Classification of Data Streams with Skewed Distribution

Pooja
M. Tech. Student
Department of Computer Science & Engineering
Mata Raj Kaur Institute of Engineering and Technology Rewari, Haryana, India

Abstract

The emerging domain of data stream mining is one of the important areas of research for the data mining community. The data streams in various real life applications are characterized by concept drift. Such data streams may also be characterized by skewed or imbalance class distributions for example financial fraud detection, Network intrusion detection etc. In such cases skewed class distribution of the stream increases the problems associated with classifying stream instances. Learning from such skewed data streams results in a classifier which is biased towards the majority class. Thus the model or the classifier built on such skewed data streams tends to misclassify the minority class examples. In case of some applications for instance, financial fraud detection the identification of fraud-ulent transaction is the main focus because here misclassification of such minority class instances might result in heavy losses, in this case financial. Increasingly higher losses due to misclassification of such minority class instances cannot be ruled out in many other data stream applications as well. The challenge, therefore, is to pro-actively identify such minority class instances in order to avoid the losses associated with the same. With an effort in this direction we propose a method using k nearest neighbours approach and oversampling technique to classify such skewed data streams. Oversampling is achieved by making use of minority class examples which are retained from the stream as the time progresses. Experimental results show that our approach shows good classification performance on synthetic as well as real world datasets.

Keywords: Data, Data Mining, Information and Knowledge

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 Jiawei Han in his book Data Mining: Concepts and Techniques 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 pre-processing. 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 via Classification, Clustering, Association rule mining etc. Our work mainly falls under the classification data mining technique.

A. 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 Unlike the traditional datasets, data stream arrive continuously yet varying speeds. Data streams are fast changing, temporally ordered, potentially infinite and massive. 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 petabytes, stream data has rightly captured our data mining needs of today. Even though complete set of data can be collected and stored its quite expensive to go through such huge data multiple times.

B. An Overview of Skewed Data Sets in the Real World

The rate at which science and technology have developed has resulted in proliferation of data at an exponential pace. This unrestrained increase in data has intensified need of various applications in data mining. This huge data in in no way necessarily equally distributed. Class skew or class imbalance refers to do- mains where in one class instances outnumber the other class instances, i.e. some classes occupy the majority of the dataset which are known as majority classes; while the other classes are

known as minority classes. The most vital issue in these kinds of data sets is that, compared to the majority of the classes, minority classes are often of much significance and interest to the user.

There are many real world applications, where in datasets contain such skewed nature. Following paragraphs gives an overview of some real world problems that exhibit such nature.

- Financial Fraud Detection: In financial fraud detection, majority of financial transactions are genuine and legitimate, and very small number of them may be fraudulent.
- Network Intrusion Detection: Here, the number of malicious activities are hidden among the voluminous routine network traffic. Usually there are thousands of access requests every day. Among all these requests, the number of malicious connections is, in most cases, very small compared to the number of normal connections. Obviously, building good model that can effectively detect future attacks is crucial so that the system can respond promptly in case of network intrusions.
- Medical Fraud Detection: In medical fraud detection, the percentage of bogus claims is small, but the total loss is significant.
- Real Time Video Surveillance: Imbalance is seen in data that arrives as video sequences.
- Oil Spillage: Oil spills detection in satellite radar images.
- Astronomy: Skewed data sets exist in astronomical field also; only 0.001% of the objects in the sky survey images are truly beyond the scope of current science and may lead to new discoveries.
- Spam Image Detection: In Spam image detection, near duplicate spam im- ages are difficult to discover from the large number of spam images.
- Text Classification: In text classification, there is imbalanced data such as text number, class size, subclass and class fold.
- Health Care: Health care domain has a classic example of class imbalance presence; the rare diseases affect very negligible amount of people, but the consequences involved are very severe. It id extremely vital to correctly detect and classify the rare diseases and the affected patients. If any errors occur in such cases it might be fatal.

II. LEARNING FROM SKEWED DATA STREAMS

In general learning from skewed data streams is challenging due to following issues.

