Simultaneous Approach of Clustering and Metric Multi-dimensional Scaling Methods with Similarity Data for Medical Diagnosis

Ryo Haruna* and Hiroaki Ishii*
Faculty of Information Science and Technology
Accepted in November 2006
Available online

Abstract

We propose a new regularization fuzzy clustering method for similarity data. It is necessary for a computer supported medical care diagnosis system to judge the group of many diseases by huge medical database consisting of the other items comprehensively. In knowledge acquisition from huge database, it is necessary to extract rules from data divided into the layer by clustering methods more than discovering general feature, and so the clustering analysis method put together from such a viewpoint on the analysis purpose is researched. Our proposed method supports medical treatment to extract the appearance of disease patterns of complication in the diabetes.

Keywords: Similarity data; Metric multi-dimensional scaling; Fuzzy clustering; Entropy regularization

1. Introduction

The detailed appearance of disease mechanism in the patient with diabetes is still hard to clarify though the patient with diabetes is assumed to be an appearance of diabetic with risk factors (environmental factor and genetic factor, etc.) that are related to the disease in complexity. In our research, the disease appearance of disease is focused by analyzing the relation of two risk factors the disease based on a new mathematical principle analysis method, and using these risk factors and it is evaluated (Brandeau et al., 2004).

Based on our special algorithm, our project develops an onset risk diagnosis tool of the diabetes complication and builds the treatment system for the complication prevention corresponding to the diagnosis result. It is very difficult to realize medical diagnosis systems on a computer (Zhang et al., 2004), since in order to judge many sick support groups synthetically one needs to use the data that consists of dozens or more items (the amount of the features). In knowledge acquisition from huge database, it is necessary to classify the database by clustering, and to extract rules from each cluster with more than the general characteristic is discovered. Cluster analysis (Jain et al., 1988) is a technique which discovers the substructure of a data set by dividing it into several clusters. It is also known as one of the data mining (Berry et al., 2004; Adriaans et al., 1996; Dick et al., 2004; Morishita et al., 2001; Lin et al., 2004; Fayyad et al., 1996) approaches (Oh et al., 2001). There have been many researches for cluster analysis. Fuzzy clustering (Klawonn et al., 1997; Frank et al., 2000; Bezdek et al., 1981) is an extension of the cluster analysis, which represents the affiliation of data points to clusters by memberships. Introducing fuzziness to clustering gives us the flexible representations of substructures of the data set.

The method of unifying the clustering analysis method with the metric multi-dimensional scaling, i.e., Quantification Type 4 (Hayashi, 1952 and Kawaguchi, 1998) method has been devised as an application of this. This paper, it explains the analysis technique for clustering by treating only the disease and two categories of the name of a disease from the data base of the diabetes. The technique introduces a fuzzy clustering and using the made similarity according to the layer. In our proposal technique, it aims to extract the appearance of disease pattern of coexisting illness in the diabetes.

2. Preliminaries

2.1 Metric MDS

Multi-dimensional Scaling (Kruskal et al., 1978), initiated by Shepard (1962) under the name proximity analysis, has been developed by Shepard, then Carroll, Kruskal, etc. MDS, like CA (Corresponding Analysis), is concerned by Euclidean representations (Le Roux and Rouanet, 2004). See Benzecri (1965), Shepard (1966, 1980). The affinities between the two methods are well recognized today, and CA is sometimes put under the heading “Multi-dimensional Scaling”. The MDS paradigm is not the contingency table, but the similarity table, therefore for comparing CA and MDS two lines can be followed:

(1) Investigating CA for analyzing similarity tables;
(2) Investigating MDS for analyzing square contingency tables.
The algorithm seeks for configuration of the objects such that the distance between the points, representing the objects, is some monotonic function of the original similarity measure.

### 2.2 Fuzzy Clustering

The Fuzzy c-means (FCM) clustering partitions the data set by introducing the membership to fuzzy clusters. s dimensional vector $v_c$ denotes prototype parameter (i.e., cluster centroid). $u_{ck}$ is membership of the data point $k$ to the cluster $c$ which represents degree of affiliation to clusters. In this algorithm, a partition matrix $U$, that reflects a measure of similarity among the elements (i.e., proximity of the data points to each of the cluster centroid), is defined as follows:

$$U = \begin{bmatrix} u_{c11} & u_{c12} & \cdots & u_{c1n} \\ u_{c21} & u_{c22} & \cdots & u_{c2n} \\ \vdots & \vdots & \ddots & \vdots \\ u_{cn1} & u_{cn2} & \cdots & u_{cn1} \end{bmatrix}.$$ 

