Estrategia basada en el aprendizaje de máquina para tratar con conjuntos de datos no etiquetados usando conjuntos aproximados y/o ganancia de información

Strategy based on machine learning to deal with untagged data sets using rough sets and/or information gain

Luis-Alexánder Calvo-Valverde[1]

Fecha de recepción: 19 de junio del 2015
Fecha de aprobación: 26 de setiembre del 2015

Calvo-Valverde, L. Estrategia basada en el aprendizaje de máquina para tratar con conjuntos de datos no etiquetados usando conjuntos aproximados y/o ganancia de información. *Tecnología en Marcha*. Edición especial. Matemática Aplicada, Mayo 2016. Pág 4-15.

1 Doctorado en Ciencias Naturales para el Desarrollo (DOCINADE), Instituto Tecnológico de Costa Rica, Centro de Investigaciones en Computación, Programa Multidisciplinar eScience. Correo electrónico: lcalvo@itcr.ac.cr

Palabras clave

Aprendizaje de máquina; minería de datos; conjuntos aproximados; entropía; ganancia de información; reducción de atributos.

Resumen

Hoy en día se recogen datos de muy diversa índole y a un bajo costo, como no se había visto antes en la historia de la humanidad; por ejemplo, sensores que registran datos a cada minuto, páginas *web* que almacenan todas las acciones que realiza el usuario, supermercados que guardan todo lo que sus clientes compran y en qué momento lo hacen. Pero estas grandes bases de datos presentan un gran reto a sus propietarios ¿Cómo sacarles provecho?, ¿cómo convertir datos en información para la toma de decisiones?

Este artículo presenta una estrategia basada en el aprendizaje de máquina para tratar con conjuntos de datos no etiquetados utilizando conjuntos aproximados y/o ganancia de información. Se propone una estrategia para agrupar los datos utilizando *k-means*, considerando cuánta información aporta un atributo (ganancia de información), además de poder seleccionar cuáles atributos son realmente indispensables para clasificar nuevos datos y cuáles son dispensables (conjuntos aproximados), lo cual es muy beneficioso pues permite tomar decisiones en menor tiempo.

Keywords

Machine Learning; Data Mining; Rough Sets; Entropy; Information Gain; Feature Reduction.

Abstract

As had been seen in the history of humanity, today data of various kinds and cheaply collected, for example sensors that record information every minute, web pages that store all the actions performed by the user on the page supermarkets that keep everything their customers buy and when to do it and many more examples like these. But these large databases have presented a challenge to their owners How to take advantage of them? How to turn data into information for decision making? This paper presents a strategy based on machine learning to deal with unlabeled datasets using rough sets and/or information gain. A method is proposed to cluster the data using k-means considering how much information provides an attribute (information gain); besides being able to select which attributes are really essential to classify new data and which are dispensable (rough sets), which is very beneficial as it allows decisions in less time.

Introducción

La aparición de grandes bases de datos (*Big Data*) en las organizaciones ha producido un auge en la búsqueda de medios que permitan sacarle provecho a estos grandes repositorios.

Es en este contexto en el que la aplicación del aprendizaje de máquina como medio para realizar minería de datos se ha utilizado con un éxito relativo. En particular, dos aspectos han llamado la atención de muchos autores, y tienen que ver con el objetivo de realizar el aprendizaje de una manera más eficiente:

- ¿Cómo etiquetar datos históricos para luego poder clasificar datos nuevos?
- ¿Cómo reducir la dimensionalidad de estas grandes bases de datos, por ejemplo, eliminando atributos redundantes o innecesarios, con el fin de hacer más eficiente el proceso?

Al respecto, la aplicación de los conjuntos aproximados ha tenido un resultado interesante al manejar conceptos de aproximación y reducción de atributos redundantes; esto se puede apreciar en trabajos como los de Bello y Verdegay (2010), César, Caicedo y Pérez (2010), Hedar, Wang y Fukushima (2008), Mahajan, Kandawal y Vijay (2012), Zdzislaw y Ziarko (1995), Rissino y Torres (2009), Thangavel, Shen y Pethalakshmi (2006), Velayutham y Thangavel (2011) y Zhang, Li y Chen (2013).

