Fuzzy Clustering Technique for Numerical and Categorical dataset

Revati Raman Dewangan , Lokesh Kumar Sharma, Ajaya Kumar Akasapu

Dept. of Computer Science and Engg. , CSVTU Bhilai(CG), Rungta College of Engineering and Tech. Bhilai(C.G.),India

Dept. of IT and MCA, CSVTU Bhilai(CG), Rungta College of Engineering and Tech. Bhilai(C.G.),India

Dept. of Computer Science and Engg. , CSVTU Bhilai(CG), Rungta College of Engineering and Tech. Bhilai(C.G.),India

revati2004@gmail.com * drlokeshksharma@gmail.com * ajaykumar.akasapu@gmail.com

Abstract— **Data clustering is a common technique for data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. In Fuzzy logic System, Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. Use of traditional fuzzy c-mean type algorithm is limited to numeric data. We present a modified description of cluster center to overcome the numeric data only limitation of Fuzzy c-mean algorithm and provide a better characterization of clusters. The fuzzy k-modes algorithm for clustering categorical data. We are going to propose new cost function and distance measure based on co-occurrence of values. The measures also take into account the significance of an attribute towards the clustering process. Fuzzy k-modes algorithm for clustering categorical data is extended by representing the clusters of categorical data with fuzzy centroids. Use of fuzzy centroids makes it possible to fully exploit the power of fuzzy sets in representing the uncertainty in the classification of categorical data. The effectiveness of the new fuzzy k-modes algorithm is better than those of the other existing k-modes algorithms.**

Keywords — **Fuzzy logic System, Fuzzy C-Mean, Fuzzy K-Modes, Fuzzy centroids.**

I. INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Data mining discovers description through clustering visualization, association, sequential analysis. This dissertation work aims to clustering numeric and categorical data using fuzzy approach.

Fuzzy logic, Type of reasoning based on the recognition that logical statements are not only true or false (white or black areas of probability) but can also range from 'almost certain' to 'very unlikely' (gray areas of probability). Software based on application of fuzzy-logic (as compared with that based on Formal Logic) allows computers to mimic human reasoning more closely, so that decisions can be made with incomplete or uncertain data. Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. Clustering involves the task of dividing data points into homogeneous classes or clusters so that items in the same class are as similar as possible and items in different classes are as dissimilar as possible. Clustering can also be thought of as a form of data compression, where a large number of samples are converted into a small number of representative prototypes or clusters.

Fuzzy C-Means Clustering (FCM)

The FCM algorithm is one of the most widely used fuzzy clustering algorithms. This technique was originally introduced by Jim Bezdek in 1981.

The FCM algorithm attempts to partition a finite collection of elements X={x_1, x_2, ... , x_n} into a collection of c fuzzy clusters with respect to some given criterion. Fuzzy set allows for degree of membership. A single point can have partial membership in more than one class. There can be no empty classes and no class that contains no data points.

Given a finite set of data, the algorithm returns a list of c cluster centers V, such that

V =v_i, i =1, 2, ..., c

And a partition matrix U such that

U =u_{ij} , i =1....c, j =1,...n

where u_{ij} is a numerical value in [0, 1] that tells the degree to which the element x_j belongs to the i-th cluster.
The following is a linguistic description of the FCM algorithm, which is implemented Fuzzy Logic.

Step 1: Select the number of clusters c ($2 \leq c \leq n$), exponential weight μ($1 < \mu < \propto$), initial partition matrix U^0 , and the termination criterion ϵ. Also, set the iteration index l to 0.

Step 2: Calculate the fuzzy cluster centers $\{v_i^l \mid i=1, 2, ..., c\}$ by using U^l.

Step 3: Calculate the new partition matrix U^{l+1} by using $\{v_i^l \mid i=1, 2, ..., c\}$.

