Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 5, May 2015, pg.945 – 954

RESEARCH ARTICLE

Analytical Study of Data Mining on Data Stream Using Skewed Distribution of Data

Hulash Chandra Barahen[1], Dr. S.M.Ghosh[2]

[1]M.Tech (SE) Scholar, CSE Department, RCET, CSVTU, Bhilai, Chhattisgarh, India

[2]Associate Professor, CSE Department, RCET, CSVTU, Bhilai, Chhattisgarh, India

[1] hulash.chandra@rediffmail.com; [2] samghosh06@rediffmail.com

Abstract— Analytical Study of Data Mining is an important data analysis tool that uses a model from historical data to predict class labels for new observations. More and more applications are featuring data streams, rather than finite stored data sets, which are a challenge for traditional Distribution algorithms. Concept drifts and skewed distributions, two common properties of data stream applications, make the task of learning in streams difficult. The authors aim to develop a new approach to classify skewed data streams that uses an ensemble of models to match the distribution over under-samples of negatives and repeated samples of positives. We study the emerging area of algorithms for processing data streams and associated applications, as an applied algorithms research agenda.

Keywords— Data stream analysis, data mining, Classification Skewed Distributions of data, K Nearest Neighbours Algorithm Data set, Clustering Algorithms.

I. INTRODUCTION

Analytical Study any real applications, such as network traffic monitoring, credit-card fraud detection, and click streams, generate continuously arriving data known as data streams. The knowledge discovery from stream data is challenging the data are usually and arrive with high speed, making either storing all the historical data or scanning it nearly impossible. Moreover, stream data often evolve considerably over time. In many applications, the response time usually must be short. Study can help people making decisions by labels for given data on the basis of past records. Researchers have analytical studied on stream data extensively in recent years, developing many interesting algorithms. We support our theoretical results with an experimental study over a large variety of real and synthetic data. We show that significant skew is present in both textual and telecommunication data. Our methods give strong accuracy, significantly better than other methods, and behave exactly in line with their analytic bounds. However, most studies on stream don't address skewed distributions, which are common in data stream applications .Here we propose an effective and efficient algorithm to data streams with skewed class distributions, employing both sampling and ensemble techniques.

II. MOTIVATION

The term fraud here refers to the profit organization system without necessarily to direct legal consequences. In a competitive environment, fraud can become a business critical problem and if the prevention procedures are not fail-safe. Fraud detection, being part of the overall fraud control, automates and helps reduce the manual parts of a screening/checking process. This area has become one of the most established industry/government data mining applications. It is impossible to be absolutely certain about the legitimacy of and intention behind an application or transaction. Given the reality, the best cost effective option is to tease out possible evidences of fraud from the available data using mathematical algorithms. Evolved from numerous research communities, especially those from developed countries, the analytical engine within these solutions and software are driven by artificial parallel computing, econometrics, expert systems, genetic algorithms, machine learning, neural networks, pattern recognition, statistics and others. There are plenty of specialized fraud detection solutions and software1 which protect businesses such as credit card, e-commerce, insurance, retail, telecommunications industries.