- Evaluation Metric: Appropriate choice of evaluation metrics is also important in this domain. Evaluation metrics play vital role in data mining; they are used to guide the algorithms to desired solution. Thus if evaluation metric does not take minority class into the consideration, the learning algorithm will not be able to cope up well with the skewed data streams. The standard evaluation metrics like overall accuracy are not valid this case, because although minority class instances are misclassified then also the overall accuracy may remain higher, primary reason being negligible amount of minority class instances.
- Lack of minority class data for training: In skewed data streams due to lack of minority class data it becomes difficult to learn class boundaries. As the number of instances available are very few. Thus training a classifier in such situations is very difficult.
- Treatment of minority class data as noise: One of the major issues is that of the noise. Noisy data in the streams affects the minority class more than that of the majority data. Furthermore, standard stream mining algorithm tend to treat the minority class data as noise.
- As stated earlier data streams are usually massive and arrive at varying speeds: which makes it difficult to store them completely and multiple scans are nearly impossible and hence learning from such streams is difficult.
- Data streams also undergo concept evolution considerably over time: As per the above points we can see that classification of the skewed data streams is and multi-fold problem. All the above issues need to be addressed so as to design a appropriate learning algorithm for skewed data streams.

A. Overview of Methods for Dealing with Skewed Data Streams -Traditional Approaches

Some of the approaches for dealing with skewed data streams are categorised under following methods.

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

Oversampling and under-sampling are sampling based pre-processing 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 penalise the classifier.

1) Oversampling

Oversampling is one of the sampling based pre-processing 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). Some of the Oversampling based approaches in the literature are discussed below.

2) 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 filtering 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. Zhang et al 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 centroids 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 centroids of the clusters of negative samples. A new classifier was created on these sampled instances. Further size of the ensemble was fixed so for all classifiers present in the ensemble along with new classifier built on sampled instances.

3) Cost Sensitive Learning.

Cost sensitive learning is one if the important technique of data mining. It assigns different values of misclassification penalties to each class. Cost sensitive learning has been incorporated into classification algorithms by taking into account the cost information and trying to optimize overall cost during the learning process. In cost sensitive classification the problem is dealt by adjusting the learning. It is done by creating costs associated with misclassification of minority class and adjusting the learner based on punishment-reward system. One of the advantage of this method is that training dataset is unchanged unlike oversampling and under- sampling where in the data distribution changes completely.

III. MOTIVATION

While working on identification of the project topic in the area of data mining we found that lot of work has been already done in the different areas of the data min- ing with respect to the static datasets. Further in the last decade class imbalance problem in static datasets has drawn the attention of the data mining community. Due to which various workshops in the different conference were dedicated to specially for problem of class imbalance. First of such workshop was organized way back in 2000 in the AAAI 2000 conference, another workshop on ”Learning from Imbalanced Datasets” was organized in ICML 2003. Recently another workshop named ”Data Mining when Classes are Imbalanced and Errors have Cost” was organized in the PAKDD 2009 conference.

Various points from the above discussion drew our attention towards the problem of class imbalance. Further we could find that considerable amount of work has been done on the class imbalance problem in terms of static datasets. Then we went on to look for the another area wherein class imbalance is in more primary concern. Meanwhile we found that the data streams is an area where class imbalance has not been thoroughly studied. After this we concentrated on various real life application where in data streams are prominent. Various applications like Network intrusion detection, Financial fraud detection etc. are various areas which are characterised by stream data.

A. Evaluation Matrix

- Confusion Matrix: The columns of the confusion matrix represent the predictions, and the rows represent the actual class. Correct predictions always lie on the diagonal of the matrix.
- Recall: Recall is a metric that gives a percentage of how many of the actual minority class members the classifier correctly identified. (TP + FN) represent a total of all minority members.
- Precision: It gives us the total the percentage of how many of minority class instances as determined by the model or classifier actually belong to the minority class. (TP + FP) represents the total of positive predictions by the classifier.
- F-Measure: It is a harmonic mean of Precision & Recall. We can say that it is essentially an average between the two percentage. It really simplifies the comparison between the classifiers.
- G-Mean: G-Mean is a metric that measures the balanced performance of a learning algorithm between the classes.
- Area Under ROC Curve: The area under ROC Curve(Receiver Operating Characteristics) gives the probability that, when one draws one majority and one minority class example at random, the decision function assigns the higher value to the minority class than the majority class sample. AUROC is not sensitive to the class distributions in the dataset. Generally it is plotted as a True Positive Rate verses False Positive Rate. AUROC was mainly used in signal detection theory and medical domain where it was said to be the plot of Sensitivity verses (1−Specificity) where Sensitivity is same as True Positive Rate and Specificity is (1 - False Positive Rate) so in general it reduces to False Positive Rate thus both definitions of AUROC are one and the same.

