Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications xxx (2007) xxx–xxx
Expert Systems with Applications
www.elsevier.com/locate/eswa

Imbalanced text classification: A term weighting approach

Ying Liu [a,*], Han Tong Loh [b], Aixin Sun [c]

[a] Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China
[b] Department of Mechanical Engineering, National University of Singapore, 9 Engineering Drive 1, Singapore 117576, Singapore
[c] School of Computer Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798, Singapore

Abstract

The natural distribution of textual data used in text classification is often imbalanced. Categories with fewer examples are under-represented and their classifiers often perform far below satisfactory. We tackle this problem using a simple probability based term weighting scheme to better distinguish documents in minor categories. This new scheme directly utilizes two critical information ratios, i.e. relevance indicators. Such relevance indicators are nicely supported by probability estimates which embody the category membership. Our experimental study using both Support Vector Machines and Naïve Bayes classifiers and extensive comparison with other classic weighting schemes over two benchmarking data sets, including Reuters-21578, shows significant improvement for minor categories, while the performance for major categories are not jeopardized. Our approach has suggested a simple and effective solution to boost the performance of text classification over skewed data sets.

Keywords: Text classification; Imbalanced data; Term weighting scheme

1. Introduction

1.1. Motivation

Learning from imbalanced data has emerged as a new challenge to the machine learning (ML), data mining (DM) and text mining (TM) communities. Two recent workshops in 2000 (Japkowicz, 2000) and 2003 (Chawla, Japkowicz, & Kolcz, 2003) at AAAI and ICML conferences, respectively and a special issue in ACM SIGKDD explorations (Chawla, Japkowicz, & Kolcz, 2004) were dedicated to this topic. It has been witnessing growing interest and attention among researchers and practitioners seeking solutions in handling imbalanced data. An excellent review of the state-of-the-art is given by Weiss (2004).

The data imbalance problem often occurs in classification and clustering scenarios when a portion of the classes possesses many more examples than others. As pointed out by Chawla et al. (2004) when standard classification algorithms are applied to such skewed data, they tend to be overwhelmed by the major categories and ignore the minor ones. There are two main reasons why the uneven cases happen. One is due to the intrinsic nature of such events, e.g. credit fraud, cancer detection, network intrusion, and earthquake prediction (Chawla et al., 2004). These are rare events presented as a unique category but only occupy a very small portion of the entire example space. The other reason is due to the expense of collecting learning examples and legal or privacy reasons. In our previous study of building a manufacturing centered technical paper corpus (Liu & Loh, 2007), due to the costly efforts demanded for human labeling and diverse interests in the papers, we ended up naturally with a skewed collection.

Automatic text classification (TC) has recently witnessed a booming interest, due to the increased availability of documents in digital form and the ensuing need to organize them (Sebastiani, 2002). In TC tasks, given that most test collections are composed of documents belonging to multiple classes, the performance is usually reported in terms of micro-averaged and macro-averaged scores (Sebastiani,

* Corresponding author. Tel.: +852 34003782.
E-mail address: mfyliu@polyu.edu.hk (Y. Liu).

0957-4174/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2007.10.042

2002; Yang & Liu, 1999). Macro-averaging gives equal weights to the scores generated from each individual category. In comparison, micro-averaging tends to be dominated by the categories with more positive training instances. Due to the fact that many of these test corpora used in TC are either naturally skewed or artificially imbalanced especially in the binary and so called "one-against-all" settings, classifiers often perform far less than satisfactorily for minor categories (Lewis, Yang, Rose, & Li, 2004; Sebastiani, 2002; Yang & Liu, 1999). Therefore, micro-averaging mostly yields much better results than macro-averaging does.

1.2. Related work

There have been several strategies in handling imbalanced data sets in TC. Here, we only focus on the approaches adopted in TC and group them based on their primary intent. The first approach is based on sampling strategy. Yang (1996) has tested two sampling methods, i.e. proportion-enforced sampling and completeness-driven sampling. Her empirical study using the ExpNet system shows that a global sampling strategy which favors common categories over rare categories is critical for the success of TC based on a statistical learning approach. Without such a global control, the global optimal performance will be compromised and the learning efficiency can be substantially decreased. Nickerson, Japkowicz, and Milios (2001) provide a guided sampling approach based on a clustering algorithm called Principal Direction Divisive Partitioning to deal with the between-class imbalance problem. It has shown improvement over existing methods of equalizing class imbalances, especially when there is a large between-class imbalance together with severe imbalance in the relative densities of the subcomponents of each class. Liu's recent efforts (Liu, 2004) in testing different sampling strategies, i.e. under-sampling and over-sampling, and several classification algorithms, i.e. Naïve Bayes, k-Nearest Neighbors (kNN) and Support Vector Machines (SVMs), improve the understanding of interactions among sampling method, classifier and performance measurement.

The second major effort emphasizes cost sensitive learning (Dietterich, Margineantu, Provost, & Turney, 2000; Elkan, 2001; Weiss & Provost, 2003). In many real scenarios like risk management and medical diagnosis, making wrong decisions are usually associated with very different costs. A wrong prediction of the nonexistence of cancer, i.e. false negative, may lead to death, while the wrong prediction of cancer existence, i.e. false positive, only results in unnecessary anxiety and extra medical tests. In view of this, assigning different cost factors to false negatives and false positives will lead to better performance with respect to positive (rare) classes (Chawla et al., 2004). Brank, Grobelnik, Milic-Frayling, and Mladenic (2003) have reported their work on cost sensitive learning using SVMs on TC. They obtain better results with methods that directly modify the score threshold. They further propose a method based on the conditional class distributions for SVM scores that works well when only very few training examples are available.

The recognition based approach, i.e. one-class learning, has provided another class of solutions (Japkowicz, Myers, & Gluck, 1995). One-class learning aims to create the decision model based on the examples of the target category alone, which is different from the typical discriminative approach, i.e. the two classes setting. Manevitz and Yousef (2002) have applied one-class SVMs on TC. Raskutti and Kowalczyk (2004) claim that one-class learning is particularly helpful when data are extremely skewed and composed of many irrelevant features and very high dimensionality.

Feature selection is often considered an important step in reducing the high dimensionality of the feature space in TC and many other problems in image processing and bioinformatics. However, its unique contribution in identifying the most salient features to boost the performance of minor categories has not been stressed until some recent work (Mladenic & Grobelnik, 1999). Yang and Pedersen (1997) has given a detailed evaluation of several feature selection schemes. We noted the marked difference between micro-averaged and macro-averaged values due to the poor performances over rare categories. Forman (2003) has done a very comprehensive study of various schemes for TC on a wide range of commonly used test corpora. He has recommended the best pair among different combinations of selection schemes and evaluation measures. The recent efforts from Zheng, Wu, and Srihari (2004) advance the understanding of feature selection in TC. They show the merits and great potential of explicitly combining positive and negative features in a nearly optimal fashion according to the imbalanced data.

