IOSR Journal of Computer Engineering (IOSRJCE)
ISSN: 2278-0661, ISBN: 2278-8727 Volume 6, Issue 2 (Sep-Oct. 2012), PP 09-13
www.iosrjournals.org

Hybrid Algorithm for Clustering Mixed Data Sets

V.N. Prasad Pinisetty[1], Ramesh Valaboju[2], N. Raghava Rao[3]
[1]*Department of CSE, DRK College Of Engineering& Technology, Hyderabad, Andhra Pradesh, India*
[2]*Department of CS, DRK Institute Of Science & Technology, Hyderabad, Andhra Pradesh, India*
Associate Professor
[3]*Department of IT, DRK Institute Of Science & Technology, Ranga Reddy, Hyderabad, Andhra Pradesh, India*

Abstract: *Clustering is one of the data mining techniques used to group similar objects into different meaningful classes known as clusters. Objects in each cluster have maximum similarity while the objects across the clusters have minimum or no similarity. This kind of partitioning of objects into various groups has many real time applications such as pattern recognition, machine learning and so on. In this paper we review a clustering algorithm based on genetic K-means [1] and compare it with GKMODE and IGKA. The algorithm works well for both numeric and discrete values. The existing genetic K-means algorithms have limitation as they can cluster only numeric data. The algorithm [1] overcomes this problem and provides a better way of characterization of clusters. The empirical results revealed that the performance of the proposed algorithm has been improved. We also make observations and recommendations on the proposed algorithm, GKMODE, and IGKA that help in future enhancements.*

Index Terms *– Generic algorithm, data mining, clustering, mixed data*

I. Introduction

Classifying a group of data objects into different classes is very essential operation in data mining. There are many applications associated with this. They include dissection [2], segmentation and aggregation, classification and machine learning. Generally the data set containing m-dimensional space might have possibility of k unique clusters and each data point within cluster is highly similar to the other objects in the same cluster. The data points when compared with objects of other clusters are not similar or dissimilar. This is the phenomenon that takes place in clustering objects. As discussed in [1] three of the problems that can be solved using clustering technique are when a similarity is measured, it is possible to compare distance between elements; a clustering algorithm helps in obtaining most similar objects into a specific cluster; a description can be derived that can characterize the objects of a cluster in a specific manner. In many traditional clustering algorithms ED (Euclidean Distance) measure is used to judge the distance between objects to be clustered [3], [4]. These algorithms work fine when attributes of data set are of numeric types. Many such algorithms including ED fail to measure distance between objects if the type of attributes is categorical or mixed in nature. K-means is another algorithm that has traditionally been used to cluster numeric values. It is widely used in the data mining domain for various kinds of applications. Still it is in the top 10 clustering algorithms [7]. It has got many variations to meet the changing requirements. It is also traditionally used to work with numeric data though it has variations. As a matter of fact, data available over Internet or academia and industries is abundant in discrete values [5]. Such data is from sectors like banking, health care, education, web-logs, biological and so on. The datasets from these domains are essentially having mixed attributes. It does mean that each and every data set is having some numeric and some categorical values. Clustering such data into meaningful groups of objects is a challenging job. The clustering algorithms ED and K-means that have been discussed are inadequate in their original form to measure distances and cluster such data containing mixed attributes into meaningful classes. This is the motivation for the invention of new clustering algorithms that can handle numeric, discrete or mixed values [6].

Handling mixed content with respect to clustering technique is the area that needs further research. There are certain strategies that can be used to work with mixed data. One of the strategies is the conversion of categorical data into numerical data as pre processing prior to the application of clustering algorithm or as part of algorithm; however, it is difficult to do this because, there are instances where the conversion does not given meaningful numeric data and does not make sense. Another strategy is the reverse process as it is quite opposite to the first strategy. The second strategy allows us to convert numerical data into categorical data and apply any known categorical clustering algorithm. However, the discretization may result in the loss of information. To overcome these problems we have used and reviewed a genetic K-means clustering algorithm [1] which can efficiently group huge datasets with categorical and numeric values as such databases are used in real time data mining applications frequently.

Our contributions in this paper include the review of algorithms such as GKMODE, IGKA and [1] as clustering algorithms that work on numeric and categorical values and observations that can help improve the existing clustering algorithms of that kind.

II. Related Work

Genetic algorithms are basically evolutionary algorithms that are used to solve many real time applications. They are stochastic in nature and can be used to solve many real world problems [5]. They can perform parallel operations on complex search spaces. Evolution strategies, evolutionary programming and genetic algorithms are all part of evolutionary algorithms. Genetic algorithms are used in complex games and applications where certain values are to be evaluated dynamically at runtime. For instance in Master Mind game some values are hidden and the player has to guess values every time. The correctly guessed values are indicated and the other values are to be guessed. This guessing may take place several times until the hidden values are correctly guessed. The same thing can be achieved by using genetic algorithm that makes use of initial population, operators and iterations. Finally it converges into guessing correct values. Thus genetic algorithms are practically used in many real world applications.

Holland [5] originally proposed genetic algorithms. They have been applied to many problems such as function organization problems. They are proved to be good for finding near optimal solutions. Genetic algorithms are domain independent and robust in nature for evolutionary problems. This has motivated many researchers to work on genetic algorithms. Genetic K-means algorithm was proposed by Krishna and Murthy which is the combination of K-means algorithm and also a genetic algorithm. For this reason this algorithm converges to the global optimum faster than only K-means algorithms or only genetic algorithms.

FGKA (Fast Genetic K-means Cluster Technique) was proposed by Lu et al. (8). It has many enhancements over GKA that includes evaluation with Total Within-Cluster Variation (TWCA). The enhancements include mutation operator simplification and avoiding illegal string elimination overhead. Moreover FGKA is faster than other algorithms such as GKA. Nevertheless, the FGKA algorithm has a significant disadvantage. When the mutation probability is small, it gives problems besides making the algorithm more expensive. In order to overcome this issue Lu et al. proposed IGKA (Incremental Genetic K-means Algorithm) which exceeds the FGKA in performance. The incremental nature of the IGKA provides such performance benefits. When compared with FGA, IGKA performs well. From this perspective the a hybrid algorithm is proposed that is the combination of FGKA and IGKA and performs exceptionally well when the mutation probability is more. Many GA based algorithms are based on K-means and they work only for numeric data. Another generic algorithm known as GKMODE works only for discrete data. Therefore in [1] a Genetic K-Means algorithm is proposed that works for both numeric and categorical data. This paper reviews [1] and related algorithms, evaluate [1] with a prototype application and compare its results with that of GKMODE and IGKA. Besides, this paper also provides observations that help enhance the genetic clustering algorithms that work for both discrete and numeric data.

III. Genetic K-Means Algorithm

This algorithm as proposed in [1] is a clustering algorithm that works for both numeric and categorical data. It has been reviewed and experimented in this paper and then compared with other such algorithms known as GKMODE and IGKA. Afterwards a set of observations are made that help in further enhancing the way of evaluation and improvement of these algorithms.

Objective Function

For the purpose of clustering with Genetic K-Means algorithm, N genes and corresponding N patterns. D dimensions vector is associated with each pattern. The goal of algorithm [1] is to make N patterns into a user-defined number of clusters. The TWCV as defined as [6] is

$TWCV = \Sigma^{N}_{n=1} \Sigma^{D}_{d=1} X^2_{nd} - \Sigma^{K}_{k=1} 1/Z_k \Sigma^{D}_{d=1} SF^2_{kd}$

The closest cluster center is computed as

$V(d_i,C_j) = \Sigma^{m\,r-1}_{t=1} (w_t(d^r_{it} - C^r_{jt}))^2 + \Sigma^{mc-1}_{t=1} \Omega (d^c_{it}, C^c_{jt})^2$

Selection Operator

For the selection process, the proportional selection is used. With z independent random experiments the population of next generation is determined. From the current population, each experiment randomly selects a solution based on the probability distribution which is computed as

$P_z = \frac{F(S_z)}{\Sigma^{Z}_{z=1} F(S_z)}$ (z=1,……..Z),

The fitness value of solution is computed as

$F(S_z) = \{ 1.5 \text{ x } V_{max} - V(S_z)$, if S_z is leagal

$e(S_z) \text{ x } F_{min}$, otherwise

The maximum V and minimum V are represented as Vmax and Vmin respectively.

Mutation Operator

Shaking the algorithm out of a local optimum into global optimum is performed by mutation operator. For this purpose the probability distribution is computed as

$$p_k = 1.5 * d_{max}(X_n) - d(X_n,c_k) + 0.5$$

$$\overline{\Sigma^{K}_{k=1} (1.5 * d_{max}(X_n) - d(X_n,c_k) + 0.5)}$$

K-Means Operator

K-means operator is introduced in order to speed up the convergence. When the solution encoded by a1..aN is given, an is replaced by an' for n=1..N concurrently where the number of the cluster whose centroid is closest to Xn in Euclidean distance. If the kth cluster is empty we define d (Xn, ck) = $+\infty$ in order to deal with illegal strings. We also give a new definition that is d (Xn, ck) = 0 if the kth cluster is empty. The new definition is motivated by the fact that this can avoid reassigning all patterns to empty clusters.

