Imbalanced Text Categorization Based on Positive and Negative Term Weighting Approach

Behzad Naderalvojoud(✉), Ebru Akcapinar Sezer, and Alaettin Ucan

Computer Engineering Department, Hacettepe University, Ankara, Turkey
{n.behzad,ebru,aucan}@hacettepe.edu.tr,
http://humir.cs.hacettepe.edu.tr

Abstract. Although term weighting approach is typically used to improve the performance of text classification, this approach may not provide consistent results while imbalanced data distribution is available. This paper presents a probability based term weighting approach which addresses the different aspects of class imbalance problem in text classification. In this approach, we proposed two term evaluation functions called as PNF and PNF^2 which can produce more influential weights by relying on the imbalanced data sets. These functions can determine the significance of a term in association with a particular category. This is a crucial point because in one hand a frequent term is more important than a rare term in a particular category according to feature selection approach, and on the other hand a rare term is no less important than a frequent term based on *idf* assumption of traditional term weighting approach. Incorporation of these two approaches at the same time is the main idea that make them superior to other weighting methods. The achieved results from experiments which were carried out on two popular benchmarks (Reuters-21578 and WebKB) demonstrate that the probability based term weighting approach yields more consistent results than the other methods on the imbalanced data sets.

Keywords: Text classification · Class imbalance problem · Term weighting approach · Machine learning

1 Introduction

In text classification, class imbalance problem typically occurs when the number of documents of some classes is higher than the numbers of the others. In the imbalanced datasets, classes containing more number of instances are known as major classes while the ones having relatively less number of instances are called as minor classes. At this point, most of standard classifiers tend towards major classes and consequently show poorly performance on the minor classes. In other words, there should be as many examples belonging to major classes as examples belonging to minor ones [1,2]. This fundamental requirement cannot be always

P. Král and V. Matoušek (Eds.): TSD 2015, LNAI 9302, pp. 325–333, 2015.
DOI: 10.1007/978-3-319-24033-6_37

met and standard applications of machine learning algorithms may not provide satisfactory results. One of the effective approaches to resolve this problem which is also useful in text mining, is *term weighting strategy* via *tfidf* method [3]. *Tfidf* weighting is used to express how much a term can be important in a certain document while documents are represented in the Vector Space Model (VSM). In text classification, VSM is used to represent documents as term vectors and *tfidf* as a traditional term weighting scheme provides an influential solution for classification of imbalanced texts in many studies [4,5]. Debole and Sebastiani [6] proposed a number of supervised variant of *tfidf* weighting by replacing *idf* with feature selection metrics and presented a category based weighting scheme for classification task. In the other study [7] the supervised term weighting, *tf.rf*, was proposed based on distribution of relevant documents. The *rf* metric indicates the relevance level of a term with respect to a category. They evaluated *tf.rf* weighting scheme using SVM and kNN algorithms over different corpora and showed it consistently preforms well. In [8] a probability based term weighting scheme which can better distinguish documents in a minor category was introduced. In another one, [4] addressed the feature selection process for solving the class imbalance problem and took into consideration the abilities and characteristics of various metrics for feature selection. They asserted that negative features make a positive influence on the classification performance. In a more recent study, [5] explored the feature selection policies in text categorization by using SVM classifier.

In this study, we tackle the class imbalance problem using a probability based weighting scheme for a better multi-class classification task. Actually, two category based functions named as PNF and PNF^2 are proposed as a global component of term weighting scheme. These functions are based on two probabilities of relevant documents distribution. PNF^2 is designed as a two-sided function which takes into account either positive or negative terms. By this way, it can indicate either the type of term relevancy or the strength of relevancy (or irrelevancy) with respect to a specific category. Conversely, PNF is known as one-sided version of PNF^2 which can only determine the power of relevancy. In fact, we can distinguish documents better either in minor or major categories by replacing *idf* with the proposed category based metrics.

