Estimación de rendimiento académico a través de técnicas para minería de datos

Ricardo A. ALAYON
Facultad ingeniería, Universidad Distrital Francisco José de Caldas
Bogotá, Bogotá 11021-110231588, Colombia.

Diego A. MONCADA
Facultad ingeniería, Universidad Distrital Francisco José de Caldas
Bogotá, Bogotá 11021-110231588, Colombia.

Victor H. MEDINA
Facultad Tecnológica, Universidad Distrital Francisco José de Caldas
Bogotá, Bogotá 11021-110231588, Colombia.

Jorge E. RODRÍGUEZ
Facultad Tecnológica, Universidad Distrital Francisco José de Caldas
Bogotá, Bogotá 11021-110231588, Colombia.

RESUMEN

Los sistemas de información y gestión educativa manejan enormes cantidades de valiosa información, la cual contiene conocimiento muy útil. Las técnicas para descubrir ese conocimiento son conocidas como minería de datos educativa. El objetivo de la minería educativa de datos (MED), es mejorar el desempeño de los estudiantes, profesores e instituciones educativas. Se sabe de trabajos que han analizado los sistemas académicos de información para mejorar las planeaciones de parte de los profesores y el desempeño de los estudiantes. La intención de este documento es obtener conocimiento desde el punto de vista de la estadística y de la minería de datos. Se analiza una muestra de 10000 registros generados aleatoriamente, con el fin de generar un modelo predictivo del rendimiento académico usando modelos de clasificación. Estos análisis pueden ayudar a mejorar el desempeño de los estudiantes prediciendo sus probabilidades de éxito o fracaso académico de manera temprana, a la vez que pueden evitar gastos a las instituciones al permitir planear adecuadamente la disposición de espacios, grupos y horarios.

Palabras Claves: Agrupación, clasificación, minería de datos educativa, aprendizaje computacional.

1. INTRODUCCIÓN

Es bien sabido que una de las mayores preocupaciones dentro del ámbito escolar es el fenómeno denominado fracaso académico. Este es un fenómeno ampliamente estudiado por diferentes ciencias entre las que pueden listarse la pedagogía, sicología, sociología e incluso la estadística. No es un tema indiferente para las ciencias de la computación, por tanto, la minería educativa de datos se ha ocupado de analizar patrones y obtener información de los sistemas académicos de información con miras a detectar regularidades y reglas que permitan inferir el comportamiento del sistema académico mediante el estudio de registros que de él se tienen. Los objetivos de investigación en minería educativa de datos pueden separarse en dos grandes grupos: la orientación administrativa y la orientación académica. La visión administrativa del problema se enfoca en el manejo de recursos y óptica institucional; la visión académica se centra en el desarrollo individual y la consecución de resultados mediante la previsión, corrección y seguimiento de procesos académicos individuales u orientados a grupos de estudiantes con características académicas similares.

El análisis de los datos almacenados de los estudiantes puede ayudar a pronosticar su rendimiento debido a la gran cantidad de datos disponibles, sin embargo, esta gran cantidad de datos no permite realizar esta tarea manualmente, por tanto, la tarea se puede lograr mediante el uso de Técnicas de minería para minimizar el problema de fracasos académicos.

.

En este artículo, el objetivo es construir un modelo de predicción del rendimiento estudiantil mediante la extracción de un conjunto de datos simulado. Para lograr esta tarea, se comparan diferentes tipos de técnicas de clasificación de DM (Naïve Bayes, Decision Tree y K-NN), y determina cuál es el más apropiado para la predicción temprana del rendimiento de los estudiantes. Los resultados del estudio podrían usarse para ayudar a docentes en la planificación para apoyar a los estudiantes con bajo rendimiento esperado.