IV. Experiments And Results

We have made experiments on the Genetic K-Means algorithm [1] with a prototype application that has been built in Java programming language. The environment used to develop the prototype application include JSE 6.0 (Java Standard Edition), Net Beans IDE that run in a PC having 2GB RAM, 2.9x GHz processor with Windows 7 OS. The experiments are done with real world datasets such as Vote, Iris, and Heart Diseases obtained from KDD cup datasets and UCI repository. To judge results accurately we compare the results from genetic k-means algorithm and compare them with the results obtained from GroundTruth technique. Comparatively good results were found. The experiments and the results done with Heart Diseases dataset are presented in this paper.

Heart diseases dataset used here contains five numeric and seven categorical attributes. The dataset contains data pertaining to healthy people and also heart patients with 164 and 139 data instances respectively. The experiments are made with values such as generation size and population size set to 100 and 50 respectively. The algorithm is run over 100 times and average values are presented in table 1. The mutation probability ranges from 0.001 to 0.1.

CLUSTER NO.	NORMAL	HEART PATIENT
1	130	29
2	34	110

Table 1 – Clusters obtained from heart diseases dataset

The result of genetic K-means algorithm is presented in table 1. The results show that the average number of data elements that are not in expected center is 46 while standard deviation for error is 3.

The clustering accuracy is found by using average convergence concept and the corresponding objective function value over many generations for different probabilities of mutation. As per the observations, in both cases, the proposed algorithm [1] converges faster when compared with other algorithms such as GKMODE and IGKA. The global optimal clustering is converged in just five generations. The convergence results are presented in fig. 1

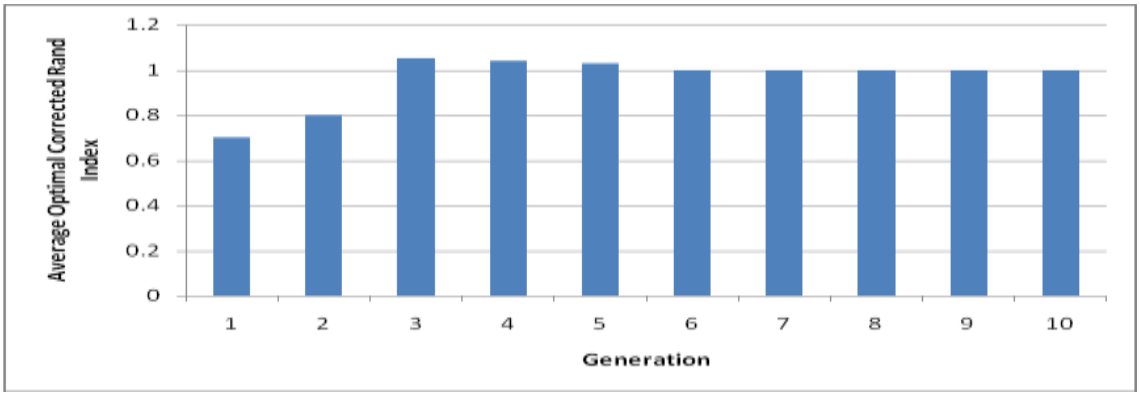

Fig. 1 – Average optimal corrected rand index changes [1]

We also review the comparison of algorithm [1] with GKMODE and IGKA based on the heart diseases dataset with population set to 50 and mutation probability is set to 0.0001.

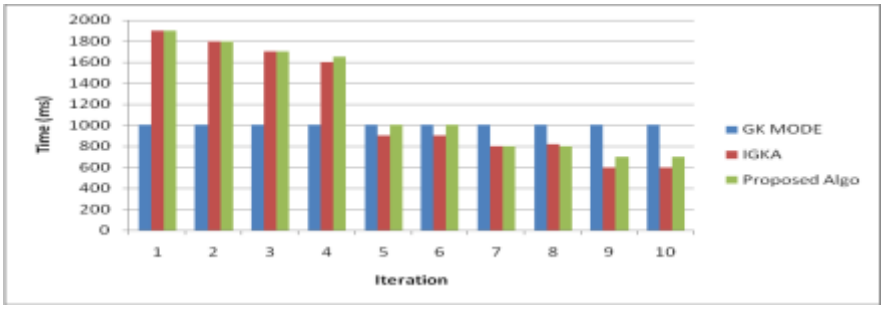

Fig. 2 – The performance comparison of [1], GKMODE and IGKA [1]

As can be seen in fig. 2, the Genetic K-means algorithm shows performance better than that of GKMODE and IGKA in terms of time taken for given iterations.

