Biomedical Text Classification with Improved Feature Weighting Method

Duong B. Nguyen[1], Mohamed Shenify[2], and Hisham Al-Mubaid[1]
[1]University of Houston - Clear Lake, Houston, TX 77058, USA
[2]University of Baha, Baha, KSA

Abstract

In bioinformatics, we are interested in new techniques and advances in classification of biomedical documents for the hope of extracting useful biomedical knowledge out of the classification task. In this paper we introduce a feature weighting method for improving biomedical text classification. The method is effective in inducing weighted features from text data for classification. The weight of a feature is induced by the difference in class probability with versus without the feature. Specifically, in a simple inductive learning setting, the difference in class probability in the presence vs absence of feature fj can be a good metric for the contribution of fj in predicting the class. This technique is suitable for biomedical text mining in particular because it gives rise to terms with low per-document frequency and such terms play a good role in predicting the class in biomedical texts. The technique gives weight for each term feature based on the distribution of the class in presence vs absence of the term without considering the term frequency in each document. The evaluation is conducted using six biomedical text datasets and compared to the *tfidf* technique and baseline with encouraging results. We further examined the predictiveness of low average frequency terms and their effectiveness in classification accuracy.

1 Introduction

Text classification task, also known as text categorization, is the task of assigning a label or class label for text documents from a pre-defined set of class labels. Text classification is an active area of research in certain fields including bioinformatics [1, 2, 12, 15, 20, 25, 26, 27, 28]. In bioinformatics, text classification can lead to improvement in other tasks like gene-disease relationships extraction, understanding gene functions, biomedical knowledge discovery, and more [12, 20, 28]. In this paper, we propose a new technique for biomedical text classification based on calibrating feature weights of word features extracted from biomedical text corpora. The proposed technique is applied to *term* features extracted from the medical documents. When documents are modelled with *bag-of-word* then class prediction task can be manifested as an induction problem with a very high dimensional data. The proposed technique relies on the probability distribution of the class conditioned on the presence versus absence of each feature. This is basically the difference in the class probability before and after seeing the feature. This is a supervised term weighting scheme that allows for calibrating the weight of each feature, *i.e., term*, according to its contribution in each class leading to improving the learning and prediction. The method assigns a weight for each term based on its occurrence in a document without giving rise to how many times that term occurred in the document, *i.e., term frequency*. This makes it more appropriate for the biomedical text classification since the occurrence of a term in a given document is significant regardless of frequency of that term in that document. That is, in biomedical text mining, the occurrence of a term, even once, can be more significant than in the normal general document classification problem. Therefore, our method leverages the occurrence of each term in each document without much focus on the term frequency. We evaluated the method using six datasets of biomedical texts. The proposed technique produced encouraging results and proved to be effective for inducing fairly strong term features and encoding text documents as vectors of features. In the experimental results, our method outperformed the baseline in all class pairs of all six datasets. We also compared the performance of the method with *tfidf* which is well-known term weighting technique. We further examined our scheme with SVM and compared it with tfidf using fairly large text collection from *PubMed* and the results are also encouraging. This method is capable of extracting strong features from text information which will have positive impact and can lead to enhancing the performance of numerous applications that rely on text mining.

2 Related Work

With the high rate of growth of biomedical texts generated from medical publications and clinical trials, it is important to organize and arrange the biomedical information so that users can access the relevant information easily and quickly. In [18], Donaldson et al. (2015) collected the

BICOB 2016, April 4-6 2016, Las Vegas, Nevada, USA
978-1-943436-03-3 / Copyright ISCA

Medline abstracts and then applied 'bag-of-words' approach with SVM classifier to categorize abstracts containing information on protein–protein interactions, prior to curating this information into their BIND database. The classification system would reduce the number of abstracts that the curators needed to read by about two-thirds. The proposed method in [19] uses Probabilistic Latent Categorizer (PLC) with Kullback-Leibler divergence to re-rank documents related to curating information in *Swiss-Prot* database. In term of precision, the results improved 25-45 percent higher than PubMed ranking methods. Liu et al carried out a classification system for figures from the full-text biological papers to search the genes and protein interacting information existing in the figures [20].

