International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161

Available at http://inpressco.com/category/ijcet

Review Article

Review: Approaches for Handling DataStream

Purva S. Gogte[Å*] and Deepti P. Theng[Å]

[Å] Department of Computer Science and Engineering, G. H. Raisoni College of Engineering, Nagpur, Maharashtra, India-441110

Accepted 10 January 2014, Available online 01 February 2014, **Vol.4, No.1 (February 2014)**

Abstract

Today, there is tremendous use of technology that causes generation of huge volume of data called as Data Stream. Data Stream are continuous, unbounded and usually come with high speed and changes with time. It has different issues such as Memory, Time, Noise, Dynamic data. There is need of handling data streams because of its changing nature, and the data stream may be labelled or it may be unlabelled. Classification is supervised it can only handle labelled data. Thus, there is need of Hybrid Ensemble Classifier in which clustering and classifier are brought together so that the labelled as well as unlabelled datastream both can be handled. This Paper describes different Approaches for Handling DataStream.

Keywords: *Data Streams, Clustering, Classification*

1. Introduction

In recent years, many sources of streaming data have been developed. Tens of applications and millions of users access the World Wide Web daily. Moreover, advances in hardware devices, like wireless sensors and mobile devices, led to an increase in the applications that generate streaming data.(Satpute Pravin C,2012).Data Stream is a sequence of continuously arriving data items at a high speed which are real time, implicitly or explicitly ordered by timestamps, evolving and uncertain in nature. Data Stream mining has recently emerged as a growing field of multidisciplinary research. It combines various research areas such as databases, machine learning, artificial intelligence, statistics, automated scientific discovery data visualization, decision science, and high performance computing thus, Data stream classification has been a widely studied research problem in recent years. The dynamic and evolving nature of data stream requires efficient and effective techniques that are significantly different from static data classification techniques. In recent years mining data streams in large real time environments has become a challenging job due to wide range of applications that generate boundless stream of data such as log records, mobile application sensors, emails, blogging, credit card, fraud detection, medical imaging, intrusion detection, weather monitoring, stock trading, planetary remote sensing etc.

There are many issues while handling with the data streams which are summarized as follows:

i) Large space: Data streams have enormous volumes of continuously incoming data.

ii) Dynamic data: Data streams are fast, changing, uncertain and require fast response to incorporate changes in data and reflect it in output.

iii) Noise: Any approach applied to data streams should be able to deal with noise and outliers.

iv) Single scan: Since data streams have infinite volume of information which is fast and changing, hence stream data should be read only once.

v) Light weight: Techniques applied to vast data streams should process stream less time and memory to should provide an optimal output

Data Stream are nothing but the Big data .The term “Big data” is used for large data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process. Big data sizes are a constantly moving target currently ranging from a few dozen terabytes to many petabytes of data in a single data set. Typical examples of big data found in current scenario includes web logs, RFID generated data, sensor networks, satellite and geo-spatial data, social data from social networks, Internet text and documents, Internet search indexing, call detail records, astronomy, atmospheric science, genomics, biogeochemical etc. Big Data has emerged because we are living in a society which makes increasing use of data intensive technologies.

There are many Big data problems such as it is difficult to use relational databases with big data. The various challenges faced in large data management include scalability, unstructured data, accessibility, real time analytics, fault tolerance and many more. In addition to variations in the amount of data stored in different sectors, the types of data generated and stored i.e., whether the data encodes video, images, audio, or text/numeric information also differ markedly from industry to industry

*Corresponding author: **Purva S. Gogte**

Clustering and Classification are the two techniques that are widely used to extract patterns from the large data stream. Clustering is one of the most vital tasks in data mining field. It groups the similar data points into one cluster and dissimilar data points into another cluster. It helps in uncovering useful structures in data that were previously unknown. Classification is a process where a system first learns from the class label of known data means training Then, it use certain rule based on which it then predicts the class of unforeseen data (testing). Classification is supervised thus it is able to work only on labeled data set. Unsupervised approach encompasses clustering in which dataset are unlabelled. The classifiers are of mainly two types Single Classifier and Ensemble Classifier. Tradionally, Single Classifier were used but, recently an ensemble classifier is used.

Ensemble Classifier is combination of multiple classifiers. An ensemble classifier is conventionally constructed from a set of base classifiers that separately learn the class boundaries over the patterns in a training set. The decision of an ensemble classifier on a test pattern is produced by fusing the individual decisions of the base classifiers. Ensemble classifiers are also known as multiple classifier systems, committee of classifiers, and mixture of experts. The main goal in ensemble learning is to build diverse base classifiers.

