ISSN 1870-4069

Selección de atributos mediante separación de centroides

Héctor Jiménez-Salazar, Alfredo Mateos-Papis, Christian Sánchez-Sánchez

Universidad Autónoma Metropolitana, Unidad Cuajimalpa,
Ciudad de México, México
{hjimenez,amateos,csanchez}@correo.cua.uam.mx

Resumen. Este trabajo presenta la aplicación de un método basado en la separación de centroides para seleccionar atributos, cuyos resultados fueron confrontados con los obtenidos a través del método de ganancia de información. Estos métodos se aplicaron al problema de determinar las materias previamente aprobadas que influyen en la aprobación de una nueva materia de un plan de estudios universitario. Los experimentos consideraron varias nuevas materias de interés y, para conocer la efectividad de cada uno de los métodos, los atributos seleccionados (materias previamente cursadas) se usaron en un clasificador de árbol de decisión. Se destaca que el uso de separación de centroides mejora la clasificación con relación a lo que se obtiene por ganancia de información pero, además, los atributos que la separación de centroides seleccionó resultaron ser los que se esperaba obtener cualitativamente.

Palabras clave: selección de atributos, clasificación, desempeño escolar.

Feature Selection through Centroids Separation

Abstract. This paper presents the application of a method based on the separation of centroids to select attributes, whose results were compared with those obtained through the information gain method. These methods were applied to the problem of determining the previously approved subjects that influence the approval of a new subject of a university curriculum. The experiments considered several new subjects of interest and, in order to determine the effectiveness of each one of the methods, the selected attributes (previously coursed subjects) were used in a decision tree classifier. It is emphasized that the use of separation of centroids improves the classification in relation to what is obtained through information gain but, in addition, the attributes selected through separation of centroids turned out to be those that were expected to be obtained qualitatively.

Keywords: feature selection, classification, school performance.

1. Introducción

La selección de atributos de instancias de una colección de datos puede ser definida como el proceso en el que se escoge el subconjunto mínimo de atributos a partir de un conjunto original [13] de tal forma que mejore el desempeño del clasificador que los usa. Este proceso sigue siendo un problema de gran importancia ya que impacta una amplia variedad de aplicaciones pero, además, porque no hay un método general que pueda ser aplicado en cualquier contexto [5]. En los problemas reales comúnmente se cuenta con poca información y este hecho impide aplicar con efectividad los métodos de selección existentes. En este trabajo se experimenta con un método basado en la separación de los centroides de las clases para identificar los atributos que influyen en la clasificación. Aquí se reporta que los atributos que separan centroides tienen ventajas con respecto al que usa árboles de decisión, basado en ganancia de información de los atributos.

El problema es identificar cuáles materias de un plan de estudios influyen en otras, es decir, si habiendo aprobado la materia A hay más seguridad de aprobar la materia objetivo B, o bien, en términos escolares, A debe ser requisito de B, pues en B se usan los conocimientos presentados en A. Si bien este problema se resuelve habitualmente con la opinión de especialistas, consideramos que hay múltiples factores que están involucrados. Particularmente, disponemos de información sobre las evaluaciones realizadas a los alumnos en un plan de estudios universitario, en el cual hay casos de materias que no están seriadas entre sí, mientras que, normalmente, en otros planes de estudio materias similares sí lo están. Así pues, con esta información puede conocerse el efecto de no seriar algunas materias; algo que no es posible en cualquier plan de estudios seriado de la forma acostumbrada.

Para ciertos casos el problema podría ser resuelto directamente por algún método bien conocido como una prueba de hipótesis. Por ejemplo, considerando como variables aleatorias la proporción de aprobados en cada grupo que toma B en dos casos: quienes han aprobado previamente A, x, y quienes no, y, la hipótesis a probar sería $\bar{x} > \bar{y}$. Este procedimiento debería ser repetido para cada par de materias donde creemos que existe relación. Otro enfoque podría tratar, más bien, de descubrir cuáles materias, de entre varias previamente aprobadas, influyen en la materia objetivo. Este último enfoque fue el seguido pues, además, existen diversas herramientas que pueden ser aplicadas y comparar su efectividad. Justamente, se utilizó la técnica de análisis basada en la separación de centroides, y el clasificador por árboles de decisión para validar los atributos obtenidos.