Several methods were reported for feature weighting to be based on such as term frequency *TF*, inverse document frequency *IDF*, category concept and other concepts [1, 2, 12, 15, 20, 25, 26, 27,]. As an example, the weight of a feature in IDF-based weighting methods has an inverse relationship with the number of documents containing that feature. When the total number of documents containing a specific feature increases, the capability of that feature in discriminating classes decreases and so its weight also decreases. Although this is a right assumption in the information retrieval *IR* domain, it needs some modifications for being used in text categorization. When the number of documents containing a specific feature t_k increases and most of those documents belong to class C_j, then feature t_k is one of the powerful features for discriminating class C_j from the other classes. Hence, feature t_k should produce a high weight in class C_j.

As a dimensionality reduction method, feature selection is a technique that attempts to reduce the number of features in the learning task by removing all unnecessary and redundant features in the feature space [1, 2, 4]. In general feature reduction techniques can be roughly categorized into three types: Feature selection, feature clustering, and feature hashing; and feature selection is the most widely used among these three types [1, 2, 22, 25]. Feature selection and reduction techniques have been used extensively in many bioinformatics problems including gene selection, classification of microarray data, biomedical document clustering, prediction of gene and protein function, gene-protein name disambiguation, biomedical term disambiguation, biomedical WSD, and more [7, 8, 12]. Moreover, a number of natural language processing NLP applications have benefited directly from feature reduction including information retrieval and text categorization [7, 8].

3 Methods and Techniques

In text classification, we are interested in assigning a class label for a given text document from a set of predefined labels based on a learning (training) step conducted on a set of pre-labeled text documents, *i.e.*, supervised learning. In this paper we use *bag-of-word* model with vector space representation with learning and prediction using Support Vector Machines SVM [2, 5, 9]. The main contribution of this work is in the way we induce the feature weights for improving the predictive power of text classification techniques. The method can encode document-term features for effective and accurate learning and prediction (we use '*term*' and '*word*' interchangeably to refer the same thing). The feature weights are calibrated based on the probability distribution of the class, rather than feature distribution, conditioned on the presence versus absence of the feature, and for all features. This allows for calculating the difference in the class probability with versus without the feature, *i.e.*, the difference in class probability distribution with presence vs absence of the feature for each feature (*see equations 1 and 2*). The features are collected from the term occurrences in documents where each feature is basically a term.

In text classification, we use the *bag-of-word* model with vector space representation of documents for learning and classification [1, 2, 12, 15, 20, 25, 26, 27, 28]. Each document is represented as a numeric vector (feature vector) where each term is a component in that *n*-dimensional vector and *n* is the total number of terms in the text collection. Let the vector X_i represent document d_i then each term t_j in d_i can be represented in vector X_i in one of three ways: (1) using term frequency tf_{ij} (raw count) of t_j in document d_i. (2) using binary features 0,1 to indicate presence vs absence of the term in the document. (3) using the most commonly adopted term weighting technique *tfidf* (equations 7 and 8). The proposed term weighting technique in this paper is evaluated and compared with (1) and (3).

Let $F = \{f_i\}_{i=1..n}$ be the set of all features (terms) in a given text dataset. Each feature f_i represents a term t_i from the text corpora after preprocessing steps and *n* is the total number of terms in the text collection vocabulary. The preprocessing steps remove all stop-words (i.e., non-content words like *the, of,…etc.*) and convert words to roots (stemming). For a given feature f_i in a two-class classification $(C, \neg C)$ we calculate the weight w_i of feature f_i based on probability distribution of class C conditioned on the presence and absence of the feature as follows:

$$w_i = P(C \mid f_i) - P(C \mid \neg f_i) \quad \ldots\ldots\ldots (1)$$

and this can also be written as:

