Ianna Journal of Interdisciplinary Studies
ISSN:2735-9883 \ E-ISSN:2735-9891

VOL -06 NO 2, 2025

USING OVERSAMPLING TO SOLVE CLASS IMBALANCE PROBLEMS WITH LARGE DATASETS

Pertik Garg[1]Jarnail Singh[2,]

[1]Associate Professor, Department of Computer Science & Engineering, Swami Vivekanand Institute of Engineering & Technology, Banur, Punjab-140601

[2]Assistant Professor, Department of Computer Science & Engineering, Swami Vivekanand Institute of Engineering & Technology, Banur, Punjab-140601

Abstract:

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. Developing techniques for the machine learning of a classifier from class-imbalanced data presents an important challenge. 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.

Keywords: Oversampling, Imbalance, Data mining, AUC, G-mean

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 constitutes 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 techniques 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 shows the classification task which maps attribute set x to its class label y.

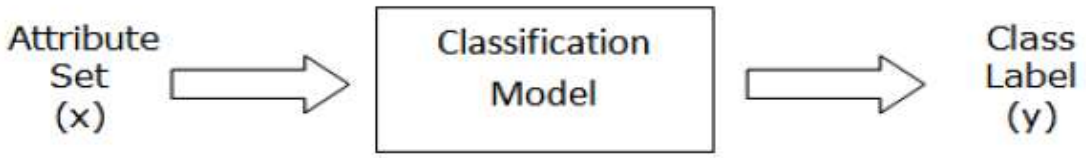

Fig. 1: Classification as a task of mapping input attribute set x into its class Label y

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 is challenge to traditional classification algorithms.

An Overview of Data Streams

Many real world applications, such as network traffic monitoring, credit card transactions, real time surveillance systems, electric power grids, remote sensors, web click streams etc, generate continuously arriving data known as data streams [2]. Unlike the traditional data sets, data streams arrive continuously at varying speeds. Data streams are fast changing, temporally ordered, potentially infinite and massive [3]. It may be impossible to store the entire data stream into memory or to go through it more than once due to its voluminous nature. Thus there is need of single scan, multidimensional, online stream analysis methods. In today's world with data explosion the data is increasing by terabytes and even peta bytes, stream data has rightly captured our data mining needs of today. Even though complete set of data can be collected and stored it’s quite expensive to go through such huge data multiple times.

Data Stream Classification

Since classification could help decision making by predicting class labels for given data based on past records, classification on stream data has been extensively studied in recent years with many interesting algorithms developed. Some of them are cited here: [2], fig 1.2 depicts the classification model in data streams. As shown in fig. 2 data chunks C1; C2;C3::::Ci arrive one by one.

