2009 Ninth IEEE International Conference on Bioinformatics and Bioengineering

AESNB: Active Example Selection with Naïve Bayes Classifier for Learning from Imbalanced Biomedical Data

Min Su Lee[a], Je-Keun Rhee[b], Byoung-Hee Kim[a], and Byoung-Tak Zhang[a,b]
[a] School of Computer Science and Engineering, [b] Graduate Program in Bioinformatics
Seoul National University
Seoul, Korea
{mslee, jkrhee, bhkim, btzhang}@bi.snu.ac.kr

***Abstract*—Various real-world biomedical classification tasks suffer from the imbalanced data problem which tends to make the prediction performance of some classes significantly decrease. In this paper, we present an active example selection method with naïve Bayes classifier (AESNB) as a solution for the imbalanced data problem. The proposed method starts with a small balanced subset of training examples. A naïve Bayes classifier is trained incrementally by actively selecting and adding informative examples regardless of the original class distribution. Informative examples are defined as examples that produce high error scores by the current classifier. We examined the performance of AESNB algorithm by using five imbalanced biomedical datasets. Our experimental results show that the naïve Bayes classifier with our active example selection method achieves a competitive classification performance compared to the classifier with sampling or cost-sensitive methods.**

Keywords-active example selection; imbalanced data problem; naïve Bayes classifier; resampling; cost-senseitive learning

I. INTRODUCTION

Classification is one of popular data mining methods in biomedical research which requires training examples to predict a target class of unseen examples. In the classification task, imbalanced data problem frequently causes highly imbalanced prediction performances among classes due to skewed class distribution [1],[2]. One of major reasons for the problem is insufficiency of the absolute amount of examples of some classes for training a classifier. This is fundamentally caused by the intrinsic rarity of the cases or by limitations on data collection process such as high cost or privacy problems.

Most biomedical data often have skewed class distribution. The examples of interesting class such as a disease and interacting protein pairs are generally rare, and information is insufficient to learn discriminating patterns of the interesting class. However, the design principle of most classification algorithms optimizes overall accuracy. It causes lower prediction performance in the minority class than the majority class. As a result, the minority class is more likely to be misclassified than the majority class and the false positive rate of the minority class can be extremely high.

To overcome the imbalanced data problem, many researchers in biomedical domain have attempted to create more balanced class distributions using various sampling techniques and ensemble methods [3]-[5]. These methods often improved prediction performances of rare class. However, they also brought about some problems (e.g. information loss or data redundancy).

In this paper, we present an active example selection strategy with naïve Bayes classifier (AESNB). To solve the imbalanced data problem, the AESNB attempts to adjust the number of examples among classes to improve classification performance rather than makes equal the number of examples for each class. The AESNB starts with small number of balanced training examples and actively adds informative examples to improve prediction performance regardless of the original class distribution. Although the first approximation may not be satisfactory, we can use this knowledge to select next set of examples which efficiently improve the current estimation. Through this incremental learning process, the final classifier is trained with more task-relevant and proper amount of examples. In this new learning scheme the classifier is trained on incrementally selected examples, rather than on all the available data. Our empirical results show that AESNB can be a more effective alternative to resampling and cost-sensitive methods for solving imbalanced learning problem.

The rest of the paper is organized as follows: Related works are described in section II. We describe the detail idea of active example selection with naïve Bayes classifier algorithm in section III. The experimental results of the proposed method comparing to other methods on selected datasets is shown in section IV. We conclude the proposed method and experimental results in the final section V.

II. RELATED WORKS

A. Imbalanced Data Problem

There have been many studies to solve the imbalanced data problems. One simple way is to make the number of examples for each class equal by under- or oversampling.

In Random Under Sampling (RUS) method, the examples of the majority class are randomly discarded. The method can reduce the time complexity since it extracts a small part of examples from the majority class. However, it is possible to remove certain significant examples and it has

978-0-7695-3656-9/09 $25.00 © 2009 IEEE
DOI 10.1109/BIBE.2009.63

a potential disadvantage of distorting the distribution of the majority class. If the patterns sampled from the majority class do not represent the original distribution, it may decrease the classification performance. The potential drawbacks come true when the number of minority class patterns is very small.