2. Literature Survey

As streaming data analysis techniques are given increasing attention in recent years, different forms of analysis are studied in the literature for different applications. Classification and Clustering are the basic techniques that are widely used for analysis of streaming data .The clustering process of streaming data targets discovering similar groups of data while the stream is flowing.

Classification is supervised learning method as in this class labels are already defined. There are major two types of classifiers.

Single Classifier: These are the traditional Classifier such as Naive Bayes, Nearest Neighbour Methods, and decision rules.

Ensemble Classifier: An ensemble Classifier is formed by combining multiple classifiers.

a) Single Classifier

VFDT (P. Domingos and G. Hulten , 2000) .It is a decision tree algorithm .It helps to overcome the long training times issue. VFDT is used for the Real-Time Data Mining of Imperfect Data Streams in a Distributed Wireless Sensor Network. It has solved the problem of long training time. The VFDT system constructs a decision tree by using constant memory and constant time per sample. VFDT algorithm is based on a decision-tree learning method combined with sub-sampling of the entire data stream.

CVFDT(Geoff Hulten, Laurie Spencer, Pedro Domingos, 2001)Concept Adapting very Fast Decision is used to handle concept drift issue by growing alternate trees and subtrees. This algorithm mines high-speed data streams under the approach of one-pass mining

Random forests(L. Breiman, 2001) are nothing but the combination of tree predictor. Each tree depends on the values of a random vector sampled independently. Suppose we are having given a training set S. Build subset Si by sampling and replacement. Choose best split from random subset of F features .Make predictions according to majority vote of the set of trees.

Decision Tree (J.S R Jang,2000) partitions the input space of a data set into mutually exclusive regions, each of which is assigned a label. The decision tree mechanism is transparent and we can follow a tree structure easily to see how the decision is made A decision tree is a tree structure consisting of internal and external nodes connected by branches. An internal node is a decision making unit that evaluates a decision function to determine which child node to visit next. The external node, on the other hand, has no child nodes and is associated with a label or value that characterizes the given data that leads to its being visited. However, many decision tree construction algorithms involve a two step process. First, a very large decision tree is grown. Then, to reduce large size and overfiting the data, in the second step, the given tree is pruned .The pruned decision tree that is used for classification purposes is called the classification tree.

Naive Bayes (G. Qiang,2000) is based on the Bayes theorem and computes class-conditional probabilities for each new example. Bayesian methods learn incrementally by nature and require constant memory. Naive Bayes is a lossless classifier.

Nearest Neighbor (T. Darrell and P. Indyk and G. Shakhnarovich,2006)Classifiers, also called instance-based learners or lazy learners, provide an accurate way of learning data incrementally. Each processed example is stored and serves as a reference for new data points. Classification is based on the labels of the nearest historical examples.

Hoeffding option Tree (P.Chaudhuri,A.K.Ghosh and H.Oja,2009) allows each training example to update a set of option nodes rather than just a single leaf. Like standard decision tree nodes it can split the decision paths into several subtrees.It makes a decision with an option tree by combining the predictions of all applicable leaves into a single result.

b) Ensemble Classifier

A substantial amount of work has also focused on Ensemble Classifier such as Fast And Light Classifier, OzaBag, OzaBoost, OzaBagADWIN, OzaASHT. There are two major types of Ensemble Classifier that are Bagging and Boosting. Bagging (L. Breiman, 2000) is sampling based approach has been proposed by Breiman. It generates multiple base classifiers by training them randomly. Finally, the decision is taken according to majority voting, Boosting has been proposed by Schapiro. It creates the data subsets for base classifier training by resampling the training data. In Boosting weight is assigned to training instances that determines how well was the instance was classified.

OzaBag (Oza N, Russell S.,2001)online bagging has been proposed by Ooze and Russel. This is used for stream data classification. Online bagging is a good approximation to batch bagging.

OzaBoost(Oza N, Russell S.,2001) is online boosting algorithm. It generates a sequence of base models using weighted training sets and the correctly classified examples are given the remaining half of the weight.

OzaBagASHT(Bifet A, Gavald R ,2007)is new bagging method which use Adaptive Size Hoeffding Tree that sets the size for each tree. If the number of split nodes of the ASHT tree is higher than the maximum value, then it deletes some nodes to reduce its size. It is bagging using trees of different size.

