TASK QUARTERLY 8 No 2 (2004), 193–198

BREAST CANCER DIAGNOSIS VIA FUZZY CLUSTERING WITH PARTIAL SUPERVISION

TOMASZ PRZYBYŁA

Division of Biomedical Electronics, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland
tobyla@zeus.polsl.gliwice.pl

(Received 12 September 2003; revised manuscript received 14 February 2004)

Abstract: A new clustering method of fuzzy c-myriad clustering with partial supervision is presented in this paper. The proposed method has been applied to breast cancer diagnosis data obtainted from the University of Wisconsin. The data set contains 699 cases of breast cancer, with each instance described by 10 features.

Keywords: robust method, fuzzy clustering, weighted myriad, partial supervision

1. Introduction

Clustering is a procedure in which objects are distiguished or classified in accordance with their similarity. There is no teacher to provide guidance, hence it is also called unsupervised classification. According to the theory of classification, clustering methods may be treated as classification methods that utilize minimal information about classified objects (their features). A data set partition can be described by a $c \times N$ partition matrix $\mathbf{U}$ (where c is the number of clusters, N is the number of objects) [1], where each element of $\mathbf{U}$ represents the membership of every input object in fuzzy clusters. Clustering results can either be used, after hardening of the partition matrix, as a final partition of the input data or be processed further by a human expert, expert systems and so on.

2. The fuzzy c-myriad clustering method with partial supervision

2.1. Weighted myriads

Let us consider a set of N independent and identically distributed observations, $\mathbf{X} = \{x_1, x_2, \cdots, x_N\}$, and a set of assigned weights, $\mathbf{U} = \{u_1, u_2, \cdots, u_N\}$. A weighted

myriad is a value, $\hat{\Theta}$, that minimizes the weighted myriad objective function defined as follows [2, 3]:

$$\hat{\Theta} = \arg\min_{\Theta\in\mathbb{R}} \sum_{k=1}^{N} \ln[K^2 + u_k(x_k - \Theta)]. \tag{1}$$

The value of a weighted myriad depends on the data set $\mathbf{X}$ and assigned weights $\mathbf{U}$ and on parameter K, called a linearity parameter. Two interesting may occur: first, when the K value tends to infinity ($K \to \infty$), the weighted myriad converges with the weighted mean, that is:

$$\lim_{K\to\infty} \hat{\Theta}_K = \frac{\sum\limits_{k=1}^{N} u_k x_k}{\sum\limits_{k=1}^{N} u_k}, \tag{2}$$

where $\hat{\Theta}_K = \text{myriad}\{u_k \diamondsuit x_k; K\}_{k=1}^{N}$. Second, where the K parameter value tends to zero ($K \to 0$), the weighted myriad is always equal to one of the most frequent values of the input data set.

2.2. The fuzzy c-myriad clustering method

A partition of an input data set can be described by a $c \times N$ matrix (where c is the number of clusters, N is the number of objects), called a partition matrix, in the following form:

$$\mathbf{U} = \begin{bmatrix} u_{11} & \dots & u_{1k} & \dots & u_{1N} \\ u_{21} & \dots & u_{2k} & \dots & u_{2N} \\ \vdots & \dots & \vdots & \dots & \vdots \\ u_{c1} & \dots & u_{ck} & \dots & u_{cN} \end{bmatrix} = [\mathbf{u}_{(1)} \quad \dots \quad \mathbf{u}_{(k)} \quad \dots \quad \mathbf{u}_{(N)}]. \tag{3}$$

For fuzzy clustering methods, the partition matrix is defined as [1]:

$$M_{fcN} = \left\{ U \in [0,1]^{c\times N} \,|\, \sum_{i=1}^{c} u_{ik} = 1, k = 1,2,\dots,N; \sum_{k=1}^{N} u_{ik} > 0, i = 1,2\dots,c \right\}. \tag{4}$$

A set of N objects, $\mathbf{O} = \{o_1, o_2, \dots, o_N\}$, is described by a set of N feature vectors, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$, in a p-dimensional feature space (where p is the number of features bescribing each object). The sum of squared errors has been choosen as an objective function for the proposed method:

$$J_m(\mathbf{U}) = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik}^m \|\mathbf{x}_k - \mathbf{v}_i\|_{\mathbf{A}}^2, \tag{5}$$

where $\mathbf{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_c\}$ is a prototype matrix $\underset{1\le i\le c}{\forall} \mathbf{v}_i \in \mathbb{R}^p$, $\mathbf{U} \in M_{fcN}$ and $\mathbf{A}$ is positive-defined matrix. In this work, the unity matrix $\mathbf{I}$ has been choosen as matrix $\mathbf{A}$, hence the norm $\|\cdot\|_I^2$ is an euclidan norm in the p-dimensional space.