The Random Over Sampling (ROS) method works in the way that examples of the minority class are randomly duplicated from the dataset to balance the number of each dataset. Since oversampling does not lose the information on whole data patterns, it can achieve relatively high classification accuracy. However, it can lead to overfitting problems and long training time, since the number of data used in training is much larger than the number of the original patterns.

It is possible to combine these two kinds of sampling methods. In this way, it produces a random subsample of a dataset by extraction with replacement under specified class distribution. Although these approaches are also interesting ideas trying to solve the imbalanced data problem, any method does not improve the classifier remarkably from imbalanced data. Besides to these methods, many other intelligent resampling techniques are proposed to solve the problem [6]-[9]. Recently, some experiments are performed to compare these resampling-based approaches to solve the imbalanced data problem. However, the performances of 'intelligent' sampling techniques are not generally superior to simple sampling techniques [10], [11].

Another way to solve the class imbalance problems is to modify leaning processes or methods. Cost-sensitive learning is a method for these purposes [12]. It dictates that misclassified examples originally belonging to the minority class receive larger penalty than those belonging to the majority class. In the learning process, it can handle the data imbalance problems without changing the original data distribution by modifying a cost function. In fact, it has been often reported that cost-sensitive learning outperforms random resampling methods in several cases [12], [13]. When data is highly imbalanced, however, its effect on the classification performance is not as good as that of undersampling method or oversampling method, since cost-sensitive learning does not modify the class distribution of the data.

That is, in spite of many approaches from various views, the limitations still remain to solve the imbalanced learning problems. Therefore, it is necessary to propose a new approach to overcome these problems.

B. Imbalanced data problems in biomedical tasks

Since available biomedical datasets for training and validating classification models often have imbalanced class distribution, there have been many related studies to handle imbalanced data problem.

One of frequently used methods is to divide the original dataset into a balanced dataset for training and an imbalanced dataset for validating or testing the trained model. The method was used to diagnose myocardial perfusion using cardiac SPECT (Single Proton Emission Computed Tomography) images and to predict polyadenylation signals in human sequences [14], [15].

Random undersampling of the majority class also can be easily applied. In discrimination task of deleterious nsSNPs (nonsynonymous Single Nucleotide Poly- morphisms) from neutral nsSNPs with imbalanced training dataset, by applying random undersampling method combined with decision tree algorithm, prediction performances were improved [5]. The undersampling method can be combined with an ensemble machine. An Ensemble of under-sampled classifiers was constructed for predicting the activity of drug molecules based on structural characteristics of compounds and for predicting glycosylation sites in genomic sequences [3], [4].

The various cost-sensitive learning methods were also applied. To predict protein β-turn structure, a cost-sensitive k-nearest neighborhood algorithm was used [16]. The method manually tuned the number of minimally required closest training fragments by considering the imbalanced ratio of the natural occurrence of β-turns and non-β-turns. Cost-sensitive decision tree algorithm was also used to improve peptide-MHC class I binding prediction [17]. For imbalanced, multi-class and multi-labeled dataset such as protein localization dataset, SVDD (Support Vector Data Description), which is a kind of one-class classification algorithm, was applied [18].

III. ACTIVE EXAMPLE SELECTION WITH NAÏVE BAYES CLASSIFIER FOR IMBALANCED DATA PROBLEM

We propose an active example selection method to solve the imbalanced data problem in biomedical classification issues. Although machine learning algorithms are useful for data classification, qualified training dataset is necessary to learn a classifier with good performance. The qualified dataset can be explained in terms of the nature and size of the training set. Because some of training examples are contradictory or redundant, it usually takes a long time to train a classifier with all the given examples. If there are no guarantees that classification performance is improved when the size of training set are increased, it means that not only training time is increased but also many of training examples are wasteful for learning. Therefore it is necessary to choose examples which are most likely helpful to solve the problem. If the training dataset contains imbalanced class labels, by extracting informative training examples from imbalanced dataset, we can improve the classification performance and, at the same time, solve the imbalanced data problems.

The active example selection strategy was originally proposed as a method to accelerate training speed of multilayer neural networks by starting learning with a small subset training data and adding critical examples incrementally [25]. We apply the active example selection strategy to solve the imbalanced data problem by adjusting

the number of training examples among classes based on classification error.

