con las variables seleccionadas de 11 técnicas. Respecto al porcentaje obtenido con todas las variables, con la BD-1 se obtuvieron 10 valores mayores y con la BD-5 se lograron 5 resultados superiores, sin embargo, pero también hay 3 valores que están por debajo. Con la BD2 se mantuvo el mismo porcentaje. Finalmente, con la BD-3 y la BD-4 que son de tipo binario, para todos los casos los porcentajes obtenidos fueron menores que al utilizar todas las variables. De acuerdo con los resultados de la Tabla VI, se puede concluir que con las técnicas que ahí se reportan y con las bases de datos de prueba utilizadas no se garantiza que se pueda mantener la calidad en la clasificación.

En la tabla VII se muestran los resultados

Tabla VI. Porcentajes de los procesos de clasificación de las variables seleccionadas (%vs) en R.

Tcas/ BD	BD-1	BD-2	BD-3	BD-4	BD-5
CFS filter	69.3	100	73.2	73.2	72.6
Chi-squared filter	67.2	100	73.2	62.6	52.5
OneR algorithm	66.3	100	62.6	62.6	22.8
Rrelief filter	73.4	100	70.0	70.0	58.8
Consistensy-based filter	67.3	100	62.6	78.9	52.2
RandomForest filter	67.2	100	60.2	62.6	66.3
Best-first search	74.1	100	65.1	62.6	71.6
Greedy search	69.5	100	95.1	70.0	61.0
Hill climbing search	74.1	100	96.8	86.2	56.5
Cutoff	73.8	100	70.0	70.0	55.4
Entropy-basedfilters	73.8	100	73.2	70.0	48.1
%tv	66.8	100	100	87.8	56.7

Tabla VII. Porcentajes de los procesos de clasificación de las variables seleccionadas (%vs) en Weka.

Tcas/ BD	BD-1	BD-2	BD-3	BD-4	BD-5
CfsSubsetEval	71.9	100	62.6	62.6	67.3
ConsistencySubsetEval	---	100	62.6	78.9	52.2
FilteredSubsetEval	69.3	100	73.2	73.2	72.6
LatentSemanticAnalysis	---	55.8	66.2	84.6	56.7
%tv	66.8	100	100	87.8	56.7

obtenidos en la clasificación al utilizar las variables obtenidas en Weka.

El comportamiento de los resultados para este caso, son similares a los obtenidos con las técnicas que incorpora la herramienta R. La única excepción es que en la BD-2 no se mantiene el porcentaje en todos los casos. De la misma manera y con base en los resultados, se puede concluir que con las técnicas que ahí se reportan y con las bases de datos de prueba no se garantiza que se pueda mantener la calidad en la clasificación.

En la tabla VIII se aprecian los resultados obtenidos en la clasificación con las variables obtenidas en Tanagra.

Nuevamente el comportamiento de los resultados se mantiene similar a los reportados en R y Weka.

Una diferencia significativa que se puede observar en la tabla IX, con respecto a las Tablas VI, VII y VIII, es que los resultados del proceso de clasificación al utilizar todas las variables, así como los resultados al emplear únicamente los conjuntos de variables seleccionadas, dan exactamente los mismos valores. Es decir, la calidad de la clasificación se mantiene al hacer una selección de variables con la Teoría de Testores.

En la tabla IX se indican los resultados de los porcentajes que se obtuvieron con la clasificación usando las variables seleccionadas con ReduceTT.

Tabla VIII. Porcentajes de los procesos de clasificación de las variables seleccionadas (%vs) en Tanagra.

Tcas/ BD	BD-1	BD-2	BD-3	BD-4	BD-5
Backward-logit	65.5	0.0	---	---	---
Fisher filtering	67.3	100	100	87.8	56.7
Forward-logit	65.5	100	---	---	---
ReliefF	59.6	100	62.6	62.6	59.8
Runs filtering	67.2	100	98.4	62.6	56.7
Stepdisc	66.5	100	62.6	62.6	55.1
%tv	66.8	100	100	87.8	56.7

Tabla IX. Porcentajes de los procesos de clasificación de las variables seleccionadas (%vs) en ReduceTT.