Un aspecto que queda claro al analizar los trabajos anteriores es que el tema del presente artículo está abierto y resta mucho por investigar, sobre todo en la búsqueda de algoritmos eficientes y no solo eficaces. En este trabajo se presentan algunas consideraciones que contribuyen a esta línea de investigación.

En primer lugar, se hace una presentación introductoria de conceptos que permiten comprender la propuesta del autor, luego se presentan los materiales y métodos utilizados y finalmente los resultados y algunas conclusiones.

Terminología y conceptos

A. Aprendizaje de máquina *(Machine learning)*

El aprendizaje de máquina (*machine learning*, como se denomina en inglés) se refiere al estudio de algoritmos de computadora que mejoran automáticamente a través de la experiencia. Este tipo de aprendizaje se ha utilizado en aplicaciones que van desde la minería de datos que descubren las reglas en grandes conjuntos de datos hasta sistemas de filtración de información que automáticamente aprenden los intereses de los usuarios. De acuerdo con Murphy (2012), el aprendizaje de máquina es un conjunto de métodos que automáticamente pueden detectar patrones en los datos y usar los patrones descubiertos para predecir datos futuros o para ejecutar otra clase de toma de decisión bajo incertidumbre, como, por ejemplo, planificar cómo recolectar más datos.

En cuanto a la clasificación de las técnicas de aprendizaje de máquina, el mismo Murphy las divide en:

- Aprendizaje supervisado: también llamado predictivo, cuyo objetivo es aprender a mapear desde X entradas a Y salidas, dado un conjunto etiquetado de N pares de entrada-salida; este conjunto se denomina *Training se*t.
- Aprendizaje no supervisado: también llamado descriptivo, cuyo objetivo es encontrar patrones interesantes en las N entradas.
- Reforzamiento del aprendizaje: se usa para conocer cómo actúa o se comporta cuando se dan ciertas señales ocasionales de premio o castigo.

En algunas publicaciones, el aprendizaje de máquina se identifica con el reconocimiento de patrones; para algunos autores, este último tiene sus orígenes en la ingeniería, mientras que el aprendizaje de máquina creció en las ciencias de la computación. Sin embargo, ambas actividades pueden verse como dos facetas del mismo campo y han experimentado un desarrollo importante en los últimos años (Bishop, 2006).

Por otra parte, ante la pregunta de ¿en qué se diferencia el aprendizaje de máquina de la minería de datos?, se puede decir que la segunda pone más énfasis en modelos interpretables, mientras que el primero le da mayor relevancia a que los modelos sean precisos (Murphy, 2012).

B. Conjuntos aproximados (*Rough Sets*)

En 1982, Pawlak introdujo la teoría de los conjuntos aproximados, siguiendo a Thangavel, Shen y Pethalakshmi (2006). En esta sección se presentan los conceptos fundamentales al respecto.

Conceptos generales

Siguiendo a Rissino y Torres (2009), se puede decir que esta teoría fue desarrollada inicialmente para un universo finito en el cual la base de conocimiento es una partición, que se obtiene por una relación de equivalencia definida en ese universo. En la teoría de los conjuntos aproximados, el dato se organiza en una tabla llamada tabla de decisión. Las filas de la tabla de decisión corresponden a los objetos y las columnas a los atributos. En el conjunto de datos, una etiqueta de clase indica la clase a la cual pertenece cada fila. La etiqueta de clase se denomina un atributo de decisión (D), el resto de los atributos son los de condición (C), donde $C \cup D = \varnothing$, y t_j indica la j^{esima} tupla de la tabla de datos. La teoría de los conjuntos aproximados define tres regiones basadas en las clases de equivalencia inducidas por los valores de atributo: aproximación por abajo, aproximación por arriba y frontera.

