Fourth LACCEI International Latin American and Caribbean Conference for Engineering and Technology (LACCET'2006)
"Breaking Frontiers and Barriers in Engineering: Education, Research and Practice"21-23 June 2006, Mayagüez, Puerto Rico

Utilización de la metodología KDSM para el descubrimiento del conocimiento en el ámbito laboral.

Jorge Rodas
Escuela de Ingeniería, ITESM Campus Chihuahua, Chihuahua, Chihuahua, México 31300

Gabriela Alvarado
Escuela de Negocios, Universidad de Barcelona, Barcelona, España 08033

Fernando Vázquez
UPIICSA Dep. Ingeniería, Instituto Politécnico Nacional, Iztacalco, México D.F. 08400

Resumen:

Objetivo: Introducción de la metodología del descubrimiento del conocimiento en medidas seriadas como una forma sencilla y óptima para el análisis de medidas seriadas, repetidas y muy cortas con cierto factor de bloque.

Método: Se utiliza la metodología aplicada al ámbito laboral y se describe siguiendo los pasos de la misma.

Resultados: Nuevo descubrimiento acerca del ámbito laboral y su comportamiento ha sido obtenido mediante KDSM al ser aplicado a este dominio específico.

Conclusión: KDSM es una metodología hibrida (estadística e inteligencia artificial) que da la posibilidad de obtener una solución optima cuando se presenta un problema de conocimiento, especialmente cuando aparentemente no existen atributos relevantes.

Palabras clave: Descubrimiento de conocimiento y Ámbito Laboral.

1 Introducción

Frecuentemente es posible encontrar que mucha información se obtiene mediante la monitorización en cierto tipo de procesos dando como resultado medidas seriadas repetidas durante el tiempo que el proceso está presente. Parece ser que las características de dichas medidas no constituyen un serio problema cuando pudieran ser analizadas mediante técnicas clásicas de series de tiempo, no obstante surge la siguiente pregunta:
¿Que sucede cuando el numero de medidas es extremadamente pequeño?

Frecuentemente en una situación como esta se tiene mucha información adicional acerca de los individuos y eventos que están involucrados en dicho proceso, sin embargo, esta información puede no ser representada mediante medidas seriadas repetidas a pesar de que estén

extremadamente relacionadas con aquello que toma lugar en el proceso. Además dichos individuos frecuentemente constituyen un factor de bloque que actúa sobre las series medidas así las siguientes preguntas surgen: ¿Es posible tomar ventaja de esta información adicional?
¿Como puede ser manejada dicha información en relación a las medidas seriadas repetidas? ¿Si esta información no se obtiene mediante la monitorización de las características de los individuos que forman parte en dicho proceso?

Para responder a estas preguntas la metodología KDSM fue establecida y permite el descubrimiento del conocimiento en dominios donde se presentan medidas seriadas muy cortas y repetidas con factor de bloque compuesto por individuos y donde la información adicional acerca de dichos individuos así como sus atributos que les caracterizan esta disponible.

La metodología KDSM básicamente se desarrolla en tres fases:

1.- **Análisis de Básales de los Individuos (BLA)**.- Inicialmente los básales, que son las primeras medidas seriadas y su relación con los atributos de los individuos son estudiadas para identificar los diferentes perfiles iniciales de dichos individuos los cuales se utilizaran como conocimiento a priori.

2.- **Análisis de los Efectos de los Eventos (EEA)**.- El conocimiento inducido de la fase previa se utiliza como entrada para el estudio del efecto de un cierto evento en un atributo de interés y para descubrir diferentes patrones en las medidas seriadas en relación a los individuos cuando se recibe un evento.
3.- **Producción de Conocimiento (KP)**.- Finalmente los resultados obtenidos de la fase previa se cruzan con los atributos de los individuos y con los eventos respectivamente para encontrar relaciones entre ellos y para determinar que atributos de los individuos y de los eventos son relevantes y así constituir los patrones.

Los fundamentos de la metodología KDSM se encuentran documentados ampliamente en (Rodas 2003) es posible señalar que dicha metodología se utilizo con gran éxito en su aplicación en el ámbito de la psiquiatría (Rodas, Gubert y Rojo, 2001)

En este documento la aplicación de la tecnología KDSM al ámbito laboral se presenta a continuación.

