Bias Analysis in Text Classification for Highly Skewed Data

Lei Tang and Huan Liu
Department of Computer Science & Engineering
Arizona State University
Tempe, AZ 85287-8809, USA
{l.tang, hliu}@asu.edu

Abstract

Feature selection is often applied to high-dimensional data as a preprocessing step in text classification. When dealing with highly skewed data, we observe that typical feature selection metrics like information gain or chi-squared are biased toward selecting features for the minor class, and the metric of bi-normal separation can select features for both minor and major classes. In this work, we investigate how these feature selection metrics impact on the performance of frequently used classifiers such as Decision Trees, Naïve Bayes, and Support Vector Machines via bias analysis for highly skewed data. Three types of biases are metric bias, class bias, and classifier bias. Extensive experiments are designed to understand how these biases can be employed in concert and efficiently to achieve good classification performance. We report our findings and present recommended approaches to text classification based on bias analysis and the empirical study.

1 Introduction

Text classification studies how to classify documents into predefined categories. Since the dimensionality of the feature space for a document is very high, feature (term) selection is often applied [5]. It is shown in [29] that after removing up to 90% features one can obtain the same or better performance than using all the features . Commonly used feature selection metrics are information gain and chi-squared. Standard classifiers for text classification are decision trees(DT), naïve bayes classifier (NBC), and support vector machines (SVMs) [28]. Dealing with highly skewed data (we follow the definition in [5] in this work: the ratio between the minor and major classes exceeds $1 : 67$), we notice that some typical feature selection metrics do not perform as expected. We therefore conduct a systematic study of feature selection metrics, representative classifiers, and their biases in dealing with highly skewed data.

Learning from skewed data has been attracting increasing attention in recent years [24]. Skewed class distributions exist in many applications including direct marketing [14], fraud detection [2, 19], text categorization [13], and network intrusion detection [10]. Classification of highly skewed data is a difficult task in data mining [8, 24]. To address the skewness problem, researchers realize accuracy may not be suitable in evaluating the performance of a classifier. Alternative measures are Receiver Operating Characteristics (ROC) analysis, the area under the curve (AUC) [20], precision, recall and F-measure [28]. Total misclassification cost is another useful measure [4] provided that we know misclassification cost for each class or example. The authors in [6] integrate the performance evaluation metric into classifier design. By optimizing Macro F-measure of a classifier, their MFoM classifier with LSI-inducted features is comparable to a linear SVM using all features.

The study of sampling methods prior to classification is another line of research tackling skewed data [12, 14, 15]: e.g., over-sample the minor class or under-sample the major class. Some heuristics can be designed to remove redundancy, noise, unsafe data points or data near the borderline while sampling [12, 1]. In [3], the authors propose interpolating artificial data points between examples of the minor class. In general, over-sampling is more secure and stable compared with under-sampling as it does not lose any information. Interestingly, random over-sampling is competitive with complex sampling methods [1]. Besides sampling, cost-sensitive learning [11, 4], one-class learning [22, 16] and many algorithm specific approaches [10, 26, 7] are also considered when dealing with skewed data.

The authors in [30] propose to divide features into positive features and negatives features, and then use a wrapper model to find the optimal ratio to combine positive features and negative features together in classification. However, no pattern of the optimal ratio between positive and negative features is found and recommended. Since a filter model usually runs much faster, information gain and chi-squared are shown to be effective for feature selection [29]. Odds ratio is suggested to handle the skewness with a naïve bayes classifier [18]. Bi-normal separation proposed in [5] improves the performance of support vector machines especially for highly skewed data compared with other metrics.

This work is to investigate how various biases associated with feature selection metrics and classification algorithms can be effectively used in text classification for highly skewed data. We first study three biases with specific examples, next examine their combinations for effective text classification, and then design experiments to extensively evaluate the effectiveness using various biases together for text classification on benchmark data sets.

2 Biases Associated with Data Skewness

We study three types of biases: feature selection metric bias, class bias and classifier bias.

2.1 Feature selection metric bias

Among many feature selection metrics, we focus on four widely used metrics: information gain (IG), chi-squared (CHI) [29], odds ratio (Odds) [18] and bi-normal separation (BNS) [5]. IG and CHI are reported as the best measure in [29]. However, when a data set is extremely skewed, typical feature selection measures may not work well. We adopt the notions of positive and negative features to study why these metrics perform differently. Here 1 denotes one word occurring in document, and 0 otherwise. Throughout the paper, pos means the minor class, and neg the major class. A feature selection metric is used to assign a score to each feature based on the contingency table as in Table 1. "tp", "fp","fn" and "tn" are frequencies of different feature values in different classes, respectively. As all features are binary (present or absent in a document), we categorize the features into three groups: (a) positive features, where $\frac{tp}{\#pos} > \frac{fn}{\#neg}$. These features have higher probability appearing in documents of the positive class; (b) negative features, where $\frac{tp}{\#pos} < \frac{fn}{\#neg}$; and (c) neutral features, where $\frac{tp}{\#pos} = \frac{fn}{\#neg}$, or the features occur in the positive and negative classes with the same probability.

