http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (20): 5482-5486, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.5482.5486

A Fuzzy Centroids Clustering Algorithm with Between-cluster Information for Categorical Data

Wang Li-Na, Liu Qian and Zhou Yuan
College of Electronic and Information Engineering,
Nanjing University of Information Science and Technology,
210044, Nanjing, China

Abstract: In this study, a new fuzzy centroids clustering for categorical data is presented. The objective function of the fuzzy k-modes algorithm is modified by adding the between-cluster information so as to simultaneously minimize the within-cluster dispersion and enhance the between-cluster separation. Due to the misclassification by using the hard centroids, a fuzzy centroids clustering with the between-cluster information for categorical data is provided. Furthermore, the dissimilarity measure between an object and the centroid at the feature level is given as 1 minus the frequency of the feature value of the object. On several real data sets from UCI, the proposed algorithm is effective and the performance of the novel algorithm outperforms the one with hard-type centroids.

Key words: Fuzzy clustering, fuzzy centroids, between-cluster information, categorical data

INTRODUCTION

Clustering algorithms are increasingly required to deal with large scale data sets containing categorical data as well as numeric data, particularly in the context of data mining. A variety of fuzzy clustering algorithms have been proposed for clustering categorical data. The k-modes algorithm was developed by extending the standard k-means algorithm with a simple matching dissimilarity measure for categorical data, and provided a frequency-based method to update centroids in the clustering (Huang, 1998). A dissimilarity measure in the k-modes objective function was introduced, where the dominant level of the mode category is considered in the calculation of the dissimilarity measure (Ng *et al*., 2007). Furthermore, a generalized version of the k-modes algorithm was introduced (Huang and Ng, 1999).The fuzzy k-modes algorithm generates the fuzzy partition matrix from categorical data with the framework of the fuzzy k-means-type algorithm. However, there are a number of challenges in clustering categorical data. The lack of an inherent order on the domains of the individual attributes prevents the definition of a notion of similarity, which measures resemblance between categorical data objects. The fuzzy k-modes clustering algorithm uses the alternating minimization method to solve a non-convex optimization problem in finding cluster solutions. The within-cluster information is as an objective function to optimize the membership matrix and cluster modes. However, the between-cluster information is not considered, which often results in the clustering results with weak between-cluster separation. A novel fuzzy clustering algorithm with between-cluster information for categorical data was proposed, which gave the detailed analysis on the importance of the between-cluster information (Bai *et al.,* 2013).

To address the misclassification in the region of doubt, fuzzy centroids can make full use of the power of fuzzy sets in representing the uncertainty in the classification of categorical data (Kim *et al*., 2004; Ji *et al*., 2012). In the study, we developed a fuzzy centroids clustering algorithm with the between-cluster information for categorical data.

RELATED WORK

Fuzzy k-modes algorithm: The fuzzy k-modes clustering for categorical data was shown as follow. Let $U = \{X_1, X_2, \ldots, X_n\}$ be a set of n data, $A = \{A_1, A_2, \ldots, A_m\}$ be a set of m attributes. Each A_l describes a domain of values denoted by $DOM(A_l) = \{a_l^1, a_l^2, \ldots, a_l^{n_l}\}$, where n_l is the number of categorical values of attribute A_l.The objective of the fuzzy k-modes algorithm is to cluster the data X_j into k clusters by minimizing the function:

Corresponding Author: Wang Li-Na, College of Electronic and Information Engineering, Nanjing University of Information Science and Technology, 210044, Nanjing, China

$$J = \sum_{i=1}^{k}\sum_{j=1}^{n} u_{ij}^{m'} d_c(X_j, V_i) \quad (1)$$

subject to:

$$0 \le u_{ij} \le 1, 1 \le i \le k, 1 \le j \le n$$
$$\sum_{i=1}^{k} u_{ij} = 1, 1 \le j \le n$$
$$0 < \sum_{j=1}^{n} u_{ij} < n, 1 \le i \le k$$

The distance measure $d_c(X_j, V_i)$ between a categorical data X_j and the centroid V_i is defined as:

$$d_c(X_j, V_i) = \sum_{l=1}^{m} \delta(x_{jl}, v_{il}) \quad (2)$$

where, δ is a simple dissimilarity measure. The formula is as below:

$$\delta(x_{jl}, v_{il}) = \begin{cases} 1, & \text{if } x_{jl} \ne v_{il} \\ 0, & \text{if } x_{jl} = v_{il} \end{cases} \quad (3)$$