2 Term Weighting Approach

To better distinguish documents in the VSM, the term weighting approach is applied to represent documents. At first, traditional methods inspired by information retrieval are used for the purpose of term weighting. Their basic assumptions can be listed as follows: (1) "multiple appearances of a term in a document are no less important than single appearance" (*tf* assumption); (2) "rare terms are no less important than frequent terms" (*idf* assumption); (3) "for the same quantity of term matching, long documents are no more important than short documents" (*normalization* assumption) [6]. *Tfidf* as a standard weighting scheme has been used in many studies [4,7–9], because it provides an effective

solution for the classification of imbalanced texts by relying on these assumptions. It has been formulated in form of multiplying term frequency (tf) by inverse document frequency (idf). The common and normalized form of *tfidf* weighting are shown in Eq. 1 [3,10]:

$$tfidf(t_i, d_j) = tf(t_i, d_j) \times \log(\frac{N}{N_{t_i}}) \qquad w_{i,j} = \frac{tfidf(t_i, d_j)}{\sqrt{\sum_{k=1}^{|T|} tfidf(t_k, d_j)^2}} \quad (1)$$

where $tf(t_i, d_j)$ denotes the number of times that term t_i occurs in document d_j, N is the number of documents in the training set, N_{t_i} denotes the number of documents in which term t_i occurs at least once and $|T|$ denotes the number of unique terms. Actually, *tfidf* method is constituted from local and global principles. The frequency of a term within a specific document (tf) provides the local principle in the term weighting scheme and inverse document frequency (idf) supplies the global principle. Even if tf is used as a term weighting scheme alone, it can perform well [3,7,10]. On the other hand, idf is considered as an unsupervised function since it does not take into account the category membership in documents.

In text classification, if labeled documents are available, the term weighting approach which uses the prior known information can be applicable and named as supervised term weighting [6]. In this approach, metrics used in the term selection phase are replaced by the idf function, because the aim of term selection phase is to associate terms with each category. In fact, supervised approach uses category based term selection metrics as global component of term weighting scheme. In this study, we use the popular term selection metrics employed in [7] for supervised term weighting scheme. These metrics are represented by information elements in Table 1 (please see Table 2 for 'a', 'b', 'c' and 'd').

Table 1. Employed metrics as the global component of term weighting scheme in the experiments

Metric name	Formula
Chi square (X^2)	$N\frac{(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$
Information gain (ig)	$\frac{a}{N}log\frac{aN}{(a+c)(a+b)} + \frac{b}{N}log\frac{bN}{(b+d)(a+b)} + \frac{c}{N}log\frac{cN}{(a+c)(c+d)} + \frac{d}{N}log\frac{dN}{(b+d)(c+d)}$
Odds ratio (or)	$log\frac{ad}{bc}$
Relevance frequency (rf)	$log(2 + \frac{a}{max(1,c)})$

3 Proposed Positive and Negative Based Term Weighting Scheme

In the supervised functions, a one-sided function like rf or or only takes relevant terms that appear mostly in the given category into consideration, whereas two-sided function like X^2 or ig takes into account the irrelevant terms that do not

mostly appear in the given category, as well as relevant ones. In this study, a two-sided function (Eq. 2) is proposed for global component of term weighting scheme based on two probabilities of relevant documents; i.e. $P(t_i|C_j)$ which is known as the probability of documents from category C_j where term t_i occurs at least once and $P(t_i|\bar{C}_j)$ which is considered as the probability of documents not from category C_j where term t_i occurs at least once. The main idea is to specify the degree of being relevant or non-relevant for a term with respect to each category where the negative documents outnumber the positive ones. To achieve this, the difference between two probabilities is computed as shown in Eq. 2. In fact, if $P(t_i|C_j)$ is bigger than $P(t_i|\bar{C}_j)$, which basically indicates that term t_i is relevant to category C_j, then the term is labeled as a positive term associated with category C_j and otherwise is assumed as negative. By dividing the difference into the summation of two probabilities, the normalized values of weights are obtained and the weights are transformed to [-1, 1] interval.

In imbalanced cases, use of conditional probabilities plays an important role in the weighting process, because it creates a balanced situation between categories. It means that the document frequency for a certain term is computed based on its distribution over classes. Moreover, PNF^2 function can assign high weights to the terms, which are rare or frequent, according to their distributions in different classes. This is a crucial point because in one hand a frequent term is more important than a rare term in a particular category according to feature selection approaches, and on the other hand, a rare term is no less important than a frequent term based on *idf* assumption. We named the proposed function as PNF^2 which is the abbreviation of *Positive Negative Features* and power of 2 symbolizes that equation is designed as two-sided.

$$PNF^2(t_i, C_j) = \frac{P(t_i|C_j) - P(t_i|\bar{C}_j)}{P(t_i|C_j) + P(t_i|\bar{C}_j)} \tag{2}$$

