RED BAYESIANA JERÁRQUICA PARA ANALIZAR DATOS EDUCATIVOS
HIERARCHICAL BAYESIAN NETWORK TO ANALYZE EDUCATIONAL DATA

Byron Oviedo, Universidad Técnica Estatal de Quevedo, https://orcid.org/0000-0002-5366-5917, boviedo@uteq.edu.ec, 120508, Quevedo-Ecuador

INFORMACIÓN DE LA INVESTIGACIÓN O DEL PROYECTO: Este proyecto fue financiado por la Universidad Técnica Estatal de Quevedo, Ecuador, bajo la convocatoria interna **00020130516**, con el propósito de analizar la deserción estudiantil en la universidad

RESUMEN.

Este documento presenta una propuesta para implementar un método de clúster que mejor acople los datos educativos. El uso de modelos gráficos probabilísticos en el campo de la educación ha sido considerado para esta investigación. Pero el problema con estos procedimientos generales de aprendizaje proviene de la presencia de un alto número de variables que miden diferentes aspectos del mismo concepto. En este caso, tenemos que todas las variables tienen algún grado de dependencia entre ellas, sin una verdadera estructura causal. Por lo tanto, se presenta un nuevo procedimiento que hace una agrupación jerárquica de los datos mientras se aprende una distribución de probabilidad conjunta. Se busca hacer una prueba para cada caso donde la probabilidad se mide en cada modelo usando algoritmos de propagación. Luego, el logaritmo de probabilidad se aplica a cada caso y los resultados se agregan en cada modelo para determinar el mejor ajuste para el propuesto. El método se aplica al análisis de un conjunto de datos educativos: evaluación de estudiantes de profesores de la Universidad Gazi en Ankara (Turquía).

PALABRAS CLAVE: Redes Bayesianas, K2, PC, EM, Clustering Jerárquico, Desempeño Académico, Evaluación Estudiantil

ANALYTICAL SUMMARY.

This document presents a proposal to implement a cluster method that best matches the educational data. The use of probabilistic graphic models in the field of education has been considered for this research. But the problem with these general learning procedures comes from the presence of a high number of variables that measure different aspects of the same concept. In this case, we have that all variables have some degree of dependence between them, without a true causal structure. Therefore, a new procedure is presented that makes a hierarchical grouping of the data while learning a joint probability distribution. We try to make a test for each case where the probability is measured in each model using propagation algorithms. Then, the probability logarithm is applied to each case and the results are added in each model to determine the best fit for the proposed one. The method is applied to the analysis of a data set of educational data: evaluation of students of professors of the Gazi University in Ankara (Turkey).

KEYWORDS: Bayesian Networks, K2, PC, EM, Hierarchical Clustering, Academic Performance, Student Evaluation

INTRODUCCIÓN

En este trabajo se presenta una propuesta para aplicar un método de clúster jerárquico para analizar datos en los que se suponga que las relaciones existentes entre los mismos se deban a la existencia de variables ocultas. Este suele ser el caso de muchos conjuntos de datos entre los que están los datos de evaluación por parte de los estudiantes a los instructores de un curso; se propone un nuevo método basado en clasificación no supervisada para la construcción de una jerarquía de variables artificiales a partir de las variables observadas $X = \{X_1,.....,X_n\}$. La idea básica es similar a AUTOCLASS (procedimiento no supervisado que tiene la ventaja de poder imputar varios atributos tras una lectura única del fichero de aprendizaje), pero en lugar de usar todas las variables, primero hacemos una clasificación de las

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2022.1.1.641
ISBN: 978-628-95207-0-5 **ISSN:** 2414-6390

variables observadas por grupos que tengan un alto grado de dependencia entre sí. Por cada grupo de variables se estima una variable oculta con AUTOCLASS [1] y se repite el proceso con las variables ocultas introducidas. Se obtiene así una jerarquía de variables ocultas que determinan las relaciones entre las variables observadas de manera simple. El resultado final será una red bayesiana B, que inicialmente contiene variables **X** y se completa con las variables ocultas [2].

Se considera que el agrupamiento o "clúster" es el hecho de clasificar a los elementos de un conjunto de observaciones formando grupos con ellas, de manera tal, que los individuos dentro de cada grupo sean similares entre sí. Se supone que se tiene un conjunto de datos D = $\{x^1,...,x^N\}$, y hay que dividirlos en categorías de tal modo que los casos dentro de un mismo grupo sean más parecidos entre sí que a los casos de otros grupos. Estos casos "parecidos" son considerados con una noción de similitud o distancia entre muestras.

El procedimiento de este trabajo se puede considerar como un método de agrupamiento multidimensional donde cada variable oculta muestra una forma de agrupar los datos. Ante un nuevo caso, se puede obtener mediante algoritmos de propagación, la probabilidad de pertenecer a cada uno de los grupos asociados a los valores de las distintas variables ocultas.

En una serie de artículos [3], proponen un modelo de "redes jerárquicas Bayesianas", aplicables al dominio del razonamiento psicológico y en específico al esquema teleológico de Csibra y Gergely.

