An updated literature review on the problem of Class Imbalanced Learning in Clustering

Ch.N.Santhosh Kumar, Dr. K.Nageswara Rao, Dr.A.Govardhan, Dr. K.Sudheer Reddy

***Abstract*— Clustering is one of the most interesting and important topics in data mining. Class imbalance is a problem that is very much critical in many real-world application domains of machine learning. When examples of one class in a training data set vastly outnumber examples of the other class(es), traditional data mining clustering algorithms tend to create suboptimal models. Researchers have rigorously studied several techniques to alleviate the problem of class imbalance, including resampling, algorithms, and feature selection approaches. This paper presentsa brief overview of imbalance learninginclustering, summarize well known clustering methods, discuss the majorchallenges and key issues in designing clustering algorithms, and point out some of the emerging and useful research directions inclustering.The paper also suggests a unified algorithmic framework for recent developments and describes the benchmark datasets and methodologies.**

***Index Terms*— Clustering, Imbalanced data, Clustering challenges, Future research directions.**

I. INTRODUCTION

Cluster analysis is a well-studied domain in data mining. In cluster analysis data is analyzed to find hidden relationships between each other to group a set of objects into clusters. One of the most popular methods in cluster analysis is k-means algorithm. The popularity and applicability of k-means algorithm in real time applications is due to its simplicity and high computational capability.

Researchers have identified several factors [1] that may strongly affect the *k*-means clustering analysis including high dimensionality [2]–[4], sparseness of the data [5], noise and outliers in the data [6]–[8], scales of the data [9]–[12], types of attributes [13], [14], the fuzzy index *m* [15]–[18], initial cluster centers [19]–[24], and the number of clusters [25]–[27].

In spite of the prevalence of such a large number of clusteringalgorithms, and their success in a number of different application domains, clustering remains a difficult problem. This can be attributed to the inherent vagueness in the definition of a cluster, and the difficulty in defining an appropriate similarity measure and objective function.The following fundamental challenges associated with clusteringwere highlighted in [53], which are relevant even to this day.

Manuscript received Feb. 17, 2014.

Ch.N.Santhosh Kumar, Research Scholar, Dept. of CSE, JNTU-Hyderabad, A.P., India

Dr. K.Nageswara Rao, Principal, PSCMR college of Engineering and Technology, Kothapet, Vijayawada**,** A.P., India.

Dr.A.Govardhan, Professor in CSE & SIT, JNTU Hyderabad, A.P., India.

Dr. K.Sudheer Reddy, Researcher, Hyderabad. A.P., India

(a) What is a cluster?
(b) What features should be used?
(c) Should the data be normalized?
(d) Does the data contain any outliers?
(e) How do we define the pair-wise similarity?
(f) How many clusters are present in the data?
(g) Which clustering method should be used?
(h) Does the data have any clustering tendency?
(i) Are the discovered clusters and partition valid?

However, further investigation is the need of the hour to better understand the efficiency of k-means algorithm with respect to the data distribution used for analysis. A good amount of research had done on the class balance data distribution for the performance analysis of k-means algorithm.

We review in this article the proposed methods till date with comparison and assessment, starting by sampling adjustment of class imbalance learning in section2 and related work in section 3. The trends in clustering in section4, while benchmark datasets in class imbalance learning are given in section 5. Section 6 describes the Future research directions, and conclusion is provided in section 7.

II. CLASS IMBALANCE LEARNING

One of the most popular techniques for alleviating the problems associated with class imbalance is data sampling. Data sampling alters the distribution of the training data to achieve a more balanced training data set. This can be accomplished in one of two ways: under sampling or oversampling. Under sampling removes majority class examples from the training data, while oversampling adds examples to the minority class. Both techniques can be performed either randomly or intelligently.

The random sampling techniques either duplicate (oversampling) or remove (under sampling) random examples from the training data. Synthetic minority oversampling technique (SMOTE) [2] is a more intelligent oversampling technique that creates new minority class examples, rather than duplicating existing ones. Wilson's editing (WE) [3] intelligently undersamples data by only removing examples that are thought to be noisy. In this study, we investigate the impact of intelligent oversampling technique on the performance of the clustering algorithms.While the impacts of noise and imbalance have been frequently investigated in isolation, their combined impacts have not received enough attention in research, particularly with respect to clustering algorithms. To alleviate

this deficiency, we present a comprehensive empirical investigation of learning from noisy and imbalanced data using k-means clustering algorithm.