OzaBagADWIN(BifetA, Gavald R,2007)lgorithm is to use a sliding window, not fixed a priori, whose size is recomputed in online according to the change rate observed from the data in window itself. The contents of the window can be used for the three tasks:

(i)To detect change

(ii)To obtain updated statistics from the recent examples

(iii) To have data to rebuild or revise the models after data has changed.

Fast and light classifier (K.Wankhade,2010)is a new ensemble method for classification of stream data. It uses adaptive windowing technique for change detection and estimation and it uses the boosting technique with hoeffding tree for building ensembles. It also deals better with concept a drift which is crucial problem of evolving data streams.

Random subspace (G. Fumera, F. Roli, and A. Serrau, 2008)is an ensemble creation method that uses feature subsets to create the different data subsets to train the base classifiers. It is used for both constructing and aggregating classifier. It solve the small sample problem.

Streaming Ensemble Algorithm (W .Nick Street, Yong Seog Kim,2001)processes the incoming stream in data chunks. The size of those chunks is an important parameter because it is responsible for the trade-off between accuracy and flexibility. Each data chunk is used to train a new classifier, which is later compared with ensemble members. If any ensemble member is "weaker" than the candidate classifier it is dropped and the new classifier takes its place.

Dynamic Weighted Majority (DWM) (J. Kolter and M. Maloof.,2007)is an ensemble approach which maintains a varying size set of online base learners. In this method, prediction is based on a weighted vote of base models and weights of those classifiers which predict wrongly are decreased. If the ensemble prediction is wrong then a new online classifier is added and those base models whose weights fall below a threshold are removed.

Accuracy Weighted Ensemble (Haixun Wang, Wei Fan, Philip S. Yu, Jiawei Han,2003) as proposed by Wang et al. In this a new classifier is trained ever time whenever there is incoming data chunk and use that chunk to evaluate all the existing ensemble members to select the best component classifiers.

Multi chunk partition multi ensemble method (M. H. Chehreghani, H. Abolhassani and M. H. Chehreghani, 2008) reduces error rate over single partition single chunk which uses simple majority voting. It keep optimally best k*v classifiers, where k is ensemble size and v is number of partitions. It uses labeled chunks to first train the classifiers.

c) Clustering Technique

There are many clustering techniques that were used such as Den Stream, r-Den Stream, Density Grid Based, and ClusStream.

In DenStream (HUANG Hai, LIU Li-xiong, 2009) clustering, dropped micro-clusters are stored on outside memory temporarily, and new chance is given to attend clustering to improve the clustering accuracy.

Density Grid Based (Amineh Amini, Teh Ying Wah, Mahmoud Reza Saybani, Saeed Reza, Aghabozorgi Sahaf Yazdi,2011) adopts a density decaying technique to capture the evolving data stream and extracts the boundary point of grid. It resolves the problem of evolving automatic clustering of real-time data streams.

Grid Based clustering (J. Han, 2005) The Grid-Based clustering method uses a multiresolution grid data structure. It forms the grid data structure by dividing the data space into a number of cells and perform the clustering on the grids. Clustering depends on the number of grid cells and independent of the number of data objects. Gird-based method could be natural choice for data stream in which the infinite data streams map to finite grid cells. The synopsis information for data streams is contained in the grid cells.

CluStream (C. C.Aggarwal,J. Han, J. Wang, and P. S. Yu ,2003) has been proposed by Aggarwal. It is used for dealing with the data having large dimension. In this method the clustering process is divided into two parts: online and offline. The online part clusters coming data divided by time window and store the results. The offline part generates the clustering results based on the observation.

HPStream (C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu,2004) is proposed by Aggarwal et al. for clustering of high dimensional data streams. It uses a Fading Cluster Structure (FCS) to stores the summary of streaming data and it gives more importance to recent data by fading the old data with time. For handling high dimensional data it selects the subset of dimensions by projecting on original high dimensional data stream

E-Stream(K.Udommanetanakit, T. Rakthanmanon, and K. Waiyamai, 2007)is a data stream clustering technique which supports following types of evolution in streaming data such as Appearance of new cluster, Disappearance of an old cluster, Split of a large cluster, merging of two similar clusters and change in the behaviour of cluster itself.

