Feature Subset Selection Bias for Classification Learning

Surendra K. Singhi SURENDRA@ASU.EDU
Department of Computer Science and Engineering, Arizona State University, Tempe, AZ 85287-8809, USA

Huan Liu HLIU@ASU.EDU
Department of Computer Science and Engineering, Arizona State University, Tempe, AZ 85287-8809, USA

Abstract

Feature selection is often applied to high-dimensional data prior to classification learning. Using the *same* training dataset in both selection and learning can result in so-called feature subset selection bias. This bias putatively can exacerbate data over-fitting and negatively affect classification performance. However, in current practice *separate* datasets are seldom employed for selection and learning, because dividing the training data into two datasets for feature selection and classifier learning respectively reduces the amount of data that can be used in either task. This work attempts to address this dilemma. We formalize selection bias for classification learning, analyze its statistical properties, and study factors that affect selection bias, as well as how the bias impacts classification learning via various experiments. This research endeavors to provide illustration and explanation why the bias may not cause negative impact in classification as much as expected in regression.

1. Introduction

Feature selection is a widely used dimensionality reduction technique, which has been the focus of much research in machine learning and data mining (Guyon & Elisseeff, 2003; Liu & Yu, 2005) and found applications in text classification, Web mining, gene expression micro-array analysis, combinatorial chemistry, image analysis, etc. It not only allows for faster model building by reducing the number of features, but also helps remove irrelevant, redundant and noisy features, thus in turn allows for building simpler and more comprehensible classification models with good classification performance.

A common practice of feature selection is to use the training data D to select features, and then conduct classification learning on the same D with selected features. The use of the same training data for feature selection and classification learning can eventuate some bias in the estimates of the classifier parameters. This bias known as 'feature subset selection bias' or 'selection bias' has been studied in the context of regression in Statistics (Miller, 2002). Statisticians often recommend (Miller, 2002; Zhang, 1992) that one should be careful about its magnitude in selecting features (or attributes) and then building the regression model. In machine learning, (Jensen & Cohen, 2000) discuss different pathologies affecting induction algorithms and how over-searching for the best model can result in biased estimates of the parameters, causing overfitting of the resultant model with deterioration in performance. This work studies *whether the current common practice of using the same training data for feature selection and classification learning is proper or not.*

We formally define feature subset selection bias in the context of Bayesian learning, study in detail the statistical properties of selection bias and various factors that affect the bias. We then discuss the relationship between selection bias and classification learning. This work provides theoretical explanations why selection bias may not have deteriorating effects as severe as expected in regression. We use both synthetic and real-world data to verify our hypotheses and findings, aiming to better understand behaviors of selection bias. We offer and evaluate some options to handle selection bias when the amount of training data is limited. Section 2 discusses the related research. Section 3 introduces and explains selection bias. Section 4 examines some of the factors that affect the bias. Section 5 experimentally investigates effects of selection bias on

Appearing in *Proceedings of the 23^{rd} International Conference on Machine Learning*, Pittsburgh, PA, 2006. Copyright 2006 by the author(s)/owner(s).

classification learning. Section 6 presents an empirical study with text data. Section 7 concludes this work with key findings.

2. Related Research

This section briefly reviews the related work on feature subset selection bias in regression, and differentiates this work from others studying related concepts.

Feature selection bias has been recognized as an important problem in the context of regression (Miller, 2002). The studied regression methods are generally based upon the biased least squares regression coefficients. In (Lane & Dietrich, 1976), the authors carried out simulation studies using 6 independent predictors, all of which had non-zero real regression coefficients. They found that for a sample size of 20, two of the six sample regression coefficients on average almost doubled their true values when those were selected. An important concern about feature selection bias in regression is about the ability to make inference (Zhang, 1992; Chatfield, 1995) with feature selection bias in the built models. This bias can make the relationship between the features and the output (class) look stronger than it should be. Feature selection bias can also adversely affect the prediction ability of the model. This is because the model overfits the data in the presence of selection bias and may not generalize well. The above findings have motivated this work. As we know, regression analysis is employed to make quantitative predictions, while classification is used to make qualitative predictions. It is therefore inappropriate to directly generalize the results from regression to classification.