La aproximación por abajo contiene todos los objetos clasificados con seguridad basados en los datos recolectados, la aproximación por arriba contiene los objetos que pueden ser clasificados probablemente y la frontera es la diferencia entre la aproximación por arriba y la aproximación por abajo.

Sea U un universo finito. Sea R una relación de equivalencia definida en U, la cual particiona a U. (U, R) es una colección de todas las clases de equivalencia, llamada espacio de aproximación. Sean $w_1, w_2, w_3, \ldots w_n$ elementos del espacio de aproximación (U, R). Esta conexión se conoce como base de conocimiento. Entonces, para cualquier subconjunto B de U, la aproximación por arriba $\overline{B}$ y la aproximación por abajo $\underline{B}$ se definen como:

$$\overline{B} = \cup \left\{ w_i / w_i \subseteq B \right\}$$

$$\overline{B} = \cup \left\{ w_i / w_i \subseteq B \neq \varnothing \right\}$$

Al par ordenado ($\underline{B}$,$\overline{B}$) se le denomina un conjunto aproximado Además se tiene (Rissino & Torres, 2009):

POS(B) = $\underline{B} \Rightarrow$ ciertamente miembro de X.

NEG(B) = U – $\overline{B} \Rightarrow$ ciertamente no miembro de X.

BR(B) = $\overline{B}$ – $\underline{B} \Rightarrow$ posiblemente miembro de X.

En la figura 1 se aprecia cómo es esta distribución

Donde POS(B) refiere a la región positiva de B, NEG(B) refiere a la región negativa de B y BR(B) refiere a la región frontera de B.

Relación de indiscernibilidad

En una tabla de decisión existen atributos de condición y atributos de decisión. Se llama concepto a un conjunto de atributos de decisión para los cuales todos los objetos tienen el mismo valor de decisión.

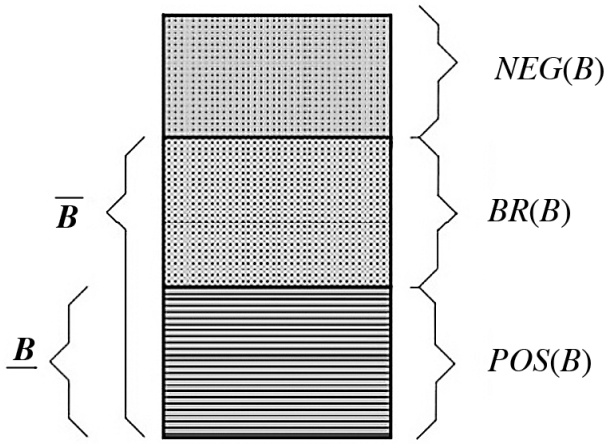

Figura 1. Representación gráfica de las regiones en un conjunto aproximado, basado en Rissino y Torres (2009).

Conjuntos indiscernibles o elementales son aquellos que no se diferencian entre sí por sus atributos. Por tanto, una relación de indiscernibilidad es realmente una relación de equivalencia. La unión de conjuntos elementales, o indiscernibles, es un conjunto definible.

Una tabla de decisión se llama inconsistente o conflictiva si para al menos dos objetos los atributos de condición son iguales pero con diferente atributo de decisión.

Si un conjunto de atributos y su superconjunto definen la misma relación de indiscernibilidad, entonces cualquier atributo que pertenece al superconjunto y no al conjunto es redundante.

El conjunto de atributos que no tienen atributos redundantes es llamado mínimo (o independiente).

P es una reducción de *Q*, si *P* es mínimo y las relaciones de indiscernibilidad definidas por *P* y *Q* son la misma.

Representado de manera más formal:

Sea *IS* una tabla de decisión donde:

$$IS = \{U, AT\}$$

donde U representa a todos los registros y AT son todos los atributos.

Sea *a* un atributo $a \in AT$.