2. SOLUCIÓN AL PROBLEMA

Los algoritmos de árbol de decisión se pueden utilizar en las notas de los estudiantes para predecir si aprueban en un examen específico. La comparación de los resultados pronosticados y los resultados reales indicó, que hubo una mejora significativa en los éstos y ayudó mucho para identificar a los estudiantes débiles y buenos, con intención de ayudarlos a mejorar. El algoritmo del árbol de decisión ID3 es el mejor en términos de eficiencia y tiempo para construir el árbol de decisión [2]. R. R. Kabra y R. S. Bichkar sugirieron que un Árbol de decisión puede usarse en estudiantes de ingeniería; datos pasados de rendimiento generan el modelo y este modelo se puede utilizar para predecir, además, permitirá identificar a los estudiantes que están en riesgo y advertirlos para generar mejoras en su rendimiento [3]. De acuerdo con Pandey y Sharma, diferentes algoritmos de árbol de decisión, como el J48, NBtree, Reptree y Simple CART se pueden usar para la predicción.

Agrupación (Cluster).
Descubrir conocimiento de un enorme volumen de datos es un reto en sí mismo. Para ello se utilizan herramientas automáticas que:

a) emplean algoritmos para descubrir principalmente patrones ocultos, asociaciones, anomalías, y/o estructuras de la gran cantidad de datos almacenados en los repositorios de información, y

b) filtran la información necesaria de las grandes bases de datos. El concepto de KDD (inciales de *Knowledge Discovery in Databases*) se ha desarrollado, y continúa desarrollándose. Los sistemas KDD incorporan teorías, algoritmos, y métodos de diversos campos. Una buena perspectiva general del KDD se puede encontrar en [5] y [6].

El algoritmo de Agrupación se refiere a la agrupación de registros, observaciones, o casos en clases de objetos similares. Para medir la similitud, se suelen utilizar diferentes formas de distancia: distancia euclídea, de Manhatan, de Mahalanobis, etc. El representar los datos por una serie de agrupaciones, conlleva la pérdida de detalles, pero consigue la simplificación de los mismos. Agrupación es una técnica más de Aprendizaje Automático, en la que el aprendizaje realizado es no supervisado.

K-Medias
Se trata de un algoritmo clasificado como Método de Agrupamiento. El método de las k-medias [7][8], es hasta ahora el más utilizado en aplicaciones científicas, e industriales. El nombre viene porque representa en k agrupaciones cada una por la media (o media ponderada) de sus puntos, es decir, su centroide. Este método únicamente se puede aplicar a atributos numéricos, y los *outliers* (o atípicos) le pueden afectar muy negativamente. Sin embargo, la representación mediante centroides tiene la ventaja de que tiene un significado gráfico y estadístico inmediato. La suma de los cuadrados de los errores se puede racionalizar, como el negativo del *log-likelihood*, para modelos mixtos que utilicen distribuciones normales. Por lo tanto, el método de las k-medias se puede derivar a partir del marco probabilístico [9]. La descripción matemática explica que dado un conjunto de observaciones $(x_1, x_2, \dots, x_n)$, donde cada observación es un vector real de d dimensiones, k-means construye una partición de las observaciones en k conjuntos $(k \leq n)$ a fin de minimizar la suma de los cuadrados dentro de cada grupo (WCSS): $S = \{S_1, S_2, \dots, S_k\}$

$$\frac{\arg\min}{s} \sum_{i=1}^{k} \sum_{x_j \in xs_i} \| x_j - \mu_i \|^2 \quad (1)$$

donde μ_i es la media de puntos en *Si*.

Existen dos versiones del método de las k-medias. La primera es parecida al algoritmo EM, y se basa en dos pasos iterativos; esta versión se conoce como algoritmo de Forgy [10]. La segunda versión [11] reasigna los puntos basándose en un análisis más detallado de los efectos causados sobre la función objetivo. La explicación matemática del algoritmo Forgy dice que dado un conjunto inicial de k centroides $m_1(1), \dots, m_k(1)$, el algoritmo continúa alternando entre dos pasos:

1. Paso de asignación: Asigna cada observación al grupo con la media más cercana

$$S_i^{(t)} = \left\{ x_p : \| x_p - m_i^{(t)} \| \leq \| x_p - m_j^{(t)} \| \ \forall\, 1 \leq j \leq k \right\} \quad (2)$$