Feature selection bias is different from 'sample selection bias'. (Zadrozny, 2004) studies the sample selection bias which refers to the fact that the data samples collected for learning may not be randomly drawn from a population. Sample selection bias is an important problem with data collected from surveys and polls where due to the nature of sampling process, data samples from some portions of the population may be over-represented, while other portions may be under-represented or hardly present. The selection bias discussed in this paper occurs due to the interaction between feature selection and classification learning.

In (Jensen & Neville, 2002), 'feature selection bias' is used in the context of relational learning to denote the property that some features may have an artificial linkage with the class, causing them to be selected due to the relational nature of the data. To our knowledge, our work is the first to study feature subset selection bias in classification learning using identical and independently distributed samples in parallel to the similar work associated with regression in (Miller, 2002).

3. Definition of Selection Bias

We now explain and define feature subset selection bias in which '**bias**' refers to an offset or the difference between the true expected value and the estimated expected value (Duda et al., 2000), that is,

$$bias = \mathcal{E}[\theta] - \mathcal{E}[\hat{\theta}]$$

3.1. Example

Consider the following thought experiment. We are given two types of crops, A and B, each with a life span of one month. We assume that the yields of both crops are identical and each has a normal distribution $\sim N(\mu, \sigma)$. Our task is to select the best crop in terms of mean yield[1] and estimate this value based on n-month observations. (1) If we were to just randomly select a crop and estimate its mean yield, then the estimate of the mean yield will be a random variable $\hat{\mu}$ with a probability density function $f(\hat{\mu}) \sim N(\mu, \sigma/\sqrt{n})$. (2) But instead of randomly selecting the crop, if we pick the best one after the n-month observation and then report its yield, then in this case the estimate of the mean yield will follow the distribution of a random variable of the second order statistic[2] for a sample of size 2, i.e., distribution $2f(\hat{\mu})F(\hat{\mu})$, where $F(\hat{\mu})$ is the cumulative distribution function for $f(\hat{\mu})$. Let the expected value of this distribution be μ'. The difference between μ' and μ is selection bias in the estimate of the yield of the crop. (3) If one repeats this experiment just to estimate the yield of the best crop without selection of the best crop, then the average yield reported during that period will again be an unbiased distribution $f(\hat{\mu}) \sim N(\mu, \sigma/\sqrt{n})$ as in (1).

Similarly, in feature selection we select features instead of crops, and selecting features that will help improve classification accuracy replaces searching for the crop that maximizes the mean yield. Subsequently, when the same dataset is used for estimating the classifier parameters, the probability estimates tend to be biased.

[1] Other parameters can also be used to decide the best crop, but for simplicity we use the mean.

[2] Given a sample of N variates $X_1, \ldots, X_N$, when they are reordered as $Y_1 < Y_2 < \ldots < Y_N$. Then Y_i is called the ith order statistic. (Bain & Engelhardt, 1991)

3.2. Definition

We now formally define selection bias for Bayesian learning. Let $\mathbf{X} = (X_1, X_2, \ldots, X_p, X_{p+1}, \ldots, X_q)$ be the set of features, where q is the dimensionality, and Y be the class variable. Also, the relationship between the class Y and the features $\mathbf{X}$ be Bayesian,

$$P(Y|\mathbf{X}) = \frac{P(X_1, \ldots, X_q|Y)P(Y)}{P(X_1, \ldots, X_q)}$$

A feature selection algorithm will select a subset of features, without loss of generality we assume that the following p features are selected.

$$\mathbf{X_A} = (X_1, X_2, \ldots, X_p),\ and$$
$$\mathbf{X_A} \subseteq \mathbf{X}$$

For Bayesian classifiers to make posteriori probability predictions, we estimate class-conditional probabilities. So, if the class of interest is $Y = \omega_j$ and the instance value for the feature subset $\mathbf{X_A}$ is

$$\mathbf{X_A} = \mathbf{v_A}$$
$$\mathbf{v_A} = (v_1, v_2, \ldots, v_p)\ where\ v_i \in Domain(X_i)$$