Cada atributo tiene un dominio de valores V_a que puede tomar el atributo *a*.

$$a : U \rightarrow Va$$

A cada objeto *x* en el universo *U* se le asigna un valor $a\ (x)$ desde V_a a cada atributo a, y a cada objeto x en el universo *U*. Si V_a contiene valores perdidos para al menos uno de los atributos a, entonces IS es llamada una tabla de información incompleta, de lo contrario es completa.

Para cualquier subconjunto de atributos $P \subset AT$, hay una relación de equivalencia (indiscernibilidad). *IND(P)* es llamada relación de indiscernibilidad de P.

$$IND(P) = \{(x, y) \in U^2 \,/\, \forall a \in P, a(x) = a(y)\}$$
$$IND(P) = \cap IND(a) \ donde \ a \in P$$

Con lo anterior, sea P un conjunto de atributos, $a \in P$, el atributo *a* es dispensable en P si:

$$IND(P) = IND(p - \{a\})$$

De lo contrario *a* es un atributo dispensable.

El conjunto de atributos A, donde $A \subset P$ es llamado reducción de P si:

$$IND(A) = IND(P)$$

Y puede tener muchas reducciones, denotado como *RED(P)*. El conjunto de todos los atributos indispensables en P es llamado el core de P y se denota como *CORE(P)*, donde:

$$CORE(P) = \cap RED(P)$$

Si $IND(P - \{a\}) = \mathrm{IND(P)}, a$ es un atributo dispensable y el conjunto de atributos constituye una reducción de P.

La intersección de todas las reducciones de P produce el core que son los atributos más importantes para clasificar correctamente. B3. Medidas de incertidumbre

Dado un conjunto de ejemplo X, no necesariamente definido por un conjunto P de atributos, una manera de estimar la calidad de la aproximación de X es la siguiente (Mahajan, Kandawal & Vijay, 2012).

$$Calidad\ Aproximada\ por\ abajo\ = \frac{\#\text{total de elementos en la Aproximaxión por abajo de X}}{\#total\ de\ ejemplos\ de\ X}$$

$$Calidad\ Aproximada\ por\ arriba\ = \frac{\#\text{total de elementos en la Aproximaxión por arriba de X}}{\#total\ de\ ejemplos\ de\ X}$$

C. Reducción de atributos

La reducción de atributos tiene dos objetivos fundamentales (Mahajan, Kandawal & Vijay, 2012):

- Disminuir el número de atributos de condición.
- Maximizar la información contenida en los atributos seleccionados.

El logro de estos dos objetivos se ve reflejado en la mejora del tiempo de respuesta de los algoritmos de aprendizaje máquina, al tener que realizar menos comparaciones y cálculos y eliminar variables que pueden generar ruido a tal punto que produzcan generalizaciones bajo supuestos incorrectos. Todo por considerar variables que aportan poca información para la toma de decisiones o no son representativas del grupo de datos en estudio.

D. Criterios de comparación de algoritmos en *k-means*

Existen varios criterios para comparar los resultados de diferentes versiones de los algoritmos *k-means*, pero hay dos en particular que son muy claros para su interpretación:

- Minimizar la suma de distancias: Se trata de sumar las distancias de todos los datos con respecto a los centros de los grupos en que quedaron clasificados y seleccionar la de menor valor.

- Maximizar el número de casos de éxito: Se trata de seleccionar la solución que tiene el mayor número de casos de éxito al relacionar la predicción con los valores reales.