The classification criterion is realized by minimizing the following objective function (the performance index) with respect to both the membership values and the cluster centroids:

$$L_{FCM} = \sum_{c=1}^{C} \sum_{i=1}^{n} (u_{ci})^\theta d_{ci}$$

where the weighting exponent $\theta$ is added for the fuzzification of memberships. The larger $\theta$ is, the fuzzier the membership assignments are. When we choose $\theta = 1$, the fuzzy c-means algorithm is a generalization of its historical predecessor, the hard c-means algorithms, where the prototypes are computed by the same formula: the memberships are assigned only one the “hard” values 0 or 1. A common choice of the fuzzifier $\theta = 2$. The clustering algorithm is based on the iterative least squares technique. The non-negative membership $u_{ci}$ satisfies the following condition:

$$\sum_{c=1}^{C} u_{ci} = 1, \quad c = 1, 2, \cdots, C \quad (2)$$

$$d_{ci} = (x_i - v_c)^T A_i^{-1} (x_i - v_c)$$

is a measure of the distance or dis-similarity from $x$ to the $i$-th cluster centroid. The Euclidean distance metric is often used where $A_i$ is a unit matrix.

The optimal $u_{ci}$ and $v_c$ for all $i$ and $c$ are sought the following solution using a fixed-point iteration scheme.

$$u_{ci} = d_{ki}^{-2(\theta-1)} \sum_{j=1}^{C} d_{ji}^{-2(\theta-1)}, \quad (3)$$

$$v_c = \frac{\sum_{i=1}^{n} (u_{ci})^\theta x_i}{\sum_{i=1}^{n} (u_{ci})^\theta} \quad (4)$$

There is one technical trick in the regular FCM. When $x_i$ and $v_c$ assume the same value and the distance $d_{ci}$ between them equals 0, then the membership $u_{ci}$ goes to infinite.

Miyamoto et al., (1997-1999) proposed another approach to fuzzy membership assignments by regularizing the objective function of crisp clustering. The fuzzification technique is called “Regularization by entropy” and the objective function of FCM is defined as follows:

$$L_{FCM2} = \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci} \|x_i - v_c\|^2 + \lambda \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci} \log u_{ci} \quad (5)$$

where the entropy term works like the weighting exponent in the standard FCM algorithm. The larger $\lambda$ is, the fuzzier the memberships are. The updating rules for memberships and cluster centroid are derived as follows:

$$u_{ci} = \frac{\exp \left( -\frac{1}{\lambda} \|x_i - v_c\|^2 \right)}{\sum_{i=1}^{C} \exp \left( -\frac{1}{\lambda} \|x_i - v_c\|^2 \right)} \quad (6)$$

$$v_c = \frac{\sum_{i=1}^{n} u_{ci} x_i}{\sum_{i=1}^{n} u_{ci}} \quad (7)$$

### 3. Regularization Fuzzy Clustering based on Similarity Coefficients

We excluded some missing data from database of diabetes disease, and only two categories are taken into account to the treatment layer. Hamming distance (Shannon, 1949 and Oike et al., 2004) is used as one kind of a similarity. We define the difference of two layers as Hamming distance. The similarity among patients of the each layer is used as relative frequency of the contraction period and the age difference. Classification is done by clustering based on a similarity by Hamming distance, and using a similarity that uses the relative frequency based on difference of between the contraction period and the age.

The quantification type 4 is a kind of MDS method, but the basic concept of data analysis is different from that of metric MDS. The similarity $s_{ij}$ shown in Table 1 represents a quantity but the quantity itself does not have strict meanings and it is somewhat vague.
Therefore the goal is to seek the Euclidian space corresponding to this vague similarity. On the other hand, in the Torgerson’s metric MDS, distance itself has rigid and significant meanings. The procedure is strictly metric in the sense that the aim is to determine a configuration of the \( n \) objects such that the solution distance is identical to the given \( d_0 \) measure for all pairs of object \( i \) and \( j \). If the strict reconfiguration is the object, the metric MDS will surely be better.

Our method of analysis introduces entropy regularization in non-linear component analysis (Ikeda et al., 1999) based on fuzzy clustering and multi-dimensional scaling, the unknown values \( x_i \) and \( x_j \) are given for each individual. In our method, it formulates it as the following optimization problems.

\[
\max \sum_{c=1}^{C} \sum_{i=1}^{N} \sum_{j=1}^{N} u_{ci} u_{cj} s_{ij} (x_i - x_j)^2
\]

(8)

\[
\sum_{c=1}^{C} u_{cj} = 1, \ u_{cj} \in [0, 1], \ i = 1, \cdots, N
\]

(9)

\[
\sum_{c=1}^{C} u_{ij} = 1, \ u_{ij} \in [0, 1], \ j = 1, \cdots, N
\]

\[
\sum_{i=1}^{N} x_i^2 = 1
\]

In eq. (8) and (9), two different memberships are determined for the objective function and constraints. Eq.(8) represents that the objects \( i \) and \( j \) whose \( s_{ij} \) value is large should be given the similar value. If those objects belong to the same cluster \( c \), then both the membership \( u_{ci} \) and \( u_{cj} \) are large, and so thus \( x_i \) and \( x_j \) take similar values. In addition, we modify eq.(8) by introducing entropy regularization .

\[
\max \sum_{c=1}^{C} \sum_{i=1}^{N} \sum_{j=1}^{N} u_{ci} u_{cj} s_{ij} (x_i - x_j)^2 + \eta \sum_{c=1}^{C} \sum_{i=1}^{N} u_{ci} \log u_{ci} + \rho \sum_{c=1}^{C} \sum_{j=1}^{N} u_{cj} \log u_{cj}
\]

(10)

where \( \eta \) and \( \rho \) are regularization coefficients. Eq. (10) is optimized based on constraints of eq. (9).

