An Investigation on Topic Maps Based Document Classification with Unbalance Classes

Maher Baloch[1], Muhammad Rafi[2]

[1]*Shaheed Zulfiqar Ali Bhutto Institute of Science & Technology (SZABIST) Karachi, Pakistan*

[2]*National University of Computer and Emerging Sciences (NU-FAST) Karachi, Pakistan*

[1]maherbaloch@gmail.com

[2]rafi.muhammad@gmail.com

Abstract— Classification of imbalanced data has become a widespread problem due to the fact that the most real world datasets are imbalanced. In a classification task, one of the challenges is to learn the feature-space of classification under class-imbalance setting. The majority classes generally have good representation of features in the learned classification function and the minority classes lack this representation; subsequently, the classification for these classes failed more often. In this paper, authors investigate the task of document classification with topic map based representation of documents under class imbalance setting. In order to measure of topic-map based representation for classification under imbalance data, authors compare three representations: Bag-of-Words, Phrases and Topic terms for three approaches (i) under-sampling, (ii) cost-adjusting, and (iii) cluster based sampling. A series of experiments are carried out and results are reported.

***Keywords*—Imbalance Classes, Document Classification, Topic Map, Sampling, Under-sampling, Over-sampling, 20-Newsgroup, Neural, Network, Naive Bayes, Support Vector Machine**

I. INTRODUCTION

The Machine learning / text mining realm a problem of document classification which is getting momentum as documents are rapidly growing from different source for example email, web blogs, social media, chat servers, news articles, magazines, books, and other sources. These sources of document required proper classification to get knowledge in the field of data mining.

Document Classification or text classification is a process of assigning document to a class; the document can be assign to one or more class in document classification. Document classification can be mapped from a set of document D : { d_1,d_2,d_3,…d_n }, to a set of category C : {c_1,c_2,c_3,….c_k} where every D -> C. The process can be carried out through different ways to assign document in class or classes [1].

There are three different types of document classification namely semi-supervised document classification, supervised document classification and unsupervised document classification. In supervised document learning approach, a tag of class is associated with each object. It is dominant paradigm of machine learning which learns from label data and make a distinction between them. Unsupervised document classification is a collection of document without any label or reference for classification. The semi supervised document classification is also known as partially supervised learning. It is used to learn with the small set of labeled example and a large set of unlabeled example. It is also used to learn with the positive and unlabeled examples. Text classification is increased because of document availability to ensure the need and organization. The Text classification may belong to one class or more than one class [1]. Widely employed learning technique is supervised learning method. The supervised learning method normally depends on amount of large labeled data in which the labor cost is higher with most labor hour's employed [2]. There were various problems with the supervised learning technique to be solved. Semi-supervised method is proposed to train small scale label data to a large amount of unlabeled data [2]. The labeled and unlabeled data assume to be balanced between the positive and negative samples in the existing semi-supervised learning methods [2]. In this case, the numbers of positive sample differs from the negative sample; therefore, the distribution of class is imbalanced in the cases of unlabeled and labeled data [2].

In machine learning, the problem of lesser positive data number of a class than the negative data class total number is termed as the Class imbalance. The Class imbalance problem is observed in various disciplines and practically, it is extremely common in medical diagnosis, anomaly detection, facial recognition, fraud detection, oil spillage detection, email filtering, managing risk, banking operation, predicting failure and etc. Machine learning algorithms provide efficient result when equal number of classes of each instance is available. But the problem arises when the number of instance classes is not equal. This problem is frequent in document classification when the numbers of instance of classes are not equal.

Fewer cases of class balance instances are observed in practical application. The imbalance classification can be distinguished in major and minor class. The major class has the more number of classes while the minor class has fewer

numbers of classes [3]. In class imbalance, the major class rule over minor classes in the classification and classification rate of the minor classes are not satisfactory [3]. The minor classes are ignored in classification; everything is predicted to be major class [3]. This is enduring problem of text classification has been a focused area of researchers. There are different methods and technique that has been discussed in this area. This problem gained attention when it has been encountered and provided balanced class distribution in standard learning methods [3].