2. Paso de actualización: Calcular los nuevos centroides como el centroide de las observaciones en el grupo.

$$m_i^{(t+1)} = \frac{1}{s_i^{(t)}} \sum_{x_j \in s_i^{(t)}} x_j \quad (3)$$

Algoritmo de Regresión
La regresión es una herramienta de aprendizaje automático que ayuda a hacer predicciones aprendiendo a partir de los datos estadísticos existentes como las relaciones entre su parámetro objetivo y un conjunto de otros parámetros [12]. La idea de la regresión es bastante simple: dados los datos suficientes, se puede observar la relación entre su parámetro de destino (la salida) y otros parámetros (la entrada), y luego aplicar esta función de relación a los datos reales observados [12].

3. MODELO DE DATOS

Este artículo utiliza una base de datos simulada con un paquete estadístico, son en total diez mil registros con tres a tributos a saber: un identificador y dos calificaciones. Por experiencia, se utilizó para la calificación 1, una distribución normal con media m=3 y desviación estándar 0,5 y para la calificación 2 una distribución normal con media m= 2.8 y desviación estándar de 0.5. Esta elección se soporta en que, dado que las mejores calificaciones suelen encontrarse en las primeras evaluaciones, su desviación es menor. Los datos usados entonces son:

Id. Un identificador de individuo, equiparable con el código de estudiante, en general un entero que permite encontrar el registro en base de datos.

N1. Nota 1, la primera calificación obtenida por el estudiante.

N2. Nota 2, la segunda calificación obtenida por el estudiante.

Np. Nota predicha, la calificación esperada en el examen, dado que se conocen sus notas preliminares de evaluaciones.

En el presente documento, por haber obtenido los datos mediante simulación se conocen los estadísticos de la muestra, sin embargo, para problemas prácticos el análisis exploratorio de datos es una necesidad. La muestra debe ser tan completa como sea posible, para intentar conocer a fondo la variación de cada dato individual. Un análisis de correlación entre las calificaciones mostrará la tendencia de las calificaciones de los individuos. A continuación, los algoritmos para la predicción.

4. ANALISIS Y PRUEBA DE RESULTADOS

Análisis del problema con Agrupación

Tener un conjunto de datos tan variable, sugiere la agrupación. Vale la pena pensar en agrupación, para relacionar los comportamientos de calificaciones de individuos y encontrar las tendencias de cada grupo. En el presente no se toman en cuenta otros factores de tipo socioeconómico, personal y étnico estudiados con detalle en documentos como [2] y [4].

Análisis Exploratorio de Datos

Como se dijo anteriormente en este caso no es necesario hacer exploración de datos dado que se conoce todo el conjunto y su distribución, por tanto, no se presenta un resumen de datos.

Simple k-means

Agrupar con simple k-means como se mencionó anteriormente, redistribuye al conjunto en 5 grupos de acumulación, los centroides de cada clase, donde cada centroide es la media ponderada de los 5 conjuntos más relevantes. Para el caso se encontraron los siguientes:

Tabla 1. Agrupación de datos k-means en 5 centroides.

Final cluster centroids:

		Cluster#				
Attribute	Full Data	0	1	2	3	4
	(10000.0)	(2739.0)	(1454.0)	(2072.0)	(1648.0)	(2087.0)
nota1	2.9967	3.3561	3.5777	2.8627	2.7044	2.4841
nota2	2.808	2.5081	3.4011	3.3893	2.0051	2.8452
media	2.9023	2.9321	3.4894	3.126	2.3547	2.6646

Sin tomar en cuenta el identificador que no presenta relevancia, lo interesante no ocurre al aplicar el k-means, sino al contrastar con los atributos originales del conjunto obteniendo clases agrupadas y distinguibles.