Finding minority class examples effectively and accurately without losing overall performance is the objective of class imbalance learning. The fundamental issue to be resolved is that the clustering ability of most standard learning algorithms is significantly compromised by imbalanced class distributions. They often give high overall accuracy, but form very specific rules and exhibit poor generalization for the small class. In other words, overfitting happens to the minority class [6], [36], [37], [38], [39]. Correspondingly, the majority class is often overgeneralized. Particular attention is necessary for each class. It is important to know if a performance improvement happens to both classes and just one class alone. Many algorithms and methods have been proposed to ameliorate the effect of class imbalance on the performance of learning algorithms. There are three main approaches to these methods.

• *Internal approaches acting on the algorithm*. These approaches modify the learning algorithm to deal with the imbalance problem. They can adapt the decision threshold to create a bias toward the minority class or introduce costs in the learning process to compensate the minority class.
• *External approaches acting on the data*. These algorithms act on the data instead of the learning method. They have the advantage of being independent from the classifier used. There are two basic approaches: oversampling the minority class and undersampling the majority class.
• *Combined approaches that are based on boosting accounting for the imbalance in the training set*. These methods modify the basic boosting method to account for minority class underrepresentation in the data set. There are two principal advantages of choosing sampling over cost-sensitive methods. First, sampling is more general as it does not depend on the possibility of adapting a certain algorithm to work with classification costs. Second, the learning algorithm is not modified, which can cause difficulties and add additional parameters to be tuned. The different imbalance data learning approaches are as follows [59]:

Table 1. Imbalanced Data learning Approaches

- ❖ SAMPLING METHODS
  - ✓ BASIC SAMPLING METHODS
    - ➢ Under-Sampling
    - ➢ Over-Sampling
  - ✓ ADVANCED SAMPLING METHODS
    - ➢ Tomek Link
    - ➢ The SMOTE approach
    - ➢ Borderline-SMOTE
    - ➢ One-Sided Selection OSS
    - ➢ Neighbourhood Cleaning Rule (NCL)
    - ➢ Bootstrap-based Over-sampling (BootOS)
- ❖ ENSEMBLE LEARNING METHODS
  - ✓ BAGGING
    - ➢ Asymmetric bagging, SMOTE Bagging
    - ➢ Over Bagging, Under Bagging
    - ➢ Roughly balanced bagging
    - ➢ Lazy Bagging
    - ➢ Random features selection
  - ✓ BOOSTING
    - ➢ Adaboost
    - ➢ SMOTEBoost
    - ➢ DataBoost-IM
  - ✓ RANDOM FORESTS
    - ➢ Balanced Random Forest BRF
    - ➢ Weighted Random Forest WRF
- ❖ COST-SENSITIVE LEARNING
  - ✓ Direct cost-sensitive learning methods
  - ✓ Methods for cost-sensitive meta-learning
  - ✓ Cost-sensitive meta-learning
  - ✓ Thresholdingmethods
  - ✓ MetCost
  - ✓ Cost-sensitive meta-learning sampling methods
- ❖ FEATURE SELECTION METHODS
  - ✓ Warpper
  - ✓ PREE (Prediction Risk based feature selection for Easy Ensemble)
- ❖ ALGORITHMS MODIFICATION
  - ✓ Proposal for new splitting criteria DKM
  - ✓ Adjusting the distribution reference in the tree
  - ✓ Offset Entropy

III. RELATED WORK

In, this section, we first review the major research about clustering in class imbalance learning. In recent years, clustering techniques have received much attention in wide areas of applicability such as medicine, engineering, finance and biotechnology. The main intention of clustering is to group data together which are having similar characteristics. The clustering can also be referred as "the art of finding groups in data". It's not fair to declare one clustering method as the best clustering method since the success of clustering method will highly depend on the type of data and the way of investigation for a specific applicability. Although many researchers attempted to make clustering process as a pure statistical technique but still largely it is regarded as an exploration procedure for finding the similar group of data.