1.1 Estructura del artículo

Este documento tiene la siguiente estructura, sección 2 introducción al dominio en el cual se desarrolla el caso de estudio, sección 3 se formula el problema, sección 4 son presentados los pasos de la metodología KDSM, en la sección 5 se da un pequeño resumen acerca de la aplicación de la metodología KDSM al estudio del ámbito laboral, sección 6 resumen con los resultados obtenidos con la tecnología KDSM, finalmente en la sección 7 se introducen las conclusiones acerca de la aplicación de la metodología así como las miras de trabajo futuro.

2 Estudio del Dominio

El servicio estatal de empleo de cada uno de los gobiernos[1] de los estados de la republica mexicana tiene como objetivo la promoción y supervisión así como el diseño y la aplicación de políticas relativas a la generación de empleo para asegurar que se llevan a cabo dichos objetivos entre otras tareas relevantes se estableció un programa de becas de entrenamiento para trabajadores

desempleados, la metodología KDSM es aplicada para obtener conocimiento de la evolución de dicho programa así como para definir cuan efectivo puede ser dicho programa analizando la información que se reúne analizando la aplicación de este programa de entrenamiento, sus características, así como las características de los municipios en que es aplicado así como los diferentes estados de la republica mexicana incluyendo las necesidades del sector productivo la información procesada por KDSM permite que el servicio estatal de empleo conozca el efecto que cada curso de entrenamiento impartido tiene sobre el municipio en el cual el programa de entrenamiento actúa el servicio podrá establecer o llevar a cabo un rol vital en el desarrollo de dicho programa de entrenamiento de acuerdo con las tendencias globales de los mercados laborales con mayor efectividad y de forma oportuna, el análisis incluye todas las características de los municipios, medidas seriadas acerca de la contratación de las personas desempleadas y todas las características de los recursos, este análisis no es trivial ya que se cuenta con conjuntos de medida para cada tipo de curso siendo esto 180 diferentes dentro del programa de entrenamiento para cada uno de los 2427 municipios que actúan a su vez como factor de bloque sobre dichas medidas seriadas y las características básicas de los cursos.[1]

2.1 Estructura de datos

La descripción genérica del fenómeno modelado se presenta en esta subseccion, la representación de la serie de individuos desde ($i_{1....}i_{n)}$ correspondiente a los municipios en el cual i_n ocurrencias para un evento dado E que son los cursos toma lugar en diferentes momento en l tiempos señalados desde ($E_{i1}...E_{in1}$), ($i_{1...}i_n$) se muestra en la figura 1 conectada para cada ocurrencia del evento existe un atributo de interés Y que es el índice de contratación el cual afecta al comportamiento del individuo. Además tenemos que el objetivo de este trabajo es estudiar el comportamiento de i durante un muy breve periodo de tiempo de $[t_{1...}\ t_r]$ inmediatamente seguido a la ocurrencia del evento E.

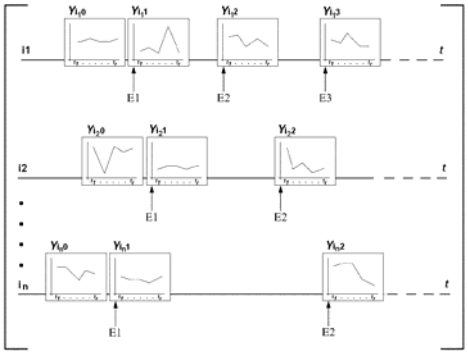

Figura 1.- Y medidas para cada *i* después de cada ocurrencia de *E*

Dado una cierta numero muy pequeño fijo de medidas (r) de Y es tomado para cada individuo y para cada ocurrencia de E, en este caso en particular en los momentos de medición donde Y es medido corresponde a un conjunto de medidas fijo para cada ocurrencia de E por lo cual siempre se empezaran a contar a partir de cada ocurrencia de dicho evento y serán las mismas para cada caso.
Este tipo de estructura invita a cualquier analista de datos a utilizar una técnica especial de metodología que le permita distinguir la información realmente importante para en consecuencia

[1] Correspondiente al gobernador de cada estado.

lograr sus objetivos principales a la hora de hacer la investigación que en este caso es la efectividad del programa de capacitación. Ya que el experto que participa en esta investigación le es familiar la información del estado de Chihuahua como primer paso se determino trabajar con dicho estado, de hecho el estudio de Chihuahua se realiza utilizando únicamente el análisis de un curso que es este caso se trata del campo textil. El servicio estatal de empleo del estado de Chihuahua esta muy interesado en dicho análisis debido a la popularidad que este curso tiene en el sector productivo.