E. Entropía y ganancia de información

Dadas dos clases *P* y *N* en un espacio muestral *S*, donde:

$$S = P \cup N$$

Las cardinalidades están dadas por:

$$|P| = p$$

$$y$$

$$|N| = n$$

El término *entropía* se refiere a la cantidad de información necesaria para decidir si una muestra de S pertenece a P ó a N. y se define como (González, 2013):

$$E(S) = \frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right)$$

Partiendo de que al seleccionar un atributo *b* el espacio muestral es dividido en subconjuntos hijos de *b*, el modo de determinar cuánta información aporta un atributo *b* en un conjunto total de atributos *A* está dado por:

$$Aporte(b) = \mathrm{E}(\mathrm{A}) - \sum(\forall los\ subconjuntos\ hijos\ de\ b)$$

Finalmente, si se tienen k clases, N instancias en el conjunto de datos, la entropía de todo el conjunto es E, la entropía de cada uno de los subconjuntos es E1 y E2, la cantidad de instancias en una clase es $k1$ y en la otra $k2$, entonces el mínimo aporte de información sería (Witten, Frank & Hall, 2001):

$$\frac{\log_2 N - 1}{N} + \frac{\log_2 3^k - 2 - k^E + k1 * E1 + k2 * E2}{N}$$

F. Algoritmo *k-means*

K-means es un algoritmo para realizar aprendizaje no supervisado (Thangavel, Shen y Pethalakshmi, 2006). Su idea general es:

- Se particiona el conjunto de datos en K grupos (*clusters*) de modo aleatorio.
- Se seleccionan aleatoriamente K puntos centrales, uno de cada grupo (centroides).
- Para cada dato se calcula la distancia del punto a cada punto central de los grupos y el dato pasa a formar parte del grupo cuya distancia es menor a su centro.
- Si el dato es más cercano a su propio grupo, se queda en su grupo, de lo contrario pasa a formar parte del grupo del centro más cercano.
- Se repite el proceso anterior hasta que ningún punto se pase de grupo.

La selección inicial de los centros puede afectar significativamente los resultados.

Materiales y métodos

Para realizar las pruebas se utilizaron conjuntos de datos provenientes de la Universidad de California (UCI) (Bache & Lichman, 2013). En el cuadro 1 se muestran más detalles de los mismos:

Cuadro 1. Conjuntos de datos utilizados de UCI.

Conjuntos de datos	Tipos de datos	Número de instancias	Número de atributos
Car Evaluation	Multivariados	1728	6
Credit Approval	Multivariados	690	15
Chess	Multivariados	3196	36
Skin Segmentation	Univariados	245057	4

Se tomaron conjuntos de datos con atributos de decisión para poder correr los algoritmos excluyendo este atributo y luego comparar los resultados obtenidos con lo que originalmente indicaban dichos atributos. Además, el tener conjuntos de datos para los cuales se conoce el atributo de decisión, permite determinar el número de grupos del algoritmo k-*means*.

Como en la mayoría de los procesos de minería de datos, cada repositorio tuvo que pasar por las etapas de: limpiar, integrar, seleccionar, transformar, minar, interpretar y presentar.

Se utilizó el IDLE de Python 2.7.2 para programar los algoritmos y se corrió en la siguiente plataforma:

- Fabricante del sistema; Hewlett-Packard
- Modelo del sistema: HP ProBook 6450b
- Tipo de sistema: PC basado en x64
- Procesador Intel(R) Core(TM) i5 CPU M 460 @
- 2.53GHz, 2534 Mhz, 2 procesadores principales, 4 procesadores lógicos
- Sistema operativo: Microsoft Windows 7 Professional.

En cuanto al criterio de comparación de los algoritmos de k-*means*, se escogió el de maximizar el número de casos de éxito, por cuanto al final el interés último es determinar qué tan bien hizo el agrupamiento.

Con el fin de comparar los resultados obtenidos, se corrieron tres procesos con los mismos conjuntos de datos. Se hicieron 100 repeticiones cuando se realizaban procesos aleatorios, con el fin de determinar el efecto medio del algoritmo. Cuando se utilizaron rangos, no tenía sentido repetirlo más de una vez pues el algoritmo es determinístico para un conjunto de datos dado. Solo en el caso de k-*means* clásico para *Skin Segmentation* se hicieron 50 corridas en lugar de 100, por motivo del tiempo computacional requerido para completar las 100, pues en realidad con 50 ya se podía estimar una media confiable del proceso aleatorio.