We use the "cora36" data [5] to illustrate the problem associated with skewed data. It consists of 36 classes and there are 50 documents in each class. First, we select "Data Mining" as the positive class and "Agents" as the negative class to obtain a balanced data set. We then generate another data set via the one-vs-all approach, that is, "Data Mining" is the positive class while all the other 35 classes are negative, and its skewness ratio is 1:35. We show the proportion of positive features selected by four feature selection metrics on the balanced data in Figure 1 and on the imbalanced case in Figure 2. The x-axis is the number of features being selected and the y-axis is the proportion of positive features. The straight line parallel to the x-axis is the proportion of positive features among all the features.

Feature Value	pos	neg
1	tp	fn
0	fp	tn

Table 1. Contingency Table

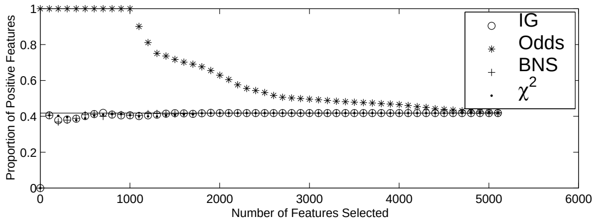

Figure 1. Features selected on balanced data

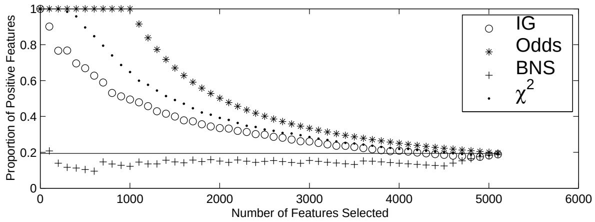

Figure 2. Features selected on skewed data

When the data is balanced, IG, CHI and BNS all select both positive and negative features and their proportions are similar to the natural distribution of positive and negative features. However, when the data is skewed, as in Figure 2, IG and CHI choose more positive features, thus are biased toward the positive features. BNS, however, still selects both positive and negative features, and the proportion of positive features of BNS is not far away from the true distribution. Odds ratio, according to its definition, selects only positive features initially in both cases as observed in both figures. We further demonstrate this metric bias issue in Figure 3 with the average result of 5 extremely skewed data sets derived from data set "wap" [5]. The skewness of data sets varies from $1:85$ to $1:311$. The bias of IG and CHI is substantial for the top 500 features.

Hence, in the context of the highly skewed data, we divide feature selection metrics into two categories:

- **Biased metrics:** IG, CHI, and Odds fall into this cat-

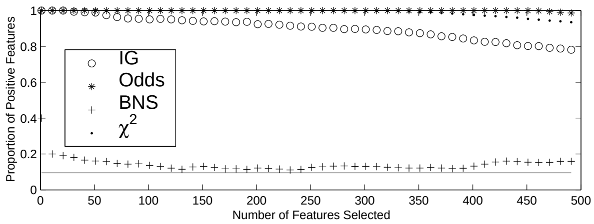

Figure 3. Metric bias of the top 500 features

egory as they are all biased toward selecting positive features especially when we select a relative small number of features (say, less than 500).

- **Unbiased metrics:** BNS selects both positive and negative features and is not biased toward either class.

BNS outperforms other metrics for highly skewed data in [5]. Hence, one way to deal with data skewness is to employ an unbiased feature selection metric such as BNS.

2.2 Class bias

For highly skewed data, the class distribution is biased toward the majority in the sense that most classifiers would predict the major class to obtain overall accuracy. However, in dealing with highly skewed data, we are more interested in predicting the minor class as to achieve a low false-negative rate while maintaining overall accuracy [24]. One way to address the class bias is to move the decision boundary of certain classifiers (NBC,SVM etc.) by changing the threshold. But it is difficult to determine how much to shift. Comparatively, over-sampling is a simple yet effective way to alleviate this bias [1].

2.3 Classifier bias

Three widely used classifiers (DT, NBC, and SVM) also exhibit different biases. As we know, DT like C4.5 has an embedded feature selection mechanism, i.e., it prefers features with high information gain. This bias leads to its selection of positive features to branch. Because of this embedded mechanism, we can anticipate that feature selection sometimes may not help much if we use DT for text classification. DT is, however, sensitive to sampling as sampling can change data local distributions. Over-sampling can make data balanced and negative features would be equally likely to be selected by IG. Thus, both positive and negative features will be used in building a decision tree.

