iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** III **Month of publication:** March 2017

DOI: http://doi.org/10.22214/ijraset.2017.3008

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

A Review of Clustering Technique Based on Different Optimization Function Using for Selection of Center Point

Kavita Firke[1], Dinesh Kumar Sahu[2]
[1]M. Tech Scholar, [2]Asst. prof., Department of CSE
Sri Satya Sai College of Engineering, RKDF University Bhopal M.P., India

Abstract: ***The selection of center point is major issue in clustering technique. the center point of cluster decides the quality and validation of clustering technique. for the better selection of clustering technique used different optimization function such as genetic algorithm, particle swarm optimization algorithm and many more algorithm used for center selection process. In this paper present the review of clustering technique for automatic validation and cluster center selection. The process of clustering basically group the data based on feature attribute of data. the selection of features attribute of data based on the process of iteration.***
Keywords: - Clustering, EA, K-Means, SVM.

I. INTRODUCTION

Bunching is the critical stride for some errands in machine learning [5]. Each algorithmic run has its own inclination owing to the upgrades of different criteria. Unsupervised machine learning is naturally an advancement assignment; one is attempting to fit the best model to a specimen of information. The meaning of "best" is unlimited; speculation show with significance to the full universe of information focuses. However machine learning calculations don't comprehend this from the earlier, and instead of depend on heuristic estimations considering the norms of their copy and confinement, for example, integrity of fit with significance to the analysis raw numbers, demonstrate niggardliness, thus on [5]. Undertaking conclusions were custom fitted at last by amplifying/limiting an objective or favorable position. The measurements and unpredictability of change issues that can be clarified in a sensible time has been progressed by the approach of exceptional processing advancements.Cluster by nature are the collection of similar objects. Each group or cluster is homogeneous, i.e., objects belonging to the same group are similar to each other. Also, each group or cluster should be different from other clusters, i.e., objects belonging to one cluster should be different from the objects of other clusters. In order to increase the classification performance for imbalance data streams, many approaches are proposed by improving traditional classification algorithms, for example, the cost-sensitive learning, the resample, the improved SVM, etc. The cost-sensitive learning takes a full consideration for the performance of the minority classification, and can resolve effectively the imbalanced classification in real life. Adel proposed an online ensemble of neural network classifiers for non-stationary and imbalanced data streams [1].Clustering is the process of grouping similar objects, and this could be hard or fuzzy. In hard clustering algorithm, each element is allocated to a single cluster during its operation; however, in fuzzy clustering method, a degree of membership is assigned to each element depending on its degree of association to several other clusters. Clustering problem for unsupervised data exploration and analysis has been investigated for decades in the statistics, image retrieval, bioinformatics, data mining and machine learning fields. Basically clustering algorithms aim to divide data objects into groups so that objects in the same group are similar to one another and different from objects in other groups. Generally, clustering is identified as an unsupervised learning method which divides data objects into designated clusters based only on the information presented in the dataset without any external background knowledge and label data. Grouping is the essential stride for some errands in machine learning. Each algorithmic lead has its own inclination owing to the upgrades of different criteria. Unsupervised machine learning is naturally an advancement assignment; one is attempting to fit the best model to an example of information. The meaning of "best" is unlimited; speculation show with significance to the full universe of information focuses. However, machine learning calculations don't comprehend this from the earlier, and instead of depend on heuristic estimations considering the principles of their copy and limitation, for example, integrity of fit with significance to the examination raw numbers, show miserliness, etc. Advancement is that the technique for getting the most straightforward outcome or benefit underneath a given arrangement of

weakening variables. Undertaking conclusions were custom-made at last by boosting/limiting an objective or preferred standpoint. The measurements and many-sided quality of change issues that can be clarified in a sensible time has been progressed by the approach of breakthrough computing technologies.

The rest of paper discuss as in section 2 discuss the Related Work. In section 3 discuss the Data Clustering. In section 4 discuss problem statement. Finally discuss conclusion & future work in section 5.

II. RELATED WORK

In this section we discuss the literature survey entitled with their author name and given references number respectively.

They discussed the method of concept drifts detection by extracting the attributive characters of imbalanced massive data streams. If the change of attributive characters exceeds threshold value, the concept drift occurs. Another side, they give Cost-sensitive extreme learning machine algorithm, and the optimal cost function is defined by the dynamic cost matrix. Build the cost-sensitive classifiers model for imbalanced massive data streams under MapReduce, and the data streams are processed in parallel. At last, the weighted cost-sensitive ensemble classifier is constructed, and the dynamic cost-sensitive ensemble classification based on extreme learning machine classification is given.