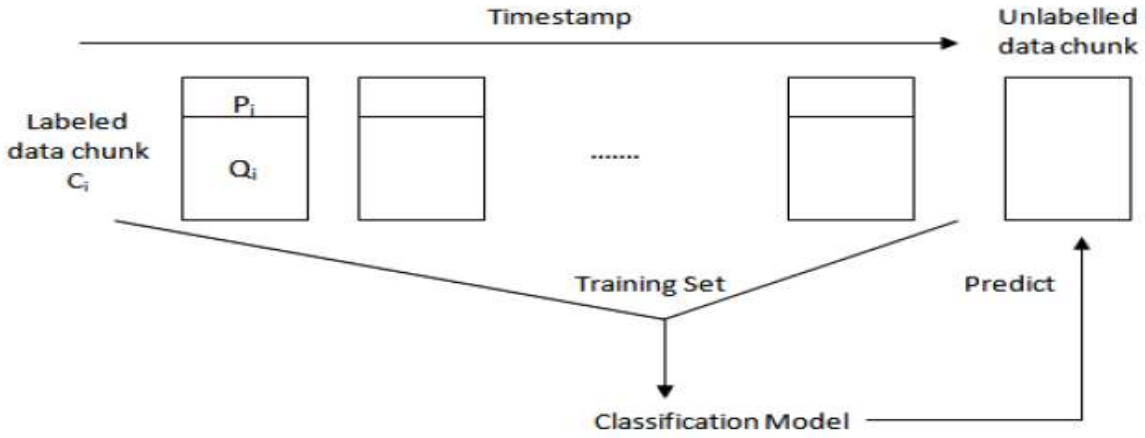

Fig. 2: Classification model in data streams

Each chunk contains positive instances Pi and negative instances Qi. Suppose C1;C2;C3::::Ci are labeled. At the time stamp m + 1, when an unlabelled chunkCm+1 arrives, the classification model predicts the labels of instances in Cm+1 on basis of previously labeled data chunks. When experts give true class labels of the instances in Cm+1, the chunk can join the training set, resulting in more and more labeled data chunks. Because of storage constraints, it is critical to judiciously select labeled examples that can represent the current distribution well. Most studies on stream mining assume relatively balanced and stable data streams. However, many applications can involve concept-drifting data streams with skewed distributions. In data with skewed distributions each data chunk has many fewer positive instances. Fig 3 shows the similar concept diagrammatically. At the same time, loss functions associated with positive and negative classes are also unbalanced. Misclassifying positive instances can have serious elects in some applications like stream of financial transactions.

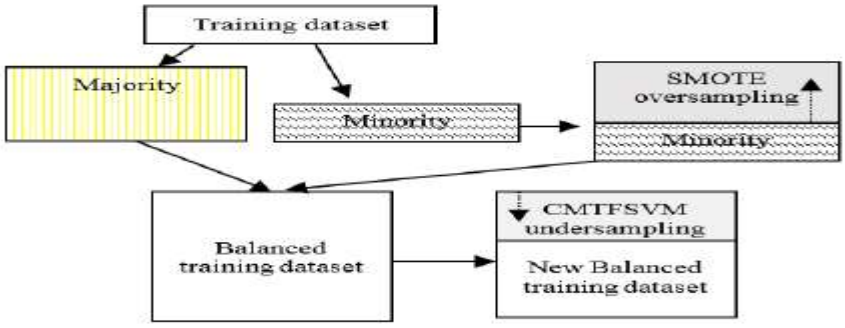

Fig. 3: Oversampling with SMOTE (Synthetic Minority Oversampling Technique) then under sampling using Folding method [4]

Over Sampling and Folding Technique

For balancing the imbalanced data stream over sampling and folding technique is the best technique. In this technique large amount of data sample will be collected for minority class. So that using classification criteria the over sample data will be reduced to applicable for minority class. This way the minority class will be converted to balanced class.

Literature Survey

Overview of Methods for Dealing with Skewed Data Streams

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.

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 techniques 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. The most commonly used method of oversampling is SMOTE (Synthetic Minority Oversampling Technique).

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 [1]** 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. 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 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 is chosen and the training chunk is balanced accordingly.

Sheng Chenet al (2017) [2] 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 was 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.

Victoria López (2017)[5] et al. discussed that training classifiers with datasets which suffer of imbalanced class distributions is an important problem in data mining. This issue occurs when the number of examples representing the class of interest is much lower than the ones of the other classes. Its presence in many real-world applications has brought along a growth of attention from researchers. We shortly review the many issues in machine learning and applications of this problem, by introducing the characteristics of the imbalanced dataset scenario in classification, presenting the specific metrics for evaluating performance in class imbalanced learning and enumerating the proposed solutions. In particular, we will describe preprocessing, cost sensitive learning and ensemble techniques, carrying out an experimental study to contrast these approaches in an intra and inter-family comparison.

Lee et al. (2002) [2] conferred temporary introduction is conferred on SVM and a number of other applications of SVM in pattern recognition issues. SVM are with success applied to variety of applications starting from face detection and recognition, object detection and recognition, written character and digit recognition, speaker and speech recognition, data and image retrieval,

prediction and etc as a result of they need yielded wonderful generalization performance on several applied math issues with none previous information and once the dimension of input house is extremely high however failed to compare the performance results for same application.

