Initialize the centers of categorical data cluster using genetic approach: A Method

Kusha Bhatt
M. Tech Scholar, Department of CSE.
Shrinathji Institute of Technology
Nathdwara, India
kusha.bhatt@gmail.com

Prof. Pankaj Dalal
Department of CSE.
Shrinathji Institute of Technology
Nathdwara, India
pkjdalal@gmail.com

Abstract: **Clustering is a challenging task in data mining technique. The aim of clustering is to group the similar data into number of clusters. Various clustering algorithms have been developed to group data into clusters. The leading partitioned clustering technique, k-modes, is one of the most computationally efficient clustering methods for categorical data. However, the performance of the k-modes clustering algorithm which converges to numerous local minima strongly depends on initial cluster centers. Currently, most methods of initialization cluster centers are mainly for numerical data. Due to lack of geometry for the categorical data, these methods used in cluster centers initialization for numerical data are not applicable to categorical data. This research proposes a novel initialization method for categorical data which is implemented to the k-modes algorithm using genetic algorithm**

Keywords: K-Modes, Genetic Algorithm, Categorical data, clustering.

I. INTRODUCTION

Clustering categorical data, i.e., data in which attribute domains consist of discrete values that are not ordered, is a fundamental problem in data analysis. Despite many advances and the vast literature in the clustering of data objects with numerical domains, clustering categorical data, where there is neither a natural distance metric nor geometric interpretation for clusters, remains a significant challenge. In addition, categorical clustering presents many of the same difficulties found in clustering numerical data, e.g., high dimensionality, large data sets and the high computational complexity associated with the discrete clustering problem. Moreover, to be effective most algorithms for clustering categorical data often require the careful choice of parameter values, which makes these algorithms difficult to use by those not thoroughly familiar with the methods

Clustering is unsupervised learning that aims at partitioning a data set into groups of similar items. The goal is to create clusters of data objects where the within-cluster similarity is maximized (intra-cluster similarity) and the between-cluster similarity is minimized (inter-cluster similarity). One of the stages in a clustering task is selecting a clustering strategy. In this stage, a particular clustering algorithm is selected that is suitable for the data and the desired clustering type. Selecting a clustering algorithm is not an easy task and requires the consideration of several issues such as data types, data set size and dimensionality, data noise level, type or shape of expected clusters, and overall expected clustering quality. Over the past few decades, many clustering algorithms have been proposed that employ a wide range of techniques such as iterative optimization, probability distribution functions, density-based concepts, information entropy, and spectral analysis

Clustering – an Overview

Prior to discussing specific algorithms for categorical data, we provide a brief discussion of the elements that are considered when designing clustering algorithms. We provide a description of data types, proximity measures, and objective functions.

Data Types

The first stage in data clustering is data collection (Jain and Dubes, 1988)[4]. In this stage, a determination of what data to collect and their initial data type is made. Each dimension represents an attribute, a feature, or an observation. The value of an attribute can be classified as follows:

- Quantitative. These attributes contain continuous numerical quantities where a natural order exists between items of the same data type and an arithmetically based distance measure can be defined. Height, weight, and length are some examples.

- Qualitative. These attributes contain discrete data whose domain is finite. We refer to the items in the domain of each attribute as categories. Qualitative data are further subdivided as:

  - ○ Nominal. Data items belonging to this group do not have any inherent order or proximity. We refer to these data as categorical data. Attributes such as color, shape, and city names are some examples of this data type.

  - ○ Ordinal. These are ordered discrete items that do not have a distance relation. Examples are ranks or ratings.

- Binary attributes are attributes that can take on only two values: 1 or 0.

Depending on the context, binary attributes can be qualitative or quantitative.

Proximity Measures

Proximity measures are metrics that define the similarity or dissimilarity between data objects for the purpose of determining how close or related the data objects are. There are various approaches to defining proximity measures. These approaches vary from one application area to another, and depend on the data type. For most algorithms, these proximity measures are used to construct a proximity matrix that reflects the distance or similarity between the data objects. These matrices are used as input for a clustering algorithm that clusters the data according to a partitioning criterion or an objective function. For example, in some of the graph-based algorithms, the input to the algorithm is the graph adjacency matrix and the goal is to partition the graph by finding the minimum cut of the graph. In this section, we discuss some of the well-known proximity measures.