Fuzzy clustering with fuzzy centroids: In fuzzy clustering with fuzzy centroids, a soft decision can be made when selecting the cluster centroids for categorical attributes. For DOM $(A_l) = \{a_l^1, a_l^2, \ldots, a_l^{nl}\}$, ,the fuzzy centroid, denoted by $\tilde{v}$, is defined as:

$$\tilde{V} = [\tilde{V}_1, \ldots, \tilde{V}_l, \ldots, \tilde{V}_m] \quad (4)$$

where $\tilde{V}_l = a_l^1 / \omega_l^1 + a_l^2 / \omega_l^2 + \ldots + a_l^r / \omega_l^r + \ldots + a_l^{n_l} / \omega_l^{n_l}$, subject to:

$$0 \le \omega_l^r \le 1, \quad 1 \le r \le n_l,$$
$$\sum_{r=1}^{n_l} \omega_l^r = 1, \quad 1 \le l \le m.$$

Each attribute $\tilde{v}_l \in \tilde{v}$ is a fuzzy category value represented as a fuzzy set $\{\alpha_l^r, \omega_l^r\}$. This is determined by the category distribution of attribute A_l for the data objects belonging to the cluster.

Fuzzy k-modes algorithm with between-cluster information for categorical data: Considering the between-cluster information, the clustering objective function is modified so as to simultaneously minimize the within-cluster dispersion and enhance the between-cluster separation. The objective function is written as follow:

$$J_\gamma = \sum_{i=1}^{k}\sum_{j=1}^{n} u_{ij}^{m'} d(X_j, V_i) + \gamma \sum_{i=1}^{k}\sum_{j=1}^{n} u_{ij}^{m'} \frac{1}{n}\sum_{p=1}^{n} s(X_p, V_i) \quad (5)$$

Here, $s(X_p, V_i)$ is a similarity measure between X_p and V_i which is defined as:

$$s(X_p, V_i) = \sum_{l=1}^{m} \phi(x_{pl}, v_{il}) \quad (6)$$

Where:

$$\phi(x_{pl}, v_{il}) = \begin{cases} 1, & \text{if } x_{pl} = v_{il} \\ 0, & \text{if } x_{pl} \ne v_{il} \end{cases}$$

The minimal solution of the objective function can be obtained in a finite number of iteration. The matrix u_{ij} and V_i are calculated according to the following two theorems. The optimal u_{ij} is decided by:

$$u_{ij} = \begin{cases} 1, & d(X_j, V_i) + \gamma \frac{1}{n}\sum_{p=1}^{n} s(X_p, V_i) = 0, \\ \left[\sum_{h=1}^{c}\left[\frac{d(X_j, V_i) + \gamma \frac{1}{n}\sum_{p=1}^{n} s(X_p, V_i)}{d(X_j, V_h) + \gamma \frac{1}{n}\sum_{p=1}^{n} s(X_p, V_h)}\right]^{\frac{1}{m'-1}}\right]^{-1} \end{cases} \quad (7)$$

The optimal V_i is shown as below:

$$v_{il} = \alpha_l^r \in \mathrm{DOM}(A_l) \quad (8)$$

Where:

$$\sum_{j=1, x_{jl}=a_l^r}^{n} u_{ij}^{m'} - \gamma \frac{1}{n}\sum_{j=1}^{n} u_{ij}^{m'} \mid X_p \mid x_{pl} = a_l^r, X_p \in U \mid$$
$$\ge \sum_{j=1, x_{jl}=a_l^q}^{n} u_{ij}^{m'} - \gamma \frac{1}{n}\sum_{j=1}^{n} u_{ij}^{m'} \mid X_p \mid x_{pl} = a_l^q, X_p \in U \mid, \quad 1 \le q \le n_l$$

for $1 \le l \le m$ and $1 \le i \le k$.

A NOVEL FUZZY CENTROIDS CLUSTERING WITH BETWEEN CLUSTER INFORMATION FOR CATEGORICAL DATA

In this section, we will introduce a new clustering algorithm integrated the between-cluster information with fuzzy centroids. The application of fuzzy centroids allows the user to fully exploit the power of fuzzy sets in representing uncertainty and imprecision. Due to good cluster criteria having high within-cluster similarity and low between-cluster similarity, the between-cluster information results in enhanced separation between clusters.

Distance measure: To solve the clustering of categorical objects, some dissimilarity measures were proposed

(Chan *et al*., 2004; Cao *et al*., 2012). The dissimilarity measure between an object and a centroid at the feature level is given as 1 minus the frequency of the feature value of the object (Ng *et al*., 2007; Lee and Pedrycz, 2009). Let $\tilde{v}_i$ and X_j be a fuzzy centroid and a data point, represented as $[\tilde{v}_{i1},...,\tilde{v}_{il},...,\tilde{v}_{im}]$ and $[x_{j1}, x_{j2},\ldots x_{jm}]$, respectively. The distance measure between $\tilde{v}_i$ and X_j is defined as:

$$d(X_j, \tilde{V}_i) = \sum_{l=1}^{m} \theta(x_{jl}, \tilde{v}_{il}) \tag{9}$$

Where:

$$\theta(x_{jl}, \tilde{v}_{il}) = \begin{cases} \sum_{r=1}^{n_l} f_{ilr}\,\delta(x_{jl}, \tilde{v}_{il}) = 1 - f_{ilq}, & \text{if } x_{jl} = a_l^q \text{ for } 1 \le q \le n_l \\ 1, & \text{otherwise} \end{cases}$$

In the above equation, δ is a simple dissimilarity measure, f_{ilr} is the frequency of α_l^r in the ith cluster and is defined as below:

$$f_{ilr} = \frac{\sum_{j=1 \wedge x_{jl}=a_l^r}^{n} u_{ij}^{m'}}{\sum_{j=1}^{n} u_{ij}^{m'}} \tag{10}$$

Note that:

$$\sum_{r=1}^{n_l} f_{ilr} = 1$$

Proposed clustering algorithm for categorical data: In the new algorithm, we modify the objective function by using fuzzy centroids so that we can make full use of fuzzy sets to obtain the logical result. To minimize the objective function with fuzzy centroids, the function is modified as below:

$$J = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m'} d(X_j, \tilde{V}_i) + \gamma \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m'} \frac{1}{n}\sum_{p=1}^{n} s(X_p, \tilde{V}_i) \tag{11}$$

Here, $s(X_p, \tilde{V}_i)$ is a similarity measure between X_p and $\tilde{V}_i$, which is defined as:

$$s(X_p, \tilde{V}_i) = \sum_{l=1}^{m} \varphi(x_{pl}, \tilde{v}_{il}) \tag{12}$$

where $\varphi(x_{pl}, \tilde{v}_{il}) = \omega_l^r$, if $x_{pl} = a_l^r$, $1 \le r \le n_l$. The parameter γ is a factor to maintain a balance between the effect of between-cluster information and that of within-cluster information in the minimization procedure of Eq. 11. As the literature has suggested, the values of γ varied in [0,1] to attain the optima (Bai *et al*., 2013). The proposed algorithm uses the fuzzy c-means-types paradigm to cluster categorical data.

Algorithm 1. Fuzzy centroids clustering algorithm with between-cluster information for categorical data:

- **Step 1:** Choose the number of clusters, c. Given chosen value of m′ and initialize the fuzzy partition membership u_{ij}. Set iterative counter t = 1
- **Step 2:** Compute the fuzzy cluster centroid $\tilde{V}_i(t+1) = [\tilde{v}_{i1},...,\tilde{v}_{ik},...,\tilde{v}_{im}]$ for i = 1, 2, …, c. For each $\tilde{v}_{i,l} = (a_l^r, \omega_l^r)$ for $1 \le l \le m$:

$$\omega_l^r = \frac{\sum_{j=1 \wedge x_{jl}=a_l^r}^{n} u_{ij}^{m'}}{\sum_{j=1}^{n} u_{ij}^{m'}} \tag{13}$$