The class imbalance problem is of crucial importance since it is encountered by a large number of domains of great environmental, virtual or commercial importance and was shown in certain cases to cause a significant bottleneck in the performance attainable by standard learning method which assumed a balanced class distribution.

In this paper, authors proposed Topic Map data structure for representation of rich semantic document and a compact. Topic Map is used to represent the information which can be found efficiently. It is a standard technology and describes the structure knowledge [4]. It is based on the formal model and on modern information management [4]. Following are the three types on which topic map is based 1) Association 2) Topic 3) Occurrence. The role of association in Topic Map is the relationship between the topics asserted by an element. The Subject in the Topic Map present in Topic on which the structure/unstructured documents are extracted (Topic Map representation is in Subject) [5].The relationship between the Topic and its subject is defined as Reification [5]. The Reification in Topic map is used to assigned characteristics of the subject in topic [5]. The Occurrence is used to represent the child element of the topic in Topic Map. The construction of Topic Map allows constructing multi-Topic topics as well as single-topic. Machine learning algorithms are used for the classification of document classification which is as follows 1) Support Vector Machine (SVM) 2) Naïve Bayes (NB).

II. LITERATURE REVIEW

The Classification can be used to improve the access of information [4]. The automatic organization of document or the classification is the need for the electronic document in rapid growth [4]. The retrievals, summarization and classification are the operations perform for document to extract information in text mining [6].The performing classification and the textual data is important in extracting information. In supervised learning technique, the label document is used for training set to set the predefined label in document for the category in the label.

As the Class imbalanced has been diagnosed in document classification, there have been several strategies proposed by researchers. The data mining and the machine learning implements many traditional algorithms to provide better approach to tackle the class imbalance problem [7]. The previous study shows that no semi-supervised method has been implemented for class imbalance but the focus was on the supervised imbalanced classification [2]. To balance multiple set for initial training, data is generated by under-sampling. A novel semi-supervised learning method was proposed which is based on the generation of random subspace to guarantee enough variation for the iteration process to dynamically generate various subspaces among the involved classifiers. The unlabeled data was successfully used in semi-supervised method that significantly static subspace generation to the generation of dynamic subspace [2]. The semi-supervised learning for the imbalanced class distribution was systematically addressed in the sentiment classification [2]. The learning point of view class is usually the instances which number of class is lowest [8]. The worst scenario of the class imbalance for two classifiers in 100 documents is 99:1 [8]. The internal method and the external method are the two categories on which the Class Imbalance Learning (CIL) method was divided. The Class Imbalance Learning (CIL) method is the existing technique for the imbalanced dataset to improve the performance of the classifier [8]. Balancing training dataset involves preprocessing in the external method while the modification of the learning algorithms in order to reduce their sensitiveness is dealt in internal methods [8]. The failing reason of standard machine learning algorithm for classification of imbalance data is the error in minority class classification commanding the error in majority class classification.

The feature selection, data level, cost sensitive level, ensemble level and algorithmic level are the techniques to handle imbalanced data problem [7, 8, 9, 10]. To save computational cost and to reduce the dimensionality of input data feature selection is key procedure. In machine learning algorithms, feature selection has been default step integrated in decision tree, k-nearest neighbors, artificial neuron network and etc. [9]. According to some measure, no predictive information or the subset of input features by eliminating features is the main idea of feature selection. There are two predictive approaches for the class imbalanced problem to adopt feature selection are 1) The beginning of new feature selection measures 2) eliminating on adapting class-probability [7, 8]. The feature is a process in which choosing a subset from the original one is frequently used as a preprocessing technique in analysis of data. The features selection has proved in enhancing result comprehensively, increasing mining accuracy, improving mining efficiency and reducing dimensionality [10]. The solution for the class is in the data level process as using oversampling algorithm for preprocessing and for balancing the data level process provides better results [7]. There are different sampling which are proposed at the data level named as over-sampling and under-sampling. The class population can be balance through eliminating the MA samples termed as under-sampling while the class population to balance through replicating the MI samples is known as over-sampling [2].

The random under sampling, random oversampling with replacement, directed under sampling, directed oversampling, oversampling with informed generation of new samples and combinations of the above techniques are the different forms of re-sampling at the data level [8, 10]. The cost sensitive learning emphasizes major effort in daily life scenarios like medical diagnosis and risk management. The different cost usually associated with the wrong decisions. In the case of cancer, the prediction of none existence may lead to the death. This could be false negative factor which shows the wrong prediction of cancer existence leading to death. While the false positive results in extra medical test and unnecessary anxiety. For positive rare classes, the cost factor of assigning to false negative and false positive will lead better performance [9]. For class imbalanced problem, many methods have been proposed like one-class classification, re-sampling and cost-sensitive learning to handle this issue. It is still unclear which method is more suitable for the classification [2]. To improve the performance of the over system, ensemble methods is used. The ensemble methods efficiency is highly reliant on the independence of the error committed by the learner. The diversity of the base learner and the accuracy depends upon the performance of the ensemble method. At the algorithmic level, specific learning algorithms, such as cost-sensitive learning, one-class learning and ensemble learning are proposed. The majority of work in feature selection for imbalanced data sets has focused on text classification or web categorization domain. Existing measures used for feature selection are not very appropriate for imbalanced data sets. They proposed a feature selection framework which selects features for positive and negative classes separately and then explicitly combines them to separates feature for positive and negative classes existing measures showed simple for converting [11].

Topic Maps [12] is a standard technology for describing knowledge structures and using them to improve the find ability of information. Topic maps are used to represent the subject or subjects extracted from the document. The TAO model of topic map is the information structure [5] which is used to structure information in topic maps formats. Topic Map has three kinds of assertions which are subject/topic name, occurrence and association. The major benefits of using topic maps are to reduce the size of document to get the related document from the document in structured format and to handle semantic topic.

III. CLASS IMBALANCE PROBLEM

To classify imbalanced data problem, standard machine learning algorithms fails because the majority rule over the minority classes. To solve imbalanced data problem, there are different techniques which are ensemble level, feature selection level, data level, algorithmic level and cost sensitive level.

A. Ensemble Level

Ensemble learning method in machine learning is a most successful approach for bagging and boosting. This method is widely used for class imbalance problems. This method improves the performance of the overall system. The ensemble method depends on the diversity of the base learner and accuracy. Manipulating the training data is to generate diverse base classifier is the easiest approach. The performance of weak classifier is improved by the boosting of ensemble learning algorithm.

B. Feature Selection Level

The feature selection in class imbalance is to eliminates feature from choosing a subset which having no predictive information to some measure or having no little information. Class imbalance problem in feature selection provides two approaches which are 1) the beginning of new feature selection measures and 2) eliminating on adapting class probability. The feature is a process in which choosing a subset from the original one is frequently used as a preprocessing technique in analysis of data. The features selection is proven technique for enhancing result comprehensibility, increasing mining accuracy, improving mining efficiency and reducing dimensionality.

C. Data Level

The Data level approach in class imbalance problems attempts to re-balance the class distributions and data space in pre-processing stage. The actual classification of the data level is self-determining and can be employed flexibly. The oversampling approach is most admired strategy in which the data space is introduces in artificial objects. The RAMO and ADASYN are the improved, most recently arrived alternative techniques but SMOTE is the unsurpassed technique. When having too much iteration, the oversampling methods may leads to other problems by slow down the class distribution.

D. Algorithm Level

The algorithm level in class imbalance is most common strategy which chooses appropriate inductive bias at the algorithmic level, specific learning algorithms, such as cost-sensitive learning, one-class learning and ensemble learning are proposed.

E. Cost Sensitive Level

The misclassifying patterns in imbalance dataset can be handling with Cost-sensitive learning method. The consequence of the classifying examples from one to another is the arithmetic illustration used in cost matrix. The objects for the minority class cost is the higher misclassification and overall learning cost is reduce in classification, in cost sensitive level the costs matrix represent the cost. The cost-sensitive drawback having lack of knowledge in setting values for cost matrices, this data is not available or not provided. In Cost-sensitive level is the method of increasing the weight of the minority classes by sampling method.

IV. CLASS IMBALANCE SOLUTION

Classifier algorithm is used to classify the instances of class. The minority class have the fever number of instances as compared to the majority classes which having large

number of patterns which is the main problem in the dealing with the classifier algorithm. It is difficult to classify the class imbalanced data because in classifying minority and majority classes which produce low accuracies. Solving the minority over majority class is to balance the class patterns to get balance dataset.

A. Sampling

The data is considered as imbalanced dataset if the distribution of the class is unequal in the classification of the class as compared to other classes. The large number of patterns is the majority classes which rule over the small number of patterns, small number represents the minority classes. The unbalance distribution of class in minority and majority classes can lead accuracies to low and high. To overcome with this problem, the class distribution of minority and majority classes can be balanced as the minority class has same number of patterns as the majority classes. Once the distribution of class is balanced then accuracies of the classification will increase. To balance class either majority class have same number of class as the minor classes or vice versa.

1) *Random Sampling*: The most sophisticated sampling is random sampling in which the equal number of class is selected for the classification. In random sampling, it is required to balance the classes which are over fitted in the distribution by removing them.

2) *Under-Sampling*: The under-sample method is used to balance the class distribution by eliminating the over fitted majority classes. This method is used to create subset for the original dataset in majority class instances by eliminating the instances. It will create the new instances from the original dataset by process of replicating.

3) *Over Sampling*: The oversampling is used to balance the dataset for the minority class data. This case is required when the dataset was balanced for the majority class and is in under-sample state. In oversampling, the imbalanced dataset problem is solved by balanced class for the accuracy in result. This method provide excellent results by distributing the equal classes and if required, it will increase the occurrence of the over fitting instances.

4) *Hybrid Method*: The Hybrid method is used to balance the imbalance data set by under and oversampling method. In other word, the hybrid method is the combination of both sampling methods. The hybrid method boosts the algorithm for the skewed training data learning.

V. EXPERIMENTAL SETUP

The proposed solution for the imbalanced dataset has been tested with the different sampling method in document classification.

A. Dataset

NEWS20: The News20 group dataset is most common dataset for research work in document classification and text classification machine learning. The document in News20 group is partitioned into 20 different groups with each group contain 1,000 documents.

B. Application

1) *Weka*: Weka is one of the tools which is used to test or train machine learning algorithms to predict data modeling, data virtualization to analyze and data preprocessing. Clustering on data, visualization, data processing, classification of the data and regression are supported by the Weka. These are assumption base technique.

2) *Wandora*: Extracting information and publishing application on Topic Maps is produced by the Wandora. The Topic Map is represented by the hyper graph which emphasize on subject which is generated by the hyper graph of Wandora. The Wandora is capable to provide graphical interface for the huge extracted information. The purpose of using this application in this research is only to generate the Topic Map for the given dataset.

3) *Open Calais*: Open Calais is a web service which integrates with ruby and Wandora. Open Calais is capable of reading any electronic document and then extract entities from document and make relation with the entities. The document which can be extracted by Open Calais are unstructured document, supplied document, news articles, HTML files, XML files, online blogs, text files, email and etc.

4) *Machine Learning Technique*: Machine learning is the process to learn from data in study of algorithms and construction dealing. The decisions or making predictions and model building are based on inputs of algorithms rather than explicitly instructed program.

This research work is focused on two machine learning techniques namely Support Vector Machine (SVM) and Naïve Bayes (NB).

C. Experiment

To generate Topic Map, the document from the News20 group dataset was selected. The document selection from the News20 group was random of any five classification of document. Topic Map was generated by using Wandora application. To generate Topic Map pass, documents were selected individually. Wandora performed well in extracting document. It also provided different extraction tools but Open Calais extractor was used. Open Calais parse the document by removing stemming, stop words and supporting word. It only keeps the object name and removes all the other irrelevant data from document. Once the irrelevant data is removed, open Calais created the association of the extracted objects. To create association, open Calais searched the object entity in Place, country, things, author, Company, person, position, email address and etc.

