Imbalanced Data Classification Using oversampling and folding Technique.

[1]Er. Hemlata,[2]Dr. Dinesh Kumar
Department of Computer Engineering
Guru Kashi University
Talwandi Sabo Bathinda,Punjab India.

Department of Computer Engineering
Guru Kashi University
Talwandi Sabo Bathinda,Punjab India.

Abstract: In current era there are large number of applications which are producing large amount of multimedia data. The collected data can be processed to generated a well represented format. So that analyzed data can be used in various types of decision making purposes. In current time various data mining techniques are in existence where useful data can be extracted from the existing repository. The collected data can prevail with various discrepancies. Like imbalanced classification. The data once will be segregated or sub divided into multiple classes each class has imbalance of the data stream. Like Airlines data of on time and late flights. The data in late class is substantially lower than the on time flights. This misclassification can lead towards the poor analysis. For proper analysis the balancing of the data classification is required most. Over sampling technique is the best technique using which the imbalanced class be balanced using over sampling technique. All the parameters using AUC and G-mean will put best results once data is balanced using oversampling.

Keyword: Oversampling, Folding, Misclassification.

I. INTRODUCTION

With the internet age the data and information explosion have resulted in the huge amount of data. Fortunately to gather knowledge from such abundant data there exist data mining techniques. As per the definition by G. Ditzler in his book Data Mining: Concepts and Techniques [1] the data mining is - Extraction of interesting, non trivial, implicit, previously unknown and potentially useful patterns or knowledge from huge amount of data. Data mining has been used in various areas like Health care, business intelligence, financial trade analysis, network intrusion detection etc.

General process of knowledge discovery from data involves data cleaning, data integration, data selection, data mining, pattern evaluation and knowledge presentation. Data cleaning, data integration constitute data preprocessing. Here data is processed so that it becomes appropriate for the data mining process. Data mining forms the core part of the knowledge discovery process. There exist various data mining techniques viz. Classification , Clustering, Association rule mining etc. Our work mainly falls under the classification data mining technique.

Classification is one of the important technique of data mining. It involves use of the model built by learning from the historical data to make prediction about the class label of the new data/observations. Formally, it is task of learning a target function f, that maps each attribute set x to a set of predefined class labels y. Classification model learned from historical data is nothing but the target function. It can serve as a tool to distinguish between the objects of different classes as well as to predict class label of unknown records. Fig 1.1 shows the classification
task which maps attribute set x to its class label y.

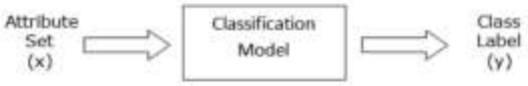

Fig. 1 Classification as a task of mapping input attribute set x into its class label y

Classification is a pervasive problem that encompasses many diverse applications, right from static datasets to data streams. Classification tasks have been employed on static data over the years. In last decade more and more applications featuring data streams have been evolving which are a challenge to traditional classification algorithms.

II. LITERATURE SURVEY

Overview of Methods for Dealing with Skewed Data Streams -Traditional Approaches We went through various methods available in the literature to deal with imbalanced datasets and portray some of the well known and most popular approaches, algorithms and methods that have been devised to deal with skewed data streams. Some of the books that we have referred to get an effective understanding of data mining concepts are Data Mining:

Introduction to Data Mining In the literature there are number of methods addressing class imbalance problem but the area of skewed data streams is relatively new to the research community. The sampling based and ensemble algorithms are the simplest yet the effective ones. Following paragraphs will provide the brief overview of the same. Some of the approaches for dealing with skewed data streams are categorized under following methods.

- Oversampling.
- Under-sampling.
- Cost Sensitive Learning.

Oversampling and under-sampling are sampling based preprocessing methods of data mining. The main idea in these methods is to manipulate the data distributions such that all the classes are represented well in the training or learning datasets. Recent studies in this domain have shown that sampling is effective method to deal with such kind of problems. Cost sensitive learning is basically associates cost of misclassifying the examples to penalize the classifier. **Oversampling:** Oversampling is one of the sampling based preprocessing technique in data mining. In oversampling the number of minority class instances in increased by either reusing the instances from the previous training/learning chunks or by creating the synthetic examples. Oversampling tries to strike the balance between ratio of majority and minority. classes. One of the advantage of this method is that using this normal stream classification methods can be used. The most commonly used method of oversampling is SMOTE(Synthetic Minority Oversampling Technique)[7]. Some of the Oversampling based approaches in the literature are discussed below. Most of the stream classification algorithms available assume that the streams have balanced distribution of classes. In the last few years few attempts have been made to address the problem to deal with skewed data streams.

