MODELO CLUSTERING PARA EL ANÁLISIS EN LA EJECUCIÓN DE PROCESOS DE NEGOCIO

Surelys Pérez Jiménez[1], Joan Jaime Puldón[2], Rafael A. Espín Andrade[3]
[1]Departamento IngenieríaIndustrial, Instituto Superior Politécnico José Antonio Echeverría, CUJAE
[2]FacultadIngenieríaInformática, Instituto Superior Politécnico José Antonio Echeverría, CUJAE
[3]Centro de Estudios de Técnicas de Dirección (CETDIR), Instituto Superior Politécnico José Antonio Echeverría.

ABSTRACT
The work is based on the proposed clustering model to support decision making based on analysis of records in the execution of processes in organizations that implement BPM technologies. An investigation is conducted that includes the study of how to integrate a model based on a data mining algorithm to business processes and how, from the results of the implementation of processes, can generate knowledge to improve performance indicators in their implementation.
The proposal includes the creation of a tool based on the implementation of the clustering technique of data mining to find useful knowledge. It is developed using open source technologies.

KEY WORDS: data minning,process mining, minningalgorithms, clustering, BPM.

MSC: 91C20

RESUMEN
El trabajo se fundamenta sobre la propuesta de un modelo de agrupamiento que apoye la toma de decisiones a partir del análisis de registros en las ejecuciones de procesos en organizaciones que implementen tecnologías BPM. Se realiza una investigación que incluye el estudio de cómo integrar un modelo basado en un algoritmo de minería de datos a los procesos de negocio y cómo, a partir de los resultados de la ejecución de los procesos, se puede generar conocimiento a fin de mejorar los indicadores de eficacia en la ejecución de los mismos.
La propuesta incluye la creación de una herramienta basada en la aplicación de la técnica de clustering de minería de datos, para encontrar conocimiento útil. Se desarrolla utilizando tecnologías de software libre.

1. INTRODUCCIÓN

El volumen de datos que se genera a diario ha permitido satisfacer las necesidades crecientes de las organizaciones, pero ha superado las capacidades humanas para analizar y transformar la información en conocimiento útil que apoye la toma de decisiones. Es conocido que los modelos matemáticos basados en principios estadísticos, se han utilizado siempre en la solución de los más variados problemas de las ciencias, con carácter empírico o teórico. Fenómenos de naturaleza inorgánica o inanimada regidos por leyes de la mecánica, de la física, o de la química, han resultado asimilables por modelos matemáticos. Esto, unido a la necesidad de perfeccionamiento diario, que sólo puede lograrse mediante la retroalimentación y el profundo conocimiento del área en que se desea mejorar, ha motivado el empleo de técnicas y herramientas que posibiliten la extracción de conocimiento a partir de datos.

La posibilidad de procesar y extraer información relevante, descubrir conocimientos y patrones en bases de datos, hacen ver a la Minería de Datos (DM, Data Mining, según sus siglas en inglés) como soporte a la toma de decisiones empresariales. El término Minería de Datos es relativamente moderno e integra varias técnicas de análisis de datos y construcción de modelos, y permite identificar patrones, describir tendencias y regularidades, predecir comportamientos y, en general, sacar partido a la información digitalizada, ver Rosete (2004), Brito (2008) y Martin (2008).

Con el fin de automatizar y administrar los procesos de negocio, muchas empresas utilizan la metodología de Gestión de Procesos de Negocio, BPM (Business Process Management, BPM según sus siglas en inglés). Los sistemas de software que implementan esta metodología son los BPMS, los cuales, en su primera etapa, modelan y simulan los procesos del negocio, en una segunda los automatizan y en una tercera emplean herramientas de

[1]sperezj@ind.cujae.edu.cu,
[2]sperezj@ind.cujae.edu.cu,
[3]sperezj@ind.cujae.edu.cu,

inteligencia de negocio para monitorear y controlar los procesos, ver Díaz (2007).
La etapa de modelado de procesos la ejecuta y gestiona un analista de proceso a partir de la información asociada al funcionamiento de la organización. Sin embargo la integridad de estos modelos en ocasiones es cuestionable.En los BPMS no se incluye de forma gratuita la posibilidad de extraer conocimiento útil de la ejecución de los modelos representados, con la que se pueda reformular el proceso y que sirva como guía a los gerentes en la toma de decisiones, trayendo consigo un mayor costo para las empresas, pues implicaría una nueva inversión.Con la idea básica de descubrir, monitorizar y mejorar los procesos reales, extrayendo conocimiento de los registros de eventos, interviene la minería de procesos.Sobre este entorno se ha visto la necesidad de requerir aplicaciones informáticas que permitan gestionar datos, dedicadas a la extracción de conocimiento sobre los registros de eventos de los procesos de negocio, basadas en técnicas y algoritmos de la minería de datos, la aplicación de estas técnicas es el trabajo de Pérez (2009)

2. MINERÍA DE DATOS (MD)

En dependencia del tipo de búsqueda empleada para obtener conocimiento, se clasifica la Minería de Datos en Directas (MDD) o Indirectas (MDI). En la MDD se conoce claramente lo que se busca, el objetivo, generalmente predecir unos ciertos datos o clases, siendo las tareas de clasificación y predicción las que pertenecen a este grupo. El agrupamiento, la asociación y la correlación se incluyen dentro de la MDI, y se emplean para descubrir patrones que describan los datos sin un objetivo concreto definido, ver Marcano y Talavera (2007).

