ISSN: 2321-7782 (Online)
Volume 3, Issue 2, February 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

A Review on Genetic Algorithm-based Text Clustering Technique

Henal Parmar[1]
Department of Computer Science
Parul Institute of Technology
Gujarat, India

Bhailal Limbasiya[2]
Assistant Professor
Department of Computer Science
Parul Institute of Technology
Gujarat, India

Abstract: ***This paper presents a review on genetic algorithms based on clustering methods. Clustering is an important form of data mining. It can be used to extract useful and hidden information from the datasets. Clustering techniques have a large area of applications including bioinformatics, web use data analysis and image analysis etc. Traditional clustering algorithms applied to datasets most of the times result in sub-optimal solution due to large search space, so evolutionary algorithms particularly genetic algorithms are best suited for the clustering tasks. The capability of Genetic algorithms is applied to find optimally disjoint partitions and proper number of clusters for a dataset. Genetic Algorithms (GAs) is applied to the clustering algorithm to enhance the performance of clustering algorithms. The capability of GAs is applied to evolve the proper number of clusters and to provide appropriate clustering. This paper present some existing GA based clustering algorithms.***

Keywords: Clustering Algorithm, Genetic Algorithm, GA Based Clustering.

I. INTRODUCTION

Data Mining is extracting knowledge from large amount of data. It is a Process of semi-automatically analyzing large databases to find patterns that are valid, novel, useful, and understandable. It is also known as Knowledge Discovery in Databases (KDD) [1]. Data mining consists of six basic types of tasks which are Anomaly detection, Association rule learning, Clustering, Classification, Regression and Summarization. Data Mining is defined as extracting the information from the large amount of data. We can also say that data mining is mining the knowledge from data. Data mining is the process of carefully studying data from different opinions and summarizing it into useful information. It allows users to carefully study data from many different dimensions or angles, label it, and summarize the relationships identified. Technically, data mining is the process of finding relationships or patterns among dozens of fields in large relational information files. Data mining tools predict future trends and behaviors, and the tool also allowing businesses to make proactive, knowledge-driven decisions.

Text Clustering (TC) is the task of automatically classifying unlabelled natural language documents into a predefined set of semantic categories. The basic idea of Text clustering is to divide the similar text into the same class. Using text clustering methods, you can find a large-scale of text set classification system, and it provides a broad view for the set of text. Text clustering is large-scale text data sets which can be grouped into several categories. It is applied at information extraction and also in Web data mining. There are many text Clustering algorithm like self organizing maps algorithm, hierarchical clustering, partitioned clustering, density-based algorithm [2].

This paper is organized as follows: Section 2 and section 3 give a brief introduction on clustering and genetic algorithms respectively. Section 4 contains overview of existing GA based clustering techniques. Section 5 discusses conclusion. Section 6 gives the references.

II. TEXT CLUSTERING METHODS

1) *Hierarchical clustering*

Hierarchical clustering is most common text clustering method, which is used to generate hierarchical nested class. Hierarchical clustering accuracy is relatively high, but when we merge each class, it needs to compare all classes similarity in the global and after that we can select the more similar of two classes, so it's relatively slow. The problem in hierarchical clustering is that once a step is merged, It can't be revoked, so wrong decision is not corrected. Hierarchical clustering methods are divided into bottom-up hierarchical clustering method and top-down hierarchical clustering method [2].

» ***Bottom-up hierarchical clustering method***

Bottom-up hierarchical clustering method merged the cluster. This method starts from a single object, in this it first takes an object as a separate category, and after that it repeatedly merges, more than two appropriate categories. Bottom-up hierarchical clustering process is the process of constructing the tree, and it contains the class hierarchy information, and the similarity among all the classes.

» ***top-down hierarchical clustering method***

Top-down hierarchical clustering method is used for splitting clusters. This method starts from the object's complete works, and it divided into more categories. The typical approach is to construct a minimum spanning tree on similar graphs, and then at each step choosing a side which in the smallest similarity of the spanning tree (or in the farthest distance of the spanning tree) and removing it. Top down method does not have much application and which is complex because it contains large number of computation

2) *Partitioned clustering*

In the partitioned clustering, the data set is divided into k disjoint point sets, so that each sub-set point as far as possible homogeneity. Partitioned clustering is applicable for carrying on the cluster to the small scale's database to discover the each cluster class regards as one cluster. Partitioned clustering methods include classification based on k-means [2].

