Fuzzy modeling and bayesian inference network

Luminita STATE, Bucharest, Romania, radus@sunu.rnc.ro
Catalina COCIANU, Bucharest, Romania, ccocianu@ase.ro
Viorica ŞTEFĂNESCU, Bucharest, Romania, anton@fmi.unibuc.ro
Panayiotis VLAMOS,Corfu Town, Greece, vlamos@vlamos.com

Data mining is an evolving and growing area of research and involves interdisciplinary research and development encompassing diverse domains. In this age of multimedia data exploration, data mining should no longer be restricted to the mining of knowledge from large volumes of high-dimensional data sets in traditional databases only. The aim of the paper is to present guidelines in fuzzy modeling, fuzzy clustering and the design of Bayesian inference networks.

Keywords: *fuzzy modeling, fuzzy clustering, c-means algorithm, Bayesian networks*

INTRODUCTION

In the old days, system analysts faced many difficulties in finding enough data to feed into their models. The picture has changed and since databases have grown exponentially, ranging in size into the terabytes within these masses of data being hidden information of strategic importance, the reverse picture becomes a daily problem-how to understand the large amount of data we have accumulated over the years. When there are so many trees, how do we draw meaningful conclusions about the forest? Research into statistics, machine learning, and data analysis has been resurrected. Unfortunately, with the amount of data and the complexity of the underlying models, traditional approaches in statistics, machine learning, and traditional data analysis fail to cope with this level of complexity. The need therefore arises for better approaches that are able to handle complex models in a reasonable amount of time.

Data mining is an evolving and growing area of research and development, both in academia as well as in industry. It involves interdisciplinary research and development encompassing diverse domains. In this age of multimedia data exploration, data mining should no longer be restricted to the mining of knowledge from large volumes of high-dimensional data sets in traditional databases only. Researchers need to pay attention to the mining of different datatypes, including numeric and alphanumeric formats, text, images, video, voice, speech, graphics, and also their mixed representations. Fuzzy sets provide the uncertainty handling capability, inherent in human reasoning, while artificial neural networks help incorporate learning to minimize error. Genetic algorithms introduce effective parallel searching in the high-dimensional problem space.

2 CLUSTER ANALYSIS

Cluster analysis is a method of grouping data with similar characteristics into larger units of analysis. Since Zadeh, 1965, first articulated fuzzy set theory which gave rise to the concept of partial membership, based on membership functions, fuzziness has received increasing attention. Fuzzy clustering, which produce overlapping cluster partitions, has been widely studied and applied in various area (Bezdek, 1999). So far, there have been proposed a relatively small number of methods for testing the existence/inexistence of a natural grouping tendency in a data collection, most of them being based on arguments coming from mathematical statistics and heuristic graphical techniques (Panayirci and Dubes,1983, Smith and Jain,1984,Jain and Dubes,1988, Tukey,1977, Everitt,1978).

The data are represented by *p*-dimensional vectors, $X = (x_1,...,x_p)^t$, whose components are the feature values of a specified attributes and the classification is performed against a certain given label set. The classification of a data collection $\aleph = \{X_1,...,X_n\} \subset \Re^p$ corresponds to a labelling strategy of the objects

of $\aleph$.

In the fuzzy approaches, the clusters are represented as fuzzy sets $(u_i, 1 \le i \le c)$, $u_i : \aleph \to [0,1]$, where $u_{ik} = u_i(X_k)$ is the membership degree of X_k to the *i*-th cluster, $1 \le i \le c$, $1 \le k \le n$. A c-fuzzy partition is represented by the matrix $U = \|u_{ik}\| \in M_{c\times n}$. The number of labels *c* has to be selected in advance, the problem of finding the optimal *c* is usually referred as cluster validation. The main types of label vectors are *crisp* N_c, *fuzzy* N_p, *and possibilistic* N_{poz}, defined as follows,

$$N_c = \left\{ y \mid y \in \Re^c, y = (y_1, y_2, ..., y_c), y_i \in \{0,1\}, 1 \le i \le c, \sum_{i=1}^{c} y_i = 1 \right\} = \{e_1, e_2, ..., e_c\} \quad (1)$$

where $(e_i)_j = \delta_{ij} = \begin{cases} 0, i \ne j \\ 1, i = j \end{cases}$ (2)

$$N_p = \left\{ y \in \Re^c \mid y = (y_1, y_2, ..., y_c), \forall i, y_i \in [0,1], \sum_{i=1}^{c} y_i = 1 \right\}, \quad (3)$$

$$N_{poz} = \left\{ y \in \Re^c \mid y = (y_1, y_2, ..., y_c), \forall i, y_i \in [0,1], \exists j, y_j \ne 0 \right\}, \quad (4)$$

Obviously, $N_{pos} \supset N_p \supset N_c$. If we denote by $U = [U_1, ..., U_n] = \|u_{ij}\|$ a partition of $\aleph$, then, according to the types of label vectors, we get the *c*-partition types M_{pos}, M_p and M_c,

$$M_{pos} = \left\{ U \mid U \in M_{c\times n}, U = [U_1, ..., U_n], \forall k, U_k \in N_{pos}, \forall i, \sum_{k=1}^{n} u_{ik} > 0 \right\} \quad (5)$$

$$M_p = \left\{ U \,\middle|\, U \in M_{pos}, \forall k, U_k \in N_p \right\} \quad (6)$$

$$M_c = \left\{ U \,\middle|\, U \in M_p, \forall k, U_k \in N_c \right\} \quad (7)$$

Note that $M_c \subset M_p \subset M_{pos}$.