En la actualidad existe un número amplio de herramientas de apoyo al análisis de datos durante un proceso de MDD. Una de las líderes entre las de libre distribución se encuentra WEKA (WaikatoEnviromentforKnowledgeAnalysis), según una entrevista realizada por KDnuggets.Constituye un entorno de experimentación de análisis de información, formado por una serie de paquetes de código abierto con diferentes técnicas de preprocesado, clasificación, regresión, asociación, y visualización de datos, ver Dunham (2003). Entre sus principales características se encuentra el poseer una interfaz gráfica de usuario compuesta de cuatro entornos (Comand-Line Interfaz (Simple CLI), Explorer, Experimenter y KnowledgeFlow) que permiten diferentes funcionalidades y formas de análisis. Dado que se trata de una herramienta bajo licencia GNU, es posible actualizar su código fuente para incorporar nuevas utilidades o modificar las ya existentes, de ahí la presencia de proyectos asociados a WEKA que permiten garantizar sucontinua evolución. La aplicación de algunas de estas herramientas se muestra en el trabajo de Molina y Garcia (2006), Brito (2008) y Martin (2008).

Dado que la MD es un campo interdisciplinar, no existe un método universal para todo tipo de aplicación. Algunas tareas pueden ser resueltas por muy diversas técnicas y, algunas técnicas pueden aplicarse para varias tareas. Cada técnica incluye diferentes algoritmos y variaciones de los mismos. Las técnicas constituyen el enfoque conceptual para extraer conocimiento y obtener modelos a partir de la información recopilada. Son muchas las técnicas utilizadas en la MD por lo que se clasifican en dos grandes categorías: supervisadas o predictivas y no supervisadas o descriptivas, ver Weijters (2009). Afortunadamente, las técnicas más diversas se encuentran implementadas en entornos software que se pueden encontrar a la venta o distribuidos de forma gratuita.

La minería de datos es relevante para la industria porque permite un análisis objetivo de los procesos basado en sus ejecuciones actuales. En este sentido, las técnicas de minería están enfocadas más en el análisis de las instancias de procesos ejecutadas. El análisis proveído por estas técnicas pueden ser utilizadas para detectar puntos de optimización para los procesos de negocios actuales. Este enfoque de la minería de datos a los procesos de negocio, se conoce como Minería de Procesos, partiendo del enfoque dado en Alves (2008).

3. MINERÍA DE PROCESOS

La minería de procesos es el “data mining” pero con una visualización fuerte de proceso de empresa. Incluso algunas de las técnicas de minería de datos más tradicionales pueden ser usadas en el contexto de la minería de proceso.Desde el punto de vista más pragmático y asociándolo directamente a las actividades de negocios, la minería de procesos es el conjunto de metodologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información no estructurada (interna y externa a la compañía) en información estructurada, para su explotación directa o para su análisis y conversión en conocimiento y así dar soporte a la toma de decisiones sobre el negocio, ver Espen y Atle (2008).

Los datos tal cual se almacenan en las bases de datos no suelen proporcionar beneficios directos. Su valor real reside en la información que podamos extraer de ellos, es decir, información que nos ayude a tomar decisiones o a

mejorar la comprensión de los fenómenos que nos rodean. Ejemplos de ello pueden ser: contrastar que todo va bien, analizar diferentes aspectos de la evolución de la empresa, presentar información de formas más intuitiva, comparar información en diferentes períodos de tiempo, comparar resultados con previsiones, identificar comportamientos y evoluciones excepcionales, confirmar o descubrir tendencias e interrelaciones, entre otras acciones.

La posibilidad de elevar los niveles de competencia de los negocios, basándose en la rapidez para identificar, procesar y extraer la información que realmente es importante, descubriendo conocimientos y patrones en bases de datos, hacen ver a la Minería de Procesos como soporte a la toma de decisiones empresariales.

La idea básica de la minería de proceso es descubrir, monitorear y mejorar procesos reales, extrayendo conocimientos de "logs" de evento. Actualmente muchas de las tareas que ocurren en procesos son soportadas o monitoreadas por sistemas de información. Sin embargo, la minería de proceso no está limitada a sistemas de información y puede también ser usada para monitorear otros sistemas o procesos operacionales (por ejemplo, servicios Web, flujos de personal en hospitales, etc). Todas estas aplicaciones tienen en común una noción de un proceso y que los acontecimientos de las tareas son grabados en "logs" de evento, ver Alves et al (2008).

Lo aprendido de las ejecuciones observadas de un proceso puede ser usado para:

1. Descubrir nuevos modelos (por ejemplo, formular una red Petri que puede reproducir el comportamiento observado).
2. Verificar la conformidad de un modelo chequeando si el comportamiento modeladose combina con el comportamiento observado.
3. Prolongar un modelo existente proyectando la información extraída de los "logs" en algún modelo inicial (por ejemplo, mostrar los obstáculos en un modelo de proceso analizando el "log" de evento).

Si se trata de construir un modelo para casos muy diferentes, entonces el modelo es probablemente demasiado complicado. Las técnicas de minería de procesos que existen, son incapaces de detectar este problema y hacerle frente. Por tanto, son creados modelos muy generalizados, donde se incluyen comportamientos no deseados en el proceso, elementos caracterizados en Alves (2008).

Las técnicas clásicas en el campo de la MD y las llamadas herramientas de Inteligencia de Negocios usadas en la industria, apuntan al descubrimiento de conocimiento, medición de desempeño, y predicción, sin ofrecer la funcionalidad de Minería de Procesos. Es por ello que el objetivo del trabajo está enfocado a la utilización de la técnica de agrupamiento de la MD para el preprocesamiento de los registros de eventos, como un paso inicial de la minería de procesos.