The active example selection (AES) strategy can be applied as a wrapper learner of classification algorithms. The AES strategy works well with classification algorithms which have following properties:

- small number of parameters to be tuned
- parameter estimation is possible with small amount of data
- short training time

In this study, we applied the active example selection with a naïve Bayes classifier. Naïve Bayes classifier is a simple probabilistic classifier based on Bayes' theorem with independence assumptions among attributes. It selects the class label c^* with the maximum probability which is calculated according to the following equation,

$$c^* = \arg\max_{c \in C} P(c) \prod_i P(a_i \mid c). \quad (1)$$

where, a_i represents the i-th attribute. In spite of its naïve design and over-simplified assumptions, naïve Bayes classifier has shown good performances in many complex real-world problems.

The proposed incremental learning procedure is summarized in Fig. 1. The total training dataset (D_{total}) is divided into training dataset (D_{train}) for training a naïve Bayes classifier and validation dataset for validating the trained classifier ($D_{validation}$). The D_{train} is initialized with randomly selected a small set of examples which includes the same number of examples from each class. The $D_{validation}$ is initialized with the rest of the total training dataset.

We start learning a naïve Bayes classifier with a small balanced subset of training dataset. Although the first approximation with the initial dataset, $D_{initial}$, may not be satisfactory, we can use the current classifier to select the next set of examples. To make up the weaknesses of the current classifier more effectively, some informative examples are selected from $D_{validation}$ and added to D_{train} actively and incrementally.

The informative examples can be selected by measuring classification error scores for each example in $D_{validation}$ with the current classifier. The classification error score of an example is conceptually defined as 'the distance from the decision boundary when it is misclassified'. In the case of naïve Bayes classifier for binary class problem, it is defined as the difference between the incorrectly predicted probability and the threshold for decision (usually 0.5). Top q examples which have high error scores are selected (D_{error}), and added to training dataset regardless of their original class labels. We can efficiently improve the overall classification performance, as the examples selected from $D_{validation}$ are added to D_{train} iteratively.

When the error rates for validation set are zero or the whole validation examples are used for training, the learning process is terminated. Through the incremental learning process, AESNB can achieve a reasonable performance using only a subset of examples.

IV. EXPERIMENTS AND EVALUATION

A. Experimental datasets

We study performances of AESNB on several real-world biomedical benchmark datasets. We consider binary classification problems in this study. The overview of the datasets is given in Table I. From the interestingness of a class, an interesting target class is called the positive class and a normal class is called the negative class. From the quantity of examples of a class, a class which has more examples than the other is called the majority class and the other is called the minority class.

The Diabetes dataset is extracted from a Pima Indians diabetes database which consists of 268 diabetes examples and 500 healthy examples [20]. The diagnostic target class indicates whether the patient shows signs of diabetes according to World Health Organization criteria. The 8 real-valued features are derived from geographic data and characteristics of patients. The WDBC is the Wisconsin diagnostic breast cancer dataset which includes 212 malignant examples and 357 benign examples [21]. The 30

Data preparation
Data for learning: D_{total}, #(D_{total}) = n
  Training data: D_{train} = {}
  Validation data: $D_{validation}$ = D_{total}
Data for test: D_{test}

Parameter setting
Initial size of training data: c x p
($1 \le p \le n_{minority}$, $n_{minority}$: # of examples in the minority class, c: # of target classes)
Incremental size of training data: q
($0 < q < n-(c \times p)$)

Data initialization
For each class in $D_{validation}$, randomly select p examples ($D_{initial}$)
$D_{train} = D_{initial}$
$D_{validation} = D_{validation} - D_{initial}$

Learning & active example selection
do {
Training
  Train a naïve Bayes classifier with the D_{train}
Validation & updating datasets
  Validate the resulting classifier with $D_{validation}$
  Sort misclassified examples by error score
  Select top q examples with high error score regardless of class labels (D_{error}). If the number of misclassified examples is less than q, select all misclassified examples as D_{error}.
  $D_{train} = D_{train} \cup D_{error}$
  $D_{validation} = D_{validation} - D_{error}$
} while ($D_{validation}$!= {} && D_{error} != {})