Based on the fuzzy partition matrix definition (4), elements u_{ik} of $\mathbf{U}$ have to satisfy the following constraints:

$$\underset{1\le k\le N}{\forall} \sum_{i=1}^{c} u_{ik} = 1, \tag{6}$$

and

$$\underset{1\le i\le c}{\forall}\sum_{k=1}^{N}u_{ik}>0. \tag{7}$$

Constraint (7) guarantees that no cluster is empty and constraint (6) ensures that the sum of membership degrees for each feature vector equals 1.

Let us now minimize the objective function (5). Noting that the columns of $\mathbf{U}$ are independent, the clustering task reads as the following constrained optimization problem:

$$\min J_k, \tag{8}$$

subject to conditions (6) and (7), where

$$J_k=\sum_{i=1}^{c}u_{ik}^m d_{ik}^2 \qquad k=1,2,\ldots,N, \tag{9}$$

where d_{ik}^2 denotes euclidan distance (*i.e.* $d_{ik}^2=\|\mathbf{x}_k-\mathbf{v}_i\|^2$).

Using the standard technique of Lagrange multipliers [4, 1], the optimization problem is converted into a form of unconstrained minimization (here J_k is written down explicitly to underline the variables taken into account in the optimization):

$$J_k(\lambda,\mathbf{u}_k)=\sum_{i=1}^{c}u_{ik}^m-\lambda\left(\sum_{i=1}^{c}u_{ik}-1\right), \tag{10}$$

with λ denoting the Lagrange multiplier and $\mathbf{u}_k$ denoting the k^{th} column of $\mathbf{U}$. The $(\lambda,\mathbf{u}_k)$ pair forms a stationary point of the optimized functional if and only if:

$$\frac{\partial J_k}{\partial\lambda}=0, \qquad \frac{\partial J_k}{\partial\mathbf{u}_k}=0. \tag{11}$$

These two derivatives yield the following relationships:

$$\frac{\partial J_k}{\partial\lambda}=\sum_{i=1}^{c}u_{ik}-1=0, \tag{12}$$

and

$$\frac{\partial J_k}{\partial u_{st}}=mu_{st}^{m-1}-\lambda=0, \tag{13}$$

$s=1,2,\ldots,c$ and $t=1,2,\ldots,N$. We begin with solving Equation (13) for u_{st}:

$$u_{st}=\left[\frac{\lambda}{m(d_{st})^2}\right]^{\frac{1}{m-1}}. \tag{14}$$

The sum of membership values, $\sum_{j=1}^{c}u_{jt}=1$, implies:

$$\sum_{j=1}^{t}u_{jt}=\sum_{j=1}^{c}\left(\frac{\lambda}{m}\right)^{\frac{1}{m-1}}\left[\frac{1}{(d_{jt})^2}\right]^{\frac{1}{m-1}}=\left(\frac{\lambda}{m}\right)^{\frac{1}{m-1}}\left\{\sum_{j=1}^{c}\left[\frac{1}{(d_{jt})^2}\right]^{\frac{1}{m-1}}\right\}=1. \tag{15}$$

Thus,

$$\left(\frac{\lambda}{m}\right)^{\frac{1}{m-1}}=\frac{1}{\sum\limits_{j=1}^{c}\left(\frac{1}{(d_{jt})^2}\right)^{\frac{1}{m-1}}}. \tag{16}$$

The left-hand side of Equation (16) may be directly substituted into Equation (14) producing the final expression for u_{st}:

$$\underset{1\le s\le c}{\forall}\ \underset{1\le t\le N}{\forall} u_{ik} = \left[\sum_{j=1}^{c}\left(\frac{d_{st}^2}{d_{jt}^2}\right)^{\frac{1}{m-1}}\right]^{-1}. \tag{17}$$

2.2.1. The fuzzy c-myriad clustering algorithm

The fuzzy c-myriad clustering method can be described as:

1. given the data set $\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\}$, where $\mathbf{x}_i\in\mathbb{R}^p$, fix the number of clusters, $c\in\{2,\ldots,N-1\}$, the fuzzyfier, $m\in[1,\infty)$, and the tolerance limit, ε. Initialize randomly the partition matrix, $\mathbf{U}$ (with respect to conditions (6) and (7)), fix $l=0$;
2. calculate the prototype values, $\mathbf{V}$, as weighted myriads. A weighted myriad has to be calculated for each component of $\mathbf{v}_i$;
3. update the partition matrix, $\mathbf{U}$, using Equation (17);
4. if $\|\mathbf{U}^{(l+1)}-\mathbf{U}^{(l)}\|<\varepsilon$, stop the clustering algorithm, otherwise $l=l+1$ and go to 2°.