4. LA TÉCNICA Y ALGORITMO DE MD APLICADA

Cuando una aplicación no es lo suficientemente madura, no tiene el potencial necesario para una solución predictiva. En ese caso hay que recurrir a las categorías descriptivas o no supervisadas, como el agrupamiento y la asociación, que descubren patrones y tendencias en los datos. El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficio o conocimiento de ellas.En las tareas descriptivas, el conjunto de observaciones no tienen clases asociadas. El objetivo es derivar características (correlaciones, clúster, trayectorias, anomalías) que describan las relaciones entre los datos. Estas tareas son comúnmente de exploración natural y frecuentemente requieren de técnicas de postprocesamiento para explicar los resultados, ver Rosete (2004).

La formación de grupos es muy utilizada como paso analítico en la minería, ya que permite analizar las variables en cada uno de los grupos. Una vez obtenidos los grupos, si se desea predecir, lo que se realiza es determinar a qué grupo pertenece con mayor probabilidad la característica que buscamos, y en función del grupo, se realiza una determinada acción sobre el mismo, ver Hernandez y Ferri (2004).

Técnicas y Algoritmos de Agrupamiento.

Existe, en la literatura, una gran cantidad de técnicas de clustering que varían de acuerdo a la arquitectura que utilizan. Una clasificación general divide los algoritmos en: agrupamiento particional, agrupamiento jerárquico, agrupamiento basado en densidad y agrupamiento basado en rejas. Para cada una de las categorías existen una variedad de sub-clasificaciones que presentan algoritmos con diferentes técnicas para encontrar clúster en los datos, ver Huang (2006).

- **Agrupamiento jerárquico**. Crea una descomposición jerárquica de un conjunto de datos, formando un dendograma, que divide recursivamente el conjunto de datos en conjuntos cada vez más pequeños. El dendograma puede ser cortado en diferentes niveles para producir los clúster deseados. Es necesaria una condición de parada.
- **Agrupamiento basado en densidad**. Se obtienen clúster basados en regiones densas de objetos en el espacio de datos que están separados por regiones de baja densidad (los elementos aislados representan ruido).
- **Agrupamiento basado en rejas**. Cuantifica el espacio en un número finito de celdas y aplica operaciones sobre dicho espacio. Recientemente los algoritmos basados en este agrupamiento han sido presentados para datos espaciales.
- **Agrupamiento de particiones**. Organiza los objetos dentro de k grupos de tal forma que sea minimizada la desviación total de cada objeto desde el centro de su grupo o desde una distribución de grupos. El problema que se presenta al utilizar algoritmos particionales es la decisión del número deseado de clúster de salida. El algoritmo de *K-medias* es comúnmente utilizado en los algoritmos de partición.

El agrupamiento de particiones es el más adecuado para utilizar en el trabajo, pues se formarían grupos con los casos o eventos más semejantes. Resolver la dificultad que tiene el algoritmo, correspondiente a la cantidad de grupos a formar, queda en manos del usuario encargado de realizar la tarea, el cual debe conocer los datos con los que se trabajará y las variaciones de los mismos. Como la MD permite la combinación de varias técnicas y algoritmos, se puede tener en cuenta, para resolver este problema, el empleo de técnicas estadísticas, donde el algoritmo Máximo Valor Esperado (ExpectationMaximization, EM por sus siglas en inglés) es una acertada solución, ver Hernández (2006) y Garre (2007).

5. BUSINESS PROCESS MANAGEMENT: BPM. GESTIÓN DE PROCESOS DE NEGOCIO

Desde sus inicios, los sistemas de software han sido usados para automatizar los Procesos de Negocio (PN). *"El proceso de negocio es un elemento primordial e intangible que está presente en todas las organizaciones"*, y que *"…consiste en la distribución lógicamente interrelacionada de tareas desarrolladas en tiempo y espacio (con comienzo y fin, con entradas y salidas definidas) y que se orienta al logro de un objetivo de negocio, generando un resultado de valor (total o parcial) para el cliente del proceso"*.

El concepto de PN y su importancia no es algo nuevo. Hace muchos años se manejan estos términos, sin embargo, hasta hace muy poco no existían las herramientas de software necesarias para modelar, implementar y controlar los PNs. La gestión de procesos de negocio comprende un conjunto de tecnologías orientadas a dar soporte a la ejecución de la lógica de negocio de organizaciones, ver Tabares, Pinera y Barrera (2008).

BPM

La metodología de Gestión de Procesos de Negocio(BPM, según sigla en inglés) se concibe para la gestión de los flujos de trabajo. Es una variedad de técnicas, actividades, tareas y tecnologías bajo un enfoque metodológico, con el fin de gestionar los procesos de negocio.

Se puede pensar en BPM como *"un conjunto de disciplinas y tecnologías aplicadas para modelar, automatizar, integrar, gestionar y optimizar los procesos, reglas, servicios, y recursos empresariales, e incrementar así la calidad de los servicios, la eficiencia de la organización, y la rentabilidad del negocio"*. BPM permite a los analistas de negocio modelar el ciclo completo de un PN (tanto los pasos manuales como los automáticos), consolidar procesos que abarcan aplicaciones existentes y desplegar procesos de principio a fin con un bajo o nulo esfuerzo en tareas de programación de software. Técnica utilizada en trabajo de Pérez (2009).

Las soluciones BPM están conformadas, por tres etapas relacionadas a su entorno de modelado y simulación, automatización- administración y monitización sobre los eventos ejecutados, retroalimentando para una mejora continua. La relación entre ellas se muestra en la Figura 1.

La implementación o automatización de estos modelos, correspondiente a una segunda etapa, se realiza a través de un motor de ejecución BPM o workflow, definiendo formularios electrónicos, integraciones con otros sistemas, y reglas o validaciones del proceso. De igual modo, la última etapa, utiliza datos de las actividades que ocurren en el sistema para la optimización del proceso, ver Sordi (2007) y Leganza (2008). En el mercado existen varias herramientas de software que dan soporte a estas tareas y que reciben el nombre de Sistemas de Gestión de Procesos Empresariales (BPMS por sus siglas en inglés).