According to author, Common approaches for dealing with the class imbalance problem involve modifying the data distribution or modifying the classifier. They choose to use a combination of both approaches. They use support vector machines with soft margins as the base classifier to solve the skewed vector spaces problem. They then counter the excessive bias introduced by this approach with a boosting algorithm. They found that this ensemble of SVMs makes an impressive improvement in prediction performance, not only for the majority class, but also for the minority class. they have discussed a new algorithm, boosting-SVMs with Asymmetric Cost, for tackling some of the problems associated with imbalanced data sets.

They discuss a balanced dataset is very important for creating a good training set. Most existing classification methods tend not to perform well on minority class examples when the dataset is extremely imbalanced. They examine the performance of over-inspecting and under-examining systems to adjust cardiovascular information. Surely understood over-examining method SMOTE is utilized and some under-testing systems are likewise investigated. An enhanced under examining procedure is talked about.

They think of one as conceivable information handling situation for the SKA, for the motivations behind an all-sky pulsar overview. Specifically, they treat the determination of promising signs from the SKA handling pipeline as an information stream characterization issue. They consider the attainability of arranging signs that arrive through an unlabeled and vigorously class imbalanced information stream, utilizing as of now accessible calculations and structures.

They give a terminology that highlights a few angles that are essential with regards to transformative information bunching. The paper missions the bunching exchange offs fanned out with far reaching Multi Objective Evolutionary Approaches (MOEAs) techniques. MOEAs have significant accomplishment over an assortment of MOP applications, from instructive multifunction improvement to genuine building plan. The review paper perceptibly composes the improvements saw in the previous three decades for EAs based met heuristics to take care of multi target streamlining issues (MOP) and to infer critical movement in decision fantastic clarifications in a solitary run.

They address this insufficiency by proposing the utilization of the Hellinger separate measure, as a quick choice tree split standard. They exhibit that by utilizing Hellinger a measurably critical change in review rates on imbalanced information streams can be accomplished, with an adequate increment in the false positive rate.

They show that MapReduce Clusters are particularly well suited for parallel parameter optimization. They use MapReduce to optimize regularization parameters for boosted trees and random forests on several text problems: three retrieval ranking problems and a Wikipedia vandalism problem. They show how model accuracy improves as a function of the percent of parameter space explored, that accuracy can be hurt by exploring parameter space too aggressively, and that there can be significant interaction between parameters that appear to be independent.

They discussed, in this implementation text clustering which is one of the most important techniques of text mining that aims at extracting useful information by processing data in textual form is addressed. An improved variant of spherical K-Means (SKM) algorithm named multi-cluster SKM is developed for clustering high dimensional document collections with high performance and efficiency. Experiments were performed on several document data sets and it is shown that the new algorithm provides significant increase in clustering quality without causing considerable difference in CPU time usage when compared to SKM algorithm.

Author describe, Support vector machine is well recognized method for data classification. For the process of support vector

machine evaluation of new feature during classification is major problem. The problem of feature evaluation decreases the performance of Support Vector Machine (SVM). For the improvement of support vector machine, particle of swarm optimization technique is used. Particle of swarm optimization controls the dynamic feature evaluation process and decreases the possibility of confusion in selection of class and increase the classification ratio of support vector machine.

The aim of this work is to shortly review the main issues of this problem and to describe two common approaches for dealing with imbalance, namely sampling and cost sensitive learning. Additionally, they will pay special attention to some open problems, in particular they will carry out a discussion on the data intrinsic characteristics of the imbalanced classification problem which will help to follow new paths that can lead to the improvement of current models, namely size of the dataset, small disjoints, the overlapping between the classes and the data fracture between training and test distribution.

III. DATA CLUSTERING

Information is for the most part in the crude shape, Information is prepared into information i.e. a semantic association gives the information a significance. Records called information things are communicated as tuples of numerical/clear cut values; every incentive in the tuple demonstrates the watched estimation of a component. The elements in an informational collection are additionally called traits or factors. Data can be consequently removed via seeking designs in the information [10]. The way toward recognizing designs in information is called gaining from information. Blackmailing verifiable, already obscure and approaching valuable data from information include the substance of the Data Mining. The algorithmic system giving programmed support to information mining is by and large called Machine Learning [5].