k-means clásico con centroides aleatorios

Se utilizó *k-means* como algoritmo de agrupamiento de modo que los grupos resultantes sirvieran luego para etiquetar los objetos en su atributo de decisión (D); utilizando el número de grupo en el que el objeto quedó agrupado como valor del atributo de decisión (D).

k-means utilizando solo los atributos con un aporte de información superior a una frontera

Se calculó la entropía de cada uno de los atributos y su ganancia de información. El método utilizado fue el siguiente:

- Sea $E(C)$ la entropía de todo el conjunto de atributos.
- Se calcula cuanta información aporta la entropía de cada uno de los c atributos de condición (C), considerando:
- Sea $E(c_i)$ la entropía del atributo de condición c_i.
- Como la selección del criterio de en qué valor, de los V_c valores, de dividir el atributo c para calcular la entropía puede ser muy diferente para cada atributo, se decide ordenar los V_c valores de menor a mayor y tomar la media como criterio de división.
- El aporte de información del atributo c es igual a: $(C) - \sum E(\forall los\ subconjuntos\ hijo\ de\ c)$
- Se utilizan los atributos de condición que aportan la mayor cantidad de información como los seleccionados para elegir de ellos los centros iniciales para el algoritmo *k-means*.

Una vez elegidos los atributos a considerar, si se desea que el atributo de decisión (D) tome V_d valores diferentes, entonces se corre *k-means* para formar V_d grupos, utilizando para el cálculo de distancias solo los atributos seleccionados por su mayor aporte de información. Se pueden inicializar los centros aleatoriamente o bien dividir el rango total de los valores del atributo c en k trozos uniformes y tomar estos valores como centros iniciales del algoritmo *k-means*. Al respecto, dado que son los atributos que más información aportan, se decidió inicializar los centros con rangos uniformes.

k-means utilizando solo los atributos seleccionados por conjuntos aproximados

También se podría utilizar la teoría de los conjuntos aproximados, para determinar qué atributos de condición son indispensables y cuales dispensables y, por lo tanto, proceder a la reducción de atributos, calculando la relación de indiscernibilidad de cada uno de ellos. Recuérdese que siendo P el conjunto de atributos, $a \in P$, el atributo a es dispensable en P si:

$$IND(P) = IND(p - \{a\})$$

De manera similar, una vez elegidos los atributos a considerar, si se desea que el atributo de decisión (D) tome V_d valores diferentes, entonces se corre *k-means* para formar V_d grupos utilizando para el cálculo de distancias solo los atributos indispensables. Parecido al caso anterior, los centros se pueden inicializar aleatoriamente o se puede dividir el rango total en k trozos uniformes; con el fin de comparar los resultados se inicializaron los centros con rangos uniformes.

Resultados

Luego de ejecutar los experimentos indicados, en el cuadro 2 se presentan los resultados obtenidos.

Cuadro 2. Resultados obtenidos para los diferentes conjuntos de datos.

Característica / Conjuntos de datos	Credit approval	Car Evaluation	Chess	Skin Segmentation
Total de registros	690	1728	3196	245057
Total de atributos incluyendo el de decisión	16	7	37	4
k-means clásico				
Media de la tasa de éxito *k-means* clásico	60.34	33.52	52.19	51.45
Desviación estándar de la tasa de éxito *k-means* clásico	3.46	8.40	02.59	9.19
Coeficiente de variación de la tasa de éxito *k-means* clásico	0.06	0.35	0.05	0.18
k-means usando ganancia de información				
Cantidad de atributos eliminados por ganancia de información	3	3	19	0
Media de la tasa de éxito usando los atributos restantes	56.81	28.41	59.73	46.67
Desviación estándar de la tasa de éxito usando solo los atributos restantes	0.00	0	0	0.00
Coeficiente de variación de la tasa de éxito usando solo los atributos restantes	0.00	0.00	0.00	0.00
Conjuntos aproximados				
Cantidad de atributos eliminados por conjuntos aproximados	12	0	5	0
Media de la tasa de éxito usando los atributos restantes	59.86	39.35	55.48	50.78
Desviación estándar de la tasa de éxito usando solo los atributos restantes	0.00	0.00	0.00	0.00
Coeficiente de variación de la tasa de éxito usando solo los atributos restantes	0.00	0.00	0.00	0.00