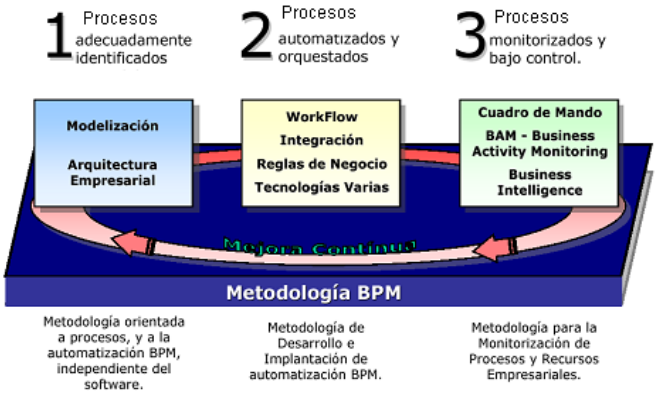

Figura 1. Representación de las etapas de trabajo con la metodología BPM.

BPMS

Los BPMS (Business Process Management Systems), representan una moderna tecnología para la gestión empresarial. Su concepción integral y adaptable por definición a los procesos específicos de la empresa. Los BPMS están diseñados para manejar procesos, instancias, versiones y variantes, componentes, reglas y participantes de procesos. En efecto, un BPMS separa el PN de la gestión del software, permitiendo una reconfiguración rápida del proceso si fuese necesario.

La Figura 2 representa la posible arquitectura de un BPMS. Muestra los elementos claves, descritos anteriormente, y varios elementos de soporte.

Embebido en los Sistemas de Gestión de Procesos se encuentra el Motor BPM o servidor BPM, responsable de ejecutar, controlar y monitorizar todos los procesos de negocio. Denominado genéricamente como motor BPM, realiza la orquestación de los eventos a través de múltiples procesos y maneja las interacciones de los empleados o usuarios, enruta el trabajo a los mismos y asegura que se complete, gestionando el estado de cada tarea. También se encarga de la coordinación de aplicaciones externas dentro del proceso y de la manipulación de datos relacionados con los procesos. Estos datos son normalmente almacenados en un Sistema Gestor de Base de Datos (SGBD). Llámense, a partir de este momento, "tabla registro de proceso" a las tablas creadas por el motor de orquestación en el SGBD. Caracterizaciones descritas en el trabajo de Hinojo (2008).

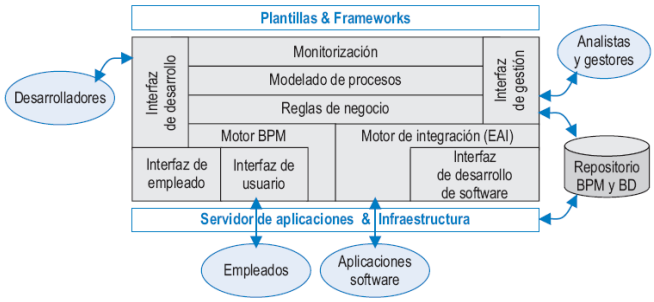

Figura 2. Arquitectura de un BPMS.

En la segunda etapa de BPM, el diseño delproceso se exporta hacia un motorde ejecución, el cual gestiona el enrutamiento y la ejecución de los procesos. El motor de ejecución gestiona el proceso de negocio de principio a fin (tanto los pasos manuales como los automáticos) y proporciona una auditoría de la ejecución, proveyendo la información requerida para el análisis y comportamiento del proceso, monitorización de la actividad de negocio (BAM) y la gestión del rendimiento de los procesos, a través de los monitores de proceso.

6. Descripción de la Solución propuesta

La necesidad de analizar las ejecuciones de los procesos de negocio, con el objetivo de extraer conocimiento útil, hace el surgimiento y desarrollo de la herramienta software ClusDataPro. El nombre se debe a la combinación de

"clustering", "data" y "process", dada la idea final como *"Agrupamiento de Datos de Procesos"*. La herramienta se alimenta de las librerías de WEKA para aplicar la técnica de agrupamiento de la minería de datos, a través de los algoritmos *K-medias* y/o *EM*, a datos de la ejecución de los procesos de negocio.

6.1. Descripción y aplicación de los algoritmos de clustering

K-medias

El nombre viene dado porque representa cada uno de los grupos por la media (o media ponderada) de sus puntos, es decir, por sus "centroides". La representación mediante "centroides" brinda un sentido estadístico inmediato, donde se puede conocer una aproximación de las características por grupos.

El algoritmo o criterio resulta más eficiente aplicándolo a atributos numéricos. Cuando esto ocurre, una medida de distancia muy común es la distancia Euclídea, definida en la ecuación 1.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i} \quad (1)$$

No solo se analizan atributos numéricos, también pueden existir instancias formadas por cadenas de caracteres. En este caso, el campo es introducido para el agrupamiento como un atributo nominal, donde las cadenas completas constituyen los posibles valores. El concepto distancia también se puede utilizar cuando los casos están formados por atributos nominales. Para estos se emplea la función delta, donde $\delta(a,b) = 0$ sí y sólo si $a = b$ y $\delta(a,b) = 1$ en caso contrario. Queda definida la distancia para atributos nominales como:

$$d(x,y) = \omega \sum_{i=1}^{n} \delta(x_i,) \quad (2)$$

donde ωes un factor de reducción. El factor se elige convenientemente cuando existen atributos nominales y numéricos. Se realiza el cálculo de las distancias para los subconjuntos de valores por separado y se combinan utilizando un factor adecuado.