En esta investigación se realiza un estudio experimental utilizando herramientas de acceso libre como Elvira y Weka para:

I. Comparación con otros algoritmos de aprendizaje con bases de datos de la UCI.
II. Datos de evaluación de encuestas realizadas por los estudiantes de la Universidad de Gazi.

Adicionalmente se hace referencia a la investigación del uso clúster jerárquico en el campo de la enseñanza para la realización del diagnóstico de estudiantes y poder determinar el problema de evaluación de estudiantes a profesores y puede ser usado también para determinar la deserción estudiantil en las universidades, el mismo que ha sido ya estudiado por algunos investigadores. Magaña Echeverría lo analiza haciendo uso de clúster agrupando a individuos u objetos en conglomerados de acuerdo a sus semejanzas, maximizando la homogeneidad de los objetos dentro de los conglomerados a la vez que maximiza la heterogeneidad entre agregados [4].

Otro caso de estudio para predecir la probabilidad de que un estudiante abandone la institución educativa se han realizado utilizando técnicas de minería de datos para lograr el objetivo; entre ellos tenemos a [5], quienes realizaron un trabajo basado en el uso del conocimiento, en reglas de descubrimiento y en el enfoque TDIDT (Top Down Induction of Decision Trees) sobre la base de datos de la gestión académica del consorcio SIU de Argentina (que reúne 33 universidades de Argentina), lo cual permite un interesante análisis para encontrar las reglas de conducta que contienen variables de ausencia.

I. REDES BAYESIANAS Y CLÚSTER JERÁRQUICO

Si tenemos un conjunto de variables denotadas $X = \{X_1 , X_2 , .., X_n \}$ donde cada variable Xi toma valores dentro de un conjunto finito Ω_{xi} ; entonces, usamos x_i para expresar uno de los valores de X_i , $xi \in \Omega_{xi}$. Ahora si tenemos un conjunto de índices denotado como I, simbolizamos x_i, $i \in I$.

Notaremos el conjunto de todos los índices N= {1, 2, .., n} y Ω_I representa a los elementos de Ω_{xi} y se les llama configuraciones de X_i que serán representadas con x o x_I.

Definición de Red Bayesiana: es un grafo acíclico dirigido que representa un conjunto de variables aleatorias y sus dependencias condicionales (véase Figura 1), donde Xi es un suceso aleatorio representado por cada nodo (conjunto de variables), y la topología del gráfico muestra las relaciones de independencia entre variables de acuerdo con el criterio d-separación [1].

Cada nodo X_i tiene una distribución de probabilidad condicional pi(Xi|Π(X_i)) para esa variable, dadas sus padres. Por lo tanto una red bayesiana determina una distribución única probabilidad conjunta tal como se puede apreciar en la ecuación 1 y la figura 1:

$$p(X = x) = \prod p_i(x_i | \prod(x_i)) \ , \forall x \in \Omega_x \qquad (1)$$

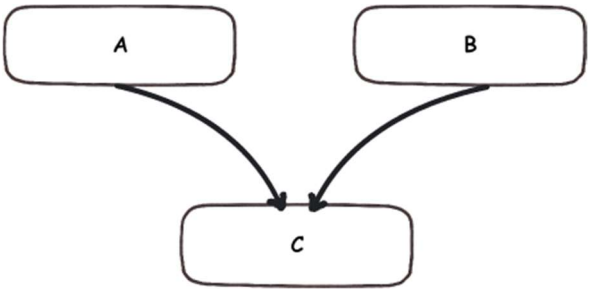


Figura 1: EJEMPLO DE REDES BAYESIANAS

Algoritmo PC: Es uno de los algoritmos utilizados para aprendizaje bayesiano. Este Algoritmo se basa en pruebas de independencias entre variables $I(X_i, X_j |A)$, donde A es un subconjunto de variables, donde se asume que se tienen los datos suficientes y que las pruebas estadísticas no tienen errores. El algoritmo PC empieza con un grafo completo no dirigido para, posteriormente, ir reduciéndolo. Primero elimina las aristas que unen dos nodos que verifican una independencia condicional de orden cero, después las de orden uno, y así sucesivamente. El conjunto de nodos candidatos para formar el conjunto separador (el conjunto al que se acondiciona) es el de los nodos adyacentes a alguno de los nodos que se pretende separar. En este trabajo usamos la implementación hecha de PC en el programa Elvira, con un nivel de significancia de 0.05. PC realiza su cálculo basado en independencia y no en optimización del score.

Require: Set of variables **X**, Independence test *I*
Initialize a complete undirected graph G'
i=0
repeat
  for $X \in$ **X do**
    for $Y \in ADJ(X)$ **do**
      for $S \subseteq ADJ(X) - \{Y\}, | S | = i$ **do**
        if $I(X, Y | S)$ **then**
          Remove the edge $X - Y$ from G'
        end if
      end for
    end for
  end for
  i=i + 1
until $| ADJ(X) | \leq i, \forall X$
Orient edges in G'

Algoritmo 2: ALGORITMO PC

Los algoritmos basados en funciones de *score+search* pretenden tener un grafo que mejor modelice a los datos de entrada, de acuerdo con un criterio específico. Cada uno de estos algoritmos utilizan una función de evaluación junto con un método que mide la bondad de cada estructura explorada en el conjunto total de estructuras. Durante el proceso de exploración, la función de evaluación es aplicada para evaluar el ajuste de cada estructura candidata a los datos. Cada uno de estos algoritmos se caracteriza por la función de evaluación y el método de búsqueda utilizados.