$$w_i = \frac{P(f_i, C)}{P(f_i)} - \frac{P(\neg f_i, C)}{P(\neg f_i)} \quad \ldots\ldots\ldots (2)$$

where $P(C \mid f_i)$ is probability distribution of class C conditioned on presence of feature f_i; and $P(C \mid \neg f_i)$ is probability of class C conditioned on absence of f_i; thus:

$$-1.0 \leq w_i \leq 1.0 \ldots\ldots\ldots\ldots (3)$$

Assuming we have *k* classes; through supervised learning process we induce *k* predictors, one for each class. Here we consider the two-class case. In equation (3) we have

$w_i \in [-1..1]$ implies that a feature f_i tends to incline to class C (resp. to class $\neg C$) as its weight w_i is approaching 1.0 (resp. –1.0). Let X_i be the numeric vector of document d_i such that

$$X_i = \{x_{ij}\}_{i=1..m}^{j=1..n}$$

where x_{ij} is the value of *jth* feature f_j in the vector X_i, n is the total number of features, and m is the total number of documents. Then, $P(C|f_j)$ can be depicted as:

$$P(C \mid f_j) = \frac{|X_i : X_i \in C \text{ and } x_{ij} \neq 0|}{|X_i : x_{ij} \neq 0|}$$

We use the term frequency tf_{ij} in the *baseline* method for un-weighted features where

$$tf_{ij} = \text{number of occurrences of term } t_j \text{ in document } d_i \quad \ldots\ldots (4)$$

then, feature weights in the baseline will be

$$x_{ij} = tf_{ij} \quad \ldots\ldots (5)$$

In the proposed method, we use:

$$x_{ij} = w_j . tf_{ij} \quad \ldots\ldots (6)$$

This proposed technique is suitable for document classification task because it is mainly based on feature appearance $\{P(C \mid f_i)\}$ versus absence $\{P(C \mid \neg f_i)\}$. Specifically, in the biomedical domain, the appearance of a word (feature) in a given document once is more significant that in the general text classification domain for general text documents.

4 Evaluation and Experiments

In the evaluation, we used six text datasets from the biomedical domain; see Table 1 and Table 2. We used LibSVM for learning and prediction with polynomial kernel [1, 5, 9] (*we also experimented with N. Bayes* [10, 11] *but SVM produced higher performance throughout all tests*). We examined the proposed feature weighting by comparing the prediction accuracy and AUC of the proposed technique with baseline and *tfidf* using the SVM and *10fcv*. The *tfidf* method is the most common approach for feature representation and for encoding word features in text mining (see section 5 for discussion on *tfidf*). We used 10-fold cross validation (*10fcv*) tallying the average of the ten runs of each experiment. We further examined the effectiveness of the low average frequency features (equation 9) in classification as the proposed technique gives leverage to such features.

Table 1: Six biomedical text datasets used in the evaluation

Dataset	Classes	Nmbr of documents
2007 Medical NLP Challenge	4	987
TREC 2006 Genomics Track	5	4,671
TextCATH.DX33	2	10,000
Farm	2	4143
BC3 – part 1 (2280 documents; 2 classes: 1140+1140)	2	2280
BC3 – part 2 (4000 documents; 2 classes:3318+ 682)	2	4000

Table 2: TREC 2006 dataset

Journal Name	No. of documents
Cerebral Cortex CC	917
Glycobiology GLY	719
Alcohol and Alcoholism AA	657
International Journal of Epidemiology IJE	1203
International Immunology II	1175