Let assume the Wandora pass the below unstructured document for open Calais

"In the current finical year Mario break along the existing records and it is also noticeable that in future it will break the other record."

Once this unstructured document was received by open Calais, it performed above mention step. In the above paragraph, open Calais can only find one entity that is company. Calais search for previously created entity, if not created, and then it will create company and make association with object as company – Mario.

In this document, Calais find the Named Entity, Facts and Events. Named Entity was detected in this case which Mario is laying in Company tag. Calais returns Company -> Mario. Calais did not detect any other tag in document so that tags are not returned.

Open Calais identified subject are then used to generate hyper plan where they are termed as Topic Maps. The topic Map select subject as topic and object as occurrences and show association with the subject. Creation of Topic map must be subject and document wise.

Pictorial representation of the above example of Topic Map is illustrated in figure 1.

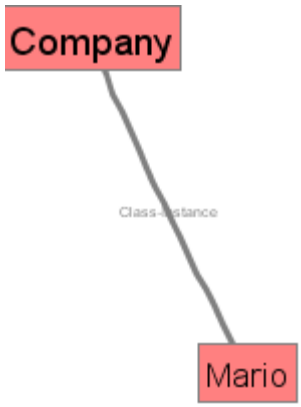

Fig. (1). Topic Map Example

Once the Topic Map of each subject document is acquired, it is required to convert them into XML form. The Conversion of Hyper map into XML will give the textual data for processing to the data. To compile the entire XML file in one CSV file, a function in written to serve the purpose. In the end of each document, it is necessary to provide the classification as supervised learning method is used. The above mentioned solution was applied to solve class imbalance problem step by step.. Machine learning technique required vector format so another function was made which converted CSV into Vector format CSV. Weka accepts (.arff) Attribute-Relation File Format file to train and test model.

VI. RESULT

The class imbalance solutions are applied on machine learning technique on the News20 group dataset. The obtain results are mentioned below.

A. Confusion Matrix

1) *Imbalance Class*: In this experiment, 24 documents were selected from the News20 group dataset and 2 classes for the C1 and C3 and 10 classes for the C2 and C4 were obtained. In this scenario, there were imbalanced classes and dataset was tested on machine learning algorithm. Table 1 and 2 are the confusion matrix for the Imbalance dataset.

Table 1. Naive Bayes Confusion Matrix

	C1	C2	C3	C4
C1	2	0	0	0
C2	0	10	0	0
C3	0	0	2	8
C4	0	2	0	8

Table 2. Support Vector Machine Confusion Matrix

	C1	C2	C3	C4
C1	0	2	0	0
C2	0	10	0	0
C3	0	2	0	0
C4	0	6	0	4

2) *Random Class*: In this second experiment, 24 random documents from the News20 group dataset were selected and there were 6 classifications for each document. In this scenario, there were balance classes and then dataset was tested on machine learning algorithm. Table 3 and 4 are the confusion matrix for the Imbalance dataset.

Table 3. Naive Bayes Confusion Matrix

	C1	C2	C3	C4
C1	6	0	0	0
C2	0	6	0	0
C3	1	0	5	0
C4	0	0	0	6

Table 4. Support Vector Machine Confusion Matrix

	C1	C2	C3	C4
C1	6	0	0	0
C2	0	6	0	0
C3	5	0	1	0

C4	2	0	0	4

3) *Under Sampling*: In this third experiment, 30 random documents from the News20 group dataset were obtained and which contain 5 documents for the C1 and C4 and 10 documents for C2 and C4. In this scenario, there were balance classes by eliminating the exceeding class from the C2 and C4 to make them balance with respect to other classes and dataset was tested on machine learning algorithm. Table 5 and 6 are the confusion matrix for the Imbalance dataset.

