A Review on diverse types of Data Mining Techniques

[1]Chanchal, [2]Rupali Zakhmi
Department of Computer Science Engineering
[1,2]SVIET,Banur,Punjab

Abstract: In today's world large number of applications which are used by millions of people. They produces large quantity of data. These large amount of data will be stored into large repositories. Later on with the help of data mining techniques these data items will be processed and relevant aspect can be identified. These relevant facts will be useful for various levels of decision making fact. This processed information can be in use for different organizations to know the behavior of people to whole organizations are directly linked with. But the performance of the data mining approach and techniques will be downgraded while we collect the data and classify the data into different pre set classes. Various researchers has worked on the techniques to remove these kind of problems occurs while data collection and storage. In nutshell it will enhance the data mining performances.

Keyword: Data Mining, Big Data, Pre-processing, Extraction

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[1].

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[1].

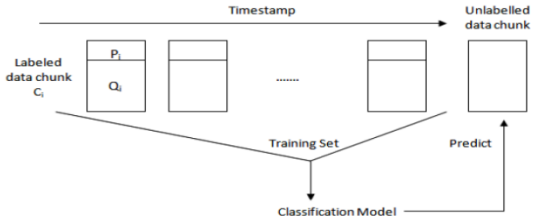

Figure 1 Classification of data

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. Figure 1 shows the classification task which maps attribute set x to its class label y[2].

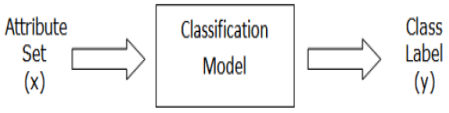

Figure 2: Classification as a task of mapping input attribute set x into its class label y

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. 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[2].

- 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[4]

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[2].

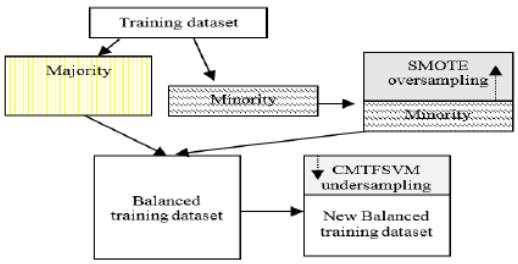

Figure 3 Training set

Data Mining Techniques

i. Association

Association is one of the best-known data mining technique. In association, a pattern is discovered based on a relationship between items in the same transaction. That's is the reason why association technique is also known as relation technique. The association technique is used in market basket analysis to identify a set of products that customers frequently purchase together[3].

ii. Classification

Classification is a classic data mining technique based on machine learning. Basically, classification is used to classify each item in a set of data into one of a predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network, and statistics[3].

iii. Clustering

Clustering is a data mining technique that makes a meaningful or useful cluster of objects which have similar characteristics using the automatic technique. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes[3].

iv. Prediction

The prediction, as its name implied, is one of a data mining techniques that discovers the relationship between independent variables and relationship between dependent and independent variables[3].

v. Sequential Patterns

Sequential patterns analysis is one of data mining technique that seeks to discover or identify similar patterns, regular events or trends in transaction data over a business period[3].

vi. Decision trees

The A decision tree is one of the most commonly used data mining techniques because its model is easy to understand for users. In decision tree technique, the root of the decision tree is a simple question or condition that has multiple answers[3].

II. LITERATURE SURVEY

[1] S. Chen et. al(2016) :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.

[2] Haibo He (2014) 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. **[3] G. Ditzler et. al(2016):** 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

[4] S Kang Li et. al(2016) 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.

[5] P. Jeatrakul et. al(2015) 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.

[6] Tom Fawcett et. al (2015) 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. CONCLUSION

On the basis of study of various research techniques based on different researches it is clear that, data mining is the critical for any organization which is having need of processed data. Due to the different levels of drawbacks in the existing techniques for data mining and data classifications the extraction will less optimized. Different types of improvements are defined and worked on by different researchers to enhance the techniques. This will enhance the quality of the extracted data. These processed data can be used for organization's decision making purposes. The data streams extracted from different large data producing applications have the defined entities of over sampling and under sampling.

IV. FUTURE WORK

Data mining for extracting useful data from large repository requires large number of pre processing and post processing steps. These steps are followed to enhance the results. Further the researches can be enhanced by considering over sampling for better classification and balanced class generation.

REFERENCES

[1] Sheng Chen Sera: "Selectively recursive approach towards non stationary imbalanced stream data mining.", In Neural Networks, 2009, pp:522-529, 2009.

[2] Haibo He, Kang Li, and S. Desai. Musera: "Multiple selectively recursive approach towards imbalanced stream data mining". In Neural Networks, PP: 1-8, 2010.

[3] G. Ditzler, R. Polikar, and N. Chawla. "An incremental learning algorithm for non-stationary environments and class imbalance". In Pattern Recognition (ICPR), pp: 2997-3000, 2010.

[4] S Kang Li. "Towards incremental learning of non stationary imbalanced data stream: a multiple selectively recursive approach". Evolving Systems, pp: 1-16, 2011.

[5] P. Jeatrakul, K. W. Wong, and C. C. Fung, "Classification of Imbalanced Data by Combining the Complementary Neural Network and SMOTE Algorithm " in 17th International Conference on Neural

[6] Tom Fawcett. Roc graphs: Notes and practical considerations for researchers. Technical report, 2004.

[7] 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.

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

[9] 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.

[10] 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.

[11] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "Smote: synthetic minority over-sampling technique.", J. Artif. Int,pp:321-357,2002.

[12] X. Fan and Z. He, "A Fuzzy Support Vector Machine for Imbalanced Data Classification," presented at the 2010 International Conference on Opto electronics and Image Processing (ICOIP), Haiko, 2010.

[13] A. Ralescu and S. Visa, "Fuzzy classifiers versus cost-based Bayes classifiers," Montreal, Que., 2006, pp. 302 - 305.

[14] S. Visa and A. Ralescu, "Fuzzy Classifiers for Imbalanced, Complex Classes of Varying Size," in The Internation Conference on Information Processing and Management of Uncertainty in Knowledge-Based System, (Perugia), 2004, pp. 393-400.

[15] R. Batuwita and V. Palade, "FSVM-CIL: Fuzzy Support Vector Machines for Class Imbalance Learning," IEEE Transactions on Fuzzy Systems, vol. 18, pp. 558-571, 2010.