4 Formulación del problema

El problema formal que se intenta resolver se fórmula de la siguiente forma:

Dados:

- $I=\{i_1 \ldots i_n\}$ un conjunto de individuos,
- $X_1 \ldots X_k$ un conjunto de atributos (cuantitativos o cualitativos) que describen a I,
- $X=[x_{ik}]_{nK}$ una matriz, $i=\{1 \ldots n\}$, $k=\{1 \ldots K\backslash\}$, x_{ik} es el valor de un atributo X_K Para el individuo i.
- $E=\{E_{ij}\}$ un conjunto de eventos, donde $i=\{1 \ldots n\}$, $j=\{1 \ldots n_i\}$, y $N=\Sigma^i_1 n_i$, ocurren en los individuos de I siendo $\{E_{ij}\}$ la j-th ocurrencia de E para el i-th individuo.
- $Z_{1 \ldots} Z_k$ un conjunto de atributos (cuantitativos o cualitativos) que describen a E, $Z=[z_{(ij)l}]_{NL}$ una matriz, $i=\{1 \ldots n\}$, $j=\{1 \ldots n_i\}$, $N=\Sigma^i_1 n_i$, $l=\{1 \ldots L\}$, $z_{(ij)l}$ siendo el valor de Z_l para el evento E_{ij}.

- $\{Y^t\}$ es una familia de atributos de interés que conforman las medidas seriadas, donde $i=\{1 \ldots n\}$, $j=\{0 \ldots n_i\}$, y $t=\{1 \ldots r\}$,
- $Y=[Y^t_{ij}]_{N'r}$ la matriz que contiene todas las medidas seriadas, $i=\{1 \ldots n\}$, $j=\{0 \ldots n_i\}$, $t=\{1 \ldots r\}$, y $N'=\Sigma^i_1 n_i + n$

y tomando en cuenta que:

- los individuos $I=\{i_{1 \ldots} i_n\}$ actúan como un factor de bloque par alas matrices Y *y* Z,
- $\{Y^t_{i0}\}$ son las series básales de medidas de los individuos tomadas anterior al primer evento E_{i1} sin que se correspondan con los valores de los atributos de Z.
- la medición de los puntos en tiempo $t=\{1 \ldots r\}$ representan una distribución *fija* en el tiempo para todas las medidas seriadas,
- Los números de observaciones por serie r es *muy pequeño*, y para cada i existe un número *variable* de medidas seriadas n_i

Se desea encontrar conocimiento para las medidas seriadas Y que provea:

- un patrón del conocimiento de las medidas seriadas $\{Y^t\}$ y
- la relación entre las medidas seriadas $\{Y^t\}$, la matriz X y la matriz Z.

Este escenario genera información que puede ser estructura da en tres matrices, los atributos de los municipios, los atributos para los 67 municipios del estado. De chihuahua, matriz X, el índice de contrataron, las medidas seriadas en el numero de personas empleadas en una posición de

trabajo, matriz Y en un periodo fijo de tres meses que es la cantidad que se mide en 6 ocasiones con un periodo de 15 días entre cada medida y las características de los cursos de entrenamiento, matriz Z para cada curso, 180 diferentes cursos donde cada municipio actúa como un paquete de factor de bloque en las medidas seriadas y para las características de los cursos, en otras palabras tenemos mediadas seriadas para cada municipio y otras para los atributos de los cursos.

La matriz X contiene la información acerca de diversos atributos de *I*, por ejemplo en los municipio de chihuahua algunos de estos atributos[2] son: la extensión territorial, la población total por municipio… la matriz Y contiene las mediada seriadas repetidas del atributo de interés que en este caso es el índice de contratación tomado durante un periodo fijo de monitores después de cada curso. El curso se ha ofrecido diversas veces para cada municipio y el índice de contratación que fue registrado para cada curso a los 15, 30, 45, 60, 75 y 90 días después de cada curso, la matriz Z contiene demasiada información valiosa para cada curso, algunos de estos atributos son la especificación de rama económica para cada curso, la actividad económica para la cual el curso se otorga, el lugar de entrenamiento en el cual el curso se imparte , la inversión en los cursos…para mas detalles en estos atributos ver (Rodas 2003).

Tabla 1 muestra la correspondencia entre los elemento formales definidos anteriormente y objetos particulares del dominio objetivo.

I	Conjunto de municipios de Chihuahua
X	Municipios característicos de la matriz
E	Perteneciente a un curso
E_{ij}	j-th curso perteneciente al municipio *i*
Z	cursos característicos de la matriz
Y	(CI)
Y^{t}_{ij}	Contracting Index of Municipality *I* after t days of j-th Course

Tabla 1.- Correspondecia entre la descripción formal y el caso real.

Finalmente es posible obsevar que la información de las 3 matrices introducida en esta sección es incompatible entre ella, (para mas detalle acera de esta incompatibilidad ver Rodas 2003)[2]

4 Metodología KDSM

Los pasos que componen a KDSM están expresados en tres fases y se presentan en esta sección (su justificación se provee en Rodas 2003).

1. Análisis de Básales Individuales (BLA)

1.1 **BLA**: Obtención de la matriz de mediciones básales Y_0.
1.2 **BLA**: Clustering jerárquico de Y_0.
1.3 **BLA**: Interpretación de *PI* de los patrones de $Y_0 \mid C$

[2] Esta información fue otorgada por el Centro de Información Social y Económica del gobierno del estado de Chihuahua.

1.4 **BLA**: Interpretación de *PI* usando la matriz X
1.5 **BLA**: Obtención del conjunto de reglas K B_0

2. Análisis de los Efectos de los Eventos (EEA)

2.1 **EEA**: Obtención de la matriz D
2.2 **EEA**: Clustering inducido por reglas (CIR) de la matriz D

3. Producción de Conocimiento (KP)

3.1 **KP**: Asociación entre P_E y el factor de bloque.
3.2 **KP**: Interpretación de P_E de los patrones de D $|$ P_E
3.3 **KP**: Interpretación de P_E utilizando la matriz X
3.4 **KP**: Interpretación de P_E utilizando la matriz Z
3.5 **KP**: Análisis de la caracterización de $P_{E.}$

5 Aplicando KDSM al Dominio Laboral

Para realizar el análisis de datos o el análisis de la información del Estado de Chihuahua se realizaron las tres fases mencionadas en la sección uno de la metodología KDSM.

1. Análisis de básales individuales (**BLA**): Se caracterizó la estructura conformada por los municipios a partir de su primer índice de contratación esto es, se establecieron las condiciones iniciales.
2. Análisis de los efectos de los eventos (**EEA**): Análisis del efecto de cada curso aislado eliminando el factor de bloque que constituyen cada uno de los municipios.
3. Producción de conocimiento (**KP**): Finalmente los resultados obtenidos en la fase previa fueron cruzados con las características de los municipios y los cursos (matrices X y Z) para encontrar las relaciones entre estos y así poder determinar que atributos son relevantes para encontrar los patrones que nos permitan obtener una descripción de su estructura y dar así una interpretación de lo que sucede.

5.1 Caracterización de la estructura de municipios (BLA)

Después de aplicar los pasos 1.1 y 1.5 de la metodología KDSM (sección 4) se constituye una base de conocimiento conformada por reglas que describen la estructura de los municipios obtenida.

Cuando los individuos, en este caso particular los municipios de dominios que son reales y complejos son descritos por atributos cuantitativos, encontrar atributos que caractericen a las diferentes clases bajo cierta partición en ocasiones resulta una tarea bastante complicada. Así, para poder realizar un análisis adecuado de estos dominios es conveniente trabajar con *atributos que parcialmente los caracterizan* (Vázquez 2002) que son el uso de atributos comunes para todas las clases. Esta situación puede ser representada por un sistema de reglas que maneje cierto grado de membresías para cada clase que establece la partición. La caracterización y el proceso de determinación de calidad de las clases resultantes de una clasificación de referencia no es una tarea trivial ya que no se cuenta con un criterio objetivo para determinar la calidad de las clases (Witten y Frank 1999). Es por esto que se requiere de un punto de vista subjetivo dado por un experto del

dominio y es esta acción o tarea usada frecuentemente para establecer la calidad de la clasificación basada en su utilidad o en el significado de las clases.

De acuerdo a este marco general de la metodología KDSM los siguientes casos se deben desarrollar para obtener *calidad* o *utilidad* de la clasificación.