Table 5. Naive Bayes Confusion Matrix

	C1	C2	C3	C4
C1	5	0	0	0
C2	0	5	0	0
C3	1	0	4	0
C4	0	0	0	5

Table 6. Support Vector Machine Confusion Matrix

	C1	C2	C3	C4
C1	5	0	0	0
C2	0	5	0	0
C3	4	0	1	0
C4	1	0	0	4

4) *Over Sampling*: In this fourth experiment, 20 random documents from the News20 group dataset were obtained which contain 3 documents for the C2 and C4 and 5 documents for C1 and C2. In this scenario, balance classes were extracted by increasing minority and by repeating the number classes make, then it was balanced with respect to other classes. The dataset was tested on machine learning algorithm. Table 7 and 8 are the confusion matrix for the Imbalance dataset.

Table 7. Naive Bayes Confusion Matrix

	C1	C2	C3	C4
C1	5	0	0	0
C2	0	5	0	0
C3	1	0	4	0
C4	0	0	0	5

Table 8. Support Vector Machine Confusion Matrix

	C1	C2	C3	C4
C1	5	0	0	0
C2	0	5	0	0
C3	4	0	1	0
C4	1	0	0	4

VII. CONCLUSIONS

In this study, the authors endeavored to find out solution for the class imbalance problem that how this problem can be overcome. For that, different samples were selected and tested on by using different method. In first experiment, imbalance dataset was tested and unacceptable results were obtained. In this regard classification result will be not good for the minority class. In this case majority classes ruled over minority class. Then class imbalance solution was proposed to be experimented on random sampling, over sampling and under sampling. All the obtained results were acceptable for the class imbalance problem as the case of majority class over minority class was discarded by equal distribution of classes.

VIII. ACKNOWLEDGMENT

The authors would like to thanks Shaheed Zulfiqar Ali Bhutto Institute of Science and Technology for the research support.

REFERENCES

[1] F. Sebastiani. “Classification of text, automatic”. In *The Encyclopedia of Language and Linguistics*, Keith Brown, Eds.,Volume 14, 2nd Edition, Elsevier Science Publishers, Amsterdam, NL, 2006, pp. 457-462.

[2] S. Li, Z. Wang, G. Zhou and S. Y. M. Lee, "Semi-supervised learning for imbalanced sentiment classification." In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*. 2011, Vol. 3, pp: 1826-1831.

[3] N. Japkowicz, and S. Stephen. "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, vol. 6, no .5, pp: 429-449, 2002.

[4] M. Rafi, M. S. Shaikh and A. Farooq. "Document Clustering based on Topic Maps." *International Journal of Computer Applications*, vol. 12, no. 1, 2010.

[5] S. Pepper and G. Moore, Ed., XML Topic Maps (XTM) 1.0; TopicMaps.Org 2001, http://www.topicmaps.org/xtm/1.0/

[6] B. Baharudin, L. H. Lee, and K. Khan. "A review of machine learning algorithms for text-documents classification." *Journal of Advances in Information Technology*, vol. 1, no. 1, pp: 4-20, 2010.

[7] D. Ramyachitra, and P. Manikandan, "Imbalanced Dataset Classification and Solutions: A Review." *International Journal of Computing and Business Research*, vol. 5, no. 4, 2014.

[8] S. Jayasree and A. A. Gavya. "Addressing imbalance problem in the class–A survey." *International Journal of Application or Innovation in Engineering & Management*, vol. 3, no. 9, 2014.

[9] Y. Liu, H. T. Loh and A. Sun. "Imbalanced text classification: A term weighting approach." *Expert systems with Applications* vol. 36, no. 1, pp: 690-701, 2007.

[10] V. Ganganwar, "An overview of classification algorithms for imbalanced datasets." *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 4, pp: 42-47, 2012.

[11] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou. "On the class imbalance problem." In *Proceedings of Fourth International Conference on Natural Computation, (ICNC'08)*, 2008, vol. 4, pp. 192-201.

[12] S. Pepper, *Topic Maps*, Encyclopaedia of Library and Information Sciences, 3rd Ed.