Lu et al.(2003) [6] conferred intimately our approach that uses SVM for classification and segmentation of an audio clip. The projected approach classifies audio clips into one in every of 5 classes: Pure speech, Music, setting sounds and silence. We've additionally projected a group of latest options to represent a 1 second sub clip, together with band regularity, LSP divergence form and spectrum flux. The experimental analysis have shown that the SVM technique yields high accuracy and with high process speed. We have a tendency to area unit extending this work to include visual data to assist video content analysis, the result's additionally terribly satisfying.

Results and Discussion

1. Performance Parameters

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:

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

2. Dataset of Airlines

Table 1: Dataset

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
AA,	2408,	LAX,	DFW,	3,	55	,	170,	0
AS,	132,	ANC,	PDX,	3,	55	,	215,	0
US,	498,	DEN,	CLT,	3,	55	,	179,	0
B6,	98,	DEN,	JFK,	3,	59	,	213,	0
CO,	1496,	LAS,	IAH,	3,	60	,	162,	0
DL,	1450,	LAS,	MSP,	3,	60	,	181,	0
CO,	507,	ONT,	IAH,	3,	75	,	167,	0
AS,	128,	FAI,	SEA,	3,	80	,	206,	0
DL,	2223,	ANC,	SLC,	3,	85	,	270,	0
AS,	112,	ANC,	SEA,	3,	90	,	200,	0
HA,	17,	LAS,	HNL,	3,	100	,	380,	1
US,	122,	ANC,	PHX,	3,	113	,	327,	1
AS,	114,	ANC,	SEA,	3,	150	,	200,	0
B6,	766,	BQN,	MCO,	3,	240	,	169,	0
B6,	768,	PSE,	MCO,	3,	262	,	180,	0
HA,	206,	HNL,	OGG,	3,	300	,	36,	1
OO,	4746,	BIS,	MSP,	3,	300	,	85,	0
OO,	6466,	IYK,	LAX,	3,	300	,	53,	0
US,	1011,	EWR,	CLT,	3,	300	,	111,	0
US,	1983,	BOS,	CLT,	3,	300	,	135,	0
HA,	108,	HNL,	KOA,	3,	302	,	41,	0
OO,	6900,	MKE,	ORD,	3,	302	,	44,	0
9E,	3975,	GFK,	MSP,	3,	305	,	68,	0
HA,	102,	HNL,	ITO,	3,	305	,	49,	0
OO,	4829,	OMA,	MSP,	3,	305	,	75,	0
OH,	6338,	GSO,	ATL,	3,	315	,	93,	1
OO,	6505,	LMT,	SFO,	3,	315	,	95,	0
US,	149,	SEA,	PHX,	3,	315	,	166,	1
US,	1640,	MCO,	CLT,	3,	315	,	97,	0
US,	908,	TPA,	CLT,	3,	315	,	96,	0
HA,	106,	HNL,	OGG,	3,	317	,	36,	0
9E,	3854,	DLH,	MSP,	3,	320	,	58,	1
DL,	1276,	MSP,	ATL,	3,	320	,	158,	0
OO,	4651,	FAR,	MSP,	3,	320	,	62,	1
CO,	1585,	MFE,	IAH,	3,	325	,	64,	0

CO,	222,	MSY,	IAH,	3,	325	,	70,	0
EV,	4966,	VPS,	ATL,	3,	325	,	69,	0

For the analysis purpose the data is taken from UCI repository. It is open repository provides the data collected from the primary sources. These sources are selected based on merits. The data provides is for educational purpose.

3. Interface

Fig. 4: Interface

It shows the snapshot of the Home Interface build using Matlab Simulator. In this snapshot by just simply clicking on the button the balancing of the data stream will be done.