2.3. The fuzzy c-myriad clustering method with partial supervision

At the first step in the fuzzy c-myriad clustering algorithm, the partition matrix, $\mathbf{U}$, has to be initialized. The assigment of each feature vector from the input data set is unknown, hence the initial values of the partition matrix are random. At the end of the clustering procedure, each row of the final partition matrix corresponds to a particular class, but the method has no way to know which row is which class. If the input data set is truly unlabeled (has no elements assigned to particular classes) this problem can only be resolved by human expert intervention. At the same time, a matrix including assigment of each input vector to a particular class could be utilized as an initial partition matrix. Unfortunately, each assigment of an input vector requires expert intervention, hence for high-dimensional, large data sets the assigment of every input vector is tedious, costly and impracticable.

The proposed approach is based on some feature vectors being selected from the input data set by expert as the best representative samples of a particular cluster [5]. Let us consider an input data set in the following form:

$$\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\}. \tag{18}$$

On the basis of the knowledge and experience of a human expert, the best representative samples can be selected from the input data set. Assuming that $\mathbf{x}_1^{(i)},\ldots,\mathbf{x}_{n_i}^{(i)}$ describes n_i samples of the i^{th} cluster selected by an expert, the input data set $\mathbf{X}$ can be denoted as:

$$\mathbf{X}=\left\{\underbrace{\mathbf{x}_1^{(1)},\ldots,\mathbf{x}_{n_1}^{(1)}}_{cluster\ 1},\underbrace{\mathbf{x}_1^{(2)},\ldots,\mathbf{x}_{n_2}^{(2)}}_{cluster\ 2},\ldots,\underbrace{\mathbf{x}_1^{(c)},\ldots,\mathbf{x}_{n_c}^{(c)}}_{cluster\ c},\mathbf{x}_1^{(u)},\ldots,\mathbf{x}_{n_u}^{(u)}\right\}=\mathbf{X}^{(s)}\cup\mathbf{X}^{(u)}, \tag{19}$$

where $\mathbf{x}_i^{(u)}$ denotes an unassigned input vector, $\mathbf{X}^{(s)}$ denotes a subset of the input data set assigned by the expert – the supervised part of input data $\mathbf{X}$, while $\mathbf{X}^{(u)}$ denotes the unassigned subset of $\mathbf{X}$ – the unsupervised part of the input data set.

In accordance with the division of the input data set, the partition matrix can also be divided into two parts: the supervised and the unsupervised, so that the partition matrix can be expressed as:

$$\mathbf{U} = \left[\left[\mathbf{u}_1^{(1)} \cdots \mathbf{u}_{n_1}^{(1)}\, \mathbf{u}_1^{(2)} \cdots \mathbf{u}_{n_2}^{(2)} \cdots \mathbf{u}_1^{(c)} \cdots \mathbf{u}_{n_c}^{(c)}\right] \left[\mathbf{u}_1^{(u)} \cdots \mathbf{u}_{n_u}^{(u)}\right]\right], \tag{20}$$

or

$$\mathbf{U} = \left[\begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} u_{11}^{(u)} & \cdots & u_{1n_u}^{(u)} \\ u_{21}^{(u)} & \cdots & u_{2n_u}^{(u)} \\ \vdots & & \vdots \\ u_{c1}^{(u)} & \cdots & u_{cn_u}^{(u)} \end{bmatrix}\right] = \left[\mathbf{U}^{(s)}\ \ \mathbf{U}^{(u)}\right], \tag{21}$$

where $\dim(\mathbf{U}^{(s)}) = c \times n_s$ and $\dim(\mathbf{U}^{(u)}) = c \times n_u$, n_s is the number of supervised (assigned) input vectors, n_u – the number of unsupervised input vectors, and the relationship $N = n_s + n_u$ is satisfied.