As the constraints are all equations, this optimization problem can be obtained membership values \((u_{ci}, u_{cj})\) by Lagrangian method. Corresponding Lagrangian function becomes as follows:

\[
L = \sum_{c=1}^{C} \sum_{i=1}^{N} \sum_{j=1}^{N} u_{ci} u_{cj} s_{ij} (x_i - x_j)^2 + \eta \sum_{c=1}^{C} \sum_{i=1}^{N} u_{ci} \log u_{ci} + \rho \sum_{c=1}^{C} \sum_{j=1}^{N} u_{cj} \log u_{cj}
\]

\[
- \sum_{c=1}^{C} \sum_{i=1}^{N} u_{ci} (\sum_{c=1}^{C} \sum_{j=1}^{N} u_{cj} - 1) - \sum_{j=1}^{N} \sum_{c=1}^{C} \sum_{i=1}^{N} u_{cj} (\sum_{c=1}^{C} \sum_{i=1}^{N} u_{ci} - 1)
\]

\[
- \lambda \left( \sum_{i=1}^{N} x_i^2 - 1 \right)
\]

(11)

Here \( \mu_c, \mu_j \) and \( \lambda \) are Lagrangian multipliers. From the necessary conditions for the optimality of the objective function \( L \), i.e., \( \partial L / \partial u_{ci} = 0 \) and \( \partial L / \partial u_{cj} = 0 \), we have the following equations.

\[
\exp \left\{ \frac{-1}{\eta} \sum_{i \neq j} u_{cj} s_{ij} (x_i - x_j)^2 \right\}
\]

(12)

\[
\exp \left\{ \frac{-1}{\eta} \sum_{i \neq j} u_{ij} s_{ij} (x_i - x_j)^2 \right\}
\]

(13)

Furthermore, \((x_i, x_j) \) to similar, from the necessary conditions for the optimality of the objective function \( L \),

\[
\sum_{j=1}^{N} \left( \sum_{c=1}^{C} u_{ci} u_{cj} (s_{ij} + s_{ji}) \right) x_i - x_j = \lambda x_i
\]

\[
\Leftrightarrow \sum_{j=1}^{N} r_{ij} (x_i - x_j) = \lambda x_i \ (i \neq j)
\]

(14)

it translates to the following eigenvalue problem.

\[
(W - \lambda I) x = 0
\]

(14)

where the elements \( w_{ij} \) of matrix \( W \) is
\[ w_{ij} = \begin{cases} r_{ij} & : i \neq j \ 
\sum_{k=1}^{N} r_{ik} & : i = j \ (k \neq i) \end{cases} \quad (15) \]

\[ r_{ij} = \sum_{c=1}^{C} u_{ci} u_{cj} (s_{ij} + s_{ji}) \quad (16) \]

Therefore, variables \((x_i, x_j)\) and membership values \((u_{ci}, u_{cj})\) can be obtained. The solution \(x_i\) and \(x_j\) are components of the eigenvector associated with the largest eigen value in eq.(14). The fixed point repetition algorithm is shown as follows.

Procedures:

Step 1: Set values of parameters \(C, \eta, \rho\) and the positive small number \(\epsilon\). Initialize memberships \((u_{ci}, u_{cj})\) randomly.

\[ \max_{c,i,j} |u_{ci}^{NEW} - u_{ci}^{OLD}| < \epsilon \]

Step 2: Update membership \(u_{cj}\) using eq. (13).
Step 3: Update membership \(u_{ci}\) using eq. (12).
Step 4: If the following conditions, is satisfied, then stop. Otherwise, return to Step 2.

4. Conclusion

In this paper, we have proposed a new entropy regularization fuzzy \(c\)-means clustering model for similarity coefficients data to which the conventional fuzzy clustering could not be applied. Eq.(14), (15) and (16) are similar to fuzzy quantification type 4 (Terano et al., 1992). The metric MDS requires solving eigen value problem which is computationally intractable, but our method needs only simple algebraic calculations.

The processing of the missing value is considered in our method, and as the problem in the future, a similarity to the classification result is defined and reclassified in addition after it classifies it by clustering.

Acknowledgments

Two anonymous referees read carefully this manuscript and helped to improve the presentation. We appreciated for them.

References