II. K-MEANS AND K-MODES ALGORITHMS

The k-means algorithm (Anderberg, 1973; Ball & Hall, 1967; MacQueen, 1967; Jain & Dubes, 1988)[23][24][25][4] is a well-known partitioned clustering algorithm which is widely used in real world applications such as marketing research and data mining to cluster very large data sets due to their efficiency. In 1997 Huang (1997, 1998)[1][2], extended the k-means algorithm to propose the k-modes algorithm whose extensions have removed the numeric-only limitation of the k-means algorithm and enable the k-means clustering process to be used to efficiently cluster large categorical data sets from real world databases. Since first published, the k-modes algorithm has become a popular technique in solving categorical data clustering problems in different application domains (Andreopoulos, An, & Wang, 2005)[3].

The k-means algorithm and the k-modes algorithm use alternating minimization methods to solve non convex optimization problems in finding cluster solutions (Jain & Dubes, 1988)[4]. These algorithms require a set of initial cluster centers to start and often end up with different clustering results from different sets of initial cluster centers. Therefore, these algorithms are very sensitive to the initial cluster centers. Usually, these algorithms are run with different initial guesses of cluster centers, and the results are compared in order to determine the best clustering results. One way is to select the clustering results with the least objective function value formulated in these algorithms, see, for instance (Huang, Ng, Rong, & Li, 2005)[5]. In addition, cluster validation techniques can be employed to select the best clustering result, see, for instance (Jain & Dubes, 1988)[4]. Other approaches have been proposed and studied to address this issue by using a better initial seed value selection for the k-means algorithm (Arthur & Vassilvitskii, 2007; Babu & Murty, 1993; Brendan & Delbert, 2007; Bradley, Mangasarian, & Street, 1997; Bradley & Fayyad, 1998; Khan & Ahmad, 2004; Krishna & Murty, 1999; Laszlo & Mukherjee, 2006, 2007; Pen, Lozano, & Larraaga, 1999)[6][7][[8][9][10][11][12][13][14][15]. For example, some experts (Babu & Murty, 1993; Krishna & Murty, 1999; Laszlo & Mukherjee, 2006, 2007)[7][11][13][14] used genetic algorithm to obtain the good initial cluster centers. Arthur and Vassilvitskii (2007)[6] proposed and studied a careful seeding for initial cluster centers to improve clustering results.

However, due to lack of intuitive geometry for categorical data, the techniques used in cluster centers initialization for numerical data are not applicable to categorical data. To date, few researches are concerned for cluster centers initialization for categorical data. However, due to the fact that large categorical data sets exist in many applications, it has been widely recognized that directly clustering the raw categorical data is important. Examples include environmental data analysis (Wrigley, 1985), market basket data analysis (Aggarwal, Magdalena, & Yu, 2002)[16], DNA or protein sequence analysis (Baxevanis & Ouellette, 2001)[18], text mining (Wang & Karypis, 2006)[18], and computer security (Barbara & Jajodia, 2002). Therefore, how to select initial cluster centers for clustering categorical data become an important research question. The k-centers clustering technique.