Introducidos los atributos para el agrupamiento y la cantidad de grupos a formar (parámetro k) continua el siguiente procedimiento:

1. Se calcula, para cada instancia x_j, el prototipo más próximo A_gy se incluye en la lista de ejemplos de dicho prototipo

$$A_g = \text{argmin}_{A_i}\{d(x_j,A_i)\}, \quad \forall i = 1,\ldots \quad (3)$$

Después de haber introducido todos los ejemplos, cada prototipo $A_p$tendrá un conjunto de instancias a los que representa.

$$l(A_p) = \{x_{p_1}, x_{p_2}, .., x_{p_m}\}$$

2. Se desplaza el prototipo hacia el centro de masas de su conjunto de instancias.

$$A_p = \frac{\sum_{i=1}^{m}}{m} \quad (4)$$

3. Se repite el procedimiento hasta que ya no se desplacen los prototipos.

Máximo Valor Esperado

El algoritmo *EM*, brinda una aproximación estadística al problema, busca el número de clúster más probable dados los datos. La base de este tipo de agrupamiento se encuentra en un modelo estadístico llamado mezcla de distribuciones, donde cada distribución representa la probabilidad de queun objeto tenga un conjunto particular de pares atributo-valor. Es utilizado el método Gaussiano finito de mezclas, asumiendo que todos los atributos son variables aleatorias independientes.

El procedimiento seguido por el algoritmo es el siguiente:

1. Calcular la probabilidad de pertenencia de un objeto a una clase, dada por la ecuación 5.

$$P(A|x) = \frac{P(x|A)P(A)}{P(x)} = \frac{f(x;\mu_A,\sigma_A)P(}{P(x)} \quad (5)$$

donde ***A***es el clúster del sistema, ***x*** la instancia de muestra, $P_A$la probabilidad de observar un ítem incluido en el clúster ***A*** y $f(x, \mu_A, \sigma_A)$la función de distribución normal de ***A***,que se expresa con la ecuación 6

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-}{2\sigma}} \quad (6)$$

2. Estimar los parámetros de distribución, considerando que son conocidas únicamente las probabilidades de pertenencia a cada grupo. Estas probabilidades actúan como pesos, con lo que el cálculo de la media y la varianza se realiza con las ecuaciones 7 y8 respectivamente.

$$\mu_A = \frac{\sum_{i=1}^{N} w}{\sum_{i=1}^{N} v} \quad (7)$$

$$\sigma^2{}_A = \frac{\sum_{i=1}^{N} w_i{}^2(x_i-\mu}{\left(\sum_{i=1}^{N} w_i\right)^2} \quad (8)$$

donde ***N***es el número total de ejemplos del conjunto de entrenamiento y $w_i$la probabilidad de que la instancia ***i*** pertenezca al clúster***A.***

3. Calcular la verosimilitud general de los datos, dada por la ecuación 9, multiplicando las probabilidades de las instancias.

$$\prod_{i=1}^{N} \sum_{j}^{clusters} P_j P(x_i| \quad (9)$$

Donde ***j*** representa cada uno de los clúster del sistemas, y $P_j$la probabilidad de dicho grupo.

4. Iterar hasta queel incremento sea menor que una cota ε >0 prefijada por el decisor.

6.2. Modelo de agrupamiento propuesto

Comúnmente existen muchas técnicas en un mismo problema para realizar el agrupamiento de los datos. Por eso en este epígrafe se define el procedimiento a seguir para lograr el objetivo.

Dentro de la fase de modelado el primer paso a aplicar consiste en seleccionar la(s) técnica(s) a utilizar para solucionar los objetivos planteados. La selección de los algoritmos *K-medias* como principal y *EM* de apoyo, se basó en un estudio previo realizado, buscando la correspondencia entre la minería de datos y la de procesos. Además se tuvo en cuenta para ello el conjunto de funcionalidades que ofrecen y las necesidades requeridas para obtener los datos resultantes.

El modelo de agrupamiento se basa en la utilización del criterio de *K-medias*, el cual, para su funcionamiento, requiere parámetros previamente especificados. Se define un orden en cuanto a la selección de los parámetros, aunque este no interfiere en el buen funcionamiento del algoritmo.

Se debe indicar, sobre qué tabla registro de proceso se desea realizar el análisis, ya que la entidad puede contener más de un proceso automatizado. De esta, se especifican las características para el agrupamiento, parámetros o atributos a seleccionar por el usuario de acuerdo a las necesidades u objetivos trazados. El último requerimiento es el establecimiento de una cantidad de grupos a formar. Para ello hay que tener en cuenta las posibles variaciones de los parámetros especificados anteriormente, por eso es necesario que el usuario conozca el funcionamiento del proceso de negocio con el que se trabaja. En muchas ocasiones el número de grupos indicado no es el óptimo, por tanto se utiliza el criterio *EM* para obtener esta cantidad de forma automática.

No solo seleccionar los atributos sobre los que se realizará el agrupamiento e indicar la cantidad de grupos a formar, son necesarios para el funcionamiento del algoritmo. Deben especificarse otros parámetros requeridos por la herramienta, aunque vienen definidos valores por defecto.

6.3. Diseño del Modelo

El modelo de agrupamiento se construye basado en las condiciones que necesitan inicialmente el algoritmo *K-medias.* En las tablas I y II son listados los parámetros requeridos por WEKA para la ejecución del algoritmo, sus valores y la explicación del por qué fueron seleccionados.

TABLA I. ESCENARIO DE PARÁMETROS DEL MODELO: CLÚSTER_CANTIDAD_CASOS

Modelo:	Algoritmo:
Cluster_Cantidad_Casos	***SimpleKMeans***
Descripción: Obtención de los datos de grupos generados por la segmentación de los registros, utilizando los atributos seleccionados del proceso, según la cantidad de grupos indicados.	
Parámetros	**Valor y Justificación**
NumClusters	Valor entero. Se utilizará el especificado por el usuario al desarrollar el modelo.
Seed	10, valor por defecto que utiliza la herramienta WEKA. Cantidad de semillas, a partir de la cual se genera el número aleatorio para inicializar los centros de los "clusters".
Características de agrupamiento	Parámetros por los que se realiza el agrupamiento. Varían en cada uno de los experimentos a realizar.

6.4. Evaluación del Modelo

Existen varias alternativas de evaluación para modelos de análisis descriptivos. La propuesta se centra en el cálculo de la desviación estándar por atributos numéricos de un grupo.

En el modelo obtenido, se representa para cada atributo de un grupo el valor de la desviación estándar. Este da una medida de relación entre los casos del clúster, analizando que mientras menor sea el valor, más relación existe entre ellos, es decir, más semejantes son. Para obtener la desviación estándar, se parte del cálculo de la varianza, que indica el grado de dispersión alrededor de la media. El valor de la media se convierte en centro del grupo para el atributo analizado. La ecuación 10 representa este cálculo:

$$c_A = \frac{\sum_{i \in A} x_i}{\|A\|}, \|A\| \tag{10}$$

donde $\|A\|$ es el número de ítems en A y $\boldsymbol{x_i}$ una instancia del grupo.

TABLA II. ESCENARIO DE PARÁMETROS DEL MODELO: CLÚSTER_CASOS_ÓPTIMOS

Modelo:	Algoritmo:
Cluster_Casos_ Óptimos	***EM***
Descripción: Segmentación de los registros, utilizando los atributos seleccionados del proceso, según la cantidad de grupos óptimos encontrados, a partir del resultado que se obtiene del algoritmo de agrupamiento *EM*.	
Parámetros	**Valor y Justificación**
NumClusters	-1, valor que indica al algoritmo calcular el número de "clusters" óptimo para los datos analizados.
Seed	100, valor por defecto que utiliza la herramienta WEKA. Cantidad de semillas, a partir de la cual se generan los números aleatorios del algoritmo.
maxIterations	100, número máximo de iteraciones del algoritmo. Representa la condición de parada si no converge antes.
minStdDesv	1.0E-6, valor por defecto que utiliza WEKA. Mínima desviación estándar admisible en las distribuciones de densidad.
debug	False. Según la opción, muestra información sobre el proceso de "clustering".
Características de agrupamiento	Parámetros del proceso por los que se realiza el agrupamiento. Varían en cada uno de los experimentos a realizar.

La varianza, como se expresa en la ecuación 11, viene dada por la sumatoria de las distancias al cuadrado de cada instancia al centro del grupo, dividida entre la cantidad de casos menos 1. El cálculo de la desviación estándar se realiza a través de la ecuación 12, siendo la raíz cuadrada de la varianza.

$$Var(A_m) = \frac{\sum_{i=1}^{n}(x_i - c_m}{n-1} \tag{11}$$

$$\sigma_A = \sqrt{Var(A_m)} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - c_m}{n-1}} \tag{12}$$

donde $A_m$es el atributo perteneciente al grupo z.

6.5. Descripción de la Herramienta software desarrollada

Como solución al problema planteado se presenta ClusDataPro, una herramienta que utiliza un enfoque orientado a objetos para realizar agrupamiento a partir de datos obtenidos en la ejecución de procesos de negocio. Estas informaciones pueden estar relacionadas con los participantes o empleados, de un proceso en la organización; con parámetros básicos de ejecución, como tiempo de duración, de afectaciones, costos de operatividad, entre otras.

ClusDataPro es una aplicación base para el análisis de los datos de las organizaciones que utilizan BPM en la gestión de sus procesos de negocio. Constituye una guía inicial para conocer el estado de los procesos en las empresas y, partiendo de ello, tomar decisiones. Con esta herramienta se pueden encontrar vínculos referentes al funcionamiento interno de los procesos, lo que da, en gran medida, una valoración del estado de la entidad.

La propuesta presenta una estructuración lógica basada en tres capas que dividen los componentes en niveles de reutilización diferentes. Estas capas están formadas por paquetes que agrupan clases teniendo en cuenta su funcionalidad.

6.6.Estructura en capas de los paquetes del sistema

En la Figura3 se muestra el diagrama de estructuración en capas de los paquetes del sistema. En la capa de Software se incluyen los paquetes de mayor nivel de reutilización, referidos a herramientas ya existentes. En el nivel General se muestran los paquetes con posibilidades de ser reutilizables y en el Particular los componentes propios del sistema, que no son reutilizables para ningún otro.

Los paquetes *java.sql* y *weka* pertenecen al nivel de Software. El primero brinda las clases necesarias para relacionarse con la base de datos, tanto para la conexión como para obtener información de la misma. *Weka* posibilita la realización delos algoritmos*K-medias* y *EM*, ofreciendo clases pertenecientes a la herramienta para aplicar la Minería de Datos WEKA.

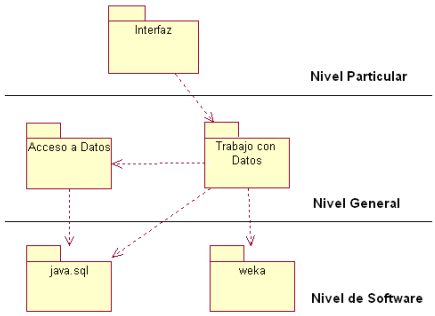

Figura 3. Diagrama de estructuración en capas de los paquetes del sistema.

Una de las funcionalidades de la herramienta es el trabajo dinámico con una base de datos, díganse el conocimiento de su estructura, la realización dinámica de consultas, entre otras. Esto se incluye en el paquete *Acceso a Datos*, donde las clases utilizan *java.sql* en su implementación. En el paquete *Trabajo con Datos* se encuentran las clases para desarrollar la técnica de agrupamiento y generar los reportes a los resultados obtenidos.