We denote the class-conditional probability as

$$P(\mathbf{X_A} = \mathbf{v_A}|Y = \omega_j)$$

The expected value of $P(\mathbf{X_A} = \mathbf{v_A}|Y = \omega_j)$ in the original populations be

$$\mathcal{E}[P(\mathbf{X_A} = \mathbf{v_A}|Y = \omega_j)] \tag{1}$$

A feature selection algorithm (FSA) selects a certain subset of features that outperform other features based upon certain selection criteria. Due to this, when the same dataset is used for selecting the feature subset A and for estimating the probability value $P(\mathbf{X_A} = \mathbf{v_A}|Y = \omega_j)$, the estimated probability values tend to get biased. This bias is conditioned upon the feature subset A and the feature selection algorithm (FSA), and hence the estimated conditional probability is represented as

$$\mathcal{E}[P(\mathbf{X_A} = \mathbf{v_A}|Y = \omega_j)|FSA\ selected\ subset\ A] \tag{2}$$

The difference between the conditional expected value in (2) and the unconditional expected value in the original population (1) is selection bias.

Definition (Selection Bias):

$$\mathcal{E}[P(\mathbf{X_A} = \mathbf{v_A}|Y = \omega_j)|FSA\ selected\ subset\ A] - \mathcal{E}[P(\mathbf{X_A} = \mathbf{v_A}|Y = \omega_j)]$$

Table 1. Properties of attributes: (a) **Discrete dataset** - class conditional probabilities (b) **Continuous dataset** - class conditional means and standard deviations

(a)

Attribute Values	+	-
Hot	0.75	0.25
Cold	0.25	0.75

(b)

	+	-
Mean	$\mu_+ = 0.1$	$\mu_- = -0.1$
Standard Deviation	$\sigma_+ = 1$	$\sigma_- = 1$

Empirically, the unconditional expected value (1) or the value without feature selection can be obtained by averaging over all **datasets** containing independent and identically distributed samples; and the conditional expected value in (2) or the biased value can be estimated by averaging over those datasets where FSA selected feature set A. In other words, when we use the same dataset for both feature selection and training, we tend to measure the biased value instead of the unconditional expected value.

4. Selection Bias and Related Issues

We create synthetic datasets with known distributions to understand selection bias, as they allow for controlled experiments and better understanding.

4.1. Synthetic Datasets

Two types of datasets with binary class were generated: one with discrete features, and the other with continuous features. For the discrete data, each feature has two possible values $\{Hot,\ Cold\}$, and the class conditional probabilities are shown in Table 1(a). For the continuous data the class conditional distribution of the attributes is normal and has values as shown in Table 1(b). All features in each type of dataset are independent and have identical properties. By design the datasets are symmetric, i.e., in discrete datasets, $P(x = hot|+) = P(x = cold|-) = 0.75$, while $P(x = cold|+) = P(x = hot|-) = 0.25$. We created discrete and continuous datasets with 100 attributes and 1000 instances ($n = n_- = n_+ = 500$ from each class).

4.2. Illustrating Selection Bias

Using the synthetic dataset a Naïve Bayes classifier (WEKA-Simple Naïve Bayes (Witten & Frank, 2005))

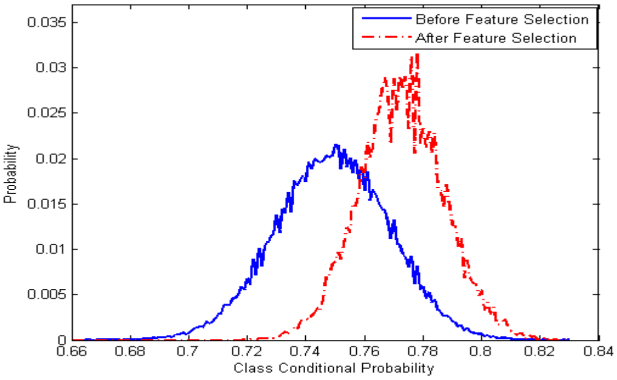


Figure 1. Illustrating selection bias on the discrete data, using the distribution of $P(X|Y)$.

was built and parameters such as class-conditional probabilities for the discrete attributes, and class conditional attribute mean and standard deviation for the continuous attributes were measured. Then, the top 10 attributes were selected using the SVM recursive feature elimination algorithm (Guyon et al., 2002). On the feature-reduced data, a Naïve Bayes classifier was once again built and the different parameters were estimated. This experiment was repeated 500 times with different, randomly generated synthetic datasets. Because of the symmetry of the original datasets, the plots of distribution of $P(x = cold|-)$ and $P(x = hot|+)$ are identical, and so we only show the distribution of $P(x = cold|-)$ before and after feature selection in Figure 1; also the distributions of the $P(x = hot|-)$ and $P(x = cold|+)$ are mirror image of this distribution and hence not shown. The before feature selection distribution implies the distribution of the class conditional probability of the feature in the datasets created from the original population. While the after feature selection distribution implies the distribution of the class conditional probability of the feature in the datasets where the feature was selected. For discrete feature the before feature selection distribution forms a binomial distribution (approximated by a normal distribution $\sim N(p, \sqrt{p(1-p)/n})$, where $p = P(X|Y)$). Also by design, because the attributes are independent and identical, for simplicity we only show the probability estimate of one attribute, but the result can be generalized to the entire set. Selection bias is the difference between the expected values of the distributions of $P(X|Y)$ before and after feature selection. For $P(x = cold|-)$ the distribution of $P(X|Y)$ after feature selection is biased on the higher side, resulting in a positive selection bias. Since the total sum of the class conditional probabilities for a given class (in both biased and unbiased case) is 1, the total selection bias of a class over all values of X is 0. This means that, for binary attributes a positive selection bias for $P(x = cold|-)$ will result in a negative selection bias of equal magnitude for the $P(x = hot|-)$.

For continuous datasets, the underlying distribution was assumed to be Normal, and maximum likelihood estimation was applied to estimate the class-conditional attribute mean and standard deviation. The distribution of values of unbiased negative (-) class-conditional mean will be $\sim N(\mu_-, \sigma_-/\sqrt{n})$, while the distribution of standard deviation will be $\sim N(\sigma_-, \sqrt{2(n-1)}\sigma_-^2/n)$. In our simulation results shown in Figure 2(a), no bias is observed in the estimate of the class-conditional standard deviations (σ_-) before and after feature selection. The distribution of σ_+ is similar. But, for a given attribute, there is a bias in the class conditional attribute means (μ_+ and μ_-); such that the distributions of μ_+ and μ_- are shifted away from each other. Since there is no bias in the expected value of class-conditional standard deviations, the selection bias for any feature value x is directly proportional to the bias in the expected value of the class-conditional attribute mean (see Figure 2(b) and 2(c)). Due to selection bias, there is an illusion that an attribute does a better job of separating the different classes than what it actually does in the original population.

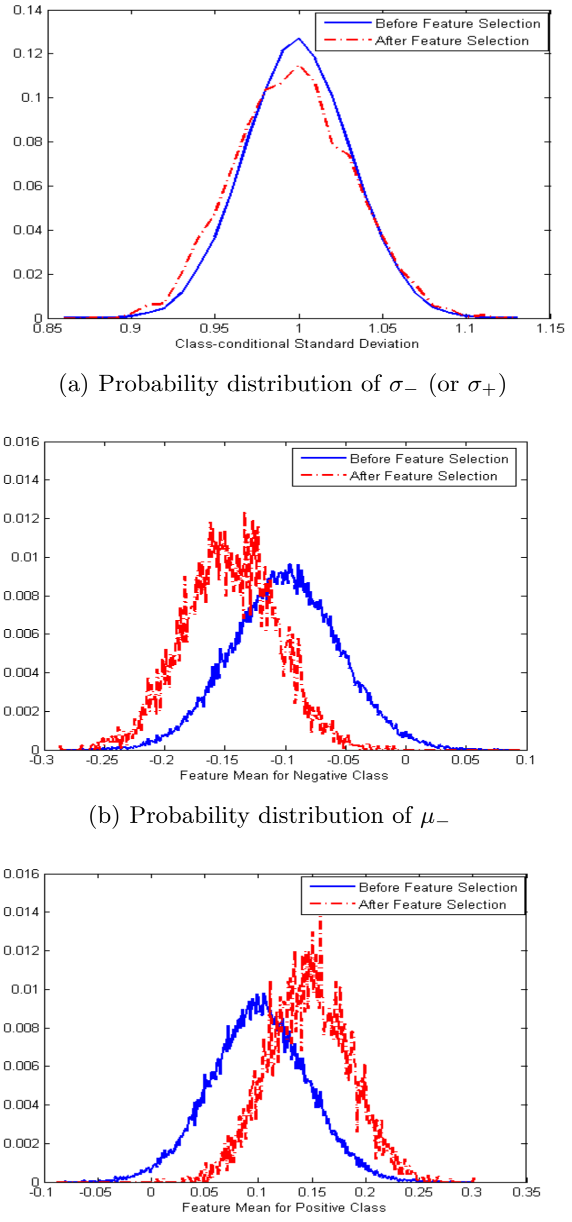
We also conducted experiments using multi-class synthetic datasets and different feature selection algorithms such as Information Gain Attribute evaluation criteria, Relief-F, One-R, Chi-Squared Attribute evaluation available in (Witten & Frank, 2005). The detailed results are not included here due to space constraint. In summary, regardless of feature selection algorithms, bias was observed with varied magnitudes. Irrespective of the number of classes, for discrete features, the attribute conditional probabilities are biased such that the different classes of instances are better separated than they should be. Attribute values having relatively high class-conditional probabilities tend to get positive bias; while attribute values with relatively low class-conditional probabilities are negatively biased. Likewise, when a continuous feature is selected, the attribute means for different classes shift away from each other such that the attribute seems to better isolate the different classes.

4.3. Factors Affecting Selection Bias

We now examine some factors affecting selection bias.

4.3.1. Number of Instances

The first factor is the effect of number of data points or instances on selection bias. We perform similar exper-

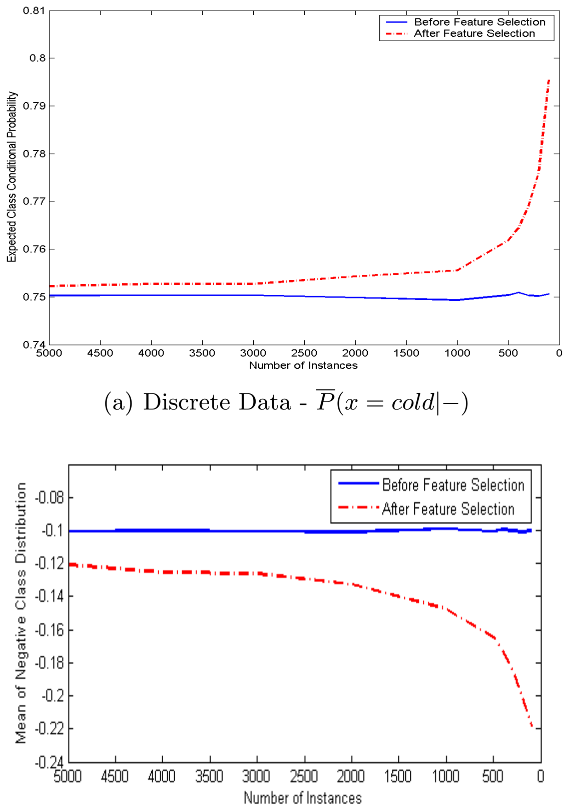


(a) Probability distribution of σ_- (or σ_+)

(b) Probability distribution of μ_-

(c) Probability distribution of μ_+

Figure 2. Illustrating selection bias on the continuous data.



(a) Discrete Data - $\overline{P}(x = cold|-)$

(b) Continuous Data - $\overline{\mu}_-$

Figure 3. Effect on $\overline{\mu}_-$ and $\overline{P}(x = cold|-)$, while varying the number of instances in the dataset.

iments as mentioned earlier by varying the number of instances $(n_+ + n_-)$ (or data size) from $5,000$ down to 100. For the discrete data, Figure 3(a) shows how selection bias increases for $P(x = cold|-)$ as the number of instances decreases. This is because as the number of instances decreases, the variance of the distribution of estimated $P(X|Y)$ increases, resulting in an increased bias. In simulation, we observe that small datasets can cause acute selection bias - an almost vertical spike near datasets with 100 instances. This result is very important especially in the context of microarray gene analysis datasets (Baxevanis & Ouellette, 2005), or text classification where one needs to select features with some hundreds of or fewer documents (Forman, 2003). Similar results are obtained for the continuous dataset as depicted in Figure 3(b).

4.3.2. Effect of σ for Continuous Data

We also observe how changing the values of σ_+ and σ_- in the continuous data affects selection bias. In this experiment, $n_+ = n_- = 500$ remains constant, but σ_+ and σ_- are varied from 0.1 to 2.9 in an increment of

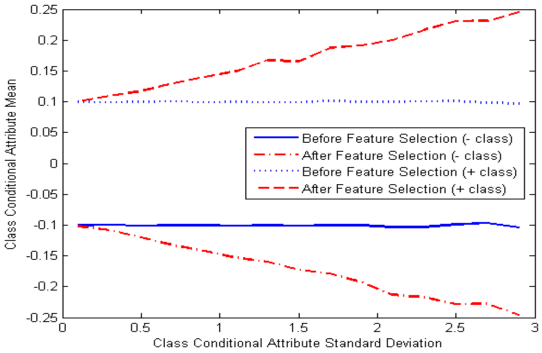


Figure 4. $\overline{\mu}_+$ and $\overline{\mu}_-$ before and after feature selection while varying the class conditional attribute standard deviation.

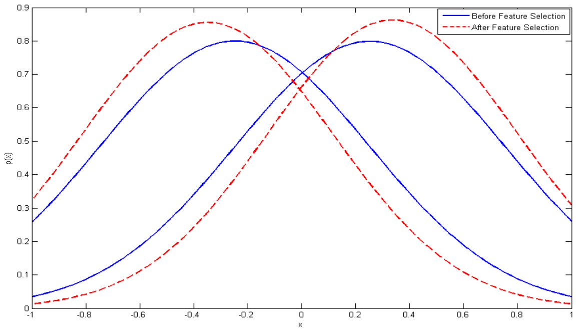


Figure 5. Decision boundaries are almost the same when $\sigma_- = \sigma_+$ and there is an equal amount of selection bias.

0.2. As seen in Figure 4, when the value of σ_+ or σ_- is big, the standard deviation of the distribution of μ_+ or μ_- is also big (unbiased $\mu_+ \sim N(\mu_+, \sigma_+/\sqrt{n_+})$ and unbiased $\mu_- \sim N(\mu_-, \sigma_-/\sqrt{n_-})$). A bigger attribute standard deviation results in more selection bias. In other words, such a feature is more likely selected as it seemingly better separates the classes during feature selection.

5. Selection Bias and Classification

Our discussions so far centered around the estimate of selection bias of a feature-value given a class. Since it is the decision boundary which matters most in classification learning, we now discuss how selection bias affects the decision boundary. Based on the results in Section 4, we evaluate two general cases.

5.1. Case 1: $\sigma_- = \sigma_+$

When the class-conditional standard deviation in both classes is the same, it is likely that there is an equal amount of selection bias in the attributes of both the classes, and hence the decision boundary will remain at its original position, resulting in no change in classification error rate. We set $\sigma_- = \sigma_+ = 0.5$, $\mu_- = -0.5$, and $\mu_+ = 0.5$. Using these parameters, we create a dataset of 200 instances with 200 continuous attributes. We then select the top 5 attributes using SVM recursive feature elimination and record their averaged estimates of μ and σ. The simulation is repeated 100 times, the averaged results are depicted in Figure 5. It shows that (1) the distributions have moved away from their original positions in the opposite direction after feature selection; and (2) the two decision boundaries are almost the same.

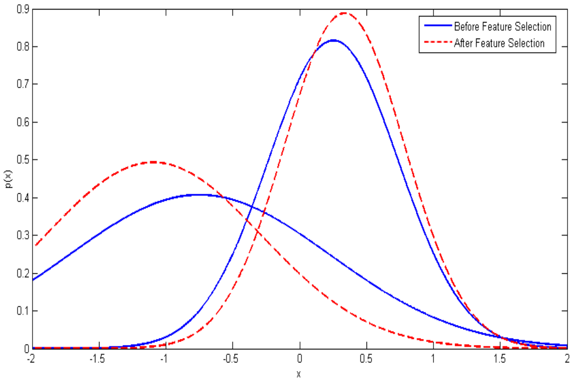


Figure 6. Decision boundaries when $\sigma_- \neq \sigma_+$.

5.2. Case 2: $\sigma_- \neq \sigma_+$

Without loss of generality we assume that $\sigma_- > \sigma_+$. We also assume that the bias in the estimate of μ is directly proportional to the value of σ. Hence, there is a bigger bias in the estimate of μ_- than that of μ_+. We set $\sigma_- = 1$, $\sigma_+ = 0.5$, $\mu_- = -0.75$, and $\mu_+ = 0.25$. Following the same procedure in Case 1, we obtain the plots in Figure 6. The decision boundary after feature selection has a more obvious shift than in Case 1 away from the decision boundary before feature selection, although the difference is still small in absolute value. This is because when the σ is high, there needs to be a larger change in the μ to shift the decision boundary.

