ELVIRA II: APLICACIONES DE LOS MODELOS GRÁFICOS PROBABILÍSTICOS

Aprendizaje de modelos gráficos probabilísticos. Aplicación al clustering con datos de expresión genética

TIC 2001-2973-C05-03

Departamento de Ciencias de la Computación e Inteligencia Artificial

Universidad del País Vasco

http://www.sc.ehu.es/isg/

Albacete · 18 de Junio 2002

A. ELVIRA: Implementación y desarrollo

A.3 Preprocesamiento. Selección y transformación de variables

- 1. Selección indirecta, selección directa, selección híbrida
- 2. Transformación de variables via construcción inductiva, via cambio de variables
- 3. Imputación en bases de datos incompletas
- 4. Categorización

M1–M12: Estudio

M13–M18: Implementación

A. ELVIRA: Implementación y desarrollo

A.4 Aprendizaje

- Aprendizaje con variables Gaussianas
  M1–M9: Estudio
  M10–M18: Implementación
- Aprendizaje de redes híbridas
  M10–M15: Estudio
  M16–M24: Implementación

1. Selección indirecta, selección directa, selección híbrida

- Motivación:
  - Grandes bases de datos con información redundante y/o irrelevante
  - Construir modelos con mejores prestaciones, mediante la selección de las variables adecuadas
- Tres aproximaciones:
  - Selección indirecta
  - Selección directa
  - Selección híbrida

1. Selección indirecta (*filter approach*)

- Las variables se seleccionan, una vez ordenadas, en base a una medida que:
  - Tiene en cuenta las propiedades intrínsecas de los datos
  - No tiene en cuenta el algoritmo de clasificación supervisada (o no supervisada) a utilizar con posterioridad
- El algoritmo de clasificación supervisada (naive–Bayes, C4.5, k–NN, Ripper, ...) o de clasificación no supervisada (k–means, jerárquico ascendente, ...) se aplica a las variables seleccionadas

1. Selección indirecta (*filter approach*)

Clasificación supervisada

- M. Ben–Bassat (1982) Use of distance measures, information measures and error bounds in feature evaluation. *Handbook of Statistics*, Vol. 2, 773–791.
- H. Liu, H. Motoda (1998) *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers.

1. Selección indirecta (*filter approach*)

Clasificación supervisada

- Medidas de información (*information gain*)

$$IG(X_i, C) = -\sum_{j=1}^{n_c} p(C = c_j) log_2 p(C = c_j)$$

$$+\sum_{r=1}^{n_i} p(X_i = x_i^r) \sum_{j=1}^{n_c} p(C = c_j \mid X_i = x_i^r) log_2 p(C = c_j \mid X_i = x_i^r)$$

Mide la reducción en incertidumbre en C motivada por el conocimiento de la variable X_i

1. Selección indirecta (*filter approach*)

Clasificación supervisada

- Medidas de distancia, separabilidad, divergencia o discriminación

$$D(p(X_i \mid C = c_1), p(X_i \mid C = c_2))$$

$$D(p(X_i), p(X_i \mid C = c_1)) + D(p(X_i), p(X_i \mid C = c_2))$$

Distancia entre dos distribuciones de probabilidad a posteriori (dada la clase) o entre dos distribuciones de probabilidad una a priori, la otra a posteriori

1. Selección indirecta (*filter approach*)

Clasificación supervisada

- Medidas de dependencia, de asociación o de correlación

$$I(X_i, C) = \sum_{r=1}^{n_i} \sum_{j=1}^{n_c} p(X_i = x_i^r, C = c_j) \log_2 \frac{p(X_i = x_i^r, C = c_j)}{p(X_i = x_i^r) \cdot p(C = c_j)}$$

Cantidad de información mútua

1. Selección indirecta (*filter approach*)

Clasificación no supervisada. Criterio Univariante

- Medida de relevancia de una variable

$$R(X_i) = \sum_{j=1, j\neq i}^{n} \frac{-N \log(1 - r^2_{ij|resto})}{n-1}$$

$r_{ij|resto}$ coeficiente de correlación parcial de X_i y X_j ajustado por el resto de las variables. $R(X_i)$ medida de relevancia de X_i. Cuanto mayor sea más relevante es la variable

- J. M. Peña, J. A. Lozano, P. Larrañaga, I. Inza (2001) Dimensionality reduction in unsupervised learning of conditional Gaussian networks. *IEEE PAMI*, Vol. 23, No. 6, 590–603

1. Selección indirecta (*filter approach*)

Clasificación no supervisada. Criterio Multivariante

- Seleccionar el conjunto S^* de variables que haga que:

$$S^* = \arg\min_S D(M(N,N), D(M_S(N,N))$$

con

- $D(M(N,N))$ matriz de distancias entre los N casos teniendo en cuenta todas las variables
- $D(M_S(N,N))$ matriz de distancias entre los N casos teniendo en cuenta las variables seleccionadas