Cada uno de los algoritmos como función de evaluación utilizan una métrica no especializada en clasificación usada comúnmente en el aprendizaje de redes bayesianas como BIC [6], BDEu [7] y K2 [8]. Por eficiencia la métrica usada debe ser descomponible, así sólo tendremos la parte que el operador modifica y no el grafo entero.

La métrica K2 para una red G y una base de datos D como lo representa la ecuación 2:

$$f(G:D) = \log P(G) + \sum_{i=1}^{n}\left[\sum_{k=1}^{S_i}\left[\log\left(\frac{\Gamma(n_{ik})}{\Gamma(N_{ik}+n_{ik})}\right)+\sum_{j=i}^{r_i}\log\left(\frac{\Gamma(N_{ijk}+n_{ijk})}{\Gamma(n_{ijk})}\right)\right]\right] \quad (2)$$

Dónde: N_{jk} es la frecuencia de las configuraciones encontradas en la base de datos D de las variables x_i; n es el número de variables, tomando su j-ésimo valor y sus padres en G tomando su k-ésima configuración; s_i es el número de configuraciones posibles del conjunto de padres r_i es el número de valores que puede tomar la variable x_i :

$$N_{ik} = \sum_{j=1}^{r_i} N_{ijk} \quad (3)$$

Γ es la función Gamma
La métrica BIC se define de la siguiente manera:

$$f(G:D) = \sum_{i=1}^{n}\sum_{j=1}^{r_i}\sum_{k=1}^{s_i} N_{ijk}\log\left(\frac{N_{ijk}}{N_{ik}}\right) - \frac{1}{2}C(g)\log(N) \quad (4)$$

Dónde: N es el número de registros de la base de datos; C(G) es una medida de complejidad de la red G, definida como, ver ecuación 5:

$$C(G) = \sum_{i=0}^{n}(r_i - 1)s_i \quad (5)$$

La métrica BDEu depende de un solo parámetro, el tamaño muestral α y se define de la siguiente manera, :

$$gBDeu(G:D) = \log(p(D)) +$$

$$\sum_{i=1}^{n}\left[\log\left(\frac{\Gamma\frac{\alpha}{q_i}}{\Gamma(N_{ij}+\frac{\alpha}{q_i})}\right)+\sum_{k=1}^{r_i}\log\left(\frac{\Gamma(N_{ijk}+\frac{\alpha}{q_i})}{\Gamma\frac{\alpha}{q_i}}\right)\right] \quad (6)$$

Algoritmo K2: Este algoritmo está basado en la búsqueda y optimización de una métrica bayesiana. K2 realiza una búsqueda voraz y muy eficaz para encontrar una red de buena calidad en un tiempo aceptable [8]. K2 Es un algoritmo heurístico codicioso basado en la optimización de una medida. Esa medida se usa para explorar, mediante un algoritmo de ascensión de colinas, el espacio de búsqueda formado por todas las redes que contienen las variables de la base de datos. Este algoritmo supone un orden entre las variables.

Require:
$nvar$ = número de variables binarias.
$nsample$ = número de cadenas binarias (para calcular g).
$xsample$ = un conjunto de cadenas binarias (para calcular g).
u = máximo número de padres permitidos.
i = índices en el orden dado.
Ensure:
Para cada nodo x_i, un conjunto de padres π_i.

```
for i = 1 to nvar do
    π_i ← ∅
    P_old ← g(i, π_i)
    OKToProceed ← true
    while OkToProceed and |π_i| < u do
        sea z el nodo en los predecesores Pred(x_i) que maximiza
        g(i, π_i ∪ {z})
        P_new ← g(i, π_i, ∪{z})
        if P_new > P_old then
            P_new = P_old
            π_i = π_i ∪
        else
            OKToProceed =false
        end if
    end while
    return Nodo: x_i, padres de este nodo: π_i
end for
```

Algoritmo 2: ALGORITMO K2

Algoritmo EM (Expectation Maximization) [9]: es un método para encontrar el estimador de máxima verosimilitud de los parámetros de una distribución de probabilidad, este algoritmo es muy útil cuando parte de la información está oculta. Para encontrar estos parámetros óptimos se deben realizar dos pasos: primero (Expectation) calcular la esperanza de la verosimilitud con respecto a la información conocida y unos parámetros propuestos, luego (Maximization) maximizar con respecto a los parámetros; estos dos pasos se repiten hasta alcanzar la convergencia [10].

