Fuzzy self-organizing maps for data mining with incomplete data sets

Shidong YU, Hang LI
College of Software
Shenyang Normal University
Shenyang 110034, China
ysd0510@sina.com

Qi XU
College of Physics
Shenyang Normal University
Shenyang 110034, China
xuqiqi1981@163.com

Xianfeng WU
Shenyang CBMP XINDA Banking
Equipment CO.,Ltd.
Shenyang 110010, China
xianfeng@163.com

***Abstract*—Self-organizing maps (SOM) have become a commonly-used cluster analysis technique in data mining. However, SOM are not able to process incomplete data. To build more capability of data mining for SOM, this study proposes an SOM-based fuzzy map model for data mining with incomplete data sets. Using this model, incomplete data are translated into fuzzy data, and are used to generate fuzzy observations. These fuzzy observations, along with observations without missing values, are then used to train the SOM to generate fuzzy maps. Compared with the standard SOM approach, fuzzy maps generated by the proposed method can provide more information for knowledge discovery.**

Keywords-fuzzy clustering; incomplete data; self-organizing maps

I. INTRODUCTION

Data mining is the process of trawling through data in the hope of identifying patterns. Data mining is different from traditional statistical analysis in that it is aimed at finding unsuspected relationships which are of interest or value to the databases owners, or data miners [1]. Due to the large number of dimensionality and the huge volume of data, traditional statistical methods have their limitations in data mining. To meet the challenge of data mining, artificial intelligence techniques have been widely used in data mining [2]. Among those artificial intelligence data mining methods, the self-organizing maps (SOM) method based on Kohonen neural network [3] has become one of the powerful techniques of data mining through cluster analysis. SOM has advantages over statistical and other non-traditional methods of cluster analysis because of their simplicity and relaxation of strict assumptions [4], and are considered as an effective method in dealing with high-dimensional data. More importantly, the SOM method provides a base for the visibility of clusters of high-dimensional data. It shifts cluster analysis from the traditional data-centred approach to the human-machine interaction approach. This unique feature makes data mining more effective in terms of incorporation of human domain interests in cluster analysis.

The standard SOM approach, however, is not able to process input data with missing values. On the other hand, in data mining, it is a rare case where the data set contains entries for all of the variables for each observation. Although there has been a large amount of work on fuzzy SOM (e.g. [5, 6]), the fuzzy SOM models reported in the literature do not address the problem of incomplete data. Also, the feature maps generated by these fuzzy SOM models are virtually crisp. This study is to add an important advantage to the SOM technique for data mining by proposing an SOM-based fuzzy map model that can effectively deal with missing data.

II. FUZZY OBSERVATIONS WITH MISSING VALUES

In data mining, one can rarely find a case where the data set contains entries for all of the variables for each observation. This is especially true when mining survey data sets such as marketing surveys and questionnaires. Commonly, surveys and questionnaires are often only partially completed by respondents. The extent of damage of missing data is unknown when it is virtually impossible to return the survey or questionnaires to the data source for completion, but it is one of the most important parts of knowledge for data mining to discover.

One of the convenient solutions to incomplete data is to eliminate from the data set those records that are missing values. This, however, ignores potentially useful information in those records. In cases where the proportion of missing data is large, the conclusions drawn from the screened data set are more likely to be misleading.

Another simple approach of dealing with missing data is to use generic ‘unknown’ for all missing data items. However, this approach does not provide any fuzzy information that might be useful for interpretation of missing data.

The third solution to dealing with missing data is to estimate the missing value in the data field. In the case of time series data, interpolation based on two adjacent data points that are observed is possible. In general cases, one may use some expected value in the data field based on statistical measures [7, 8]. However, in data mining, survey data are commonly of the types of ranking, category, multiple choices, and binary. Interpolation and use of an expected value for a particular missing data variable in these cases are generally inadequate.

Next, we apply the fuzzy sets theory [9] and develop a fuzzy sets method to treat missing values. Without loss of generality, we assume that an observation $X(x_1, x_2,...x_m)$ represents a discrete event in the sample space, where each variable x_j (j=1, 2... m) takes discrete values. In cases where a variable takes continuous values, one may make bins based on a certain criteria of intervals. The number of bins might affect computational time, but does not particularly influence knowledge discovery in data mining.