The Online Divisive-Agglomerative Clustering ODAC (Y. Chen and L. Tu,"Density-based clustering for real-time stream data,2007) clustering technique combines both divisive and agglomerative hierarchical clustering approaches to support the concept evolution. ODAC maintains a tree-like hierarchy of clusters, using a top-down strategy.

POD-Clus Probability and Distribution-based Clustering(P. P. Rodrigues, J. a. Gama, and J. Pedroso,,2008)is a model based clustering technique for streaming data. It is applicable to both clustering by example and clustering by variable scenarios.
For summarizing the cluster information and update it incrementally, it uses a cluster synopsis which comprises the mean, standard deviation, and number of points for each cluster.

COMET-CORE Clustering over Multiple Evolving streams by Correlations and Events (P. Chaovalit and A. Gangopadhyay, 2009)is proposed for clustering multiple data streams in online manner. It uses weighted correlation as similarity measurement.

SPE-Cluster clustering algorithm for multiple data streams (E. J. Keogh, S. Chu, D. Hart, and M. J. Pazzani, ,2001) It measures correlations between data streams using auto-regressive modelling technique. For this it finds frequency spectrum to extract relevant features from streaming data. Each stream is approximated by the spectral components, and the correlation is measured component-wise.

k-median algorithm for streaming data(S. Guha, N. Mishra, R. Motwani, and L. O'Callaghan,2000)The data is divided into chunks; each of them is clustered separately into weighted canters. Then, all the centers are clustered again to generate the final centers. COBWEB (D. Fisher, 2008)is an incremental clustering technique intends to discover understandable patterns in data. It uses a category function to create a tree. COBWEB keeps a hierarchical clustering model in the form of classification tree. Each node contains a probabilistic description of the concept that summarizes objects classified under that node.

STREAM (L. O'callaghan, N. Mishra, A. Meyerson, S. Guha, and R. Motwani, 2011)uses moving window and produces centers at each particular stage. This approach does not need to specify in advance the number of clusters expected at the output and instead evaluates the performance by using a combination of sum of squared distances and the number of centers used.

Divide and conquer technique (Edwin Lughofer,2002) is used to devise dynamic split and merge method which would be able to handle dynamic nature of data. Merging is performed using weighted averaging of cluster centers based on homogeneity condition between two clusters. Splitting is done when clusters are shrinking inside one large cluster to form separate clusters using Bayesian information criterion.

3. Discussion

Thus, several important algorithms for mining data streams were discussed such as Single Classifier, Ensemble Classifier and Clustering techniques they have different advantages and disadvantages.

Single Classifier like VFDT requires less memory but has various drawbacks such as whenever the size of the training set is small, the performance of this approach can be unsatisfactory. It cannot handle the concept drift.

CVFDT takes more space and the accuracy of this model is not greater than the best sliding window model. Random Forests has drawback such as the feature selection process is not explicit. Feature fusion is also less obvious on small size training data. It has weaker performance. Hoeffding Option Tree is time consuming.

Ensemble Classifiers are comparatively more accurate, easy and react better to concept drift than single classifiers. Classifier ensembles are rapidly gaining popularity in data mining community. Ensemble Classifier like OzaBag cannot handle gradual and sudden concept drift and it requires more memory space. Oza Boost cannot handle gradual and sudden concept drift. OzaBagASHT cannot handle abrupt concept drift, OzaBagADWIN cannot handle abrupt concept drift. Fast and light classifier is more accurate, in terms of time and memory in classifying both synthetic and real data sets. It is not able to accurately separate and measure training and testing time SEA is easy to implement and quickly adapts to changes in concept. Its minor disadvantage is that it requires multiple scans which slightly affect time complexity in case of vast data sets. Weighted ensemble can efficiently handle concept drift but requires more space and time. It has good Multi chunk partition multi ensemble method has good accuracy and can handle concept drift.