Test
Test the resulting classifier with D_{test}

Figure 1. Pseudo-code of AESNB learning procedure

TABLE I. OVERVIEW OF DATASETS

Dataset	# of Examples	# of Features	Class Distribution	Imb. Ratio
Diabetes	768	8	Positive 268 Negative 500	1:1.87
WDBC	569	30	Positive 212 Negative 357	1:1.68
Parkinson's disease	194	22	Negative 47 Positive 147	1:3.13
Colon	62	2000	Negative 22 Positive 40	1:1.82
Promoter	7,047	10	Positive 1,839 Negative 5,208	1:2.83

real-valued features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. The Parkinson's disease is the Oxford Parkinson's disease detection dataset which includes 47 healthy examples and 147 Parkinson's disease examples [22]. The Parkinson's disease can be classified by voice disorder detection. The 22 real-valued features are composed of biomedical voice measurements. The Colon is a colon cancer microarray dataset which has 2000 features and includes 40 tumor examples and 22 normal examples [23]. The Diabetes, WDBC, Parkinsons, and Colon dataset are benchmark datasets from the popular UCI Machine Learning Repository. The Promoter dataset is the core human promoter prediction dataset to develop the PromSearch system by artificial neural networks [24]. The dataset includes 1,839 promoter examples and 5,208 negative examples and has a set of nominal and real-valued features which represent probabilities of motives around the core promoter region. We have considered datasets with diversity in the number of examples. The smallest dataset has 62 examples, while the largest dataset has 7,047 examples.

B. Experiments and Evaluation

We conducted experiments to compare the proposed AESNB with default naïve Bayes classifier algorithm (NB), naïve Bayes classifier with three random sampling techniques (NB with Sampling), and cost-sensitive naïve Bayes classifier (CSNB). Since it is known that intelligent sampling techniques show inferior performances than simple random sampling in general [10], [11], we only considered simple sampling methods which are random oversampling (ROS), random undersampling (RUS), and mixture of random over and undersampling (ROUS). ROUS produces a resampled example set with a uniform class distribution and an equal number of examples of original dataset by undersampling for the majority class and oversampling for the minority class.

For all dataset except Promoter dataset, we set p, the number of initial training example per class as 1 and the incremental chunk size (q) as 2 in AESNB. In the case of the Promoter dataset, which is the largest dataset, we set p as 10 and q as 50 to speed up learning.

To evaluate the performance of classification algorithms, overall accuracy, AUC (Area Under the ROC Curve), geometric mean, true positive rate, and F-measure were calculated. When dataset has highly skewed class distribution, the overall accuracy is not an appropriate measure. Since the overall accuracy tends to be overwhelmed by the prediction power for the majority class, the performance comparison with overall accuracy is very misleading in the imbalanced data learning case. Hence we used the AUC and the geometric mean which give balanced evaluation by incorporating measures of both positive and negative classes with equal weights. These measures give higher values only when the classifier predicts accurately on both classes. We also use the true positive rate (TPR) and F-measure as the balanced evaluation measures which represent the classification performances per class.

More specifically, AUC measures the area under the ROC curve. The ROC curve is a technique for visualizing, organizing, and selecting classifiers based on their performance [25]. The AUC incorporates the trade-off relation between a true positive rate and a false positive rate into a single value. In the imbalanced data problem, the AUC have been widely used. Other measures are defined as following in the binary classification case:

$$\text{Geometric Mean} = \sqrt{\text{TPR} \times \text{TNR}} \tag{2}$$

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4}$$

$$\text{F} - \text{meausre} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

where TP is the number of true positive, FP is the number of false positive, FN is the number of false negative, and TN is the number of true negative. Geometric mean is a geometric average of true positive rates and true negative rates which are related to a point on the ROC curve. Precision is a measure of exactness, and recall is a measure of completeness. F-measure is a harmonic average between prediction and recall which can measure the goodness of a learning algorithm on the current class. Note that TPR and precision are same measures with the same meaning with 'positive accuracy'.