978-1-4244-7237-6/10/$26.00 ©2010 IEEE

If an observation without missing data is considered crisp, then an observation with missing data becomes fuzzy due to the uncertainty of the missing value. Nevertheless, in discrete cases, the possible values the missing data variable might take are known. Also, one may estimate the possibility of the missing value for the missing data variable based on general knowledge or distributions of available data. The fuzzy event is then perceived a set of crisp observations weighted on the possibility of the missing values, called fuzzy observations. For instance, suppose we have a fuzzy observation X(x_1, x_2, …, x_c, …, x_m) where x_c is the variable with missing value, and it is known that x_c takes one of the values from{1, 2, 3, 4, 5}. If the available data show that the chances of the five values are equal, then the fuzzy observation can be perceived the sum of five weighted observations: Σ_i 0.2X(x_1, x_2, …, i,⋯, x_m), (i = 1,…,5). These five perceived observations can be used for training the SOM to generate fuzzy maps, as explained later in this article. Next, we use formal descriptions to define fuzzy observations.

Let A be a non-fuzzy event, $A \in \Omega$ where Ω is the data mining sample space. The membership function of A of an observation X(x_1, x_2, …, x_m) without missing values is defined to be 1; that is,

$$X(x_1, x_2, \ldots, x_m) \rightarrow A \qquad (1)$$

and

$$\mu_A(X) = 1 \qquad (2)$$

Suppose the observation has a missing value for the variable x_c, and $x_c \in R$, where R ={r_1, r_2, …, r_p} is the support of x_c. The observation becomes fuzzy; that is

$$X(x_1, x_2, \ldots, x_c, \ldots, x_m) \rightarrow A \sim \qquad (3)$$

where $A \sim$ is a fuzzy event. Let μ_j denote the fuzzy membership of the fuzzy observation belong to a possible crisp observation

$$\mu_j = P_j(x_c = r_j)\ (j = 1, 2, \ldots ,p) \qquad (4)$$

That is, μ_j is the possibility with which the missing data variable takes value r_j, or, in fuzzy sets terms, the fuzzy membership with which the missing data variable belong to r_j. The fuzzy event $A \sim$ can be written as

$$A \sim = \mu_1[A|(x_c = r_1)] \cap \mu_2[A|(x_c = r_2)] \ldots \cap \mu_p[A|(x_c = r_p)] \quad (5)$$

Where μ_j is the fuzzy membership of the fuzzy observation belonging to a possible corresponding crisp observation, and $\Sigma_j \mu_j = 1$. In this study, we use fuzzy membership instead of probability to specify missing data, because the data miner can have his/her subjective judgement on the possible outcomes of the missing values. Our objective of dealing with missing data is to generate self-organizing maps with fuzzy clusters for the data miner.

If an observation has more than one missing data variable, one can always work on one missing data variable each time, and repeat the procedure to generate fuzzy observations for it. Apparently, the more missing data variables an observation has, the fuzzier the generated fuzzy event would be.

To determine μ_j, one must select a fuzzy membership function for the missing data variable based on general knowledge or available data. For instance, suppose a survey question regarding teaching evaluation requires five-point ranks. One might be able to find out a fuzzy membership function, such as
{0.3/5, 0.3/4, 0.2/3, 0.13/2, 0.07/1}
for the missing value based on the complete data records in the database.

Generally, in data mining with discrete survey data, rectangular, trapezoidal, and triangular are common types of fuzzy membership functions for missing values, as shown in Fig. 2.