Clustering technique like Den Stream cannot distinguish clusters which have different levels of density and there is Loss of knowledge points. It is more accurate than Den Stream, rDestream needs more memory space, because it needs external disk to memorize historical outliers Density Grid Based has better scalability in processing large-scale and high dimensional stream data but It cannot find arbitrary shaped clusters with noise. CluStream fails to handle changing data, thus leads to lost of knowledge point HpStream cannot discover the cluster of arbitrary shapes and require domain knowledge for specifying the number of clusters and average number of projected dimensions parameters. EStream has better performance than HPStream algorithm but it requires many parameters to be specified by user. The Online Divisive-Agglomerative Clustering (ODAC) cannot Handle Data Fading and arbitrary Shape Clusters K means have various advantages like it is simple, results is easy to interpret, but it is sensitive to noise and requires more space. COMET-CORE cannot handle arbitrary shaped clusters. SPE-Cluster cannot handle arbitrary shaped clusters and fading data. COBWEB can manage outlier relatively well in this method but because of the tree structure it includes overhead for managing tree. STREAM has various drawbacks which include failure to handle dynamic data, requires more time and space. It has poor accuracy and it inherits all drawbacks of K means. Divide and Merge requires more space and processing time. It copes up with dynamic data to provide high purity and cleaner cluster. However it should be used in combination of some other technique to refine the results.

Conclusion

Thus, there are different techniques that can handle the data stream but none of the clustering technique can handle the concept drift as the datastream is huge and

changing continously it may be labeled or unabelled . thus, there is need of Hybrid Ensemble Classifier so that datastream can be handled more efficiently

References

Brijesh Verma and Ashfaqur Rahman(2012),Cluster-Oriented Ensemble Classifier: Impact of Multicluster Characterization on Ensemble Classifier Learning, IEEE Trans. on Knowledge And Data Engineering,pp.1156-1167.

P.Chaudhuri,A.K.Ghosh and H.Oja (2009),Classification Based on Hybridization of Parametric and Non-Parametric Classifiers,vol.31, IEEE Trans. Pattern Analysis and Machine Intelligence,pp.1153-1164.

Reza,O.Pujol,D.Masip(2009), Geometry Based Ensemble:Toward Structural Characterization of the Classification Boundary,IEEE Trans. Pattern Analysis and Machine Intelligence ,pp.1140-1146

Geoff Hulten, Laurie Spencer, Pedro Domingos(2001).Mining time changing data streams, ACM,pp.97-106.

Bieft A, Holmes G, Pfahringr B, Kirkby R,Gavalda R (2009),New ensemble methods for evolving data streams,, KDD,pp.139–148

Satpute, Pravin C., and Deepti P. Theng. (2012),A Study of Data Mining Techniques for WSN Based Intellectual Climate System

Bifet A, Gavald‘a R (2007),Learning from time changing data with adaptive windowing.,In: SIAM Int Conf Data Mining. pp.443–448

C. C.Aggarwal,J. Han, J. Wang, and P. S. Yu(2003),A framework for clustering evolving data streams, In Proc. Of VLDB,pp. 81-92

Feng Cao et al, Martin Ester,Weining Qian, and Aoying Zhou (2006),Density-based clustering over an evolving data stream with noise.,In SDM

L. Breiman, (2001),Random Forests,Machine Learning, vol.45, no. 1, pp.5-32.

HUANG Hai, LIU Li-xiong (2009).,rDenStream,A Clustering Algorithm over an Evolving Data Stream, The National High Technology Research and Development Program(“863” Program)

Oza N, Russell S (2001).Online bagging and boosting,In:Artificial intelligence and statistics, Morgan Kaufmann. ,pp.105–112

P. Domingos and G. Hulten(2000),Mining High-Speed Data Streams,In Proceedings of the Association for Computing Machinery Sixth International Conference on Knowledge Discovery and Data Mining.

Oza N, Russell S. (2001),Experimental comparisons of online and batch versions of bagging and boosting, pp. 359–364,In:ACM SIGKDD.

Amineh Amini, Teh Ying Wah, Mahmoud Reza Saybani, Saeed Reza, Aghabozorgi Sahaf Yazdi,(2011),A Study of Density-Grid based Clustering Algorithms on Data Streams, Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD).

T. Windeatt (2006) ,Accuracy/Diversity and Ensemble MLP ClassifierDesign, IEEE Trans. Neural Networks,pp. 1194-1211,

G.M. Munoz, D.H. Lobato, and A. Suarez(2009),An Analysis of Ensemble Pruning Techniques Based on Ordered Aggregation,IEEE Trans. Pattern Analysis and Machine Intelligence.,pp.245-259

L. Breiman (2000),Bagging Predictors,Machine Learning, pp.123-140.

G. Fumera, F. Roli, and A. Serrau(2008).“A Theoretical Analysis of Bagging as a Linear Combination of Classifiers,” IEEE Trans.Pattern Analysis and Machine Intelligence.pp.1293- 1299.