Guhaet al.[37] early proposed to make use of multiple representative points to get the shape information of the "natural" clusters with nonspherical shapes and achieve an improvement on noise robustness over the single-link algorithm. Liu *et al.* [38], proposed a multiprototype clustering algorithm, which applies the *k*-means algorithm to discover clusters of arbitrary shapes and sizes. However, there are following problems in the real applications of these algorithms to cluster imbalanced data. 1) These algorithms depend on a set of parameters whose tuning is problematic in practical cases. 2) These algorithms make use of the randomly sampling technique to find cluster centers. However, when data are imbalanced, the selected samples more probably come from the majority classes than the minority classes. 3)

The number of clusters *k* needs to be determined in advance as an input to these algorithms. In a real dataset, *k* is usually unknown. 4) The separation measures between subclusters that are defined by these algorithms cannot effectively identify the complex boundary between two subclusters. 5) The definition of clusters in these algorithms is different from that of *k*-means. Xiong*et al.* [33] provided a formal and organized study of the effect of skewed data distributions on the hard *k*-means clustering. However, the theoretic analysis is only based on the hard *k*-means algorithm.

Haitaoxiang et al., [39]have proposed a localclustering ensemble learning method based on improved AdaBoost (LCEM) for rare class analysis. LCEM usesan improved weight updating mechanism where the weights of samples which are invariably correctlyclassified will be reduced while that of samples which are partially correctly classified will be increased. The proposed algorithm also perform clustering on normal class and produce sub-classes with relatively balanced sizes.AmuthanPrabakar et al., [40] have proposed a supervised network anomaly detection algorithm by the combination of k-means and C4.5 decision tree exclusively used for portioning and model building of the intrusion data. The proposed method is used mitigating the Forced Assignment and Class Dominance problems of the k-Means method.

Li Xuan et al., [41]have proposed two methods, in first method they applied random sampling of majority subset to form multiple balanced datasets for clustering and in second methodthey observed the clustering partitions of all the objects in the dataset under the condition of balance and imbalance at a different angle. Christos Bouraset al., [42]have proposed W-k meansclustering algorithm for applicability on a corpus of news articles derived from major news portals. The proposed algorithm is an enhancement of standard k-means algorithm using the external knowledge for enriching the ‘‘bag of words’’ used prior to the clustering process andassisting the label generation procedure following it.

P.Y. Mok et al., [43]have proposed a new clustering analysis method that identifies the desired cluster number and produces, at the same time, reliable clustering solutions. It first obtains many clustering results from a specific algorithm, such as Fuzzy C-Means (FCM), and then integrates these different results as a judgment matrix. An iterative graph-partitioning process is implemented to identify the desired cluster number and the final result.

Luis A. Leivaet al., [44]have proposed Warped K-Means, a multi-purpose partition clustering procedure that minimizesthe sum of squared error criterion, while imposing a hard sequentiality constraint in theclassification step on datasets embedded implicitly with sequential information.The proposed algorithm is also suitable for online learning data, since the change of number of centroids and easy updating of new instances for the final cluster is possible.M.F.Jianget al., [45] have proposed variations of k-means algorithm to identify outliersby clustering the data the initial phase then using minimum spanning tree to identify outliers for their removal.

Jie Cao et al., [46] have proposed a Summation-bAsed Incremental Learning (SAIL) algorithm for Information-theoretic K-means (Info-Kmeans) aims to cluster high-dimensional data, such as images featured by the bag-of-features (BOF) model, using K-means algorithm with KL-divergence as the distance. Since SAIL is a greedy scheme it first selects an instance from data and assigns it to the most suitable cluster. Then the objective-function value and other related variables are updated immediately after the assignment. The process will be repeated until some stopping criterion is met. One of the shortcomings is to select the appropriate cluster for an instance.Max Mignotte[47] has proposed anew and simple segmentation method based on the K-means clustering procedure for applicability on image segmentation.The proposed approach overcome the problem of local minima, feature space without considering spatial constraints and uniform effect.

IV. TRENDS IN CLUSTERING

Information explosion is not only creating large amounts of databut also a diverse set of data, both structured and unstructured. Unstructured data is a collection of objects that do not follow a specific format. For example, images, text, audio, video, etc. On the other hand, in structured data, there are semantic relationships within each object that are important. Most clustering approaches ignore the structure in the objects to be clustered and use a feature vector based representation for both structured and unstructured data. The traditional view of data partitioning based on vector based feature representation does not always serve as an adequate framework. Examples include objects represented using sets of points [54], consumer purchase records[55], data collected from questionnaires and rankings [56], social networks [57], and data streams[58]. Models and algorithms are being developed to process huge volumes of heterogeneous data. A brief summary of some of the recent trends in data clustering is presentedbelow.