El agrupamiento jerárquico se lo desarrolla en varias fases tal como se indica a continuación

Variables de Agrupamiento: Se supone un conjunto de variables **X** = (X_1,……,X_n) y un conjunto de observaciones D = x^1,...,x^N. Se realiza el cálculo de una matriz A, n x n, donde a_{ij} representa el grado de dependencia de las variables X_i y X_j. Este grado de dependencia es calculado con una métrica BDEu. Si D es el conjunto de datos observados, se tiene:

a_{ij} = Dep(X_i, X_j | D) = Score(X_i, {X_j} | D – Score (X_i, ∅ | D) (7)

Donde Score es calculado con BDEu dado un parámetro S. Conocidas las propiedades de BDEu (grafos equivalentes que tienen igual score), tenemos que a_{ij} = a_{ji}; es decir, la matriz A es simétrica. Luego, definimos la siguiente relación en ***X***:

$$X_i \rightarrow X_j \; si \; y \; solo \; si \; X_j = arg\acute{m}\acute{a}x_k a_{ik}, \; i \neq j, \; y \; a_{ij} > 0$$

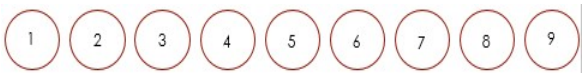


Figura 2: CONJUNTO DE VARIABLES

Una vez que hemos calculado el grado de dependencia se une las variables iniciales con aquella variable que tiene mayor grado de dependencia siempre que esta sea mayor que 0. Los grupos van a ser las componentes conexas del grafo resultante que se obtiene uniendo cada variable con su variable de máxima dependencia, tal y como se presenta en la figura 2. [2]

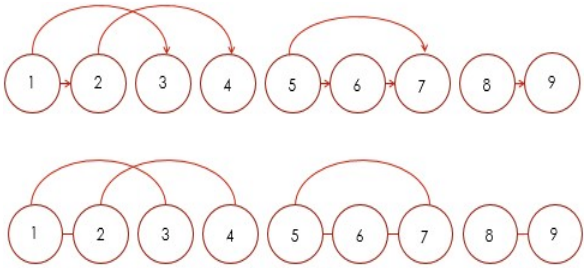


Figura 3: VARIABLES JUNTO A LA DE MÁXIMA DEPENDENCIA

Para llevar a cabo el cálculo de forma efectiva de estos grupos de variables se usa el siguiente algoritmo:

1: Conjunto $\mathscr{P} \leftarrow \{\{X_1\},\ldots,\{X_n\}\}$
2: **for all** par de variables X_i, X_j $(i \neq j)$ **do**
3: **if** $a_{ij} = \max_{k \neq j} a_{ik}$ y $a_{ij} > 0$ **then**
4: Sea $\mathbf{C_1} \in \mathscr{P}$, tal que $X_i \in \mathbf{C_1}$
5: Sea $\mathbf{C_2} \in \mathscr{P}$, tal que $X_j \in \mathbf{C_2}$
6: **if** $\mathbf{C_1} \neq \mathbf{C_2}$ **then**
7: $\mathbf{C} \leftarrow \mathbf{C_1} \cup \mathbf{C_2}$
8: $\mathscr{P} \leftarrow (\mathscr{P} \cup \{\mathbf{C}\}) \setminus \{\mathbf{C_1}, \mathbf{C_2}\}$
9: **end if**
10: **end if**
11: **end for**
12: **return** $\mathscr{P}$

Algoritmo 3: AGRUPAMIENTO JERÁRQUICO

Adición de variables ocultas artificiales: Una vez calculados los grupos, se añade una variable para cada grupo de variables ocultas y se conecta con cada una de las variables de su grupos y se ejecuta un algoritmo recursivo donde se vuelve a aplicar el procedimiento pero con las variables ocultas (aux) añadidas y para hacer está misma operación se requiere una matriz de dependencia como se puede apreciar en la figura 4.

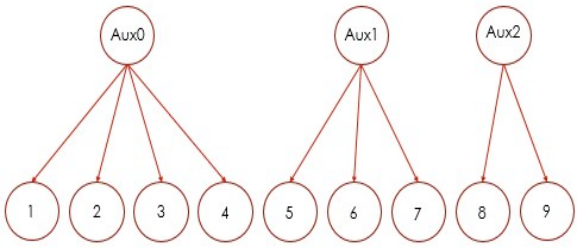


Figura 4: ADICIÓN DE VARIABLES AUXILIARES

Cálculo Recursivo: Como las variables aux son no observadas no puedo hacer un cálculo directamente de esa dependencia basado en score, pero se basa en un grado de dependencia heurística entre las variables ocultas que se calcula como la media de los grados de dependencia de las variables de cada uno de los grupos asociados a las variables asociadas, como se aprecia en la figura 5.

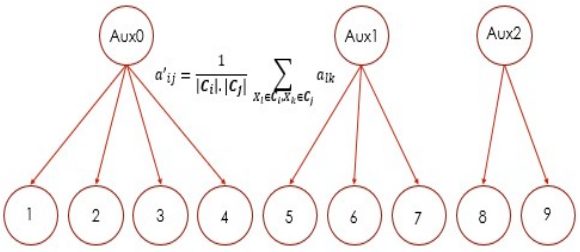


Figura 5: RESULTADO DEL CÁLCULO RECURSIVO

Con esta matriz de dependencia se vuelve a aplicar el mismo procedimiento y se va añadiendo variables ocultas en la parte superior. Al final se obtiene un árbol o bosque de árboles. En este clúster jerárquico si una variable está sola en un grupo no se añade variable auxiliar, en la figura 6 se puede apreciar el grafo resultante.

