International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 6, Issue 4, July- August 2017 ISSN 2278-6856

Performance Evaluation of Soft Computing using Clustering Techniques

Dr. Suman Kumar Mishra[1], Mr. Shobhit Shukla[2]

[1]Assistant Professor, Faculty of Engineering, Dr. Shakuntala Misra National Rehabilitation University, Mohan Road, Lucknow, India

[2]Assistant Professor, Faculty of Computer Science and IT, Dr. Shakuntala Misra National Rehabilitation University, Mohan Road, Lucknow

Abstract

Soft Computing can be considered as a tool to handle inexactness and uncertainty. The main concept of soft computing is to exploit the forbearance for inexactness, uncertainty, fractional truth and approximation to accomplish tractability, robustness and inexpensive solutions. It is used to solve real life problems full of uncomfortable features due to fractional, unclear, noisy and partial information. Clustering gets its name as a metaphor for the soft computing. Clustering aims to divide data groups into subgroups called clusters. Clustering is still searching for killer applications that not only takes advantage of its promise of "high performance and short development lead time". This paper focuses on the evaluation of the Genetic Algorithm concept of Soft Computing and measures its performance using various clustering techniques.

Keywords: Soft Computing, Genetic Algorithms, Clustering, K-means and K - means++.

1. INTRODUCTION

1.1. Soft Computing

Soft computing is a collection of methodologies, which aim to exploit tolerance for imprecision, uncertainty and partial truth to achieve tractability, robustness and low solution cost [1]. It is the amalgamation of methods intended to model and permit solutions to real world problems which are difficult to model mathematically. The core of soft computing consists of Fuzzy Systems, Neural Networks and Genetic Algorithms. The principles of Fuzzy Logic were introduced several decades ago by Lotfi Zadeh in 1965. It is a mathematical device aimed at dealing with vagueness, inexactitude and data granularity. A neural network is a technique that mimics the working of the human brain (nerves and neurons), and contains densely interconnected computer processors working simultaneously. Genetic algorithm is a tool that imitates the natural evolution in such a way so as to resolve problems. Its objective is to cognize the occurrence of “adaptation” as it ensues in nature and to cultivate methods which help in introducing the tools of natural adaptation into computer systems. Apart from these three techniques, various hybrid soft computing techniques are also employed for more accurate and successful disaster management activities. Soft computing varies from conventional computing in that, unlike conventional computing, it accepts data which contains inexactitude, vagueness, partial truth, and estimate. Conventional computing requires a precisely stated analytical model and often a lot of computation time but the real world problems are pervasively imprecise and uncertain thus, necessitating the use of soft computing in many real world scenarios. The core of soft computing consists of Fuzzy Systems, Neural Networks and Genetic Algorithms. Soft computing methods have applied to many real-world problems like signal processing, business forecasting, pattern recognition, speech processing, quality assurance, credit rating, adaptive process control, robotics control, natural language processing, prediction etc.

1.2. Clustering

Clustering is the unsupervised grouping of patterns (observations, data items, or feature vectors) into groups (clusters) [2]. The concept of a cluster involves taking a group of various application based servers working together to run a common set of applications and organize them to work together to provide clustering in data applications. Performance of clustering using soft computing based data has been a topic of much recent interest, motivated by applications of this field on data analytics. It should be noted that clusters are not only being used for high performance computation, but increasingly as a platform to provide highly available services for application such as soft computing based data. Clusters are used in many scientific disciplines, including environmental (ecology, genetics), engineering (turbo-fan design, automobile design) and high-energy physics (nuclear- weapons simulation). The problem of clustering is detailed as: Given a set of data objects, the problem of clustering is to partition data objects into groups in such a way that objects in the same group are similar while objects in different groups are dissimilar according to the predefined similarity measurement. Therefore, clustering analysis can help us to gain insight into the distribution of data [3].

2. FRAMEWORK OF CLUSTERING AND SOFT COMPUTING

2.1. K-Means Clustering

The K- Means is the simplest and most commonly used algorithm employing a squared error criterion [4]. It is a

partitional clustering algorithm which tries to find a user specified number of clusters which are represented by their centroids. In K-means algorithm firstly, K initial centroids are chosen which correspond to the number of clusters required. Each point in the dataset is then assigned to the closest centroid such that each group of points assigned to a centroid forms a cluster. The centroid of each cluster is then updated based on the points assigned to each cluster. The above steps are repeated until the clusters become constant and their respective centroids do not change. The K-means clustering algorithm formally described as follows:

BEGIN
STEP 1: Initialize Cluster Centers to coincide with randomly select points inside the population or hyper volume containing the pattern set.
i.e. $\{C_1, C_2, C_3, \ldots\ldots\ldots C_k\}$

STEP 2: Assign each object to the closet cluster center.

STEP 3: After assigning of the objects, then recalculates the difference between the cluster centers or k centroids.

STEP 4: If shortest distance does not comes, then repeat Step 2 and Step 3.

END

2.2. K-Means ++ Clustering

In the k-means clustering algorithm, an integer k and a set of n data points $X \subset R$, are given. The goal is to choose k centers so as to minimize the sum of the squared distances between each point and its closest center. From a theoretical stance, k-means is not a good based on efficiency or quality: the running time can be exponential in the worst case [5, 6] and though the final solution is locally optimal, it can be very far away from the global. A significant step was taken by Ostrovsky et al. [7] and Arthur and Vassilvitskii [8], who displayed a simple technique called k-means++, that both leads to good assurances for the quality of the solution, and, by benefit of a good initial point, improves the running time of K-means in practice. This algorithm selects the initial center uniformly at random from the data. Each successive center is nominated with a probability proportional to its contribution to the overall error given the previous selections. Intuitively, the initialization algorithm uses the fact that a good clustering is relatively spread out, thus when selecting a new cluster center, preference is given to those further away from the previously selected centers [9]. The K- Means++ clustering algorithm formally described as follows:

BEGIN
STEP1: Take one center c_1, chosen uniformly at random from χ.

STEP 2: Take a new center c_i, choosing $x \in \chi$ with probability $\frac{D(x)2}{\sum_{x \in \chi} D(x)^2}$

STEP 3: Repeat step ii until we have taken k centers altogether.
END

2.3. Genetic Algorithm

Genetic algorithms are computer programs that mimic the processes of biological evolution in order to solve problems and to model evolutionary systems (10, 14). They were proposed by John Holland in the 1960s and further developed by his team at University of Michigan in the 1960s and 1970s. Holland's objective was to cognize the occurrence of "adaptation" as it occurs in nature and to cultivate techniques through which natural adaptation can be used in computer systems. This algorithm is a technique for travelling from one generation of "chromosomes" to a new generation, using 'selection' along with the evolution motivated operators of 'crossover', 'mutation' and 'inversion'. Each chromosome consists of "genes" (e.g., bits), with individual gene representing a particular "allele" (e.g., 0 or 1). The 'selection' operator selects those chromosomes in the current generation that would be permitted to reproduce. The 'crossover' exchanges certain parts of two chromosomes, imitating natural recombination between two single-chromosome entities. The 'mutation' arbitrarily changes the values of some sites in the chromosome and the 'inversion' operator reverses the order of some sections of the chromosome. Each repetition of this method is called a "generation". A genetic algorithm typically repeated for about 50 to 500 or more generations. At the end of these repetitions, there are one or more very fitting chromosomes in the generation. The basic procedure for Genetic Algorithm is as follows:
1. Initialize the population. Call this the current population.
2. Repeat Step 3 through Step 5 till termination condition is satisfied.
3. Apply selection operation on the current population to obtain the mating pool.
4. Apply crossover and mutation operators on the mating pool to generate the new population.
5. Replace the current population by the new population.
6. Return the best solution of the current population.

3. EXPERIMENTAL SETUP

To implement the genetic algorithm based evaluation using clustering techniques. The system uses an Intel Core i5 Processor with 4 GB DDR3 RAM. Windows 7 or higher version of operating system can use and SQL server 2005 or higher version can be used as a database. Versant Software is use for the clustering.

3.1. Datasets

We use four datasets to evaluate the performance estimation of k- means++. The first dataset is the East Zone in Uttar Pradesh. To generate the dataset, we sampled 25 and 50 clusters from a 2, 00 sample points (Size). The second dataset is the West Zone in Uttar Pradesh. To

generate the dataset, we sampled 25 and 50 clusters from a 2, 00 sample size. The third dataset is the North Zone in Uttar Pradesh. To generate the dataset, we sampled 25 and 50 clusters from a 2, 00 sample points (Size) and the fourth dataset is the South Zone in Uttar Pradesh. To generate the dataset, we sampled 25 and 50 clusters from a 2, 00 sample points (Size).

4. INTEGRATION OF SOFT COMPUTING AND CLUSTERING

Having discussed the different clustering based techniques, we now move to the discussion of these techniques and features based on a practical approach. These approach involves the implementation of two clustering techniques (K – means and K- means ++), and testing each one of them on a set of disaster based data related to flood in Uttar with different clusters and sample points. The data set consist of more than 500 cases. The data set partitioned into four datasets i.e. East zone, West Zone, North Zone and South Zone.

4.1. Datasets

In this part we now focus that k- means++ is faster than k-means when implemented to run in parallel. We assume that dimension is 10. The below table present the running time of the k- means and k-means++.

Table 1: The Total Time result for 10 iterations

Dataset	K	Size	Total Time	
			K -means	K – means++
East	35 50	2,00	0.14 0.17	0.04 0.06
West	35 50	2,00	0.13 0.20	0.05 0.09
North	35 50	2,00	0.16 0..18	0.07 0.08
South	35 50	2,00	0.15 0.19	0.05 0.09

In the above table, we can say that K- means++ total execution time is less than k –means, so K- means++ is faster than k-means.

4.2. Average Distance Calculation (ADC)

In this part we now present the overall average distance calculations. We assume that dimension is 10. The below table present the average distance calculations of the k-means and k-means++ algorithms. The number of distance calculations required is very small.

Table 2: The Average Distance Calculations result for 10 iterations

Dataset	K	Size	Total Time	
			K - means	K – means++
East	35 50	2,00	2.31 5.55	1.98 2.76
West	35 50	2,00	2.75 1.67	1.98 1.06
North	35 50	2,00	2.06 2.04	1.05 1.10
South	35 50	2,00	3.08 4.04	2.06

In the above table, we can say that K- means++ total average distance is less than k –means so K- means++ is using less distance that is increasing the execution speed in comparison to k-means.

5. CONCLUSION

From the reflected research works and explanation, lot of space for research and innovations provided by working framework of soft computing and clustering. In this paper, we have discussed how the disaster system is extending to innovative to find and control disaster system with the help clustering approach. The presented work can easily extended by implementing this into a working scenario of other disaster system like earthquake. The comparative analysis is done by using two performance metrics known as running time and average distance calculation. The performance measures for two clustering techniques have been done on Versant software. Disaster system involves high number of dimensions with complicated relationship between the variables in the input data. But in this case, using k – means clustering is used to look over the performance of the other clustering techniques but when the number of clusters is not known, and then K – means cannot handle this type of problem. After implementation of K- means++ the good results are obtained as compared to the k-means.

Acknowledgment

Authors are grateful to Dr. Puneet Misra , Senior Assistant Professor, Department of Computer Science,Lucknow University, Lucknow, Uttar Pradesh for providing the excellent facility in the computing lab of Lucknow University, Lucknow, Uttar Pradesh, India. Thanks are also due to Dean (Academic) of PSIT, Kanpur, Uttar Pradesh, Prof. Harsh Dev for providing his valuable guidance in above mentioned directions.

References

[1] Kurhe A.B., Satonkar S.S., Khanale P.B. and Shinde Ashok. 2011. Soft Computing and its Applications. BIOINFO Soft Computing Volume 1, Issue 1, 2011, pp-05-07.

[2] A. K. Jain, M. N. Murty, and P. J. Flynn. 1999. Data clustering: a review. ACM Comput. Surv. 31, 3 (September 1999), 264-323. DOI=http://dx.doi.org/10.1145/331499.331504.

[3] H. L. Chen, M. S. Chen and S. C. Lin, "Catching the Trend: A Framework for Clustering Concept-Drifting Categorical Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 5, pp. 652-665, May 2009.

[4] McQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 281–297.
[5] D. Arthur and S. Vassilvitskii. How slow is the k-means method? In SOCG, pages 144-153, 2006.
[6] A. Vattani. k-means requires exponentially many iterations even in the plane. DCG, 45(4):596-616, 2011.
[7] R. Ostrovsky, Y. Rabani, L. J. Schulman, and C. Swamy. The effectiveness of Lloyd-type methods for the k-means problem. In FOCS, pages 165-176, 2006.
[8] D. Arthur and S. Vassilvitskii. K-means++: The advantages of careful seeding. In SODA, pages 1027-1035, 2007.
[9] Bahman Bahmani, Benjamin Moseley, Andrea Vattani, Ravi Kumar, and Sergei Vassilvitskii. 2012. Scalable k-means++. Proc. VLDB Endow. 5, 7 (March 2012), 622-633. .
[10] D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, New York, 1989.
[11] L. Davis (Ed.), Handbook of Genetic Algorithms, Van Nostrand Reinhold, New York, 1991.
[12] Z. Michalewicz, Genetic Algorithms+Data Structures=Evolution Programs, Springer, New York, 1992.
[13] J.L.R. Filho, P.C. Treleaven, C. Alippi, Genetic algorithm programming environments, IEEE Comput. 27 (1994) 28-43.

AUTHORS

Dr. Suman Kumar Mishra recived the Ph. D. (Computer Science & Engg.) in 2013 and M.C.A. Degree in 2005 from Baba Sahab BhimRao Ambedkar University (A Central University), Lucknow. He has 12 years of teaching experience and research experience in the field of Object Oriented Analysis and Design, E – Commerce and Data Mining and Warehouse.

Shobhit Shukla received the B.Tech degree in Computer Science & Engineering from Jaypee University of Information Technology and M.Tech degree in Software Engineering from Manipal Institute of Technology in 2010 and 2013, respectively. He is now working as an Assistant Professor in Dr. Shakuntala Misra National Rehabilitation University, Lucknow, India.