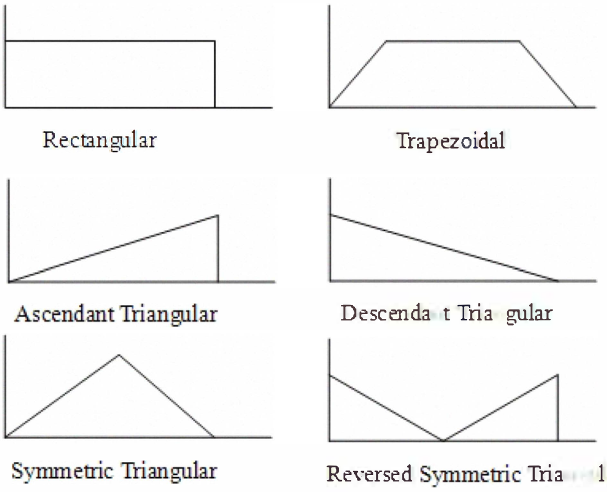

Figure 1. Typical fuzzy membership functions for missing values

We use simple examples to explain the procedure of the generation of fuzzy observations. Suppose we have an observation with missing data X = [1, z, 2], where z is the variable with missing values and z = {0.2/2, 0.8/3}. Then the fuzzy observations for the fuzzy event are $A \sim$ = {0.2[1, 2, 2] and 0.8[1, 3, 2]}.

When more than one variable has missing values, we can work in one missing data variable each time, and repeat the procedure to the next missing data variable to generate a set of fuzzy observations for the fuzzy event. For example, suppose we have an observation with missing data X = [1, z, y], where z and y are the variables with missing values, and z = {0.2/2, 0.8/3} and y ={0.7/4, 0.3/5}. Then the four fuzzy observations for the fuzzy event are $A \sim$ = {0.14[1, 2, 4] and 0.06[1, 2, 5] and 0.56[1, 3, 4] and 0.24[1, 3, 5]}.

Note that the generation of fuzzy observations is not an objective of data mining; rather, this process will make it possible to utilize those observations with missing data, and thus add more information to the data mining than what is provided by the observations without missing data. The selection of fuzzy membership functions for missing values is subject to the data miner's subjective judgement. Discrete probability distributions might provide a base for the selection of fuzzy membership functions; however, the use of probability distributions often introduces biases. Missing information might be more properly handled in the framework of possibility theory. Yet this study has not investigated this issue in any detail.

III. Self-organizing fuzzy maps

The proposed method is to transform observations with missing values to fuzzy observations and then use them to train the SOM in order to generate a fuzzy map. After observations with missing data have been converted into fuzzy observations, each of the generated fuzzy observations becomes an observation with complete data, but possesses its fuzzy membership. All crisp and fuzzy observations are then used to train the SOM. This SOM-based fuzzy map model is different from the ordinary SOM approach. It remembers the fuzziness of an observation if it is derived from a fuzzy observation, and cumulates the value of fuzzy membership for each of the activated output node for the generation of a fuzzy map. This method does not change the training algorithm of SOM *per se*, and the connections of the SOM represent the training observations which are all with complete data. However, after the training, the fuzzy map is depicted based not only on the trained SOM, but also on the fuzzy memberships of these training observations. The procedure of the proposed method is summarised below.

Step 1. Generate fuzzy observations.

(1.1) Let S be the number of available observations, S1 the number of observations without missing values, and S2 the number of observations with missing values. For f-th (f =1…S2) observation with missing values, generate F_f fuzzy observations, and $F_f = \{p_1, p_2, ..., p_q\}$, where p_g (g = 1,…,q) is the number of possible values of x_g, and q is the number of the variables that have missing values. The total number of observations becomes $S' = S1 + (F_1 + F_2 + \cdots + F_{S2})$.

(1.2) Let s be the index of all crisp and fuzzy observations. Each observation for training SOM is X_s (s = 1,2,...,S'). Let A_s be the array which holds the fuzzy membership of each X_s (s = 1,2,...,S'). A_s = 1, if X_s is a crisp observation or $A_s = A$ ~,if X_s is a derived fuzzy observation (see (5)).

Step 2. Initialize the SOM training procedure.

(2.1) Set time t = 0.

(2.2) Set the SOM with m input nodes and N output nodes, where m is the dimensionality of input X, and N is the number of bins of the map. Initialize the weights w_{ij} (t = 0), from input node i (i = 1,2,…, m) to output node j (j = 1, 2,…, N), to small random numbers.

(2.3) Set the initial neighbourhood $v(t = 0) = k$, and k is an arbitrary integer (e.g. $k = N$ in this study).

(2.4) Set the initial learning rate $\eta(t = 0) = p$, where $0< p <1$.

(2.5) Set the index of observations s = 1.

Step 3. Train the SOM.

(3.1) If $s < S'$, set $s = s + 1$; else set s = 1. Present observation X_s.