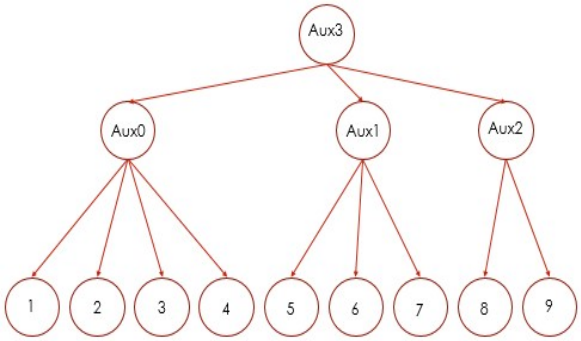


Figura 6: ÁRBOL GENERADO

De esta manera puede que no se obtenga un árbol sino un bosque de árboles. ¿Cuándo ocurre que una variable está sola en un grupo?, pues cuándo la variable de máxima dependencia con ella tiene una dependencia negativa.

Una vez q tengo calculado un grafo, se estima los parámetros y un paso previo es determinar el número de casos de cada variable oculta que no vienen determinado.

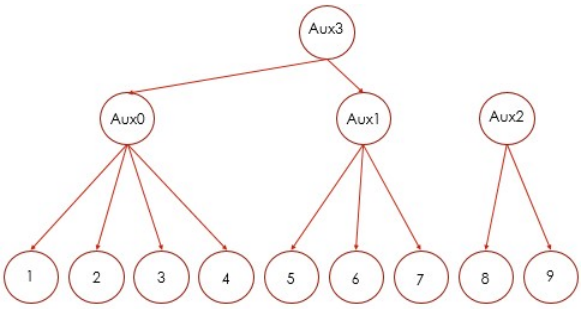


Figura 7: DETERMINACIÓN VARIABLES OCULTAS

Estimación de parámetros y optimización de casos:

- Paso previo: Este paso asigna a cada variable no observada una variable observada. La variable que más dependencia tiene con la otra variable (es la suma de los grados de dependencia de cada variable con las demás y calcula la variable que tiene mayor valor de suma). Este procedimiento previo sirve para establecer un número inicial de casos para esta variable Aux.
-

El número inicial de casos de cada variable no observada es el mismo número de casos que la variable observada asociada. La idea es que Aux0 resuma las variables 1..4 y busca la variable más significativa, si la variable más significativa es 2 entonces el número de casos de Aux0 es el mismo número de casos que la variable 2.

Para estimar los parámetros y calcular la optimización de números de casos de una variable oculta vamos a asignarle de manera previa una variable observada a cada variable oculta. ¿Cómo? A cada variable oculta de 1er nivel está asociado a un grupo de variables, en ese grupo de variables calculamos la más representativa, ¿Cuál es?, la que maximiza grado de dependencia de esa variable con el resto de variables, la más dependiente de acuerdo a la matriz A.

En un nivel superior algo parecido, se le asigna una variable auxiliar de nivel inferior y la variable observada que se le asigna a la del nivel superior es la misma que la variable observada asociada a la del nivel inferior.

Para cada variable de nivel superior (Aux3), con el mismo procedimiento de antes se asigna una variable.

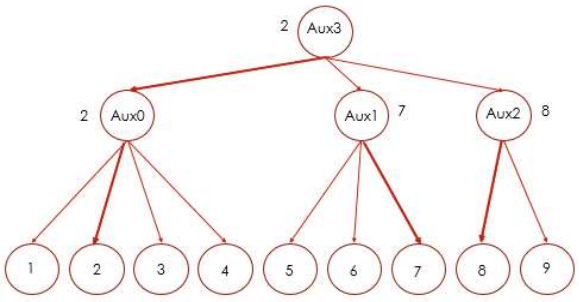


Figura 8: DETERMINACIÓN VARIABLE MÁS SIGNIFICATIVA

- Estimando las probabilidades condicionadas: inicial

Como se tiene un número de casos inicial para cada variable entonces se aplica el algoritmo EM para los parámetros. El algoritmo EM es un algoritmo de óptimo local que cuyo comportamiento depende de los parámetros iniciales. Ya que se tiene asociadas las variables a cada variable oculta se hace la estimación inicial de los parámetros. ¿Cómo se hace esto? midiendo las frecuencias en la base de datos, lo que pasa que aquí tenemos variables no observadas, para este tipo de variables lo que se hace es sustituirlas por las variables observadas. Pero en este caso para evitar dependencias funcionales demasiados grandes se debe suavizar usando una suavización que es más importante que la corrección de Laplace. Se suma log (N+1) para todos los casos y un factor aleatorio para incluir algo de diversidad.

$$\theta_{ik} = \left(\theta_{i1k}, \dots, \theta_{ir_{ik}}\right) \; donde \; \theta_{ijk} = P\left(Z_i = Z_i^j \mid \pi_{ik}\right) \quad (8)$$

$$\theta_{ijk}^0 = \frac{N_{ijk} + \log(N+1) + \left(R_{ijk}\right)}{\sum_j \left(N_{ijk} + \log(N+1) + \left(R_{ijk}\right)\right)} \quad (9)$$