Some recent work simply adapting existing machine learning techniques and not even directly targeting the issue of class imbalance have shown great potential with respect to the data imbalance problem. Castillo and Serrano (2004) and Fan, Yu, and Wang (2004) have reported the success using an ensemble approach, e.g. voting and boosting, to handle skewed data distribution. Challenged by real industry data with a huge number of records and an extremely skewed data distribution, Fan's work shows that the ensemble approach is capable of improving the performance on rare classes. In their approaches, a set of weak classifiers using various learning algorithms are built up over minor categories. The final decision is reached based on the combination of outcomes from different classifiers. Another promising approach which receives less attention falls into the category of semi-supervised learning or weakly supervised learning (Blum & Mitchell, 1998; Ghani, 2002; Goldman & Zhou, 2000; Lewis & Gale, 1994; Liu, Dai, Li, Lee, & Yu, 2003; Nigam, 2001; Yu, Zhai, & Han, 2003; Zelikovitz & Hirsh, 2000). The basic idea is to identify more positive examples from a large amount of unknown data. These approaches are especially

Please cite this article in press as: Liu, Y. et al., Imbalanced text classification: A term weighting approach, *Expert Systems with Applications* (2007), doi:10.1016/j.eswa.2007.10.042

viable when unlabeled data are steadily available. The last effort attacking the imbalance problem uses parameter tuning in kNNs (Baoli, Qin, & Shiwen, 2004). The authors expect to set k dynamically according to the data distribution, in which a large k is granted given a minor category.

In this paper, we tackle the data imbalance problem in text classification from a different angle. We present a new approach assigning better weights to the features from minor categories. After a brief review of the classic term weighting scheme, e.g. *tfidf*, in Section 2 and inspired by the analysis of various feature selection methods in Section 3, we introduce a simple probability based term weighting scheme which directly utilizes two critical information ratios, i.e. relevance indicators, in Section 4. These relevance indicators are nicely supported by the probability estimates which embody the category membership. The setup of experimental study is explained in Section 5. We carry out the evaluation and comparison of our new scheme with many other different weighting forms over two skewed data sets. We report the experimental findings and discuss their performance in Section 6. Section 7 concludes as well as highlights some future work.

2. Term weighting scheme

Text classification (TC) is such a task to categorize documents into predefined thematic categories. In particular, it aims to find the mapping ξ, from a set of documents D: $\{d_1, \ldots, d_i\}$ to a set of thematic categories C: $\{C_1, \ldots, C_j\}$, i.e. $\xi : D \to C$. In its current practice, which is dominated by supervised learning, the construction of a text classifier is often conducted in two main phases (Debole & Sebastiani, 2003; Sebastiani, 2002):

- Document indexing – the creation of numeric representations of documents:
  - Term selection – to select a subset of terms from all terms occurring in the collection to represent the documents in a better way, either to faster computing or to achieve better effectiveness in classification.
  - Term weighting – to assign a numeric value to each term to weight its contribution which helps a document stand out from others.
- Classifier induction – the building of a classifier by learning from the numeric representations of documents.

In information retrieval and machine learning, term weighting has long been formulated in a form as term frequency times inverse documents frequency, i.e. *tfidf* (Baeza-Yates & Ribeiro-Neto, 1999; Salton & Buckley, 1988; Salton & McGill, 1983; van-Rijsbergen, 1979). The more popular "ltc" form (Baeza-Yates & Ribeiro-Neto, 1999; Salton & Buckley, 1988; Salton & McGill, 1983) is given by

$$tfidf(t_i, d_j) = tf(t_i, d_j) \times \log\left(\frac{N}{N(t_i)}\right) \tag{1}$$

and its normalized version is

$$w_{i,j} = \frac{tfidf(t_i, d_j)}{\sqrt{\sum_{k=1}^{|T|} tfidf(t_k, d_j)^2}}, \tag{2}$$

where N and $|T|$ denote the total number of documents and unique terms contained in the collection, respectively, and $N(t_i)$ represents the number of documents in the collection in which term t_i occurs at least once, and

$$tf(t_i, d_j) = \begin{cases} 1 + \log(n(t_i, d_j)), & \text{if } n(t_i, d_j) > 0, \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where $n(t_i, d_j)$ is the number of times that term t_i occurs in document d_j. In practice, the summation in Eq. (2) is only concerned about the terms occurred in document d_j.

The significance of the classic term weighting schemes in Eqs. (1) and (2) is that they have embodied three fundamental assumptions of term frequency distribution in a collection of documents (Debole & Sebastiani, 2003; Sebastiani, 2002). These assumptions are:

- Rare terms are no less important than frequent terms – *idf* assumption.
- Multiple appearances of a term in a document are no less important than single appearance – *tf* assumption.
- For the same quantity of term matching, long documents are no less important than short documents – normalization assumption.

Because of these, the "ltc" and its normalized form have been extensively studied by many researchers and show their good performance over a number of different data sets (Sebastiani, 2002). Therefore, they have become the default choice in TC.

3. Inspiration from feature selection

Feature selection serves as a key procedure to reduce the dimensionality of input data space to save computational cost. It has been integrated as a default step for many learning algorithms, like artificial neuron network, k-nearest neighbors, decision tree, etc. In the research community of machine learning, the computation constraints imposed by the high dimension of the input data space and the richness of information available to maximally identify each individual object is a well known tradeoff. The ability of feature selection to capture the salient information by selecting the most important attributes, and thus making the computing tasks tractable has been shown in information retrieval and machine learning research (Forman, 2003; Ng, Goh, & Low, 1997; Ruiz & Srinivasan, 2002; Yang & Pedersen, 1997). Furthermore, feature selection is also beneficial since it tends to reduce the over-fitting problem, in which the trained objects are tuned to fit very well the data upon which they have been built, but performs poorly when applied to unseen data (Sebastiani, 2002).

Please cite this article in press as: Liu, Y. et al., Imbalanced text classification: A term weighting approach, *Expert Systems with Applications* (2007), doi:10.1016/j.eswa.2007.10.042

Table 1
Several feature selection methods, and their functions

Feature selection method	Mathematical form
Information gain	$P(t_k,c_i)\log\frac{P(t_k,c_i)}{P(t_k)\cdot P(c_i)}+P(\bar{t}_k,c_i)\log\frac{P(\bar{t}_k,c_i)}{P(\bar{t}_k)\cdot P(c_i)}$
Mutual information	$\log\frac{P(t_k,c_i)}{P(t_k)P(c_i)}$
Chi-square	$\frac{N\cdot[P(t_k,c_i)\cdot P(\bar{t}_k,\bar{c}_i)-P(t_k,\bar{c}_i)\cdot P(\bar{t}_k,c_i)]^2}{P(t_k)\cdot P(\bar{t}_k)\cdot P(c_i)\cdot P(\bar{c}_i)}$
Odds ratio	$\log\frac{P(t_k\|c_i)\cdot(1-P(t_k\|\bar{c}_i))}{(1-P(t_k\|c_i))\cdot P(t_k\|\bar{c}_i)}$

t_k denotes a term; c_i stands for a category; $P(t_k,c_i)$ denotes the probability of documents from category c_i where term t_k occurs at least once; $P(t_k,\bar{c}_i)$ denotes the probability of documents not from category c_i where term t_k occurs at least once; $P(\bar{t}_k,c_i)$ denotes the probability of documents from category c_i where term t_k does not occur; $P(\bar{t}_k,\bar{c}_i)$ denotes the probability of documents not from category c_i where term t_k does not occur.

In TC, several feature selection methods have been intensively studied to distill the important terms while still keeping the dimension small. Table 1 shows the main functions of several popular feature selection methods. These methods are evolved either from the information theory or from the linear algebra literature (Sebastiani, 2002; Yang & Pedersen, 1997).