Del cuadro 2 se pueden resaltar varios aspectos:

1. Como es bien sabido, el algoritmo *k-means* clásico es muy dependiente de la selección de los centros iniciales. La inicialización de centros aleatorios tiende a tener desviaciones estándar altas y, por consiguiente, coeficientes de variación también altos.
2. Si se utiliza entropía y ganancia de información y solo se usan los atributos que aportan más información que el conjunto, y se usan rangos uniformes para los centroides en lugar de centros aleatorios, el proceso se vuelve determinístico para el mismo conjunto de datos; por eso la desviación estándar y el coeficiente de variación se muestran en cero.
3. Una vez etiquetados los datos, o si se dispone de conjuntos de datos ya etiquetados, y aunque la determinación de los atributos indispensables y los dispensables utilizando conjuntos aproximados es un proceso caro en tiempo computacional, una vez determinados, la reducción de atributos beneficia igualmente el proceso de clasificación a futuro. De hecho, se corrió *k-means* solo con los atributos indispensables y utilizando

rangos uniformes para los centroides, lo que mostró un rendimiento aceptable para todos los conjuntos de datos.

4. Si bien k-*means* clásico con centros aleatorios mostró que en algunos casos obtenía mayor tasa de éxito que los otros, el problema es que su desviación estándar es alta y, por tanto, como el caso promedio no siempre se dará, perfectamente se puede dar el peor de los casos, o casos cercanos a este, y en estos escenarios su rendimiento es menor que cuando se usa ganancia de información o conjuntos aproximados.

En el fondo hay un tema de conveniencia en la selección del método, ¿se prefiere una probable mayor tasa de éxito (k-*means* clásico con valores aleatorios) o un valor determinístico que no siempre será mejor que el aleatorio, pero sin el riesgo de que la aleatoriedad lleve al peor de los casos (media menos desviación estándar) o a casos cercanos a este?

En el cuadro 3 se muestran los cálculos en que se comparan el mejor, el peor y el caso promedio para los conjuntos de datos en estudio y con las diferentes opciones de cálculo.

Cuadro 3. Comparación de las tasas de éxito para el peor, el mejor y el caso promedio.

Tipo / Conjuntos de datos	Credit approval	Car Evaluation	Chess	Skin Segmentation
k-means clásico – aleatorio – mejor caso	63.80	41.92	54.78	60.64
k-means clásico – aleatorio – peor caso	56.88	25.12	49.60	42.26
k-means clásico – aleatorio – caso promedio	60.34	33.52	52.19	51.45
k-means usando ganancia de información – rangos – determinístico	56.81	28.41	59.73	46.67
k-means usando conjuntos aproximados – rangos – determinístico	59.86	39.35	55.48	50.78

Como se aprecia en el cuadro 3, "k-*means* clásico – aleatorio – mejor caso" supera a los otros esquemas en su tasa de éxito para *Credit Approval* (alcanza un 63,80%), para *Car Evaluation* (obtiene un 41,92%) y para *Skin Segmentation* (con un 60,61%); solo para el conjunto de datos *Chess, el valor obtenido* 54,78%, es inferior a cuando se usa ganancia de información (59,73%) y cuando se utilizan conjuntos aproximados (55,48%).

Conclusiones

La estrategia de aprendizaje de máquina propuesta para tratar con los conjuntos de datos no etiquetados se puede resumir de la siguiente manera:

Utilizar entropía y ganancia de información para seleccionar de cuáles atributos calcular los centros de k-*means*. La idea es utilizar solo los atributos que aportan mayor información.

Utilizar k-*means* con solo los atributos seleccionados del paso anterior para etiquetar los datos en su atributo de decisión.