1. Estadísticas descriptivas que proveen información preliminar acerca de toda la información que deberá ser analizada para ser posteriormente utilizada. Una vez que la clasificación se haya realizado otra descripción de las clases de la información se debe realizar.
2. El conocimiento a *priori* del experto deberá ser incluido para obtener restricciones semánticas (reglas) de la partición resultante y así poder ser empleadas para encontrar la utilidad del significado.
3. Para obtener atributos relevantes estadísticamente significativos una vez que se han caracterizado las clases utilizamos el método de (1) Kruskal – Wallis X^2 de independencia o el método de CIADEC (Vázquez 2002). Es obvio, en este caso que las pruebas estadísticas fallan desde que no se han obtenido atributos significativos estadísticamente hablando. Es así que CIADEC es utilizado para producir una base de conocimiento parcial *FKB* que caracterise a las clases y que provea la base de conocimiento requerida para el siguiente paso.
4. La calidad de las clases es determinada en términos del mismo experto acerca del *significado* y la *utilidad* que las clases obtenidas a partir de las reglas se conformaron en el paso anterior. Si después de un proceso de clasificación se descubre que la estructura no es útil entonces en todo el proceso incluyendo la primera fase BLA de la metodología KDSM deberá repetirse nuevamente para considerar una nueva clasificación con nuevas restricciones semánticas (reglas) pudiendo o no incluir conocimiento adicional del experto o reglas de membresía compuestas por una combinación de atributos que reúnan la idea de una nueva estructura y con un mejor conocimiento acerca del objetivo de estudio de la anterior estructura. Esto no es nada más que aplicar el método científico para investigación en el sentido de que cuando los resultados aparentemente son de baja calidad el modelo se debe formular de nuevo para hacerlo que se ajuste más a la realidad.
5. Los resultados obtenidos y la estructura descubierta de la información se interpretan esto es claramente un nuevo conocimiento, puede ser usado para la toma de decisiones y detener el proceso o continuarlo hasta que la clasificación sea útil para el experto o hasta que el analista de la información decida detener el proceso.

Desafortunadamente en este caso específico el punto tres descrito anteriormente no ofreció información útil, de cualquier forma el experto pudo determinar que algunos atributos son de su especial interés para el objetivo que persigue investigar, así el experto previó de un conjunto de reglas difusas para dar la utilidad a las clases obtenidas. Para conformar este conjunto el experto definió rangos de valores para los atributos en tres etiquetas lingüísticas, pequeñas, medianas y grandes, de acuerdo con su propia experiencia. Los atributos de interés que él tomo en cuanta fueron (a) La Extensión Territorial (TS), pues esta refleja una oportunidad abierta para el sector de negocios de establecerse con nuevas empresas o mantener el mínimo número necesario de empresas existentes. Este atributo es un indicador de cuantas y que clase de cursos son apropiados para ofrecer por el Servicio Estatal de Empleo (SES). (b) Población Económicamente Activa Desempleada (EAUP), este indica el número de gente desempleada.

Así, el conjunto de datos obtenido es el siguiente:

KB_0

Si (x_{iTS} → s ∧ $x_{i\,EAUP}$ → m) → U
Si (x_{iTS} → s ∧ $x_{i\,EAUP}$ → l) → U
Si (x_{iTS} → m ∧ $x_{i\,EAUP}$ → m) → U
Si (x_{iTS} → m ∧ $x_{i\,EAUP}$ → l) → U
Si (x_{iTS} → l ∧ $x_{i\,EAUP}$ → m) → U
Si (x_{iTS} → l ∧ $x_{i\,EAUP}$ → l) → U

donde x_{iTS} es el valor para el atributo de Extensión Territorial (TS) para el *i-th* municipio y x_{iEAUP} es el valor para el atributo EAUP de la población económicamente activa para el *i-th* municipio.

Una vez que el analista de datos ha obtenido el conjunto de reglas el experto ha valorado su nivel de representación de estructura y determinado cuan útil es para su objetivo de investigación finalmente las reglas se usan como sesgo para la siguiente fase de la metodología KDSM (mayores detalles se pueden encontrar en Rodas, Alvarado y Vázquez 2002).

5.2 Análisis del efecto de cada curso aislado (EEA)

La segunda fase de la metodología KDSM consiste de los pasos 2.1 y 2.2 y se corresponde al análisis del efecto de cada curso aislado utilizando las *diferencias* para eliminar el factor de bloque constituido por los municipios sobre los cursos.

En la figura 2 el dendrograma nos muestra la clasificación CIR sobre los efectos de los cursos basados en el índice de contratación (CI). El dendrograma sugiere un corte en 13 clases y el experto estuvo de acuerdo en ello ya que se ve como el corte más apropiado, dos de estas clases (U y D) son inducidas por medio de las reglas y resultan interesantes y útiles desde el punto de vista del experto para lograr sus objetivos de investigación. Los resultados de esto fueron importantes y novedosos y se mostrarán posteriormente. Las clases residuales fueron 11 y se han convertido en solo una ya que no representan mayor interés para el experto. Además estas clases representan únicamente el 20% de los datos que para este particular estudio son irrelevantes.

