Clustering Categorical Data Using a Swarm-based Method

Hesam Izakian, Ajith Abraham
Machine Intelligence Research Labs
MIR Labs
Auburn, Washington 98071-2259, USA
hesam.izakian@gmail.com, ajith.abraham@ieee.org

Václav Snášel
Faculty of Electrical Engineering and Computer Science
VSB-Technical University of Ostrava
Ostrava, Czech Republic
vaclav.snasel@vsb.cz

Abstract—**The K-Modes algorithm is one of the most popular clustering algorithms in dealing with categorical data. But the random selection of starting centers in this algorithm may lead to different clustering results and falling into local optima. In this paper we proposed a swarm-based K-Modes algorithm. The experimental results over two well known Soybean and Congressional voting categorical data sets show that our method can find the optimal global solutions and can make up the K-Modes shortcoming.**

Keywords-clustering; categorical data; swarm based optimization

I. Introduction

Cluster analysis is the process of grouping objects into sets of disjoint classes called clusters so that within the same class the characteristics of objects are similar to one another, while in different classes the characteristics of objects are dissimilar. Clustering techniques are applied in many application areas such as data mining, pattern recognition, bioinformatics etc.

Generally in clustering algorithms, the data objects are represented in Euclidean space and the clustering objective is to minimize the sum of squared distance from all objects in a cluster domain to the cluster center. K-Means is one of the most popular clustering algorithms and is efficient in dealing with a large amount of data. This algorithm partitions objects into k clusters where the number of clusters, k, is decided in advance according to application purposes. The K-Means algorithm aims at minimizing the sum of dissimilarities (squared distances) between all objects in each cluster to the cluster center. This algorithm is based on alternating two procedures. The first is the assignment of objects to the clusters. This is done by randomly generating k cluster centers as starting centers. Then an object is usually assigned to the cluster which is closest in the Euclidean sense. The second procedure is the calculation of new cluster centers based on the assignments. The mean of objects which belong to each cluster will be considered as new cluster centers. The process terminates when no movement of an object to another cluster will reduce the within-cluster sum of squares. The K-Means algorithm is sensitive to selection of initial cluster centers and it may converge to local optima.

The K-Means algorithm only can work with numeric values and it can not be used to cluster real world data including categorical values. For solving this shortcoming, Huang [1] extends the K-Means algorithm to categorical domains called K-Modes algorithm. The K-Modes algorithm uses a simple matching dissimilarity measure to deal with categorical objects, replaces the means of clusters with modes, and uses a frequency-based method to update modes in the clustering process to minimize the clustering cost function. These extensions have removed the numeric-only limitation of the K-Means algorithm and enable it to be used to efficiently cluster large categorical data sets from real-world databases. The K-Modes algorithm is sensitive to the selection of initial cluster centers similar to the K-Means and may converge to local optima.

Nowadays meta-heuristic algorithms have gained much attention as global optimization tools and are used in wide application areas. Some of the most popular meta-heuristic algorithms are genetic algorithm (GA), simulated annealing (SA), particle swarm optimization (PSO), and ant colony optimization (ACO). Recently some of these algorithms are used to make up the shortcoming of K-Means and other clustering algorithms.

In this paper we propose a swarm-based K-Modes algorithm to make up the shortcoming of K-Modes. The rest of the paper is organized in the following manner. In Section 2, we investigate the related works and section 3 introduces K-Modes clustering algorithm, in Section 4 the swarm intelligence based methods is briefly described, in Section 5 our proposed swarm-based K-Modes algorithm is discussed, and section 6 reports the experimental results. Finally section 7 concludes this work.

II. Related Works

Yuqing et al. [2] used ACO to improve the K-Means algorithm. In this algorithm, the ants move objects in the 2D board frequently according to similarity. The proposed method uses the capability of ant colony algorithm to avoid clustering getting into local optimality, and it also avoids sensibility of the initial partition of K-Means algorithm. Zhenkui et al. [3] used the global optimization ability of PSO to avoid convergence the K-Means algorithm to local optima.

