A Review of Threshold Based Clustering Technique on Different Optimization Function

Anuradha Paliwal , Mr.Himanshu Yadav, Mr.Anurag jain

Abstract --Threshold based clustering technique is new generation of clustering technique. The threshold based clustering technique gives the principle of optimal cluster generation process. The optimal cluster generation process gives the feasible number of cluster during the generation of clustering technique. The concept of optimal clustering technique is a new area of pattern generation and analysis process. In this paper present the review of threshold based clustering technique used in different clustering domain such as hard clustering, soft clustering, hierarchical clustering and mountain clustering technique. For the selection of threshold function used various heuristic and meta-heuristic functions along with neural network. The family of heuristic function gives the verity of optimization algorithm. The neural network based threshold selection process gives optimization as well as error minimization during the cluster generation.

Keywords-- Data Mining, Clustering, Threshold Function, Heuristic Function

1 INTRODUCTION

Clustering play an important role in discovery of unknown pattern for large database analysis. In large database have multiple features and multiple features generate multiple views of data. In multi-view data used two clustering approach one is centralized and other is distributed approach. Centralized algorithms make use of multiple representations simultaneously to discover hidden patterns from the data. Most of the existing work in multi-view clustering follows the Centralized approach with extensions to existing clustering algorithms. The generation of clustering technique proceeds in manner of distributed and partition clustering technique. Using a partition clustering technique to generate centralized clustering process by k-means technique, but the k-means clustering technique not support multiple feature of data because it not assigned random center for cluster generation. Now in current research trend used variable weighted clustering technique for improving performance of clustering technique. In the journey of improvement of clustering technique used variable weighting clustering technique. For the more extension of clustering technique used two level weighted clustering techniques. In this dissertation proposed fuzzy based two level weighted cluster technique for multi-view data [21]. Variable weighting clustering has been important research topic in cluster analysis. The principle of optimal clustering gives the concept of threshold selection for the better and efficient generation of cluster generation. The optimal process adapts heuristic based threshold function for the constraints validation and improvement of quality of clustering technique. The heuristic function gives various optimization algorithms such as particle swarm optimization, ant colony optimization, glowworm optimization algorithm and teacher based learning optimization algorithm. These entire algorithm by nature support dynamic population based condition. The neural network inspired clustering algorithm such as SOM based clustering technique and RBF based clustering technique. The SOM based clustering technique gives very useful in terms of pattern generation and pattern analysis. The SOM network unsupervised neural network model and process on the basis of iteration for the grouping of regular pattern. The RBF neural network clustering technique is also very useful technique for analysis of clustering process. The rest of paper organized in the form of section II discusses related work in the field of clustering. In section III discuss the optimization technique. In section IV discuss the problem formulation of current work and finally discuss the conclusion and future work in section V.

2 RELATED WORK

In this section discuss the related work in the field of threshold based clustering technique. The threshold based clustering technique adapts various optimization algorithm and neural network model for the generation and validation of cluster index. Here presents some review in format of tabular fashion in terms of demerits of dedicated work in the field of clustering algorithm.

Sr. No.	Author Name	Paper Title	Approaches Used	Publication/ Year	Demerits

[1]	Nishchal K. Verma, Abhishek Roy	Self-Optimal Clustering Technique Using Optimized Threshold Function	Optimal Clustering	IEEE, 2013.	The optimized threshold function based on genetic algorithm, the process of genetic algorithm is very efficient for limited data instance.
[2]	Li Xuan, Chen Zhigang, Yang Fan	Explorploring of clustering algorithm on class imbalbalanced Data	Clustering Algorithm	IEEE, 2013.	Imbalanced data distribution still remains an unsolved problem in machine learning.
[3]	Ramachandra Rao Kurada, K Karteeka Pavan, AV Dattareya Rao	A preliminary survey on optimized multiobjective metaheuristic methods for data clustering using evolutionary approaches	Meth Heuristic Methods	IJCSIT, 2013.	The multiple instance based threshold function faced a problem of confusion during the selection of threshold function.
[4]	R. J. Lyon, J. M. Brooke, J. D. Knowles	A Study on Classification in Imbalbalanced and Partially-Labeled Data Streams	Classification Techniques	IEEE, 2013.	Future work will expand on the results of this investigation, and test other data stream classifiers not exclusively based on the Hoeffding bound.
[5]	Rushi Longadge, Snehlata S. Dongre, Latesh Malik	Multi-Cluster Based Approach for skewed Data in Data Mining	Data Mining Methods	IOSR, 2013.	Performance will degrade as the class imbalance ratio increased.
[6]	Rukshan Batuwita, Vasile Palade	Class imbalbalance learning methods for support vector machines	Support Vector Machines	John Wiley & Sons, 2012.	The class imbalance problem faced during the process of classification.
[7]	M. Mostafizur Rahman and D. N. Davis	Addressing the Class Imbalbalance Problem in Medical Datasets	Sampling Techniques	International Journal of Machine Learning an	The medical data faced a problem of multiple instance of new feature point always.