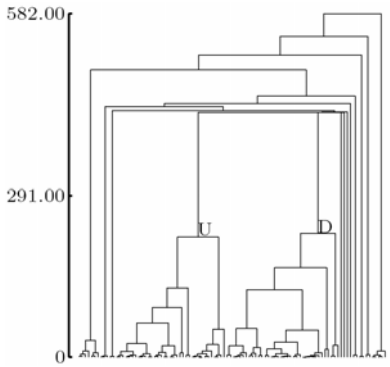

Figura 2. Estructura de cursos

5.3 Identificación de características relevantes (KP)

La tercera y última fase de la metodología KDSM consistentes en los pasos 3.1 y 3.5 corresponden a la búsqueda de atributos relevantes en los municipios y cursos en relación a la partición obtenida en la última fase.

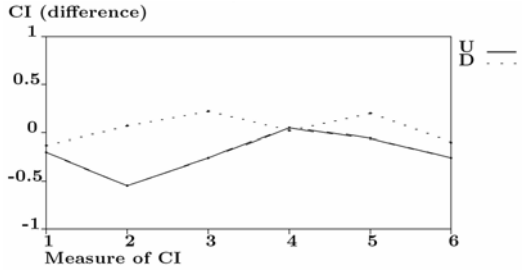

Figura 3. Representación de curvas de 2 clases.

La figura 3 representa la tendencia general del efecto del curso en los índices de contratación para cada clase. Cada curva representa una clase y esta se obtiene a partir del promedio de todas las curvas que existen en la clase. En esta representación la variabilidad dentro de las clases no se muestra porque esta en realidad no es evaluada. Sin embargo esta clase de gráfico nos da una idea intuitiva que los efectos de los cursos en el Índice de Contratación siguen en estas clases. Así, el experto puede analizar el comportamiento general del Índice de Contratación en las diferentes clases para poder encontrar una relación entre ellos.

En la figura 3 la clase U (línea seguida) muestra valores negativos con una única excepción que es un mes después del curso. Esto significa que el número de contratación se incrementó en esta medición específica. La clase D (línea de puntos) muestra valores positivos. Además estos valores son discretos, indican que en general la contratación es estable durante el tiempo de monitorización por el programa de capacitación.

Por otra parte se obtiene una clara descripción de estas dos clases que tienen una semántica obvia para el dominio del experto.

- U. Cursos que producen contratación en un punto específico en el tiempo, 30 días después del curso.
- D. Cursos que mantienen una contratación estable entre el curso actual y el siguiente.

La tendencia de la clase U (línea seguida), de la figura 3 revela un restablecimiento lento hacia la cuarta medición (60 días) del Índice de Contratación. Esta tendencia esta engranada muy de cerca con la modalidad de curso (ver figura 4), el cual muestra que la mayoría de estos cursos son de *autoempleo.* Esta es la razón por la cual los participantes de este curso cuando se gradúan requieren de mayor tiempo para encontrar un trabajo o para establecer su propio negocio.

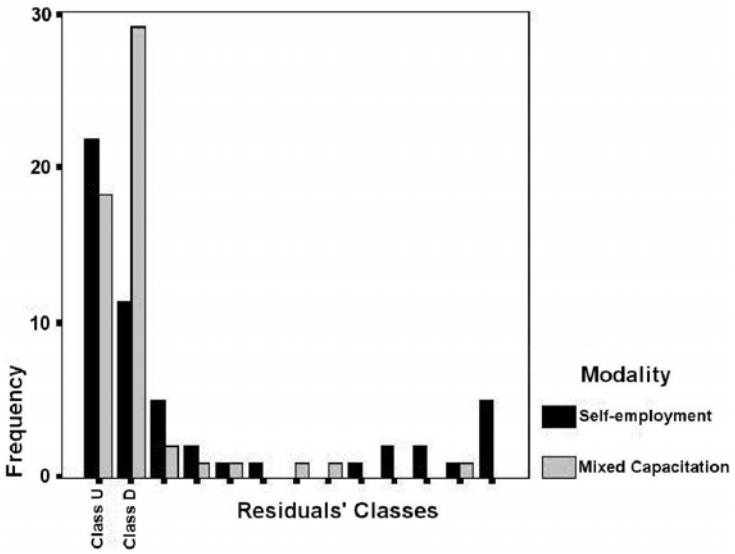

Figura 4. Modalidad de cursos por clase.