V. Observations And Recommendations

Keeping the importance of usefulness of clustering data objects with mixed content such as numeric and categorical, we have studied many algorithms including [1], GKMODE and IGKA. We have also built a prototype application that facilitates to evaluate the performance of Genetic K-means algorithm [1] and compare it with other algorithms mentioned. Based on experiments, in this paper we present the following observations.

- Currently the K-means algorithm performance COMPARISON is being made in terms of time taken for given iterations. The only comparison (Time) may not be sufficient for complete analysis of the performance of the algorithm. The introduction of additional measures like number of computations performed, and memory consumption by each algorithm can give more efficient and advanced analysis on the performance of the algorithm when compared with GKMODE and IGKA algorithms. These measures are useful for cost analysis, along with time.
- The algorithms can be enhanced to adapt to various benchmark datasets and the comparison results can be analyzed, and reported to provide more meaningful insights into the study.(e.g. Various clusters formed by the algorithm can be shown in a graphical report for better representation of data sets).

VI. Conclusion

Real world databases always contain numeric and discrete values. The numeric values are generally used in mathematic operations and categorical values can't be used. This is the reason many clustering algorithms fail in working with large datasets that contain attributes of mixed type. The popular Euclidian Distance (ED) and K-means clustering algorithms in their original form fail to work with mixed content. This is the motivation behind taking up this work. In this paper we review many algorithms and strategies such as [1], GKMODE and IGKA with respect to finding their performance in clustering mixed data objects. The clustering algorithm proposed in [1] has been tested with a prototype application and compared with others. This genetic K-means algorithm [1] is more efficient than GKMODE and IGKA. The enhanced cost function and modified cluster center is used in genetic K-means algorithm in order to capture cluster characteristics effectively and efficiently. We also reviewed the additional features encountered in [1] such as minimizing or avoiding string elimination overhead, mutator operator simplification and TWCVs. The observations made here are that the performance comparison can also be made in terms of number of computations performed by the algorithm, and memory consumption for cost analysis (in addition to existing analysis on the Time) by Genetic K-Means Clustering Algorithm when compared with GKMODE and IGKA algorithms. Also the; another observation is that the algorithms can be enhanced to adapt to various benchmark datasets and the comparison results can be analyzed, and reported to provide more meaningful insights into the study. These two observations can pave the way for the future work that can be made on clustering data objects possessing both categorical and numeric data.

Acknowledgements

The authors thank to R.V Krishnaiah, Principal- DRK Institute of Science & Technology. Also we would like to thank K. Praveen, Head of the IT Department for their valuable support.

References

[1] Dharmendra K Roy & Lokesh K Sharma. Genetic K-Means Clustering Algorithm for Mixed Numeric and Categorical Data Sets. International Journal of Artificial Intelligence & Applications (IJAIA), Vol. 1, No. 2, April 2010.

[2] A. Ahmad and L. Dey, (2007), A k-mean clustering algorithm for mixed numeric and categorical data', Data and Knowledge Engineering Elsevier Publication, vol. 63, pp 503-527.

[3] G. Gan, Z. Yang, and J. Wu (2005), A Genetic k-Modes Algorithm for Clustering for Categorical Data, ADMA , LNAI 3584, pp. 195–202.

[4] J. Z. Haung, M. K. Ng, H. Rong, Z. Li (2005) Automated variable weighting in k-mean type clustering, IEEE Transaction on PAMI 27(5).

[5] K. Krishna and M. Murty (1999), 'Genetic K-Means Algorithm'**,** IEEE Transactions on Systems, Man, and Cybernetics vol. 29, NO. 3, pp. 433-439.

[6] Jain, M. Murty and P. Flynn (1999), 'Data clustering: A review', ACM Computing Survey., vol. 31, no. 3, pp. 264–323.

[7] Chaturvedi, P. Green and J. Carroll (2001), k-modes clustering. Journal of Classification, vol 18, pp. 35-55.