Subject to:

$$\sum_{r=1}^{n_l} \omega_l^r = 1$$

- **Step 3:** Update the ith fuzzy membership u_{ij} for each X_j as follow:

$$u_{ij} = \begin{cases} 1, & d(X_j, \tilde{V}_i) + \gamma \frac{1}{n}\sum_{p=1}^{n} s(X_p, \tilde{V}_i) = 0, \\ \left[\sum_{h=1}^{c}\left[\frac{d(X_j, \tilde{V}_i) + \gamma \frac{1}{n}\sum_{p=1}^{n} s(X_p, \tilde{V}_i)}{d(X_j, \tilde{V}_h) + \gamma \frac{1}{n}\sum_{p=1}^{n} s(X_p, \tilde{V}_h)}\right]^{\frac{1}{m'-1}}\right]^{-1} \end{cases} \tag{14}$$

- **Step 4:** If there is no improvement in J,then stop; otherwise, return to Step 2

Let us consider the time complexities of the proposed algorithm. First, computing the frequency of each categorical value of each attribute in U and it takes:

$$O(n\sum_{l=1}^{m} n_l)$$

operations, which can be used to compute the distance of each object to the centroids. The time complexities required mainly depends on the updates of the fuzzy centroids and partition matrix in each iteration. The computational costs of updating the fuzzy centroids and partition matrix are O (cnmR) and O (cnm), respectively. Where c is the number of clusters, n is the number of objects, m is the number of attributes, R (= max (n_l)) is the

maximum number of categories for $1 \le l \le m$. If one needs t_d iterations to obtain a local minimal solution of Eq. 11 for each γ_d (d = 1, 2, …,e), thus the total computational complexity of the proposed algorithm is:

$$O((n\sum_{l=1}^{m} n_l + cnm(R+1))\sum_{d=1}^{e} t_d)$$

EXPERIMENT ANALYSIS

A through experimental evaluation of the clustering performance of our algorithm are provided by using some UCI repository datasets (Frank and Asuncion, 2010). To assess the quality of the obtained clustering results, we utilize the provided ground truth labeling and employ some objective criteria that measure the overlap between an obtained clustering and the corresponding ground truth classifications. Specially, let us consider a dataset comprising c clusters and its clustering result obtained by application of a considered clustering algorithm. Let η_i denote the number of data points correctly assigned to the ith class by the applied clustering algorithm. Let err_i denote the data points that are incorrectly assigned to the ith class and θ_i denote the data points which were incorrectly rejected from the cluster. Then, the precision p_i of the considered clustering algorithm on the ith class of the clustered dataset is defined as:

$$p_i = \eta_i / (\eta_i + err_i) \quad (15)$$

while the recall r_i for the ith class is defined as:

$$r_i = \eta_i / (\eta_i + \theta_i) \quad (16)$$

Based on these measures, the right rate (micro-right) of an evaluated clustering algorithm can be derived, it can be shown that the measures is given (Yang, 1999):

$$micro-right = \frac{1}{n}\sum_{i=1}^{c} \eta_i \quad (17)$$

Considering the value of the degree of fuzziness parameter in FCM-type fuzzy clustering is usually conducted heuristically using a performance-based criterion. Typically, it is suggested to use the values in the interval (1, 3]. Here, as has suggested, we set the optimal value of fuzzy degree m' as 1.1 (Huang and Ng, 1999). However, the appropriate setting of γ depends on the domain knowledge of the data sets. As the corresponding literature suggested, the parameters γ is gradually reduced from 1 to 0 as the step 0.1(Bai *et al.*, 2013). At the same time, set error threshold ϵ as 0.0001 and iteration counter t_{max} as 10.

Table 1: Performance of the evaluated algorithms on Mushroom dataset

Method	η_1	η_2	err_1	err_2	Micro-right
Fuzzy k-modes	2745	1258	894	743	0.7098
Bai, *et al.*'s method	3389	1348	804	99	0.8399
New method	3434	1373	779	54	0.8523

Table 2: Performance of the evaluated algorithms on Statlog (heart disease) dataset

Method	η_1	η_2	err_1	err_2	Micro-right
Fuzzy k-modes	115	87	33	35	0.7481
Bai *et al.*, method	121	90	36	23	0.7815
New method	121	100	20	29	0.8185

Table 3: Performance of the evaluated algorithms on credit approval dataset.

Method	η_1	η_2	err_1	err_2	Micro-right
Fuzzy k-modes	225	268	89	71	0.7550
Bai *et al.*, method	217	295	62	79	0.7841
New method	254	279	78	42	0.8162

Mushroom dataset: First, the Mushroom data from the UCI repository are considered. This dataset comprises 22 categorical attributes of gilled mushrooms in the Agaricus and Lepiota Family. It contains 8124 data points. Data points are divided into two classes, namely edible and poisonous. Due to the missing of the data, 5640 data can be used here. The edible mushrooms contains 3488 data points, the poisonous contains 2152 data points. We run the proposed algorithms 100 times, each time with different(random) initializations and we compute its average performance in terms of the number of correctly classified data points and the resulting micro-right values. The corresponding analysis is presented in Table 1.

Statlog (Heart disease) dataset: Second, the Statlog (heart) dataset from the UCI repository is considered. This collection consists of 270 samples comprising mixed categorical and numeric attributes; there are eight categorical and five numeric attributes. In the test, only categorical attributes are considered. The available data points are classified into 2 classes: normal, including 150 data points and heart patient, including 120 data points. As in the previous experiment, we run the algorithms 100 times, each time with different (random) initializations and we evaluate its performance and the obtained results are provided in Table 2.

Credit approval dataset: Finally, we consider the Credit approval dataset from the UCI repository. This dataset contains 690 data points comprising eight categorical and six numberic attributes. The numberic attributes are removed in the experiment.37 cases have one or more missing values. Date points have 653 in practice and are divided into 2 classes: Approval, including 296 samples and denial, including 357 samples. In our investigations, the obtained results are provided in Table 3.

CONCLUSION

In this study, we proposed a novel fuzzy centroids clustering algorithm for categorical data. In the method, the objective function contains the within-cluster and the between-cluster information and thus the clustering result is produced with high within-cluster similarity and low between-cluster similarity. Here, the within-cluster dissimilarity between an object and the centroid at the feature level is given as 1 minus the frequency of the feature value of the object. By using the fuzzy centroids, the algorithm can preserve the uncertainty inherence in data sets for longer time before decisions are made and it is prone to the global optima in comparison with other algorithms. The updating formulas of the fuzzy centroids and membership matrix are derived from the optimal procedure. The experimental evaluation on real data sets from UCI repository has shown that our method outperforms fuzzy k-modes clustering with between-cluster information for categorical data.

ACKNOWLEDGMENTS

The study would like to thank for the support by Natural Science Foundation of China under the Grant 61203273,61302048 and 61202137.The author also thank for the support by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

REFERENCES

Bai, L., J. Liang, C. Dang and F. Cao, 2013. A novel fuzzy clustering algorithm with between-cluster information for categorical data. Fuzzy Sets Syst., 215: 55-73.

Cao, F., J. Liang, D. Li, L. Bai and C. Dang, 2012. A dissimilarity measure for the k-Modes clustering algorithm. Knowl. Based Syst., 26: 120-127.

Chan, E.Y., W.K. Ching, M.K. Ng and Z.J. Huang, 2004. An optimization algorithm for clustering using weighted dissimilarity measures. Pattern Recogn., 37: 943-952.

Frank, A. and A. Asuncion, 2010. UCI machine learning repository Irvine. University of California, School of Information and Computer Science, USA, http://archive.ics.uci.edu/ml/

Huang, Z. and M.K. Ng, 1999. A fuzzy k-modes algorithm for clustering categorical data. IEEE Trans. Fuzzy Syst., 7: 446-452.

Huang, Z., 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining Knowledge Discovery, 2: 283-304.

Ji, J., W. Pang, C. Zhou, X. Han and Z. Wang, 2012. A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data. Knowl. Based Syst., 30: 129-135.

Kim, D.W., K.H. Lee and D. Lee, 2004. Fuzzy clustering of categorical data using fuzzy centroids. Pattern Recogn. Lett., 25: 1263-1271.

Lee, M. and W. Pedrycz, 2009. The fuzzy C-means algorithm with fuzzy P-mode prototypes for clustering objects having mixed features. Fuzzy Sets Syst., 160: 3590-3600.

Ng, M.K., M.J. Li, J.Z. Huang and Z. He, 2007. On the impact of dissimilarity measure in k-Modes clustering Algorithm. Trans. Pattern Anal. Mach. Intell., 29: 503-507.

Yang, Y., 1999. An evaluation of statistical approaches to text categorization. Inform. Retrieval, 1: 69-90.