En el caso de los cursos de *capacitación mixta*, las compañías privadas que participan en el programa de entrenamiento, tiene un compromiso del 90% de contratación de las personas entrenadas. Esto es establecido por un acuerdo entre el Gobierno del estado de Chihuahua y las compañías participantes. En la clase D (línea de puntos, en la figura 3) muestra las diferencias que son positivas desde el punto de vista del Índice de Contratación que está directamente relacionado con la modalidad del curso y mayormente con esos cursos denominados de *capacitación mixta* (ver figura 4). En otras palabras la mayoría de los participantes en esta clase de cursos cuando se gradúan tienen un trabajo en alguna de las empresas que participan en este proceso de capacitación mixta.

Finalmente queda clara la diferencia de tendencias entre ambas clases, denotada por la modalidad del curso, donde cada curso tiene un efecto particular el cual promete un análisis de los atributos respecto a su especial influencia en el comportamiento descubierto. El Servicio Estatal de Empleo (SES) así, puede realizar una planeación a futuro con base en esta información.

6 Resultados

La metodología KDSM permite un nuevo conocimiento al ser descubierto cuando se presenta medidas seriadas muy cortas y repetidas en un conjunto de individuos y su aplicación al ámbito laboral ha sido satisfactoria y ha presentado resultados importantes.

En este estudio tres bases de datos del ámbito laboral se han utilizado. El primero es relativo a información de los municipios (matriz X), la segunda es el resultado de las medidas de los índices de contratación tomadas después de cada curso aplicado (matriz Y), y la tercera contiene información sobre los cursos (matriz Z). En un esfuerzo para mejorar este estudio y verificar la utilidad de la metodología KDSM como herramienta de ayuda para el análisis de esta tipo de programas se valoró la creación de un primer programa de aplicación y utilizar únicamente un solo tipo de curso (manufactura textil) en 17 municipios del estado de Chihuahua.

Es conveniente señalar que con estos resultados positivos posteriormente se aplicará para los 180 diferentes tipos de cursos en los 2427 municipios de toda la República Mexicana.

Un análisis de tres fases fue desarrollado con la metodología KDSM con los siguientes datos:

1.- *Caracterización de los municipios* (BLA). El primer índice de contratación se analizó debido a que representa las condiciones iniciales de cada municipio, esto es antes de iniciar el programa de entrenamiento, el objetivo antes de iniciar esta fase es encontrar información útil para descubrir los posibles perfiles de cada municipio antes de que el primer curso se de en este mismo y para tratar de encontrar cualquier estructura *a priori* que permite determinar las diferencias en los efectos del efectos del programa de entrenamiento. Nueve reglas (ver expresión 1, sección 5) fueron derivadas de dicho análisis de los grupos de los municipios, entre ellos tenemos diferentes extensiones territoriales e indicadores de población económicamente activa desempleada que los delimita.
2.- *Análisis del efecto de cada curso aislado* (EEA).- El clustering inducido por reglas fue desarrollado utilizando los resultados del análisis previo como una base de conocimiento. Las diferencias entre los indicadores de contratación fueron utilizadas para eliminar el *efecto de municipio* (características del municipio) y para estudiar el *efecto del curso* de forma aislada. Este estudio mostró la evolución de los municipios después de cada curso durante el periodo del programa de entrenamiento. En esta presente aplicación se puede observar que el clustering inducido por reglas mejora los resultados de interpretación, además cualquier técnica simple de clustering nunca podría haber incorporado reglas de membresía como criterio de clustering en el análisis del índice de contratación ya que los atributos utilizados por esta regla no están presentes en la matriz Y.
Finalmente, dos clases fueron detectadas (ver figura 3). Por una parte la clase U (línea seguida) que contiene la mayoría de los cursos de modalidad de *autoempleo* y por otra parte la clase D (línea a puntos) que nos muestra que la mayoría de las diferencias son positivas y por lo cual la mayoría de estos cursos pertenecen a la modalidad de *capacitación mixta.*
3.- Identificación de características relevantes (KP). Finalmente todos los atributos de la matriz X y Z se proyectaron en las dos clases obtenidas en la fase tres. Algunos atributos relevantes fueron identificados usando esta proyección. Así se puede mostrar claramente una tendencia entre ambas clases denotada principalmente por el atributo de modalidad de curso, donde cada curso tiene un efecto particular que determina que el análisis de los atributos tiene una influencia especial en el comportamiento descubierto. Con base a estos resultados el Servicio Estatal de Empleo podrá realizar una planeación futura para el programa de entrenamiento y como resultado ahorrar en cuanto a inversiones se refiere.