1. Selección directa (*wrapper approach*)

Clasificación supervisada

- Ligado a un método de clasificación predefinido
- La selección de variables vista como un proceso de búsqueda
- Cardinalidad del espacio de búsqueda: 2^n
- Cada subconjunto de variables candidato se evalúa por medio de un método de validación (*K-fold, boostrapping, ...*)
- Búsqueda: voraz (hacia adelante, hacia at´ás), estocástica

1. Selección directa (*wrapper approach*)

Clasificación no supervisada

- Planteamiento del problema
  - Encontrar el subconjunto de variables que nos proporcione un *clustering* lo más parecido posible al que se obtiene con todas las variables
- Representación
  - Cada subconjunto como un array binario de dimensión n
- Bondad de cada selección
  - Porcentaje de coincidencia entre los *clusterings*

1. Selección híbrida

- 2 fases (*filter + wrapper*)
  - Ordenar las variables según un criterio. Escoger las k mejores ($k < n$). La determinación de k es arbitraria, a no ser que el criterio siga una ley de probabilidad conocida (test de hipótesis)
  - Efectuar una selección *wrapper* sobre las k variables anteriores

1. Selección híbrida

- R. Blanco, P. Larrañaga, I. Inza, B. Sierra (2001) Selection of highly accurate genes for cancer classification by estimation of distribution algorithms. *Workshop of Bayesian Models in Medicine. AIME2001*, 29–34
- S. Das (2001) Filters, wrappers and a hybrid: evaluating feature selection algorithms. *ICML2001*, 74–81
- E. P. Xing, M. I. Jordan, R. M. Karp (2001) Feature selection for high–dimension genomic microarray data. *ICML2001*, 601–608

Construcción inductiva

- Crear variables que sean producto cartesiano de 2 o más variables
- Incorporar dichas variables (*supernodos* en un modelo de clasificación (supervisada o no supervisada) que extienda el naive–Bayes

2. Construcción inductiva. Clasificación supervisada

Pazzani, M. J. (1996) Searching for dependencies in Bayesian classifiers. *Learning from Data: AI and Statistics V. Lecture Notes in Statistics 112*. D. Fisher, H.-J. Lenz (eds.), Springer-Verlag.

- 2 procedimientos voraces de búsqueda:
  - *Backward sequential elimination and joining (BSEJ)*: parte del naive–Bayes con todas las variables, y en cada paso junta o elimina
  - *Forward sequential selection and joining (FSSJ)*: parte de *nada* y en cada paso selecciona o junta
- Scores
  - Porcentaje de bien clasificados
  - Verosimilitud marginal

2. Construcción inductiva. Clasificación no supervisada

Peña, J. M., Lozano, J. A., Larrañaga, P. (1999) Learning Bayesian networks for clustering by means of constructive induction. *Pattern Recognition Letters*, 20 (11–13), 1219–1230

- Adaptación de las ideas de Pazzani al clustering
- Búsqueda: BSEJ, FSSJ
- Score: verosimilitud marginal

2. Cambio de variable

H. Liu, H. Motoda (1998) *Feature extraction, construction and selection. A data mining perspective*. Kluwer Academic Publishers.

- Dado un conjunto de N condiciones:
  - $M-of-N$ representación: será *true* si al menos M de las N condiciones son verificadas por el caso
  - $1-of-N$ representación: disyunción
  - $N-of-N$ representación: conjunción
  - $X-of-N$ representación: será *true* si exactamente X de las N condiciones son verificadas por el caso

3. Imputación de bases de datos incompletas

Algoritmo EM (*Expectation–Maximization*)

- Demspter, A. P., laird, N. M., Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39(1): 1–38
- McLachlan, G. J., Krishnan, T. (1997) The EM Algorithm and Extensions. *John Wiley and Sons*
- Lauritzen S. L. (1995) The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis*, 19: 191–201

3. Algoritmo EM (*Expectation–Maximization*)

- X variable multinomial con cinco posibles valores: x_1, x_2, x_3, x_4, x_5
- $p(X = x_1) = 1/2$
- $p(X = x_2) = \theta/4$
- $p(X = x_3) = (1 - \theta)/4$
- $p(X = x_4) = (1 - \theta)/4$
- $p(X = x_5) = \theta/4$

con $0 < \theta < 1$

3. Algoritmo EM (*Expectation–Maximization*)

- En una muestra de tamaño N, se obtiene que los 5 valores anteriores se han dado: N_1, N_2, N_3, N_4, N_5 veces. $N = \sum_{i=1}^{5} N_i$
- La verosimilitud de la muestra es proporcional a:

$$(1/2)^{N_1}(\theta/4)^{N_2}((1-\theta)/4)^{N_3}((1-\theta)/4)^{N_4}(\theta/4)^{N_5}$$