III. DATA STREAMS ORIGINATION TECHNIQUES

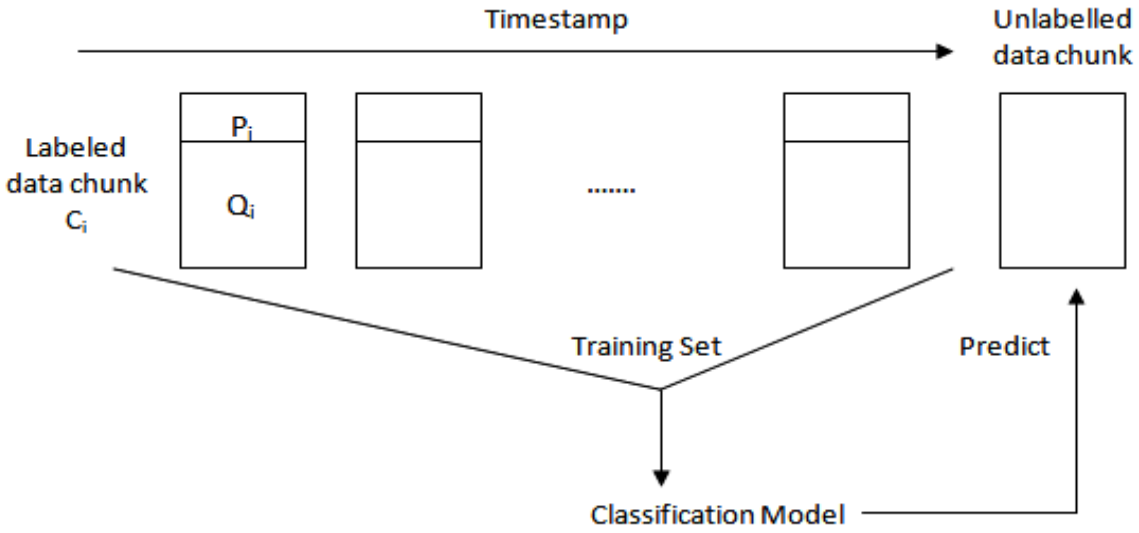

Fig 1 Data Streams Techniques

The data streams applied to the problem of credit-card fraud detection. As data chunks *C*1, *C*2… *Ci* arrives one by one, and each chunk contains positive instances *Pi* and negative instances *Qi*. Suppose *C*1, *C*2… *Cm* are labeled1, when an unlabelled data chunk *Cm*+1 arrives, the labels of instances in *Cm*+1 on the basis of previously labelled data chunks. When experts give the true class labels of instances in *Cm*+1, the chunk can join the training set, resulting in more and more labelled data chunks. Because of the storage constraints, it's critical to wisely select labelled examples that can represent the current distribution well. Therefore the system can add the transactions that received their labels into the set of labelled data chunks to update the model.

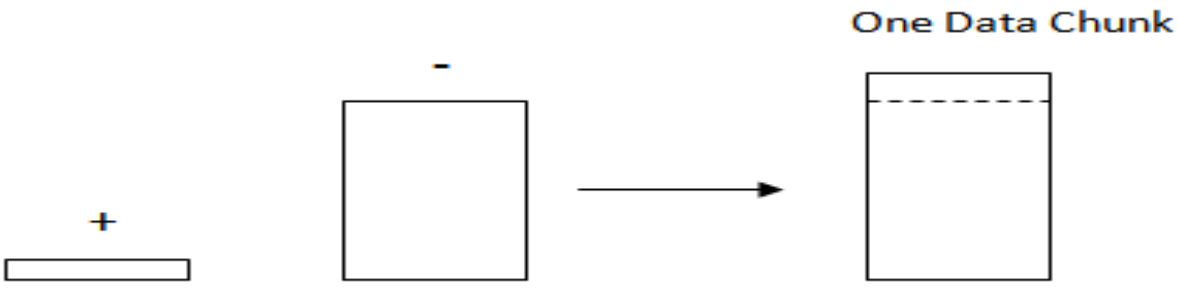

Fig 2 Skewed Distributions, each data chunk has fewer positive examples than negative examples

IV. STREAM ENSEMBLE WORK

We propose a simple strategy that can effectively classify data streams with skewed distributions. In stream applications, the incoming data stream arrives in sequential chunks, *C*1, *C*2… *Cm*, where *Cm* is the most up-to-date chunk. The data chunk that arrives next is *Cm*+1, and, for simplicity, we denote it as *T*. The aim is to train a classifier on the basis of the data that has already arrived to predict how likely each example in *T* falls into one of the predefined categories. We further assume that data come from two classes (positive and negative) and that the number of examples in the negative class is much greater than the number in the positive class. We need only estimate the probability of an example **x** belonging to the positive class, $P(+|\mathbf{x})$; that of the negative class would then be $1 - P(+|\mathbf{x})$. For accurate probability estimation, we propose using both sampling and ensemble techniques in the work.