Step 4: Calculate the new partition matrix $\Delta = \|U^{l+1} - U^l\| = \max_{i,j} |u_{ij}^{l+1} - u_{ij}^l|$. If $\Delta > \epsilon$, then set l = l + 1 and go to step 2. If $\Delta \leq \epsilon$, then stop.

To demonstrate the FCM clustering algorithm, we will create a data set that consists of four groups of data in a row.A lot of work has been done in this direction. Different researcher proposed their view to find out the cluster areas for different types data items.

Most of the techniques used in the literature in clustering symbolic data are based on the hierarchical methodology, which utilizes the concept of agglomerative or divisive methods as the core of the algorithm [1]. The main contribution of this paper is to show how to apply the concept of fuzziness on a data set of symbolic objects and how to use this concept in formulating the clustering problem of symbolic objects as a partitioning problem. Finally, a fuzzy symbolic c-means algorithm is introduced as an application of applying and testing the proposed algorithm on real and synthetic data sets.

The clustering of data set into subsets can be divided into hierarchical and non-hierarchical or partitioning methods. The general rationale behind partitioning methods is to choose some initial partitioning of the data set and then alter cluster memberships so as to obtain better partitions according to a predefined objective function. Hierarchical clustering procedures can be divided into agglomerative methods, which progressively merge the objects according to some distance measure in such a way that whenever two objects belong to the same cluster at some level they remain together at all higher levels and divisive methods, which progressively subdivide the data set.

SYMBOLIC OBJECTS

Various definitions and descriptions of symbolic objects and distance measures are found in the literature.

A. Feature Types

The symbolic object can be written as the Cartesian product of specific values of its features Ak's as equation (1)

$A=A_1 * A_2 *$ ……..* A_d……… (1)

The feature values may be measured on different scales resulting in the following types: 1) quantitative features, which can be classified into continuous, discrete. and interval values and 2) qualitative features, which can be classified into nominal (unordered), ordinal (ordered), and combinational.

B. Dissimilarity

Many distance measures are introduced in the literature for symbolic objects. The dissimilarity between two symbolic objects and is defined as as equation (1.1.1.2)

$D(A,B) = \sum_{k=1}^{d} D(A_k, B_k)$……(2)

For the th feature $D(A_k, B_k)$, is defined using the following three components.

1) $D_p(A_k, B_k)$ due to position p;
2) $D_s(A_k, B_k)$ due to position s;
3) $D_c(A_k, B_k)$ due to position c;

Fuzzy Clustering for Symbolic Objects

Fuzzy c-means clustering for numerical data is the algorithm that attempts to find a solution to the mathematical program as defined in equation (3)

$$J(Z,W,X) = \sum_{i=1}^{c} \sum_{i=1}^{n} \mathrm{w_{ij}}^{\mathrm{m}}\ \mathrm{d}^2(\mathrm{Xi},\mathrm{Zj}) \ldots 3$$

Where

n = number of patterns

C $\equiv$ *number of cluster*

m$\equiv$ $a\ scalar, m > 1$

Zj$\equiv$center of cluster j

w_{ij} $\equiv$degree of membership of pattern i in cluster j

Z = Cluster center matrix

S $\equiv$ dimension of the feature space

W$\equiv$ membership matrix

Xi $\equiv$ pattern i

[3]This correspondence describes extensions to the fuzzy k-means algorithm for clustering categorical data. By using a simple matching dissimilarity measure for categorical objects and modes instead of means for clusters, a new approach is developed, which allows the use of the k-means paradigm to efficiently cluster large categorical data sets. A fuzzy k-modes algorithm is presented and the effectiveness of the algorithm is demonstrated with experimental results.