In [4] the SA is used. The proposed method views the clustering as optimization problem. At first the K-Means splits the dataset into k clusters, and then runs simulated annealing algorithm using the sum of distances between each pattern and its centre based on K-Means as the aim function. Also in [5] a genetic k-Modes algorithm that finds a globally

978-1-4244-5612-3/09/$26.00 ©2009 IEEE

optimal partition of a given categorical data set into a specified number of clusters is proposed in which a K-Modes operator in place of the normal crossover operator is introduced.

In [8] authors presented the genetic fuzzy k-Modes algorithm for clustering categorical data sets. They treated the fuzzy k-Modes clustering as an optimization problem and used GA to solve the problem in order to obtain globally optimal solution. To speed up the convergence process of the algorithm, a one-step fuzzy k-Modes algorithm in the crossover process instead of the traditional crossover operator is used.

III. K-Modes Algorithm for Clustering Categorical Data

Assume the set of data objects that is to be clustered is defined by a set of attributes $A_1, A_2, \ldots, A_d$. Each attribute $A_j\ (1 \le j \le d)$ describes a domain of n_j categorical values denoted by $DOM(A_j) = \{a_{j1}, a_{j2}, \ldots, a_{jn_j}\}$. An object X can be represented as a conjunction of attribute-value pairs $[A_1 = x_1] \wedge [A_2 = x_2] \wedge \ldots \wedge [A_d = x_d]$, where $x_j \in DOM(A_j)$. Let $D = \{X_1, X_2, \ldots, X_n\}$ is a set of n objects that is to be clustered. Object X_i is represented as $[x_{i1}, x_{i2}, \ldots, x_{id}]$ and the cluster centers are represented by $Z_l = \{z_{l1}, z_{l2}, \ldots, z_{ld}\}$ for $1 \le l \le k$ where k is the number of clusters. Assume X and Y are two categorical data objects in D, the simple matching distance measure between X and Y is defined as

$$d_c(X, Y) = \sum_{j=1}^{d} \delta(x_j, y_j) \tag{1}$$

where x_j and y_j are the jth component of X and Y respectively and

$$\delta(x_j, y_j) = \begin{cases} 0 & \text{if } x_j = y_j \\ 1 & \text{otherwise} \end{cases} \tag{2}$$

then the objective of K-Modes clustering is to find W and Z that minimize

$$F_c(W, Z) = \sum_{l=1}^{k} \sum_{i=1}^{n} w_{li} d_c(X_i, Z_l), \tag{3}$$

subject to (4) , (5), and 6.

$$w_{li} \in \{0, 1\}, \quad 1 \le l \le k \ , \quad 1 \le i \le n \tag{4}$$

$$\sum_{l=1}^{k} w_{li} = 1, \quad 1 \le i \le n \tag{5}$$

$$0 < \sum_{i=1}^{n} w_{li} < n , \quad 1 \le l \le k \tag{6}$$

To obtain the $W = (w_{li}), 1 \le l \le k,\ 1 \le i \le n$ which minimizes the $F_c(W, Z)$ we have

$$w_{li} = \begin{cases} 1 & \text{if } d_c(Z_l, X_i) < d_c(Z_h, X_i) \\ & \forall\ h \in \{1,2,...,k\}\ and\ l \ne h \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

This means that each object belongs to the cluster which is nearest to its center based on matching distance measure (1). Also to update cluster centers given $W = (w_{li}), 1 \le l \le k,\ 1 \le i \le n$ for minimizing $F_c(W, Z)$ we have $z_{lj} = a_{jr} \in DOM(A_j)$, where

$$\left|\{w_{li} \mid x_{ij} = a_{jr}, w_{li} = 1\}\right| \ \ge\ \left|\{w_{li} \mid x_{ij} = a_{jt}, w_{li} = 1\}\right|, \quad 1 \le t \le n_j, \quad 1 \le j \le d \tag{8}$$

Here, $|X|$ denotes the number of elements in the set X. The K-Modes clustering algorithm can be stated as follows [1].