Basically, there are two distinct ways to rank and assess the features, i.e. globally and locally. Global feature selection aims to select features which are good across all categories. Local feature selection intends to differentiate those terms that are more distinguishable for certain categories only. The sense of either 'global' or 'local' does not have much effect on the selection of method itself, but it does affect the performance of classifiers built upon different categories. In TC, the main purpose is to address whether document belongs to a specific category. Obviously, we prefer the salient features which are unique from one category to another, i.e. a 'local' approach. Ideally, the salient feature set from one category does not have any items overlapping with those from other categories. If this cannot be avoided, then how to better present them has become an issue.

While many previous works have shown the relative strengths and merits of these methods (Forman, 2003; Ng et al., 1997; Ruiz & Srinivasan, 2002; Sebastiani, 2002; Yang & Pedersen, 1997), our experience with feature selection over a number of standard or ad hoc data sets shows the performance of such methods can be highly dependent on the data. This is partly due to the lack of understanding of different data sets in a quantitative way, and it needs further research. From our previous study of all feature selection methods and what has been reported in the literature (Yang & Pedersen, 1997), we noted when these methods are applied to text classification for term selection purpose, they are basically utilizing four fundamental information elements shown in Table 2.

Table 2
Fundamental information elements used for feature selection in text classification

	c_i	$\bar{c}_i$
t_k	A	B
$\bar{t}_k$	C	D

A denotes the number of documents belonging to category c_i where the term t_k occurs at least once; B denotes the number of documents not belonging to category c_i where the term t_k occurs at least once; C denotes the number of documents belonging to category c_i where the term t_k does not occur; D denotes the number of documents not belonging to category c_i where the term t_k does not occur.

Table 3
Feature selection methods and their formations as represented by information elements in Table 2

Method	Mathematical form represented by information elements
Information gain	$-\frac{A+C}{N}\log\frac{A+C}{N}+\frac{A}{N}\log\left(\frac{A}{A+B}\right)+\frac{C}{N}\log\left(\frac{C}{C+D}\right)$
Mutual information	$\log(AN/(A+B)(A+C))$
Chi-square	$N(AD-BC)^2/(A+C)(B+D)(A+B)(C+D)$
Odds ratio	$\log(AD/BC)$

These four information elements have been used to estimate the probability listed in Table 1. Table 3 shows the functions in Table 1 as presented by these four information elements A, B, C and D.

4. A probability based term weighting scheme

4.1. Revisit of tfidf

As stated before, while many researchers believe that term weighting schemes in the form as *tfidf* representing those three aforementioned assumptions, we understand *tfidf* in a much simpler manner, i.e.

- Local weight – the *tf* term, either normalized or not, specifies the weight of t_k within a specific document, which is basically estimated based on the frequency or relative frequency of t_k within this document.
- Global weight – the *idf* term, either normalized or not, defines the contribution of t_k to a specific document in a global sense.

If we temporarily ignore how *tfidf* is defined, and focus on the core problem, i.e. whether this document is from this category, we realize that a set of terms is needed to represent the documents effectively and a reference framework is required to make the comparison possible. As previous research shows that *tf* is very important (Leopold & Kindermann, 2002; Salton & Buckley, 1988; Sebastiani, 2002) and using *tf* alone can already achieve good performance, we retain the *tf* term. Now, let us consider *idf*, i.e. the global weighting of t_k.

The conjecture is that if term selection can effectively differentiate a set of terms t_k out of all terms t to represent category c_i, then it is desirable to transform that difference into some sort of numeric values for further processing. Our approach is to replace the *idf* term with the value that reflects the term's strength of representing a specific category. Since this procedure is performed jointly with the cat-

Please cite this article in press as: Liu, Y. et al., Imbalanced text classification: A term weighting approach, *Expert Systems with Applications* (2007), doi:10.1016/j.eswa.2007.10.042

egory membership, this basically implies that the weights of t_k are category specific. Therefore, the only problem left is how to compute such values.

4.2. Probability based term weights

We decide to compute those term values using the most direct information, e.g. A, B and C, and combine them in a sensible way which is different from existing feature selection measures. From Table 2, two important ratios which directly indicate terms' relevance with respect to a specific category are noted, i.e. A/B and A/C:

- A/B: if term t_k is highly relevant to category c_i only, which basically indicates that t_k is a good feature to represent category c_i, then the value of A/B tends to be higher.
- A/C: given two terms t_k, t_l and a category c_i, the term with a higher value of A/C, will be the better feature to represent c_i, since a larger portion of it occurs with category c_i.

In the following of this paper, we name A/B and A/C relevance indicators since these two ratios immediately indicate the term's strength in representing a category. In fact, these two indicators are nicely supported by probability estimates. For instance, A/B can be extended as $(A/N)/(B/N)$, where N is the total number of documents, A/N is the probability estimate of documents from category c_i where term t_k occurs at least once and B/N is the probability estimate of documents not from category c_i where term t_k occurs at least once. In this manner, A/B can be interpreted as a relevance indicator of term t_k with respect to category c_i. Surely, the higher the ratio, the more important the term t_k is related to category c_i. A similar analysis can be made with respect to A/C. The ratio reflects the expectation that a term is deemed as more relevant if it occurs in the larger portion of documents from category c_i than other terms.

Since the computing of both A/B and $A/$C has its intrinsic connection with the probability estimates of category membership, we propose a new term weighting factor which utilizes the aforementioned two relevance indicators to replace *idf* in the classic *tfidf* weighting scheme. Considering the probability foundation of A/B and A/C, the most immediate choice is to take the product of these two ratios. Therefore, the proposed weighting scheme is formulated as

$$tf \cdot \log\left(1 + \frac{A}{B}\frac{A}{C}\right). \tag{4}$$

5. Experiment setup

Two data sets were tested in our experiment, i.e. MCV1 and Reuters-21578. MCV1 is an archive of 1434 English language manufacturing related engineering papers which we gathered by the courtesy of the Society of Manufacturing Engineers (SME). It combines all engineering technical papers published by SME from year 1998 to year 2000. All documents were manually classified (Liu & Loh, 2007). There are a total of 18 major categories in MCV1. Fig. 1 gives the class distribution in MCV1.

Reuters-21578 is a widely used benchmarking collection (Sebastiani, 2002). We followed Sun's approach (Sun, Lim, Ng, & Srivastava, 2004) in generating the category information. Fig. 2 gives the class distribution of the Reuters data set used in our experiment. Unlike Sun et al. (2004), we did not randomly sample negative examples from categories not belonging to any of the categories in our data set, instead we treated examples not from the target category in our data set as negatives.

We compared our probability based term weighting scheme with a number of other well established weighting schemes, e.g. TFIDF, 'ltc' and normalized 'ltc', on MCV1 and Reuters-21578. We also carried out the benchmarking experiments between our conjecture and many other feature selection methods, e.g. chi-square (ChiS), correlation coefficient (CC), odds ratio (OddsR), and information gain (IG), by replacing the *idf* term with the feature selection value in the classic *tfidf* weighting scheme. Therefore, schemes are largely formulated in a form as $tf \cdot (feature\ value)$ (TFFV). Table 4 shows all eight weighting schemes tested in our experiments and their mathematic formations. Please note that basically the majority of TFFV schemes are composed of two items, i.e. the normalized term frequency, $tf(t_i, d_j)/\max[tf(d_j)]$, and the term's feature value, e.g. $N(AD - BC)^2/(A + C)(B + D)(A + B)(C + D)$, in the chi-square scheme, where $tf(t_i, d_j)$ is the frequency of term t_i in the document d_j and $\max[tf(d_j)]$ is the maximum frequency of a term in the document d_j. The only different ones are TFIDF weighting, 'ltc' form and the normalized 'ltc' form as specified in Table 4.

