**Respuestas**
Año 14
No. 1
Junio 2009
ISSN 0122-820X

# Técnicas de minería de datos aplicadas a la valoración de ambientes creativos

Germán Augusto Osorio Z.[1] | Luis Gonzalo Sánchez G.[2] | Néstor Darío Duque M.[3]

Recibido:
10 de abril de 2009

Aceptado:
9 de junio de 2009

## Resumen

*En este trabajo se presentan los resultados de la aplicación de algunas técnicas de Minería de Datos (DM) sobre la información de una investigación realizada por la Universidad Nacional de Colombia - Sede Manizales en los años 2004 y 2005 en veintitrés empresas de la ciudad, con el fin de determinar el estado de los ambientes creativos para la innovación, desde las dimensiones psicosocial, didáctica y física. Se consideran las técnicas de Análisis Discriminante Lineal (LDA), Análisis de Componentes Principales (PCA) y Análisis de Varianza Multivariado (MANOVA) de un camino para tareas de orden predictivo y descriptivo.*

***Palabras clave:*** *Minería de datos, creatividad, análisis discriminante, análisis de varianza, análisis de componentes principales*

## Abstract

*This paper presents the results of applying some techniques of Data Mining (DM) on the information an research project conducted by the Universidad Nacional de Colombia - Sede Manizales in 2004 and 2005 in twenty-three companies in the city, in order to determine the status of the creative environment for innovation, from the psychosocial, didactical and physical dimensions. Considered techniques are Linear Discriminant Analysis (LDA), Principal Components Analysis (PCA) and Multivariate Analysis of Variance (MANOVA) of a way to order tasks predictive and descriptive.*

***Keywords:*** *Data Mining, Creativity, Linear Discriminant Analysis (LDA), Multivariate Analysis of Variance (MANOVA), Principal Components Analysis (PCA)*



## 1. Introducción

Recientemente, el grupo de trabajo académico PROCREA de la Universidad Nacional de Colombia - Sede Manizales cuyo objeto de estudio es la creatividad y la innovación, inició un trabajo de investigación con el objetivo de "*determinar el estado del ambiente creativo para*

[1] 1 Universidad Nacional de Colombia Sede Manizales, gaosorioz@unal.edu.co
[2] Universidad Nacional de Colombia Sede Manizales, lgsanchezg@unal.edu.co
[3] Universidad Nacional de Colombia Sede Manizales, ndduqueme@unal.edu.co

Respuestas
Año 14
No.1
Junio 2009
ISSN 0122-820X



*la innovación, desde las dimensiones psicosocial, didáctica y física, en las empresas de Manizales y formular lineamientos estratégicos para su fortalecimiento"* [1]. Para ello, se desarrolló una encuesta como instrumento de medición del estado del ambiente creativo aplicada en 23 empresas de Manizales consideradas como estratégicas para el desarrollo económico local. El análisis de los datos obtenidos de estas encuestas se fundamentó en la utilización de técnicas estadísticas univariadas. Sin embargo, la evaluación de la creatividad es muy compleja y ha de ser polivalente respecto a las estrategias utilizadas, por lo tanto, se hace necesario recurrir a estrategias variadas, ya que a través de ellas se recaban informaciones complementarias [2]. Por este motivo, con el fin de dar continuidad al esfuerzo previamente realizado, en este trabajo, se introducen técnicas de minería de datos, especialmente análisis multivariado como herramienta alternativa y de gran potencial para la extracción de información relevante contenida en las encuestas previamente mencionadas.

Con el creciente flujo de información y recopilación de datos, las técnicas de minería de datos han venido ganando un creciente interés dentro de diversas disciplinas del conocimiento. El estudio de la creatividad no ha sido ajeno a esta tendencia. En [3], se plantea la creación de un modelo cognoscitivo de la creatividad, inspirado en redes conexionistas y en los conceptos de activación y conectividad, propio de la teoría de procesamiento de la información. Para el análisis de resultados, utiliza diagramas de dispersión, análisis de correlación y análisis de factores con rotación oblicua. Por otra parte, en [4], se presentan los resultados obtenidos en la evaluación de un programa de desarrollo de la creatividad, puesto en marcha en el contexto normal de clase y del currículo ordinario en los niveles de educación infantil y primer ciclo de educación primaria. Los datos fueron evaluados usando análisis factorial de varianza entre e intragrupo. También, en [5], se plantean las múltiples formas en que las pruebas de "papel y lápiz" fueron desarrolladas basadas en el Test de Pensamiento Creativo de Torrance (TTCT).

Con el método propuesto se disminuye la cantidad de tiempo requerido para la administración y puntuación de las pruebas de creatividad. El estudio utiliza, como un método alternativo, un modelo de ecuación estructural (SEM) con múltiples indicadores para examinar la validez y confiabilidad del nuevo test de creatividad. La evaluación de los resultados se hizo con estadísticos descriptivos multivariados. Con la aplicación del método de modelo de variable latente, se observaron mejoras sustanciales en la consistencia interna y en los coeficientes de validez concurrente. Esto en parte porque el método trata con variables latentes, las cuales están libres de error de medida y parte en la capacidad de modelos SEM de usar múltiples indicadores.

En [6], se plantea la existencia de propuestas para interrelacionar la creatividad con la innovación tecnológica, como la de Buglioni y Abran (2001) que busca identificar el qué y el cómo medir la creatividad, analizando la creatividad corporativa y no solo el individuo diseñador de productos y para ello involucra aspectos organizativos tales como métodos, técnicas y herramientas que permitan mejorar los procesos empresariales para lograr mejores resultados y la identificación de la trayectoria evolutiva seguida durante la adopción progresiva de tales métodos.

En el presente trabajo, se utilizan tres técnicas de análisis multivariante como parte de la minería de datos, en la interpretación de la información obtenida en [1]. En particular, se consideran el análisis de componentes principales y el análisis discriminante lineal en conjunto con el análisis de varianza multivariado. A continuación, se presenta una introducción acerca de la minería de datos y de las técnicas consideradas para este trabajo. Seguidamente, en la sección 3 se describen los datos empleados en el análisis y los resultados obtenidos. Finalmente, en la sección 4 se presenta las conclusiones derivadas de los análisis y algunas ideas para trabajo futuro.

Respuestas
Año 14
No. 1
Junio 2009
ISSN 0122-820X

# 2. Marco teórico

## 2.1 Minería de datos

La DM integra técnicas de análisis de datos y extracción de modelos. Se basa en varias disciplinas, algunas de ellas más tradicionales como la estadística y el aprendizaje automático, se diferencia de ellas en la orientación más hacia el fin que hacia los medios [7]. La minería de datos tiene como objetivo analizar los datos para extraer conocimiento. Este conocimiento puede ser en forma de relaciones, patrones o reglas inferidos de los datos y (previamente) desconocidos, o bien en forma de una descripción más concisa (es decir, un resumen de los mismos). Estas relaciones o resúmenes constituyen el modelo de los datos analizados.

Dentro de la minería de datos se distinguen dos tipos de tareas: predictivas donde se tratan problemas en los que hay que predecir uno o más valores para uno o más ejemplos; y descriptivas, que buscan describir y arrojar luces a la interpretación de los datos. Para la solución de dichas tareas, se emplean diferentes métodos y técnicas entre ellas: *las técnicas algebraicas y estadísticas, técnicas bayesianas, técnicas basadas en árboles de decisión y sistemas de aprendizaje de reglas, técnicas basadas en redes neuronales artificiales y difusas*. También, se pueden combinar las técnicas ya mencionadas, para aprovechar las ventajas que pueda ofrecer cada una de ellas.

## 2.2 Análisis de componentes principales

El Análisis de Componentes Principales ha sido la tendencia dominante para el análisis de datos en un gran número de aplicaciones [8]. Su atractivo recae en la simplicidad y capacidad de reducción de dimensión, minimizando el error cuadrático de reconstrucción que se obtiene a partir de una combinación lineal de variables latentes conocidas como componentes principales. Los parámetros del modelo pueden ser calculados directamente de la matriz de datos centralizada X bien sea por descomposición en valores singulares o la diagonalización de la matriz de covarianza [3]. Sea $x_i$ el iésimo vector de observación de longitud $p$, siendo $x = (x_1, x_2, ...., x_n)^T$ , la matriz de rotación U con la que se calculan las $p'$ componentes principales denotadas por $z$ que resumen $x$, como sigue.

$$z = U^T x \qquad (1)$$

U puede ser calculada a partir de los primeros $p'$ valores propios de $X^TX$, esto es,

$$X^TXU = U\Lambda \qquad (2)$$

## 2.3 Análisis discriminante lineal

El objetivo del análisis discriminante es el establecimiento de reglas sobre las cuales un objeto observado se asigna a una población dada. Se pretende asignar un objeto observado a una u otra población, partiendo de la suposición de que las poblaciones en cuestión generan distribuciones gaussianas.

Sea $f_1$ y $f_2$ distribuciones normales con distintos vectores de medias pero idéntica matriz de varianzas, en las cuales se desea clasificar [9] un elemento genérico $x$, que pertenece a la población $i = 1, 2$ que tiene función de densidad:

$$f_i(x) = \frac{1}{(2\pi)^{\frac{p}{2}}|V|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(X - \mu_i)^T V^{-1}(X - \mu_i)\} \qquad (3)$$

En donde $V$ viene dado por el estimador insesgado de la matriz de covarianza común

$$S_{pl} = \frac{1}{n_1 + n_2 - 2}((n_1 - 1)S_1 + (n_2 - 1)S_2) \qquad (4)$$

Siendo $S_1$ y $S_2$ los estimados de la covarianza para las clases 1 y 2, respectivamente [10].

La partición óptima, es clasificar en la población $P_2$ si:

$$\frac{f_2(x)\pi_2}{c(2|1)} > \frac{f_2(x)\pi_1}{c(1|2)} \qquad (5)$$

Respuestas
Año 14
No.1
Junio 2009
ISSN 0122-820X



Como ambos términos son siempre positivos, tomando logaritmos y sustituyendo $f_i(x)$ por su expresión, la ecuación anterior se convierte en:

$$\{-\frac{1}{2}(X-\mu_2)^T V^{-1}(X-\mu_2)\}+\ln\frac{\pi_2}{c(2|1)} > \{-\frac{1}{2}(X-\mu_1)^T V^{-1}(X-\mu_1)\}+\ln\frac{\pi_1}{c(1|2)} \quad (6)$$

Llamando $D_i^2$ a la distancia de Mahalanobis entre el punto observado, $x$, y la media de la población $i$, se tiene:

$$D_1^2 - \ln\frac{\pi_1}{c(1|2)} > D_2^2 - \ln\frac{\pi_2}{c(2|1)} \quad (7)$$

### 2.4 Análisis de varianza (MANOVA) de un camino

El objetivo principal del análisis de varianza es la evaluación de la hipótesis acerca de la igualdad de poblaciones para $k$ muestras. Particularmente, cada muestra corresponde a una clase. La evaluación de la hipótesis nula se realiza por medio de la comparación de dos estimadores de varianza para la población: un estimador derivado de la diferencia de las medias de cada muestra (clase), y como segundo estimador se examina la varianza intra clase (dentro de cada muestra). Esta evaluación corresponde al siguiente modelo generativo para los datos:

$$x_{kj}=m_k+\varepsilon_{kj}, m_k=m+\alpha_k \quad (8)$$

siendo $j$ la observación y $k$ la clase, $m_k$ es el vector de media para cada clase, y $\varepsilon_{kj}$ es la perturbación del modelo, m es la media global $\alpha_k$ es la perturbación de la media. La hipótesis evaluada es

$H_0 : m_1 = m_2 = \cdots = m_L$
$H_1 :$ Al menos uno de los $m_j$ difiere del resto

donde $L$ corresponde al número de clases.

# 3. Marco experimental

De la investigación del PROCREA se tiene el resultado de la encuesta que consta de 54 preguntas, aplicada a 1438 trabajadores en los niveles operativo, táctico y estratégico de 23 empresas de Manizales de los sectores Alimentos y Bebidas, Insumos, Productos de consumo y Servicios.

Los investigadores del PROCREA asignaron valores numéricos decrecientes para cada categoría. El valor asignado a cada respuesta era adicionado al total de cada característica específica en las dimensiones analizadas en la encuesta, como son: didáctica, física y psicosocial.