3 *C-MEANS* MODEL

In fuzzy clustering, the fuzzy c-means clustering algorithms are the best know and most powerful methods used in cluster analysis (Bezdek, 1981).

The variational problem corresponding to *c-means* model is given by

$$\min_{(U,V)} \left\{ J_m(U,V;w) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^m D_{ik}^2 + \sum_{i=1}^{c} w_i \sum_{k=1}^{n} (1-u_{ik})^m \right\} \quad (8)$$

where $U \in M_c/M_p/M_{pos}$, $V = (v_1, ..., v_c) \in M_{c\times p}$, v_i is the centroid of the *i*-th cluster, $w = (w_1, ..., w_c)^T$ is the penalties vector corresponding to the cluster system, $m \ge 1$ is the fuzzyfication degree, and $D_{ik}^2 = \|x_k - v_i\|^2$. Let $(\hat{U}, \hat{V})$ be a solution of (8). Then,

1.The crisp model:

$(U,V) \in M_c \times M_{c\times p}; w_i = 0, 1 \le i \le c$,

$$\hat{u}_{ik} = \begin{cases} 1, D_{ik} \le D_{ij}, i \ne j \\ 0, \text{otherwise} \end{cases} \quad (9)$$

$$\hat{v}_i = \frac{\sum_{k=1}^{n} \hat{u}_{ik} x_k}{\sum_{k=1}^{n} \hat{u}_{ik}}; \quad 1 \le i \le c, 1 \le k \le n \quad (10)$$

2.The fuzzy model:

$(U,V) \in M_p \times M_{c\times p}; m > 1, w_i = 0, 1 \le i \le c$

$$\hat{u}_{ik} = \left[\sum_{j=1}^{c} \left(\frac{D_{ik}}{D_{jk}} \right)^{\frac{2}{m-1}} \right]^{-1}, \qquad \hat{v}_i = \frac{\sum_{k=1}^{n} u_{ik}^m x_k}{\sum_{k=1}^{n} u_{ik}^m}; \quad 1 \le i \le c, 1 \le k \le n \quad (11)$$

3.The possibilistic model:

$(U,V) \in M_{pos} \times M_{c\times p}; \forall i, w_i > 0$,

$$\hat{u}_{ik} = \left[1 + \left(\frac{D_{ik}^2}{w_i} \right)^{\frac{1}{m-1}} \right]^{-1},$$

$$\hat{v}_i = \frac{\sum_{k=1}^{n} u_{ik}^m x_k}{\sum_{k=1}^{n} u_{ik}^m}; \quad 1 \le i \le c, 1 \le k \le n \quad (12)$$

The general scheme of a cluster procedure $\wp$ is ,

t←0

repeat

t←t+1; $U_t \leftarrow F_\wp(V_{t-1})$; $V_t \leftarrow G_\wp(U_{t-1})$

until $(t = T \quad or \quad \|V_t - V_{t-1}\| \le \varepsilon)$

$(U,V) \leftarrow (U_t, V_t)$

where *c* is the given number of clusters, *T* is upper limit on the number of iterations, *m* is the weight parameter, $1 \le m < \infty$, *C* is the ter-

minal condition, w is the system of weights $\forall i, w_i > 0$, $V_0 = (v_{1,0},...,v_{c,0}) \in M_{c\times p}$ is the initial system of centroids and $F_{\wp}$, $G_{\wp}$ are the updating functions.

4 BAYESIAN NETWORKS

The causal relationships are weighted by intensity factors assigned to the edges of a causal network. One of the major limitations of the computational model is that it does not include suitable features in case of feedback relationships, therefore it can be applied only to acyclic causal networks. The structure of a Bayesian network is given by the following components: a directed acyclic graph (DAG), to each variable corresponds a finite set of states such that the set of states are pairwise disjoint. For each variable A, the table of the values corresponding to the condition probabilities $P(A/B_1, B_2,..., B_k)$ is available, where $B_1, B_2,..., B_k$ are the parental variables of A. In case A has no parental variables, then the table P(A) of the values of the probability distribution defined on the state set of A is available.

Because the structure of a Bayesian network is not directly related to causality, that is the links between the variables do not express necessarily causal impact relationships, D-separation properties have to be imposed. In particular, this implies that if A and B are separated on the basis of the evidence ξ, then P(A|ξ)=P(A|B, ξ) holds. Let $U = \{A_1, A_2,..., A_n\}$ the universe of variables. In case the joint probability distribution $P(U) = P(A_1, A_2,..., A_n)$ is known, then, for any evidence ξ, the marginal probability distributions P(A_i) and P(A_i|ξ), $1 \le i \le n$ can be computed.