In sum, selection bias has limited impact on the change of decision boundary in classification.

6. Empirical Study with Text Data

The experimental results in the previous two sections indicate that (1) the increase of the number of instances usually decreases the selection bias; (2) bigger attribute variance leads to bigger selection bias; and

(3) selection bias has different impacts on classification and on regression. This section focuses on experiments with benchmark text data in (Forman, 2003) containing 19 datasets. For each dataset, we divide it equally into 3 parts (A, B, and C). Parts A and B are used for feature selection and learning. Part C is used for test.

For the first set of experiments, we investigate if using separate datasets for feature selection and classification learning will make a difference. Hence, we compare two models: a *biased* model (**M1**) that uses one part of data (say Part B) for both feature selection and learning; and an *unbiased* model (**M2**) that uses one part of data (say Part A) for feature selection and the other part of data (say Part B) for classification learning. The same process is repeated with the roles of Part A and Part B swapped and the test results on Part C are averaged. The experiment is repeated 25 times, and the results are averaged. Clearly, if there is any difference between the two models, it should be solely due to the selection bias. The test results on Part C are measured using the error rates, the micro and macro F-measures[3] (Witten & Frank, 2005), as the latter two are commonly used criteria for evaluating learning algorithms on text data (Forman, 2003). The macro-averaged F-measure is the arithmetic mean of the F-measure computed for each class, while micro-averaged F-measure is an average weighted by the class distribution. To compare the performance of the biased and unbiased models, we employ the 'corrected resampling t-test' (Nadeau & Bengio, 2003), instead of the 'paired t-test with resampling' which can have unacceptably high Type I error (Dietterich, 1998). Out of the 19 datasets, only 5 datasets were observed to have statistically significant differences at $\alpha = 0.05$. The results are summarized in Table 2. The values after the ± sign are the standard deviations. This first set of experiments confirms that selection bias exists but has limited impact on classification learning contrasting its effect in regression (Miller, 2002).

The above experiments inspired us to ask the following: (1) when the training data is limited, should we use all of it for both feature selection and classification learning? and (2) if we need to stick to the principle that separate data should be used for feature selection and classification learning, do we have an alternative?

One solution achieving some separation between data for feature selection and learning is using one bootstrap sample from the entire dataset for feature selection and another bootstrap sample for classification learning. To verify the efficacy of this bootstrap-based approach, we conducted a simulation experiment with discrete data as the one in Section 4.2 adding the bootstrap model. Figure 7 shows the slight reduction of selection bias. The bootstrap distribution has its mean shifted slightly to the left of the mean of the biased distribution (with bigger variance), resulting in a lower expected value. Hence, it slightly reduces selection bias. We then design the second set of experiments on the 5 datasets that exhibit the effect of selection bias using two *additional* models. One is the *bootstrap model* (**M3**), which uses one bootstrap sample from combined Parts A and B for feature selection, and another bootstrap sample for classification learning. The other model is called the *biased complete model* (**M4**) that uses Parts A and B as one part for both feature selection and learning. The experiment is also repeated for 25 times and averaged test results on Part C are reported and summarized in Table 3. In sum, the averaged values of M4 are consistently better, but the two models are not statistically significantly different based on the corrected resampling t-test ($\alpha = 0.05$). Both M3 and M4 are consistently better than the unbiased model (M2) (in Table 2). Combining the results in the two sets of experiments, we obtain the following: (1) selection bias indeed exists; and (2) under normal circumstances, one does not need to use separate data for feature selection and learning as recommended in the Statistics literature.

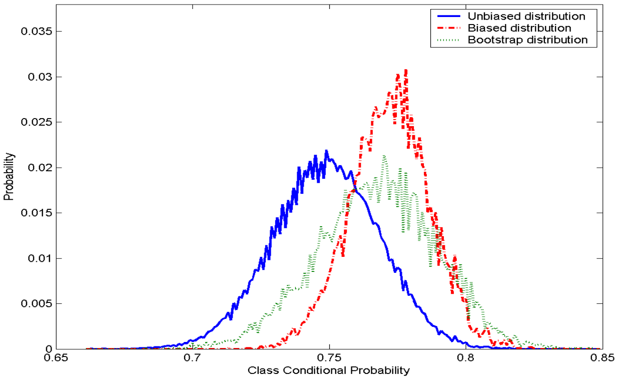


Figure 7. Illustrating the effect of bootstrap in reducing selection bias.