Datasets: The six datasets are summarized in Table 1 and Table 2. The first dataset is the 2007 *Medical NLP Challenge* dataset which is used for the clinical text classification task [21, 22]. Documents in the dataset are categorized using the ICD-9-CM codes [17]. ICD codes system is the international standard codes for classifying diseases. Each document can have multiple ICD codes [17]. To use the dataset for binary classification task, we separated and labeled the dataset according to distinct non-overlapping codes. The first set includes the documents which have a specific ICD code and the second set contains the remaining documents. In addition, the idea of labeling the dataset is described in [29]. The second dataset, *TREC 2006 Genomics Track*, contains a collection of full-text biomedical journal articles to answer the topic questions [30]. From a total of 49 journals and more than 162,259 documents we used articles from 5 journals; Table 2 shows the five journals and the number of documents from each journal. Dataset *TextCATH.DX33* is downloaded from [13] from the CATH project [12], and contains 10,000 document and two categories. Farm dataset contains 4143 documents and two categories obtained from UCI machine learning data repository [14]; also the last two dataset from UCI repository [14].

In the experiments, we used the *prediction accuracy* as evaluation metric which is the fraction of correctly classified documents to the total number of tested documents. We also used *Area Under ROC Curve* (AUC)

Table 3: Performance results of three weighting methods using SVM and 10fcv.

Dataset	Accuracy			AUC		
	Baseline	Tfifd	proposed	Baseline	Tfidf	proposed
2007 NLP	96.2	95.1	96.8	96.2	95.3	97.4
TREC 2006	92.6	92.5	95.9	91.8	91.6	97.9
TextCATH. DX33	79.1	74.9	86.0	79.8	76.9	87.5
Farm	88.8	89.5	92.6	89.3	89.4	94.8
BC3-p1	82.9	83.2	85.7	83.1	84.5	89.7
BC3 – p2	79.0	79.0	79.8	79.3	81.4	84.7

Table 4: Performance results for the feature weighting technique with the *2007 NLP Medical Challenge* dataset

Class pair	# of documents	Accuracy		
		Baseline	Tfidf	proposed
786_all	978	95.5	93.0	96.2
599_all	978	97.9	97.4	98.9
593_all	978	96.4	96.1	96.7
780_all	978	95.0	93.7	95.2
Average		96.2	95.1	**96.8**

in the evaluation. The receiver operating characteristics (ROC) is a curve of the false positive rate *FPR* (x-axis) and true positive rate *TPR* (y-axis); and the area under the ROC curve (AUC) has been used as a reliable indicator of the performance of a classifier or predictor and widely adopted in bioinformatics and machine learning. All the experiments are conducted with *all* word features and using *10fcv*. The evaluation results are summarized in Table 3 in terms of prediction accuracy and AUC for the six datasets using baseline (unweighted features), *tfidf*, and the proposed method. Table 4 and Table 5 contain the detailed results for *2007NLP* and *TREC-2006* datasets. Figures 1 and 2 illustrate the accuracy results of the 2007*NLP* and *TREC2006* datasets. In all datasets, the proposed method outperformed the baseline and *tfidf*, and sometimes with significant improvement. We did not include the detailed results for all classes of all datasets for the interest of space.

5 Discussion and Conclusions

In machine learning, converting the data into attributes is a crucial step that significantly affects the performance and behavior of the learner. In the baseline we use raw count (term frequency) of each term to represent that term in the document vector. We also examined the *tfidf* technique which is the most successful and most common term weighting technique in text classification. Our term weighting approach faired very well in the experiments and produced promising results in all experiments.

Effectively, feature weighting is more comprehensive and more powerful than feature selection. For example, in feature weighting we can assign the weight 0 to unselect features. As mentioned earlier, most of the research in this domain is focused on feature selection rather than feature weighting [8, 10, 11]. The most widely used feature weighting technique is the *tfidf* which can be stated as:

$$tfidf_{ij} = tf_{ij}.idf(f_j) \ldots\ldots\ldots (7)$$

$$idf(f_j) = log(N/df_j) \ldots\ldots\ldots (8)$$

and df_j is basically number of documents (document frequency) in which feature f_j occurs.

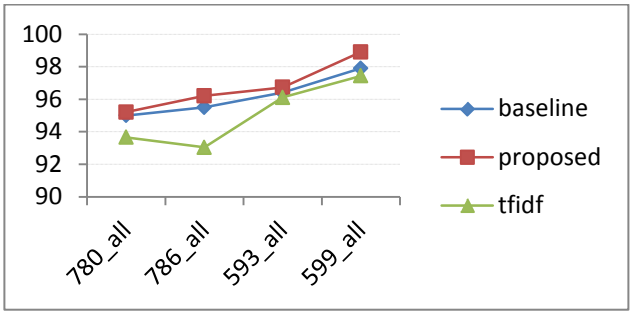

Figure 1: Accuracy of each pair in the NLP dataset.

Table 5: Performance results using *TREC-2006* data

Class pair	Accuracy		
	Baseline	Tfidf	proposed
CC_GLY	93.3	93.5	96.4
CC_AA	87.2	87.9	95.3
CC_IJE	95.7	95.6	96.9
CC_II	92.3	91.6	94.9
GLY_AA	94.2	94.9	96.4
GLY_IJE	96.8	96.7	97.9
GLY_II	95.6	95.9	97.4
AA_IJE	94.6	94.8	95.6
AA_II	86.4	85.5	92.0
IJE_II	90.3	88.4	95.9
Average	92.63	92.47	95.86

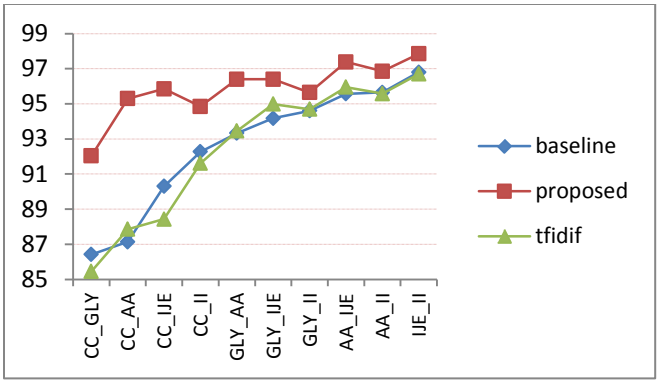

Figure 2: Accuracy of each pair in the *TREC-2006* dataset.

Using SVM, which is a state-of-the-art learner, *tfidf* did not produce any significant improvement over the baseline feature count; see Table 3, Table 4, and Table 5. The best performance of the proposed over the other two techniques is in TextCATH dataset (10,000 documents) with 10.6% higher in AUC than *tfidf* (Table 3); and on average the proposed technique achieved 7.1% (*micro-avg*) higher AUC than *tfidf* in all six datasets.

Contribution: Why does the proposed term weighting scheme work well? In the experimental evaluation, the proposed term weighting scheme performed fairly well compared with the raw term frequency and *tfidf* which is the most popular term weighting scheme in text mining. In text classification, if a term t_j occurs 10 times in a document d_i (*i.e.*, $tf_{ij} = 10$) our approach will account for this as 1. So every feature (term) is either 1 or 0 (*present or absent*). This is important in text mining of biomedical documents in particular because strong medical terms tend to occur with very low frequency and such terms have high predictive power. For example, if a medical term t_q occurs only once in a document d_i then our proposed scheme will give this term more contribution into the classification of d_i than the other two methods. Furthermore, two terms t_p and t_q with the same document frequency (*i.e.* $df(t_p) \sim= df(t_q)$) but ~90% of document containing t_p are in class *C* whereas 90% of the documents containing t_q are in class ¬C then these two terms will be exploited in the proposed term weighting more effectively than in *tfidf*. In *tfidf* both t_p and t_q will be weighted similarly. To confirm these points, we examined the proposed weighting scheme and *tfidf* with various experiments using documents downloaded from PubMed (http://www.ncbi.nlm.nih.gov/pubmed) [16] and using LibSVM. We wanted to check the effectiveness of utilizing low frequency words in biomedical texts. Specifically, we downloaded documents from PubMed using search terms like *Alzheimer disease*, *stem cells*, *hemoglobin*, *hypertension*, *gene function*, *Parkinson disease*, *epigenetics*, *carcinoma*, .etc. We created five experimental settings (see Table 6) where each setting contains two classes (*2 non-overlapping sets*) of documents form these downloads. We divided the documents into two non-overlapping sets to create two classes; the five settings (datasets) are shown in Table 6. We extracted all terms that have frequency of 3 or less in any document (we call it *frq3* setting). Firstly, we examined the classification accuracy using only *frq3* setting with *10fcv* and LibSVM; and the results are shown in Table 7. The average improvement of the proposed over *tfidf* is ~ 15% – 16% over text classification experiment involving 187,000 documents (Table 7). These results also are illustrated in Figure 3.