- El EMV para θ es: $\hat{\theta} = \frac{(N_2+N_5)}{(N_2+N_3+N_4+N_5)}$

3. Algoritmo EM (*Expectation–Maximization*)

- Supongamos ahora que no tenemos N_1 y N_2 y que en su lugar nos dan el valor de $M_{1+2} = N_1 + N_2$. Denotamos: $M_3 = N_3, M_4 = N_4, M_5 = N_5$
- Utilizando el siguiente procedimiento iterativo, podemos buscar una aproximación para $\hat{\theta}$
  - 1. Dado un valor inicial para θ, usamos los valores observados M_{1+2}, M_3, M_4, M_5 para estimar los valores *esperados* para N_1 y N_2:

$$N_1 = M_{1+2} \frac{1/2}{1/2 + \theta/4}$$

$$N_2 = M_{1+2} \frac{\theta/4}{1/2 + \theta/4}$$

  - 2. Usamos la estimación de los datos completos para revisar nuestra estimación sobre θ, por medio de la fórmula de su *estimación* máximo verosímil
  - 3. Alternamos los dos pasos anteriores hasta condición de parada

3. Algoritmo EM (*Expectation–Maximization*)

Partiendo de $\theta = 0{,}5$ y $M_{1+2} = 125, M_3 = 18, M_4 = 20, M_5 = 34$ se obtiene:

Iteración	θ
0	0.500000000
1	0.608247423
2	0.624321051
3	0.626488879
4	0.626777323
5	0.626815632
6	0.626820719
7	0.626821395
8	0.626821484

3. Algoritmo EM (*Expectation–Maximization*)

- Paso de *expectation*: estimar los datos incompletos a partir de sus valores esperados obtenidos a partir de los estimadores actuales de los parámetros
- Paso de *maximization*: usando las complexiones de los datos incompletos, tratarlos como completos y estimar por máxima verosimilitud los parámetros

3. Algoritmo EM (*Expectation–Maximization*)

Propiedades:

- La sucesión de estimadores obtenidos con el EM genera una sucesión de verosimilitudes creciente. Si dicha sucesión está acotada superiormente, converge hacia cierto valor
- Bajo ciertas condiciones de regularidad dicho valor es un mx́imo local de la función de verosimilitud

3. Algoritmo EM (*Expectation–Maximization*)

- En redes Bayesianas el paso E (*expectation*) trae como consecuencia el tener que partir el caso
- Lo habitual suele ser completarlo bien con:
  - El valor modal (versión determinista)
  - Un valor obtenido por simulación (versión estocástica)
- Experiencia en el grupo en la aplicación del EM al problema del clustering con modelos gráficos probabilísticos. Variable *hidden* no observada

3. Algoritmo BC (*Bound and Collapse*)

M. Ramoni, P. Sebastiani (1997) Learning Bayesian networks from incomplete databases. *UAI'97*, 401–408

- Método determinista que actúa en dos fases:
  - Primera fase: acota el conjunto de posibles valores para el parámetro en un intervalo. El máximo y el mínimo valor del mismo deben de ser consistentes con los datos
  - Segunda fase: colapsa el máximo y el mínimo en un punto, por medio de una combinación convexa de dichos puntos extremos

4. Categorización

Clasificación de métodos de discretización desde una triple perspectiva:

- Global versus local
- Supervisados versus no supervisados. En función de que usen o no la variable clase
- Estáticos versus dinámicos. En los estáticos el número de intervalos para cada variable se fija de antemano o bien se determina después de una única lectura de la base de datos

4. Categorización

Fayyad, U. M., Irani, K. B. (1993) Multi–interval discretization of continuous–valued attributes for classification learning. *IJCAI'93*, 1022–1027

- Particionamiento recursivo de mínima entropía
  - Local, supervisado, estático
  - Usa la entropía de la variable clase, calculada en las particiones candidatas, para seleccionar las fronteras de la discretización
  - $T^* = \arg\min \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2)$
  - El método se aplica recursivamente
  - Criterio de parada basado en ganancia de información en conjunción con MDL

4. Categorización

Propuesta: global, no supervisado, estático

- Fijar desde el inicio el número de intervalos para cada variable
- Buscar los puntos de corte de todas las variables al unísono, por medio de un heurístico estocástico
- Evaluar cada posible combinación de puntos de corte por medio de la verosimilitud penalizada que se obtiene al aprender una red Bayesiana (con el algoritmo K2) sobre el fichero de casos categorizados obtenido