To estimate the probabilities of Eq. 2, four information elements shown in Table 2 are used. In Table 2, C_j denotes the class corresponding to the j^{th} category in the dataset; t_i is the i^{th} term in the vocabulary set; $a_{i,j}$, $b_{i,j}$, $c_{i,j}$ and $d_{i,j}$ denote the document frequencies associated with the corresponding conditions. Therefore, the probabilities are calculated by using Eq. 3:

$$P(t_i|C_j) = \frac{a_{i,j}}{a_{i,j}+b_{i,j}} \qquad P(t_i|\bar{C}_j) = \frac{c_{i,j}}{c_{i,j}+d_{i,j}} \tag{3}$$

If PNF^2 is used as a global component of term weighting scheme, either positive or negative values are assigned to terms. When PNF^2 computes a

Table 2. Fundamental information elements which are used in feature selection functions

	Containing term t_i	Not containing term t_i
Belonging to class C_j	$a_{i,j}$	$b_{i,j}$
Not belonging to class C_j	$c_{i,j}$	$d_{i,j}$

negative value for a term, it shows not only the term is irrelevant for given category but also it has a negative effect for that category as much as its absolute value. To eliminate the negative effect, the one-sided form of PNF^2 (Eq. 4) is defined as another alternative for the global component of term weighting scheme. In fact, we transform PNF^2 to one-sided function abbreviated as PNF and compare it with the performance of PNF^2.

$$PNF = 1 + PNF^2 \tag{4}$$

PNF function does not produce any negative weights for terms and it assigns just low positive values to non-relevant terms instead of negative. Thus, PNF function does not transform the trend of weighting to the negative space. Since the weighting scheme is employed for only training data, this approach becomes plausible.

4 Empirical Observation of Term Weighting and Feature Selection Approaches

In this part, we try to make a comparative explanation by using a realistic example. First, the scores of terms in the *grain* category of Reuters dataset are calculated by using two popular feature selection metrics i.e. ig, X^2 and proposed PNF metrics; then the scores of terms are sorted in descending order to select top 4 terms of each metric. Actually, *grain* is a minor category with 41 documents and Table 3 lists a, c and idf values of the selected top 4 terms. At this point, we want to emphasize the differences between feature selection and term weighting approaches. Feature selection means the identification of more representative terms, and selected features should represent the most number of documents. As a result, they ignore rare terms. On the other hand, a term weighting scheme which uses idf as a global component, gives higher score to terms with low document frequency. As can be seen from Table 3, idf values of terms selected by PNF are higher than the idf values of other terms. The difference between term weighting and feature selection approaches can be obviously proven with c values. Although, most of terms selected by ig and X^2 metrics have high document frequency in non-*grain* categories (i.e. high c values), terms selected by PNF metric have 0 values for the c parameter. Since use of feature selection metrics for category based weighting purposes has been preferred in the previous studies [6–8], we have to evaluate our proposed metrics by comparing with them. The last point is that, proposed PNF metric has closer approach to idf than the others, but unlike idf, PNF is proposed for category based weighting.

5 Experiments

In this study, all experiments were conducted on two different benchmarks such as Reuters-21578 and WebKB. The Reuters-21578 dataset has been widely used in text classification researches as an imbalanced collection [6,8,9,11,12]. The R8

Table 3. The characteristics of top 4 terms selected by different manners for *grain* category in Reuters-21578 dataset

Terms	X^2			IG			PNF		
	a	c	idf	a	c	idf	a	c	idf
t_1	36	15	6.75	36	15	6.75	14	0	8.61
t_2	14	0	8.61	24	52	6.17	3	0	10.84
t_3	11	5	8.42	14	0	8.61	3	0	10.84
t_4	24	52	6.17	11	5	8.42	3	0	10.84

version of Reuters dataset which was used in the experiments [13], consists of two major categories called as *earn* and *acq* with almost 52% and 30% class distributions respectively and 6 minor categories with almost 3% class distributions. WebKB dataset consists of four categories of web pages collected from computer science departments of four universities[13]. This dataset contains two minor categories called as *project* and *course* with almost 10% and 20% class distributions respectively and two major categories with 30% and 40% class distributions. For both datasets, experiments were performed on the original training and test sets obtained from benchmarks [13]. Standard text preprocessing steps were applied and all features were used in classification.