El paquete de *Interfaz*, como su nombre lo indica, hace referencia a las interfaces que ofrece el sistema al usuario para dar respuesta a sus solicitudes. Este utiliza al paquete *Trabajo con Datos* para hacer al sistema independiente y vulnerable a cambios.

La implementación de ClusDataProfue realizada utilizando como IDE el Eclipse en su versión 3.4.0 en una plataforma de trabajo Windows XP Professional. Fue necesaria la incorporación de las librerías JDK 6.0 para el trabajo con el paquete *java.sql* y, weka-src.jar y weka.jar para la utilización del paquete *weka*.

7. RESULTADOS Y DISCUSIONES

Para la aplicación de la herramienta, se tienen en cuenta los modelos anteriormente diseñados. Se realizan dos experimentos asociados, cada uno, a un modelo de agrupamiento, y se detallan los resultados de los mismos. Las pruebas se desarrollan condatos obtenidos de la ejecución del proceso de "Aprobación de Contratos", automatizado en la empresa Tecnomática.

Como resultado de los experimentos, se brinda el centro de cada grupo generado, que ofrece una idea de las características de los registros que pertenecen al clúster. Con el valor de desviación estándar por atributos, como forma de evaluación, se puede conocer cuán compacto es el grupo y arribar a conclusiones sobre el comportamiento observado.

Señalar que es respetada la terminología utilizada por la empresa para nombrar los procesos, tareas, usuarios, estados, etc.

Experimento # 1

Los modelos de agrupamiento que se obtienen a continuación corresponden con el diseño del modelo *Clúster_Cantidad_Casos*. Se requiere una cantidad deseada de grupos y los parámetros para el agrupamiento. Se tienen en cuenta para la prueba los parámetros *duración* y *estado* correspondientes a la tabla *Aprob_Contrato*, donde están registrados los nuevos contratos.

Se desea obtener la relación entre los *estados* y tiempos de *duración* en el proceso *Aprobación de Contratos*, para conocer en cuál estado el proceso se tarda más.

La tabla III resume los parámetros y resultados obtenidos. Se definen 4 grupos a formar porque el parámetro *estado* varía en esa cantidad. El análisis de los centros de grupos obtenidos arroja una incongruencia, asigna al estado "Solicitado" dos tiempos promedios de duración muy diferentes. Al grupo 3 pertenecen la mayoría de las instancias, lo que justifica sea el que mejor resume las características de los procesos "Solicitados". El grupo 4 está compuesto por procesos "Solicitados" que describen un comportamiento distinto al esperado.

En sentido general y cumpliendo el objetivo de la prueba, se puede conocer que en la mayoría de los casos los contratos fueron "Aprobados". Un proceso completo, definido por este estado, tiene un tiempo de *duración* promedio equivalente a los 16 días. Esta *duración* varía en aproximadamente cuatro días, es decir, pueden existir procesos que demoren en ser aprobados 12 días y otros 20 días.

TABLA III
PARÁMETROS Y RESULTADOS DEL EXPERIMENTO # 1

Tabla registro de proceso	*Aprob_Contrato*			
Parámetros seleccionados	*duración, estado*			
Cantidad de Grupos	*4*			
No.	**Características**		**Cantidad de casos**	**Desviación Estándar**
	Duración	**Estado**		
1	9.5707	No Aprobado	7 (21%)	1.67
2	15.8832	Aprobado	16 (48%)	3.87
3	5.5524	Solicitado	8 (24%)	1.54
4	19.134	Solicitado	2 (6%)	0.06

Experimento # 2

Los modelos de agrupamiento que se muestran corresponden con el diseño del modelo *Clúster_Casos_Óptimos*, donde se obtiene el número de grupos de la ejecución del algoritmo *EM*. Esta cantidad es tomada como parámetro para realizar el agrupamiento a través del criterio de *K-medias.* Con el procedimiento se crearon las pruebas del experimento, utilizando las características de agrupamiento señaladas en la prueba anterior. Se brindan los datos generales por grupo generado, para arribar a conclusiones en cuanto a la eficiencia del algoritmo.

En la ejecución del criterio de *EM* se obtiene tres como número óptimo de grupos. Los resultados de realizar el

agrupamiento a través del algoritmo *K-medias*, se resumen a continuación:

TABLA IV
PARÁMETROS Y RESULTADOS DEL EXPERIMENTO # 2

Tabla registro de proceso	*Aprob_Contrato*			
Parámetros seleccionados	*duración, estado*			
No.	**Características**		**Cantidad de casos**	**Desviación Estándar**
	Duración	**Estado**		
1	9.5707	No Aprobado	7 (21%)	1.67
2	15.8832	Aprobado	16 (48%)	3.87
3	8.2687	Solicitado	10 (30%)	5.88

Los resultados arrojan una caracterización del tiempo de *duración* según el estado de los procesos. Se obtiene como conclusión que los contratos solicitados a la empresa tienen una mayor probabilidad de ser "Aprobados", con un tiempo promedio de 15 días. La desviación estándar es alta por el comportamiento de los procesos aprobados. Existen procesos que demoran 20 días y otros 12.

7.1. Análisis de los Resultados

Para el desarrollo de este epígrafe se tienen en cuenta los modelos generados en los Experimentos 1 y 2. De ellos se analizan los resultados obtenidos y el aporte brindado.

Desde un inicio existe el problema de seleccionar la cantidad de grupos a generar, parámetro requerido por el Experimento 1. Si no se conoce la relación entre variación de los tiempos de *duración*y los demás atributos, es muy difícil definir un número óptimo de grupos. Se generan tantos grupos como se indican, donde de cada uno se extrae información, pero el resultado en general es muy amplio. Este problema es resuelto en el Experimento 2, donde se ponderan las *duraciones* según un análisis de densidad realizado por el criterio de *EM*. Los resultados de la nueva ejecución de *K-medias*, con este número de clúster, se centran más en el objetivo que define la prueba.

