*Pakistan Social Sciences Review* *P-ISSN 2664-0422*
*March 2021, Vol. 5, No. I [491-504]* *O-ISSN 2664-0430*


Pakistan Social Sciences Review
www.pssr.org.pk

**RESEARCH PAPER**

# Expert System for the Identification of Review Papers Using Ensemble Learning

Dr. Ghulam Mustafa[1] Dr. Naveed Jhamat*[2] Dr. Khurram Shahzad[3]

1. Assistant Professor, Department of Information Technology, University of the Punjab, Gujranwala Campus, Punjab, Pakistan
2. Assistant Professor, Department of Information Technology, University of the Punjab, Gujranwala Campus, Punjab, Pakistan
3. Assistant Professor, Punjab University College of Information Technology, University of the Punjab, Lahore, Punjab, Pakistan

**PAPER INFO**

**Received:**
February 06, 2021
**Accepted:**
March 01, 2021
**Online:**
March 15, 2021

**Keywords:**
Class Imbalance Learning,
Ensemble Learning
Expert System,
Machine Learning
Medline Abstract Classification,
Performance Evaluation

***Corresponding Author**
naveed.jhamat@pugc.edu.pk

**ABSTRACT**

Creating an effective classifier in the presence of imbalanced data is a challenging task. The objective of this work was to apply machine learning technique to automatically identify review articles given the imbalance representation of publications types in publications. As a contribution in that direction; we develop a hybrid ensemble algorithm, called Balanced MultiBoost (BMB). The presented algorithm provides an efficient alternative to existing algorithms, by combines the strengths of Multiboost ensemble with the sampling technique. In order to demonstrate the effectiveness of BMB, we compared its performance with five existing algorithms, based on established metrics, precision, recall, $F_1$-measure and AUC-ROC. For the comparison, we used two customized datasets extracted from Medline citations database. These datasets contain 19,299 examples for 2005 and 19,200 examples for 2006 with imbalance ratio 1:6 and 1:7, respectively. The results show, BMB is a powerful ensemble solution for identifying minority examples in a text corpus.

## Introduction

A review article incorporates the pertinent published literature depicting the eminent search evidence that satisfy a particular research question (Sackett et al., 1996) (Greenhalgh et al.,2014) (Boynton et al.,1998) and a non-review article aims to develop/validate an artefact. The ever-growing research articles make it difficult and time consuming task to segregate review articles from non-review articles, hence hindering the process of searching review articles. An alternate approach to detection of review articles is automatic classification of articles into review and non-review articles. In addition to detection of review articles, there are several benefits of automatic classification of articles. It includes, a) reduced effort for classifying a

large collection of articles,b) enhanced precision in searching of review articles,c) consistent annotation of review articles in digital libraries i.e. independent of bias that might be induced in manual annotation. Besides, classifying research articles, there are several other areas in which text classification can be applied, such as spam filtering (Cormacket al.,2011), sentiment classification (Panget al.,2002), news classification (Kallipolitis, et al.,2012), word sense disambiguation (Escudero, et al.,2000) and abstract classification (Trieschnigg et al.,2009), etc.

Creating an effective classifier is a non-trivial task because a number of challenges are associated with it. One such challenge is the presence of imbalanced data that is used for training a classification model (Heet al., 2009). Class imbalance refers to a case where one class outnumbers another class (Chawla, et al., 2011). The former is called majority class and the latter is called minority class. For example, a text corpus of research articles which is imbalanced in the sense that there are fewer review articles (the minority class) compared to non-review articles (the majority class). A classification model trained on imbalanced dataset gives effective results for the majority class; however, the results of the classification model may not be equally effective for the minority class (Weiss, et al., 2001). The underlying reason is, during training, the classification model over represents the majority class and under represents the minority class. It becomes further challenging if the imbalanced data is a collection of research articles. It is because research articles are textual data with high dimensionality, higher probability of noise and class imbalance (Wu, et al.,2014).