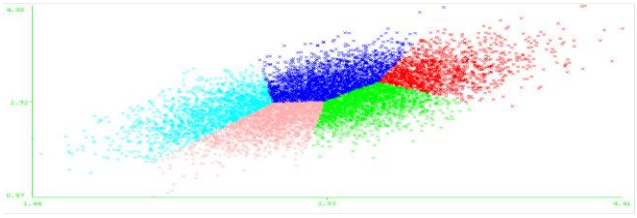

Figura 1. Nota 1, contra media asociada.

En la anterior gráfica se ven 5 conjuntos de datos completamente asociados, distinguibles contrastando la media asociada en el eje horizontal y el atributo nota 1 en el eje vertical.

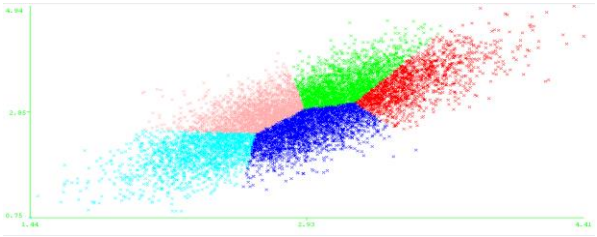

Figura 2. Nota 2 contra media asociada

La anterior gráfica muestra los clusters obtenidos por k means.
Al realizar la gráfica de clusters se identifican específicamente las tendencias centrales de cada uno.
Ahora, estos clusters pueden asociarse usando k means nuevamente para buscar patrones de asociación. Los datos están centrados en 5 clusters con marcas de clase que permiten discriminarlos. Las marcas de clase son los promedios mas frecuentes de las notas 1 y 2.

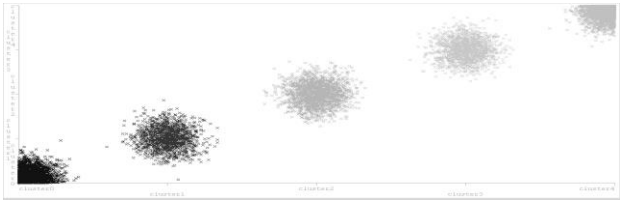

Figura 3. Diagrama de clusters.

Habiendo obtenido estos datos, y los clusters de notas que se mostraron arriba, puede entonces procederse a buscar la correlación entre las notas 1 y 2, de manera que pueda estimarse una correlación entre las notas conocidas y el promedio. Saber dicha correlación permitirá arriesgar un estadístico de la predicción de aprobación, basado en la media. Puede suponerse de esta manera que:

*0.7*media estimada + 0.3*nota faltante = nota final*

Lo que visto de otro modo, significa que se debe estimar la probabilidad de que, la nota faltante sea igual a la necesaria para aprobar en cada cluster, por supuesto, esto es forzar a que la nota definitiva sea mayor o igual que 3.0, es decir, se requiere hacer regresión sobre un nuevo atributo inducido que se llamó: *"nota faltante"*, para estimar la probabilidad de que un estudiante de cada cluster obtenga una definitiva mayor o igual a tres. Debe entenderse que el ejercicio de este artículo es académico, y la intención de este es encontrar la efectividad de predicción de las técnicas de minería para hacer este tipo de predicción.

Al discretizar los datos iniciales en 5 categorías

Desde la previsualización se evidencia que con los datos discretizados hay dos conjuntos de datos en cada nota que son los menos probables: el de valores cercanos a cero y el de valores cercanos a 5. Esto por la suposición de que los datos se distribuían normalmente.
El promedio de estos también se distribuirá normalmente, y su desviación estandar se reducirá siguiendo un patrón cuadrático.
Aplicando k means sobre los datos disdretizados se obtiene una predicción del promedio así:

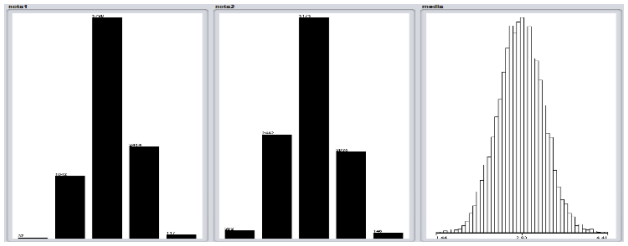

Figura 4. Resumen de histogramas *discretizados*.