IJSER

				d Com putin g, 2013.	
[8]	Nenad Tomasev, Dunja Mladeni	Hub Co-occurrence Modeling for Robust High-dimensional kNN Classification	Hidden Naive Bayes (HNB) model	IEEE, 2009.	The nature of clustering technique gives the process of homogeneity for the collection of data.

3 THRESHOLD BASED FUNCTION AND OPTIMIZATION ALGORITHM

In this section discuss the threshold function in terms of selection of as threshold in clustering algorithm. the threshold function act as center value of cluster and index of quality index of cluster. Here discuss some heuristic function and neural network model as threshold function. Here discuss three algorithm genetic algorithm, particle swarm optimization and ant colony optimization.

Genetic algorithm

Genetic algorithm is dynamic population based searching technique. The dynamic population searching technique defines the constraints for the selection of input data for the process of optimization. The process of genetic algorithm define in multiple domain such as SGA(simple genetic algorithm) MOGA(multi-objective genetic algorithm). The genetic algorithm consists of multiple steps in terms of ionization, selection, crossover, mutation and finally results. The basic concept of GAs is designed to simulate processes in natural system necessary for evolution, specifically those that follow the principles first laid down by Charles Darwin of survival of the fittest. GAs have been widely studied, experimented and applied in many fields of sciences. Not only does GAs provide an alternative method to solving problem, it consistently outperforms other traditional methods in most of the problems. Many of the real world problems involved finding optimal parameters, which might prove difficult for traditional methods but ideal for GAs