In this study, we propose a hybrid ensemble called Balanced MultiBoost (BMB) as an effective algorithm for text corpus in order to distinguish review articles from non-review articles. A key feature of BMB is that, it combines MultiBoosting (Webbet al.,2000) with random undersampling. MultiBoosting is a combination of boosting and wagging (a variant of bagging). Since, boosting is effective in reducing bias and variance in error but less sensitive to noise, while bagging is more effective in reducing variance in error (Webbet al.,2000), so the combined effective of both methods reduce the noise. This in turn enhances the performance of MultiBoosting. Random Undersampling on the other hand, is more effective as compared to oversampling because it increases the processing speed. Although, undersampling generally results in loss of information however in our case the examples left out in current iteration have chances to get selected in the other iterations and thus it minimizes the information loss problem.

**Literature Review**

A basic classification model, such as CART, trained on imbalanced textual corpus may not produce effective results for review articles. Two key techniques to handle the class imbalance problem are data sampling and ensemble learning (Batista et al.,2004). Data sampling strategies resolve the issue by arbitrarily changing the sample either by reducing the majority-class sample or by increasing the minority-class sample. The former is called undersampling and the latter is called oversampling. For the research articles corpus, undersampling would

randomly remove the examples of non-review articles from the training dataset and oversample would randomly add example of review articles. A key advantage of undersampling is reduction of training time and disadvantage is loss of information. The key advantage of oversampling is no information lost and disadvantage is inclusion of duplicate examples which may lead to over fitting the model (Batista et al.,2004), which decreases precision.

A common approach to further improve the effectiveness of classical algorithms is to build an ensemble of models i.e. an approach for combining multiple weak learners to build a strong learner. Ensemble learning algorithms require balanced class representation i.e. the algorithms are not sufficiently effective in the presence of imbalanced dataset. For the research articles corpus, where the data is imbalanced the ensemble learning algorithm may not be effective for the minority class. To address the class imbalanced problem, data resampling techniques are incorporated into ensemble learning algorithms. There are three categories of these algorithms (Galar et al.,2011), boosting based, bagging based, and hybrid.

- Boosting based algorithms adjust the allocation of weight to the minority class in order to train the succeeding learner. For instance, SMOTEBoost (Chawla , 2003), RUSBoost (Seiffert et al.,2009), and EUSBoost (Galaret al.,2013) change the class distribution to favour the minority class. A key benefit of this class of algorithms is sampling techniques introduce more diversity which thereby improves the ensemble performance (Wanget al.,2009). A key limitation of these algorithms is, they may introduce the over fitting or under fitting problem due to the inclusion of synthetic examples or random removal of examples, respectively (Liuet. al.,2008).

- In bagging based algorithms, class distribution is changed when examples are drawn from the original sample. Over Bagging and SMOTE Bagging (Wang, et al.,2009) are the two bagging based algorithms that incorporate random oversampling and SMOTE sampling techniques to change the final class representation. The key benefit of this class of algorithms is increased diversity in ensembles (Wang et al.,2009).

- Hybrid ensembles combine both bagging and boosting with data sampling techniques to form hierarchical ensembles. For instance, EasyEnsemble (Liu et al.,2008) employs bagging with exploratory undersampling as the primary ensemble and AdaBoost as secondary ensemble, thus forming an ensemble of ensembles. The benefit of combining ensembles is to inherit the benefit of both bagging and boosting.

SMOTEBoost (Chawla, et al.,2003) and RUSBoost (Seiffert,2009) are the two algorithms especially designed for the imbalance dataset by combining boosting and data sampling techniques. SMOTEBoost combines SMOTE with boosting and RUSBoost combines RUS with boosting. The advantage of SMOTEBoost is that it intelligently applies oversampling and the disadvantage is that it escalates the

drawback of SMOTE (Chawlaet al.,2002) i.e. increased training time, complexity and over fitting problem. On the other hand, RUSBoost is a simpler, faster and less complex alternative to SMOTEBoost, and it surpasses various predecessors (Chawlaet al.,2003). A key limitation of these two algorithms is that they do not perform well in the presence of textual data. This justifies the need for an algorithm that is (an alternative to SMOTEBoost and RUSBoost) robust and works well in the presence of textual data.

**Materials and Methods**

**Balanced MultiBoost (BMB)**

The primary aim of ensemble learning algorithms is to improve the classification efficiency (in terms of accuracy) by combining diverse and weak learning algorithms. It is because; the ensemble methods often perform better than their base learning algorithms. Therefore, we develop an ensemble of ensembles for the classification of a text corpus of research articles that is imbalanced and textual data.

For the Balanced MultiBoost (BMB), we propose to use balanced sampling in MultiBoosting instead of continuous Poisson distribution. Unlike other data sampling techniques, examples from training data are not strategically omitted. Instead, it would randomly delete examples from the dominant class before a balanced distribution is achieved.As a result of this change in sampling technique (from continuous Poisson distribution to balanced sampling technique using undersampling), certain information may be lost during the iteration of MultiBoosting. However, it is likely that the lost information will be included during the other iterations. The suggested change in sampling technique is highly suitable for BMB ensemble, because ensemble learning algorithms requires prolonged learning time; however, the use of random undersampling, which is a simpler and faster sampling approach, reduces the learning time of the ensemble algorithm while improving the performance.

BMB combines the data sampling technique with MultiBoost in such a manner that it modifies and bias the weight distribution towards the minority class in the iterations. BMB discards the examples from the majority class using random undersampling and gets a balanced representation of minority and majority class examples in the iterations. This approach does not assign new weights to the examples instead it only normalizes the weights of the remaining examples in the new dataset with regard to their total sum of weights. After developing and evaluating a hypothesis original examples weights are updated and then other iterations are applied to modify the weights. This introduces the balance and diversity in the ensemble and thus increases the performance of the ensemble to improve the performance of the minority class.

BMB takes committee size $T$ as an input argument and compute the size and number of subcommittees, by$\sqrt{T}$. The number of subcommittees represents the

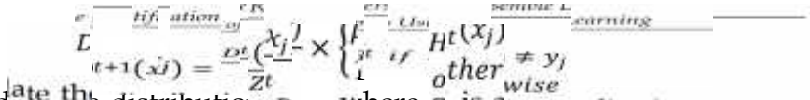
number of times a balanced sample is generated from the imbalanced dataset, by undersampling. The AdaBoost constituent shall be called upon with each subcommittee having a size equal to that of the subcommittee. A weak hypothesis is established and tested for each iteration of AdaBoost. The AdaBoost Sub-Committee shall be disbanded if the defect is too significant or zero, and the new Sub-Committee shall be appointed of an expanded scale to compensate for the premature termination of the previous Sub-Committee. At the end of the day, all subcommittees are aggregated into a weighted ballot. Algorithm 1 demonstrates the BMB algorithm.

**Corpus**

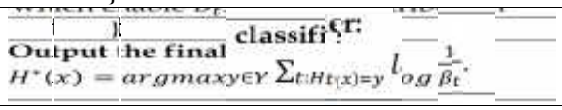
Medline is a bibliographic database of the U.S. National Library of Medicine (NLM) that have over 24 million citations (NLM: MEDLINE ,2015). It is a rich source of information and it is widely used for information exploration and knowledge discovery. The 2013 release of these citations is in the form of 714 files of bibliographic data in XML format. These files were parsed and transferred into a relational database. From the database we extracted data about two large subsets of articles published for the year 2005 and 2006. There were 589540 and 627599 articles respectively in the two years with an imbalance ratio of 1:6 and 1:7, respectively. We randomly extracted a sample set of citations, for our experiments.

**Algorithm 1 Balanced MultiBoost (BMB) Algorithm**

| | |
|---|---|
| **Input** | et, S. A sequence of be**Boost (BMB**les (.., it labels: Data s weak learning algo n, lled exam of iterations, , V specifying the number of iterations for each subcommittee. |
| **Output:** ensemble H* | |
| 1 | S' = S with example weights assigned to be 1. %initialize the weight distribution |
| 2 | Set k = 1. |
| 3 | For t = 1 to T { |
| 4 | If n |
| 5 | ther g data set wit balanced distribution by % Create temporary trainir undersampling themajority. |
| 6 | Normalize um to 1. |
| 7 | Set k=k+1. |
| 8 | in a base learner m dataset |
| 9 | % train a % calculate the error on the training set |
| 10 | If 0.5 or 0 then |
| 11 | Go to step 5. |
| 12 | % calculate the weight of |
| 13 | For each % |

14 
% update the distribution $D_{t+1}$, where $Z_t$ is a normalization constant which enable $D_{t+1}$ to be a distribution

15 }



It is important to note, that the Medline citations includes several data values about each article, such as, title, authors, journal name, publication year, article type and abstract, etc. However, for experimentation, we necessarily require three data values about each article. These are, a) title, to uniquely identify each article, b) abstract, necessary to generate a textual corpus of articles, c) type, to distinguish between review and non-review articles. From the 2005 and 2006 dataset it was observed the three data values were not available for all articles. Therefore, 'during the choice of sample', only those articles were randomly selected for experimentation where title, abstract and type of article were available. Accordingly, the data samples selected for the experiments contained, 19,299 and 19,200 citations for the year 2005 and 2006, respectively, keeping the imbalance ratios of the two years (i.e. 1:6 and 1:7). Further statistics of data sets are given in Table 1.

**Table 1**
**Specification of the datasets**

| Data Set | Total review articles | Total non-review articles | The Selected Sample | | | IR |
|---|---|---|---|---|---|---|
| | | | Total articles | Review articles | non-review articles | |
| Medline 2005 | 84797 | 504743 | 19299 | 2757 | 16542 | 1:6 |
| Medline 2006 | 75015 | 552584 | 19200 | 2400 | 16800 | 1:7 |

The size of the data samples considered for the experimentation, should be seen in the following context, a) the existing studies, such as RUSBoost and SMOTEBoost use nominal or numeric data which reduces the computation complexity. In contrast, the type of data used in our experiments is a corpus of Medline research articles, which has large number of features, higher probability of noise and involves higher computation complexity, b) the existing studies use datasets of smaller sizes having lesser number of features (called attributes). The sizes of dataset vary between 214 and 11183 and numbers of features vary between 7 and 43. In contrast, our selected sample contains, 19299 and 19200 citations; and it has 6,671 distinct features for the year 2005 and 6,790 for the year 2006.

**Pre-processing the Corpus**

In contrast to nominal or numeric data used by existing studies (Chawla et al.,2003) (Seiffert et al.,2009), the textual corpus requires multifaceted pre-processing,

before it can be used for experimentation. Each textual document (a concatenation of title and abstract) is represented as a vector model; where each dimension of the vector model corresponds to a separate term, called feature. Altogether, the collection of features is called feature space. In the feature space, some words are more informative than others. Besides that, some words are little informative, but they have higher frequency. Thus, these words contribute little if used in text classification and should therefore be removed. These words are called *stopwords*. We used PubMed stopwords list (PubMed Help,2015) to filter out these words from our corpus.

Furthermore, stemming is employed to trim the words to their roots in order to increase the effectiveness of retrieval. For that, all text documents were converted into lowercase before applying *Snow ball Stemmer*. Subsequently, the generated text was tokenized to generate individual words (called tokens) which are thereafter used as features. If a term appears in a text, the vector value may not be zero. Term frequency inverse document frequency (tf-idf) weighting is widely used as a tool. Finally, we computed tf-idf based on formula (1+logtf)*logN/df, where N is the total number of documents, tf is the frequency of the word and df is the frequency of the document.

**Evaluation Measures**

Precision, recall, precision and $F_1 - \text{measure}$ are the most critical metrics used to compare classification performance. Among these measures, accuracy is argued to be unsuitable for imbalanced data (Fawcettet al.,2006). The underlying reason is, accuracy shows a cumulative score of majority and minority classes, consequently, the effectiveness of classification for minority cannot be measured explicitly. Recall, the challenge in hand is, a classification model trained on imbalanced dataset gives effective results for the majority class, however the results of the classification model may not be equally effective for the minority class. Therefore, for true evaluation of our ensemble model explicit measures that depict the performance of minority class are needed. Henceforth, we use precision, recall and F1-measure. The three measures are defined below:

$$\text{precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (1)$$

$$\text{recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (2)$$

$$F_1 - \text{measure} = \frac{2 \times \text{precison} \times \text{recall}}{\text{precison} + \text{recall}} \quad (3)$$

In addition, (Ian ,2005) proposed the use of the Receiver Operator Characteristic (ROC) curve when dealing with the problem of binary classification. The curve illustrates how the number of correctly classified examples differs from

the number of falsely classified examples using a false positive rate and a true positive rate pair. This curve is an effective output indicator for imbalanced data sets. Therefore, in our experiments, we used the area under the ROC curve (AUC-ROC) which is a scalar representation of the ROC curve.

**Experimental Design**

In order to evaluate the effectiveness of our proposed ensemble algorithm, we compared the performance of five existing methods that share some features with our proposed technique. These are MultiBoost, SMOTE, Balanced Random Forests (BRF) (Chen et al.,2004), SMOTEBoost, and RUS Boost. MultiBoost is chosen for comparison, because we adapt the MultiBoosting algorithm to form BMB. So, the comparison will establish that its modifications have enhanced the effectiveness of the algorithm. SMOTE is chosen for comparison because it is a start-of-the-art effective boosting algorithm for classification of imbalanced data. BRF is chosen for comparison, because similar to BMB, BRF utilizes balanced samples of training data to form a random tree ensemble. Also, we use SMOTEBoost and RUSBoost, because, similar to BMB, they are hybrid ensemble algorithms for classification.

For experimentation, the committee size of MultiBoost with parameter T was set to nine, because it gives optimal results at this value (Webbet al.,2000). For BRF, the number of iterations was also set to nine to ensure symmetric number of iterations for each algorithm, for a fair comparison. Similar number of iterations was set for SMOTEBoost and RUSBoost, with an exception is SMOTE. For SMOTE, the $k$ nearest neighbour parameter was set to five, which is a default optimal value. For the experiments, all the algorithms were implemented using WEKA tool (Hallet al.,2009). We used 66% data as training data and 34% data as test data in our experiments, due to larger datasets and more demand of computational resources in our experiments.For each algorithm, all experiments were repeated five times to eliminate the bias that might appear due to sampling. The target class distribution for review and non-review articles was set to 50:50.

**Results and Discussion**

The experiments were performance on two customized Medline datasets and the values of the five evaluation metrics (precision, recall, F1-measure, AUC and ROC) were computed. Table 2 presents the average results of each algorithm, averaged across five repetitions. It is important to note that the table only contains the performance of the minority class (reviews articles) i.e. the results of the majority class (non-review articles) are excluded. We used the Wilcoxon signed-rank test to show the statistical significance of our approach. Asterisk sign (*) in the table represents that our proposed method is statistically significant over all other methods (Wilcoxon signed-rank test, $p<0.05$).

**Table 2**
**Comparison of five ensemble algorithms with BMB**

| Data Set | Method | Precision | Recall |  | AUC |
|---|---|---|---|---|---|
| **Medline 2005** | MultiBoost | 0.6612 | 0.6111 | 0.6351 | 0.8236 |
| | SMOTE | 0.8991 | 0.8043 | 0.8490 | 0.9093 |
| | BRF | 0.8731 | 0.7225 | 0.7904 | 0.8916 |
| | RUSBoost | 0.8321 | 0.7575 | 0.7929 | 0.8234 |
| | SMOTEBoost | 0.8749 | 0.7692 | 0.8182 | 0.8861 |
| | BMB | 0.9619* | 0.7842 | 0.8639* | 0.9285* |
| **Medline 2006** | MultiBoost | 0.6742 | 0.6078 | 0.6392 | 0.8212 |
| | SMOTE | 0.9085 | 0.8139 | 0.8584 | 0.9155 |
| | BRF | 0.8554 | 0.7235 | 0.7839 | 0.8775 |
| | RUSBoost | 0.8489 | 0.7856 | 0.8159 | 0.8396 |
| | SMOTEBoost | 0.8907 | 0.7706 | 0.8261 | 0.8870 |
| | BMB | 0.9712* | 0.7907 | 0.8716* | 0.9387* |

From Table 2 it can be observed, BMB outperformed all the existing ensemble algorithms i.e. it scored higher F1-measure on Medline data set 2005 and 2006. This indicates that, compared to other algorithms, BMB is more effective in classifying minority class in an imbalanced text corpus.

Another observation is, the BMB results are equally effective for Medline 2005 and Medline 2006 dataset; and BMB drastically outperformed MultiBoost using F1-measure. Their respective scores for year 2006 are 0.8716 and 0.6392, and for year 2005 are 0.86392 and 0.6351. It is because; in MultiBoost the minority class is significantly less represented in the dataset than the majority class in both datasets, indicating that MultiBoost does not address the imbalance problem in data effectively. However, by introducing the balanced dataset instead of continues Poisson distribution; BMB drastically improved the performance on minority class

Furthermore, the performances of BRF, RUSBoost and SMOTEBoost are comparable to each other while SMOTEBoost has an edge over BRF and RUSBoost in *precision* performance evaluation metric for both datasets. While, RUSBoost has an edge over other two methods in *recall* on 2006 data set but SMOTEBoost score higher $F_1$-measure. For AUC measure, BMB performed exceptionally well as compared to the baseline method, MultiBoost. BMB performed better than all other methods on AUC measure. The SMOTE and SMOTEBoost methods are closest contender of BMB on AUC measure.

From Figure 1 and Figure 2, it can be seen that the ROC values of BMB are better than all other methods on both datasets. The values of ROC for both datasets demonstrate that BMB effectively classify review and non-review publications. In identification of reviews publications, the difference of the BMB with all other

learning methods is not by chance, instead, BMB performance is significantly better on *precision*, $F_1$-measure and AUC metrics than all other methods.

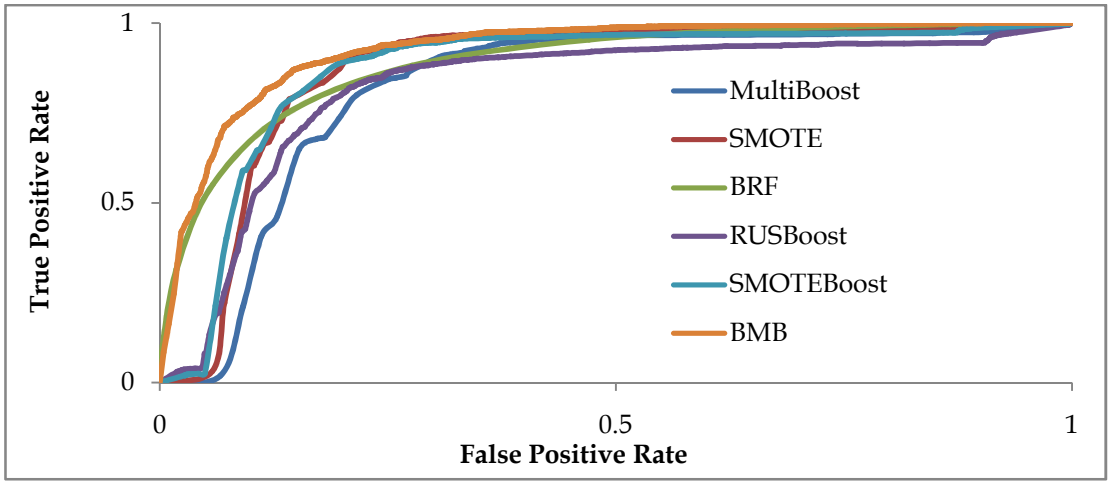



Figure 1. Comparison between the ROC space performance of BMB ensemble learning and other learners trained using Medline 2005 dataset.