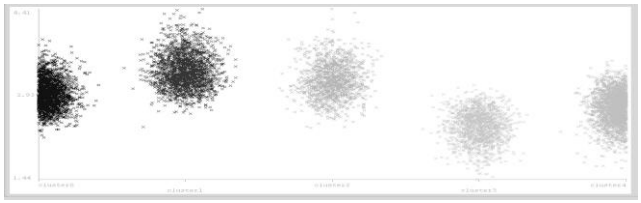

Figura 5. K means sobre el conjunto discretizado.

En la gráfica, el eje horizontal tiene los clusters, y el eje vertical el promedio ponderado. Por descarte las probabilidad de aprobación de los tres primeros clusters será mas alta que la de los dos últimos que es evidentemente mas baja. Resta entónces, estimar un estadístico que prediga la *nota faltante*, este atributo, por ser combinación lineal de distribuciones normales, ha de distribuirse normalmente, por lo que principalmente interesará conocer sus estadísticos de primer órden, y su correlación con los dos atributos conocidos; dicha correlación ha de ser alta a pesar de que los datos iniciales se hayan supuesto independientes. La correlación con la media de las dos notas, permite estimar la dependecia del resultado final con las calificaciones iniciales en conjunto. Hacer el análisis por cluster nos dará una región de factibilidad en cada uno, permitiendo de este modo estimar la probabilidad de aprobación y la confianza con que un dato estará dentro de la región, que será el criterio para predecir si un individuo aprueba o reprueba. Finalmente separar estas dos categorías y su relación con la problación total

será la predicción obtenida, que podrá darse en un modelo pseudo lineal mediante una curva de regresión. La expresion nota faltante resultó estar normalmente distribuída con media 2,89 y desviación estandar 0,86

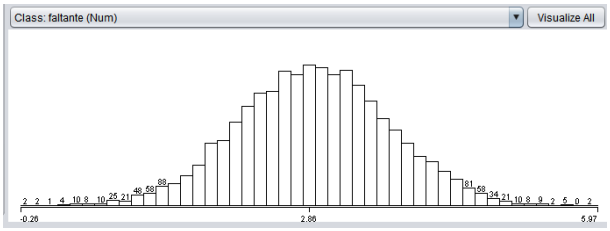

Figura 6. Histograma del atributo inducido "*nota faltante"*

Y como muestra la gráfica, está mucho mas disperso, lo que era de esperarse dado que las desviaciones de los datos originales estaban bastante justas alrededor de la media. Ahora, al confrontar las variables, promedio en el eje x y nota faltante predicha por el perceptron multicapa, se obtiene la siguiente clasificación:

.

Figura 7. Clasificación de nota faltante contra promedio

Donde queda claro que discretizando en 5 clusters, el perceptron estima 5 salidas, asociando a los promedios inferiores a la media, una nota faltante extremadamente alta; es de notar que los estudiantes incluidos en los dos clusters mas cercanos al cero en la media, pueden ser inmediatamente predichos como notas inferiores, luego su probabilidad de aprobación es muy baja. La red neuronal realiza el proceso de clasificación de manera eficiente en datos lejanos de la media.
Sin embargo hay un margen de error de predicción muy elevado en el cluster 2, que es el que contiene al 3, y debería poderse sugerir una regla que prediga cuáles son las probabilidades individuales de aprobación para los individuos aquí presentes. El descarte de las clases 0, 1,3 y 4, se da naturalmente dado que las clases 0 y 1 necesitan notas muy cercanas a 5 (incluso mas!) para aprobar con una nota de 3. Las clases 3 y 4 pueden considerarse aprobadas pues la nota faltante está muy por debajo de la media esperada. Ahora, se requeriría discretizar de nuevo, en una cantidad de clusters mucho mayor, para aproximar de una manera mas precisa la probabilidad de aprobación de los individuos de la clase 2. Asociando con una red bayesiana, se nota como el estimador de la faltante discrimina eficientemente.

Tabla 2. Grupos clasificados por perceptron multicapa y matriz de confusión.

```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,895    0,000    1,000      0,895    0,944      0,945    1,000     1,000     '(-inf-0.983699]'
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     '(0.983699-2.231502]'
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     '(2.231502-3.479305]'
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     '(3.479305-4.727108]'
               1,000    0,001    0,917      1,000    0,957      0,957    1,000     0,991     '(4.727108-inf)'
Weighted Avg.  0,999    0,000    0,999      0,999    0,999      0,999    1,000     1,000

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
   85    0    0    0   10 |    a = '(-inf-0.983699]'
    0 1948    0    0    0 |    b = '(0.983699-2.231502]'
    0    0 5585    0    0 |    c = '(2.231502-3.479305]'
    0    0    0 2262    0 |    d = '(3.479305-4.727108]'
    0    0    0    0  110 |    e = '(4.727108-inf)'


=== Summary ===

Correctly Classified Instances        9990               99.9    %
Incorrectly Classified Instances        10                0.1    %
Kappa statistic                          0.9983
K&B Relative Info Score             989343.5116 %
K&B Information Score                15340.1411 bits      1.534  bits/instance
Class complexity | order 0           15495.0545 bits      1.5495 bits/instance
Class complexity | scheme              154.9134 bits      0.0155 bits/instance
Complexity improvement     (Sf)      15340.1411 bits      1.534  bits/instance
Mean absolute error                      0.0038
Root mean squared error                  0.0184
Relative absolute error                  1.5877 %
Root relative squared error              5.331  %
Total Number of Instances            10000
```

Probability Distribution Table For promedio

faltante	'\'(-inf-2.034711]\''	'\'(2.034711-2.628902...	'\'(2.628902-3.223094...	'\'(3.223094-3.817286...	'\'(3.817286-inf)\''
'\'(-inf-0.983699]\''	0,005	0,005	0,005	0,005	0,979
'\'(0.983699-2.231502]\''	0	0	0	0,999	0
'\'(2.231502-3.479305]\''	0	0	1	0	0
'\'(3.479305-4.727108]\''	0	0,999	0	0	0
'\'(4.727108-inf)\''	0,982	0,004	0,004	0,004	0,004

Lo interesante de esta tabla es, que garantiza que la probabilidad de quedar dentro del intervalo que contiene a la media de la nota faltante, es 1 cuando el promedio se encuentra dentro del intervalo que contiene a la media. Esto confirma que pueden descartarse inmediatamente las otras cuatro clases, pero deja un problema mayor; es que aproximadamente el 50% de los datos están dentro del intervalo que contiene al promedio. Una posible solución sería, eliminar las cuatro categorías descartadas e iterar el procedimiento con el nuevo conjunto de datos hasta obtener un umbral adecuado. Se intentó clasificar con EM, pero este tipo de agrupación dejó alrededor del 70% de los datos mal clasificados.

5. RESULTADOS

Tras haber clasificado y asociado con cuatro métodos diferentes, todos permiten determinar conjuntos de individuos con tendencias marcadas de aprobacion y reprobación. Clusterizar los datos evita el trabajo de aplicar regresión y la estimación de estadísticos.
Al contrastar el promedio, con la nota faltante se ve directamente que el tipo de relación es de dependencia lineal salvo por valores aislados. Sin embargo esto se debe a dos razones pricipales:

- La independencia forzada en la creación de los dos conjuntos de notas
- Los estadísticos encontrados y faltantes, son combinaciones lineales de distribuciones normales, lo que acaba siendo una distribución normal con media predicha por combinación y media reducida,

agrupando mayor cantidad de datos cerca a la media.
Las dos razones mencionadas sugieren que debe mostrarse una correlación entre las variables conocidas y la variable predictora. La última tarea de minería aquí aplicada es la que normalmente debería aplicarse desde el principio: la selección de atributos. Al aplicar aquí selección por correlación, permitió ver la relevancia de cada variable para la predicción del atributo faltante.