To analyze the effect of different weighting methods on the imbalanced data classification problem, we need a simple classifier which makes term weighting scheme as the most effective factor in the learning process. In the proposed classifier algorithm which is inspired by Rocchio, after representing documents in VSM and applying a term weighting scheme, learning process is realized by combining training document vectors $\vec{d}$ into a vector $\vec{c}_j$ for each category. The vector $\vec{c}_j$ is computed for category C_j by dot dividing of two vectors as Eq. 5:

$$\vec{c}_j = \frac{1}{\vec{a}_j} \sum_{\vec{d} \in C_j} \vec{d} \tag{5}$$

In Eq. 5 the $\vec{a}_j$ is the vector yielded from the document frequency of terms with respect to category C_j (as shown in Table 2) and $\sum_{\vec{d} \in C_j} \vec{d}$ yields the summation of document vectors which belong to category C_j. Consequently, the set of $\vec{c}_j$ vectors which are computed for each category, represent the learned model. This model is used to classify document d^t which has never seen before. This test document is represented by the vector $\vec{d}^t$ which has only tf values as weights. In order to classify the test document, cosine similarity is computed between two vectors such as $\vec{d}^t$ and each of $\vec{c}_j$. Finally, the vector $\vec{d}^t$ is assigned to the category which has the highest similarity with $\vec{d}^t$ as indicated in Eq. 6.

$$F(\vec{d}^t) = \arg\max_{c_j \in C} \frac{\vec{c}_j}{||\vec{c}_j||} \cdot \frac{\vec{d}^t}{||\vec{d}^t||} \tag{6}$$

Precision (P), Recall (R) and F-measure were used to evaluate the performance of classification.

Table 4. The F-measure values of different term weighting schemes for Reuters-21578 dataset

Categories	The term weighting schemes						
	$tf.idf$	$tf.X^2$	$tf.ig$	$tf.or$	$tf.rf$	$tf.PNF^2$	$tf.PNF$
earn	0.771	0.512	0.845	0.945	**0.981**	0.950	**0.981**
acq	0.450	0.654	0.831	0.921	0.957	0.952	**0.961**
crude	0.698	0.896	0.887	0.867	0.902	0.835	**0.945**
trade	0.542	0.867	0.886	0.771	**0.906**	0.802	0.898
money-fix	0.646	0.789	0.781	0.798	0.719	0.834	**0.868**
interest	0.754	0.792	0.779	0.852	0.776	0.838	**0.881**
ship	0.539	0.831	0.679	0.781	0.806	0.794	**0.845**
grain	0.667	0.889	0.889	**0.900**	0.800	0.750	**0.900**
Macro Average	0.633	0.779	0.822	0.854	0.856	0.844	**0.910**
Micro average	0.687	0.639	0.836	0.912	0.945	0.925	**0.958**

Achieved F-measure values for the different weighting methods employed in Reuters-21578 benchmark are listed in Table 4. As can be seen, $tf.PNF$ term weighting method consistently outperforms all other methods for all categories except one case.The results obtained from $tf.PNF^2$ can be competitive with the other methods. The $tf.PNF$ weighting scheme, which eliminates the negative impact considered in $tf.PNF^2$, significantly improves the performance of the classification. The superiority of $tf.PNF$ scheme can also be seen by micro and macro averaged F-measure values. Another point is that the $tfidf$ weighting scheme cannot provide a good distinction between categories and consequently performs weakly on the whole categories.

In WebKB benchmark, the superiority of $tf.PNF^2$ and $tf.PNF$ can be observed among the other methods as shown in Table 5. Although the $tf.PNF$ is known as the best weighting scheme by possessing the highest micro and macro averaged F-measure values, $tf.PNF^2$ gives better results for minor categories. It can be also observed that the performance $tf.ig$, $tf.X^2$ and $tf.rf$ are degraded in contrast with their previous results on the Reuters benchmark and cannot keep their relative goodness. At this point, it can be said that they cannot perform well on different imbalanced circumstances and may not yield consistent results. Conversely, $tf.PNF$, $tf.PNF^2$ and $tf.or$ can provide more reliable results since they can make a relative minimum range of fluctuation in their results.