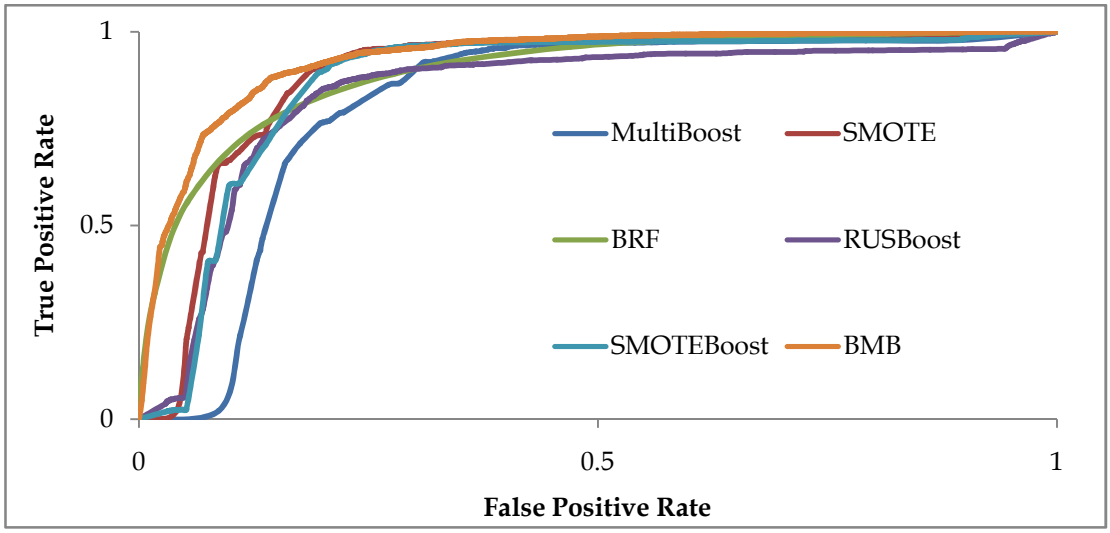



Figure 2. Comparison between the ROC space performance of BMB ensemble learning and other learners trained using Medline 2006 dataset.

A balanced resampling approach with MultiBoost is adapted to improve the effectiveness of classification performance in an imbalanced dataset. Our proposed approach, BMB outperformed all other competing methods on different metrics. The closest competing method to our approach is Synthetic Minority Oversampling Technique (SMOTE). SMOTE randomly creates artificial examples by k nearest neighbours of the minority class example. SMOTE generalized the decision boundaries for the majority class and thus deals with over fitting problem which could be created in simple oversampling techniques. This broaden in decision

boundaries not only increase the recall of the minority class but also increase the precision of the minority class and hence overall increase in the F1-measure.

As compared with SMOTE, BMB correctly classify more review articles, however SMOTE returns more review articles than BMB. In other words, SMOTE expands the review articles concept space and thus the recall increases. However, at the same time SMOTE has less impact on precision due to the false positive examples. The disadvantage of SMOTE is computationally expensive in the sense that it takes more time for training, while creating synthetic examples, and therefore not suitable for large and high dimension data sets. Thus, SMOTE is more useful when users are interested in retrieving several results, and it is followed by a more rigorous assessment of classifying articles. On the other hand, BMB uses undersampling the majority class and balances the class distribution, which takes less time for training. It achieves higher F1-measure because it decreases the error due to bias and variance; and for every subcommittee new examples from majority class are randomly selected. Consequently, more diverse classifiers are created in the BMB ensemble. Hence, it is useful when users are interested in accessing only relevant data. One such application is a question answering system.

To summarize, the performance of BMB on precision metric is better than all algorithms, indicating that BMB returned considerably more relevant articles (review articles) than irrelevant articles (non-review articles).

**Conclusion and Recommendations**

Reviews contain concise information about a specific domain but are difficult to identify from a large collection of publications. In this paper, we outlined Balanced MultiBoost (BMB) algorithm to automatically identify review articles using class imbalanced Medline datasets. We adapted the MultiBoost ensemble and used undersampling technique to balance the class distribution to address the class imbalance problem. Experimental findings show that the classification efficiency of our proposed system is higher than any other competing machine learning method.Based on the experiments, we conclude that BMB is computationally more economical with large data sets as compared to SMOTE, and therefore it is effective for automated classification of text corpus. Our future work includes extending our work to other domains. It is suggested that neural network techniques may be used to identify the review articles.