V. SAMPLING

We split each chunk *C* into two parts: *P*, which contains positive examples in *C*, and *Q*, which contains negative examples in *C*. The size of *P* is much smaller than that of *Q*. Also, note that some uncertainties are associated with classification in certain problems. An example **x** could appear in both the positive and negative sets several times. Then, the count of **x** in each class will contribute to the calculation of the probability $P(+|\mathbf{x})$. In stream classification, we can't use all the observed data chunks as the training data. The stream data are huge, so it's usually impossible to store them all. Moreover, a huge training set will slow classification. Also, the concepts that stream data carry are usually changing, thus a model built on the data sets with out-of date concepts can't make accurate predictions for the new data. A model trained on the most recent data chunk reduces the classification error. However, in the skewed stream-classification problem, the positive examples in the most recent data chunk are far from sufficient to train an accurate model. Therefore, such a classification model will perform poorly on the positive class. To enhance the set of positive examples, we propose collecting all positive examples and keeping them in the training set. Specifically, the positive examples in the training set are {*P*1, *P*2… *Pm*}. Conversely, we randomly under-sample the negative examples from the last data chunk Qm *to* balance the class distribution. we form a new data set *Ts* that contains all the positives and sampled negatives.

VI. SAMPLING COMPARISON

The fading sampling method samples more examples in the more recent data chunks, whereas the fixed sampling method ignores the temporal information. We applied these sampling techniques on only positive examples; negative examples are sampled from the latest data chunk. In the two sampling schemes, the number of sampled examples remains the

same. We compared the performance of these two methods with the optimal performance in which the system doesn't do any sampling and keeps all the positive examples for training. In reality, this doesn't work because we use sampling when not all the examples can be stored, and keeping all of them violates the storage requirements. However, in this simulation study, we used this method as a benchmark to test the performance of the two sampling schemes (Fade and Fix versus All).

VII. ENSEMBLING

Instead of training a single model on this selected training set *Ts*, we propose generating multiple samples from the training set and computing multiple models from these samples. Suppose we generate *k* data sets from *Ts*.

Each negative example in *Ts* randomly propagates to exactly one of the *k* sets; hence the negative examples in the *k* data sets are completely disjointed. As for positive examples, they propagate to all the *k* sets because positives are rare and must be used to balance the distribution. Then, we train a series of classifiers on the *k* data sets, each of which outputs $f_i(\mathbf{x})$, which approximates $P(+|\mathbf{x})$. We use simple averaging to combine the outputs from *k* models:

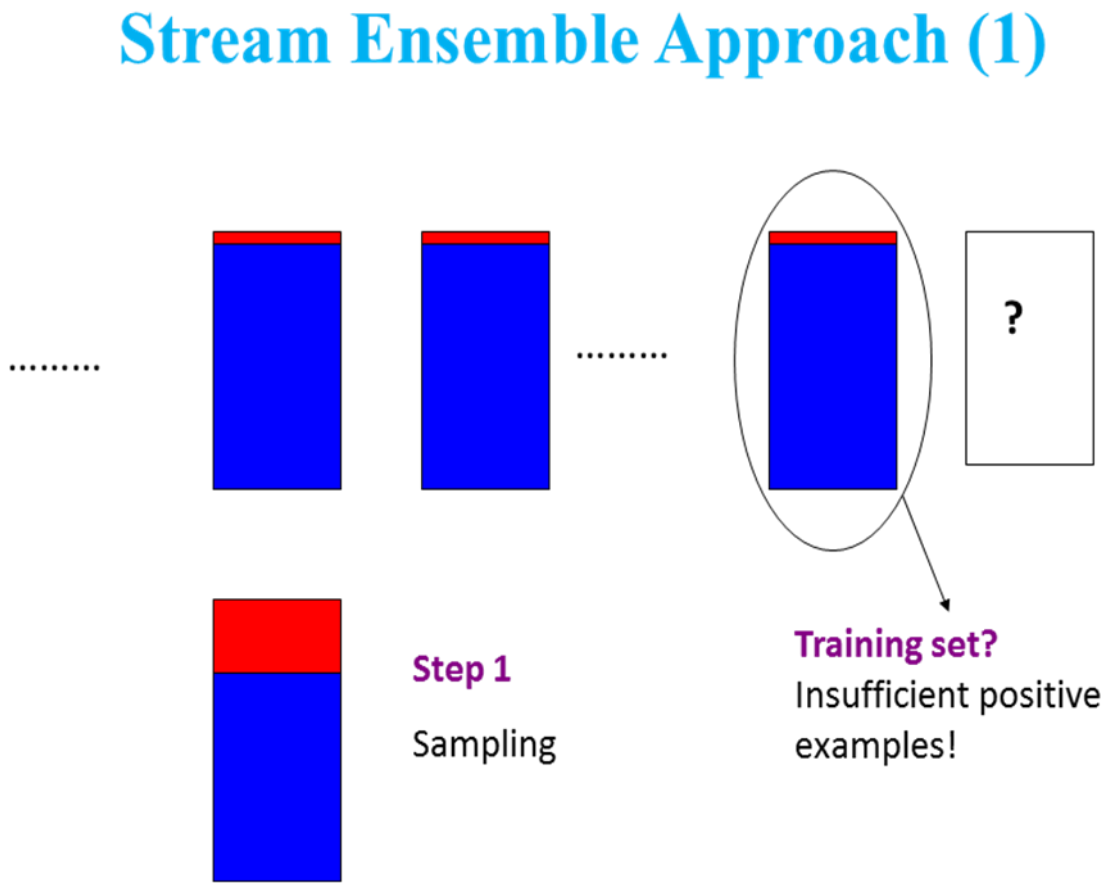

Fig 3 Stream Ensemble Approach(1)

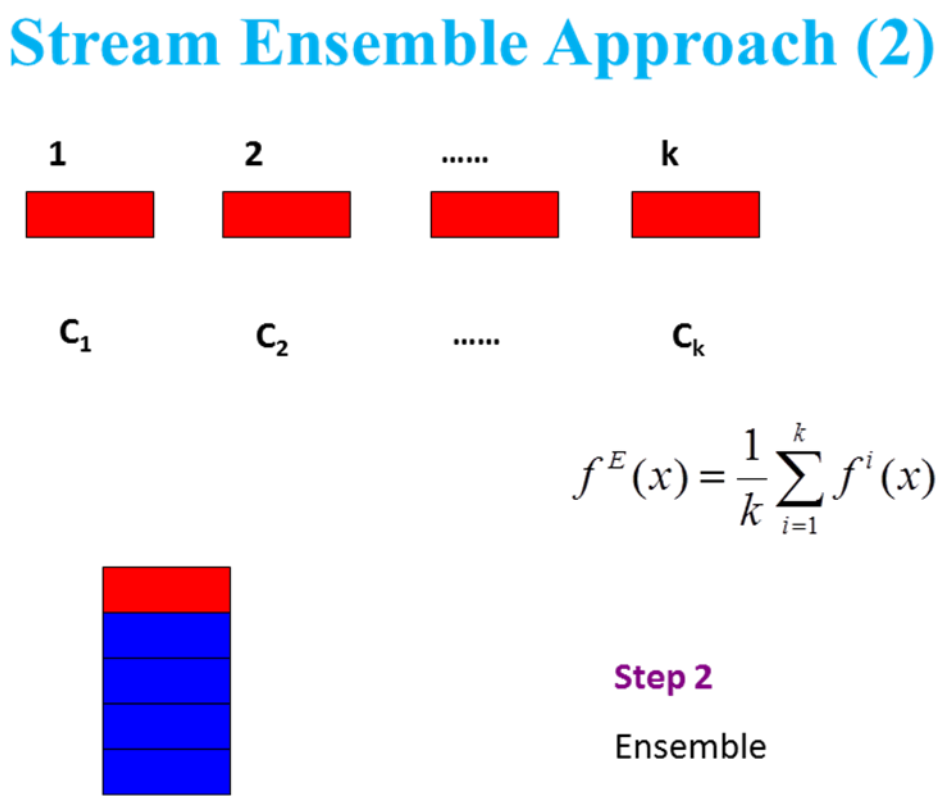

Fig 4 Stream Ensemble Approach(2)

TABLE I
DATA SET MANAGEMENT

Data set	Two classes	Number of instances	Number of features	Number of rare class instances	Number of chunks	Number of examples in a chunk
Thyroid1	Class1 vs. Class3	7000	21	66	6	2000
Thyroid2	Class2 vs. Class3	9000	23	88	8	3000
Opt	Each class vs. rest	6000	64	500	6	1000
Letter	Each class vs. rest	20000	16	200	6	1500
Covtype	Class2 vs. Class 4	28000	54	350	11	2500

