DEPSO Model For Efficient Clustering Using Drifting Concepts

Dr.N.Rajathi ,Dr.J.S.Kanchana, P.Suganthi Malarvizhi

Abstract: In datamining, knowledge is gained from data. The extracted information or knowledge can be used for market analysis and customer retention. As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe the characteristics of each cluster.Clustering process groups, the abstract objects into classes of similar objects.An iterative optimization algorithm is used in the existing system for clustering the data objects with drifting concepts and using some cluster validity function toassess the efficiency of the clustering model while each new input data subset is flowing. The Proposed system uses Differential Evolutionary Particle Swarm Optimization (DEPSO) model for effectively clustering the several real data sets with drifting concepts.

Index Terms: Cluster analysis, Differential evolutionary, Particle Swarm Optimization, Drifting Concepts ,Data mining , Cluster validity

1 INTRODUCTION

Data mining is helpful for the extracting hidden predictive information from large databases, it is a powerful technology with great potential to help companies focus on the most what matters most important in their data warehouses. Data mining tools predict future trends and behaviors, allowing business to make proactive, knowledge based decisions. The automated, prospective tools typical of decision support systems. Data mining tools can answer business questions that have been too time consuming to resolve. The techniques can be quickly implemented on existing software and hardware platforms to enhance the value of existing informationresources, and can be integrated with new products and systems as they are brought on-line. The objective of clustering is to group similar data items into groups called clusters so that items in the same cluster are highly similar but are dissimilar from items in other clusters. The process of clustering the entire dataset without performing drifting concepts are in considered. Such a process not only reduces the quality of the cluster, but also disregards users' expectations because they only need the current clustering result. The volume of data stream is very huge, so storing and retrieving the entire data stream is expensive. Hence, a well-defined technique for efficient clustering of datasets with drifting concepts is used. In this paper we use differential evolutionary particle swarm optimization algorithm for clustering the dataset with drifting concepts.

2 RELATED WORK

Daniel Barbara et. al [1] discussed about an entropy- based algorithm for clustering categorical data called COOLCAT.

- *Department of Information Technology, Kumaraguru College of Technology Coimbatore, Tamil Nadu, India*
- *Department of Information Technology, K.L.N. College of Engineering, Pottapalayam,Tamil Nadu, India*
- *Department of CSE, Velalar College of Engineering and Technology , Erode,Tamil Nadu, India*

It is inferred that COOLCAT is well-equipped for clustering data streams (continuous incoming data point streams), as it is an incremental algorithm that allows new data points to be grouped without having to search at every point that has been clustered so far. Fuyuan Cao et. al [2] proposed a framework for clustering categorical time-evolving data and found that the problem of clustering categorical time-evolving data remains a challenge task. In this proposed work, a generalized clustering framework for categorical time-developing data is used. The analysis of time complexity shows that these proposed algorithms are effective for large data sets.They experimented with the real dataset and proved that the proposed framework accurately detects the drifting concepts and also attains better clustering results. The authors Liang et al. [3] proposed an optimized model for clustering categorical data streams. The authors used a cluster validity function as an objective function to assess the effectiveness of the clustering model.L.Bai et. al [4] has introduceda new initialization method for clustering categorical data . This novel algorithm taking into account the distance between objects and the density of the object. The initialization method used in this work is adopted to the k-mode algorithm and the fuzzy k-modi algorithm in the proposed approach. Liang Bai et. al [5] used one of the computationally efficient clustering methods called k-modes for clustering categorical data streams. In this algorithm, a cluster is represented by a mode which is composed of the attribute value that occurs most frequently in each attribute domain of the cluster. The methodology proposed not only identifies the good initial cluster centers, but also provides a criterion for finding the candidates for the number of clustersHO SS et. al [6] proposed a methodto detect the change of data generating model in data streams by testing the exchangeability property of the observed data. Keke Chen and Ling Liu [7] presented a framework for detecting the change of primary clustering structure in categorical data streams. This is indicated by the change of the best number of clusters (Best K) in the data stream.The proposed framework used a Hierarchical Entropy Tree structure (HE-Tree) to capture the entropy characteristics of clusters. Hung-Leng Chen [8] has discussed "Catching the trend: A framework for clustering conceptually drifting categorical data" and identified that sampling has been recognized as an important technique for improving clustering efficiency. They proposed a mechanism called Maximal Resemblance Data Labeling (abbreviated as MARDL) to allocate each unlabeled data point into the corresponding appropriate cluster based on the novel

categorical clustering representative, namely, N-Node set Importance Representative (abbreviated as NNIR) that represents clusters by the importance of the combinations of attribute values.

3 PROPOSED SYSTEM

The flow diagram given in figure 1 shows the various functional operations involved in the proposed system.

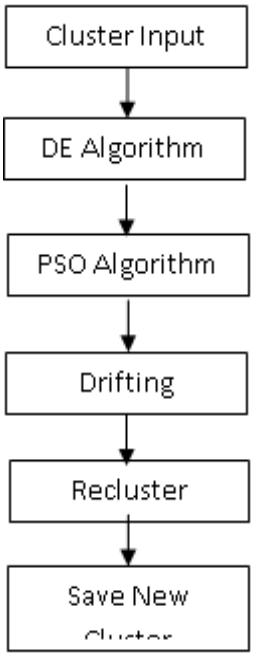

Fig 1. *Working of the Proposed System*

3.1 Clustering Framework

In this work a clustering framework based on the sliding window technique is proposedto cluster categorical data streams and identify the drifting concepts.The sliding window technique conveniently removes the obsolete records and stores only the clustering modelsused in several previous works to grouptime- varying data. Based on this technique, the latest data objects in the current window are clustered and capture the evolution trend of cluster structures in the data stream. The different partition results can lead to differentrecognition results of drifting concepts. Therefore, an optimization model is required to find out the optimal partition result for the new subset of input data.

3.2 DESPO Alogorithm

DE-PSO works like the usual DE algorithm up to the point where the trial vector is generated. If the trial vector satisfies the conditions, it is included in the population otherwise the algorithm enters the PSO phase and generates a new candidate solution. This method is repeated iteratively till the optimum value is reached. The inclusion of PSO phase leads a perturbation in the population, which in turn helps to sustain the diversity of the population and to find a good optimal solution.

3.3 The Drifting Concept Detection

After clustering the new input data subsets, the change situation between the new and the last clustering model, to determine whether the drift concepts are occurring to be analyzed. While the concepts have shifted, the last clustering model is not used to create the new clustering model. In this case, regroup the new subset of input data. The following two factors are considered to find out the drift concepts such as distribution variation and safety variation. The distribution variation can reduce or increase the uncertainty of the cluster model for cluster representatives. When the uncertainty is reduced, it is considered that the concepts do not drift, though the distribution variations are large.

3.4 Clustering Accuracy Evaluation

To evaluate the performance of clustering algorithms in this work, three validity measures accuracy, precision and recall are considered, and their corresponding formulas are as given below.

Accuracy(AC)

$$AC = \frac{1}{n} \sum_{i=1}^{k} \max_{j=1}^{k'} n_{ij}$$

Precision (PE)

$$PE = \frac{1}{k} \sum_{i=1}^{k} \frac{\max_{j=1}^{k'} n_{ij}}{b_j}$$

Recall (RE)

$$RE = \frac{1}{k} \sum_{i=1}^{k} \frac{\max_{i=1}^{k} n_{ij}}{d_j}$$

4 EXPERIMENTAL SETUP

The KDD-CUP'99 data stream is used for the experiment. Table 1 shows the computational times against the numbers of clusters. According to the results presented in table 1, the DEPSO algorithm requires more computational times than iterative algorithms.

TABLE 1
Computational Times (Seconds) of Clustering Algorithms for Different Numbers of Clusters

No. of Iteration	*Iterative Algorithm*	*DEPSO Algorithm*
5	*169.12*	*172.7*
10	*256.88*	*262.34*
15	*371.69*	*385.52*
20	*458.69*	*467.21*

The figure 2 shows the computational time of iterativealgorithm and DEPSO algorithm for different number of.clusters. The DEPSO algorithm shows the better computational time when compared with iterative algorithm

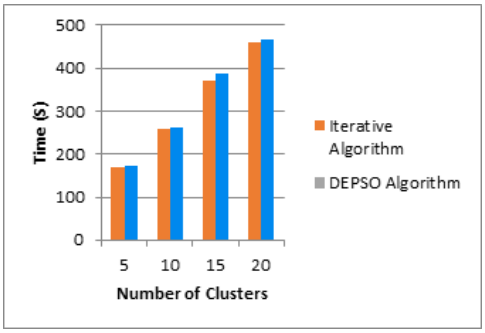

***Fig.2**. Computational Times (Seconds) of Clustering Algorithms*

for Different Numbers of Clusters

5 CONCLUSION

In this paper, Differential evolutionary particle swarm optimization algorithm for effectively clustering with drifting concepts was presented. Finally, the performance of the proposed algorithm is tested and the experimental results have shown that the proposed algorithm is effective in clustering the data streams and the detection results based on the proposed method are reliable. This work has a very vast scope in future and it can be implemented on other new heuristic algorithms in future. It can be updated in near future as and when requirement for the same arises, as it is very flexible in terms of efficiency

6 REFERENCES

[1] Daniel Barbara,Julia Couto and Yi Li, “COOLCAT: An entropy-based algorithm for categorical clustering” , CIKM,ACM,2002.

[2] F. Cao, J. Liang, X.Zhao and L. Bai, “A framework for clustering categorical time-evolving data,” IEEE Trans. Fuzzy Syst., vol. 18, no. 5, pp. 872–885, Oct. 2010.

[3] Liang Bai, Xueqi Cheng, Member, IEEE, Jiye Liang, and Huawei Shen., “An Optimization Model for Clustering Categorical Data Streams with Drifting Concepts”, IEEE Transactions on Knowledge and Data Engineering, Vol. 28 , Issue. 11, Nov. 2016.

[4] L. Bai, J. Liang, C. Dang, and F. Cao, “The impact of cluster representatives on the convergence of the K-modes type clustering,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 6, pp. 1509– 1522, Jun. 2013.

[5] L. Bai, J. Liang, and C. Dang, “An initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data,” Knowl.-Based Syst., vol. 24, no. 6, pp. 785–795, 2011.

[6] S. Ho and H. Wechsler, “A martingale framework for detecting changes in data streams by testing exchangeability,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 20, pp. 2113–2127, Dec 2010.

[7] Nagaraj, Balakrishnan, and Ponnusamy Vijayakumar. "Tuning of a PID controller using soft computing methodologies applied to basis weight control in paper machine." Journal of Korea Technical Association of The Pulp and Paper Industry 43, no. 3 (2011): 1-10.

[8] H. Chen, M. Chen, and S. Lin, “Catching the trend: A ramework for clustering concept-drifting categorical data,” IEEE Trans. Knowl. Data Eng., vol. 21, no. 5, pp. 652–665, May 2009.

[9] K. Chen and L. Liu, “HE-tree: A framework for detecting changes in clustering structure for categorical data streams,” Int. J. Very Large Data Bases, vol. 18, no. 5, pp. 1241–1260, 2009.