In Table 2, we show the effects of feature selection and over-sampling for the 5 highly skewed data sets over classifier DT. Column A is the numbers of positive and negative features found in a decision tree built from the original data and positive features are usually selected. Column B is similar to Column A but the tree is built from the data after over-sampling; it can be seen that both positive and negative features are used in the built trees. Column C shows the positive and negative features found in a decision tree built using only 50 features selected by BNS (an unbiased metric) and DT selects only positive features. Clearly, over-sampling increases the complexity of the tree and allows for many negative features to be used in the built trees. This observation confirms our hypothesis above that DT is sensitive to sampling but insensitive to feature selection.

	A		B		C	
Skew Ratio	pos	neg	pos	neg	pos	neg
1:85	3	0	10	10	3	0
1:103	4	0	8	7	4	0
1:119	2	0	5	5	2	0
1:141	3	0	10	10	2	0
1:311	1	1	1	19	2	0

Table 2. Positive/Negative features in a tree

NBC has different bias from that of DT. Feature selection can have a significant impact on NBC [18]. In addition, over-sampling changes NBC's prediction. NBC predicts the class label of an instance proportional to the class distribution. As over-sampling changes the global class distribution, the prior class probability also changes. Therefore, NBC is sensitive to both sampling and feature selection.

Feature selection also affects SVM's performance [5]. But random over-sampling affects SVM moderately and becomes ineffective when the number of features is large. As shown in [27], SVM's prediction is biased against the minority. The authors attribute this to the relatively small sample size of the minor class: Compared with their counterparts, positive instances tend to reside far away from the "actual boundary" when the training data is severely skewed. So the constructed decision boundary of SVM invades the actual space of the minor class. Random over-sampling cannot change this since no new data is generated.

Another factor also contributes to SVM's prediction bias. When those positive instances near the "actual boundary" are surrounded by some negative instances or noise, the decision boundary will be adapted to the noise but ignore the errors in the minor class. In this case, over-sampling, by increasing the error penalty for the minor class, can protect these positive instances from being overwhelmed. However, if no error occurs in the minor class during training, sampling is ineffective. This becomes true when the feature dimensionality is large, as we can easily find a perfect hyperplane to separate the majority (negative class) and the minority (positive class). Therefore, over-sampling can moderately influence SVM only when the number of features is small.

3 Relationship Between Biases

3.1 Bias Analysis

We conducted a pilot study to evaluate the effect of over-sampling on feature selection. Based on the definitions of Odds and BNS, over-sampling should not have significant impact on feature selection using Odds and BNS as it does not change the probability of one word's occurrence in classes. We compared the effect of sampling on feature distribution on 5 extremely skewed data sets from "wap" data and showed the results in Figure 4. We just show the legend corresponding to over-sampling (OS) before feature selection. The remaining symbols are the same as in Figure 3. The results suggest that over-sampling causes IG and CHI to select more negative features, but BNS can generate a more balanced subset of positive and negative features. Therefore, over-sampling before feature selection can alleviate the metric bias of IG and CHI, but not much.

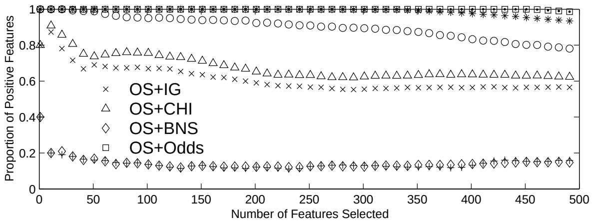

Figure 4. Features selected after sampling

In order to overcome data skewness, we can do over-sampling before or after feature selection; For classifiers, we consider DT, NBC and SVM; Concerning the class bias, we can do over-sampling or not; With feature selection, we can use a biased or unbiased metric or just select all the features. There could be a total of $2 \times 3 \times 2 \times 3 = 36$ different ways to deal with data skewness, corresponding to the four steps: 1. over-sampling; 2. feature selection; 3. over-sampling; and 4. classification. Based on the above analysis and the pilot study, we understand that not all 36 approaches are effective in addressing data skewness. If one step makes little difference (e.g., feature selection for DT), we just set "No" as default to save computation time. Table 3 lists the 12 promising approaches to tackle data skewness. The approaches in Table 3 are derived from bias analysis. We now further evaluate them through comparative experiments to investigate whether they can improve performance of classifiers for text classification, and which one is more appropriate for highly skewed data. The more interesting question is whether three types of biases can work in concert to achieve better performance.