4. G-Mean Comparison of Imbalanced Data and Balanced Data Stream

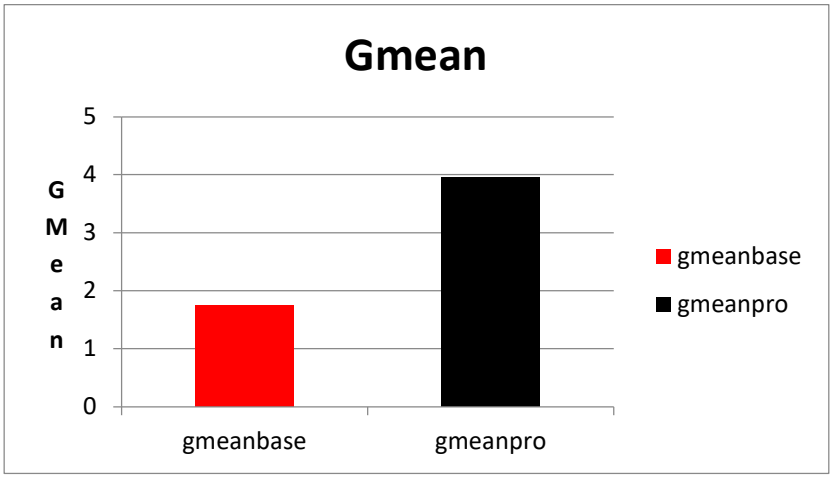

Fig. 5: G-Mean Comparison

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. AUC (AREA UNDER CURVE) Comparison

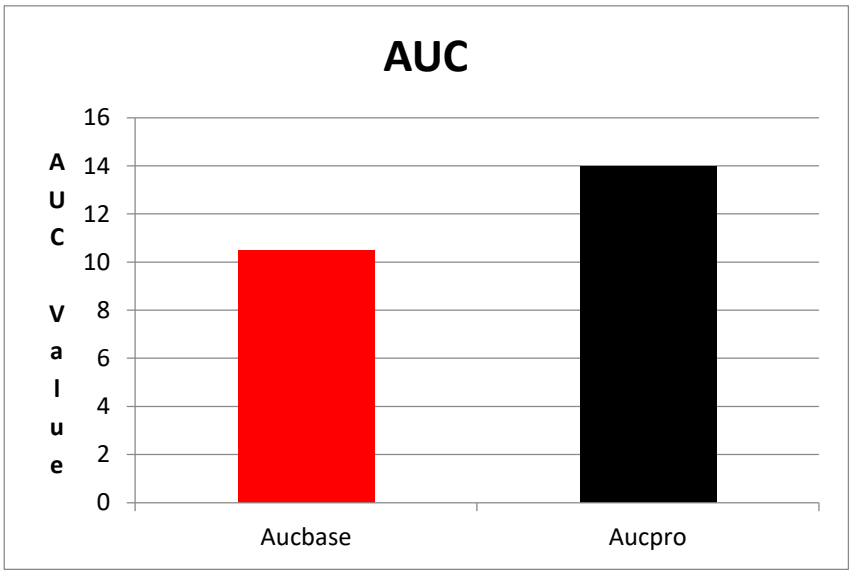

Fig. 6: AUC Comparison

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.

Conclusion

Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. 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.

References

[1] Chen 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, ACM, pp: 503-508, 2004.

[2] Stephen Lee, Krishna Kumaraswamy, Markus G. Anderle, Rohit Kumar, and David M. Steier. "Large scale detection of irregularities in accounting data",IEEE,pp:75-86, Washington, 2002.

[3] Yuleizhang, Taghi M. Khoshgoftaar, Jason Van Hulse, "Improving Learner Performance with Data Sampling and Boosting", IEEE, 2009.

[4] Sheng Chen, Pradeep Sinha, and KapilWankhade. "A fast and light classifier for data streams. Evolving Systems", vol. 1, pp:199-207,2010.

[5] Victoria López, ShivnathBabu, MayurDatar, Rajeev Motwani, and Jennifer Widom. "Models and issues in data stream systems". ACM, pp:116-134, USA, 2017.

[6] Nadeem,Lu "Effect Of Feature Selection, Synthetic Minority Over-sampling (SMOTE) And Undersampling On Class imbalance Classification", ACM, 2003.