(3.2) Compute

$$D_J = \sum_{i=1}^{M} (x_i(t) - w_{ij}(t))^2$$

where $x_i(t)$ is the input to node i at time t and $w_{ij}(t)$ is the weight from input node i to output node j at time t.

(3.3) Select the winner output node J such that
$D_J = \min_J \{D_J\}$.

(3.4) Define the set

$$v_J(t) = \{j \mid \max[1,(J - v(t))] \le j \le \min[(J + v(t)), k\}$$

Update weights as follows:

$$w_{ij}(t+1) = [w_{ij}(t) + \eta(t)(x_i(t) - w_{ij}(t))]$$
$$for\ j \in v_J(t), and\ 1 \le i \le M$$

$w_{ij}(t+1) = w_{ij}(t)$ otherwise.

(3.5) If $\eta(t) < \varepsilon$, where ε is a predetermined small positive constant (e.g. 0.0001), then Step 4;
else calculate
$\eta(t) = \delta_1\eta(t)$ and $v(t) = \delta_2 v(t)$
where $0 < \delta_1 < 1$ and $0 < \delta_2 < 1$, and δ_1 and δ_2 are so designed
that $v(t) \rightarrow 1$ when $\eta(t) \rightarrow \varepsilon$;
Set $t = t + 1$; and go to Step (3.1).

Step 4. Generate the fuzzy map.

(4.1) Create a two-dimension co-ordinate. The horizontal axis J has N bins arranged in order of the N output nodes of the SOM. The vertical axis A represents the strength of the generated clusters by the trained SOM.

(4.2) For each observation X_s (s = 1,2,...,S'), perform the following sub-steps.

(4.2.1) Present X_s to the trained SOM. It is then organized by the trained SOM to the output node j (j = 1,…,N).

(4.2.2) Locate bin j on the J axis. On the A direction of bin j, add a black bar (with height = A_s) for the non-fuzzy observation, plus a grey bar (with height = A_s) on the top for a generated fuzzy observation.

As shown in the above fuzzy map generation procedure, the trained SOM along with fuzzy memberships of individual generated observations are used to depict the fuzzy map. To depict crisp clusters as well as fuzzy clusters, one-dimensional maps are used for the trained SOM. The one-dimensional map looks almost the same as histogram in statistics, as shown in Fig. 3. The horizontal axis represents locations of output nodes of the SOM. The height of a bar indicates the number of observations which activate the output node of the SOM at the corresponding location. A black bar indicates the frequency of the appearance of crisp observations (i.e. observations without missing data) at the particular position on the map, and the grey bar on the top of a black bar indicates the cumulated frequency of the weighted observations derived from fuzzy observations at that position. Clusters are then identified on the map.

To illustrate the proposed fuzzy map method for incomplete data, we present a simple example as follows. Suppose we have three real observations $X_1 = [1, 4]$, $X_2 = [2, 4]$, and $X_3 = [z, 4]$, where *z* is a missing value. The variable with the missing value has the following fuzzy membership function:
{0.8/1, 0.2/2}
Then the set observations for training SOM becomes $X_1 =[1, 4]$, $X_2 = [2, 4]$, $X_3 = [1, 4]$, and $X_4 = [2, 4]$. Suppose the trained SOM has two active output nodes: one is for X_1 and X_3, and the other is for X_2 and X_4. Based on the trained SOM, we depict two bars on the histogram-style fuzzy map. Since the fuzzy memberships of these four observations are: $A_1 = 1$, $A_2 = 1$, $A_3 = 0.8$, and $A_4 = 0.2$, one bar is 1.8 unit tall with 1 unit black and 0.8 unit grey on the top, and the other bar is 1.2 unit tall with 1 unit black and 0.2 unit grey on the top.

Figure 3 depicts three hypothetical cases. Cluster A is a crisp cluster which is generated by observations without missing values. Cluster B shows the case where fuzzy observations enlarge a crisp cluster, indicating that the cluster based only on observations without missing data is valid. Cluster C is a very fuzzy cluster. As shown by the black bars, the cluster can hardly be identified based only on crisp observations. The contribution of fuzzy observations to the SOM in this case is the disclosure of the potential clusters.