Sampling before FS	Classifier	Sampling after FS	Feature Selection(FS)
Yes	NBC	Yes	biased
		No	biased
	SVM	Yes	biased
		No	biased
No	DT	Yes	No
	NBC	Yes	No
			biased
			unbiased
		No	unbiased
	SVM	Yes	biased
			unbiased
		No	unbiased

Table 3. Promising approaches

3.2 Experiment Setting

Since we want to reduce false negatives for the minor class without sacrificing the performance of the major class, we use Macro F-measure [5] as the performance measure.

Benchmark data sets: They are chosen based on those used in [5]. All the attributes are binary with 1 representing a word's occurrence in a document and 0 non-occurance. We change all the multi-class documents into binary-class data sets via the one-vs-all approach. We concentrate on highly skewed data sets with a ratio exceeding 1:67. Excluding those data sets with very few (less than 10) instances in the minor class, we have 18 data sets.

Classifiers: C4.5, NBC and SVM are typical classifiers for text categorization [9, 28, 5]. We use the default settings in WEKA [25] for all the classifiers. As we focus on data sets with binary attributes, the NBC we employed is multi-bernoulli model [17].

Feature selection metrics: IG, CHI, Odds are all biased metrics. CHI always yields the same trend as IG in our previous analysis. CHI performs similarly as IG and the two have correlated failures [5, 29]. Hence, we chose IG and Odds to represent biased metrics. In our experiments, we shall examine both biased metrics (IG, Odds) and unbiased metrics (BNS). We select 2, 4, 6, 8, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 1000, or all features. Over-sampling can be applied before and/or after feature selection. After feature selection, a classier can be built with original or over-sampled data. We perform 5×5-fold cross validation to obtain F-measure results.

3.3 Results and Discussions

The results for DT, NBC, and SVM with different numbers of selected features are shown in Figures 5, 6 and 7. "OS+ $\cdots$" and "$\cdots$ + OS" represent over-sampling before and after feature selection, respectively. The results are obtained by averaging over the 18 sets. For each classifier, we check whether and when over-sampling or feature selection can improve the performance. The classification result based on original data without feature selection or sampling (the straight line in the figures) is considered the baseline.

Most of the results are consistent with our bias analysis. 10 out of 12 methods work well in dealing with skewed data, except for two approaches - combine over-sampling with metric (BNS) for both NBC and SVM. Over-sampling always improves the performance using biased metrics including IG , Odds, or OS+IG, but not so when using an unbiased metric. In order to understand why, we investigate false negative rate and error rate. We find that over-sampling after we use BNS to select features will make the false negative rate very low but significantly increase the total error rate. There is a tradeoff between over-sampling

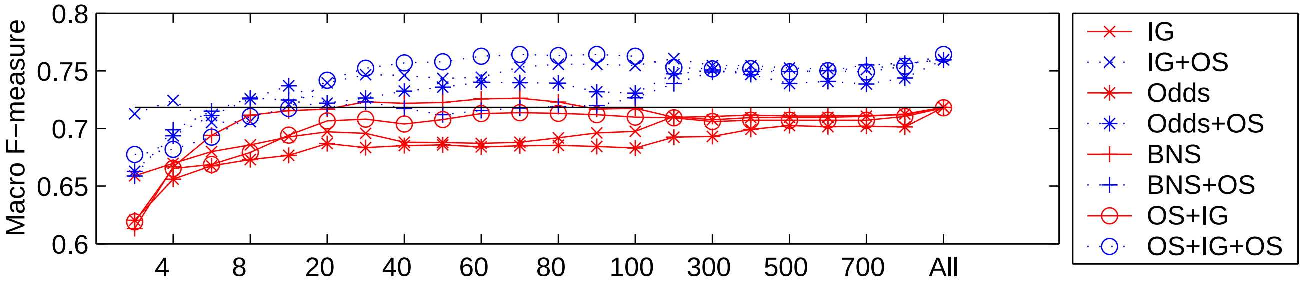

Figure 5. Performance of C4.5. Over-sampling alone improves the performance significantly. Little difference is observed for feature selection.