R.E.Schapire (2000),The Strength of Weak Learnability,Machine Learning, pp.197-227.

R.E.Banfield,L.o.Hall,K.W.Bowyer,W.P.Kegelmeyer, (2003).,A New Ensemble Diversity Measure Applied to Thinning Ensembles,In Proc. Fourth Int’l Workshop Multiple Classifier Systems (MCS ’03),pp.306-316

W .Nick Street, Yong Seog Kim (2001),A streaming ensemble algorithm (sea) for large-scale classification ,In Proceedings of the seventh ACM SIGKDD International Conference on Knowledge discovery and data mining, New York, NY, USA,pp.377-382

Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, and Bhavani Thuraisingham(2009),A Multi-partition Multi-chunk Ensemble Technique to Classify Concept-Drifting Data Streams, pp.363–375,Springer-Verlag, Berlin Heidelberg.

J. Kolter and M. Maloof (2007),Dynamic weighted majority: An ensemble method for drifting concepts", Journal of Machine Learning &Research,pp.2755-2790

Haixun Wang, Wei Fan, Philip S. Yu, Jiawei Han(2003) ,Mining Concept Drifting Data Streams Using EnsembleClassifiers, SIGKDD’03,ACM.pp.226-235,

M. H. Chehreghani, H. Abolhassani and M. H. Chehreghani (2008.),Improving density-based methods for hierarchical clustering of web pages,Data & Knowledge Engineering, pp. 30-50

J. Han (2005),Data Mining: Concepts and Techniques. San Francisco, USA: Morgan Kaufmann Publishers Inc.

C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu(2004) ,A framework for projected clustering of high dimensional data streams,In Proceedings of the Thirtieth international conference on Very large data bases –Volume 30, ser. VLDB ’04. VLDB Endowment, pp. 852–863

K.Udommanetanakit, T. Rakthanmanon, and K. Waiyamai (2007),E-stream: Evolution-based technique for stream clustering,In Proceedings of the 3rd international conference on Advanced Data Mining and Applications, ,Berlin, Heidelberg: Springer-Verlag,pp.605–615

Y. Chen and L. Tu (2007),Density-based clustering for real-time stream data, In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, ser. KDD ’07, New York, USA.pp. 133–142.

P. P. Rodrigues, J. a. Gama, and J. Pedroso,(2008) ,Hierarchical clustering of time-series data streams, IEEE Trans. on Knowl. and Data Eng.,pp. 615–62

P. Chaovalit and A. Gangopadhyay (2009),A Method for clustering transient data streams,In Proceedings of the ACM symposium on Applied Computing, ser. SAC ’09. New York, NY, USA: ACM ,pp.1518-1587

E. J. Keogh, S. Chu, D. Hart, and M. J. Pazzani (2001),An online algorithm for segmenting time series,In Proceedings of the IEEE International Conference on Data Mining, ser. ICDM ’01. Washington, DC, USA: IEEE Computer Society. ,pp.289–296

S. Guha, N. Mishra, R. Motwani, and L. O’Callaghan,(2000),Clustering data streams,Foundations of Computer Science, Annual IEEE Symposium. ,pp 359-3

D. Fisher (2008),Iterative optimization and simplification of hierarchical clustering", IEEE Trans. on Knowl. and Data Eng., pp.615–62

J.S R Jang ,ANFIS Adaptive Network Based Fuzzy Inference System,Vol. 23,IEEE Transaction on Systems Man and Cybernetics. ,pp 665-6

G. Qiang (2000),An Effective Algorithm for Improving the Performance of Naive Bayes for Text Classification,pp.699-701.

T. Darrell and P. Indyk and G. Shakhnarovich (2006), Nearest Neighbor Methods in Learning and Vision: Theory and Practice. MIT Press

Edwin Lughofer (2011),Dynamic Evolving Cluster Models using On-line Split and-Merge Operations,In proceedings of 10th International Conference on Machine Learning and Applications,pp.20-26

L. O’callaghan, N. Mishra, A. Meyerson, S. Guha, and R. Motwani (2002),Streaming-data algorithms for high-quality clustering,In Proceedings of the 18th International Conference on Data Engineering (ICDE.02), pp.685–694.

Satpute, P.C.; Theng, D.P.,(2013),Intellectual Climate System for Monitoring Industrial Environment,Third International Conference on Advanced Computing and Communication Technologies (ACCT),pp. 36,39, 6-7.