IV. APPROACH TO DEAL WITH SKEWED DATA STREAMS

This section gives the brief description of our approach to deal with classification of data streams with skewed distribution. Our approach in general is as follows. Instead of using a single model from single training set, it is proposed to use multiple models from different sets. We use the ensemble of classifiers/models to classify the data streams with skewed distributions. The use of ensemble of classifiers has been proven to be the effective one to deal with concept drifting data streams. In our approach we have also used the oversampling approach to so as to balance the training chunk contents.

A. Experimental Setup

This section gives the abstract of the experimental setup on which we have worked to perform our experiments and test the effectiveness of our algorithm. For the implementation of our algorithm we have used MOA. MOA is a open source framework for mining data streams. It is implemented in java. It has collection of various data stream mining algorithms. MOA being an open source framework we could easily add and integrate our algorithm into it. Details of other hardware and software that we have used is as follows. We used MOA release 20110124 along with it we have also used WEKA version 3.7.5 on Windows 7 Profession operating system running on Dual core AMD Opteron processor 2210 @1.8 GHz with 2GB RAM.

The size of the ensemble is maintained to 10. As the size of ensemble goes beyond this specified limit then classifiers with best AUROC (Area Under ROC Curve) values are chosen. Thus as the stream is continuously coming in the ensembles continuously updated. Meanwhile the test examples are tested by taking the predictions from the ensemble. The predictions are made by taking weighted majority voting among the classifiers.

V. DATASETS

To evaluate the performance of our algorithm we have tested it on various synthetic as well as real world datasets. We have used various datasets available at UCI repository. Data set available at MOA data set repository.Also we have used various synthetically generated datasets. These datasets were chosen such that they model real world scenarios, have variety of features and largely vary in size and class distribution.

A. Synthetically Generated Datasets

This subsection presents the details of the synthetic datasets that we have generated. Later in this section we have briefly explained the details of these datasets. The main reason for generating these two datasets was that they focus on one of the main characteristic of data a stream that is concept drift. The characteristics of synthetically generated datasets. As the number of instances, number of majority class instances, number of minority class instances, number of attributes in the data set, chunk size of the data sets and the ratio of majority class to minority class.

B. Spinning Hyper Plane Dataset

The SPH (Spinning Hyper plane Dataset) in generated to model the Gradual Concept Drift. The SPH dataset was proposed by Wang et al. The SPH dataset defines a class boundary in n dimensions by coefficients as $\alpha 1, \alpha 2, \alpha 3 \ldots \alpha n$. An in- stance d= (d1,d2,d3...,dn) is generated by randomizing each attribute in the range 0 to 1.

The following part of the section provides information about the datasets used. We provide the common name of the dataset followed by the actual name of the dataset in the description.

- Letter: Letter Recognition Dataset - Here the Objective is to identify each one of the large number of black and white displays as one of the English alphabet capital letter. The character images taken were based on almost 20 different fonts. Each letter within these 20 fonts was randomly distorted to produce file of 20,000 unique stimuli. Each of these stimulus were converted to 16 primitive numerical attributes (edge counts and statistical moments) which were further scaled to fit the range of integers values through 0 to 15.
- Adult: In case of this dataset main task is to predict whether the income exceeds $50,000 per year or not based on the census data. The extraction of this dataset was done by Barry Becker from the 1994 census database .
- Connect-4: Connect-4 Opening Dataset - This dataset contains all the legal 8 ply positions in the game of Connect-4, in which none of the two players has won yet, and in which next move is not forced.

REFERENCES

[1] Moa dataset repository: http://moa.cs.waikato.ac.nz/datasets/.
[2] 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.
[3] Vahida Attar, Pradeep Sinha, and Kapil Wankhade. A fast and light classifier for data streams. Evolving Systems, 1:199–207, 2010.
[4] Brian Babcock, Shivnath Babu, Mayur Datar, Rajeev Motwani, and Jennifer Widom. Models and issues in data stream systems. In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, PODS '02, pages 1–16, New York, NY, USA, 2002. ACM.
[5] Stephen Bay, Krishna Kumar as wamy, 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.

[6] Albert Bifet, Geoff Holmes, Richard Kirkby, Bernhard Pfahringer, and Mikio Braun. Moa: Massive online analysis.
[7] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. J. Artif. Int. Res., 16:321–357, June 2002.
[8] S. Chen and H. He. Towards incremental learning of no stationary imbalanced data stream: a multiple selectively recursive approach. Evolving Systems, pages 1–16, 2011.
[9] Sheng Chen and Haibo He. Sera: Selectively recursive approach towards no stationary imbalanced stream data mining. In Neural Networks, 2009. IJCNN 2009. International Joint Conference on, pages 522–529, june 2009.
[10] 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.