Huang in Huang (1998)[20] suggested to select the first k distinct objects from the data set as the initial k modes or assign the most frequent categories equally to the initial k modes. Though the methods are to make the initial modes diverse, an uniform criteria is not given for selecting k initial modes in Huang (1998)[2]. Sun, Zhu, and Chen (2002) [19] introduces an initialization method which is based on the frame of refining. This method presents a study on applying Bradley’s iterative initial-point refinement algorithm (Bradley & Fayyad, 1998)[10] to the k-modes clustering, but its time cost is high and the parameters of this method are plenty which need to be asserted in advance. In Coolcat algorithm (Barbara, Couto, & Li, 2002)[20] , the MaxMin distances method is used to find the k most dissimilar data objects from the data set as initial seeds. However, the method only considers the distance between the data objects, by which outliers maybe be selected. Cao, Liang, and Bai (2009)[21] and Wu, Jiang, and Huang (2007)[22] integrated the distance and the density together to propose a cluster centers initialization method, respectively. The difference between the two methods is the definition of the density of an object. Wu used the total distance between an object and all objects from data set as the density of the object. Due to the fact that the time complexity ofcalculating the densities of all objects is O(n^2), it limits the process in a sub-sample data set and uses a refining framework. But this method needs to randomly select sub-sample, so the sole clustering result cannot be guaranteed. Cao et al. (2009) defined the density of an object based on frequency of attribute values. In this paper, we prove that Cao's density is equivalent to Wu's density, which means that Cao's method is equivalent to Wu's method. Although the two methods can avoid to select outliers as the cluster centers by the density, they have some shortcomings: (1) the object with the maximum density is taken as the first cluster center. Due to the fact that they only considered the factor of density in the selection of the first cluster center, it is possible that the selected object is a boundary point among clusters, which is proved in this paper; (2) one real object in a cluster is selected as the cluster center. But in most cases, the center of a cluster is not a real object but a virtual object, which means that a real object could not sufficiently represent the cluster. In summary, there are no universally accepted method for obtaining initial cluster centers currently. Hence, it is very necessary to propose a new initialization method for categorical data which overcomes shortcomings of the existing initialization methods

The k-means algorithm has the following important properties:

1. It is efficient in processing large data sets.

2. It often terminates at a local optimum.

3. It works only on numeric values.

4. The clusters have convex shapes.

The barriers of can be removed by making the following modifications to the k-means algorithm which help in formulation of k-mode algorithms:

1. Using a simple matching dissimilarity measure for categorical objects,

2. Replacing means of clusters by modes, and

3. Using a frequency-based method to find the modes to solve problem.

III. GENETIC ALGORITHMS

In a genetic algorithm, a population of candidate solutions (called individuals, creatures, or phenotypes) to an optimization problem is evolved toward better solutions. Each candidate solution has a set of properties (its chromosomes or genotype) which can be mutated and altered; traditionally, solutions are represented in binary as strings of 0s and 1s, but other encodings are also possible.

The evolution usually starts from a population of randomly generated individuals and happens in generations. In each generation, the fitness of every individual in the population is evaluated, the more fit individuals are stochastically selected from the current population, and each individual's genome is modified (recombined and possibly randomly mutated) to form a new population. The new population is then used in the next iteration of the algorithm. Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population.

A typical genetic algorithm requires:

- Genetic representation of the solution domain
- Fitness function to evaluate the solution domain

The basic genetic algorithm is as follows:

- **Start** - Genetic random population of n chromosomes (suitable solutions for the problem)
- **Fitness** - Evaluate the fitness f(x) of each chromosome x in the population
- **New population** - Create a new population by repeating following steps until the New population is complete
- **Selection -** Select two parent chromosomes from a population according to their fitness (the better fitness, the bigger chance to get selected).
- **Crossover** - With a crossover probability, cross over the parents to form new offspring (children). If no crossover was performed, offspring is the exact copy of parents.
- **Mutation** - With a mutation probability, mutate new offspring at each locus (position in chromosome)
- **Accepting** - Place new offspring in the new population.
- **Replace** - Use new generated population for a further sum of the algorithm.
- **Test** - If the end condition is satisfied, stop, and return the best solution in current population.
- **Loop -** Go to step2 for fitness evaluation.

IV. CONCLUSION

Categorical data are ubiquitous in real-world databases. The development of the k-modes algorithm was motivated to solve this problem. However, the clustering algorithm need to rerun many times with different initializations in an attempt to find a good solution. Moreover, this works well only when the number of clusters is small and chances are good that at least one random initialization is close to a good solution. In this work, a new initialization method for categorical data clustering has been proposed by optimizing the

distance between the objects and the density of the object and overcomes shortcomings of the existing initialization methods using genetic algorithm. Furthermore, the time complexity of the proposed method will also have to analyse. We have to test the proposed method using seven real world data sets from UCI Machine Learning Repository and experimental results of the proposed method will superior to other initialization methods in the k- modes algorithm.

REFERENCES

[1] Huang, Z.X. (1997). A fast clustering algorithm to cluster very large categorical data sets in data mining. In Proceeding SIGMOD workshop research issues on data mining and knowledge discovery (pp. 1–8).,

[2] Huang, Z. X. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining Knowledge Discovery, 2(3), 283–304.

[3] Andreopoulos, B., And, A., & Wang, X. (2005). Clustering the internet topology at multiple layers. WSEAS Transactions on Information Science and Applications, 2, 1625–1634.

[4] Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data. Prentice Hall.

[5] Huang, Z. X., Ng, M., Rong, H., & Li, Z. (2005). Automated variable weighting in k-means type clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(5), 657–668.

[6] Arthur, D., & Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. In Proceedings 18th annual ACM-SIAM symposium on discrete algorithms (SODA'07) (pp. 1027–1035).

[7] Babu,G.P.,&Murty,M.N.(1993). A near-optimal initial seed value selection for k-means algorithm using genetic algorithm. Pattern Recognition Letters, 14, 763–769.

[8] Brendan, J. F., & Delbert, D. (2007). Clustering by passing messages between data points. Science, 15(16), 972–976.

[9] Bradley, P. S., Mangasarian, O. L., & Street, W. N. (1997). Clustering via concave minimization. In M. C. Mozer, M. I. Jordan, & T. Petsche (Eds.). Advances in neural information processing system (Vol. 9, pp. 368–374). MIT Press.

[10] Bradley, P. S., & Fayyad, U. M. (1998). Refining initial points for k-means clustering

[11] Khan, S. S., & Ahmad, A. (2004). Cluster center initialization algorithm for k-means clustering. Patter Recognition Letters, 25, 1293–1302.

[12] Krishna, K., & Murty, M. N. (1999). Genetic k-means algorithm. IEEE Transactions on Systems, Man, and Cybernetics, 29(3), 433–439.

[13] Laszlo, M., & Mukherjee, S. (2006). A Genetic algorithm using hyper-quad-trees for low-dimensional k-means clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(4), 533–543.

[14] Laszlo, M., & Mukherjee, S. (2007). A genetic algorithm that exchanges neighbouring centers for k-means clustering. Pattern Recognition Letters, 28(16), 2359–2366.

[15] Pen, J. M., Lozano, J. A., & Larraaga, P. (1999). An empirical comparison of four initialization methods for the k-means algorithm. Pattern Recognition Letter, 20, 1027–1040.

[16] Aggarwal, C. C., Magdalena, C., & Yu, P. S. (2002). Finding localized associations in market basket data. IEEE Transactions on Knowledge and Data Engineering, 14(1), 51–62.

[17] Dordrecht: Kluwer. Baxevanis, A., & Ouellette, F. (2001). Bioinformatics: A practical guide to the analysis of genes and proteins (2nd Ed.). NY: Wiley.

[18] Wang, J., & Karypis, G. (2006). On efficiently summarizing categorical databases. Knowledge and Information Systems, 9(1), 19–37.

[19] Sun, Y., Zhu, Q. M., & Chen, Z. X. (2002). An iterative initial-points refinement algorithm for categorical data clustering. Pattern Recognition Letters, 23, 875–884. UCI Machine Learning Repository (2010). <http://www.ics.uci.edu/mlearn/MLRepository.html>.

[20] Barbara, D., Couto, J., & Li, Y. (2002). COOLCAT: An entropy-based algorithm for categorical clustering. In Proceedings of the eleventh international conference on information and knowledge management (pp. 582–589).

[21] Cao, F. Y., Liang, J. Y., & Bai, L. (2009). A new initialization method for categorical data clustering. Expert Systems with Applications, 33(7), 10223–10228.

[22]Wu, S., Jiang, Q. S., & Huang, Z. X. (2007). A new initialization method for categorical data clustering. Lecture Notes in Computer Science, 4426, 972–980.

[23] Anderberg, M. R. (1973). Cluster analysis for applicationsAcademic.

[24] Ball, G. H., & Hall, D. J. (1967). A clustering technique for summarizing multivariate data. Behavioural Science, 12, 153–155.

[25] MacQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of fifth symposium on mathematical statistics and probability (Vol. 1, pp. 281–297).