Analysis

- Error Reduction
  - Sampling

$$f_c(x) = P(c \mid x) + \beta_c + \eta_c(x)$$

$$\sigma_b^2 = (\sigma_{\eta+}^2 + \sigma_{\eta-}^2)/s^2$$

  - Ensemble

$$f_C^E(x) = P(c \mid x) + \overline{\beta}_c + \overline{\eta}_c(x)$$

$$\sigma_{b^E}^2 = \frac{1}{k^2}\sum_{i=1}^{k}\sigma_{b^i}^2$$

- Efficiency Analysis
  - Single model

$$O(d(n_p + kn_q)\log(n_p + kn_q))$$

  - Ensemble

$$O(dk(n_p + n_q)\log(n_p + n_q))$$

  - Ensemble is more efficient

Fig 5 Analysis

Experimental Results

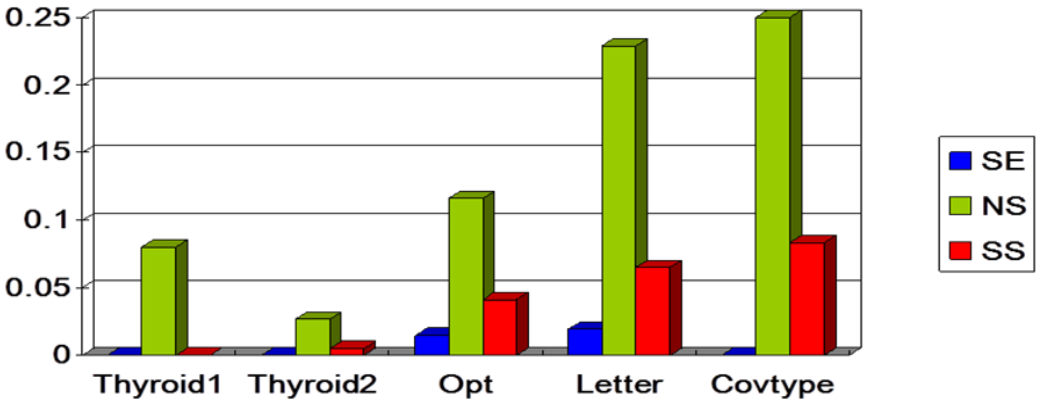

Fig 6 Mean Squared Error on Real Data

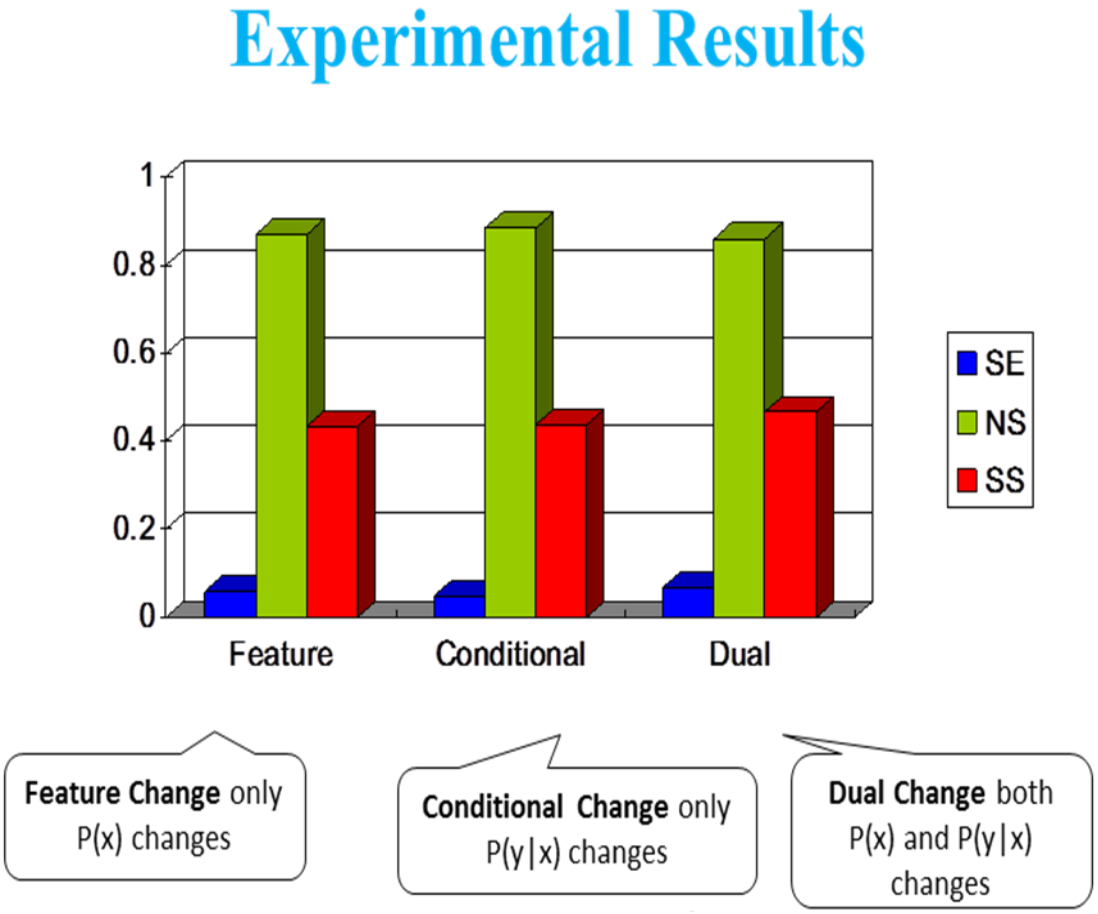

Fig 7 Mean Squared Error on Synthetic Data

Given that sampling and ensemble techniques could help reduce the classification error, the baseline methods we compared are the following:

A. No Sampling + Single Model (NS):

This method trains with only the current data chunk which is highly skewed. The method trains a single model on the training set.

B. Sampling + Single Model (SS):

The training set is the same as that used in the proposed ensemble method and is generated by keeping all positive examples seen so far and under-sampling negative examples in the current data chunk. But the method trains only a single model on this training set.

C. Sampling + Ensemble (SE):

This set denotes our proposed model, which adopts both sampling and ensemble techniques. In the experiments, the base learners include both parametric and nonparametric classifiers: decision tree, naïve, and logistic regression. We use the implementation in the package, a publicly available data mining software package.

VIII. DATA COLLECTION

Our experiments involved synthetically generated data sets as well as a series of real data sets.

IX. SYNTHETIC DATA

The generated synthetic data streams with different kinds of concept changes. The data is of the form (**x**, *y*), where **x** is a multidimensional feature vector, and $y \in \{0, 1\}$ is the class label. We then simulated $P(\mathbf{x})$, $P(y|\mathbf{x})$, and their changes.

X. REAL DATA