In another evaluation, we conducted text classification experiments on the same five datasets of biomedical documents (Table 6) with and without using terms that occur only once in any document (*i.e., terms occurring 0 or 1 in each document;* we call these *target* features) and the results are shown in Table 8. For example, as shown in Table 8, in the experimental setting *Exp1* (containing 2,000 documents and 2 classes) we found that there are 58 terms occur 0 or 1 in each of these 2000 documents; we ran classification task with and without these 58 terms (target features). As it shows in Table 8, in our scheme the accuracy improved from 69.9% without these 58 terms to 77.3% (7.4% improvement) with these terms which means that these terms improved the accuracy quite well; whereas in *tfidf* there is very slight improvement (0.3%) in accuracy when these 58 terms used in the classification.

Table 6: Five text collections downloaded from PubMed

Experimental setting	Total nmbr of documents	Percentage of class 1	Total number of features
Exp1	2,000	~50%	~960
Exp2	10,000	~50%	~1150
Exp3	25,000	~50%	~1220
Exp4	50,000	~50%	~1700
Exp5	100,000	~50%	~1800

Table 7: Accuracy results of two weighting methods using only terms having frequency ≤ 3 in any document (tested features are those terms that have at most three occurrences in any document).

Experimental setting	Total Nmbr of documents	Total Number of features	Number of tested features	Accuracy		
				Tfidf	Proposed	Imprv
Exp1	2,000	~960	129	63.8	77.3	13.5
Exp2	10,000	~1150	143	59.9	76.5	16.6
Exp3	25,000	~1220	166	54.3	76.7	22.4
Exp4	50,000	~1700	169	58.8	75.5	16.7
Exp5	100,000	~1800	188	59.1	72.5	13.4
			Micro-avg	58.5	74.1	15.6

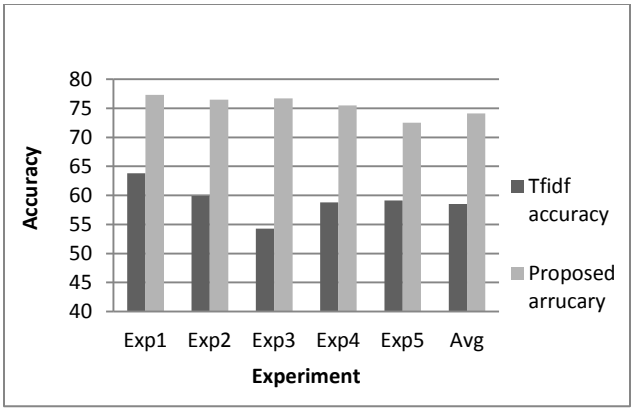

Figure 3: Illustration of the accuracy of the proposed and tfidf methods in 5 experiments with 10fcv using only having frequency ≤ 3 in any document.

Table 8: Classification accuracy using 10fcv and LibSVM on five medical datasets with and without target features. Target features are the terms that have frequency of 0 or 1 in any document.

Experimental setting	Total Nmbr of documents	Number of target features	Accuracy without – with target features	
			Tfidf	Proposed
Exp1	2,000	58	63.5 – 63.8	69.9 – 77.3
Exp2	10,000	66	59.8 – 59.9	69.4 – 76.5
Exp3	25,000	89	54.2 – 54.3	68.7 – 76.7
Exp4	50,000	108	58.8 – 58.8	69.1 – 75.5
Exp5	100,000	155	59.1 – 59.1	67.3 – 72.5

Table 9: Performance of two methods using top 10% and lowest 10% *atf* terms.