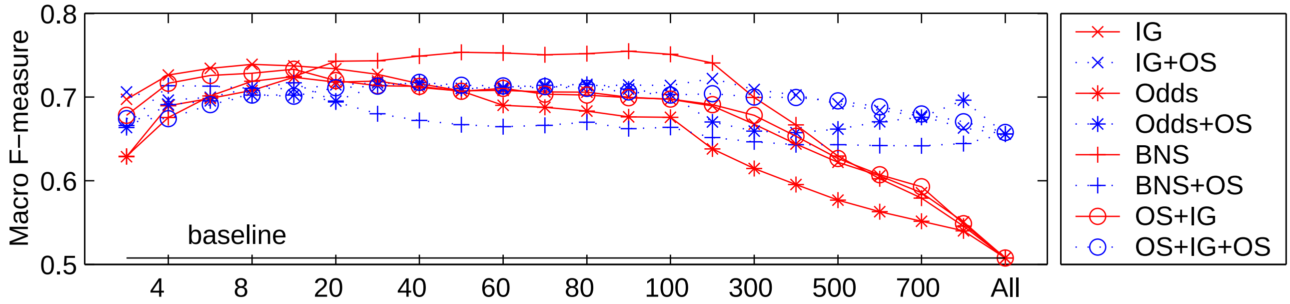

Figure 6. Performance of Naïve Bayes Classifier. Over-sampling helps a lot. Sampling plus biased feature selection methods can even achieve better result. More interestingly, unbiased metric BNS peaks when only 40 to 200 features are selected. But sampling counteract the optimality of unbiased metric a lot making it not much difference from sampling alone.

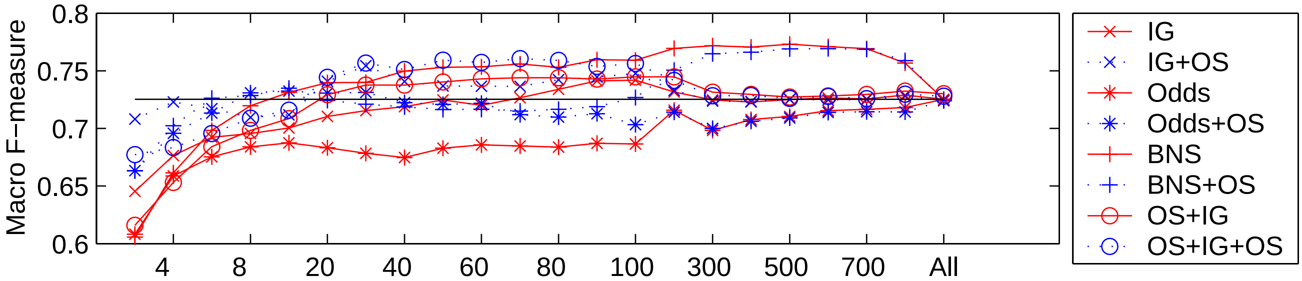

Figure 7. Performance of Support Vector Machine. When dimensionality is large (say, greater than 200), over-sampling makes no difference. But different feature selection methods lead to different final result. BNS, again, is the best. Odds, the most biased feature selection metric, even gets worse performance than the baseline. When the feature number decreases, no obvious winner exists. Generally, over-sampling can always heave the performance of biased metrics. But it's not the case for unbiased metric.

and metric bias. With NBC or SVM, we can address the data skewness either from the class bias or metric bias but not both.

Comparing all 4 feature selection methods: BNS, OS+IG, IG and Odds with increasing bias, BNS is the best in most cases without over-sampling. Odds is usually the worst. Notice that over-sampling before feature selection using IG is always better than using IG alone. This is because the former can select more negative features. When we select very few features (say less than 10), biased metrics are preferred as they can protect the minor class from being overwhelmed by the major class. As the number of features increases, negative features can help. This explains why BNS excels when a large number (more than 100 for SVM and 30 for NBC) of features are selected.

In sum, we can address the skewness using metric bias or class bias. Directly combining sampling with unbiased metric does not necessarily achieve better performance. Metric bias is in general more effective than class bias.

4 Heuristics of Metric Bias

As mentioned above, using an unbiased feature selection metric, i.e., by selecting both positive and negative features we can usually increase the discriminability of the classifier. This agrees with the results of [5] and [30]. However, the uncertainty of each class should also be considered. We have the following theorem from statistics:

Theorem 1 *Assume each document $\{d_1, d_2, \cdots, d_m\}$ in one class can be considered as a sample from certain innate population with mean μ and standard deviation σ. Then, the mean of the sampling distribution of $\bar{d} = \frac{1}{m}\sum_{i=1}^{m} d_i$, denoted by $\mu_{\bar{d}}$ and $\sigma_{\bar{d}}$, resectively, are*

$$\mu_{\bar{d}} = \mu, \quad \sigma_{\bar{d}} = \sigma/\sqrt{m}$$

Actually, $\bar{d}$ is used to estimate the probability of a word appearing in one class. Clearly, the uncertainty (variance)of this statistic is in reverse proportion to the number of instances in the class during training. Based on this theorem, it is not difficult to follow that the estimation of the probability of word occurrence in the minor class is associated with more uncertainty compared with that in the major class. Therefore, our feature selection method should bias toward the negative features to reduce uncertainty. In [30], the authors tried to find the optimal ratio of positive and negative features but in vain. Based on the observation of superiority of BNS in the experiments and the theorem above, we have the following conjecture:

Conjecture *The ratio between positive and negative features should be close to a smoothed class distribution.*

The class distribution generally does not follow the optimal ratio for highly skewed data, as the minor class often

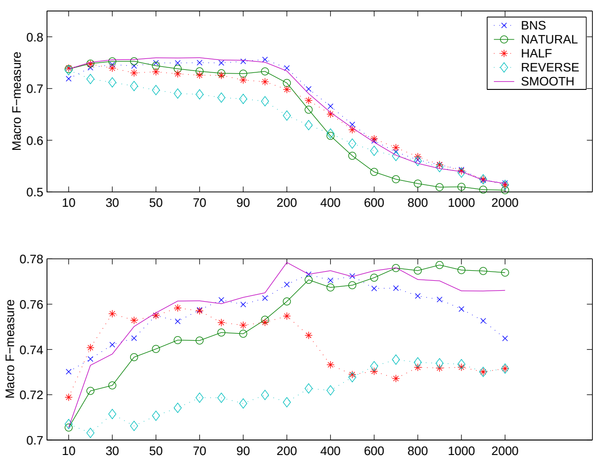

Figure 8. Various Ratios Using NBC and SVM

contains about one percent of the total training documents. Thus, most of the time, only one positive feature is selected. To achieve high discriminability, we need to smooth the distribution accordingly to select more positive features.

Smoothing function In our experiments, if the class percentage is p, then we select $\frac{1}{1+exp(-\alpha(p-0.5))}$ features out of the total number for this class. Here, α is a parameter to control the degree of smoothing. Typically, α between 4 to 7 works fine. Here, we just set α to 6. We verify the conjecture by following the strategy in [30]: Group positive and negative features first, and then use a biased metric (we adopt IG as a reference) to select a specified number of positive and negative features, respectively. Here we just show the results of four representative ratios. The first three ratios of positive and negative features are equal (NATURAL), reverse (REVERSE) proportional to the class distribution[1], or 1:1 (HALF). And the forth ratio is our conjecture to select features according to a smoothed class distribution (SMOOTH).

Figure 8 shows the average performance of NBC and SVM, respectively, under various feature selection methods: BNS, NATURAL, HALF, REVERSE, SMOOTH. Again, 5×5 cross validation is conducted on 18 data sets. The numbers of selected features (x-axis) are 10, 20, 30, $\cdots$, 90, 100, 200, $\cdots$, 900, 1000, 1500, and 2000, respectively. Clearly, SMOOTH beats other ratios often and is similar to BNS.

We also include in Figures 9 and 10 the T-test results comparing different feature selection ratios for NBC and SVM, respectively. Specially, SMOOTH vs. NATURAL, HALF, and REVERSE. The continuous line represents the count that Heuristic A beats Heuristic B; the dashed line denotes the reversed situation and the dotted line signals the cases of "tie". Obviously, selecting features with a smoothed class distribution outperforms the two ratios: HALF and Reverse dramatically, especially when we select

[1] We select negative features if there are no enough positive features.

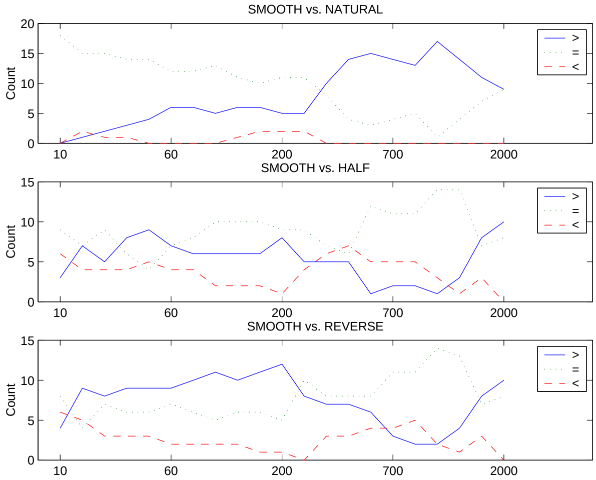

Figure 9. T-test result of NBC

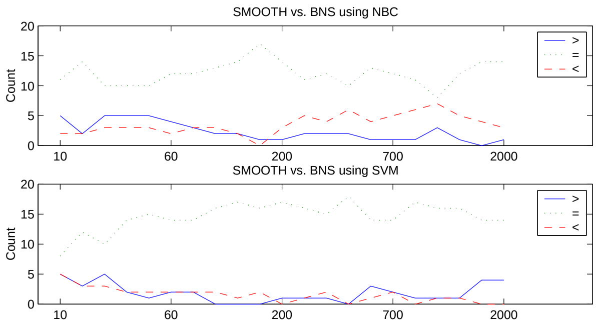

Figure 11. SMOOTH vs. BNS

200 more features for SVM and 50-200 features for NBC. *Keep in mind that those cases are when the best average performance is achieved.* Comparing smoothed distribution (SMOOTH) and non-smoothed distribution (NATURAL), a huge difference can be found for NBC. For SVM, only when we select huge numbers of features(>700) can NATURAL outperform SMOOTH. Meanwhile, SMOOTH performs comparable to or even better than BNS in Figure 11 except when we select 200 more features for NBC, that is, when NBC's performance goes down sharply in Figure 8.