SERA(Selectively Recursive Approach) framework was proposed by **Chen and He [9]** in this framework they selectively absorbed minority examples from previous chunks into current training chunk to balance it. Similarity measure used to select minority examples from previous chunks was great distance. He [10] was their further work after SERA to deal with imbalanced data stream classification. In MuSeRA balancing of training chunk is done in the similar way by using large distance as similarity measure to accommodate minority samples accumulated from all the previous training chunks. In MuSeRA a hypothesis is built on every training chunk, thus a set of hypothesis is built over time as opposed to SERA which maintains only single hypothesis. Here set of hypothesis at time-stamp i will be used to predict the classes for instances in test chunk at time-stamp i. In their further work in similar area **Chen and He [8]** proposed an approach named REA(Recursive Ensemble Approach), in which when next training chunk arrives, it is balanced by adding those positive instances from previous chunks which are nearest neighbors of the positive instances in the current training chunk, then it is used to build a soft typed hypothesis. In REA for every training chunk a new soft typed hypothesis is built. It then uses weighted majority voting to predict the posterior probabilities of test instances, here the weights are assigned to different hypothesis based on their performance on current training chunk.

Under-sampling : Under-sampling is another sampling based method which solves the problem by reducing the number of majority class instances. This is generally done by altering out the majority class instances or by randomly selecting the appropriate number of majority class examples. under-sampling is mostly carried out using the clustering method. Using clustering the best representative from the majority class are chosen and the training chunk is balanced accordingly. Some of the under-sampling based approaches in the literature are discussed below.

Sheng Chen et al [9] proposed another algorithm to deal with skewed data streams. They used clustering sampling algorithm to deal with skewed data streams. Sampling was carried out by using k-means algorithm to form clusters of negative examples in the current training chunk and then they used the centroid of each of the clusters formed to represent each of those clusters. Number of clusters formed were equal to the number of positive examples in current training batch and thus current training batch was updated by taking all positive examples along with centroid of the clusters of negative samples.

III. FLOWCHART

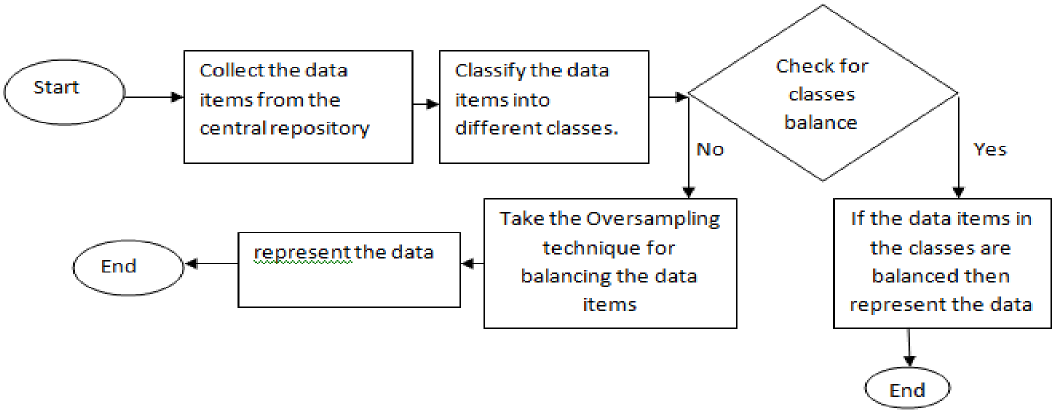
Start
Collect the data items from the central repository
Classify the data items into different classes.
Check for classes balance
No
Yes
Take the Oversampling technique for balancing the data items
If the data items in the classes are balanced then represent the data
represent the data
End
End

Fig. 2 Flowchart

IV. PSEUDO CODE

```
Step1 Collect the central repository of the Airlines data.
Step2 Perform the classification of the dataset items.
Step3 Perform the balancing of the dataset items.
Step4 Checks for class balance.
Step5 Represents the balanced classes.
```

V. ALGORITHM

Step1 Collect the data from the dataset repository. This dataset will be regarding the Airlines data.