Experimental setting	Total Nmbr of documents	Accuracy *with top 10% atf terms*		Accuracy *with lowest 10% atf terms*	
		Tfidf	Proposed	Tfidf	Proposed
Exp4	50,000	50.8	68.7	32.4	67.6
Exp5	100,000	49.1	66.8	26.7	65.3

This indicates clearly that *tfidf* almost ignores the terms that have low occurrence frequency in any document but might have good documents frequency. We reviewed these terms that have low frequency in any document (but may occur in many documents) and found out that most of these terms are strong content-subject terms in the medical/biology fields and carry more meaningful weight than many other high frequency terms. In the last evaluation, we examined the effect of terms having low average frequency. We ordered all terms based on average term frequency *atf* as follows:

$$atf_j = \frac{\sum_i tf_{ij}}{df_j} \quad \ldots\ldots\ldots\ldots\ldots\ldots (9)$$

where tf_{ij} is the occurrence frequency of term t_j in document d_i, and df_j is the document frequency of term t_j. For example, term t_k appears in q documents ($df_k = q$) and $atf_k = 3.7$ then this means that term t_k occurs on average 3.7 times in these q documents. We experimented on the two settings Exp4 and Exp5 (the two largest settings) using the top 10% atf terms and the lowest 10% atf terms and the classification accuracy results are shown in Table 9. As we can see in Table 9, the proposed method performed fairly similar with top and lowest atf features indicating that the terms with really low average frequency can be as good as the terms with top frequency and that the proposed method utilized both of them. On the other hand, *tfidf* performed significantly worse with using only lowest *atf* terms.

For future work, we plan to investigate the concept of normalizing the weight based on the probability of the feature (feature frequency) as follows:

$$w(f_i) = \frac{P(C|f_i) - P(C|\neg f_i)}{1 - \frac{logP(f_i)}{2}}$$

Conclusion: The proposed technique is effective in calibrating feature weights of all term features in a given text classification task in the biomedical domain. We presented the experimental results of the proposed technique with six text datasets and the results show that the method is quite promising. We showed that the method outperformed the baseline and the *tfidf* technique on all datasets. With its competitive results, the method can have significant contribution onto other bioinformatics research problems that will benefit from document classification and text categorization.

References

[1]. V Pekar, M Krkoska, S Staab. "Feature Weighting for Co-occurrence-based Classification of Words". Proceedings of the 20th Conference on Computational Linguistics, COLING-2004

[2] George Forman, "An Extensive Empirical Study of Feature Selection Metrics for Text Classification". The ACM Journal of Machine Learning Research Volume 3, 2003, pp 1289-1305

[3] Y. Saeys, I. Inza, and P. Larranaga. A review of feature selection techniques in bioinformatics. Bioinformatics, Vol. 23 no.19, 2007.

[4] H. Al-Mubaid and S.A. Umair. A New Text Categorization Technique Using Distributional Clustering and Learning Logic. IEEE Trans on Knowledge and Data Eng. vol.18, no. 9, 2006.

[5] V. Vapnik, The Nature of Statistical Learning Theory. Springer, 1995

[6] H. Al-Mubaid, "Data mining and analysis of material data using feature clustering for radiation shielding", proc. of Int'l conf on Comp. and App. in Industry and Eng. CAINE-2010, Nov. 2010.

[7] H. Al-Mubaid and S. Gungu. A Learning Based Approach for Biomedical Word Sense Disambiguation. TSWJ journal, vol. 2012, PMID: 22666174; 2012.

[8] H. Al-Mubaid. A Learning-Classification Based Approach for Word Prediction. International Arab Journal on Information Technology IAJIT, Vol.4 No.3, July 2007.

[9] LibSVM http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[10] N.A. Zaidi, J. Cerquides, M.J. Carman, and G.I. Webb. Alleviating Naive Bayes Attribute Independence Assumption by Attribute Weighting. JMLR 14 (2013) 1947-1988.
[11] J. Wu and Z. Cai. Attribute Weighting via Differential Evolution Algorithm for Attribute Weighted Naive Bayes (WNB). Journal of Computational Information Systems 7:5 (2011) 1672-1679.
[12] A. Koussounadis, O.C. Redfern and D.T. Jones. Methodology article Open Access Improving classification in protein structure databases using text mining. BMC Bioinformatics 10:129, 2009.
[13] CATH project webpage: {retrieved October 2015} http://bioinfadmin.cs.ucl.ac.uk/downloads/textCATH/
[14] Lichman, M. (2013). UCI Machine Learning Repository Irvine, CA: University of California, School of Information and Computer Science https://archive.ics.uci.edu/ml/datasets.html
[15] K. Rajeswari, S. Nakil, N. Patil et al., "Text Categorization Optimization By A Hybrid Approach Using Multiple Feature Selection And Feature Extraction Methods". Int'l Journal of Engineering Research and Applications, Vol. 4, 2014.
[16] National Library of Medicine, NIH, Medline database, PubMed (http://www.ncbi.nlm.nih.gov/pubmed)
[17] International Classification of Diseases (ICD). World Health Organization. Archived from the original on 12 February 2014. Retrieved 14 March 2014.
[18] Donaldson, Ian et al. "PreBIND and Textomy – Mining the Biomedical Literature for Protein-Protein Interactions Using a Support Vector Machine." BMC Bioinformatics 4 (2003): 11. PMC. Web. 11 Nov. 2015.
[19] Dobrokhotov PB, Goutte C, Veuthey AL, Gaussier E. "A probabilistic information retrieval approach to medical annotation in SWISS-PROT". Stud Health Technol Inform. 2003
[20] Wang J, Bo TH, Jonassen I, Myklebost O, Hovig E. "Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data". BMC Bioinformatics. 2003
[21] K. Kira and L. Rendell. A practical approach to feature selection. Proceedings of 9th int'l workshop on Machine learning, pp.249–256, San Francisco, CA, USA, 1992.
[22] M. Bhasin and G.P. Raghava. 2004. ESLpred:SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic AcidsRes.32,W414-9.
[23] PlantLoc: an accurate web server for predicting plant protein subcellular localization by substantiality motif, Nucleic Acids Research, 2013, 1–7
[24] N.Y. Yu, J.R. Wagner, M.R. Laird, G. Melli, S. Rey, R. Lo, P. Dao, S.C. Sahinalp, M. Ester, L.J. Foster, F.S.L. Brinkman "PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes", Bioinformatics, (2010), 26 (13):1608-1615.
[25] M. Lan, C-L Tan, and H-B Low. Proposing a New Term Weighting Scheme for Text Categorization. AAAI 2006.
[26] Y. Yang and T. Joachims. "Text categorization". Scholarpedia, 2008, 3(5):4242.
[27] F. Sebastiani. "Machine learning in automated text categorization". ACM Computing Surveys, 34(1):1-47, 2002.
[28] Sasaki, Yutaka; Rea, Brian; Ananiadou, Sophia, "Multi-topic Aspects in Clinical Text Classification," in Bioinformatics and Biomedicine, 2007. BIBM 2007. IEEE International Conference on , vol., no., pp.62-70, 2-4 Nov. 2007
[29] Garla, Vijay N, and Cynthia Brandt. "Knowledge-Based Biomedical Word Sense Disambiguation: An Evaluation and Application to Clinical Document Classification." Journal of the American Medical Informatics Association : JAMIA 20.5 (2013): 882–886. PMC. Web. 1 Dec. 2015.
[30] W. Hersh, E. Voorhees. "TREC genomics special issue overview". Journal Information Retrieval. Volume 12 Issue 1, Springer February 2009 , p 1 - 15.