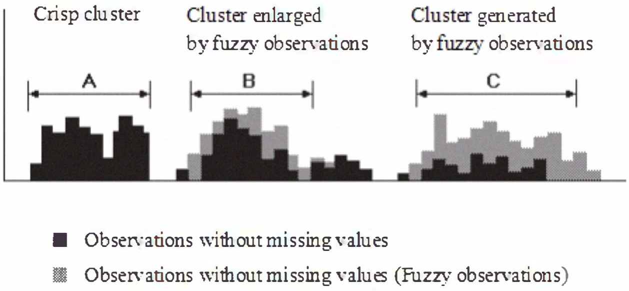

Figure 2. Depicted histogram-style fuzzy map

Clearly, when observations with missing values are transformed to fuzzy observations, the size of the training sample becomes large, depending on the number of missing values and their categories. This transformation approach might not be feasible for statistical methods such as k-means algorithm, but can be applied to SOM because of the efficiency of SOM learning process. The fuzzy clusters can add more information to the non-fuzzy results generated by imputation models (e.g. the one reported in [2]) for missing data. Furthermore, based on these clusters identified on the SOM, one can apply a formal evaluation method, such as k-means algorithm, to evaluate the validity for the found clusters [10].

IV. DATA MINING WITH INCOMPLETE DATA SETS: A CASE STUDY

In this section, we use a real-world example to demonstrate the method described. Student opinion survey methods are widely used at universities to evaluate the teaching performance of professors. The data used in this example came from student's course evaluation at Shenyang Normal University, China. In this case, 20 questions describe the characteristics of a teacher's performance. Each question is rated on a five-point scale by the students. A high rank for a question indicates a positive answer. The sets of multivariate data are used to judge effectiveness of teaching. One task for the teaching centre of the university is to identify problems in teaching so that it can take relevant measures (e.g. education seminars and workshops) to tackle these problems.

Twenty variables consisted of the multivariate dimensions for the SOM. Our sample data of 2820 student evaluations were collected to identify the problems in teaching at this university. Among these session records, 1918 (68%) were complete, and 902 (32%) had missing data. A multivariate linear regression analysis on the complete data set indicated that no linear relationship exists between any of the variables and effective teaching. Thus, the SOM cluster analysis method was applied to identify the problems. The computer program in C# implemented the SOM.

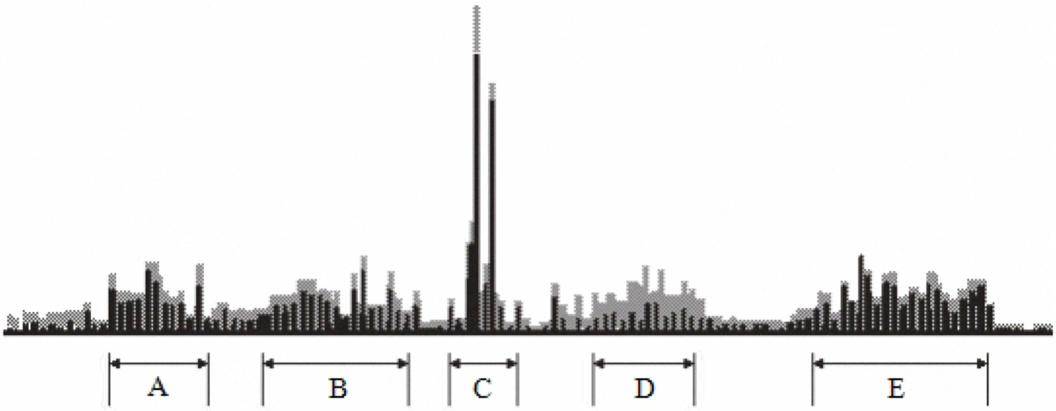

Figure 3. The fuzzy map of the student opinion survey example

TABLE I. A SUMMARY OF THE DATA MINING RESULTS OF THE EXAMPLE

Clusters	Variables receiving low ranks	Interpretation
A	V13: Where appropriate, helpful comments are provided when student work is graded.	Students need more convincing comments on the test results.
	V16: Test and assignments provide adequate feedback on student progress.	Tests and assignments do not provide feedback for students.
B	V15: Tests and assignments are reasonable measures of student learning	Tests and assignments should be better designed
C	None	Excellent teaching.
D (Fuzzy cluster)	V2: The instructor explains difficult concepts clearly and understandably.	Difficult concepts are not well described.
	V20: The text book(s) and course material are useful.	Specifically, the textbooks do not give much help.
E	Nearly all Variables	The class sessions in this group have too many problems in teaching.