HARD AND FUZZY -MEANS ALGORITHMS

Let X be a set of n objects described by m numeric attributes. The hard and fuzzy -means clustering algorithms to cluster X into k clusters can be stated as the algorithms, which attempt to minimize the cost function as defined in equation (4)

$$F(W,Z) = \sum_{i=1}^{k} \sum_{i=1}^{n} \mathrm{w_{ij}}^{\alpha}\ \mathrm{d}\,(\mathrm{Zt},\mathrm{Xi}) \quad \ldots\ldots(4)$$

Subject to

$$0 \le \mathrm{w_{li}} \le 1, \qquad 1 \le l \le k, \qquad 1 \le i \le n$$

$$\sum_{i=1}^{k} \mathrm{w_{li}} = 1, \qquad 1 \le \mathrm{i} \le \mathrm{n}$$

and $0 \le \sum_{i=1}^{k} \mathrm{w_{li}} < 1, \quad 1 \le \mathrm{l} \le \mathrm{k}$

Where

k($\le$) *is a known number of clusters*,

$\alpha\ \varepsilon(1,\infty)$ *is a weighted exponent*,

$W = [w_{li}]$ *is a k − by − n real matrix* ,

$Z = [Z1, Z2, Z3, \ldots \ldots, Z_K]$ *and* $d(Z_t, X_i)$ (≥ 0) *is some dissimilarity*

HARD AND FUZZY -MODES ALGORITHMS

The hard K-modes algorithm, made the following modifications to the K-means algorithm using a simple matching dissimilarity measure for categorical objects and replacing the means of clusters with the modes. These modifications have removed the numeric-only limitation of the K-means algorithm but maintain its efficiency in clustering large categorical data sets .

Let X and Y be two categorical object represented by $[x_1,x_2,\ldots.,x_m]$ and $[y_1,y_2,\ldots.,y_m]$. The simple matching dissimilarity measure between X and Y is defined as follows in equation (5)

$$d_c(X,Y) \equiv \sum_{j=1}^{m} \delta(x_j, y_j) \quad ..(5)$$

Where

$$\delta(x_j, y_j) \equiv \begin{cases} 0, & xj = yj \\ 1, & nj \ne yj \end{cases}$$

It is easy to verify that the function dc defines a metric space on the set of categorical objects. Traditionally, the simple matching approach is often used in binary variables which are converted from categorical variables. We note dc that is also a kind of generalized Hamming distance . The K -modes algorithm uses the K -means paradigm to cluster categorical data. The objective of clustering a set of categorical objects into clusters K is to find W and Z that Minimize from equation (6)

$$F_c(W,Z) = \sum_{t=1}^{k} \sum_{i=1}^{n} w_{li}^{\alpha} d_c Z_t, X_t) \ldots (6)$$

Here, Z represents a set of K- modes for K clusters.1 We can still use Algorithm 1 (K means) to minimize Fc(W,Z).

Algorithms for clustering mixed data

With the advent of very large databases containing mixed set of attributes, the data mining community has been on the look-out for good criterion function for handling mixed data, since the algorithms discussed earlier work well on either categorical or numeric valued data. In order to overcome this problem, some of the strategies that have been employed are as follows:

(1) Categorical and nominal attribute values are converted to numeric integer values and then numeric distance measures are applied for computing similarity between object pairs. However, it is very difficult to give correct numeric values to categorical values like colour, etc.

(2) Another approach has been to discretize numeric attributes and apply categorical clustering algorithm. But the discretization process leads to loss of information.

Huang’s cost function for k-prototypes clustering algorithm.

Huang [8] defined a cost function for clustering mixed data sets with n data objects and m attributes (mr numeric attributes, mc categorical attributes, m = mr + mc) as equation (7)

$$\zeta = \sum_{i=1}^{n} v(d_i, c_j) \ldots\ldots(7)$$

where v(di,Cj) is the distance of a data object di from the closest cluster center Cj. v(di,Cj) is defined as equation in (8)

$$v(d_i, C_j) = \sum_{t=1}^{m} (d_{it}^{r} - c_{jt}^{r})^2 + \gamma_j \sum_{t=1}^{m} (d_{it}^{c} - C_{jt}^{c}) \ldots\ldots..(8)$$

d_r it are values of numeric attributes and dc it are values of categorical attributes for data object di. Here Cj = (C_{j1},C_{j2}, . . . ,C_{jm}) represents the cluster center for cluster j. Cc jt represents the most common value (mode) for categorical attributes t and class j. C_r jt represents mean of numeric attribute t and cluster j. For categorical attributes, d(p,q) = 0 for p = q and d(p,q) = 1 for p 5 q. cj is a weight for categorical attributes for cluster j. Cost function f is minimized for clustering mixed data sets

- For categorical attributes, the cluster center is represented by the mode of the cluster rather than the mean. While this allays the problem of finding the mean for categorical values, there is information loss since the true representation of the cluster is not obtained. Only one attribute value represents the cluster, even though there may be close seconds or thirds.
- Binary distance between two categorical attribute values p and q is taken as d(p,q) = 0 for p = q and d(p,q) = 1 for p 5 q. This does not reflect the real situation appropriately. Stanfill and Waltz suggested that for supervised learning though it is observed that d(p,q) = 0 for p = q, but it is not necessarily true that d(p,q) = 1 for p 5 q.

According to them d(p,q) is mostly different for different attribute value pairs and depends on the relative frequencies of value pairs within a class. This works even for clustering since it is usually not one attribute that determines the clusters but rather a collection of attributes. Thus, during clustering, attribute value co-occurrences among different attributes should be considered to compute d(p,q). The distance measure in that case can take care of significance of an attribute.

- In Huang’s cost function weight of all numeric attributes is taken to be 1. The weight of categorical attributes is a user-defined parameter cj. However, in a real data set all numeric attributes may not have the same effect on clustering. Incorrect user-given values of cj may also lead to inaccurate clustering.

Computing distance between two categorical values

The similarity and dissimilarity of two objects obviously depend on how close their values are for all attributes. While it is easy to compute the closeness for numeric attributes, it becomes difficult to capture this notion for categorical attributes. Computation of similarity between categorical data objects in unsupervised learning is an important data mining problem. Most of the existing distance measures do not consider the distribution of values in the data set while computing the distance between any two categorical attribute values, something that is naturally captured for numerical attributes.

FUZZY K-MODE

This describes extensions to the fuzzy k-means algorithm for clustering categorical data. By using a simple matching dissimilarity measure for categorical objects and modes instead of means for clusters, a new approach is developed, which allows the use of the k-means paradigm to efficiently cluster large categorical data sets.

With the advent of very large databases containing mixed set of attributes, the data mining community has been on the look-out for good criterion function for handling mixed data, since the algorithms discussed earlier work well on either categorical or numeric valued data. In order to overcome this problem, some of the strategies that have been employed are as follows:

(1) Categorical and nominal attribute values are converted to numeric integer values and then numeric distance measures are applied for computing similarity between object pairs. However, it is very difficult to give correct numeric values to categorical values like colour, etc.

(2) Another approach has been to discretize numeric attributes and apply categorical clustering algorithm. But the discretization process leads to loss of information.

The clusters produced by the k-means procedure are sometimes called "hard" or "crisp" clusters, since any feature vector **x** either is or is not a member of a particular cluster. This is in contrast to "soft" or "fuzzy" clusters, in which a feature vector **x** can have a degree of membership in each cluster.

The fuzzy-k-means procedure of Dunn and Bezdek allows each feature vector **x** to have a degree of membership in Cluster i:

```
•  Make initial guesses for the means $\mathbf{m}_1$, $\mathbf{m}_2$,..., $\mathbf{m}_k$
•  Until there are no changes in any mean:
     o  Use the estimated means to find the degree of membership u(j,i) of $\mathbf{x}_j$ in Cluster  i;
        for example, if a(j,i) = exp(- || $\mathbf{x}_j$ - $\mathbf{m}_i$ ||$^2$ ), one might use u(j,i) = a(j,i) / sum_j a(j,i)
     o  For i from 1 to k
          ▪  Replace $\mathbf{m}_i$ with the fuzzy mean of all of the examples for Cluster i –
             $m_i = \frac{\sum_j u(j,i)^2 x_j}{\sum_j u(j,i)^2}$
        end_for
•  end_until
```

While it is easy to compute the closeness for numeric attributes, it becomes difficult to capture this notion for categorical attributes. Computation of similarity between categorical data objects in unsupervised learning is an important data mining problem.