Two popular classification algorithms were tested, i.e. Complement Naïve Bayes (CompNB) (Rennie, Shih, Teevan, & Karger, 2003), and Support Vector Machine (SVM) (Vapnik, 1999). The CompNB has been recently reported that it can significantly improve the performance of Naïve Bayes over a number of well known data sets, including Reuters-21578 and 20 Newsgroups. Various correction steps are adopted in CompNB, e.g. data transformation, better handling of word occurrence dependencies and so on. In our experiments, we borrowed the package implemented in Weka 3.5.3 Developer version (Witten & Frank, 2005). For SVM, we chose the well known implementation $\text{SVM}^{\text{Light}}$ (Joachims, 1998, 2001). Linear kernel has been adopted, since previous work has shown its effectiveness in TC (Dumais & Chen, 2000; Joachims, 1998). As for the performance measurement, *precision*, *recall* and their harmonic combination, i.e. the F_1-value, were calculated (Baeza-Yates & Ribeiro-Neto, 1999; van-Rijsbergen, 1979). Performance was assessed based on fivefold cross validation. Since we are very concerned about the performance of every category, we report the overall performance in macro-averaged manner, i.e.

Please cite this article in press as: Liu, Y. et al., Imbalanced text classification: A term weighting approach, *Expert Systems with Applications* (2007), doi:10.1016/j.eswa.2007.10.042

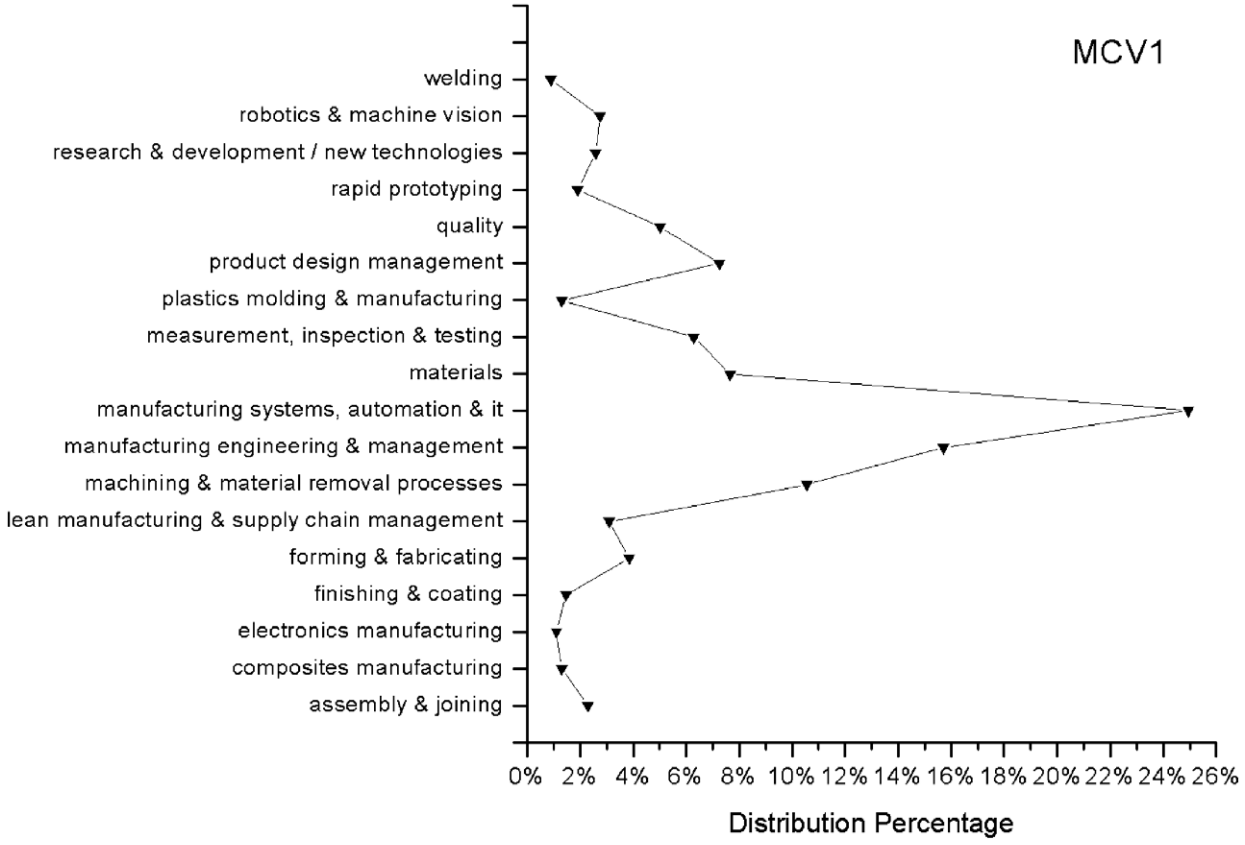

Fig. 1. Class distribution in MCV1.

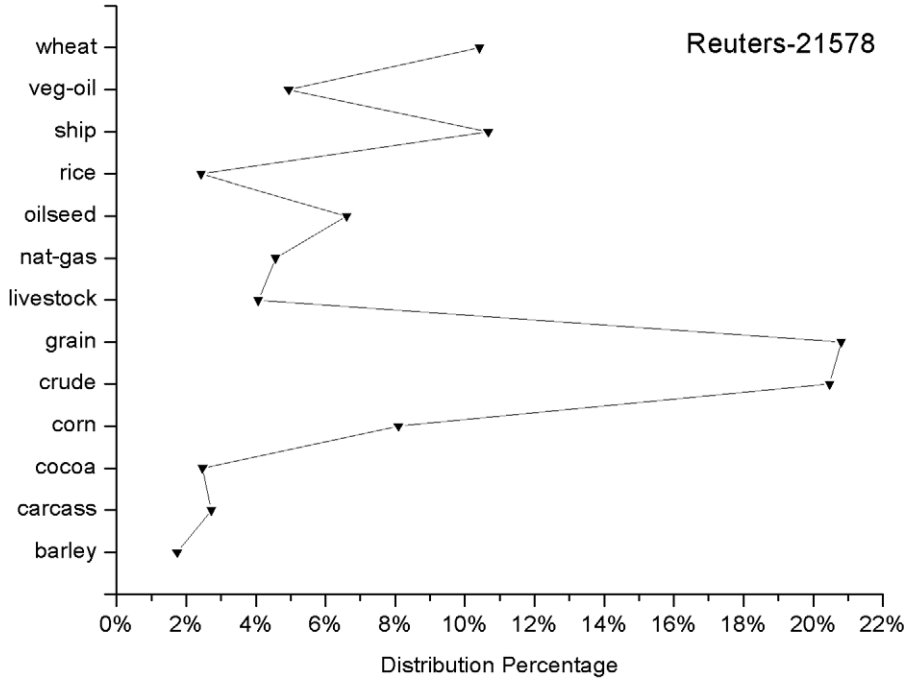

Fig. 2. Class distribution in Reuters-21578.

macro-average F_1, to avoid the bias for minor categories in imbalanced data associated with micro-averaged scores (Sebastiani, 2002; Yang & Liu, 1999).

Major standard text preprocessing steps were applied in our experiments, including tokenization, stop word and punctuation removal, and stemming. However, feature

Please cite this article in press as: Liu, Y. et al., Imbalanced text classification: A term weighting approach, *Expert Systems with Applications* (2007), doi:10.1016/j.eswa.2007.10.042

Table 4
All eight weighting schemes tested in the experiments and their mathematic formations, where the normalized term frequency *ntf* is defined as $tf(t_i,d_j)/\max[tf(d_j)]$

Weighting scheme	Name	Mathematical formations
$tf\cdot$ chi-square	ChiS	$ntf\cdot N(AD-BC)^2/(A+C)(B+D)(A+B)(C+D)$
$tf\cdot$ correlation coef.	CC	$ntf\cdot[\sqrt{N}(AD-BC)/\sqrt{(A+C)(B+D)(A+B)(C+D)}]$
$tf\cdot$ odds ratio	OddsR	$ntf\cdot\log(AD/BC)$
$tf\cdot$ info gain	IG	$ntf\cdot(\frac{A}{N}\log\frac{AN}{(A+B)(A+C)}+\frac{C}{N}\log\frac{CN}{(C+D)(A+C)})$
TFIDF	TFIDF	$ntf\cdot\log(\frac{N}{N(t_i)})$
$tfidf-$ ltc	ltc	$tf(t_i,d_j)\cdot\log(\frac{N}{N(t_i)})$
Normalized ltc	nltc	$\frac{tfidf_{ltc}}{\sqrt{\sum tfidf^2_{ltc}}}$
Probability based	Prob.	$ntf\cdot\log\left(1+\frac{A}{B}\frac{A}{C}\right)$

selection was skipped and all terms left after stop word and punctuation removal and stemming were kept as features.

6. Experimental results and discussion

6.1. Overall performance

Fig. 3 shows the overall performance of eight weighting schemes tested over MCV1 and Reuters-21578 using SVM and CompNB. They are reported in terms of macro-averaged F_1-values.

Our first observation is that all TFFV weighting schemes, e.g. $tf\cdot$ chi-square, $tf\cdot$ information gain and our probability based one, outperform classic ones, i.e. TFIDF, 'ltc', and normalized 'ltc' schemes. The TFIDF's performance on Reuters-21578 is in line with the literature (Sun et al., 2004; Yang & Liu, 1999). This has demonstrated the overall effectiveness of TFFV based schemes. In general, the performance patterns of eight weighting schemes on MCV1 and Reuters-21578 using two classification algorithms match very well. For example, our probability based term weighting scheme always take the lead in all eight schemes including the TFFV ones, and the normalized 'ltc' performs always the worst. When compared to TFIDF, the prevailing choice for term weighting in TC, our weighting strategy improves the overall performance from 6% to more than 12%, shown in Table 5. We also observe that when our scheme is adopted, CompNB has delivered a result which is very close to the best one that SVM can achieve using TFIDF scheme in Reuters-21578. This has demonstrated the great potential of using CompNB as a state-of-the-art classifier.

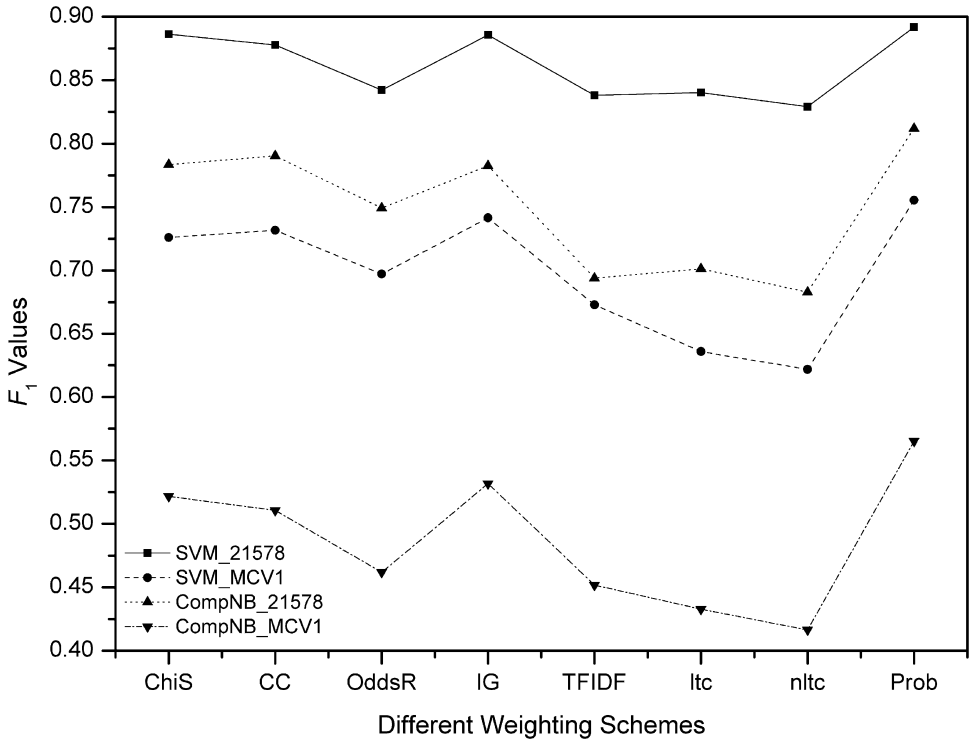

Fig. 3. The macro-averaged F_1-values of eight weighting schemes tested over MCV1 and Reuters-21578 using both SVM and CompNB.

Table 5
Macro-averaged F_1-values of TFIDF and probability based term weights on MCV1 and Reuters-21578

Classifier	MCV1		21578	
	TFIDF	Prob.	TFIDF	Prob.
SVM	0.6729	0.7553	0.8381	0.8918
CompNB	0.4517	0.5653	0.6940	0.8120

Among the three global based classic weighting schemes, i.e. TFIDF, 'ltc', and normalized 'ltc' form, none of them can generate comparable results over either MCV1 or Reuters-21578. A close look into their performance reveals that classifiers built for minor categories, e.g. composite manufacturing, electronic manufacturing and others in MCV1 or rice, natgas, cocoa and others in Reuters-21578, do not produce satisfactory results. As a result, this has largely affected the overall performance negatively. Among all TFFVs, surprisingly, odds ratio does not perform as expected, since in literature odds ratio is mentioned as one of the leading feature selection methods for TC (Ruiz & Srinivasan, 2002; Sebastiani, 2002). This implies that it is always worthwhile to reassess the strength of a term selection method for a new data set, even if it tends to perform well.

6.2. Gains for minor categories

As shown in Figs. 1 and 2, both MCV1 and Reuters-21578 are skewed data sets. While MCV1 possesses 18 categories with one major category occupying up to 25% of the whole population of supporting documents, there are six categories where each owns only 1% of MCV1, and other 11 categories falling below the average, i.e. 5.5%, if MCV1 is evenly distributed. The same case also happens to Reuters-21578 data set. While it has 13 categories, grain and crude, the two major categories, share around half of the population, there are eight categories in total whose shares falling below the average. Previous literature did not report successful stories over these minor categories (Sebastiani, 2002; Sun et al., 2004; Yang & Liu, 1999). Note, this imbalance situation is even worse when the training examples are arranged in the so called "one-against-all" setting for the induction of classifiers, i.e. examples from the target category (a minor category in this case) are considered as positive while examples from the rest categories are all deemed as negative. Nevertheless, given the nature of TC is to answer whether the document belongs to this particular category or not, the "one-against-all" setting is still the prevailing approach in TC, owing much to the fact that it dramatically reduces the number of classifiers to be induced.

Since the proposed probability based weighting scheme is the best in the benchmarking test over both MCV1 and Reuters-21578, we intend to examine why this is the case. Therefore, we plot its performances in detail against TFIDF in Figs. 4 and 5, respectively. This is largely because TFIDF is the best among the three classic approaches as well as the default choice for TC in its current research and application (Sebastiani, 2002).

A close examination of Figs. 4 and 5 shows that the probability based scheme produces much better results

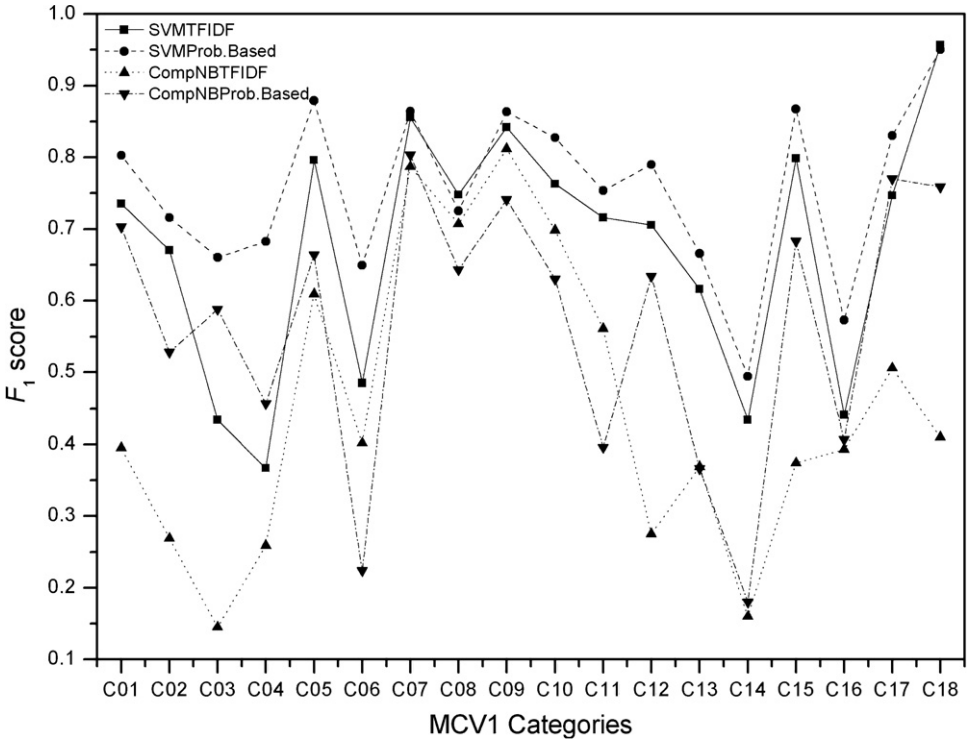

Fig. 4. F_1 scores of TFIDF and the probability based term weighting scheme tested over MCV1 using both SVM and CompNB.

Please cite this article in press as: Liu, Y. et al., Imbalanced text classification: A term weighting approach, *Expert Systems with Applications* (2007), doi:10.1016/j.eswa.2007.10.042

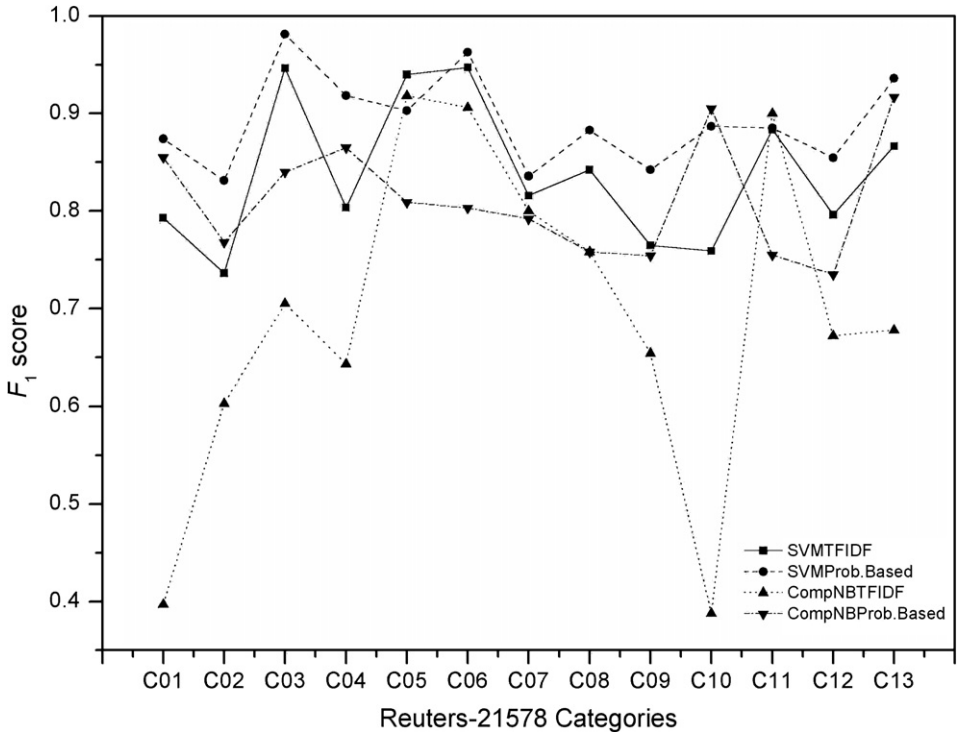

Fig. 5. F_1 scores of TFIDF and the probability based term weighting scheme tested over Reuters-21578 using both SVM and CompNB.

over minor categories in both MCV1 and Reuters-21578, regardless of classifiers used. For all minor categories shown in both figures, we observed a sharp increase of performance occurs when the system's weighting method switch from TFIDF to the probability one.

Table 6 reveals more insights with respect to the system performance. In general, we observe that using the probability based term weighting scheme can greatly enhance the systems' recalls. Although it falls slightly below TFIDF in terms of precision using SVM, it still improves the systems' precisions in CompNB, far superior to those TFIDF can deliver. For SVM, while the averaged precision of TFIDF in MCV1 is 0.8355 which is about 5% higher than the probability's, the averaged recall of TFIDF is 0.6006 only, far less than the probability's 0.7443. The case with Reuters-21578 is even more impressive. While the averaged precision of TFIDF is 0.8982 which is only 1.8% higher than the other, the averaged recall of probability based scheme reaches 0.9080, in contrast to TFIDF's 0.7935. Overall, the probability based weighting scheme surpasses TFIDF in terms of F_1-values over both data sets.

Table 6
Macro-averaged *precision* and *recall* of TFIDF and probability based term weights on MCV1 and Reuters-21578

Data	Classifier	*precision*		*recall*	
		TFIDF	Prob.	TFIDF	Prob.
MCV1	SVM	0.8355	0.7857	0.6006	0.7443
	CompNB	0.4342	0.6765	0.4788	0.5739
21578	SVM	0.8982	0.8803	0.7935	0.9080
	CompNB	0.5671	0.7418	0.9678	0.9128

6.3. Significance test

To determine whether the performance improvement gained by the probability based scheme and other TFFVs over these two imbalanced data sets are significant, we performed the macro-sign test (S-test) and macro-t-test (T-test) on the paired F_1-values. As pointed out by Yang and Liu (1999), on the one hand, the S-test may be more robust in reducing the influence of outliers, but at the risk of being insensitive or not sufficiently sensitive in performance comparison because it ignores the absolute difference between F_1-values; on the other hand, the T-test is sensitive to the absolute values but could be overly sensitive when F_1-values are highly unstable, e.g. for the minor ategories. Therefore, we adopt both tests here to give a comprehensive understanding of the performance improvement.