7. Conclusions

This work is motivated by the research on selection bias in regression. We observe selection bias in the context of classification. However, we arrive at the different conclusion: selection bias has less negative effect in classification than that in regression due to the disparate functions of the two. We formally define the feature subset selection bias, and design experi-

[3] F-measure is the harmonic mean of precision and recall.

Table 2. The results of the 5 datasets in which the unbiased model is significantly better than the biased model. Boldfaced entries indicate those with significant difference at $\alpha = .05$.

Dataset	*Biased Model* (M1)			*Unbiased Model* (M2)		
	Error-rate	Micro	Macro	Error-rate	Micro	Macro
la1	19.95±4.59	74.74±15.42	79.23±16.30	**18.59±4.23**	**76.38±15.70**	**80.60±16.54**
la2	18.97±4.26	76.01±15.63	80.57±16.52	**18.04±4.03**	**77.14±15.85**	**81.50±16.70**
tr12	38.11±8.58	52.51±11.46	61.09±12.97	**33.30±7.58**	**56.40±11.94**	**66.34±13.90**
tr31	15.63±3.73	60.85±12.90	84.08±17.29	**14.34±3.54**	62.07±13.19	**85.63±17.61**
wap	28.20±5.97	46.98±9.69	69.46±14.26	27.62±5.89	47.85±9.89	**70.10±14.39**

Table 3. The bootstrap model vs. the biased complete model. No significant difference between the two at $\alpha = .05$.

Dataset	*Bootsrap Model* (M3)			*Biased Complete Model* (M4)		
	Error-rate	Micro	Macro	Error-rate	Micro	Macro
la1	17.53±4.04	78.36±16.16	82.09±16.84	16.62±3.60	79.28±16.25	82.90±16.96
la2	16.07±3.68	80.01±16.46	84.00±17.20	15.84±3.51	80.32±16.48	84.00±17.20
tr12	29.96±8.91	61.93±14.15	69.68±15.60	28.38±7.36	65.36±13.96	71.48±15.19
tr31	11.34±3.47	73.18±15.44	89.30±18.46	10.42±2.61	74.75±15.39	90.30±18.49
wap	22.85±5.03	55.70±11.66	72.26±15.48	22.08±4.86	56.39±11.75	75.97±15.62

ments to study its statistical properties using synthetic datasets and benchmark datasets. This work provides evidence that the current practice of using the same dataset for feature selection and learning is not inappropriate, and provides illustrations why selection bias does not degrade the classification performance as it does in regression.

Acknowledgments

We thank all the reviewers and senior PC for their valuable suggestions. We also thank Subbarao Kambhampati and George Runger for reviewing initial versions of this work and giving valuable suggestions.

References

Bain, L. J., & Engelhardt, M. (1991). *Introduction to probability and mathematical statistics.* Duxbury Press. 2nd edition.

Baxevanis, A., & Ouellette, B. (2005). *Bioinformatics - a practical guide to the analysis of genes and proteins.* Wiley. 3rd edition.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series A*, *158*, 419–466.

Dietterich, T. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*, 1895–1924.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification.* Wiley. 2nd edition.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, *3*, 1289–1305.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*, 389–422.

Jensen, D., & Neville, J. (2002). Linkage and autocorrelation cause feature selection bias in relational learning. *ICML'02*. Morgan Kaufmann.

Jensen, D. D., & Cohen, P. R. (2000). Multiple comparisons in induction algorithms. *Machine Learning*, *38*, 309–338.

Lane, L. J., & Dietrich, D. L. (1976). Bias of selected coefficients in stepwise regression. *Proceedings of Statist. Comput. Section* (pp. 196–200). Americal Statistical Association.

Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans. Knowl. Data Eng.*, *17*, 491–502.

Miller, A. (2002). *Subset selection in regression.* Chapman & Hall/CRC.

Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, *52*, 239–281.

Witten, I., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques.* Morgan Kaufmann. 2nd edition.

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. *ICML'04*. ACM.

Zhang, P. (1992). Inference after variable selection in linear regression models. *Biometrika*, *79*, 741–746.