Although these data sets don't correspond directly to skewed data mining problems, we can convert them into rare class problems by taking one small class as the rare class and the remaining records, or the biggest remaining class, as the majority class. For the "thyroid" data set, we selected either Class 1 or Class 2 as the rare class and Class 3 as the majority class. Similarly, for the "covtype" data set, the biggest class (Class 2) is the majority class and the smallest class (Class 4) is the rare class. Data sets "letter" and "opt" are used to recognize 26 letters and 10 digits, respectively. For these two data sets, we simply chose one class to represent the rare class and collapsed the remaining classes into one majority class. So, from the original data sets, we generated skewed data sets and averaged the results over these data sets. To simulate a data stream, we randomly partitioned the data into several chunks with skewed distribution maintained in each chunk. Table 1 summarizes the data sets.

XI. CONCLUSION

In this paper we have come up with an approach to deal with skewed data streams using oversampling and the k nearest neighbour approach. We have illustrated our algorithm using various real world as well as the synthetic datasets with various features and the imbalance levels. Results obtained indicate that our approach shows deals well with skewed data streams; In particular, our approach has shown comparable and in some cases slightly better performance. As seen earlier various real life data stream applications like numerical fraud detection, network intrusion detection are characterized by the skewed data streams and in such cases this approach would help identify and classify minority class instances appropriately.

XII. FUTURE WORK

This throws light on the future enhancements that can be carried out. Some of the further enhancements would be to implement the approach for parallel computing platform which would help reduce the time required for the approach. There are various parts of the approach where in parallelism can be introduced. We are also working on an approach to combine two different stream classification approaches so as to get best out of the two algorithms. By combining our approach with another approach which handles general data streams or data streams with balanced distribution, we would like to extend the scope of our approach further such that it's applicable in general to data streams with any kind of distribution.

REFERENCES

[1] Dr. Lotos Malik, Mr.Rushi Langadge, "Class Imbalance Problem in Data Mining", Review – IJCSN- 2013.

[2] Jiawei Han, Jiayong, "Classification of Data Stream with Skewed Distribution", IEEE 2012.