Es bueno destacar el uso del Experimento 1 si se desea conocer sobre el comportamiento real de los parámetros. Mientras mayor números de grupos, más particiones resultarán, con una desviación estándar menor, lo que indica una mejor descripción del comportamiento dentro del grupo, es decir, más semejanza en los datos. Con los modelos obtenidos con este experimento se encontraron anomalías existentes en los datos, es decir, comportamientos no deseados. Estos resultados son un buen punto de análisis para la empresa. Conociendo dónde se encuentran los errores, es más fácil corregirlos y tomar decisiones para que no vuelvan a ocurrir.

RECEIVED AUGUST 2011
REVISED MAY 2012

REFERENCIAS

[1]ALVESANA KARLA (2008): **Minería de Procesos. Más allá de los Modelos de Procesos**. Available from: http://gerenciainnovacion.blogspot.com/2008/08/minera-de-procesos.html.

[2] BRITO RAYCOS (2008): **Minería de Datos aplicada a la Gestión Docente del Instituto Superior Politécnico JoséAntonio Echeverría, Facultad de Ingeniería Informática. Instituto Superior Politécnico "José A. Echeverría"** , Ciudad de la Habana. Tesis de maestría, Edición XIII Informática Aplicada.

[3]DUNHAM M. H (2003): **Data Mining. Introductory and Advanced Topics**. Prentice Hall.

[4]ROSETE SUAREZ ALEJANDRO (2004) **Minería de Datos: el camino de la academia a la realidad cotidiana.** Centro de Estudios de Ingeniería de Sistemas (CEIS), Instituto Superior Politécnico "José Antonio Echeverría" (CUJAE). Ciudad de La Habana, Cuba.

[5]HERNANDEZ RAMIREZ M. J y FERRI RONALD (2004): **Introducción a la minería de datos.** Madrid. Ed. PEARSON EDUCACIÓN, S.A.: Universidad Politécnica de Valencia, Departamento de Sistemas Informáticos y Computación.

[6]MOLINA yGARCIA H (2006): **Técnicas de Análisis de Datos. Aplicaciones Prácticas utilizando Microsoft Excel y WEKA.** 2006, Universidad Carlos III, Madrid.

[7]MARTIN DIANA (2008): **Aplicación de técnicas de minería de datos en la Inteligencia Criminal en Cuba**, Facultad de Ingeniería Informática,CEIS. 2008, Instituto Superior Politécnico "José Antonio Echeverría": Ciudad de la Habana.

[8]WEIJTERS TON (2009): **Introduction to Data Mining**. Technische Universiteit Eindhoven, Holanda.

[9]MARCANO y TALAVERA ROSALBA (2007): **Minería de Datos como soporte a la toma de decisiones empresariales**. Opción **23**, 104-118.

[10]ALVES ANA KARLA, et al (2008): **Process Mining Based on Clustering: A Quest for Precision.** BPM 2007 Workshops, editors. ter Hofstede, A., Benatallah, B., and Paik, H.-Y.17-29.

[11]ESPEN JOHN andATLE G (2008): **Preprocessing Support for Large Scale Process Mining of SAP Transactions.** BPM 2007 Workshops, ed. ter Hofstede, A., Benatallah, B., and Paik, H.-Y.p30-41.

[12]HERNANDEZ EDNA (2006): **Algoritmo de clustering basado en entropía para descubrir grupos en atributos de tipo mixto.** 2006: Ciudad México, D.F.

[13]HUANG X (2006): **Clustering Analysis and Algorithms. Encyclopedia of Data Warehousing and Mining**. Idea Group REFERENCE, York University, York.

[14]GARRE MIGUEL, et al (2007): **Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software.** REICIS: Revista Española de Innovación, Calidad e Ingeniería del Software,ISSN (version en linea) 1885-4486, vol3,no 001,p. 7-19. Madrid, España. Disponible en: http://redalyc.uaemex.mx/src/inicio/ArtPdfRed

[15]DIAZ JUAN CARLOS, et al (2007): **Business Process Management. El negocio en el centro de los sistemas.** Available from: http://www.atosorigin.com/WhitePaper/BPMWhitePaper.pdf.

[16]PEREZ JIMENEZ SURELYS (2009): **Obtención de requisitos funcionales a partir del Análisis de Procesos de Negocio.** Facultad de Ingeniería Informáica, Instituto Superior Politécnico "José Antonio Echeverría": Ciudad de la Habana.

[17] SORDI JOSE OSVALDO (2007) **Análise de componentes da tecnologia de Business Process Management System (BPMS) sobre la perspectiva de um caso prático. Business Process Management Systems technology practic components analysis.** JISTEM - Journal of Information Systems and Technology Management.On-line version, ISSN 1807-1775. JISTEM J.Inf.Syst. Technol. Manag. (Online), vol4, no1, p. 1-8, São Paulo, Brazil.
disponible en: http://dx.doi.org/10.4301/

[18] LEGANZA Gene,Vollmer Ken (2008): **Using BPM to Improve Operational Efficiency**. Enterprise Architecture Professionals. Articulo disponible en. http://www.ebizq.net/topics/bpm_platforms_suites/features/10043.html

[19]HINOJO FRANCISCO JOSE (2008): **BPM: Una herramienta de competitividad.** México: Process & Project Health Services.

[20]TABARES, PINERA y BARRERA (2008): **Un patrón de interacción entre diagramas de actividades UML y Sistemas Workflow.** Revista EIA. Escuela de Ingeniería de Antioquia, Medellín, 10, 105-120.