Tabla 3. Correlación de las variables de entrada con el atributo predicho

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 5 faltante):
        Correlation Ranking Filter
Ranked attributes:
 0.70224   4 promedio
 0.25295   3 nota2
 0.17299   2 nota1
 0.00843   1 indice

Selected attributes: 4,3,2,1 : 4
```

El atributo que mas influencia tiene sobre la nota faltante es el promedio, con una correlación de 70%, lo cuál es natural dado que si el promedio es cercano a la media, el faltante estará en la clase mas frecuente. Sin embargo, resulta curioso ver que los atributos nota 1 y nota 2 estén correlacionados con la variable predictora en proporciones diferentes. Este fenómeno se puede entender como un vínculo entre las correlaciónes resultantes entre el promedio y las dos notas iniciales (que son altísimas). La nota 2 tiene mayor correlación con la faltante que la nota 1, o sea, dada la nota 1, la probabilidad conjunta de que la suma de ambas notas esté sobre 3; depende en mayor medida del valor de la segunda nota. Los vínculos establecidos por la correlación quedan ilustrados por el siguiente esquema:

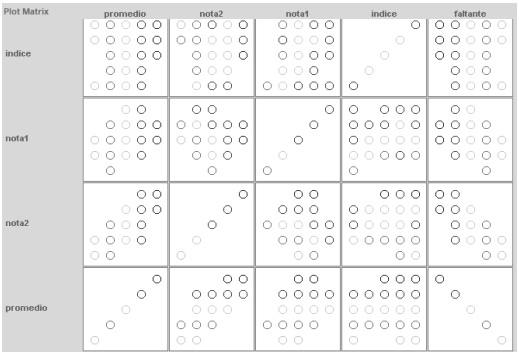

Figura 8. Tipo de relación entre clusters

Hay que notar la ultima fila, ilustra las correlaciones entre las variables del problema y la predictora.

6. CONCLUSIONES

Como se puede observar, formar clusters aumenta la capacidad de decisión sobre un conjunto tan grande de datos. Relacionar individuos permite realizar análisis cualitativos que ahorran trabajos cuantitativos complejos. Una gran fortaleza de este tipo de análisis es permitir inferir modelos que a su vez permiten conclusiones rápidas.
El análisis de datos de este tipo por regresión es el camino usual, sin embargo, una regresión permite ver datos puntuales de la evolución de un individuo como si se tratara de series de tiempo, pero la cantidad tan pequeña de calificaciones de que se dispone, no permite amoldar el modelo a ese diseño. El camino sugerido sería entonces una distribución bivariada, de la que se conocen probabilidades marginales; ahora, suponiendo independencia como se hizo en el presente trabajo, es sencillo hacer el producto de las distribuciones y calcular estadísticos que serán en realidad de poco interés, las correlaciones serán muy cercanas a cero y el modelo no brindará mayor información al respecto de las variables consideradas. Sin embargo, para poder aplicar herramientas de minería de datos acá utilizadas para conjuntos reales de datos, como discretización, agrupación y selección, sería útil conocer las distribuciones o intentar conocerlas al menos, pues es de esperarse con alto grado de certeza, que las variables no se relacionaran linealmente como se supuso, y es allí donde un análisis exploratorio de datos se hace necesario, no tanto para estimar media y varianza como es usual; más bien para intentar inferir estadísticos de orden superior. En ese caso, sería de gran utilidad una regresión que permita analizar los datos mediante un modelo pseudo-lineal y discretizar como aquí se hizo.
En la literatura se muestra que una herramienta que permite hacer predicción sobre la posibilidad de éxito académico, son los árboles de decisión; sin embargo, los datos allí utilizados son mucho menores en extensión y tienen grandes cantidades de atributos. Discretizar el problema genera ruido en los datos y evita la estimación de correlaciones inducidas de manera clara en el caso individual, pero perceptible y apreciable de manera mucho más útil en el caso grupal.