[3] "Different Algorithm for Data Stream", IEEE 2011.

[4] Haibo He Shcne Chen, King Li, "Incremental Leering from Stream Data", IEEE-2011.

[5] Juan Zhang, Xuegang Hu,Yuhong Zhang, and Pei-Pei Li., "Ancient ensemble method for classifying skewed data streams", In De-Shuang Huang, Yon Gan, Prashan Premaratne, and Kyungsook Han, editors, ICIC (3), volume -6 Lecture Notes in Computer Science, pages 144{151. Springer, 2011}.

[6] Yanling Li,,Guoshe Sun, and Yehang Zhu, "Data Imbalance, imbalance problem in text", In Proceedings of the 2010 Third International Symposium on. Information Processing, ISIP '10, pages 301,305, Washington, DC, USA 2010. IEEE Computer Society.

[7] Yong Wang Lijun Cai Longo Zhango, "Classification skewed Data stream Based on reusing", International Conference 2010.

[8] Sheng Chen, Haibo He, Kang Li, and S. Desai. Musera: "Multiple selectively recursive approach towards imbalanced stream data mining, In Neural Networks", (IJCNN), The 2010 International Joint Conference on, pages 1, 8, july2010.

[9] Sheng Chen and Haibo He. Sera: "Selectively recursive approach towards nonstationary imbalanced stream data mining, In Neural Networks", 2009 IJCNN 2009. International Joint Conference on, pages 522{529, June 2009}.

[10] B. Babcock et al., "Models and Issues in Data Stream Systems," *Proc. Symp. Principles of Database Systems* (PODS 02), ACM Press, 2002, pp. 1–16.

[11] *Aggarwal, Data Streams: Models and Algorithms, Springer, 2007.*

[12] M.M. Gaber, A. Zaslavsky, and S. Krishnaswamy, "Mining Data Streams: A Review," *ACM SIGMOD* Record, vol. 34, no. 2, 2005, pp. 18–26.

[13] S. Muthukrishnan, "Data Streams: Algorithms an Applications", *Proc. ACM/SIAM Symp. Discrete Algorithm* (SODA 03), Soc. for Industrial and Applied Mathematics, 2003, p. 413.

[14] K.Tumer and J. Ghosh, "Analysis of Decision Boundaries in Linearly Combined Neural Classifiers," *Pattern Recognition*, vol. 29, no. 2, 1996, pp. 341–348.

[15] P. Domingos, "A Unified Bias-Variance Decomposition and Its Applications" , *Proc. Int'l Conf. Machine Learning* (ICML 00), Morgan Kaufmann, 2000, pp. 231–238.

[16] H. Wang et al., "Mining Concept-Drifting Data Streams Using Ensemble Classifiers", *Proc. Conf. Knowledge Discovery in Data* (KDD 03), ACM Press, 2003.

[17] I.H. Witten and E. Frank, *Data Mining: Practical Learning Tools and Techniques*, 2nd ed., Morgan Kaufmann, 2005.

[18] Sheng Chen and Haibo He. Sera: "Selectively recursive approach towards non stationary imbalanced stream data mining, In Neural Networks", 2009. IJCNN 2009. International Joint Conference on, pages 522{529, June 2009}.

[19] Sheng Chen, Haibo He, Kang Li, and S. Desai. Musera: "Multiple selectively recursive approach towards imbalanced stream data mining, In Neural Networks", (IJCNN), The 2010 International Joint Conference on, pages 1{8, July 2010}.

[20] Peng Liu, Yong Wang, Lijun Cai, and Longbo Zhang, "Classifying skewed data streams based on reusing data", In Computer Application and System Modelling (ICCASM), 2010 International Conference on, volume 4, pages V4 90 V4 93, Oct. 2010.

[21] Qun Song, Jun Zhang, and Qian Chi, "Assistant detection of skewed data streams in cloud security", In Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference on, volume 1, pages 60{64, Oct. 2010.}

[22] Juan Zhang, Xuegang Hu,Yuhong Zhang, and Pei-Pei Li, "An ancient ensemble method for classifying skewed data streams", In De-Shuang Huang, Yong Gan, Prashan Premaratne, and Kyungsook Han, editors, ICIC (3), volume 6840 of Lecture Notes in Computer Science, pages 144{151. Springer, 2011}.