(i) Clustering Ensembles:

The basic idea is that by taking multiple looks at the same data, one can generate multiple partitions (clustering ensemble) of the same data. By combining the resulting partitions, it is possible to obtain a good data partitioning even when the clusters are not compact and well separated.

(ii) Semi-supervised clustering:

Clustering is inherently an ill-posed problem where the goal is to partition the data into some unknown number of clusters based on intrinsic information alone. The data-driven nature of clustering makes it very difficult to design clustering algorithms that will correctlyfind clusters in the given data. Any external or side information available along with the ‘n x d’ pattern matrix or the ‘n x n’ similarity matrix can be extremely useful in finding a good partition of data. Clustering algorithms that utilize such side information are said to be operating in a semi-supervised mode

(iii) Large-scale clustering:

Large-scale data clustering addresses the challenge of clusteringmillions of data points that are represented in thousands of features.

(iv) Multi-way clustering:

The co-clustering framework was extended to multi way clusteringto cluster a set of objects by simultaneously clustering their heterogeneous components. Indeed, the problem is much more challenging because different pairs of components may participate in different types of similarity relationships.

(v) Heterogeneous data:

Heterogeneous data refers to the data where the objectsmay not be naturally represented using a fixed length feature vector.

V. Benchmark Datasets in Class Imbalance Learning

Table 2summarizes the benchmark datasetsused in almost all the recent studies conducted on class imbalance learning for clustering. The details of the datasets are given in table 2. For each data set, the number of examples (#Ex.), number of attributes (#Atts.), class name of each class (minority and majority) and IR is given. This table is ordered by the IR, from low to high imbalanced data sets.

TABLE 2
SUMMARY OF BENCHMARK IMBALANCED DATASETS

Datasets	# Ex.	# Atts.	Class (_,+)	IR
Glass1	214	9	(build-win-non_float-proc;remainder)	1.82
Ecoli0vs1	2207		(im;cp)	1.86
Wisconsin	683	9	(malignant;benign)	1.86
Pima	768	8	(tested-positive; tested-negative)	1.90
Iris0	150	4	(Iris-Setosa;remainder)	2.00
Glass2	214	9	(build-win-float-proc;remainder)	2.06
Yeast1	1484	8	(nuc;remainder)	2.46
Vehicle1	846	18	(Saab;remainder)	2.52
Vehicle2	846	18	(Bus;remainder)	2.52
Vehicle3	846	18	(Opel;remainder)	2.52
Haberman	306	3	(Die; Survive)	2.68
Glass3	214	9	(non-window glass;remainder)	3.19
Vehicle0	846	18	(Van;remainder)	3.23
Ecoli1	336	7	(im;remainder)	3.36
Thyroid2	2155		(hypo;remainder)	4.92
Thyroid1	2155		(hyper;remainder)	5.14
Ecoli2	336	7	(pp;remainder)	5.46
Segment0	230819		(brickface;remainder)	6.01
Glass6	214	9	(headlamps;remainder)	6.38
Yeast3	1484	8	(me3;remainder)	8.11
Ecoli3	336	7	(imU;remainder)	8.19
Page-blk15	472	10	(remainder;text)	8.77
Ecoli4	200	7	(p,imL,imU;om)	9.00
Yeast2	5148		(cyt;me2)	9.08
Ecoli05	222	7	(cp,omL,pp;imL,om)	9.09
Ecoli06	202	7	(cp,imS,imL,imU;om)	9.10
Glass4	172	9	(build-win-non_float-proc, tableware,build-win-float-proc; ve-win-float-proc)	9.12
Yeast4	506	8	(mit,me1,me3,erl; vac,pox)	9.12
Yeast5	1004	8	(mit, cyt,me3,vac,erl;me1,exc,pox)	9.14
Yeast6	10048		(mit, cyt,me3,exc;me1,vac,pox, erl)	9.14
Ecoli7	203	6	(cp,imU,omL;om)	9.15
Ecoli8	244	7	(cp,im;imS,imL,om)	9.17
Ecoli9	224	7	(cp,imS,omL,pp;imL,om)	9.18
Glass5	929	(build-win-float-proc,containers; tableware)		9.22
Ecoli10	205	7	(cp,imL,imU,omL;om)	9.25
Ecoli11	257	7	(cp,imL,imU,pp;om,omL)	9.28
Yeast7	528	8	(me2;mit,me3,exc, vac, erl)	9.35
Ecoli12	220	6	(cp,omL,pp;om)	10.00
Vowel	988	13	(hid;remainder)	10.10
Glass6	192	9	(ve-win-float-proc; build-win-float-proc, build-win-non_float-proc,headlamps)	10.29
Glass7	2149		(Ve-win-float-proc;remainder)	10.39
Ecoli13	336	7	(cp,im,imU,pp;imS,imL,om,omL)	10.59
Led7digit	4437		(0, 2, 4, 5, 6, 7, 8, 9;1)	10.97
Glass7	108	9	(build-win-float-proc,headlamps; tableware)	11.00
Ecoli14	240	6	(cp,im;om)	11.00
Glass8	205	9	(build-win-float-proc, containers,headlamps, build-win-non_float-proc; ve-win-float-proc)	11.06
Ecoli15	332	6	(cp,im,imU,pp;om,omL)	12.28
Cleveland	177	13	(0; 4)	12.62
Ecoli16	280	6	(cp,im,imU,omL;om)	13.00
Ecoli17	336	7	(om;remainder)	13.84
Yeast8	459	8	(nuc; vac)	13.87
Shuttle1	1829	9	(Rad Flow;Bypass)	13.87
Glass4	214	9	(containers;remainder)	15.47
Page-blk47	210		(graphic; horiz.line,picture)	15.85
Abalone7	731	8	(18; 9)	16.68
Glass9	184	9	(tableware; build-win-float-proc, build-win-non_float-proc, headlamps)	19.44
Shuttle2	129	9	(FpvOpen;Bypass)	20.5
Yeast9	693	8	(vac; nuc,me2,me3,pox)	22.10
Glass10	214	9	(tableware;remainder)	22.81
Yeast10	482	8	(pox;cyt)	23.10
Yeast11	1484	8	(me2;remainder)	28.41
Yeast12	947	8	(vac; nuc,cyt,pox,erl)	30.56
Yeast13	14848		(me1;remainder)	32.78
Ecoli18	281	7	(pp,imL;cp,im,imU,imS)	39.15
Yeast14	14848		(exc;remainder)	39.15
Abalone19	4174	8	(19;remainder)	128.87