» ***k-means algorithm***

k-means algorithm is based on the input parameters k. In k-means algorithm the data set is divided into k clusters. The iterative update method is used by this algorithm. In every round, based on k point of reference points were grouped around k clusters, each cluster centroid will be used as a reference point next round of iteration. The clustering effect is getting better with the help of iteration which makes the selected reference point closer to the true cluster centroid.

» ***k- medoid algorithm***

In the k- medoid algorithm First, it randomly select k objects as initial representative points of k clusters, according to the distance of remaining objects and the representative point object, remaining objects are assigned to the nearest cluster. After that, the repeatedly use of non representative point take the place of the representative point, check whether the quality of clustering is improved or not.

3) *Density-based algorithm*

Density-based algorithm can discover the arbitrary shape; simultaneously the algorithm has the natural resisting effect on the noise. This algorithm is considering the density, connectivity and boundary areas of data space. The most common algorithm of the density-based clustering is DENCLUE algorithm. DENCLUE algorithm is certain influence relationship. The influence relationship can be described as a mathematical function between each data point and other adjacent data points [2].

4) Organizing Maps algorithm

As a high-dimensional of clustering and visualization unsupervised learning algorithm, self-organizing maps algorithm is simulates the characteristic of the human brain to the signal processing developed an artificial neural network [2].

III. GENETIC ALGORITHMS

Genetic algorithms are heuristic optimization methods whose mechanisms are analogous to biological evolution. The solutions are called individuals in the Genetic Algorithm, or it is also called chromosomes. First the initial population is generated randomly, after that the selection and variation function are executed in a loop until some termination criterion is reached. Each run of the loop is called a generation. The selection operator is calculated to improve the average quality of the population by giving individuals of higher quality a higher probability to be copied into the next generation. The quality of an individual is measured by a fitness function.

There are two operators for genetic algorithm namely crossover and mutation operators. It is used to generate the off spring of the existing population. First parents have been selected for evolution to the next generation after that, genetic operators are applied to the produce next generation we use crossover and mutation algorithm. User can changed e probability of deploying crossover and mutation operators [3].

IV. OVERVIEW OF GA BASED CLUSTERING ALGORITHMS

Cluster analysis is a technique, which is used to discover patterns and associations within data. More specifically, it is a multivariate statistical procedure that starts with a data set containing information on some variables and attempts to reorganize these data cases into relatively homogeneous groups. One of the major problems encountered by researchers, with regard to cluster analysis that different clustering methods can and do generate different solutions for the same data set. What is needed is a technique that has discovered the most `natural' groups in a data set.

The process of grouping a set of physical or abstract objects into classes of *similar* objects is called clustering. The cluster analysis represents a group of methods whose aim is to classify the investigated objects into clusters. The founders of cluster analysis were Tryon, Ward and James. One of the major problems encountered by researchers while using different clustering methods is that each method can generate different solutions for the same data. Due to this problem we need algorithms which can discover the most `natural' groups in a data set.

K.Krishna and M.N.Murty proposed a novel hybrid genetic k means algorithm (GKA) [4] to find a globally optimal partition of a given data into a specified number of clusters. The proposed GA circumvent expensive crossover operator used to generate valid child chromosomes from parent chromosomes. It hybridized the GA by using a classical gradient descent algorithm used in clustering viz., K-means algorithm. In *genetic K means algorithm* (GKA), K-means operator was defined and used as a search operator instead of crossover. It defined a biased mutation operator specific to clustering called distance-based-mutation. The authors used finite Markov chain theory to prove that the proposed GKA converges to the global optimum. It was also observed that GKA searches faster than some of the other evolutionary algorithms used for clustering.

Fast Genetic K-means Algorithm (FGKA) is an improved version of GKA was proposed in [5]. Experiments indicates that K-means algorithm might converge to a local optimum, both FGKA and GKA always converge to the global optimum. FGKA initializes the population to P0 and obtains the next population by applying selection, crossover and mutation operators and it keeps on evolving until some termination condition is met. Illegal strings are permitted in FGKA during initialization phase, but were considered as the most undesirable solutions by defining their total within cluster variation (*TWCV*s) as infinity (+∞). By allowing illegal strings the overhead of illegal string in the evolution process was avoided and thus improved the time performance of the algorithm as compared to GKA.

Incremental Genetic K-means Algorithm (IGKA) [6] was an extension to previously proposed clustering algorithm, the Fast Genetic K-means Algorithm (*FGKA*). IGKA outperforms FGKA when the mutation probability was small. The main idea of IGKA was to calculate the objective value Total Within-Cluster Variation (TWCV) and to cluster centroids incrementally whenever the mutation probability was small. IGKA inherits the salient feature of FGKA of always converging to the global optimum.

LI Jie, G. Xinbo in [7] has presented a novel clustering algorithm for mixed data sets by modifying the common cost function, trace of the within cluster dispersion matrix. The genetic algorithm was used to optimize the new cost function to obtain valid clustering result.

A hybrid genetic based clustering algorithm, called *HGA-clustering* was proposed in [8] to explore the proper clustering of data sets. This algorithm has achieved harmony between population diversity and convergence speed with the cooperation of tabu list and aspiration criteria

G*enetic Weighted K-means Algorithm* (GWKMA), which is a combination of a genetic algorithm (GA) and a weighted K-means algorithm (WKMA),proposed by Fang-Xiang et al.[9] GWKMA encodes each individual by a partitioning table which uniquely determines a clustering, and employs three genetic operators as selection, crossover, Mutation and a WKMA operator.

SPMD (Single Program Multiple Data) algorithm proposed in [10] combines GA with local searching algorithm – uphill. The hybrid parallel method not only improves the convergence of GA but also accelerates the convergence speed of GA. The SPMD algorithm exploits the parallelism of GA, at the same time, overcomes the premature and poor convergence properties of GA. The algorithm was applied on typical multiple local minima functions, TSP problem and an engineering computation problem QCBED on author developed cluster system THNPSC-1.

V. Conclusion

This paper presents a review on genetic algorithms proposed for the clustering task of data mining. Many GA based clustering algorithms are studied. It is evident from the review done that GAs are highly capable of performing successful clustering and their capabilities of GAs were applied for evolving the proper number of clusters and providing appropriate clustering. The GAs proposed so far have been applied to different types of datasets, small data sets as well as large data sets, simple datasets as well as multivariate datasets. GA based clustering techniques can be used in many application areas like document clustering , image compression, gene expression analysis and text clustering etc. GA was applied on Clustering algorithms like K-means and fuzzy c-means which are mostly distance based clustering algorithms. GA is yet to be applied to other clustering algorithm.

Acknowledgment

I would like to thank Mr. Bhailal Limbasiya (Assistant Professor, CSE Department, PIT) for the valuable guidance and advice. His invaluable guidance has proved to be a key to my success in overcoming challenges that I faced during the course of the research work and report preparation. He inspired me greatly to prepare the research. His willingness to motivate me contributed tremendously to my report preparation. I express very sincere thanks to his for showing me some examples related to the topic of my research.

References

1. Jawai Han and M. Kamber, “Data Mining Concepts and Techniques”, second edition, Elsevior.
2. Fasheng Liu, Lu Xiong “Survey on Text Clustering Algorithm”, 2011 IEEE
3. D.E. Goldberg, “Genetic Algorithms in Search, Optimization and Machine Learning”, Addison-Wesley, New York, 1989.
4. K. Krishna and M. N. Murty, “Genetic K-Means Algorithm”, IEEE Transaction On Systems, Man, And Cybernetics—Part B:CYBERNETICS, Vol. 29, No. 3, June 1999.

5. Yi Lu, Shiyong Lu, FarshadFotouhi , "FGKA: A Fast Genetic K-means Clustering Algorithm", SAC'04 Nicosia, Cyprus. , March 2004 ACM 1-58113-812-1/03/04
6. Yi Lu1, Shiyong Lu1, Farshad Fotouhi1, Youping Deng, d. Susan, J. Brown, "an Incremental genetic K-means algorithm and its application in gene expression data analysis", BMC Bioinformatics 2004
7. LI Jie, G. Xinbo, "A GA-Based Clustering Algorithm for Large Data Sets With Mixed Numeric and Categorical Values",IEEE, Proceedings of the Fifth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'03) 0-7695-1957-1/03, 2003
8. Y. Liu, Kefe and X. Liz, "A Hybrid Genetic Based Clustering Algorithm", Proceedings of the Third International Conference on Machine Leaming and Cybernetics, Shanghai,26-29 August 2004
9. Fang-Xiang Wu, Anthony J. Kusalik and W. J. Zhang, "Genetic Weighted K-means for Large-Scale Clustering Problems", University of Saskatchewan, CANADA
10. Zhihui D., Meng D., Sanli Li, Shuyou Li, Mengyue Wu and Jing Zhu, "Massively Parallel SPMD Algorithm for Cluster Computing: Combining Genetic Algorithm with Uphill"

AUTHOR(S) PROFILE

Henal Parmar, received the B.E degree in Computer science & Engineering From North Maharashtra University and Currently pursuing M.E in Computer Science & Engineering in Gujarat Technological University, India.