El mayor problema de esto radica en que los casos que aquí generaron mayor correlación pueden perfectamente ser eventos independientes: de hecho, ¡lo son!, pues provienen de simulaciones diferentes. El caso contrario es más interesante: un conjunto de datos reales debe tener algún nivel de dependencia y podría llegar a ocurrir que no presente correlación estadísticamente significativa, por ello vale la pena sugerir un análisis exploratorio de datos previo al procesamiento con herramientas de minería, de hecho, y para encerrar en bucle con la primera conclusión, el ruido y la dispersión que presenten los datos de un repositorio real, serían sin duda, los datos de mayor valor para obtención de información del conjunto.

7. TRABAJOS FUTUROS

En la posteridad sería provechoso medir la correlación de datos, e intentar obtener reglas de asociación entre atributos altamente correlacionados. La cantidad de atributos deberá ser mayor, cuantos más se tengan en cada registro, la tarea de agrupación se hará más sencilla. En este trabajo las herramientas k-means y red neuronal multicapa fueron de enorme provecho, pero se esperaba un resultado mejor de EM. Con seguridad quien disponga de una base de datos real, podría medir la eficiencia y precisión reales de este algoritmo.

8. REFERENCIAS

[1] Prabha, S.L. and D.A.M. Shanavas, Educational data mining applications. Operations Research and Applications: An International Journal (ORAJ), 2014. 1(1)..

[2] Kumar S. Anupama and Dr. Vijayalakshmi M.N. (2011). Efficiency of Decision Trees in Predicting Students Academic Performance. Computer Science & Information Technology 02, pp. 335–343.

[3] Cristobal Romero And Sebastian Ventura (2010), - Educational Data Mining: A Review Of The State Of The Art‖,IEEE transactions on systems, man, and cybernetics-part c: applications and reviews, vol. 40, no. 6, November 2010
[4] Charanjit Bambrah, Minakshi Bhandari, Nirali Maniar, Prof. Vandana Munde―Mining Association Rules in Student Assessment Data‖ International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 3, March 2014 Copyright to IJARCCE www.ijarcce.com 5340
[5] S. M. a. T. Acharya, «Data mining: multimedia, soft computing and bioinformatics» de Data mining: multimedia, soft computing and bioinformatics, John Wiley & Sons, 2003.
[6] M. J. R.-Q. J. Hernández-Orallo, Introducción a la Minería de Datos, Prentice Hall / Addison-Wesley, 2004.
[7] Hartigan, J., Agrupación Algorithms, John Wiley & Sons, New York, 1975
[8] Hartigan, J. y Wong, M., "Algorithm AS139: A k-means agrupación algorithm", Applied Statistics, Vol. 28, 1979, pp. 100-108.
[9] Mitchell, T. M., Machine Learning. McGraw-Hill, 1997.
[10] Forgy, E., "Cluster analysis of multivariate data: Efficiency versus interpretability of classification", Biometrics, Vol. 21, 1965, pp. 768-780.
[11] Duda, R. y Hart, P., Pattern Classification and Scene Analysis, Wiley & Sons, 1973.
[12] Armstrong, J. Scott. «Illusions in Regression Analysis». International Journal of Forecasting (forthcoming) 28 (3): 689 (2012)
[13] Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. (Eds.). (2010). Handbook of educational data mining. CRC press.
[14] Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601-618.
[15] Mohamad, S. K., & Tasir, Z. (2013). Educational data mining: A review. Procedia-Social and Behavioral Sciences, 97, 320-324.
[16] Iten, L., Arnold, K., & Pistilli, M. (2008). Mining real-time data to improve student success in a gateway course. Eleventh Annual TLT Conference, Purdue University, March 4.
[17] Lonn, S., & Teasley, S. D. (2009). Saving time or innovating practice: Investigating perceptions and uses of Learning Management Systems. Computers & Education, 53(3). 686-694.