The imbalance ratio (IR)is obtained by dividing the number of positive samples over the number of negative samples. A dataset is termed balance if the imbalance ratio is one. The complete details regarding all the datasets can be obtained from Victoria López*et al.* [38] and Machine Learning Repository [52].

VI. Future Research Directions

Clustering has numerous success stories in data analysis. Inspite of this, machine learning and pattern recognition communities need to address a number of issues to improve our understanding of data clustering. Below is a list of problems and research directions that are worth focusing in this regard.

(a) There needs to be a suite of benchmark data (with ground truth) available for the research community to test and evaluate clustering methods. The benchmark should include data sets from various domains (documents, images, time series, customer transactions, biological sequences, social networks, etc.). Benchmark should also include both staticand

dynamic data (the latter would be useful in analyzing clusters that change over time), quantitative and/or qualitative attributes, linked and non-linked objects, etc. Though the idea of providing a benchmark data is not new (e.g., UCI ML and KDD repository), current benchmarks are limited
to small, static data sets.

(b) We need to achieve a tighter integration between clustering algorithms and the application needs. For example, some applications may require generating only a few cohesiveclusters (less cohesive clusters can be ignored), while othersmay require the best partition of the entire data. In mostapplications, it may not necessarily be the best clustering algorithm that really matters. Rather, it is more crucial to choose the right feature extraction method that identifies the underlying clustering structure of the data.

(c) Regardless of the principle (or objective), most clustering methods are eventually cast into combinatorial optimization problems that aim to find the partitioning of data that optimizes the objective. As a result, computational issue becomes critical when the application involves large-scale data. For instance, finding the global optimal solution for K-means is NP-hard. Hence, it is important to choose clustering principles that lead to computationally efficient solutions.

(d) A fundamental issue related to clustering is its stability or consistency. A good clustering principle should result in a data partitioning that is stable with respect to perturbations in the data. We need to develop clustering methods that lead to stable solutions.

(e) Choose clustering principles according to their satisfiabilityof the stated axioms. Despite Kleinberg's impossibility theorem, several studies have shown that it can be overcome by relaxing some of the axioms. Thus, maybe one way to evaluate a clustering principle is to determine to what degree it satisfies the axioms.

(f) Given the inherent difficulty of clustering, it makes more sense to develop semi-supervised clustering techniques in which the labeled data and (user specified) pair-wise constraints can be used to decide both (i) data representation and (ii) appropriate objective function for data clustering.