**References**

Boynton, J., Glanville, J., McDaid, D., & Lefebvre, C. (1998). Identifying systematic reviews in MEDLINE: developing an objective approach to search strategy design. *Journal of Information Science*, 24(3), 137-154.

Batista, G. E., Prati, R. C., &Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1), 20-29.

Chawla, N. V., Japkowicz, N., &Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, 6(1), 1-6.

Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003, September). SMOTEBoost: Improving prediction of the minority class in boosting. In *European conference on principles of data mining and knowledge discovery* (107-119). Springer, Berlin, Heidelberg.

Chawla, N. V., Bowyer, K. W., Hall, L. O., &Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357.

Chen, C., Liaw, A., &Breiman, L. (2004). Using random forest to learn imbalanced data. University of California, Berkeley, 110(1-12), 24.

Chawla, N. V., Bowyer, K. W., Hall, L. O., &Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357.

Cormack, G. V., Smucker, M. D., & Clarke, C. L. (2011). Efficient and effective spam filtering and re-ranking for large web datasets. *Information retrieval*, 14(5), 441-465.

Choi, J. M. (2010). A selective sampling method for imbalanced data learning on support vector machines.

Escudero, G., Marquez, L., &Rigau, G. (2000, May). Boosting applied to word sense disambiguation. In *European Conference on Machine Learning* (129-141). Springer, Berlin, Heidelberg.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, 27(8), 861-874.

Greenhalgh, T. (2014). *How to read a paper: the basics of evidence-based medicine*. John Wiley & Sons.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). A review on ensembles for the class imbalance problem: bagging-, boosting-, and

hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* 42(4), 463-484.

Galar, M., Fernández, A., Barrenechea, E., & Herrera, F. (2013). EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. *Pattern recognition*, 46(12), 3460-3471.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9), 1263-1284.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1), 10-18.

Ian, H. W., & Eibe, F. (2005). Data mining: Practical machine learning tools and techniques.Morgan Kaufmann, 578.

Kallipolitis, L., Karpis, V., &Karali, I. (2012). Semantic search in the World News domain using automatically extracted metadata files. *Knowledge-Based Systems,* 27, 38-50.

Sackett, D. L., Rosenberg, W. M., Gray, J. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't.*BMJ*, 312(71).

Pang, B., Lee, L., &Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques.*arXiv preprint cs/0205070.*

Trieschnigg, D., Pezik, P., Lee, V., De Jong, F., Kraaij, W., &Rebholz-Schuhmann, D. (2009). MeSH Up: effective MeSH text classification for improved document retrieval. *Bioinformatics*, 25(11), 1412-1418.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2009). RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 40(1), 185-197.*

Liu, X. Y., Wu, J., & Zhou, Z. H. (2008). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2), 539-550.

Wu, Q., Ye, Y., Zhang, H., Ng, M. K., & Ho, S. S. (2014). ForesTexter: an efficient random forest algorithm for imbalanced text categorization. *Knowledge-Based Systems,* 67, 105-116.

National Library for Medicine.Retrieved from https://pubmed.ncbi.nlm.nih.gov/

PubMed Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); (2005). [Table, Stopwords] {www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T43/}

Vanwersch, R. J. B., Shahzad, K., Vanderfeesten, I., Vanhaecht, K., Grefen, P. W. P. J., Pintelon, L., &Reijers, H. A. (2013). *Methodological support for business process redesign in healthcare: a systematic literature review*. Beta Work Pap, 437, 20.

Webb, G. I. (2000). Multiboosting: A technique for combining boosting and wagging. *Machine learning*, 40(2), 159-196.

Weiss, G. M., & Provost, F. (2001). The effect of class distribution on classifier learning: an empirical study.

Wilczynski, N. L., & Haynes, R. B. (2009). Consistency and accuracy of indexing systematic review articles and meta-analyses in medline. *Health Information & Libraries Journal*, 26(3), 203-210.

Wang, S., & Yao, X. (2009, March). Diversity analysis on imbalanced data sets by using ensemble models. In 2009 *IEEE symposium on computational intelligence and data mining* (pp. 324-331). IEEE.