Thanks to the knowledge and experience of the human expert, the superivised part of the partition matrix, $\mathbf{U}^{(s)}$, is constant during the clustering process; only the unsupervised part, $\mathbf{U}^{(u)}$, is changing. Choosing an objective function similar to Equation (5), the optimal values of the unsupervised part of the partition matrix can be obtained from:

$$\underset{1\le i\le c}{\forall}\ \underset{1\le k\le n_s}{\forall} u_{ik} = \left[\sum_{j=1}^{c} \left(\frac{d_{ik}^2}{d_{jk}^2}\right)^{\frac{1}{m-1}}\right]^{-1}, \tag{22}$$

where $\underset{1\le i\le c}{\forall} \mathbf{v}_i$ are weighted myriads as cluster prototypes.

2.3.1. The fuzzy c-myriad clustering algorithm with partial supervision

The proposed method can be described as follows:

1. given the data set $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, where $\mathbf{x}_i \in \mathbb{R}^p$, fix the number of clusters, $c \in \{2, \ldots, N-1\}$, the fuzzyfier, $m \in [1, \infty)$, and the tolerance limit, ε. Initialize the partition matrix, $\mathbf{U}$: the supervised part, $\mathbf{U}^{(s)}$, by an expert, and the unsupervised part, $\mathbf{U}^{(u)}$ – randomly satysfying Equations (6) and (7);
2. calculate the prototype values, $\mathbf{V}$, as weighted myriads. A weighted myriad has to be calculated for each component of $\mathbf{v}_i$;
3. update the partition matrix, $\mathbf{U}$, using Equation (22);
4. if $\|\mathbf{U}^{(l+1)} - \mathbf{U}^{(l)}\| < \varepsilon$, stop the clustering algorithm, otherwise $l = l + 1$ and go to 2°.

3. A numerical experiment

A breast cancer database of the University of Wisconsin has been choosen as test data. Each instance in the database has been described by 11 features. The first parameter is an ID, the next nine parameters describe the cell nucleus, the eleventh parameter describes the type of cancer – malignant or benign; the latter parameter

is a diagnosis made by an expert (a physician). The database contains 699 instances, but in 16 of them some parameters are missing and these have been excluded from our analysis. Probably the most popular fuzzy c-means clustering method has been choosen as a reference method.

For the input data set with $p=9$ and $N=683$, the following parameters have been fixed:

- number of clusters $c=2$, (malignant cancer or benign cancer),
- the tolerance limit $\varepsilon=10^{-5}$,
- the fuzzyfier $m=2$,
- linearity weighted myriad parameter $K\in\{50,20,10,5,2\}$,
- number of supervised samples $n_s\in\{10,20,50,100,200\}$.

4. Results

For each value of the K parameter and for each value of n_s, the clustering algorithm has been performed five times (to exclude local minima solutions). A sample mean of misclassified instances, n_w, and their percentage share have been presented in Table 1.

Table 1. Number of misclassified instances

	$K=50$		$K=20$		$K=10$		$K=5$		$K=2$	
n_s	n_w	[%]	n_w	[%]	n_w	[%]	n_w	[%]	n_w	[%]
10	24	3.54	25	3.60	23	3.43	22	3.16	28	4.04
20	24	3.54	24	3.54	23	3.43	23	3.43	22	3.16
50	23	3.43	23	3.43	23	3.43	22	3.19	16	2.31
100	22	3.16	22	3.16	22	3.16	19	2.81	17	2.43
200	19	2.81	19	2.81	18	2.66	18	2.58	17	2.43

For the reference method, $n_w=30$ (4.39%) misclassified instances have been obtained. For the proposed method without supervision (*i.e.* $n_s=0$) for $K=20$, the number of misclassified samples equals $n_w=24$ (3.54%).

5. Conclusions

Including the knowledge and experience of an expert allows us to obtain more reliable results: the number of misclassified instances decreases. In the proposed method, results obtained for the number of supervised samples $n_s=10$ are very similar to those for the number of supervised samples $n_s=50$. For almost all cases (except for $n_s=10$ and $K=2$), the share of misclassified instances was lower than 4.00%.

The proposed method thus offers improved results.

References

[1] Bezdek J C 1981 *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press
[2] Arce G R and Kalluri S 2001 *IEEE Trans. Signal Proc.* **49** (11) 2721
[3] Arce G R and Kalluri S 1998 *IEEE Trans. Signal Proc.* **46** (2) 322
[4] Bertsekas D P 1982 *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press
[5] Hall L and Bensaid A M 1996 *Pattern Recognition* **29** (5) 859