Step2 Classify the data into multiple classes. One class can be of late flights and one class is of the on time flights.

Step3 Check for the classes size. If the class size has large differentials then goto step4 else goto step 6

Step4 Take oversampling of the dataset items. Put data in to the minority class for balancing the minority class.

Step5 Perform the representation of the balanced data classification for the analysis purpose.

Step6 End

VI. RESULTS AND DISCUSSIONS

6.1 Performance parameters

6.1.1 AUC(Area Under Curve).

That's the whole point of using AUC - it considers all possible thresholds. Various thresholds result in different true positive/false positive rates. As you decrease the threshold, you get more true positives, but also more false positives. The relation between them can be plotted:

6.1.2. G-Mean.

Analysis of data or data analytics, is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains

6.2 Dataset Of Airlines

CO,	269,	SFO,	IAH,	3,	15	,	205,	1
US,	1558,	PHX,	CLT,	3,	15	,	222,	1
AA,	2400,	LAX,	DFW,	3,	20	,	165,	1
AA,	2466,	SFO,	DFW,	3,	20	,	195,	1
AS,	108,	ANC,	SEA,	3,	30	,	202,	0
CO,	1094,	LAX,	IAH,	3,	30	,	181,	1
DL,	1768,	LAX,	MSP,	3,	30	,	220,	0
DL,	2722,	PHX,	DTW,	3,	30	,	228,	0
DL,	2606,	SFO,	MSP,	3,	35	,	216,	1
AA,	2538,	LAS,	ORD,	3,	40	,	200,	1
CO,	223,	ANC,	SEA,	3,	49	,	201,	1

DL,	1646,	PHX,	ATL,	3,	50	,	212,	1
DL,	2055,	SLC,	ATL,	3,	50	,	210,	0

6.3 G-Mean Comparison of Imbalanced data and balanced Data Stream

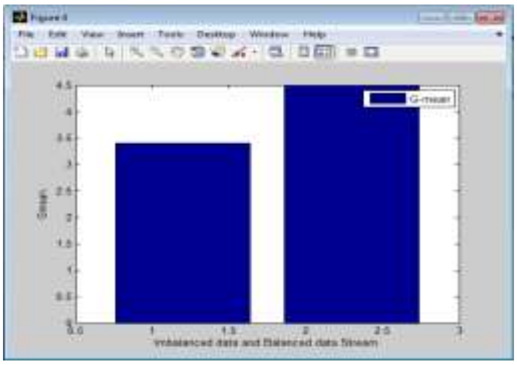

Fig. 3 G-Mean Comparison

This snapshot shows the G-mean Comparison of the Imbalanced and Balanced data Stream. The G-mean for the imbalanced data stream is lower compared to the balanced data Stream. That means G-Mean is improving for Balanced data Stream.

5.4 AUC(Area Under Curve) Comparison

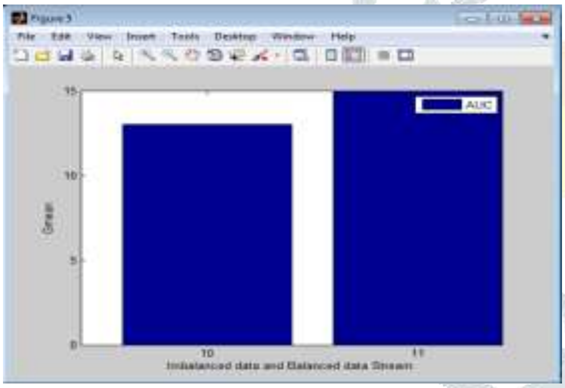

Fig. 4 AUC Comparison

This graph shows the AUC for both Imbalanced data Stream and balanced data Stream. The Area under Curve covers more area in Balanced data Stream compared to balanced data Stream. This will enhance the results. For Balanced data Stream.