VII. Conclusion

In this paper, the state of the art methodologies to deal with class imbalance problem in clustering has been reviewed. The road of challenges in clustering and the future research directions are also provided for the interesting research community. However, there was a lack of framework where each of the clustering algorithms could be modeled; for this reason, a taxonomy where they can be placed has been taken as our future work. Finally, we have concluded that intelligence based algorithms are the need of the hour for improving the results that are obtained by the usage of data preprocessing.

References

[1] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*.Reading, MA: Addison-Wesley, 2005.
[2] Z. X. Huang, M. K. Ng, H. Rong, and Z. Li, "Automated variable weightingin *k*-means type clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*,vol. 27, no. 5, pp. 657–668, May 2005.
[3] E. Y. Chan, W. K. Ching, M. K. Ng, and Z. X. Huang, "An optimizationalgorithm for clustering using weighted dissimilarity measures," *PatternRecognit.*, vol. 37, no. 5, pp. 943–952, 2004.
[4] Y. H. Qian, J. Y. Liang, W. Pedrycz, and C. Y. Dang, "Positive approximation:An accelerator for attribute reduction in rough set theory," *Artif.Intell.*, vol. 174, no. 5–6, pp. 597–618, 2010.
[5] L. P. Jing, M. K. Ng, and Z. X. Huang, "An entropy weighting *k*-meansalgorithm for subspace clustering of high-dimensional sparse data," *IEEETrans. Knowl. Data Eng.*, vol. 19, no. 8, pp. 1026–1041, Aug. 2007.
[6] J. S. Zhang and Y. W. Leung, "Robust clustering by pruning outliers,"*IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 33, no. 6, pp. 983–998,Dec. 2003.
[7] A. Zhou, F. Cao, Y. Yan, C. Sha, and X. He, "Distributed data streamclustering: A fast EM-based approach," in *Proc. 23rd Int. Conf. DataEng.*, 2007, pp. 736–745.
[8] M. Breunig, H. P. Kriegel, R. Ng, and J. Sander, "LOF: Identifying densitybased local outliers," in *Proc. Int. Conf. ACM Special Interest GroupManag. Data*, 2000, pp. 427–438.
[9] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithmfor discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. ACM Special Interest Group Knowl. Discovery Data Mining*,1996, pp. 226–231.
[10] P. Bradley, U. Fayyad, and C. Reina, "Scaling clustering algorithms tolarge databases," in *Proc. 4th Int. Conf. ACM Special Interest GroupKnowl. Discovery Data Mining*, 1998, pp. 9–15.
[11] F. Murtagh, "Clustering massive data sets," in *Handbook of Massive DataSets*. Norwell, MA: Kluwer, 2000.
[12] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient dataclustering method for very large databases," in *Proc. ACMSpecial InterestGroup Manag. Data*, 1996, pp. 103–114.
[13] Z. X. Huang, "Extensions to the *k*-means algorithm for clustering largedata sets with categorical values," *Data Mining Knowl. Discov.*, vol. 2,no. 3, pp. 283–304, 1998.
[14] F. Y. Cao, J. Y. Liang, L. Bai, X. Zhao, and C. Dang, "A frameworkfor clustering categorical time-evolving data," *IEEE Trans. Fuzzy Syst.*,vol. 18, no. 5, pp. 872–882, Oct. 2010.
[15] J. C. Bezdek, "A physical interpretation of Fuzzy ISODATA," *IEEETrans. Syst., Man, Cybern.*, vol. SMC-6, no. 5, pp. 387–390, May 1976.
[16] L. O. Hall, A. M. Bensaid, and L. P. Clarke, "A comparison of neural networkand fuzzy clustering techniques in segmenting magnetic resonanceimages of the brain," *IEEE Trans. Neural Netw.*, vol. 3, no. 5, pp. 672–682,Sep. 1992.
[17] R. L. Cannon, J. V. Dave, and J. C. Bezdek, "Efficient implementationof the fuzzy c-means clustering algorithms," *IEEE Trans. Pattern Anal.Mach. Intell.*, vol. PAMI-8, no. 2, pp. 248–255, Mar. 1986.
[18] J. Yu and M. S. Yang, "Optimality test for generalized FCM and itsapplication to parameter selection," *IEEE Trans. Fuzzy Systems*, vol. 13,no. 1, pp. 164–176, Feb. 2005.
[19] F. Y. Cao, J. Y. Liang, and G. Jiang, "An initialization method for the*k*-means algorithm using neighborhood model," *Comput. Math. Appl.*,vol. 58, no. 3, pp. 474–483, 2009.
[20] M. Laszlo and S. Mukherjee, "A genetic algorithm using hyper-quadtreesfor low-dimensional k-means clustering," *IEEE Trans. Pattern Anal.Mach. Intell.*, vol. 28, no. 4, pp. 533–543, Apr. 2006.
[21] D. Arthur and S. Vassilvitskii, "K-means++: the advantages of carefulseeding," in *Proc. 18th Annu. ACM-SIAM Symp. Discrete Algo.*, 2007,pp. 1027–1035.
[22] A. Likas, M. Vlassis, and J. Verbeek, "The global *k*-means clusteringalgorithm," *Pattern Recognit.*, vol. 35, no. 2, pp. 451–461, 2003.
[23] A. M. Bagirov, "Modified global *k*-means algorithm for minimum sum-ofsquaresclustering problems," *Pattern Recognit.*, vol. 41, no. 10, pp. 3192–3199, 2008.