7 Conclusiones

Ahí que tener en cuenta que una de las funciones del Servicio Estatal de empleo es dar seguimiento a los resultados obtenidos con el programa de entrenamiento por medio de la medición de la eficacia mediante una relación de costo-beneficio, esto es, la inversión económica con respecto a la gente que se ha graduado del programa de entrenamiento dentro del sector laboral. Por esta razón la metodología KDSM se aplicó obteniendo importantes resultados donde se pueden remarcar las siguientes conclusiones:

La inversión económica del gobierno es pequeña cuando la modalidad del curso es de *capacitación mixta* debido a la participación directa de las empresas. Así el SES puede optimar el programa de entrenamiento encontrando la mejor oportunidad y nicho donde en coordinación empresa y gobierno puedan realizar inversión y capacitación.

Así, si la modalidad del curso, las características de municipio, como su relación con la realizar una planeación de capacitación óptima determinando cuales de las modalidades de cursos diferentes que existen en su programa de capacitación pueden atender mejor las necesidades económicas de cada municipio así como del sector empresarial, maximizando la relación de costo beneficio entre todos los sectores involucrados.

Finalmente el SES en base a todo lo anterior podría estar habilitado para enfrentarse a una situación para la cual el mercado no pueda contratar a toda la gente que se ha graduado del programa de entrenamiento y el balance de la relación costo-beneficio se rompe. Para finalizar el SES puede utilizar este conocimiento provisto por los atributos relativos a la población del total número de matriculaciones de cada curso específico y la gente graduada así como la contratada.

En suma las conclusiones obtenidas por la metodología KDSM aplicadas al *dominio laboral* son las siguientes:

Soporte al proceso de toma de desiciones.

KDSM lleva a cabo un proceso de retroalimentación mediante el conocimiento adquirido del programa de conocimiento de forma que el SES pueda actuar en concordancia.

Asistencia en la optimización y tareas de planeación.

KDSM fundamenta las desiciones sobre el tipo de curso y otras características que el SES deberá seguir para mejorar el desempeño del programa de entrenamiento de forma constante y permanente.

Proveer de nuevo conocimiento.

KDSM provee al SES con conocimiento obtenido acerca de la monitorización de los cursos para que este pueda tener un buen soporte cuando quiera enfrentar la necesidad en una situación de no balance en la relación de costo-beneficio.

Finalmente, dos conclusiones acerca de la metodología KDSM podemos obtener de este trabajo de investigación:

1. La metodología KDSM integra la estadística y la inteligencia artificial para dar una posible solución a un problema cuando no existen atributos relevantes.
2. El uso de reglas difusas constituyen una combinación de atributos a partir de la experiencia del experto permitiendo realizar una partición útil que se puede obtener del objetivo de estudio del experto.

8 Trabajo futuro

Este trabajo representa una primera etapa de la aplicación de la metodología KDSM en el dominio laboral, en el cual varias tareas se han identificado para ser desarrolladas.

- Formalizar un mecanismo para mejora del proceso de adquisición de la calidad de clase en términos de la utilidad.
- Definir el mecanismo de obtención de reglas cuando ninguno de los atributos caracteriza completamente a las clases y además no existe ninguno estadísticamente relevante y
- Establecer un sistema de caracterización para descripciones conceptúales de clase a partir de una partición de referencia así como un modelo de generación automática de interpretación de clases a partir de la partición de referencia.

Agradecimientos

Agradecemos la cooperación mostrada por la ESIC del Gobierno del estado de chihuahua por proveer toda la información para este estudio.

Referencias

J. Rodas (2003). Knowledge Discovery in repeated and very short serial measures with a blocking factor. PhD. thesis of Technical University of Catalonia. Barcelona, Spain.

J. Rodas, G. Alvarado, and F. Vázquez (2002). Applying KDSM to an specific domain where very short and repeated serial measures with a blocking factor are present. In a research report of Technical University of Catalonia (LSI-02-54-R). Barcelona, Spain.

J. Rodas, K. Gibert, and J.E. Rojo (2001). Electroshock Effects Identification Using Classification Techniques. In Crespo, Maojo and Martin (Eds.), *Springer's Lecture Notes of Computer Science Series*, 238--244. Madrid, Spain.

F. Vázquez (2002). Automatic Characterization and Interpretation of Conceptual Descriptions in Ill-Structured Domains using Numerical Attribute. In a research report of Technical University of Catalonia (LSI-02-51-R). Barcelona, Spain.

I.H. Witten, and E. Frank (1999). Data Mining.Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers.