Métodos de Investigación
Colegio Marymount
Mayo 26, 2005

Método Causal Probabilístico para Predicción de Área Académica

Rodrigo Rubio, Rodrigo Salmón, Diego Reyes y Miguel Pardinas

Resumen

La razón por la que comenzamos este proyecto, fue la de construir un modelo que ayude a alumnos de segundo de preparatoria a decidir a que área es probable se dirija. Utilizamos el programa "Elvira" de computación con el lenguaje Java. En el programa se modelaron los diferentes factores que afectan la toma de decisión de área académica de un alumno de segundo de preparatoria. Se elaboraron encuestas en las que cada pregunta pertenecía a una variable o factor que afecta la toma de decisión del alumno. Se elaboró una base de datos en *Wordpad* conteniendo las variables de las encuestas. El programa "Elvira" automáticamente procesó esta base de datos y modeló una red bayesiana (modelo gráfico que representa la relación entre un conjunto de variables aleatorias) que simulaba el efecto de cada variable sobre la otra variable. Por medio de gráficas se obtiene la probabilidad que el encuestado se dirija a un área académica determinada. La función del proyecto es encontrar la probabilidad de que un alumno de segundo de preparatoria se dirija a un área académica determinada.

Introducción y Antecedentes

Nosotros elegimos este proyecto porque nos identificábamos más con el área de físico matemático, ya que nosotros estamos cursándola en estos momentos. Elegimos el proyecto ya que queríamos ver que factores influencian en la decisión de un estudiante. Además de ver factible si existe un modelo que prediga el área académica que va a elegir un estudiante.

El modelo causal probabilístico representa las relaciones causales entre variables. Estas relaciones están cuantificadas en base a la probabilidad. Nuestras variables comprenden los factores externos e internos que afectan la decisión de elección de área académica de un alumno de preparatoria.

Las redes bayesianas surgieron en la década de los 80 como modelo probabilístico para el razonamiento con incertidumbre en inteligencia artificial. En pocos años experimentaron una notable expansión: formandose grupos especializados en las universidades más importantes (UCLA, Stanford, MIT, Carnegie-Mellon...) y en las grandes compañías (IBM, Microsoft, Digital, etc.). (1)

Una red bayesiana es un modelo gráfico que representa las dependencias entre un conjunto de variables aleatorias. Las redes bayesianas sirven para la predicción y el diagnostico de opciones. También sirve para construir modelos de cualquier dominio en el que haya dependencias de variables. (2)

Para nuestro proyecto, una red bayesiana es útil ya que relaciona todas las variables en una forma de árbol que se afectan mutuamente.

Lo que se necesita en una red bayesiana para hacer un análisis estadístico o para captar la información que nosotros buscamos en nuestro experimento son las probabilidades condicionales, determinadas por los datos que introduciremos en el programa.

Actualmente el área que elige un preparatoriano se determina con dos tipos de factores:

1.-Factores externos- rendimiento académico, carrera de los familiares, estereotipos, factor social, económico, género, campo de trabajo.

2.- Factores internos- intereses, valores, aptitudes, rasgos en la personalidad, área que se le facilita al estudiante.

Hipótesis

El modelo causal probabilístico predice el área académica que un alumno de segundo de preparatoria elige.

Objetivos

1. Construir un modelo que predice el área que un alumno de segundo de preparatoria elige, con la ayuda de una encuesta.

2. Con la ayuda de una segunda encuesta, validar el modelo y comprobar que éste predice correctamente el área que un alumno elige.

Justificación

Con los resultados de este proyecto nosotros queremos ver que factores influencian en la decisión de área de un estudiante y ver que tan factible es que un modelo prediga el área que un alumno de segundo de preparatoria elige. Ya que un modelo que prediga el área puede llegar a ser muy útil y se podría emplear para asesorar a alumnos de segundo de preparatoria.

Materiales y Equipo

- Programa "Elvira" .0.11. (Graphical User Interface)

En España surgieron grupos de investigadores en varias universidades, los cuales decidieron unirse para solicitar un *Proyecto Coordinado de I+D* financiado por la CICYT, que se desarrollo entre los años 1997 y 2000. En él participaron 25 profesores de 8 universidades españolas, agrupados en cuatro subproyectos: Granada, Almería, País Vasco y UNED. El programa resultante se llamó ***Elvira***, tomando el antiguo nombre de la ciudad de Granada, a cuya Universidad están vinculados en mayor o menor medida varios de los investigadores del proyecto. Por la misma razón, el proyecto de la CICYT asociado fue denominado entre sus participantes como ***Proyecto Elvira***. (1)

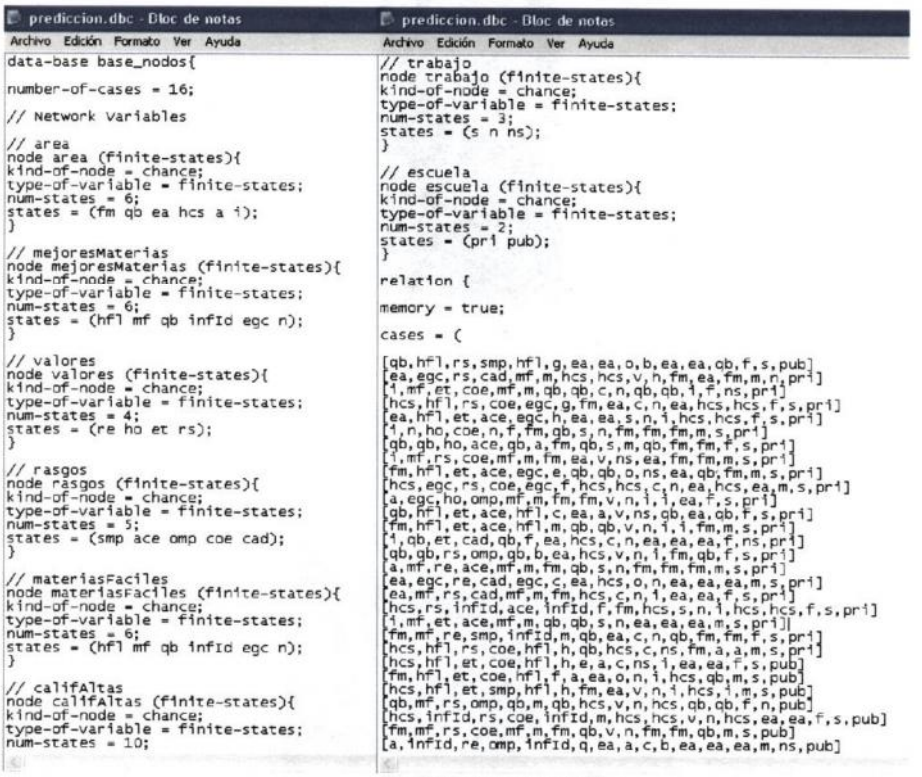
- Computadora. Una PC de disco local. Con un sistema de archivos NTFS. Con un espacio utilizado de 10.1 GB, y un espacio libre de 9.35 GB

- Encuestas (Anexo 1)

Metodología

- Realización de las encuestas y elaboración de éstas en forma de código:

En una hoja de *Wordpad* en forma de código (fig.1) modelamos 50 de las 60 encuestas, cada una de ellas con sus respectivos resultados.



prediccion.dbc - Bloc de notas

Archivo Edición Formato Ver Ayuda

```
data-base base_nodos{

number-of-cases = 16;

// Network Variables

// area
node area (finite-states){
kind-of-node = chance;
type-of-variable = finite-states;
num-states = 6;
states = (fm qb ea hcs a i);
}

// mejoresMaterias
node mejoresMaterias (finite-states){
kind-of-node = chance;
type-of-variable = finite-states;
num-states = 6;
states = (hfl mf qb infId egc n);
}

// valores
node valores (finite-states){
kind-of-node = chance;
type-of-variable = finite-states;
num-states = 4;
states = (re ho et rs);
}

// rasgos
node rasgos (finite-states){
kind-of-node = chance;
type-of-variable = finite-states;
num-states = 5;
states = (smp ace omp coe cad);
}

// materiasFaciles
node materiasFaciles (finite-states){
kind-of-node = chance;
type-of-variable = finite-states;
num-states = 6;
states = (hfl mf qb infId egc n);
}

// califAltas
node califAltas (finite-states){
kind-of-node = chance;
type-of-variable = finite-states;
num-states = 10;
```

prediccion.dbc - Bloc de notas

Archivo Edición Formato Ver Ayuda

```
// trabajo
node trabajo (finite-states){
kind-of-node = chance;
type-of-variable = finite-states;
num-states = 3;
states = (s n ns);
}

// escuela
node escuela (finite-states){
kind-of-node = chance;
type-of-variable = finite-states;
num-states = 2;
states = (pri pub);
}

relation (

memory = true;

cases = (

[qb,hfl,rs,smp,hfl,g,ea,ea,o,b,ea,ea,qb,f,s,pub]
[ea,egc,rs,cad,mf,m,hcs,hcs,v,n,fm,ea,fm,m,n,pri]
[i,mf,et,coe,mf,m,qb,qb,c,n,qb,qb,i,f,ns,pri]
[hcs,hfl,rs,coe,egc,g,fm,ea,c,n,ea,hcs,hcs,f,s,pri]
[ea,hfl,et,ace,egc,h,ea,ea,s,n,i,hcs,hcs,f,s,pri]
[i,n,ho,coe,n,f,fm,qb,s,n,fm,fm,fm,m,s,pri]
[qb,qb,ho,ace,qb,a,fm,qb,s,m,qb,fm,fm,f,s,pri]
[i,mf,rs,coe,mf,m,fm,ea,v,ns,ea,fm,fm,m,s,pri]
[fm,hfl,et,ace,egc,e,qb,qb,o,ns,ea,qb,fm,m,s,pri]
[hcs,egc,rs,coe,egc,f,hcs,hcs,c,n,ea,hcs,ea,m,s,pri]
[a,egc,ho,omp,mf,m,fm,fm,v,n,i,i,ea,f,s,pri]
[qb,hfl,et,ace,hfl,c,ea,a,v,ns,qb,ea,qb,f,s,pri]
[fm,hfl,et,ace,hfl,m,qb,qb,v,n,i,i,fm,m,s,pri]
[i,qb,et,cad,qb,f,ea,hcs,c,n,ea,ea,ea,f,ns,pri]
[qb,qb,rs,omp,qb,b,ea,hcs,v,n,i,fm,qb,f,s,pri]
[a,mf,re,ace,mf,m,fm,qb,s,n,fm,fm,fm,m,s,pri]
[ea,egc,re,cad,egc,c,ea,hcs,o,n,ea,ea,ea,m,s,pri]
[ea,mf,rs,cad,mf,m,fm,hcs,c,n,i,ea,ea,f,s,pri]
[hcs,rs,infId,ace,infId,f,fm,hcs,s,n,i,hcs,hcs,f,s,pri]
[i,mf,et,ace,mf,m,qb,qb,s,n,ea,ea,ea,m,s,pri]]
[fm,mf,re,smp,infId,m,qb,ea,c,n,qb,fm,fm,f,s,pri]
[hcs,hfl,rs,coe,hfl,h,qb,hcs,c,ns,fm,a,a,m,s,pri]
[hcs,hfl,et,coe,hfl,h,e,a,c,ns,i,ea,ea,f,s,pub]
[fm,hfl,et,coe,hfl,f,a,ea,o,n,i,hcs,qb,m,s,pub]
[hcs,hfl,et,smp,hfl,h,fm,ea,v,n,i,hcs,i,m,s,pub]
[qb,mf,rs,omp,qb,m,qb,hcs,v,n,hcs,qb,qb,f,n,pub]
[hcs,infId,rs,coe,infId,m,hcs,hcs,v,n,hcs,ea,ea,f,s,pub]
[fm,mf,rs,coe,mf,m,fm,qb,v,n,fm,fm,qb,m,s,pub]
[a,infId,re,omp,infId,q,ea,a,c,b,ea,ea,ea,m,ns,pub]
```

Fig.1 Encuestas en código

- Introducción de la base de datos al programa "Elvira"

El contenido de las encuestas son las variables independientes que son los factores externos (rendimiento académico, carrera de los familiares, estereotipos, factor social, económico, género, campo de trabajo) y lo que son los factores internos (intereses, valores, aptitudes, rasgos en la personalidad, área que se le facilita al estudiante).

El programa "Elvira" al conocer algunas de las variables, estima las probabilidades de las demás variables y una vez introducidas las variables "Elvira" basado en el *teorema de bayes* (Anexo 2) calcula las probabilidades de todas las variables no conocidas. (Fig.2). *El teorema de bayes* parte de una situación en la cual es posible conocer las probabilidades de que ocurran en una serie de eventos *Ai*. A esta se le aumenta una variable *B* cuya salida proporciona cierta información, ya que las probabilidades de salida *B* son distintas según el suceso *Ai* que haya pasado. Conociendo que ha ocurrido el suceso *B*, la fórmula del teorema de Bayes nos indica como modifica esta información las probabilidades de los sucesos *A.i.* (3)

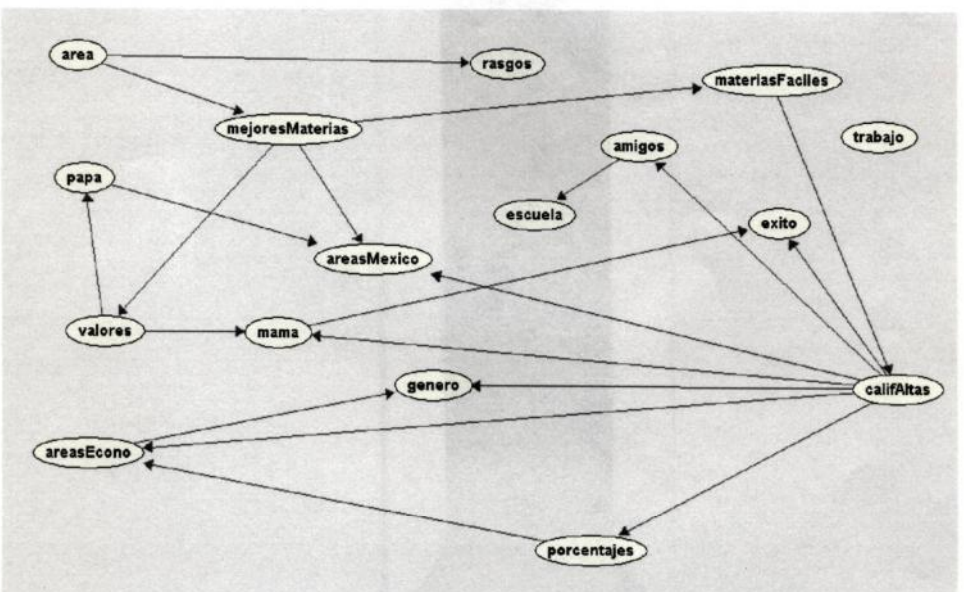


Fig. 2. Red Bayesiana del Proyecto

El programa jerarquizó las variables que son más influyentes en la decisión del alumno (Ej. Calif Altas) y otras de las variables se conectan a ésta en forma de ramificaciones para así llegar al resultado. Con las diez encuestas sobrantes se compararon con las respuestas de la primera pregunta de la encuesta (Anexo 1) para ver si realmente predecía el área.

Resultados

Los resultados que se obtuvieron (Fig. 3.) pudimos observar que la hipótesis sí fue correcta ya que predice el área de una manera probabilística. (*El modelo causal probabilístico predice el área académica que un alumno de segundo de preparatoria elige*). El modelo predice probabilísticamente que área tiene mayor probabilidad para

que el alumno la ingrese. El hecho de que el modelo nos de probabilidades es porque es un modelo probabilístico y lo que nosotros vimos fue si el área coincidía con la que el alumno quería.

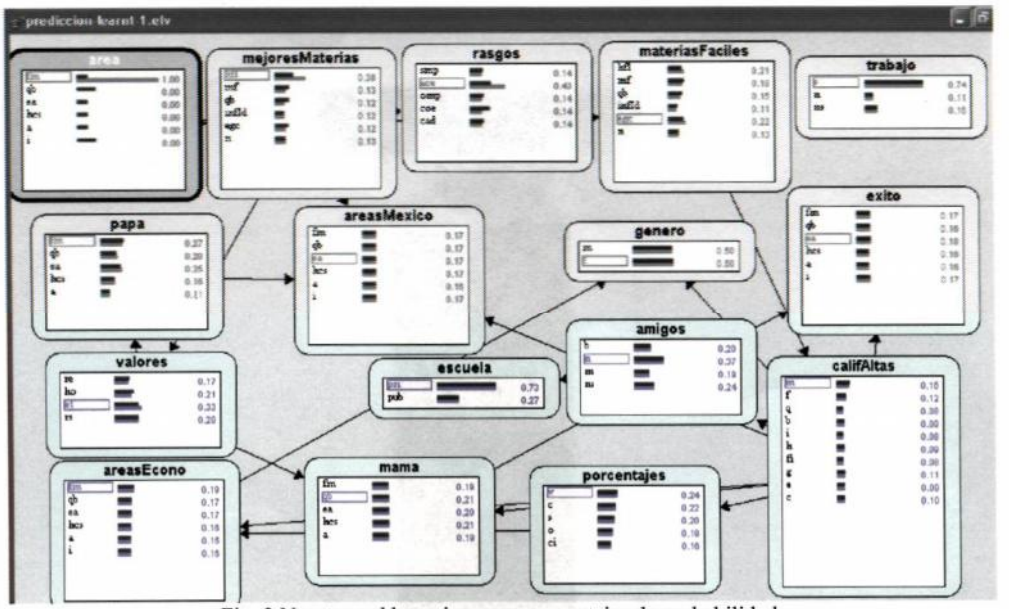


Fig. 3 Nuestra red bayesiana con porcentajes de probabilidad.

En la primera pregunta de la encuesta preguntamos qué área era a la cual éste se inclinaría más para estudiar. En nuestros resultados, el área con la probabilidad más alta es comparada con el área que el alumno de segundo de preparatoria haya puesto en la encuesta previamente. Fue así que, con esa pregunta (y con la probabilidad que salió en el programa) nosotros validamos nuestro proyecto. Sacando diez encuestas al azar, las modelamos una por una en la red que nosotros diseñamos (Fig.3.). Cada resultado de la encuesta lo comparábamos con la primera pregunta de la encuesta (Anexo 1). Y así pudimos observar que el 86 % de las encuestas si nos daban igual a la primera pregunta.

Conclusión

Lo que hace nuestro modelo de red bayesiana en el programa "Elvira" es predecir probabilísticamente el área académica que un alumno de segundo de preparatoria va a elegir, con un porcentaje de acierto del 86 %. Nos dimos cuenta que el programa tiene un porcentaje del 14 % de error.

Agradecimientos

Agradecemos al Dr. Enrique Sucar, a Alberto Reyes del ITESM Campus Cuernavaca y a la Maestra Verónica Ortega de Orientación Vocacional del Colegio Marymount por su apoyo a lo largo de nuestro trabajo de investigación.

Bibliografía

1) "Razonamiento Bayesiano." (09 Diciembre 2004) Díez J.
http://www.ia.uned.es/~fjdiez/docencia/razbayes/
Fecha de consulta: 07 de Febrero 2005.

2) "Bayesian Networks and Decision-Theoretical Reasoning for Artificial Intelligence."
Breese J. and Koller D. http://research.microsoft.com/users/breese/tutorial/sld001.htm
Fecha de consulta: 03 de Febrero 2005.

3) Probabilistic Reasoning in Expert Systems, R.E. Neapolitan. Wiley, 1990

Anexo 1:

Encuesta Para Predicción de Área de Estudio
Onceavos

1. ¿Qué área académica te interesa más? (Aunque no se la que más se te facilite)
   a. Físico-matemático
   b. Químico-biológico
   c. Económico Administrativo
   d. Humanidades y Ciencias Sociales
   e. Artes
   f. Indeciso
2. Jerarquiza los siguientes valores, 4 como el más alto.
   a. Respeto
   b. Honestidad
   c. Ética
   d. Responsabilidad
3. ¿Qué materias se te facilitan más?
   a. Historia, filosofía, literatura
   b. Matemáticas, Física
   c. Química, Biología
   d. Informática e idiomas
   e. Estructura Socioeconómica de México, Geografía
   f. Ninguna de las anteriores
4. ¿Qué rasgos crees que representes mejor tu responsabilidad?
   a. Sistemático, meticuloso, perseverante
   b. Analítico, creativo, expresivo
   c. Creativo, observador, emotivo
   d. Organizado, metódico, persuasivo
   e. Comunicativo, asertivo, disciplinado
5. ¿Qué materias se te facilitan más en lo académico?
   a. Historia, filosofía, literatura
   b. Matemáticas, Física
   c. Química, Biología
   d. Informática e idiomas
   e. Estructura Socioeconómica de México, Geografía
   f. Ninguna de las anteriores
6. ¿Cuáles son tus calificaciones más altas en lo que llevas de preparatoria? (Siendo 10 tu calificación más alta, elije tres)
   a. Matemáticas
   b. Física
   c. Química
   d. Biología
   e. Idiomas
   f. Historia
   g. Filosofía
   h. Geografía
   i. Estructura Socioeconómica de México
7. ¿En que área se especializó tu papa?
   a. Físico-matemático
   b. Químico-biológico
   c. Económico Administrativo
   d. Humanidades y Ciencias Sociales
   e. Artes
8. ¿En que área se especializó tu mama?
   a. Físico-matemático
   b. Químico-biológico
   c. Económico Administrativo
   d. Humanidades y Ciencias Sociales
   e. Artes
9. ¿Qué porcentaje de tus amigos estudiarán la misma área que tu?
   a. 20%
   b. 40%
   c. 60%
   d. 80%
   e. 100%
10. ¿Cómo te sentirías si tus amigos no estudian la misma área que tu?
    a. Bien
    b. No me importa
    c. Mal
    d. No se

11. ¿En que área crees que una persona pueda ser más exitosa?
    a. Físico-matemático
    b. Químico-biológico
    c. Económico Administrativo
    d. Humanidades y Ciencias Sociales
    e. Artes
    f. Indeciso
12. ¿Cuáles son las áreas más importantes en México para ti?
    a. Físico-matemático
    b. Químico-biológico
    c. Económico Administrativo
    d. Humanidades y Ciencias Sociales
    e. Artes
    f. Indeciso
13. ¿Qué áreas te atraen económicamente, o crees que te pueda ir mejor?
    a. Físico-matemático
    b. Químico-biológico
    c. Económico Administrativo
    d. Humanidades y Ciencias Sociales
    e. Artes
    f. Indeciso

14. ¿En tu área crees que influya el género; masculino o femenino?
    a. Si
    b. No
    c. No se

15. ¿En el área que quieras elegir, crees que puedas trabajar en algo relacionado a esto?
    a. Si
    b. No
    c. No se

Anexo 2:

$$p(A_i|B) = \frac{p(B|A_i)p(A_i)}{\sum_{j=1}^{n} p(B|A_j)p(A_j)} \quad \text{para } i = 1,\ldots n$$

Sugerencias y comentarios de nuestro asesor:

Nuestro asesor, el Dr. Enrique Sucar nos comentó que debíamos ahora validar el proyecto de la manera en que lo explicamos, o sea tomando diez de las encuestas y comparando si el resultado de la primera pregunta coincidía con los resultados del programa. Así sacamos el porcentaje de qué tan preciso era el modelo.

También nos comentó que nuestra hipótesis no estaba incorrecta como nosotros lo habíamos comentado, sino que el modelo predice, pero predice probabilísticamente los resultados.