[24] Z. C. Lai and T. J. Huang, "Fast global *k*-means clustering using clustermembership and inequality," *Pattern Recognit.*, vol. 43, no. 5, pp. 1954–1963, 2010.
[25] G. Hamerly and C. Elkan, "Learning the *k* in *k*-means," in *Proc. 17thAnn. Conf. Neural Inf. Process. Syst.*, Dec. 2003, pp. 1–8.
[26] J. J. Li, M. K. Ng, Y. M. Cheng, and Z. H. Huang, "Agglomerative fuzzy*k*-means clustering algorithm with selection of number of clusters," *IEEETrans. Knowl. Data Eng.*, vol. 20, no. 11, pp. 1519–1534, Nov. 2008.
[27] M. Halkidi and M. Vazirgiannis, "A density-based cluster validity approachusing multi-representatives," *Pattern Recognit. Lett.*, vol. 29,pp. 773–786, 2008.
[28] H. Xiong, J. J. Wu, and J. Chen, "K-means clustering versus validationmeasures: A data-distribution perspective," *IEEE Trans. Syst., Man, Cybern.B, Cybern.*, vol. 39, no. 2, pp. 318–331, Apr. 2009.
[29] W.-Z. Lu and D.Wang, "Ground-level ozone prediction by support vector machine approach with a cost-sensitive classification scheme," *Sci. Total. Enviro.*, vol. 395, no. 2-3, pp. 109–116, 2008.
[30] Y.-M. Huang, C.-M. Hung, and H. C. Jiau, "Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem," *Nonlinear Anal. R. World Appl.*, vol. 7, no. 4, pp. 720–747, 2006.
[31] D. Cieslak, N. Chawla, and A. Striegel, "Combating imbalance in network intrusion datasets," in *IEEE Int. Conf. Granular Comput.*, 2006, pp. 732–737.
[32] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," *Neural Netw.*, vol. 21, no. 2–3, pp. 427–436, 2008.
[33] A. Freitas, A. Costa-Pereira, and P. Brazdil, "Cost-sensitive decision trees applied to medical data," in *Data Warehousing Knowl. Discov. (Lecture Notes Series in Computer Science)*, I. Song, J. Eder, and T. Nguyen, Eds.,
[34] K.Kilic¸,O¨ zgeUncu and I. B. Tu¨rksen, "Comparison of different strategies of utilizing fuzzy clustering in structure identification," *Inf. Sci.*, vol. 177, no. 23, pp. 5153–5162, 2007.
[35] M. E. Celebi, H. A. Kingravi, B. Uddin, H. Iyatomi, Y. A. Aslandogan, W. V. Stoecker, and R. H. Moss, "A methodological approach to the classification of dermoscopy images," *Comput.Med. Imag. Grap.*, vol. 31, no. 6, pp. 362–373, 2007.
[36] X. Peng and I. King, "Robust BMPM training based on second-order cone programming and its application in medical diagnosis," *Neural Netw.*, vol. 21, no. 2–3, pp. 450–457, 2008.Berlin/Heidelberg, Germany: Springer, 2007, vol. 4654, pp. 303–312.
[37] S. Guha, R. Rastogi, and K. Shim, "Cure: An efficient clustering algorithm for large databases," in *Proc. Int. Conf. ACM Special Interest Group Manag. Data*, 1998, pp. 73–84.
[38] M. H. Liu, X. D. Jiang, and A. C. Kot, "A multi-prototype clustering algorithm," *Pattern Recognit.*, vol. 42, pp. 689–698, 2009.
[39] Haitaoxiang , Yi yang, Shouxiangzhao. "Local Clustering Ensemble Learning Method Based on Improved AdaBoost for Rare Class Analysis", Journal of Computational Information Systems 8: 4 (2012) 1783{1790, pp,no:1783 – 1790.