II. Conclusions

Fuzzy c-means (FCM) is a data clustering technique wherein each data point belongs to a cluster to some degree that is specified by a membership grade. Using Fuzzy C means, we can find cluster for numeric data items. We can also find the center of the cluster for numeric data items. While it is easy to compute the closeness for numeric attributes, it becomes difficult to capture this notion for categorical attributes.

In future I will implement the fuzzy k mode algorithm for categorical data and combine the tow implementation that is Fuzzy C Means and Fuzzy K modes for mixture of data items set and that will find out the cluster area and center of cluster.

Implement the Fuzzy C Mean (FCM) algorithm for numeric data. Implement the Fuzzy k Mode algorithm for categorical data. Combine the both implementation for mixture of data items that is numeric as well as categorical data items.

References

[1] Y. El Sonbty and M. A. Ismail, “Fuzzy Clustering for Symbolic Data”, IEEE Transactions on Fuzzy Systems, Vol. 6, No. 2, pp. MAY 1998.

[2] Dae-Won Kim, Kwang H. Lee, and Doheon Lee, “Fuzzy clustering of categorical data using fuzzy centroids”.

[3] Zhexue Huang and Michael K. Ng, “A fuzzy K mode algorithm for clustering categorical data”,IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 7, NO. 4, AUGUST 1999.

[4] Binu Thomas,Sonam Wangmo and Raju G. “A Modified Fuzzy C-Means Algorithm for Natural Data Exploration” ,World Academy of Science, Engineering and Technology 49 2009.

[5] Amir Ahmad a,*, Lipika Dey b, “ A k-mean clustering algorithm for mixed numeric and categorical data”,

[6] F. Can, E. Ozkarahan, “A dynamic cluster maintenance system for information retrieval”, Proceedings of the Tenth Annual International ACM SIGIR Conference, 1987, pp. 123–131.

[7] A new fuzzy k-modes clustering algorithm for categorical data Int. J. Granular Computing, Rough Sets and Intelligent Systems, Vol. 1, No. 1, 2009.

[8] M. Eissen, P. Spellman, P. Brown, D. Bostein, “Cluster analysis and display of genome- wide expression patterns”, Proceeding of National Academy of Sciences of USA, vol. 95, 1998, pp. 14863–14868.

[9] Hesam Izakian, Ajith Abraham, “Fuzzy clustering using hybrid fuzzy c-means and fuzzy particle swarm Optimization” Machine Intelligence Research Labs(MIR Labs).

[10] A.K. Jain, R.C. Dubes, “Algorithms for Clustering Data”, Prentice Hall, Englewood Cliff, New Jersey, 1988.

[11] Z. Huang, M.K. Ng, “A fuzzy k-modes algorithm for clustering categorical data”, IEEE Transactions on Fuzzy Systems 7 (4) (1999) 446–452.

[12] C. Do¨ ring, C. Borgelt, R. Kruse, “Fuzzy clustering of quantitative and qualitative data”, Proceedings of NAFIPS, Banff, Alberta, 2004.

[13] G. Biswas, J. Weingberg, D.H. Fisher, “A conceptual clustering algorithm for data mining, IEEE Transactions on Systems”, Man, and Cybernetics 28C pp. 219–230, 1998.

[14] S. Guha, R. Rastogi, S. Kyuseok, “ROCK: A robust clustering algorithm for categorical attributes”, Proceedings of 15th.

[15] M. A. Woodbury and J. A. Clive, “Clinical pure types as a fuzzy partition,” *J Cybern.*, vol. 4-3, pp. 111–121, 1974