VII. CONCLUSION

Data is the important component for any organization decision making purposes. Various applications are producing the multimedia data in millions of bytes. For better analysis of the data there requires better data mining techniques. These techniques will extract the relevant data from the large repository. But while analysis the datasets there can be misclassification of the data items. One class can have large data compared to the other class. Like in current research the late flights has substantially lower amount of data compared to on-time flights data. It in results leads to the poor analysis. The oversampling technique is the best technique for balance the minority class. Both classes then will be having balanced classes. All the performance factors like G-mean and AUC(Area under Curve) are giving better results compared to imbalanced classes.

VIII. FUTURE WORK

Current research is based on balancing of the imbalanced classes. The imbalance is to the data size of the classes comes after the classification of the dataset items. In future Fuzzy based technique can be used for balancing the classes for better performance in terms of processing time.

IX. REFERENCES

[1] Charu C. Aggarwal, Jiawei Han, Jianyong Wang, and Philip S. Yu. “On demand classification of data streams”. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '04, pages 503{508, New York, NY, USA, 2004. ACM.

[2] Vahida Attar, Pradeep Sinha, and Kapil Wankhade. A fast and light classifier for data streams. Evolving Systems, 1:199{207, 2010.

[3] Brian Babcock, Shivnath Babu, Mayur Datar, Rajeev Motwani, and Jennifer Widom. Models and issues in data stream systems. In Proceedings of the twenty-rst ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, PODS '02, pages 116, New York, NY, USA, 2002. ACM.

[4] Stephen Bay, Krishna Kumaraswamy, Markus G. Anderle, Rohit Kumar, and David M. Steier. Large scale detection of irregularities in accounting data. In Proceedings of the Sixth International Conference on Data Mining, ICDM '06, pages 75{86, Washington, DC, USA, 2006. IEEE Computer Society.

[5] Nadeem Qazi,Kamran Raza, “Effect Of Feature Selection, Synthetic Minority Over-sampling (SMOTE) And Undersampling On Class imbalance Classification”, 14th International Conference on Modeling and Simulation-2012.

[6]. Nitesh V. Chawla, Nathalie Japkowicz, Aleksander Ko lcz “Special Issue on Learning from Imbalanced Data Sets” Volume 6, Issue 1 - Page 1-6.

[7] S¸ eyda Ertekin1, Jian Huang, L´eon Bottou, C. Lee Giles “Active Learning in Imbalanced Data Classification”

[8] Saumil Hukerikar, Ashwin Tumma, Akshay Nikam, Vahida Attar “SkewBoost: An Algorithm for Classifying

Imbalanced Datasets" International Conference on Computer & Communication Technology (ICCCT)-2011.

[9] Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse, "Improving Learner Performance with Data Sampling and Boosting" 2008 20th IEEE International Conference on Tools with Artificial Intelligence.

[10] Benjamin X. Wang and Nathalie Japkowicz "Boosting Support Vector Machines for Imbalanced Data Sets" Proceedings of the 20th International Conference on Machine Learning-2009.

[11] Lei Zhu, Shaoning Pang, Gang Chen, and Abdolhossein Sarrafzadeh, "Class Imbalance Robust Incremental LPSVM for Data Streams Learning" WCCI 2012 IEEE World Congress on Computational Intelligence June, 10- 15,2012 - Australia.

[12] David P. Williams, Member, Vincent Myers, and Miranda Schatten Silvious, "Mine Classification With Imbalanced Data", IEEE Geosciences And Remote Sensing Letters, Vol. 6, No. 3, July 2009.

[13] Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse, Amri Napolitano "A Comparative Study of Data Sampling and Cost Sensitive Learning" , IEEE International Conference on Data Mining Workshops. 15-19 Dec. 2008.

[14] Mikel Galar,Fransico, "A review on Ensembles for the class Imbalance Problem: Bagging,Boosting and Hybrid-Based Approaches" IEEE Transactions On Systems, Man, And Cybernetics—Part C: Application And Reviews, Vol.42,No.4 July 2012

[15] Yuchun Tang, Yan-Qing Zhang, Nitesh V. Chawla, , and Sven Krasser "Correspondence SVMs Modeling for Highly Imbalanced Classification" IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, Vol. 39, No. 1, February 2009