En lo que resta de este artículo, se exponen algunos trabajos relacionados con el tema, en la sección 2; los métodos utilizados, en la sección 3; los experimentos, en la sección 4; y las conclusiones del presente trabajo en la sección 5.

2. Trabajos relacionados

Dentro de los trabajos cercanos a la identificación de factores que influyen en la aprobación de una materia, hay trabajos relacionados con el contenido

de los cursos impartidos y la generación de rutas flexibles (ordenamiento de los contenidos) que se adapte al perfil de los alumnos. Por ejemplo. Idris et al. [9] ofrece la secuencia personalizada de temas de un curso de Java. Los autores usaron una red neuronal. Los atributos evaluados por la red neuronal son el conocimiento del alumno de algunos temas de programación y el conocimiento sobre el dominio. Por otro lado, Chen [4] propuso un sistema multi-agente e-learning basado en algoritmos genéticos que permite guiar a los alumnos sobre una serie de temas relacionados a un curso. Años después Dwi & Basuki [2] propusieron un sistema similar para el aprendizaje del idioma inglés.

Más allá de los contenidos de los cursos Kovacic [11] utilizó la clasificación basada en árboles de decisión para prodecir, tempranamente, el éxito de los estudiantes basándose en atributos socio-demográficos (edad, etnicidad, género, y educación, entre otros) y de su entorno de estudio (cursos del programa y bloque de cursos). El conjunto de datos provino de una universidad politécnica de Nueva Zelanda. De cuatro árboles de decisión construidos, el que obtuvo mejor desempeño refiere una exactitud del 60.5 %, baja, argumentan, debido a que hay otros elementos, no contenidos en el conjunto de datos que pueden influir en el éxito de los estudiantes. En este trabajo hay muchos factores que influyen en la aprobación de materias. Incorporar este tipo de información, a las colecciones utilizadas en los experimentos expuestos, podría ayudar a determinar de manera más precisa cuál es el grado de influencia de los conocimientos previos y descubrir la importancia de otros factores.

Por otro lado, Barrak et al. [1] presentan un enfoque para estimar la Calificación Promedio Final (Grade Point Average) mediante Arboles de Decisión. Como herramienta principal utilizaron WEKA. Fueron generados Arboles de Decisión (algoritmo J48) por periodo, considerando los principales cursos. Este es un trabajo muy parecido al que se ha tratado en el presente trabajo. Desafortunadamente no contiene datos sobre su desempeño.

Es importante mencionar que aún cuando los datos académicos sean vastos puede que dentro de ellos se encuentren algunos atributos que no aporten mucho al análisis y procesamiento, aunque siempre es necesario tener el conocimiento del dominio sobre el que está trabajando, existen algunas técnicas que ayudan a seleccionar atributos. A continuación se ofrece mayor información al respecto.

El problema de selección de atributos es un problema que sigue siendo de interés pues no hay solución general. Según Kohavi [10] obtener el mejor conjunto de atributos puede ser un problema intratable y muchos problemas relacionados a la selección de atributos han mostrado tener una complejidad alta (NP-hard) [3]. La adecuada selección de atributos no solo tiene consecuencias dentro de la solución del problema, además, impacto en la problemática de grandes áreas de desarrollo [7]. Algunas de las técnicas de selección de atributos más usadas son: *Information Gain*, *Gain Ratio, Symmetrical Uncertainty* [8], Chi-Squared [12], Gini-Index y el Análisis de componentes Principales (PCA por sus siglas en inglés), la cual es una técnica que reduce la dimensionalidad y ayuda a identificar los atributos más influyentes. Aunque algunos algoritmos de clasificación están diseñados para escoger las características más relevantes y dejar fuera las irrele-

vantes, como los árboles de decisión, y las redes neuronales multicapa, estos se benefician de métodos complementarios para elegir atributos. Además, en el caso de los árboles de decisión, es posible proceder con conjuntos de datos pequeños y pocos atributos, como los documentos cortos: se ha constatado que la selección de atributos beneficia los resultados de la clasificación de documentos [16].

En relación a la técnica de separación de centroides hay poco uso reportado al respecto. Un conjunto de técnicas derivadas del trabajo de R. Fisher [6] sobre análisis discriminante, como los clasificadores lineales (LSM, SVM, etc.) se han usado con la implícita referencia a la selección de atributos que separan clases de instancias. Estos métodos exigen, en cada caso, condiciones sobre la colección (distribución o tamaño, por ejemplo) y cálculos que pueden ser largos. El antecedente del presente trabajo se encuentra en G. Salton [15] quien reporta el uso del centroide para calcular la densidad de documentos en función de los términos que mejor los distinguen, lo cual es una simplificación de los métodos derivados del análisis discriminante.

3. Metodología

En las subsecciones siguientes se expone cada uno de los métodos utilizados.

3.1. Árboles de decisión

Un algoritmo de clasificación es capaz de extraer patrones de un conjunto de entrenamiento, que son útiles para crear un modelo para clasificar nuevos datos. El algoritmo de Arboles de Decisión crea un modelo en la forma de un árbol, y funciona dividiendo el conjunto de datos en subconjuntos más pequeños, por medio de evaluar cada uno de los atributos de acuerdo con la Ganancia de Información(GI) para formar el árbol [14]. La GI de un atributo considera: la entropía de la colección:

$$H(C) = \sum_{clases} \Pr_{clase} \log_2(\frac{1}{\Pr_{clase}}), \tag{1}$$

la entropía relativa a un atributo:

$$H(A) = \sum_{v \in A} \frac{|C_{v \in A}|}{|C|} H(C_{v \in A}), \tag{2}$$

y la ganancia de un atributo:

$$A : GI(A) = H(C) - H(A). \tag{3}$$

Algunas de las razones por las que se eligió este algoritmo son las siguientes: Dentro del diseño de este algoritmo se seleccionan solo atributos con un umbral base, $GI > 0$, y los demás quedan fuera, haciendo implícitamente una selección de atributos. También permite lidiar con conjuntos de datos que tengan clases

desbalanceadas, ya que busca patrones que caracterizan todas las clases, lo cual se expresa por reglas de clasificación que facilitan el análisis visual, en forma de un árbol.

Con cada conjunto de datos se generó un modelo utilizando el Algoritmo de Árboles de Decisión en Python seleccionando el método de GI para formar el árbol. Cada modelo fue evaluado mediante validación cruzada, partiendo el conjunto de datos original en conjuntos de entrenamiento y prueba de manera aleatoria un número determinado de veces; en este caso 10 veces. Cada vez que se realiza la clasificación se evalúa la exactitud y, finalmente se promedio la exactitud de los resultados. La exactitud se calcula con la fórmula: *ACC =(número de predicciones correctas del clasificador)/(número total de predicciones).*

Con la finalidad de afinar el modelo se fue incrementando el umbral de GI, para ir dejando fuera uno por uno de los atributos y calcular la exactitud de cada modelo.

3.2. Separación de centroides

Este método, es intuitivamente muy simple. En él se considera el conjunto de datos representado en el modelo de espacio vectorial (MEV), luego para cada clase de instancias puede calcularse su centroide y, con ellos, conocer su similitud original, esto es con todos los atributos. Se procede entonces, eliminando vorazmente los atributos y midiendo nuevamente la similitud entre los centroides: si se elimina un atributo p y la similitud entre centroides aumenta, significa que p es buen atributo para representar las instancias pues ayuda a distinguir entre instancias de diferentes clases. A continuación se exponen algunas fórmulas referentes a los conceptos mencionados.

Representación en el MEV. Sea el conjunto de instancias, C, $n = |C|$, con vocabulario de atributos V_C, y $m = |V_C|$, donde para cada $I_j \in C$, tf_{ij} representa la frecuencia del elemento $t_i \in V_C$ en la instancia I_j, $df_i = |\{k \in C | i \text{ ocurre en } k\}|$ e $idf_i = \log_2(\frac{2n}{df_i})$. La representación de una instancia de C será un vector de dimensión $|V_C|$, donde cada dimensión se asocia con un orden dado a los elementos de V_C: la entrada i de $I_j \in C$ está dada por el valor: $tf_{ij} \cdot idf_j$. Los vectores de C se usan normalizados: multiplicado cada uno por el inverso de su magnitud, $|\boldsymbol{x}| = \sqrt{\sum_{k=1}^{m} x_k^2}$.

Centroide y similitud. Dado un conjunto de instancias representadas por vectores $B = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_n}\}$, su centroide se calcula mediante la fórmula:

$$\bar{B} = \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{v_k}. \tag{4}$$

Asimismo, la similitud de dos vectores está dada por:

$$sim(\boldsymbol{a}, \boldsymbol{b}) = \sum_{k=1}^{m} a_k \cdot b_k. \tag{5}$$

Separación entre clases. Supongamos que C se compone de dos clases de instancias: C_a y C_b con centroides $\bar{C_a}$ y $\bar{C_b}$. Si al eliminar $p \in V_C$ de C, obtenemos $C'_{(p)}$ y $sim(\bar{C_a}, \bar{C_b}) < sim(\bar{C_a}', \bar{C_b}')$ significa que el atributo p ayudaba a separar las clases C_a y C_b. En suma, podemos calcular para cada atributo el aporte de separación entre clases y seleccionar solamente los que separan, esto es cuando:

$$SC(p) = sim(\bar{C_a}', \bar{C_b}') - sim(\bar{C_a}, \bar{C_b}) > 0. \quad (6)$$

4. Experimentos

Antes de presentar la forma en que se realizaron los experimentos, se describirán los datos utilizados.

4.1. Conjuntos de datos

Se tuvo acceso al kardex, específicamente al registro de las evaluaciones de los alumnos de la carrera Tecnologías y Sistemas de Información de la UAM-C del periodo 2007 a 2014. La información que pudo integrarse para el análisis se constituyó por registros formados por 4 valores: Id, identificador del alumno; U, clave de la materia evaluada; C, calificación obtenida; y T, periodo en que se realizó la evaluación. Con el fin de homogenizar la información se extrajo, en primer lugar, la "población media", alumnos cuyo número de materias aprobadas en el primer año está a lo más a una desviación estándar de la media ($[3, 8]$): de los registros originales, 7,144, se trabajó únicamente con 4,174 registros. Es importante aclarar que un alumno puede tener varias evaluaciones en una misma materia en tanto no la apruebe. A partir de esta información, en cada materia objetivo, se extrajeron los registros de todos los alumnos que fueron evaluados por primera vez. A cada uno de estas instancias se asociaron las materias que habían sido previamente aprobadas y que podían tener o no relación con la materia objetivo.

Haremos referencia a las siguientes materias: ARQ: Arquitectura de Computadoras. ED: Estructura de Datos, HYC: Historia y Cultura de la Computación, IPM: Introducción al Pensamiento Matemático, LPL: Lógica y Programación Lógica, MD1: Matemáticas Discretas I, MD2: Matemáticas Discretas II, PE: Programación Estructurada, PWE: Programación Web Estático, SO: Sistemas Operativos, SUST: Seminario de Sustentabilidad, TLENG: Taller de Lenguaje (Literacidad), y TMATE: Taller de Matemáticas,

Para ejemplificar los resultados, las materias objetivo que aquí se exponen son: ED, LPL, y SO. Para cada una de ellas se generó una colección con atributos presumiblemente influyentes y otros presumiblemente no influyentes. A continuación se muestran los atributos usados para cada una de las tres materias mencionadas:

ED: MD1, PWE, HYC, TLENG, TMATE.
LPL: ED, MD2, PE, MD1, HYC, TLENG, SUST, TMATE.
SO: ED, ARQ, PE, MD1, HYC, TLENG, SUST, TMATE.

Debemos agregar que ninguna de las materias que aparece en la lista de cada materia objetivo constituye uno de sus requisitos en el plan de estudios. Así, la colección para ED, por ejemplo, se formó con registros de evaluaciones de alumnos que por primera vez fueron evaluados en ED, hayan o no aprobado (clase A o R) y, en cada instancia, aparecerá MD1, solo si aprobó dicha materia; de igual forma PWE, y las demás de su lista. La instancia 'IDk PWE TMATE HYC A' de la colección ED ejemplifica que el alumno con identificador IDk aprobó ED y previamente aprobó PWE, TMATE, y HYC, pero no aprobó TLENG ni MD1. La tabla 1 expone los tamaños de las clases de las colecciones de ejemplo.

Tabla 1. Número de instancias en cada clase para ED, LPL y SO.

Colección	**#Clase-A**	**#Clase-R**	**#Total**
ED	46	37	83
LPL	56	38	94
SO	66	43	109

De acuerdo con lo descrito se tiene un problema de clasificación binaria. Este problema fue tratado con árboles de decisión y la aplicación voraz de la separación de clases. Notemos que la clasificación se emplea como indicador de la bondad de los atributos, pues el propósito es identificar los atributos relevantes: materias previamente aprobadas que influyen en la aprobación de una materia objetivo. En la siguiente subsección se presentan los resultados obtenidos y se hace la interpretación de los mismos.