Since for both data sets, TFIDF performs better than the other two classic approaches, we choose it as the representative of its peers. For both the S-test and T-test, we actually conduct two sets of tests over two data sets, respectively. One is to test all TFFV schemes, including the probability one, against TFIDF and the other one is to test the probability scheme against others. While the first aims to assess the goodness of schemes in the form of TFFVs, the second intends to test whether the probability based scheme does generate better results. Table 7 summarizes p-values in the S-test for TFFV schemes against TFIDF and the probability one against others over two data sets. We consider two F_1-values to be the same if their difference is not more than

Table 7
p-Values of pairwise S-test on MCV1 and Reuters-21578, where two F_1-values are the same if their difference is not more than 0.01

Test	TFIDF	ChiS	CC	OddsR	IG	Prob.
MCV1						
SVM						
XX vs. TFIDF	–	4.813E−02	2.090E−03	5.923E−02	1.544E−02	6.561E−04
Prob. vs. XX	6.561E−04	6.363E−03	3.841E−02	6.561E−04	1.051E−01	–
CompNB						
XX vs. TFIDF	–	4.813E−02	4.813E−02	2.403E−01	4.813E−02	2.452E−02
Prob. vs. XX	2.452E−02	4.813E−02	2.452E−02	6.363E−03	1.544E−02	–
Reuters-21578						
SVM						
XX vs. TFIDF	–	3.174E−03	3.174E−03	1.938E−01	5.859E−03	3.174E−03
Prob. vs. XX	3.174E−03	2.744E−01	1.938E−01	1.929E−02	3.872E−01	–
CompNB						
XX vs. TFIDF	–	3.271E−02	3.271E−02	1.133E−01	7.300E−02	3.271E−02
Prob. vs. XX	3.271E−02	1.133E−01	1.133E−01	5.859E−03	1.334E−01	–

Table 8
t-Values of pairwise T-test on MCV1 and Reuters-21578, where $\alpha = 0.001$

Test	TFIDF	ChiS	CC	OddsR	IG	Prob.
MCV1						
SVM, t-critical = 3.354						
XX vs. TFIDF	–	1.963E+01	2.151E+01	8.343E+00	2.588E+01	3.017E+01
Prob. vs. XX	3.017E+01	1.347E+01	1.069E+01	2.400E+01	6.571E+00	–
CompNB, t-critical = 3.354						
XX vs. TFIDF	–	2.135E+01	2.049E+01	4.597E+00	2.419E+01	3.127E+01
Prob. vs. XX	3.127E+01	9.468E+00	1.043E+01	2.649E+01	8.192E+00	–
Reuters-21578						
SVM, t-critical = 3.467						
XX vs. TFIDF	–	2.516E+01	1.957E+01	1.692E+00	2.435E+01	2.889E+01
Prob. vs. XX	2.889E+01	3.682E+00	8.587E+00	2.262E+01	3.993E+00	–
CompNB, t-critical = 3.467						
XX vs. TFIDF	–	2.157E+01	1.946E+01	5.318E+00	2.130E+01	3.064E+01
Prob. vs. XX	3.064E+01	4.926E+00	9.127E+00	2.167E+01	6.128E+00	–

0.01, i.e. 1%. Table 8 summarizes t-values of T-test for the identical comparison settings, where α is 0.001.

From the results we can summarize the strength of different schemes. Consider the merits evaluated based on TFFVs against TFIDF, TFFVs have shown that they are the better approach in handling imbalanced data. Among various TFFVs, our proposed scheme claims the leading performance tested in both MCV1 and Reuters-21578, regardless of the classifier used. However, the approach based on the odds ratio is not much superior to TFIDF. With respect to the evaluation based on the merits of the probability scheme against others, it is not surprising to see that the new scheme still takes the lead. It manages to perform better than other approaches in TFFV, e.g. information gain and chi-square, where the absolute difference of F_1-values is considered. Finally, the results of information gain, chi-square and correlation coefficient shown in our tests are compatible with those in literature (Forman, 2003; Yang & Pedersen, 1997). In general, the more minor categories the data set possesses, the better overall performance can be elevated if the probability based weighting scheme is chosen.

7. Conclusion and future work

Handling of imbalanced data in TC has become an emerging challenge. In this paper, we introduce a new weighting paradigm which is generally formulated as $tf \cdot (feature\ value)$ (TFFV) to replace the classic TFIDF based approaches. We propose a probability based term weighting scheme, which directly makes use of two critical information ratios, as a new way to compute the term's weight. These two ratios are deemed to possess the most salient information reflecting the term's strength in associating a category. Their computation does not impose any extra cost compared to the conventional feature selection methods. Our experimental study and extensive comparisons based on two imbalanced data sets, MCV1 and Reu-

Please cite this article in press as: Liu, Y. et al., Imbalanced text classification: A term weighting approach, *Expert Systems with Applications* (2007), doi:10.1016/j.eswa.2007.10.042

ters-21578, show the merits of TFFV based approaches and their ability to handle imbalanced data. Among the various TFFVs, our probability based scheme offers the best overall performance in both data sets regardless of classifier used. Our approach has suggested an effective solution to improve the performance of imbalanced TC.

Start from the work reported in this paper, there are a few immediate tasks awaiting us. Since the probability scheme is derived from the understanding of feature selection, the $A/B \cdot A/C$ itself can also be considered as a new feature selection method that reflects the relevance of terms with respect to different thematic categories. It is interesting to further explore its joint application with other algorithms in TC. As for the slight decrease of precision noted, we intend to remedy the situation by switching the linear kernel with a string kernel in SVM (Lodhi, Saunders, Shawe-Taylor, Cristianini, & Watkins, 2002). Another challenge we are facing is to handle the situation where the critical information needed, e.g. A and B, cannot be easily secured, i.e. in text clustering. One potential direction is to infer these critical values from a small collection of labeled data and then test how robust these values or this probability approach could be, what strategies we can propose to accommodate the variation of term occurrence in the unlabeled documents, and how to modify the critical values accordingly. The whole idea falls into the emerging paradigm of semi-supervised learning. We will report our study when the results become more solid.

Acknowledgement

The work described in this paper was partially supported by a grant from the Research Grants Council of the Hong Kong Polytechnic University, Hong Kong Special Administrative Region, China (Project No. G-YF59).

References

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Boston, MA, USA: Addison-Wesley.

Baoli, L., Qin, L., & Shiwen, Y. (2004). An adaptive *k*-nearest neighbor text categorization strategy. *ACM Transactions on Asian Language Information Processing (TALIP), 3*(4), 215–226.

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *COLT: Proceedings of the workshop on computational learning theory* (pp. 92–100).

Brank, J., Grobelnik, M., Milic-Frayling, N., & Mladenic, D. (2003). *Training text classifiers with SVM on very few positive examples*. MSR-TR-2003-34.

Castillo, M. D. D., & Serrano, J. I. (2004). A multistrategy approach for digital text categorization from imbalanced documents. *ACM SIGKDD Explorations Newsletter, 6*(1) [Special issue on learning from imbalanced datasets].