- Estimando las probabilidades condicionadas: Aplicación EM

Al aplicar EM se calcula el valor esperado de los estadísticos suficientes por los parámetros dado unos parámetros iniciales, luego se vuelve a calcular los parámetros

$$N_{ijk}^t = E\left[N_{ijk} \middle| \theta^t\right] = \sum_{x \in D} P\left(Z_i = Z_i^j, \Pi_i = \pi_{ik} \middle| X = x\right) \quad (10)$$

$$\theta_{ijk}^{t+1} = \frac{N_{ijk}^t + 1}{\sum_{j=1}^{r_i} \left(N_{ijk}^t + 1\right)} \quad (11)$$

- Estimando las probabilidades condicionadas: Optimización de números de casos

Hasta ahora se ha estimado unos parámetros para un número de casos de cada variable artificial que han sido calculados de forma heurística y ahora lo toca es optimizar ese número de casos con apoyo de una métrica como Bic o Akaike. Mientras haya cambios en el score se va aumentando el número de casos en esta métrica y cuando ya no haya mejoras va disminuyendo. Aquí lo que se ha hecho es calcular mediante métrica cuál es el número de casos óptimos de la variable aux.

$$BIC = \sum_{x \in D} \log\left(P(x|G)\right) - \frac{1}{2} Dim(G) \log(N) \quad (12)$$

$$AKAIKE = \sum_{x \in D} \log\left(P(x|G)\right) - Dim(G) \quad (13)$$

- Estimación Inicial de Probabilidades tras un Cambio

Como cada vez que se cambia el número de casos en una variable aux hay que realizar una estimación en el algoritmo EM; Se determinó un procedimiento para empezar con unos parámetros que estén cercanos al otro para que la convergencia sea más rápida. Si se cambia los valores de números de casos de una variable Aux; por ejemplo, si la variable auxiliar no está involucrada en una probabilidad condicionada entonces se dejan los mismos parámetros que se habían obtenido en la anterior optimización. Y si se cambia el número de casos, se determina unos parámetros iniciales que dependerán si la variable Aux está en Z_i o si aumenta o disminuye, por ejemplo si la variable que cambia el número de casos es la variable Z_i: si aumenta el número de casos, **se** asigna unos parámetros en los que los casos antiguos juegan un papel similar a los de antes y el nuevo caso juega un papel aleatorio (están calculadas las probabilidades con números aleatorios y los casos anteriores juegan un papel similar)

$$\theta^0_{ijk} = \begin{cases} \theta_{ijk}\frac{r_i}{r_i+2R_{ik}} & if\ j<r_i \\ \frac{2R_{ij}}{r_i+2R_{ik}} & if\ j=r_i+1 \end{cases} \quad (14)$$

Si disminuye el número de casos, lo que se hace es que el último caso de la nueva variable resuma los 2 casos anteriores de acuerdo con la fórmula.

$$\theta^0_{ijk} = \begin{cases} \theta_{ijk} & if\ k<r_{i-1} \\ \left(\theta_{ijk}+\theta ij(k+1)\right) & if\ k=r_i-1 \end{cases} \quad (15)$$

Si la variable que cambia el número de casos es la variable que aparece como padre, si aumenta el número de valores, el nuevo valor va a tener un papel que sea equivalente a la media de las distribuciones de los anteriores valores y una componente aleatoria

$$\theta^0_{ijk} = \begin{cases} \theta_{ijk} & if\ j<r_i \\ \frac{0{,}5\frac{\sum_{k=0}^{r_i}\theta_{ijk}}{r_j}+R_j}{0{,}5+\sum_j R_j} & if\ k=r_i+1 \end{cases} \quad (16)$$

Si disminuye el último nuevo valor va a resumir los dos últimos valores anteriores, y la media de las anteriores probabilidades.

$$\theta^0_{ijk} = \begin{cases} \theta_{ijk} & if\ k<r-1_i \\ 0{,}5(\theta_{ijk}+\theta_{ij(k+1)}) & if\ k=r_i-1 \end{cases} \quad (17)$$

EXPERIMENTACIÓN

Se empezará el análisis experimental, dónde se va a contrastar este modelo con otros modelos de aprendizaje de redes bayesianas en función de la capacidad de representar la distribución de probabilidad conjunta. Como es la distribución conjunta se calcula el logaritmo de la probabilidad en la validación cruzada como medida de comportamiento. Se indica las características básicas del análisis experimental.

Se compara los resultados obtenidos con nuestro método con los algoritmos de aprendizaje K2, PC, Naive Bayes y Autoclass. Las bases de datos son:

- Bases de datos de la UCI de encuestas a los estudiantes sobre el profesorado en la Universidad de Gazi.

Como dato adicional se indica que se trabaja con una validación cruzada de 10 tandas y sr utiliza un test no paramétrico como es Friedman con un nivel de significancia de 0,05 el mismo que nos da un valor de p=4.96E^{-6}.