Affiliation lead mining goes for recognizing any relationship among the components. Grouping goes for anticipating the estimation of an ostensible component called the class variable. Grouping is called directed learning on the grounds that the learning plan is given an arrangement of characterized cases from which it is relied upon to take in a method for ordering inconspicuous cases. Information Clustering is executed in two distinct modes called fresh and fluffy. Arrangement is regulated as a directed learning on the grounds that the learning plan is given an arrangement of characterized cases from which it is expected to find a method for ordering concealed illustrations. In fresh bunching, the groups are disjoint and non-covering; any specimen may fit into one and just a single classification. In fluffy grouping, a specimen may have a place with more than one class with a specific fluffy enrollment positioning.

IV. PROBLEM STATEMENT

For the purpose of self-optimal data clustering various machines learning algorithm are applied, such as clustering, weighted clustering, and regression. Two of the most critical and well generalized problems of multi-category data are its new evolved feature and concept-drift. Since a multi-category data is a fast and continuous event, it is assumed to have infinite length. Therefore, it is difficult to store and use all the historical data for training. The most discover alternative is an incremental learning technique. Several incremental learners have been proposed to address this problem. In addition, concept-drift occurs in the multi-category when the underlying concepts of the multi-category change over time. Concept-evolution occurs when new classes evolve in the data. On the re-category process we found some important problem in cluster oriented multi-category data clustering. These problems are given below.

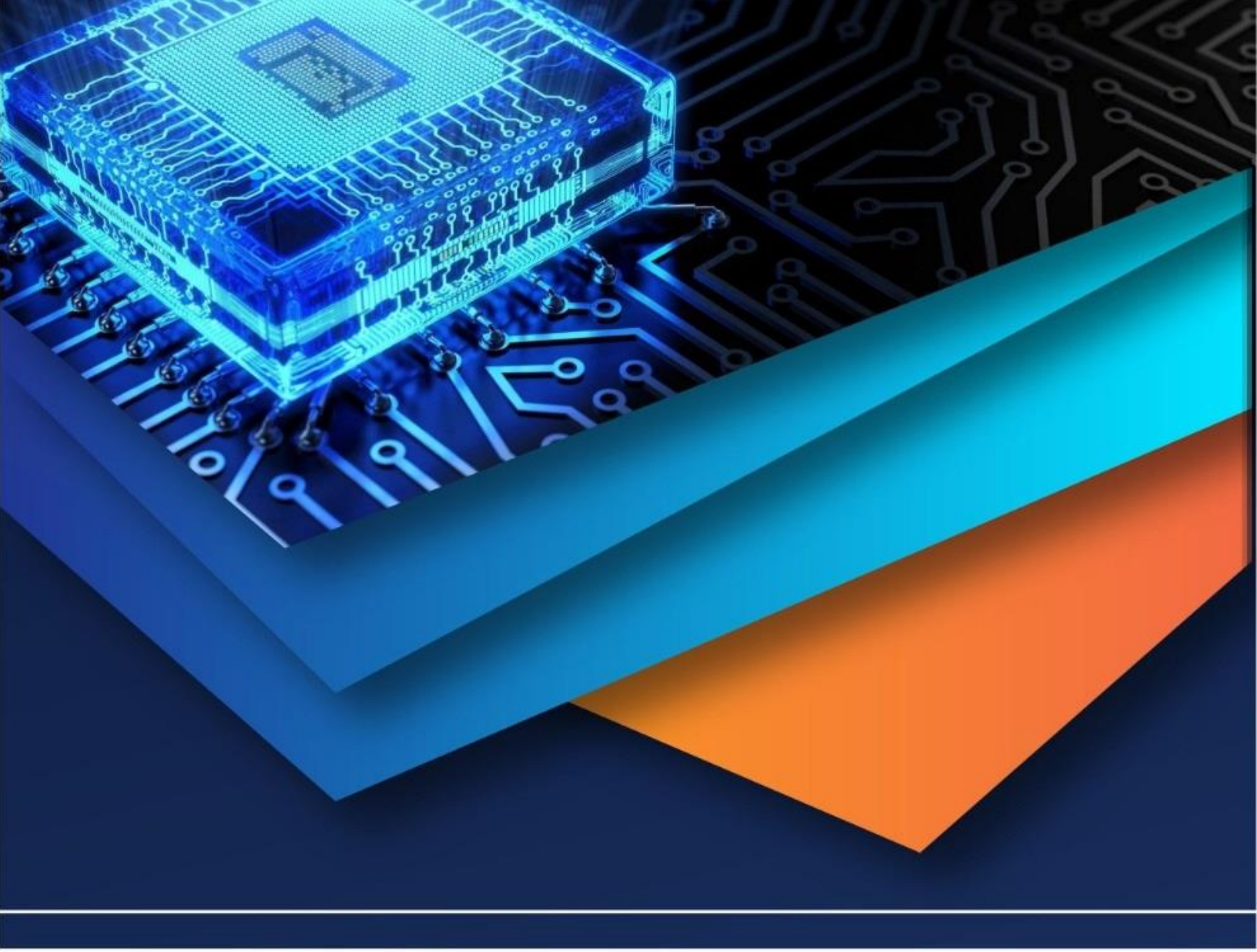
Multi-category data clustering suffered from multiple feature evaluation,
Selection of number of cluster for multi-level[1,3,6]
Diversity of feature selection process. [12]
Boundary value of cluster [9,13]

V. CONCLUSION AND FUTURE WORK

Clustering play an important role in discovery of unknown pattern for large database. 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. in this paper proposed fuzzy based two level weighted cluster technique for multi-view data. The self-optimal clustering technique faced a problem of index generation and validation of data control. For the validation of data used swarm based optimization technique. the family of swarm intelligence gives better optimal value of index for the process of cluster generation.

REFERENCES

[1] Yuwen Huang "Dynamic Cost-sensitive Ensemble Classification based on Extreme Learning Machine for Mining Imbalanced Massive Data Streams", SERSC, 2015, Pp 333-346.
[2] Benjamin X. Wang and Nathalie Japkowicz "Boosting support vector machines for imbalanced data sets", Springer, 2009, Pp 1-20.
[3] M. Mostafizur Rahman and D. N. Davis "Addressing the Class Imbalance Problem in Medical Datasets", International Journal of Machine Learning and Computing, 2013, Pp 224-228.
[4] R. J. Lyon, J. M. Brooke, J. D. Knowles and B. W. Stappers "A Study on Classification in Imbalanced and Partially-Labelled Data Streams", IEEE, 2013, Pp 1-6.
[5] Ramachandra Rao Kurada, Dr. K Karteeka Pavan, and Dr. AV Dattareya Rao "A PRELIMINARY SURVEY ON OPTIMIZED MULTIOBJECTIVE METAHEURISTIC METHODS FOR DATA CLUSTERING USING EVOLUTIONARY APPROACHES", IJCSIT, 2013, Pp 57-77.
[6] R. J. Lyon, J. M. Brooke, J. D. Knowles and B. W. Stappers "Hellinger Distance Trees for Imbalanced Streams", International Conference on Pattern Recognition, 2014, Pp 1-6.
[7] Yasser Ganjisaffar, Thomas Debeauvais, Sara Javanmardi, Rich Caruana and Cristina Videira Lopes "Distributed Tuning of Machine Learning Algorithms using MapReduce Clusters", ACM, 2011, Pp 1-8.
[8] Volkan Tunali, Turgay Bilgin, and Ali Camurcu "An Improved Clustering Algorithm for TextMining: Multi-Cluster Spherical K-Means", International Arab Journal of Information Technology, 2015, Pp 12-19.
[9] Rahul Malviya, Astt Prof.sushil Tiwari and Prof.S.R.Yadav "A Survey of Modified Support Vector Machine using Particle of Swarm Optimization for Data Classification", Journal of Advanced Computing and Communication Technologies, 2015, Pp 27-32.
[10] A. Fernandez, S. Garcia, and F. Herrera "Addressing the Classification with Imbalanced Data: Open Problems and New Challenges on Class Distribution", Springer, 2012, Pp 1-10.
[11] Shafiq Alam, Gillian Dobbie, Yun Sing Koh, Patricia Riddle and Saeed Ur Rehman "Research on particle swarm optimization based clustering: A systematic review of literature and techniques", Elsevier, 2014, Pp 1-13.
[12] LUO Xin "Chinese Text Classification Based on Particle Swarm Optimization", NCEECE, 2015, Pp 53-58.
[13] Rukshan Batuwita and Vasile Palade "Efficient Resampling Methods for Training Support Vector Machines with Imbalanced Datasets", IEEE, 2010, Pp 1-8.
[14] Angus Thomas and Yaochu Jin "Reconstructing Biological Gene Regulatory Networks: Where Optimization Meets Big Data", Springer, 2014, Pp 1-15.
[15] Mikel Galar, Alberto Fern´andez, Edurne Barrenechea, Humberto Bustince and Francisco Herrera "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches", IEEE, 2012, Pp 462-484
[16] Nishchal K. Verma, Abhishek Roy "Self-Optimal Clustering Technique Using Optimized Threshold Function" IEEE SYSTEMS JOURNAL, IEEE 2013. Pp 1-14.
[17] 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.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY