“Implementación de un Sistema basado en Minería de Datos para la obtención de las Preferencias Estudiantiles de nivel superior para la planificación de Materias”

Altamirano, M.[1] : Cevallos, L. [2] ; González, I.[3] ; Echeverría P.[4]
[1] Ingeniero en Computación en Sistemas de Información.
[2] Ingeniero en Computación en Sistemas de Información.
[3] Ingeniero en Computación en Sistemas Tecnológicos.
[4] Director de Tesis, profesor de la ESPOL.
Facultad de Ingeniería en Electricidad y Computación
Escuela Superior Politécnica del Litoral
Campus Prosperina, Km. 30.5 vía Perimetral, Guayaquil, Ecuador

Resumen

En el ámbito de modernización que viven los registros académicos y siendo una necesidad la obtención de preferencias estudiantiles con respecto a la buena planificación de las materias.

Este trabajo luego de un análisis y viendo la necesidad del estudiante universitario utiliza los modelos discriminantes para el desarrollo de la clasificación de las preferencias, en el mismo se define los datamarts donde se concentran los datos operativos, valores de hecho y conocimientos extraídos de los cuestionarios electrónicos. Además implementamos una interfaz Web necesaria para: ingreso de datos de los estudiantes, operación de los datamarts y ambiente de toma de decisiones.

Palabras claves: minería de datos, datamarts, análisis discriminante, interfaz web.

Abstract

In the area of upgrading the living being and academic records need to obtain a student preferences with regard to good planning matters.

This paper analyzes and after seeing the need for student university uses the discriminant models for the development of ranking of preferences, it is defined in the datamartswhere are the operational data, facts and values knowledge extracted from the electronic questionnaires. Moreover implement a Web interface to: data entry of students, datamarts the operation and decision-making environment.

Keywords: data mining, datamarts, discriminant analysis, web interface.

1. Antecedentes:

Demarcar un tipo de sistema que proporcione la extracción del conocimiento de los estudiantes con respecto a las preferencias académicas a medida que va transcurriendo su ciclo de rol estudiantil.

Reflejar las tendencias de elección por parte de los alumnos; es una herramienta útil que facilite la toma de decisiones para los respectivos coordinadores. Para un futuro semestre; de que materias y que profesores los alumnos desean tomar se podría determinar con un grado de confiabilidad la cantidad de alumnos que desean tomar una materia para el próximo semestre tomando en cuenta dependencias estadísticas del mismo ya sea porque esta tomando una materia que le permite tomar otra después o por el status del mismo alumno si se encuentra a prueba o esta es una materia que se encuentra en un nivel superior al siguiente semestre; sin embargo conocer las tendencias del alumnado por parte de los coordinadores facilitaría la labor de los mismos al abrir un tope de materias que se demandan.

El presente trabajo presenta un sistema que servirá de soporte para el coordinador a la toma de decisiones de este tipo de procesos con una respectiva toma de información y procesamiento de la misma para dar una selección óptima y eficiente de la cantidad demandada de materias y profesores para un siguiente semestre.

2. Recolección de Datos:

El sistema será un anexo en el sistema CENACAD que hará referencia a un formulario personal para cada alumno que ingrese su respectivo numero de matricula, el sistema presentara el respectivo flujo del alumno señalando de forma puntual las materias que ya ha tomado y las materias que esta tomando en el actual semestre permitiendo visualizar las materias que no ha tomado aun; así disminuiremos el ingreso de elecciones erróneas por parte del alumno, dentro del formulario una vez escogido las materias se le pedirá al usuario que escoja con que profesor desearía tomar cada una de las materias que eligió con el horario que le parezca mas conveniente para el.

Hay que tomar en cuenta que éste módulo del sistema no pondrá restricciones en el ámbito de la asignación de profesor y el horario que decida el usuario ya que nuestro objetivo es conocer también que horarios desea tomar el alumno para su respectivo análisis.

3. Análisis de los datos ingresados

Una vez que se hayan tomado un rango de muestras se procederá a su respectivo análisis. El análisis que se planteara para este proceso será el análisis discriminante basado en clasificadores lineales, cuadráticos o difusos.

Que responde al Análisis Discriminante….

¿Es posible predecir con antelación si un cliente que solicita un préstamo a un banco va a ser un cliente moroso?
¿Se puede predecir de antemano si un recluso que ha solicitado un permiso carcelario, huirá?
¿Se puede predecir si una empresa va a entrar en bancarrota?
¿Cuáles son las razones que llevan a un consumidor a preferir una determinada marca sobre otras existentes en el mercado?
¿Existe discriminación por razones de sexo o de raza en una empresa o en un colegio?

El análisis discriminante estudia las técnicas de clasificación de sujetos en grupos ya definidos.

Clasificación Supervisada es una técnica que sirve para indicar que conocemos una muestra de elementos bien clasificados que sirven de modelo para la clasificación de las siguientes observaciones.

Ejemplos de aplicaciones de análisis discriminante:

Clasificar un retrato entre dos posibles pintores y de acuerdo a :

- La profundidad del trazo y la proporción que ocupa el retrato en la superficie del lienzo de ambos pintores.
- Asignar un texto escrito de procedencia desconocida a uno de varios autores por las frecuencias de utilización de palabras,
- Una empresa esta en riesgo de quiebra o no,
- Un paciente esta enfermo de cáncer o no, etc.

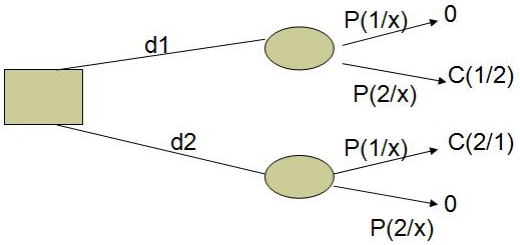
Enfoque de análisis discriminante clásico de acuerdo a Fisher

Basado en la normalidad multivariante de las variables involucradas y es optimo bajo el supuesto de que si todas las variables son continuas es frecuente que aunque los datos originales no sean normales es posible transformar las variables para que lo sean.

- **Planteamiento del problema:** Tenemos P1 y P2 poblaciones y tenemos una variable aleatoria x la cual es continua y las funciones de densidad de las poblaciones son f1 y f2 son conocidas. Deseamos estudiar el problema de clasificar un nuevo elemento Xo.
- Conocemos también que las probabilidades a priori son $\pi 2$ y $\pi 1$ y $\pi 1 + \pi 2 = 1$, es decir que el elemento Xo venga de cada una de las dos poblaciones
- En general supondremos que las posibles decisiones en el problema son únicamente dos: asignar en P1 o en P2.
- Si las consecuencias de un error pueden cuantificarse podemos incluirlas en la solución del problema.

- C(2/1) y c(1/2) donde c(i/j) es el coste de clasificación en Pi de una unidad que pertenece a Pj.
- El decisor debe maximizar su función de utilidad minimizando el coste esperado.

- Con estas dos hipótesis la mejor decisión es la que **minimiza los costes esperados.**



- Si clasificamos al elemento en el grupo 2 las posibles consecuencias son**:**
  - acertar, con probabilidad P(2/xo) en cuyo caso no hay ningún coste de penalización
  - Equivocarnos, con probabilidad P(1/x) en cuyo caso incurrimos en e coste asociado c(2/1).

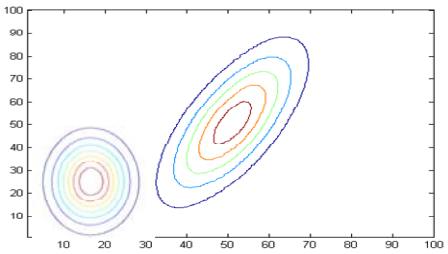
Función lineal discriminante

- **Poblaciones Normales:**

Supuestos: Tenemos dos poblaciones f1 y f2 con distribución normal, con distinto vector de medias pero idéntica matriz de varianzas. Y un elemento genérico Xo que si pertenece a la población 1 y 2.

Distancia de Mahalanobis: Di2 = (x-ui)'V-1(x-ui)

Distancia observada entre el punto observado X y la media de la población.



Es decir clasificar en P2 si D12 >D22 en otras palabras

Clasificar la observación en la población en donde la distancia sea mínima.

Un buen método es aplicar la función discriminante a las n observaciones y clasificarlas. En el caso de dos grupos, obtendríamos la tabla.

		Clasificado	
		P_1	P_2
Realidad	P_1	n_{11}	n_{12}
	P_2	n_{21}	n_{22}

Donde nij es el número de datos que viniendo de la población i se clasifica en j. El error aparente de la regla es:

$$\text{Error} = \frac{n_{12} + n_{21}}{n_{11} + n_{22}} = \frac{\text{Total mal clasificados}}{\text{Total bien clasificados}}.$$

Este método tiende a subestimar las probabilidades de error ya que los mismos datos se utilizan para determinar los parámetros y para valuar la regla resultante. Una de lo mejores métodos es de clasificar cada elemento con una regla que no se ha construido.

Para ello construimos n funciones discriminantes con las n muestras de tamaño n-1 que resulta eliminar uno a uno cada elemento de la población y clasificar después cada dato con la regla construida sin el. Este método se conoce como validación cruzada y conduce a una mejor estimación de error.
Si el número de observaciones es muy alto el costo computacional de la validación cruzada es alto y una solución mas rápida es subdividir la muestra en K grupos iguales y realizar la validación cruzada eliminando un grupo en vez de una observación.

a. Los grupos deben ser mutuamente excluyentes;
b. Las variables discriminantes se construyen en escala de intervalo

Asimismo, la técnica contiene supuestos implícitos:

a. Se cuenta con dos o más grupos y al menos dos casos por grupo;
b. El número de variables discriminantes debe ser menor que el total del número de casos menos dos;
c. Las variables discriminantes deben ser medidas en escala de intervalo;
d. Ninguna variable discriminante debe presentar una combinación lineal respecto a las demás;
e. Las matrices de covarianza de cada grupo deben ser iguales;
f. Cada grupo debe ser diseñado a partir de una población en donde las variables discriminantes presentan una distribución normal multivariada.

El análisis discriminante permite conocer las diferencias entre los grupos y ofrece una media para asignar cualquier caso en el grupo con el que este caso se asemeja más estrechamente. Su interpretación permite conocer qué tanto los grupos difieren, o su capacidad para "discriminar", y para eso se adopta la siguiente fórmula:

$$f_{km} = u_0 + u_1 X_{1m} + u_2 X_{2m} + \ldots\ldots + u_p X_{pm}$$

Donde:

f_{km} = es el valor de la función discriminante canónica para un caso m en el grupo k

X_{um} = es el valor de la variable discriminante X_i para el caso m en el grupo k; y

u_i = son los coeficientes que producen las características deseadas en la función

Se consideran las variables discriminantes como ejes que definen un espacio p-dimensional, y se calcula el "centroide" de cada grupo: un punto imaginario que tienen coordenadas que representan la media del grupo para cada variable y también representan la posición espacial típica para este grupo.

4. Conclusiones.

Como resultado de haber desarrollado este proyecto de tesis tenemos un sistema administrador de proyecciones; se combina con un esquema para establecer preferencias de estudiantes tomando en cuenta sus necesidades y opiniones. Para este sistema nos hemos enfocado en tratar de optimizar el uso de la información que se dispone de los estudiantes como perfiles establecidos.

Con el fin de cumplir con el objetivo de asegurar la independencia de plataforma que debe poseer este proyecto, la solución se implementó en herramientas que aseguraron la portabilidad y escalabilidad del mismo.

El código de la aplicación ha sido diseñado y estructurado de tal manera que facilita la implementación de nuevos módulos al sistema de forma rápida y sencilla.

La interfaz del sistema ha sido diseñada con el fin de brindar al usuario facilidad en la navegación del sitio y en el entendimiento de las tareas que puede realizar.

La base de datos fue desarrollada de tal forma que se logró mantener la consistencia de datos y la integridad relacional de las tablas.
El dividir el proyecto en dos modelos de datos, uno de estudiantes y otro de administrador, permite separar los aspectos relacionados con cualquier usuario. Todo lo relacionado con la presentación de contenidos son aplicados independientemente del usuario.

Los niveles establecidos en la administración de contenidos, aseguraron una división apropiada de las tareas para cada usuario y a la vez proporcionaron la seguridad necesaria para evitar la manipulación de información restringida por parte de usuarios no autorizados.

Es recomendable generar un módulo que recorra todos los registros que se encuentren inactivos e inutilizados durante un determinado período, para eliminarlos físicamente de la base de datos, puesto que en esta tesis se utilizó solo el concepto de eliminación lógica de registros.

Para versiones futuras, se recomienda el uso de procedimientos estándares de seguridad en la transferencia de los registros que están siendo sincronizados, con el fin de que no se transmitan como texto plano, debido a que esto podría ser vulnerable a ataques extremos.

5. BIBLIOGRAFÍA

[1] Piatetski-Shapiro et al., 1991; Chen et al., 1996; Mannila, 1997.

[1] Richard Creeth, Nigel Pendse, Winkipedia, Enciclopedia Libre, http://es.wikipedia.org/wiki/Data_mining . C-27-IN-6012-010 Documento básico de minería de datos.

[3] T. Hastie, R. Tibshirani, J. Friedman "The elements of Statistical Learning: Data Mining, Inference, and Prediction".

[4] Michalski, R. Bratko, I. Kubat, M eds.1998. Machine Learning and Data Mining, Methods and Applications, John Wiley & Sons Ltd, West Sussex, Inglaterra.

[5] Reportes Técnicos en Ingeniería del Software. 7(1): 26-29, ISSN 1667-5002. © CAPIS-EPG-ITBA,(http://www.itba.edu.ar/capis/rtis) 28.

[6] Luis R. Rivera Fernández, Inteligencia de negocio y bodegas de datos,

[7] Hand, David; Mannila, Heikki; Smyth, Padric (2001). Principles of

data mining. Cambridge: Massachussetts Institute of Technology.

[7] http://www.daedalus.es, C-27-IN-6012-010 - Noviembre de 2002, DAEDALUS – Data, Decisions and Language, S.A.

[8] Michalski, R.S., Bratko, I., Kubat M. 1998. Machine Learning and Data Mining.

[9] Methods and Applications. Wiley & Sons Ltd., EE.UU

[10] Jhon Wiley Alan Simon and Sons. Data Warehouse, Data Mining and OLAP. USA, 1997.

[11] Stephen Haag et al.. Management Information Systems for the information age, págs. pp 28. ISBN 0-07-095569-7.

[12] Xingquan Zhu, Ian Davidson (2007). Knowledge Discovery and Data Mining: Challenges and Realities, págs. pp 18. Hershey, New Your. ISBN 978-1-59904-252-7.

[13] http://users.dsic.upv.es/~flip/LibroMD/, Información asociada al libro "Introducción a la Minería de Datos". (C) José Hernández Orallo, M.José Ramírez Quintana, César Ferri Ramírez. Editorial Pearson, 2004. ISBN: 84 205 4091 9.