Algorithm 1. K-Modes clustering algorithm.

```
1: Initialize the cluster centers $Z = \{Z_1, Z_2, \ldots, Z_k\}$.
2: Determine $W = (w_{li}), 1 \le l \le k,\ 1 \le i \le n$ using (7).
3: Determine new cluster centers using (8).
4: If not converged go to step 2.
```

As mentioned in the previous section, the K-Modes algorithm is sensitive to initialization and the random selection of starting centers may lead to different clustering results and convergence to local optima.

IV. Swarm Intelligence

Studies of the social behavior of organisms in swarms prompted the design of very efficient optimization algorithms. Swarm intelligence (SI) is a type of artificial intelligence based on the collective behavior of decentralized, self-organized systems which has gained much attention and wide applications especially in optimization problems. Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) are two most popular swarm based algorithms. The first is inspired from the social behavior of the bird flocks and fish schools to find food [6] and the second is inspired from real ant colonies behavior to find the shortest path from a food source to the nest without using visual cues by exploiting pheromone information [7]. In general, the swarm based methods are based on generation number of potential solutions called individuals (or referred to as particles in PSO and ant in ACO). These algorithms try to achieve more efficient solutions in some iterations/generations. To estimate the efficiency of the generated solutions generally an objective function is used which is related to the problem being solved. Each individual is adjusted in each iteration/generation based on some parameters, such as its own experience (for example in PSO, the best solution found by the individual and in ACO, using a predefined heuristic) or the swarm experience (in PSO, the best solution found by the swarm and in ACO, the deposited pheromone) or both. These algorithms are terminated after some predefined iterations/generations or based on some other termination conditions.

V. Proposed Swarm Based K-Modes Algorithm For Clustering Categorical Data

One of the key issues in designing a successful swarm based algorithm is the representation step which aims at finding an appropriate mapping between problem solution and individuals. Since we aim at finding the cluster centers for minimizing the objective function of the K-Modes algorithm, each individual contains *k* cluster centers. Here an individual is a vector of categorical values of dimension $k \times d$, where *k* is the number of clusters and *d* is the dimension of categorical data to be clustered. An individual can be shown as Figure 1 with the following constraints:

$$\begin{aligned} &z_{11}, z_{21}, \cdots, z_{k1} \in DOM(A_1) \\ &z_{12}, z_{22}, \cdots, z_{k2} \in DOM(A_2) \\ &\vdots \\ &z_{1d}, z_{2d}, \cdots, z_{kd} \in DOM(A_d) \end{aligned} \tag{9}$$

z_{11}	z_{12}	•••	z_{1d}	•••	z_{k1}	z_{k2}	•••	z_{kd}

Figure 1. The representation of an individual

In the first step of the algorithm, *P* individuals are randomly generated based on constraint (9). We termed each individual as *Individual* in which $Individual^{q}_{lj}$ is the element z_{lj} (*j*th element in *l*th cluster center) in *q*th individual. To estimate the objective function value for each individual $W = (w_{li}), 1 \le l \le k,\ 1 \le i \le n$ should be obtained using (7) and then we use Eq (3) as objective function. The smaller the objective function is, the better solution is found. After estimating the objective function value for each individual, the best one (individual with the lowest objective function value) is selected as the best individual termed *Best* in which $Best_{lj}$ donates the *j*th element in *l*th cluster center in individual *Best*. Similar to the K-Modes algorithm, Eq (8) can be used for updating the individuals in each iteration/generation and therefore the new cluster centers can be generated. But this leads to falling into local optima the same as K-Modes algorithm. In our algorithm the following method is used to update individuals in each iteration/generation:

First $W = (w_{li}), 1 \le l \le k,\ 1 \le i \le n$ is obtained for each individual using Eq. (7). Then the elements of each individual should be updated. Consider z_{lj} in *q*th individual ($Individual^{q}_{lj}$). For updating this element, r_{lj} which is a random number in range (0, 1) should be generated. If $r_{lj} \le r0$ then $Individual^{q}_{lj}$ should be replaced with $Best_{lj}$ (the corresponding element in individual *Best* which found in previous iteration/generation). Otherwise the probability of selecting each categorical value in $DOM(A_j)$ should be obtained and one of them should be selected using roulette wheel selection. The probability of selecting $a_{jr} \in DOM(A_j)$ can be obtained using (10).