In conclusion, when facing highly skewed data, it is more appropriate to select features according to a smoothed class distribution to achieve better performance.

5 Tradeoff between Metric Bias and Sampling Ratio

In Section 4, we notice that there should be a trade-off between metric bias and sampling. In [23] the authors suggest that it is not necessarily the natural distribution or a balanced distribution after sampling will obtain optimal performance. However, feature selection bias is not investigated in that paper. We now investigate a proper sampling ratio with different feature selection metric bias.

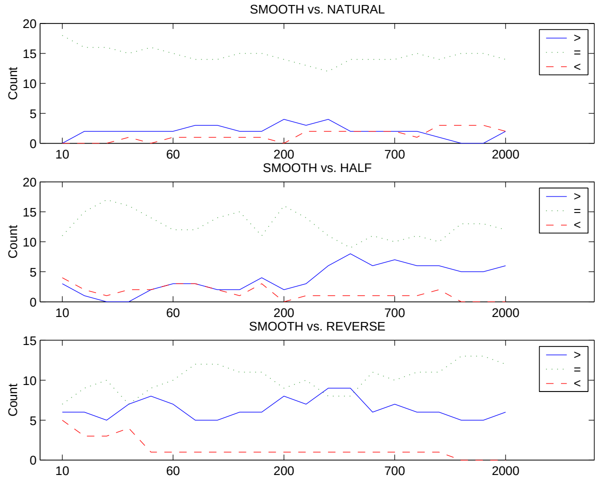

Figure 10. T-test result of SVM

In order to observe a general trend, we select 100 features according to various feature ratio(#positive feature/100) and sampling ratio (the skew ratio after sampling, i.e., #positive instances: #negative instances). The feature ratio ranges among 0, 0.01, 0.02, 0.03,$\cdots$,0.09, 0.1, 0.2, $\cdots$, 0.8, 0.9, 1.0 and the sampling ratio varies among $\frac{1}{10}, \frac{2}{10}, \cdots, \frac{8}{10}, \frac{9}{10}, \frac{10}{10}, \frac{10}{9}, \frac{10}{8}, \cdots, \frac{10}{2}, \frac{10}{1}$.

From Figure 12 and 13, we could see pretty the same trend for both NBC and SVM. When positive features are minority, sampling always decrease the performance. Only when positive features dominate can sampling contribute some improvements, and various sampling ratios always yield similar performance. The best performance is ob-

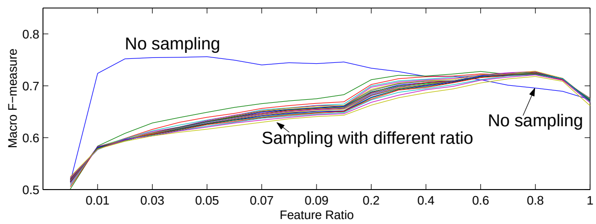

Figure 12. NBC performance

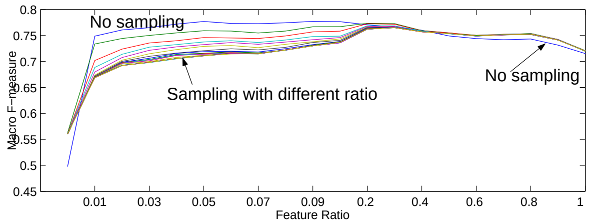

Figure 13. SVM performance

tained when no sampling is used and the feature ratio is between 0.02 to 0.1, which covers, most of the time, the smoothed class distribution in our previous experiments. This also suggests that our method is very insensitive to the parameter α as long as the smoothed distribution falls in a certain interval.

6 Conclusions

To handle highly skewed data with high dimensionality, we discuss three types of biases: class bias, feature selection metric bias, and classifier bias. Over-sampling is an effective way to address the class bias. BNS is a good unbiased metric, and IG, CHI and Odds are biased metrics. This work provides a systematic bias analysis and performs an extensive empirical study to evaluate various combinations to improve performance of text classification using typical classifiers such as DT, NBC, and SVM. Experimental results suggest that: (1) Sampling before feature selection can cause selection of more negative features, which explains why over-sampling improves the performance of decision trees on highly skewed data. (2) It is more effective to select good features than to change the class distribution for SVMs and NBC in discrimination. (3) With different uncertainty associated with majority and minority classes, we propose a heuristic to select positive and negative features according to a smoothed class distribution, which is shown to beat other feature ratios and perform as well as BNS. This also suggests that it is not the ranking method but the feature ratio that matters. (4) If a feature selection measure is biased, over-sampling can improve classification performance. But when a feature selection measure is not biased, over-sampling decreases the performance a lot. (5) Performance is insensitive to the sampling ratio if we do sampling after feature selection.

References

[1] G. Batista, R. C. Prati, and M. C. Monard. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.*, 6(1):20–29, 2004.

[2] P. K. Chan and S. J. Stolfo. Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. In *Knowledge Discovery and Data Mining*, pages 164–168, 1998.

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *JAIR*, 16:321–357, 2002.

[4] C. Elkan. The foundations of cost-sensitive learning. In *Proc. IJCAI*, pages 973–978, 2001.

[5] G. Forman. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, 3:1289–1305, 2003.

[6] S. Gao, W. Wu, C.-H. Lee, and T.-S. Chua. A maximal figure-of-merit learning approach to text categorization. In *Proc. SIGIR*, pages 174–181. ACM Press, 2003.

[7] H. Guo and H. L. Viktor. Learning from imbalanced data sets with boosting and data generation: the DataBoost-IM approach. *SIGKDD Explor. Newsl.*, 6(1):30–39, 2004.

[8] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intell. Data Anal.*, 6(5):429–449, 2002.

[9] T. Joachims. Text categorization with suport vector machines: Learning with many relevant features. In *Proc. ECML*, pages 137–142. Springer-Verlag, 1998.

[10] M. V. Joshi, R. C. Agarwal, and V. Kumar. Mining needle in a haystack: classifying rare classes via two-phase rule induction. In *Proc. SIGMOD*, pages 91–102, 2001.

[11] G. J. Karakoulas and J. Shawe-Taylor. Optimizing classifers for imbalanced training sets. In *Proc. NIPS*, pages 253–259, 1998.

[12] M. Kubat and S. Matwin. Addressing the curse of imbalanced training sets: one-sided selection. In *Proc. ICML*, pages 179–186. Morgan Kaufmann, 1997.

[13] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proc. SIGIR*, pages 3–12. Springer Verlag, Heidelberg, DE, 1994.

[14] C. X. Ling and C. Li. Data mining for direct marketing: Problems and solutions. In *Knowledge Discovery and Data Mining*, pages 73–79, 1998.

[15] M. Maloof. Learning when data sets are imbalanced and when costs are unequal and unknown. In *Workshop on Learning from Imbalanced Data Sets II, ICML*, 2003.

[16] L. M. Manevitz and M. Yousef. One-class svms for document classification. *J. Mach. Learn. Res.*, 2:139–154, 2002.

[17] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In *In AAAI-98 Workshop on Learning for Text Categorization*, 1998.

[18] D. Mladenic and M. Grobelnik. Feature selection for unbalanced class distribution and naive bayes. In *Proc. ICML*, pages 258–267. Morgan Kaufmann Publishers Inc., 1999.

[19] C. Phua, D. Alahakoon, and V. Lee. Minority report in fraud detection: classification of skewed data. *SIGKDD Explor. Newsl.*, 6(1):50–59, 2004.

[20] F. Provost and T. Fawcett. Robust classification for imprecise environments. *Mach. Learn.*, 42(3):203–231, 2001.

[21] B. Schölkof, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson. Estimating the support of a high-dimensional distribution. Technical report, Microsoft Res., 1999.

[22] G. Weiss and F. Provost. Learning when training data are costly: The effect of class distribution on tree induction. *JAIR*, 19:315–354, 2003.

[23] G. M. Weiss. Mining with rarity: a unifying framework. *SIGKDD Explor. Newsl.*, 6(1):7–19, 2004.

[24] I. H. Witten and E. Frank. *Data mining: practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann Publishers Inc., 2000.

[25] G. Wu and E. Y. Chang. Adaptive feature-space conformal transformation for imbalanced-data learning. In *Proc. ICML*, pages 816–823, 2003.

[26] G. Wu and E. Y. Chang. Class-boundary alignment for imbalanced dataset learning. In *Workshop on Learning from Imbalanced Data Sets II, ICML*, 2003.

[27] Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proc. SIGIR*, pages 42–49. ACM Press, 1999.

[28] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Proc. ICML*, pages 412–420. Morgan Kaufmann Publishers Inc., 1997.

[29] Z. Zheng, X. Wu, and R. Srihari. Feature selection for text categorization on imbalanced data. *SIGKDD Explor. Newsl.*, 6(1):80–89, 2004.