Las características por dimensión son las siguientes:

*Dimensión didáctica:* formación, aprendizaje, comunicación, lúdica

*Dimensión física:* Simbólico estético, técnico, funcional

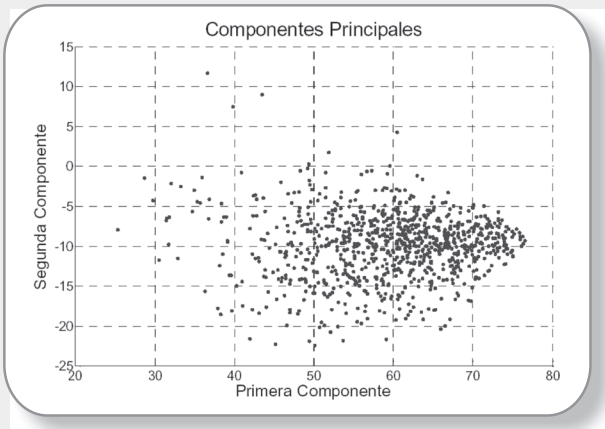
*Dimensión psicosocial:* motivación y satisfacción, reconocimiento, apertura, afiliación, logro, poder, flexibilidad, identidad

En la encuesta también se preguntó y valoró la innovación, la cual no es considerada en este análisis, ya que lo que se quiere evaluar en este trabajo, es específicamente la creatividad en sus tres dimensiones.

El proceso se inició con la integración en una sola base de datos, de las encuestas que estaban agrupadas una, por el sector servicios; y otra, por los demás sectores. Posteriormente se calculó el valor de cada característica dentro de la dimensión, la cual, se obtenía de sumar los valores de las respuestas asociadas a la característica dada. Esas características totalizadas, fueron las que se utilizaron en el análisis.

Respuestas
Año 14
No. 1
Junio 2009
ISSN 0122-820X

**Figura 1** Proyección de los datos sobre las dos primeras componentes principales




El uso de PCA con matriz de covarianza, permitió determinar variables latentes dentro de la encuesta, por ejemplo, la primera componente, según sus características de componente de tamaño, puede ser vista como determinante del nivel de ambiente creativo (Figura 1). Esta situación puede constatarse con los valores del primer vector propio que representa una componente de tamaño ya que todas las variables originales aportan en la misma dirección sobre esta variable latente.

Gráficamente (Figura 2) se puede observar como los coeficientes de los tres primeros vectores propios, permiten deducir la consistencia de la encuesta en cuanto a las variables a medir y su ámbito (el nivel de similitud entre el grupo de variables de la dimensión dada: didáctico, físico, psicosocial).

Debido al número de clases correspondientes a las empresas analizadas (veintitrés), el análisis visual sobre las tres primeras componentes, no permite observar claramente, la presencia de patrones, sin embargo, se pueden observar diferencias entre algunas empresas. Es muy posible que las nubes de puntos, generadas por cada una de la empresas, no sean fácilmente separables dada la posibilidad de la existencia de ambientes creativos similares entre ellas.

**Figura 2** Coeficientes de los autovectores sobre tres componentes principales

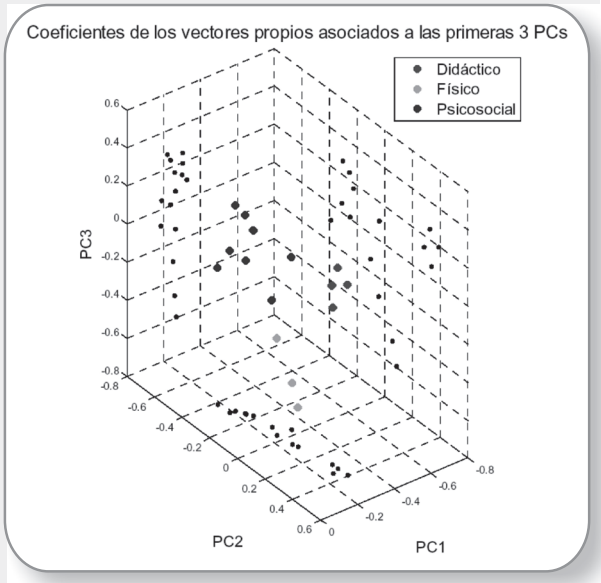
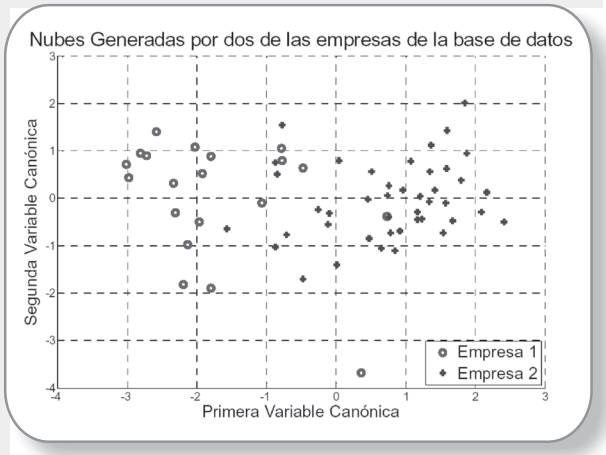



**Figura 3** Nubes de datos en dos empresas




Dado que algunas de las empresas poseen un número de observaciones limitado, es decir, el número de observaciones no es suficiente para construir clasificadores de tipo Bayesiano directamente. Entonces, la construcción de los clasificadores se llevó a cabo utilizando el espacio proyectado utilizando las tres primeras componentes principales.

En la figura 3 se muestra uno de los mapeos a variables canónicas entre un par de empresas, en el cual se puede observar la presencia de diferencia entre las respuestas entre una y otra.

Respuestas
Año 14
No.1
Junio 2009
ISSN 0122-820X



Para efectos de determinar la diferencia de ambientes creativos entre pares empresas, se utilizó un clasificador lineal discriminante de covarianza promediada (ver sección 2).

**Figura 4** Error de entrenamiento para el clasificador


Partiendo de él, la figura 4 contrasta los resultados obtenidos por la prueba de hipótesis al presentar el error de entrenamiento para el clasificador. En la figura 5, se representa la tabla de resultados de la prueba de hipótesis de igualdad de medias entre pares de empresas (los cuadros blancos muestran para la pareja el rechazo de la hipótesis de igualdad), sin embargo, es necesario ser cuidadoso al momento de rechazar o aceptar debido a que el número de observaciones no es balanceado por la empresas (grandes diferencias entre el número de encuestados para las empresas).

# 4. Conclusiones

Se presentaron dos técnicas relacionadas a la solución de tareas predictivas y descriptivas en la minería de datos, aplicadas sobre las encuestas de ambientes creativos en empresas de Manizales. Se pudo observar como las variables propuestas por el estudio presentan alta similitud dentro de las dimensiones analizadas (didáctico, físico, psicosocial). También, el análisis discriminante permitió vislumbrar la existencia de diferencias entre empresas con respecto a sus ambientes creativos y de la misma forma la similitud entre algunas de ellas.

Como trabajo futuro se propone la exploración de métodos como el Análisis de Correspondencias Múltiples para la representación de las encuestas a partir de variables categóricas. Partiendo de estos resultados, utilizar otras técnicas de DM para las tareas ya abordadas (predicción, descripción).

**Figura 5** Rechazo de hipótesis sobre igualdad de medias


# 5. Bibliografía

[1] C. A. González and A. Vargas, "Estado del ambiente creativo para la innovación en las empresas de Manizales y lineamientos estratégicos para su fortalecimiento." Universidad Nacional de Colombia, Sede Manizales, 2005.

[2] S. de la Torre and V. Violant, Comprender y Evaluar la Creatividad Tomo II: Cómo investigar y evaluar la Creatividad. Ediciones Aljibe, 2006.

[3] M. Molina, "Propuesta para la creación de un test sicométrico para la medición de la

Respuestas
Año 14
No. 1
Junio 2009
ISSN 0122-820X

creatividad," Actualidades en Psicología, vol. 18, no. 105, pp. 49–69, 2002.

[4] M. D. P. et al, "Evaluación de un programa de desarrollo de la creatividad," Psicothema, vol. 14, no. 2, pp. 410–414, 2002.

[5] J. Abedi, "A latent-variable modeling approach to assessing reliability and validity of a creativity instrument," Creativity Research Journal, vol. 14, no. 2, pp. 267–276, 2002.

[6] J. Chaur-Bernal, "Diseño conceptual de productos asistidos por ordenador: Un estudio analítico sobre aplicaciones y definición de la estructura básica del nuevo programa," Ph.D. dissertation, Universidad Politécnica de Cataluña, 2004.

[7] J. H.-O. et al, Introducción a la Minería de Datos. Pearson Prentice Hall, 2004.

[8] G. Daza, L. G. Sánchez, and J. F. Suárez, "Selección de características orientada a sistemas de reconocimiento de granos maduros de café," Scientia et Technica, Agosto 2007.

[9] D. Peña, Análisis de Datos Multivariantes. McGraw Hill, 2002.

[10] A. Rencher, Methods of Multivariate Analysis. Wiley- Interscience, 2002