ALGORITMO	PUNTUACIÓN
PC	**6.5000**
K2	4,64285
NBayes	4.07142
Aclass2	4.21428
AClass5	**2.42857**
HNB-A	2.49996
HNB-B	3.64285

ALGORITMO	HOLM
PC	**0.00833**
K2	**0.01000**
NBayes	0.01666
Aclass2	0.01250
HNB-A	0.05000
HNB-B	0.02500

TABLA I: RANKING CON BASE DATOS UCI

Se puede observar que se tienen diferencias significativas y que el algoritmo AClass5 es el que tiene mejor puntuación y por lo tanto se lo considera como algoritmo de control.

Los algoritmos de la investigación se encuentran en 2do y 3er lugar en referencia a ranking. Una vez determinado que hay diferencias significativas se toma como algoritmo de control el mejor aclass5 y se realiza unos tests estadísticos para ver con que otro procedimiento tiene diferencias significativas. Se puede determinar que es significativamente mejor que los algoritmos PC y K2 (valor p< valor de Holm correspondiente). Para el resto no hay diferencias significativas. Y de manera particular la propuesta de algoritmo jerárquico tiene resultados muy similares.

A continuación se realiza el cálculo con conjunto de datos de la UCI, Universidad de Gazi en Ankara (Turquía) sobre cuestionario de preguntas típicas que se hacen a los estudiantes para evaluar a los docentes sobre un curso determinado. Aquí se podrá encontrar preguntas relacionadas con la actitud del docente, desarrollo del curso.

TABLA II: PREGUNTAS UNIVERSIDAD DE GAZI EN ANKARA

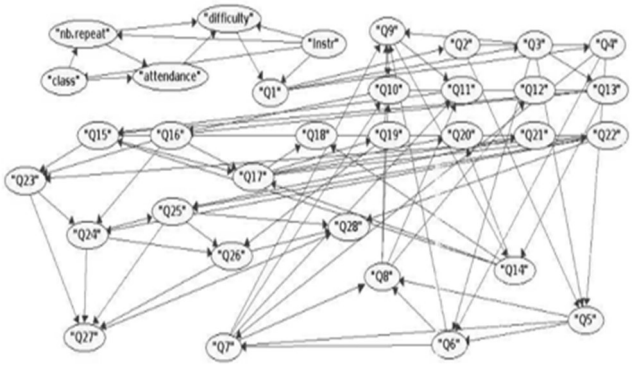
Variables	Descripción
Instr	Identificación del Instructor ; se toman valores 1,2,3
Clase	Códigos de Recursos, se toman variables del 1 a 13
Repeticiones	Número de veces que el estudiante toma este curso
Asistencia	Código de nivel de asistencia; se toman valores 0,1,2,3,4
Dificultad	Dificultad del curso según el estudiante valores 1,2,3,4,5
Q1	El contenido, método de enseñanza y sistema de evaluación, se proporcionó al principio
Q2	El curso y sus objetivos estaban claramente establecidas al inicio del periodo,
Q3	El curso merecía la cantidad de créditos asignados
Q4	El curso fue impartido de acuerdo al syllabus entregado al primer día
Q5	Las discusiones en casa, las tareas asignadas y resultados fueron satisfactorios
Q6	Los libros y otro Recursos del curso fueron suficiente y actualizados
Q7	El curso permitió trabajo en campo Aplicaciones en laboratorio análisis y otros
Q8	Las pruebas, tareas proyectos y exámenes contribuyeron a ayudar al aprendizaje
Q9	Disfrute muchos las clases y deseaba participar activamente en las conferencias
Q10	Mis expectativas Iniciales sobre el curso se cumplieron al final del periodo o año
Q11	El curso fue pertinente y beneficioso para mi desarrollo profesional
Q12	El curso me Ayudo a ver la Vida y el mundo con una nueva vida
Q13	Los conocimiento del instructor eran relevantes y actualizados
Q14	El instructor vino preparado a las clases
Q15	El instructor enseño de Acuerdo al plan de estudios entregado
Q16	El instructor estaba comprometido con el curso y era comprensible
Q17	El instructor llego a tiempo para las clases.
Q18	El instructor tiene voz suave y fácil de entender
Q19	El instructor hizo uso efectivo de las horas de clases
Q20	El instructor explico el curso y estaba dispuesto a ayudar a los estudiantes
Q21	El instructor demostró un enfoque positivo a los estudiantes
Q22	El instructor estaba abierto y respetuoso a las opiniones de los estudiantes
Q23	El instructor Incentivo la participación en el curso
Q24	El instructor entrego tareas/proyectos y mantuvo ayuda guiada a estudiantes
Q25	El instructor respondió a las preguntas sobre el curso dentro y fuera del Aula de Clases
Q26	El sistema de evaluación permite medir de manera efectiva los objetos del curso
Q27	El instructor proporciono la solución de los exámenes y los discutió e clases
Q28	El instructor trato a todos los estudiantes de manera correcta y Objetiva

Se puede observar que para este caso nuestro modelo HNBayesEM-A (akaike) obtiene los mejores resultados y con la métrica Bic es muy competitivo. Se puede también determinar que Autoclass mejora a mayor cantidad de clases y que con 12 ya supera a K2.