Obviously, the size of the memory required to retain the values of the probability distribution P(U) depends exponentially on *n*, therefore the design of a Bayesian network assumes finding compact ways to represent the information needed in the evaluation process of the components of the probability distribution P(U) and its marginals. Several remarks can prove useful in such attempts.

Remark 1. Let BN be a Bayesian network for $U = \{A_1, A_2,..., A_n\}$. If we denote by parent(A) the set of parental variables of A, then for each variable A_i, the relation $P(U) = \prod_{i=1}^{n} parent(A_i)$ holds.

Remark 2. If A, B, C are serial connected, then $P(C|A, B) = P(C|B)$.

Remark 3. (convergent connection) If parent(C)={A,B}, then $P(A|B) = P(A)$.

In case additional information is acquired, a Bayesian network allows to update the values of the probability distributions. For instance, assume that the size of the state set of A is n, $P(A) = \{p_1, p_s,..., p_n\}$ and the acquired information is e="the variable A can be either in state *i* or in state *j*". This means that all states of A different from i,j become impossible, hence

$$P(A|e) = \left(0,0,...0,\frac{p_i}{p_i + p_j},0,...,0,\frac{p_j}{p_i + p_j},0,...0\right) \quad (13)$$

The computation of $P(A, e) = (0,0,...0, p_i, 0,...,0, p_j, 0,...0)$ can be carried out as a product of the vectors P(A) and $\left(0,0,...0,\underset{i}{1},0,...,0,\underset{j}{1},0,...0\right)$. Using $P(A|e) = P(A, e)/P(e)$, the a priori distribution P(e) can be computed as, $P(e) = p_i + p_j = \sum P(A, e)$.

If A is a variable having n states, a table with n entries belonging to {0,1} is referred as restriction imposed on A. Obviously, a restriction imposed on a variable expresses the possibility/impossibility of its states. Usually, a restriction is denoted by $\underline{e}$.

If P(U) is the table representing the joint probability distribution and $\underline{e}$ is a restriction imposed on the variable A, we denote by $P(U, \underline{e})$ the table obtained from P(U) by replacing by 0 all components whose counterpart in $\underline{e}$ equals 0. P(U,e)=P(U)$\underline{e}$. Note that $P(e) = \sum_U P(U, e) = \sum_U P(U)\underline{e}$.

Remark 4 If $\underline{e}_1, \underline{e}_2,..., \underline{e}_n$ are restrictions imposed on the variables of a Bayesian network BN of universe U, then

$P(U,e)=\prod_{A\in U} P(A|parent(A))\prod_{i=1}^{n} e_i$ and, for each $A\in U$, $P(A|e)=\frac{\sum_{U-\{A\}} P(U,e)}{P(e)}$ (14)

The updated values of the probability distributions in BN can be computed using the chaining rule to P(U).

REFERENCES

[1] Al Sultan K.S., Selim,S.Z., 1993. Global Algorithm for Fuzzy Clustering Problem, *Patt.Recogn*. 26,1375-1361

[2] Cooper,G.,Herskovits,E. 1992. *A Bayesian Method for the induction of probabilistic networks from data*, Machine Learning

[3] Gath,J., Geva,A.B., 1989. Unsupervised optimal fuzzy clustering, *IEEE Trans. Pattern.Anal.Machine Intell*, 11, 773-781

[4] Huang,C., Shi,Y. 2002. *Towards Efficient Fuzzy Information Processing*, Physica-Verlag, Heidelberg

[5] Jensen,F.V., 2001. *Bayesian Networks and Decision Graphs*, Springer Verlag

[6] Jin,Y., 2003. *Advanced Fuzzy Systems Design and Applications*, Physica-Verlag, Heidelberg

[7] Krishnapuram, R., Keller,J.M., 1993. A possibilistic approach to clustering, *IEEE Trans. Fuzzy Syst.,* 1(2)

[8] Neapolitan,R.E., 2003. *Learning Bayesian Networks*, Illinois University Press

[9] Pal,N.,R., Bezdek,J.C., 1995. On Cluster validity for the Fuzzy c-Means Model, *IEEE Trans. On Fuzzy Syst.* ,Vol. 3,no.3

[10]Pedrycz,W., 2005. *Knowledge-Based Clustering:From Data to Information Granules*, Wiley

[11] State, L., Cocianu, C., Vlamos, P., Stefanescu, V.,2006. *PCA-based Data Mining Probabilistic and Fuzzy Approaches with Applications in Pattern Recognition*, Proceedings of ICSOFT 2006, September 11-14, Setubal, Portugal,

[12] Wu,K-L., Yang,M-S, 2005. A Cluster validity index for fuzzy clustrering, *Patt.Recog.Lett.* 26, 1275-1291

[13] Zahid,N., Abouelala,O., Limouri,M., Essaid,A., 1999. Unsupervised fuzzy clustering, *Patt.Recog.Lett*., 20,1