Tcas/ BD	BD-1	BD-2	BD-3	BD-4	BD-5
TT-1	66.8	100	100	87.8	56.7
TT-2		100	100		
TT-3		100			
TT-4		100			
TT-5		100			
TT-6		100			
TT-7		100			
%tv	66.8	100	100	87.8	56.7

Una diferencia significativa que se puede observar en la tabla IX, con respecto a las tablas VI, VII y VIII, es que los resultados del proceso de clasificación al utilizar todas las variables, así como los resultados al emplear únicamente los conjuntos de variables seleccionadas, dan exactamente los mismos valores. Es decir, la calidad de la clasificación se mantiene al hacer una selección de variables con la Teoría de Testores.

Conclusiones

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. Otra situación que vale la pena destacar es, que al ser una técnica de enfoque combinatorio permite encontrar la solución, si es que ésta existe, ya que realiza una búsqueda completa. Las otras técnicas al ser de tipo ranking, no exploran todas las combinaciones y eso hace que sea posible que no exploren la mejor solución. Con base en los resultados, se puede decir que, la teoría de testores mantuvo los mismos resultados de clasificación para todas las bases de datos y *Fisher Filtering* mantuvo los mismos resultados de clasificación en las bases de datos de la 2 a la 5, de las otras técnicas ninguna mantuvo en todas las bases de datos el mismo porcentaje de clasificación. Sin embargo, de estas dos técnicas, es el enfoque lógico combinatorio el que obtiene los subconjuntos de variables más pequeños.

Referencias

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

Brodley, C. 1990. Datos de Imágenes Segmentadas, Grupo de Visión, Universidad de Massachussets, ftp://ftp.ics.uci.edu/pub/machine-learning-databases/image/CRAN. 2022. Página web oficial de The Comprehensive R Archive Network. FSelector package (Paquete para selección de atributos). http://cran.r-project.org/

Contreras, J., E. Molina & P. Arteaga. 2010. Introducción a la Programación Estadística con R para Profesores, Universidad de Granada, España, 161 pp.

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

GNU. 2022. Página web oficial de GNU (Free Software Foundation). http://www.gnu.org/licenses/gpl-2.0.html

García, D. 2006. Manual de Weka, https://matema.ujaen.es/jnavas/web_recursos/archivos/weka%20master%20recursos%20naturales/tutorial2%20weka%20mediano.pdf: fecha de consulta: 20 de junio de 2022.

Ochoa, J. 2013. ReduceTT - Reduce Typical Testors, Centro Nacional de Investigación y Desarrollo Tecnológico (cenidet), Laboratorio de Inteligencia Artificial, Cuernavaca, Morelos.

Ochoa, 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, J., Valdés-Marrero, M. A., Moctezuma Cantorán, I. & C. Ayala. 2007. Dimension Reduction in Images Databases using the Logical Combinatorial Approach, Innovations Advanced Techniques in Computer and Information Sciences and Engineering, Springer, ISBN: 9781402062674.

Rakotomalala, R. 2005. TANAGRA: un logiciel gratuit pour l'enseignement et la recherche", in Actes de EGC'2005, RNTI-E-3, 2: 697-702.

R. 2022. Página web oficial de R Proyect (Roasted Marshmallows). http://www.r-project.org/

Ruiz, J., A. Guzmán & F. Martínez. 1999. Enfoque Lógico Combinatorio al Reconocimiento de Patrones, Avance en Reconocimiento de Patrones, Centro de Investigación en Computación, Instituto Politécnico Nacional, México, ISBN: 9701823841.

Santos, J., A. Carrasco & J.F. Martínez-Trinidad. 2004. Feature Selection Using Typical Testors Applied to Estimation of Stellar Parameters, Computación y Sistemas, CIC-IPN 8(1): 15-23.

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

Weka. 2020. Página oficial de Waikato Environment for Knowledge Analysis, The University of Waikato, Hamilton, New Zealand, http://www.cs.waikato.ac.nz/ml/weka/downloading.html: fecha de consulta: 15 de junio de 2022.