TABLA II: RESULTADOS LUEGO DE LA EXPERIMENTACIÓN

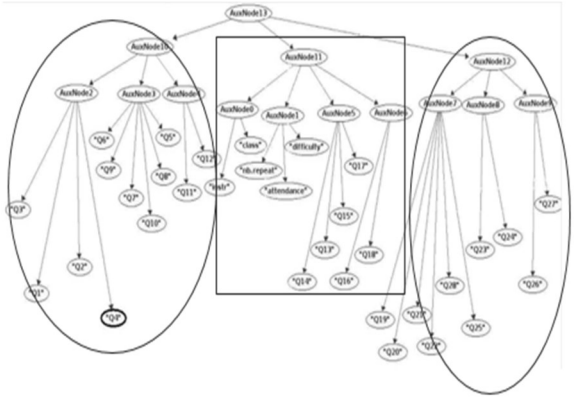
PC	K2	NBayes	AClass2
-263425.28	-130190.69	-171900.24	-228961.76
AClass6	**AClass12**	**HNBayesEM-A**	**HNBayesEM-B**
-140751.66	-128164.73	**-**113482.01	-114998.16

FIGURA 9. Red Bayesiana Con K2



Esa es la red aprendida por K2, que como se dijo al inicio presenta un tipo de red dónde tosas las variables están relacionas entre sí en forma confusa, esto se debe a que todas las variables miden un grado de satisfacción del estudiante como con la materia.

FIGURA 10. Red Bayesiana Con Clúster Jerárquico



En esta figura se visualiza el árbol obtenido por nuestro modelo, donde podemos indicar fácilmente que:

AuxNode2 resume las variables Q1; Q2; Q3; Q4, (información que se le da al estudiante).
AuxNode3 resume las variables Q5, Q6, Q7, Q9, Q10 (materiales usados por instructor)
AuxNode4 está relacionada con las variables Q11 y Q12 (beneficio esperado)
AuxNode2; AuxNode3; AuxNode4 forman un nuevo grupo con AuxNode10.
AuxNode0 variable artificial relacionada con información curso e instructor.
AuxNode1 dificultad del curso
AuxNode5 resume las variables Q13, Q14, Q15, Q17 (preparación docente)
AuxNode6 Habilidades pedagógicas
AuxNode11 Agrupa las variables ocultas de los nodos 0, 1, 5, 6
AuxNode7 cubre las variables Q19, Q20, Q21, Q22, Q25, Q28 (actitud instructor con los estudiantes)
AuxNode8 asociada a Q23, Q24 (participación estudiante)
AuxNode9 asociada a Q26, Q27 (evaluación estudiante)
AuxNode12 Agrupa las variables ocultas de los nodos 7, 8, 9

III.CONCLUSIONES

- Se ha desarrollado un método para aprender grafos cuando hay variables muy relacionadas entre sí, que dependen de distintas variables ocultas que están también relacionadas entre sí.
- De manera general en el problema de evaluación a los estudiantes que se utilizó en este artículo, nuestro método de clúster jerárquico es superior a los otros métodos de estimación con los que se compararon los resultados obtenidos.
- Adicionalmente se puede indicar que se llegó a obtener un modelo mucho más fácil de interpretar y más rápido de calcular en virtud de contar con una forma jerárquica y presentar un tipo de red dónde todas las variables están relacionas entre sí de manera directa y otras variables que también influyen de manera indirecta a través de un camino más largo.

Se pretende en un futuro considerar el problema de deserción como un problema de clasificación no supervisada, adaptando nuestro procedimiento jerárquico para clasificación y comparando el desarrollo de éste con otros modelos.

IV. REFERENCIAS

[1] PETER CHEESEMAN, J. K. (1993). Autoclass: a bayesian classification system. In Readings in know-ledge acquisition and learning. Morgan Kaufmann Publishers Inc.

[2] OVIEDO, B. M. (2016). A hierarchical clustering method: Applications to educational data. Intelligent Data Analysis, 933-951.

[3] BAKER, C. L., TENENBAUM, J. B., & SAXE, R. R. (2007). Goal inference as inverse planning. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 29, No. 29).

[4] MAGAÑA, O. A. (2006). Análisis de la evolución de los resultados obtenidos por los profesores en las

evaluaciones esdeped y las realizadas por los estudiantes. *Revista de la Educación Superior*, 29-48.

[5] RAMÓN GARCÍA MARTÍNEZ, H. K. (2010). Identificación de causales de abandono de estudios universitarios. *TE & ET*.

[6] SCHWARZ, G. (1978). Estimating the dimension of a model. The annals of statistics, 461–464.

[7] DAVID HECKERMAN, C. K. (2007). Leveraging information across hlaallelessupertypes improves epitope prediction. Journal of Computational Biology, 736–746.

[8] HERSKOVITS, G. F. (1992). A bayesian method for the induction of probabilistic net- works from data.

[9] ARTHUR P DEMPSTER, N. M.-B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, 1-38.

[10] ARAUJO, B. S. (2006). Aprendizaje automático: conceptos básicos y avanzados.