According to the achieved results from two benchmarks (Tables 4 and 5), the proposed two functions as a global component of term weighting scheme yield better results than the others. Moreover, the category based term weighting schemes outperform the traditional $tfidf$ in most cases. In other words, $tfidf$ cannot make any clear distinction between documents of the different classes in multi-class classification task. As mentioned in section 4, ig and X^2 are successful for feature selection task [5] but they cannot consistently perform well as a global component of term weighting scheme in imbalanced text classification.

Table 5. The F-measure values of different term weighting schemes for WebKB dataset

Categories	The term weighting schemes						
	$tf.idf$	$tf.X^2$	$tf.ig$	$tf.or$	$tf.rf$	$tf.PNF^2$	$tf.PNF$
student	0.636	0.587	0.588	0.636	0.735	0.705	**0.852**
faculty	0.372	0.236	0.224	0.688	0.673	0.750	**0.757**
course	0.608	0.014	0.006	0.859	0.662	**0.887**	0.860
project	0.088	0.403	0.424	**0.649**	0.443	**0.649**	0.617
Macro Average	0.426	0.310	0.311	0.708	0.628	0.747	**0.772**
Micro average	0.549	0.452	0.454	0.703	0.683	0.749	**0.805**

To determine the statistical significance of the results, we performed ANOVA test on the F-measure values gained by the methods for categories. According to results of ANOVA test for Reuters-21578 and WebKB benchmarks, since the P-value of the test is less than 0.05 for each case (P-value equals 0.0000 for Reuters and 0.0028 for WebKB), there are statistically significant differences between the macro-averaged F-measure values of $tf.PNF$ with the other different schemes at the 95.0% confidence level.

6 Conclusion

In this study, we tackled the class imbalance problem by category based term weighting approach and PNF^2 and PNF were proposed as a global component of term weighting scheme based on the probabilities of relevant documents frequency. Experiments were made with several methods on two different benchmarks. According to our findings, the $tf.PNF$ term weighting scheme is the best in all experiments and can provide the best tradeoff between precision and recall. Despite the wide range of fluctuation in the results of $tf.ig$ and $tf.X^2$, $tf.PNF^2$ as a two-sided method achieves more expectable results with high F-measure values. Additionally, one-sided functions (i.e. or, rf and PNF) consistently perform better than the two-sided ones (i.e. ig and X^2), however, PNF^2 presents competitive results in contrast with or, rf functions. As a result, the PNF and PNF^2 functions as a global component of term weighting scheme are recommended for imbalanced classification task.

References

1. Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. Intelligent Data Analysis **6**(5), 429–449 (2002)
2. Chawla, N.V., Japkowicz, N., Kotcz, A.: Editorial: special issue on learning from imbalanced data sets. ACM Sigkdd Explorations Newsletter **6**(1), 1–6 (2004)
3. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management **24**(5), 513–523 (1988)
4. Ogura, H., Amano, H., Kondo, M.: Comparison of metrics for feature selection in imbalanced text classification. Expert Systems with Applications **38**(5), 4978–4989 (2011)

5. Taşcı, Ş., Güngör, T.: Comparison of text feature selection policies and using an adaptive framework. Expert Systems with Applications **40**(12), 4871–4886 (2013)
6. Debole, F., Sebastiani, F.: Supervised term weighting for automated text categorization. In: Sirmakessis, S. (ed.) Text Mining and its Applications. STUDFUZZ, vol. 138, pp. 81–97. Springer, Heidelberg (2004)
7. Lan, M., Tan, C.L., Su, J., Lu, Y.: Supervised and traditional term weighting methods for automatic text categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(4), 721–735 (2009)
8. Liu, Y., Loh, H.T., Sun, A.: Imbalanced text classification: A term weighting approach. Expert Systems with Applications **36**(1), 690–701 (2009)
9. Ren, F., Sohrab, M.G.: Class-indexing-based term weighting for automatic text classification. Information Sciences **236**, 109–125 (2013)
10. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys (CSUR) **34**(1), 1–47 (2002)
11. Sun, A., Lim, E.P., Liu, Y.: On strategies for imbalanced text classification using svm: A comparative study. Decision Support Systems **48**(1), 191–201 (2009)
12. Erenel, Z., Altınçay, H.: Nonlinear transformation of term frequencies for term weighting in text categorization. Engineering Applications of Artificial Intelligence **25**(7), 1505–1514 (2012)
13. Cachopo, A.M.d.J.C.: Improving Methods for Single-label Text Categorization. PhD thesis, Universidade Técnica de Lisboa (2007)