Una vez etiquetados los objetos con los pasos anteriores, se pueden usar conjuntos aproximados para determinar qué atributos son dispensables y cuales indispensables y, por tanto, proceder a la reducción de atributos.

Una de las tareas futuras es la experimentación con otras fuentes de datos y determinar si otras técnicas adicionales a la de ganancia de información y a la de conjuntos aproximados pueden mejorar los resultados presentados en este trabajo.

Un aspecto que se podría argumentar es que el cálculo de la entropía, la ganancia de información y los conjuntos aproximados requiere un esfuerzo computacional previo al cálculo del k-*means*, que el método clásico (sin reducir atributos) se ahorra. Pero ante este argumento hay dos aspectos muy importantes, el primero es que si se cuenta con muchos atributos, la reducción de atributos recuperará el tiempo invertido al calcular el k-means con mucho menos atributos; y el segundo es que se sabe que la reducción de atributos colabora en la reducción del sobreajuste (en inglés *overfitting*) en caso de presentarse.

Agradecimientos

El autor desea agradecer al Dr. Carlos González Alvarado, profesor del curso Minería de Datos impartido en el Doctorado en Ciencias Naturales para el Desarrollo (DOCINADE), por sus orientaciones en la preparación del presente documento. Al Dr. Pablo Alvarado Moya, tutor de la tesis del doctorando, y al Instituto Tecnológico de Costa Rica, por su financiamiento para realizar los estudios doctorales en el DOCINADE.

Bibliografía

Bache, K., & Lichman, M. (2013). *School of Information and Computer Science.* UCI Machine Learning Repository,University of California. Obtenido de http://archive.ics.uci.edu/ml

Bello, R., & Verdegay L.J. (2010). Los conjuntos apróximados en el contexto de softcomputing Rough sets in the Soft Computing context. *4(1-2 ENERO- JUNIO)*, 5-24. España: Revista del Departamento de Ciencias de la Computación, Universidad Central de Las Villas.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning, New York: Springer Science+Business Media. *LLC*.

César, J., Caicedo, C., & Pérez, N. (2010). Servicio web inteligente para la clasificación de imágenes digitales utilizando conjuntos aproximados. *Ingenieria e Investigación, 30*, 45-51.

González, C. (2013). *Material del Curso Data Mining.* San José: Doctorado en Ciencias Naturales para el Desarrollo.

Hedar, A., Wang, J., & Fukushima, M. (2008). Tabu Search for Attributes Reduction in Rough Set Theory. *Soft Computing, 12*(9), 909-918.

Mahajan, P., Kandawal, R., & Vijay, R. (2012). Rough Set Approach in Machine Learning: A review. *International Journal of Computer Applications (0975 – 8887), 56*(10), 1-12.

Murphy, K. P. (2012). *Machine Learning: A probabilistic perspective.* Massachusetts: MIT Press.

Pawlak, Z., Grzymala-Busse, J., & Slowinski, R. (1995). Rough Sets. *Communications of the ACM*, 88-95.

Rissino, S., & Lambert-Torres, G. (2009). Rough Set Theory Fundamental Concepts , Principals , Data Extraction , and Applications. En *Data Mining and Knowledge Discovery in Real Life Applications* (pág. 438). Vienna: Julio Ponce and Adem Karahoca.

Thangavel, k., Shen, Q., & Pethalakshmi, A. (2006). Application of Clustering for Feature Selection Based on Rough Set Theory Approach. *AIML Journal*, 19-27.

Velayutham, C., & Thangavel, K. (2011). Unsupervised Quick Reduct Algorithm Using Rough Set Theory. *Journal of electronic science and technology, VOL. 9, NO. 3*, 193-201.

Witten, I., Frank, E., & Hall, M. (2001). *Data Mining.* USA: ELSEVIER.

Zhang, J., Li, T., & Chen, H. (2013). Composite rough sets for dynamic data mining. *Information Sciences. Vol 257*, 81-100.