Chawla, N., Japkowicz, N., & Kolcz, A. (Eds.). (2003). *Proceedings of the ICML'2003 workshop on learning from imbalanced data sets*.

Chawla, N., Japkowicz, N., & Kolcz, A. (Eds.). (2004). *ACM SIGKDD Explorations Newsletter, 6*(1) [Special issue on learning from imbalanced data sets].

Debole, F., & Sebastiani, F. (2003). Supervised term weighting for automated text categorization. In *Proceedings of the 2003 ACM symposium on applied computing* (pp. 784–788). Melbourne, Florida, USA.

Dietterich, T., Margineantu, D., Provost, F., & Turney, P. (Eds.). (2000). In *Proceedings of the ICML'2000 workshop on cost-sensitive learning*.

Dumais, S., & Chen, H. (2000). Hierarchical classification of Web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval (SIGIR2000)* (pp. 256–263). Athens, Greece.

Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the 17th international joint conference on artificial intelligence (IJCAI'01)* (pp. 973–978).

Fan, W., Yu, P. S., & Wang, H. (2004). Mining extremely skewed trading anomalies. In *Advances in database technology – EDBT 2004: Ninth international conference on extending database technology* (pp. 801–810). Heraklion Crete, Greece.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research, 3*, 1289–1305 [Special issue on variable and feature selection].

Ghani, R. (2002). Combining labeled and unlabeled data for multiclass text categorization. In *International conference on machine learning (ICML 2002)*, Sydney, Australia.

Goldman, S., & Zhou, Y. (2000). Enhancing supervised learning with unlabeled data. In *Proceedings of 17th international conference on machine learning* (pp. 327–334). San Francisco, California, USA.

Japkowicz, N. (Ed.). (2000). *Proceedings of the AAAI'2000 workshop on learning from imbalanced data sets*, AAAI Tech Report WS-00-05, AAAI.

Japkowicz, N., Myers, C., & Gluck, M. A. (1995). A novelty detection approach to classification. In *Proceedings of the 14th international joint conference on artificial intelligence (IJCAI-95)* (pp. 518–523).

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98, 10th European conference on machine learning* (pp. 137–142). Berlin, Germany.

Joachims, T. (2001). A statistical learning model of text classification with support vector machines. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 128–136). New Orleans, Louisiana, United States.

Leopold, E., & Kindermann, J. (2002). Text categorization with support vector machines – How to represent texts in input space. *Machine Learning, 46*(1–3), 423–444.

Lewis, D. D., & Gale, W. A. (1994). A sequential algorithm for training text classifiers. In *Proceedings of {SIGIR}-94, 17th ACM international conference on research and development in information retrieval* (pp. 3–12). Dublin, Ireland.

Lewis, D. D., Yang, Y., Rose, T. G., & Li, F. (2004). RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research, 5*, 361–397.

Liu, A. Y. C. (2004). *The effect of oversampling and undersampling on classifying imbalanced text datasets*. Masters thesis, University of Texas at Austin.

Liu, Y., & Loh, H. T. (2007). Corpus building for corporate knowledge discovery and management: A case study of manufacturing. In *Proceedings of the 11th international conference on knowledge-based and intelligent information and engineering systems, KES'07, Lecture notes in artificial intelligence, LNAI, Vol. 4692* (pp. 542–550). Vietri sul Mare, Italy.

Liu, B., Dai, Y., Li, X., Lee, W. S., & Yu, P. (2003). Building text classifiers using positive and unlabeled examples. In *Proceedings of the third IEEE international conference on data mining (ICDM'03)*, Melbourne, Florida.

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., & Watkins, C. (2002). Text classification using string kernels. *The Journal of Machine Learning Research, 2*, 419–444.

Manevitz, L. M., & Yousef, M. (2002). One-class SVMS for document classification. *The Journal of Machine Learning Research, 2*, 139–154.

Mladenic, D., & Grobelnik, M. (1999). Feature selection for unbalanced class distribution and Naive Bayes. In *Proceedings of the 16th international conference on machine learning, ICML'99* (pp. 258–267).

Ng, H. T., Goh, W. B., & Low, K. L. (1997). Feature selection, perception learning, and a usability case study for text categorization. In *ACM SIGIR forum, Proceedings of the 20th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 67–73). Philadelphia, Pennsylvania, United States.

Nickerson, A., Japkowicz, N., & Milios, E. (2001). Using unsupervised learning to guide re-sampling in imbalanced data sets. In *Proceedings of the eighth international workshop on AI and statistics* (pp. 261–265).

Nigam, K. P. (2001). *Using unlabeled data to improve text classification*. PhD thesis, Carnegie Mellon University.

Raskutti, B., & Kowalczyk, A. (2004). Extreme re-balancing for SVMs: A case study. *ACM SIGKDD Explorations Newsletter, 6*(1), 60–69 [Special issue on learning from imbalanced datasets].

Rennie, J. D. M., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of Naive Bayes text classifiers. In *Proceedings of the 20th international conference on machine learning* (pp. 616–623). Washington, DC, USA.

Ruiz, M. E., & Srinivasan, P. (2002). Hierarchical text categorization using neural networks. *Information Retrieval, 5*(1), 87–118.

Salton, G., & Buckley, C. (1988). Term weighting approaches in automatic text retrieval. *Information Processing and Management, 24*(5), 513–523.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York, USA: McGraw-Hill.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR), 34*(1), 1–47.

Sun, A., Lim, E.-P., Ng, W.-K., & Srivastava, J. (2004). Blocking reduction strategies in hierarchical text classification. *IEEE Transactions on Knowledge and Data Engineering (TKDE), 16*(10), 1305–1308.

van-Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.). London, UK: Butterworths.

Vapnik, V. N. (1999). *The nature of statistical learning theory* (2nd ed.). New York: Springer-Verlag.

Weiss, G. M. (2004). Mining with rarity: A unifying framework. *ACM SIGKDD Explorations Newsletter, 6*(1), 7–19 [Special issue on learning from imbalanced datasets].

Weiss, G. M., & Provost, F. (2003). Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research, 19*, 315–354.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd ed.). San Francisco, CA, USA: Morgan Kaufman.

Yang, Y. (1996). Sampling strategies and learning efficiency in text categorization. In *Proceedings of the AAAI spring symposium on machine learning in information access* (pp. 88–95).

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 42–49). Berkeley, California, United States.

Yang, Y., & Pedersen, J. O. (1997). A Comparative study on feature selection in text categorization. In *Proceedings of ICML-97, 14th international conference on machine learning* (pp. 412–420).

Yu, H., Zhai, C., & Han, J. (2003). Text classification from positive and unlabeled documents. In *Proceedings of the 12th international conference on information and knowledge management (CIKM 2003)* (pp. 232–239). New Orleans, LA, USA.

Zelikovitz, S., & Hirsh, H. (2000). Improving short text classification using unlabeled background knowledge. In *Proceedings of the 17th international conference on machine learning (ICML2000)*.

Zheng, Z., Wu, X., & Srihari, R. (2004). Feature selection for text categorization on imbalanced data. *ACM SIGKDD Explorations Newsletter, 6*(1), 80–89 [Special issue on learning from imbalanced datasets].

Please cite this article in press as: Liu, Y. et al., Imbalanced text classification: A term weighting approach, *Expert Systems with Applications* (2007), doi:10.1016/j.eswa.2007.10.042