4.2. Resultados

Se procedió entonces a identificar los atributos que eran relevantes para dar respuesta al problema de requisitos para cada una de las materias objetivo. El resultado de los métodos de ganancia de información y separación de centroides se presentan en la tabla 2.

La tabla 2 presenta tres secciones: para las colecciones correspondientes a ED, LPL y SO. Cada colección tiene asociadas en tres columnas los atributos de la materia, la ganancia de cada uno de ellos (**GI**), y el valor de separación entre clases (**SC**). Se han marcado los valores destacados en negritas, indicando que son los mejores atributos de acuerdo con el criterio de ganancia o separación de clases. Observamos que, en la mayoría de casos coinciden los atributos con valores más altos en GI con los que separan las clases A y R, excepto en LPL y SO. En LPL, el atributo MD1 tiene un valor GI más bajo que TLENG y SUST, pero separa clases; en tanto TLENG y SUST, de mayor GI que MD1, no separan clases. En SO, el atributo HYC tiene el valor máximo de GI pero

Tabla 2. Ganancia de información y valor de separación de centroides para atributos de ED, LPL y SO.

Atr-ED	**GI**	**SC**	**Atr-LPL**	**GI**	**SC**	**Atr-SO**	**GI**	**SC**
TLENG	0.11	-0.0002	TLENG	0.08	-0.0094	TLENG	0.05	-0.0002
TMATE	**0.42**	**0.0041**	SUST	0.10	-0.0094	SUST	0.00	-0.0026
PWE	0.12	-0.0013	TMATE	0.03	-0.0083	TMATE	0.04	-0.0035
HYC	0.05	-0.0011	HYC	0.00	-0.0145	HYC	**0.39**	-0.0028
MD1	**0.27**	**0.0002**	PE	0.01	-0.0110	PE	0.00	-0.0034
			ED	**0.30**	**0.0243**	MD1	0.01	-0.0017
			MD1	0.04	**0.0014**	ED	**0.21**	**0.0055**
			MD2	**0.41**	**0.0258**	ARQ	**0.28**	**0.0095**

no separa clases (es negativo). Ambos resultados llaman la atención, pues el resultado obtenido por separación de centroides coincide con los previamente determinados por especialistas de la disciplina.

Con el fin de comprobar la efectividad de los atributos obtenidos por el cálculo de ganancia de información, se utilizaron los atributos en el clasificador de árbol de decisión. Una vez obtenida la ganancia de cada uno de los atributos se probaron diferentes umbrales de mínima ganancia para elegir atributos y realizar la clasificación con este método. Los resultados se exponen en la tabla 3.

Tabla 3. Exactitud de la clasificación con árboles de decisión, variando el umbral de GI para las materias ED, LPL y SO.

Atr-ED	**GI**	**ACC**	**Atr-LPL**	**GI**	**ACC**	**Atr-SO**	**GI**	**ACC**
HYC	0,05	0.59	HYC	0	0.60	PE	0	0.55
TLENG	0.11	0.61	PE	0.01	-	SUST	0	-
PWE	0.12	0.62	TMATE	0.03	0.62	MD1	0.009	-
MD1	0.27	0.62	MD1	0.04	-	TMATE	0.04	0.59
TMATE	0.42	0.41	TLENG	0.08	0.65	TLENG	0.05	0.54
			SUST	0.10	0.65	ED	0.21	0.56
			ED	0.30	0.66	ARQ	0.28	0.63
			MD2	0.41	0.55	HYC	0.39	0.65

Utilizando todos los atributos, es decir con GI$>$ 0, se tendrá un valor de exactitud menor que el que se obtiene al descartar atributos con menor ganancia. Este procedimiento aumentará la exactitud de la clasificación al tomar atributos con ganancia cada vez mayor. La tabla 3 presenta también tres columnas para cada colección: la primera los atributos ordenados descendentemente de menor a mayor según GI cuyo valor aparece en la segunda columna, y la exactitud de la clasificación eligiendo atributos con valor de ganancia mínima igual a la del atributo correspondiente. Se observa que el anterior procedimiento no es monótono; por ejemplo, en las columnas última y penúltima, correspondientes a los requisitos de la materia SO, puede verse que al tomar atributos de GI mayor

a 0.05 la exactitud desciende de 0.59 a 0.54 y, posteriormente, vuelve a ascender. Este comportamiento de la exactitud, tomando atributos con GI cada vez mayor, significa que no podríamos confiar en un umbral basado en GI. En tanto, para los atributos elegidos por el algoritmo de separación de centroides los resultados de exactitud son: 0.62, para la colección ED; 0.77, para la colección LPL; y 0.71, para la colección SO.