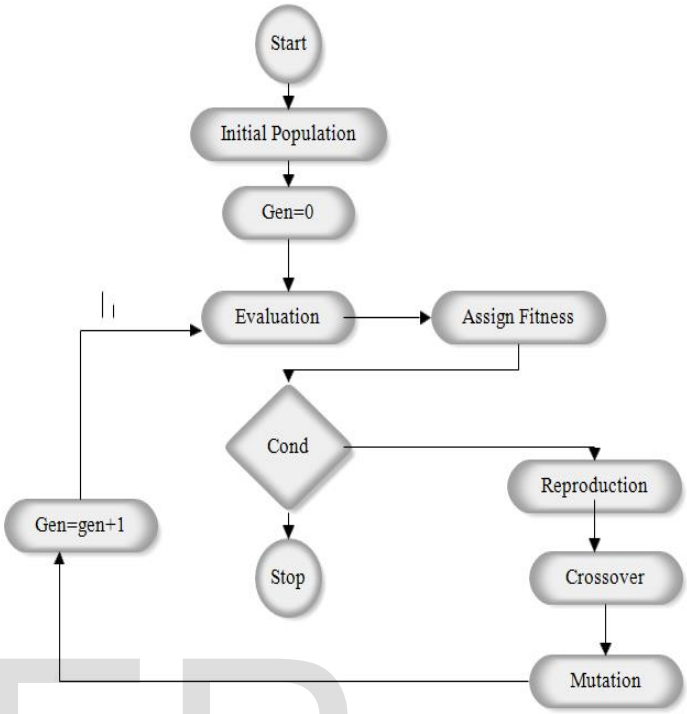

Figure 1 working block diagram of working principle of genetic algorithm

Particle swarm optimization

Particle swarm optimization is also dynamic population based optimization technique. The particle swarm optimization inspired by bird of fork fly in the sky. The process of data optimization deals with two function one id Gbest and Pbest. The Pbest is local optimal function and Gbest is global optimal function. In PSO, the population is the number of particles in a problem space. Particles are initialized randomly. Each particle will have a fitness value, which will be evaluated by a fitness function to be optimized in each generation. Each Particle knows its best position pbest and the best position so far among the entire group of particles gbest. The pbest of a particle is the best result (fitness value) so far reached by the particle, whereas gbest is the best particle in terms of fitness in an entire population. In each generation the velocity and the position of particles will be updated as in Eq 4.1 and 4.2, respectively. The heuristic optimizes the cost of task-resource mapping based on the solution given by particle swarm optimization technique.

$$v_i^{K+1} = \omega v_i^k + c1rand1 \times \left(pbseti - x_i^k\right) + c2rand2 \times \left(gbest - x_i^k\right) \dots\dots\dots\dots\dots (1)$$

$$x_i^{K+i} = x_i^k + v_i^{k+1} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (2)$$

Ant colony optimization

The ant colony optimization is inspired by biological ant. The behavior of ants finds the shortest path for the searching of food. The searching of food finds optimal process of path. The ant colony optimization process finds the dissimilar ants on the selection of path.

Let F is a feature set and N is the total artificial ants and possibility of ant selection is s1,s2…….sn , now find the selection possibility of two ants in given solution is

$$SP(i,j) = \frac{1}{si - sj} \dots\dots\dots\dots\dots\dots (1)$$

Where si and sj is the dissimilar probability of two different ants. Now estimate the value of appetence of ants is

$$ACP(i+j) = \frac{\alpha i + \beta i}{N} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (2)$$

Where αi and βi is ants whose selection possibility is maximum in terms of another ants the ratio of selection of ants is defined as $\frac{100}{N}$

On the basis of selection possibility estimate the value of artificial phenomenon value

$$\Delta\tau i = \frac{A.\,si}{ACP(i+j)} \dots\dots\dots\dots\dots\dots\dots\dots (3)$$

Where A is constant phenomenon value

Now each iteration of pheromone value is increment and decrement according to their selection probability. The derivation of universal appetence probability is

$$p_{ii}^{k} = \begin{cases} [\tau_{ij(t)}] \\ 0 \end{cases} \quad \begin{matrix} \alpha.[k_{ij}]\beta \\ \text{otherwise} \end{matrix} \quad \text{if } j \in j \dots\dots\dots\dots (4)$$

Where kij gives the information of heuristic search space and measure the selection possibility of artificial ants

And finally getting the optimal word feature of text document for the processing of optimization

4 PROBLEM FORMULATIONS

In this section discuss the problem formulation of threshold based clustering technique. The threshold based clustering technique faced a problem of selection process. The selection of threshold function based on the process of data nature. The process of clustering technique used for the unknown nature of clustering data. In the process of threshold used heuristic based function. The heuristic based function provides multiple searching space technique. The process of data clustering faced a problem of seed selection and weight ratio specification for the better generation of cluster. The cluster raised a problem of content validation and maximum number of error. Some problem is discussed here.
initial seed selection[4]
Calculation of weight value for the mapping [12].
Maximum number of iteration[9]
Maximum number of iteration[10]
Accuracy value is compromised[15]

5 CONCLUSION & FUTURE SCOPE

In this paper present the review of threshold based clustering technique. The threshold based clustering technique used various optimization algorithm for the selection of threshold function. For the selection of threshold function used genetic algorithm, particle swarm optimization and ant colony optimization. All selection algorithms perform very well but faced a problem of data diversity and new feature attribute during the clustering process. The process of review estimates the accuracy of clustering technique and maximization of iteration. The maximization of number of iteration faced a problem of data loss. Now in future used multi-genetic algorithm for the selection of threshold and measure quality index of cluster.

REFERENCES

[1] Nishchal K. Verma, Abhishek Roy "Self-Optimal Clustering Technique Using Optimized Threshold Function" IEEE SYSTEMS JOURNAL, IEEE 2013. Pp 1-14.