General observations indicated that the distributions of five-point ranks were trapezoidal with the peak value at ranks 4 and 5. Hence, trapezoidal fuzzy membership functions with the following values were applied to the variables with missing values in generating fuzzy observations.

{0.30/5, 0.30/4, 0.20/3, 0.13/2, 0.07/1}

The SOM topology was set to 200 output nodes, 200 nodes for the initial neighbourhood, the initial learning rate of 0.01, and 2000 learning iterations. Figure 4 shows the trial result of the fuzzy map. In Fig. 4, the core part (black) represents the data records without missing values, and the fuzzy part (grey) on the top of the core represents the fuzzy observations (i.e. data records with missing values). Using rule of thumb that 'three to five clusters for 80% of points', five clusters (marked A, B, C, D, and E respectively in Fig. 4) were identified on the fuzzy map. Among the five clusters, A, B, C, and E were relatively clear, while D was fuzzy. Cluster D was contributed mainly by fuzzy observations.

The original multivariate data corresponding to the clusters were extracted. The variables with low ranks (below the average) were detected, and problems were then identified, as summarised in Table 1. In cluster C, few variables had low ranks, and thus this cluster represented the group with excellent teaching. In cluster E, virtually every variable had low ranks, and this cluster represented the group with much difficulties in teaching. According to this fuzzy clustering result, two specific problems were considerably clear for improving teaching. That is, teachers often neglected helpful comments on tests and assignments and overlooked the issue of design of tests and assignments. Another problem was relatively fuzzy. This cluster might not be significant if data records with missing values were excluded in training the SOM. The interpretation of this cluster was that many teachers did not fully use (or did not carefully select) textbooks and course material to explain difficult concepts of the course.

V. CONCLUSIONS

In the data mining field, SOM are considered a useful technique in allowing the data miner to view high-dimensional data. Compared with the traditional statistical techniques, SOM are more capable of clustering high-dimensional data, and is one of the effective techniques in data mining. However, the standard SOM approach is not able to process incomplete data. This article proposes an SOM-based fuzzy map model for data mining with incomplete data. This model has two key components: translation of observations with missing data into fuzzy observations, and histogram-style fuzzy maps. Through the real-world case, this article has demonstrated that SOM can be more useful in data mining with incomplete data if the proposed fuzzy model is applied.

REFERENCES

[1] D. J. HAND , "Data mining: Statistics and more?," The American Statistician, Vol. 52, No. 2, May 1998, pp.112–118.

[2] F. Fessant, S. Midenet, "Self-organising map for data imputation and correction in surveys," Neural Computing & Applications, Vol. 10, No. 4, Apr. 2002, pp.300–310.

[3] T. Kohonen, Self-Organization and Associative Memory, 3rd ed., Berlin: Springer-Verlag, 1989.

[4] G. Deboeck, T. Kohonen, Visual Explorations in Finance with Self-Organizing Maps, London, UK: Springer-Verlag, 1998.

[5] P. Vuorimaa, "Fuzzy self-organizing map," Fuzzy Sets and Systems, Vol. 66, Issue 2, Sept. 1994, pp.223–231.

[6] S. Mitra, S. K. Pal, "Self-organizing neural network as a fuzzy classifier," IEEE Transactions on System, Man, and Cybernetics, Vol. 24, No. 3, March 1994, pp.385–399.

[7] A. P. Dempster, N. M. Laird, D.B. Rubin DB, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society, Series B (Methodological), Vol. 39, No. 1, 1997, pp.1–38.

[8] J. A. Hartigan, Clustering Algorithms, New York: Wiley, 1995.

[9] L. A. Zadeh, "Fuzzy sets as a basis for a theory of possibility," Fuzzy Sets and Systems, Vol. 100, 1999, pp.9–34.

[10] J. Vesanto, E. Alhoniemi, "Clustering of the self-organizing map," IEEE Transactions on Neural Network, Vol. 11, No. 3, May 2000, pp.586–600.