$$p(a_{jr}) = \frac{\left|w_{li} \mid x_{ij} = a_{jr}, w_{li=1}\right|}{\sum_{t=1}^{n_j} \left|w_{li} \mid x_{ij} = a_{jt}, w_{li=1}\right|} \tag{10}$$

Where $|X|$ denotes the number of elements in the set X. Also $r0$ is a number in range [0, 1] and can be obtained using (11) in each iteration/generation.

$$r0(t) = r0(1) + \frac{r0(n_t) - r0(1)}{n_t} \times t \tag{11}$$

Where $r0(1)$, $r0(n_t)$, and $r0(t)$ are the initial value of $r0$, the final value of $r0$, and the value of $r0$ at the iteration/generation *t* respectively. Also n_t is the number of iterations/generations.

The value of $r0$ is very important to ensure convergent behavior and to optimally tradeoff exploration and exploitation. It starts with a small value (e.g. 0.1) which increases over time to larger values so that in the last iteration/generation it ends to a large value (e.g. 0.9). In doing so, individuals are allowed to explore in the initial search steps, while favoring exploitation as time increased. The pseudo code of the swarm-based K-Modes algorithm is shown in Figure 2.

VI. Experimental Results And Discussion

The K-Modes algorithm is coded in C++ programming language as well as our proposed method. Two data sets are used to test the efficiency and feasibility of our algorithm. Also we assume that the number of clusters, k, is known in advance by K-Modes and our proposed method (supervised clustering).

```
Initialize the parameters. $P$: number of individuals, $n_t$ : number of iteration/generation, $r0(1)$ : initial value of $r0$ , $r0(n_t)$ : final value of $r0$ ;
Create $P$ individual each having $k \times d$ element (Figure 1) and initialize each of which randomly and based on constraint (9);
for each iteration/generation $t = 1,2,...,n_t$ do
    for each individual $q = 1,2,...,P$ do
        Calculate $W = (w_{li}), 1 \le l \le k,\ 1 \le i \le n$ using (7);
        Calculate objective function of $Individual^{q}$ using (3);
    end
    Select the individual with minimum objective function value as individual Best;
    for each individual $q = 1,2,...,P$ do
        for each element $z_{lj}, 1 \le l \le k,\ 1 \le j \le d$ do
            $r_{lj}$ =rand (0, 1);
            if $r_{lj} \le r0(t)$ then
                $Individual^{q}_{lj} = Best_{lj}$ ;
            else
                for each categorical value $a_{jr}, 1 \le r \le n_j$ in $DOM(A_j)$ do
                    Calculate $p(a_{jr})$ using Eq (10);
                end
                Select a value in $DOM(A_j)$ using roulette wheel selection;
            end
        end
    end
    Calculate $r0(t)$ using Eq (11);
end
```

Figure 2. The pseudo code of the swarm-based K-Modes algorithm

A. Parameter settings

In order to optimize the performance of the proposed method fine tuning has been performed. Experimental results show that our proposed method performs best under following settings: P=10, $n_t = 400$, $r0(1) = 0.1$, and $r0(n_t) = 0.9$.

B. Data sets

The first data set is the soybean data set which has 47 objects each one is described by 35 attributes. Since there are 14 attributes that have only one category, we only selected other 21 attributes for the clustering. The records are labeled as one of the four diseases: diaporthe stem rot, charcoal rot, rhizoctonia root rot all having 10 instances each, and phytophthora rot which has 17 instances.

The second data set is the Congressional voting data set which includes votes for each of the US House of Representatives Congressmen on the 16 key votes identified by the CQA. It has 435 objects (267 democrats, 168 republicans) each of which is described by 16 binary attributes. Also some of the objects have missing values, we denote the missing value by "?" and treat it as an additional category for that attribute [8]. These two categorical data sets are available at ftp://ftp.ics.uci.edu./pub/machine-learning-databases/.