5. Conclusiones

En este trabajo se ha aplicado a una muestra de evaluaciones de estudiantes de una carrera universitaria técnicas de análisis de atributos con la finalidad de conocer la influencia de las materias previamente aprobadas en otra que es de interés. Se conformó para cada materia objetivo una colección de instancias con atributos referentes a las materias aprobadas previamente que pueden ser influyentes en la aprobación de la materia objetivo. Para cada atributo de cada colección referente, se calculó la separación entre sus centroides de la clase de aprobados y reprobados antes y después de eliminar dicho atributo y, así, conocer su importancia en la representación de las instancias. Asimismo, se utilizó el algoritmo de árboles de decisión que permitió contrastar la importancia de los atributos por ganancia de información con el valor de separación de clases que provee cada atributo. Una desventaja del algoritmo para determinar atributos discriminantes es su complejidad ($O(nm)$, con n instancias y m atributos); en este caso fue viable su aplicación por la cantidad de datos y atributos. En contraste con lo anterior se observaron varias ventajas:

- Vemos que en muchos problemas es necesario fijar umbrales para tomar decisiones. El uso de la medida de separación entre centroides no requiere utilizar umbrales. Además, los atributos así obtenidos ayudaron a mejorar la exactitud en la clasificación por árboles de decisión.
- También, con el conjunto de datos utilizado en los experimentos, el comportamiento de separación de centroides es estable: a mayor valor de separación de clases, mejor el atributo o con más efectividad para clasificar.
- Los resultados obtenidos, materias identificadas como influyentes en la materia objetivo, coinciden con las determinadas previamente en forma cualitativa por especialistas en el campo de la computación.

A partir de estos resultados es necesario aplicar los métodos a otras colecciones de datos para conocer su desempeño con más amplitud y precisión, asimismo comparar con otros enfoques.

Referencias

1. Al-Barrak, M.A., Al-Razgan, M.: Predicting students final GPA using decision trees: a case study. International Journal of Information and Education Technology 6(7), 528 (2016)

2. Basuki, A.: Personalized learning path of a web-based learning system. International Journal of Computer Applications 53(7) (2012)
3. Blum, A., Rivest, R.L.: Training a 3-node neural network is NP-complete. Advances in neural information processing systems, pp. 494–501 (1989)
4. Chen, C.M.: Intelligent web-based learning system with personalized learning path guidance. Computers & Education 51(2), 787–814 (2008)
5. Domingos, P.: A Few Useful Things to Know about Machine Learning. Communications of the ACM 55(10), 78–87 (2012)
6. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annals of Eugenics 7(2), 179–188 (1936)
7. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), pp. 1157–1182 (2003)
8. Hall, M.A., Smith, L.A.: Practical feature subset selection for machine learning. (1998)
9. Idris, N., Yusof, N., Saad, P.: Adaptive course sequencing for personalization of learning path using neural network. Int. J. Advance. Soft Comput. Appl 1(1), 49–61 (2009)
10. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial intelligence 97(1-2),273–324 (1997)
11. Kovacic, Z.: Early prediction of student success: Mining students' enrolment data. (2010)
12. Liu, H., Setiono, R.: Chi2: Feature selection and discretization of numeric attributes. In: Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence, pp. 388–391 (1995)
13. Novakovic, J.: Toward optimal feature selection using ranking methods and classification algorithms. Yugoslav Journal of Operations Research 21(1) (2016)
14. Quinlan, J.R.: C4. 5: programs for machine learning. Elsevier (2014)
15. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM 18(11), 613–620 (1975)
16. Sánchez-Sánchez, C., Jiménez-Salazar, H.: An effect of term selection and expansion for classifying short documents. Journal of Research in Computing Science, vol. 123, pp. 99–109 (2016)