[40] AmuthanPrabakarMuniyandi, R. Rajeswari, R. Rajaram. Network Anomaly Detection by Cascading K-Means Clustering and C4.5 Decision Tree algorithm, International Conference on Communication Technology and System Design 2011, Procedia Engineering 30 (2012) 174 – 182.
[41] Li Xuan, Chen Zhigang, Yang Fan. "Exploring of clustering algorithm on class-imbalanced data".
[42] C. Bouras, V. Tsogkas, A clustering technique for news articles using WordNet, Knowl. Based Syst. (2012), http://dx.doi.org/10.1016/j.knosys.2012.06.015.
[43] P.Y. Mok, H.Q.Huang,Y.L.Kwok,J.S.Au. "A robust adaptive clustering analysis method for automatic identification of clusters", Pattern Recognition 45 (2012) 3017–3033.
[44] Luis A. Leiva, Enrique Vidal." Warped K-Means: An algorithm to cluster sequentially-distributed data", Information Sciences 237 (2013) 196–210.
[45] M.F.Jaing, S.S.Tseng and C.M. Su, "Two Phase Clustering Process for Outlier Detection", pattern recognition letters 22 (2001) pp no: 691-700.
[46] Jie Cao, ZhiangWu, JunjieWu and WenjieLiu, "Towards information-theoretic K-means clustering for image indexing", Signal Processing 93 (2013) 2026–2037.
[47] Mignotte, M. A de-texturing and spatially constrained K-means approach for image segmentation. Pattern Recognition Lett. (2010), doi:10.1016/j.patrec.2010.09.016
[48] O. Maimon, and L. Rokach, *Data mining and knowledge discovery handbook*, Berlin: Springer, 2010.
[49] J. R. Quinlan, *C4.5: Programs for Machine Learning*, 1st ed. San Mateo, CA: Morgan Kaufmann Publishers, 1993.
[50] http://www.keel.es/
[51] Witten, I.H. and Frank, E. (2005) Data Mining:Practical machine learning tools and techniques.2nd edition Morgan Kaufmann, San Francisco.
[52] Blake C, Merz CJ (2000) UCI repository of machine learning databases. Machine-readable data repository. Department of Information and Computer Science, University of California at Irvine, Irvine. http://www.ics.uci.edu/mlearn/MLRepository.html
[53] Jain, Anil K., Dubes, Richard C., 1988. Algorithms for Clustering Data. Prentice Hall.
[54] Lowe, David G., 2004. Distinctive image features from scale-invariant keypoints.Internat. J. Comput. Vision 60 (2), 91–110.
[55] Guha, Sudipto, Rastogi, Rajeev, Shim, Kyuseok, 2000. Rock: A robust clusteringalgorithm for categorical attributes. Inform. Systems 25 (5), 345–366.
[56] Critchlow, D., 1985. Metric Methods for Analyzing Partially Ranked Data. Springer
[57] Wasserman, S., Faust, K., 1994. Social Network Analysis: Methods and Applications.Cambridge University Press.
[58] Guha, Sudipto, Mishra, Nina, Motwani, Rajeev, 2003. Clustering data streams. IEEE Trans. Knowledge Data Eng. 15 (3), 515–528.
[59] Mohamed Bekkar and Dr. TaklitAkroufAlitouche, 2013.Imbalanced Data Learning Approaches Review. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.4, July 2013

Ch.N.Santhosh Kumar, Research Scholar, Dept. of CSE, JNTU-Hyderabad, A.P., India
Dr. K.Nageswara Rao, Principal, PSCMR college of Engineering and Technology, Kothapet, Vijayawada**,** A.P., India.
Dr.A.Govardhan, Professor in CSE & SIT, JNTU Hyderabad, A.P., India.
Dr. K.Sudheer Reddy, Researcher, Hyderabad. A.P., India