C. Results

The experimental results over 100 independent runs for K-Modes and 10 independent runs for Swarm-based K-modes are given in Table 1 based on the objective function value (Eq (3)). As shown in this Table our method is stable over these data sets and in all cases achieved the best results. Table 2 shows the average accuracy of the compared methods to clustering categorical objects. Also Table 3 and table 4 show the misclassification matrix of our proposed method over Soybean and Voting data sets respectively.

TABLE I. THE SUMMARY OF RESULTS FOR COMPARED METHODS BASED ON OBJECTIVE FUNCTION (EQ. (3))

Data set	K-Modes			Swarm-based K-Modes		
	Worst	*Average*	*Best*	*Worst*	*Average*	*Best*
Soybean	227	207.4	199	199	199	199
Voting	1706	1702	1701	1701	1701	1701

TABLE II. THE AVERAGE ACCURACY OF COMPARED METHODS IN CLUSTERING OBJECTS

Data set	K-Modes	Swarm-based K-Modes
Soybean	87.9%	100%
Voting	85.5%	86.4%

TABLE III. THE MISCLASSIFICATION MATRIX OF THE PROPOSED METHOD ON THE SOYBEAN DATA SET

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Diaporthe stem rot	10	0	0	0
Charcoal rot	0	10	0	0
Rhizoctonia root rot	0	0	10	0
Phytophthora rot	0	0	0	17

TABLE IV. THE MISCLASSIFICATION MATRIX OF THE PROPOSED METHOD ON THE VOTING DATA SET

	Cluster 1	Cluster 2
Democrat	153	15
Republican	44	223

Table 3 shows that our method dose not misclassifies the Soybean data set objects. Also we see from Table 4 that 44 + 15= 59 out of 435 objects are misclassified.

VII. Conclusion

The K-Modes algorithm is one of the most popular algorithms for clustering categorical data because of its feasibility and efficiency in dealing with large data sets. But the random selection of starting centers in this algorithm may lead to different clustering results and falling into local optima. To make up the shortcoming of this method we proposed a swarm-based K-Modes algorithm. This method uses the ability of swarm-based algorithms to explore the global optimal solutions. Experimental results over two well known Soybean and Congressional voting categorical data sets show that our method can find the optimal global solutions and cluster these data sets more efficient than K-Modes algorithm.

References

[1] Z. Huang, Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values, Data Mining and Knowledge Discovery 2, 283–304 (1998).

[2] P. Yuqing, H. Xiangdan, L. Shang, The K-Means Clustering Algorithm Based on Density and Ant Colony, IEEE Int. Conf. Neural Networks & Signal Processing, pp. 457-460, 2003.

[3] P. Zhenkui, H. Xia ,H. Jinfeng, The Clustering Algorithm Based on Particle Swarm Optimization Algorithm, International Conference on Intelligent Computation Technology and Automation, IEEE CS Press, pp.148-151, 2008.

[4] J. Dong and M. Qi, K-means Optimization Algorithm for Solving Clustering Problem, Second International Workshop on Knowledge Discovery and Data Mining, IEEE Press, pp. 52-55, 2009 .

[5] G. Gan, Z. Yang, J. Wu, A Genetic k-Modes Algorithm for Clustering Categorical Data, Springer-Verlag Berlin Heidelberg, pp. 195-202, 2005.

[6] J. Kennedy, R.C. Eberhart, Particle swarm optimization, in: Proceedings of the IEEE International Conference on Neural Networks (1995) 1942–1948.

[7] M. Dorigo, Optimization, Learning and Natural Algorithms, PhD Thesis, Politecnico di Milano, Italy,1992.

[8] G. Gan, J. Wu, Z. Yang, A genetic fuzzy k-Modes algorithm for clustering categorical data, Expert Systems with Applications 36 (2009) 1615–1620.