[2] Li Xuan, Chen Zhigang, Yang Fan "Exploring of clustering algorithm on class imbalanced Data" The 8th International Conference on Computer Science & Education IEEE ,2013. Pp 89-94.

[3] Ramachandra Rao Kurada, K Karteeka Pavan, AV Dattareya Rao "A preliminary survey on optimized multiobjective metaheuristic methods for data clustering using evolutionary approaches" International Journal of Computer Science & Information Technology (IJCSIT) Vol 5, 2013. Pp 57-78.

[4] R. J. Lyon, J. M. Brooke, J. D. Knowles "A Study on Classification in Imbalanced and Partially-Labelled Data Streams" IEEE 2013. Pp 451-457.

[5] Rushi Longadge, Snehlata S. Dongre, Latesh Malik " Multi-Cluster Based Approach for skewed Data in Data Mining" IOSR Journal of Computer Engineering (IOSR-JCE) vol 12, 2013. Pp 66-73.

[6] Rukshan Batuwita, Vasile Palade "Class imbalance learning methods for support vector machines" John Wiley & Sons, Inc. 2012. Pp 1-20.

[7] M. Mostafizur Rahman and D. N. Davis "Addressing the Class Imbalance Problem in Medical Datasets" International Journal of Machine Learning and Computing, Vol. 3,2013. Pp 224-229.

IJSER

[8] Nenad Tomasev,Dunja Mladeni "Hub Co-occurrence Modeling for Robust High-dimensional kNN Classification" IEEE 2009. Pp 125-141.

[9] Dech Thammasiri , Dursun Delen , Phayung Meesad , Nihat Kasap "A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition" Expert Systems with Applications, Elseivet ltd 2013. Pp 1220-1230.

[10] Hualong Yu, Shufang Hong, Xibei Yang" Recognition of Multiple Imbalanced Cancer Types Based on DNA Microarray Data Using Ensemble Classifiers" Hindawi Publishing Corporation BioMed Research International Volume 2013. Pp 201-214.

[11] V. Garc,J. S. S_anchez,R. Mart ,elez,R. A. Mollineda" Surrounding neighborhood-based SMOTE for learning from imbalanced data sets" Institute of New Imaging Technologies,2010. Pp 1-14.

[12] Mohammad Behdad, Luigi Barone, Mohammed Bennamoun and Tim French "Nature-Inspired Techniques in the Context of Fraud Detection" in IEEE transactions on systems, man, and cybernetics—part c: applications and reviews, vol. 42, no. 6, november 2012.

[13] Alberto Fernandez, Maria Jose del Jesus and Francisco Herrera "On the influence of an adaptive inference system in fuzzy rule based classification system for imbalanced datasets" in Elsevier Ltd. All rights reserved 2009.

[14] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Macia-Fernandez and E.Vazquez "Anomaly-based network intrusion detection: Techniques, Systems and challenges" in Elsevier Ltd. All rights reserved 2008.

[15] Terrence P. Fries "A Fuzzy-Genetic Approach to Network Intrusion Detection" in GECCO 08, July12–16, 2008, Atlanta, Georgia, USA.

[16] Zorana Bankovic, Dusan Stepanovic,Slobodan Bojanic and Octavio Nieto-Taladriz "Improving network security using genetic algorithm approach" in Published by Elsevier Ltd 2007.

[17] Mrutyunjaya Panda and Manas Ranjan Patra "network intrusion detection using naive bayes" in IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.12, December 2007.

[18] Animesh Patcha and Jung-Min Park "An Overview of Anomaly Detection Techniques: Existing Solutions and Latest Technological Trends" in Computer networks 2007.

[19] Ren Hui Gong, Mohammad Zulkernine and Purang Abolmaesumi "A Software Implementation of a Genetic Algorithm Based Approach to Network Intrusion Detection" in IEEE 2005.

[20] Jonatan Gomez and Dipankar Dasgupta "Evolving Fuzzy Classifiers for Intrusion Detection" in IEEE 2002.

IJSER

IJSER

IJSER