To estimate general performances of AESNB, for each combination of 5 datasets and 6 learning strategies, 10 different runs of 10-fold cross validation were executed. Each training experiment starts with randomly selected initial training data. The performances of total 100 runs for each combination are evaluated and averaged. The experimental results are shown in Table II ~ Table V. Due

TABLE II. COMPARISON OF AUC

Dataset	NB	NB with Sampling			CSNB	AESNB
		ROS	*RUS*	*ROUS*		
Diabetes	0.817	0.818	0.814	0.811	0.817	**0.828**
WDBC	0.982	0.982	0.983	0.983	0.982	**0.990**
Parkinson's disease	0.857	0.858	0.853	0.853	0.857	**0.875**
Colon	0.632	0.628	0.670	0.647	0.632	**0.755**
Promoter	0.984	0.984	0.984	0.984	0.984	**0.985**

TABLE III. COMPARISON OF GEOMETRIC MEAN (%)

Dataset	NB	NB with Sampling			CSNB	AESNB
		ROS	*RUS*	*ROUS*		
Diabetes	70.9	73.1	73.4	73.2	73.2	**73.7**
WDBC	92.6	92.6	92.6	92.7	92.6	**95.0**
Parkinson's disease	75.1	74.8	74.9	74.2	74.9	**77.2**
Colon	58.9	58.5	61.4	61.8	58.9	**70.7**
Promoter	93.0	**93.7**	**93.7**	**93.7**	**93.7**	93.0

TABLE IV. COMPARISON OF TRUE POSITIVE RATE PER CLASS (%)

Dataset	Class	NB	NB with Sampling			CSNB	AESNB
			ROS	*RUS*	*ROUS*		
Diabetes	Pos	59.8	68.5	68.8	**69.7**	68.9	66.4
	Neg	**84.1**	78.0	78.2	76.8	77.8	81.8
WDBC	Pos	89.7	89.9	90.2	90.4	90.0	**92.4**
	Neg	95.5	95.4	95.1	95.1	95.3	**97.9**
Parkinson's disease	Neg	91.0	91.7	89.7	87.8	**92.1**	90.0
	Pos	62.0	61.1	62.5	62.7	61.0	**66.2**
Colon	Neg	72.2	69.7	**75.3**	71.7	72.2	68.5
	Pos	48.0	49.3	50.3	53.3	48.0	**73.3**
Promoter	Pos	91.0	**94.1**	94.0	**94.1**	**94.1**	90.4
	Neg	95.0	93.3	93.3	93.3	93.3	**95.6**

TABLE V. COMPARISON OF F-MEASURE PER CLASS (%)

Dataset	Class	NB	NB with Sampling			CSNB	AESNB
			ROS	*RUS*	*ROUS*		
Diabetes	Pos	63.0	65.3	65.7	65.5	65.5	**66.2**
	Neg	81.8	80.0	80.2	79.5	79.9	**81.9**
WDBC	Pos	90.9	90.9	90.9	91.0	91.0	**94.1**
	Neg	94.7	94.7	94.7	94.7	94.7	**96.6**
Parkinson's disease	Neg	59.1	58.9	58.7	58.1	59.0	**61.4**
	Pos	74.7	74.0	74.9	74.7	74.0	**77.7**
Colon	Neg	53.1	52.1	55.2	55.1	53.1	**62.6**
	Pos	56.2	56.8	58.1	59.7	56.2	**75.0**
Promoter	Pos	88.8	88.3	88.3	88.3	88.4	**89.1**
	Neg	95.9	95.5	95.5	95.5	95.5	**96.1**

to the space limitations we only provide several figures of the results with Diabetes dataset.

Table II and Table III summarize overall classification performances in terms of AUC and geometric mean, respectively. AESNB outperforms other methods for the five benchmark datasets. Table IV and Table V show classification performances per class. For each dataset, upper line and under line are corresponding to minority class and majority class respectively. In Parkinson's disease and Colon datasets, positive classes are majority classes, while in Diabetes, WDBC, and Promoter datasets, positive classes are minority classes. However, in the Parkinson's disease and Colon datasets, prediction performances of majority class with naïve Bayes classifier algorithm are lower than those of minority class in terms of true positive rates (Table IV). In such case, AESNB makes remarkable improvement in true positive rates of the majority class. Table V shows outstanding performances of AESNB in F-measure.

Fig. 2 demonstrates an example of training, validation and test curves of AESNB. The plot is drawn with total training data (D_{total}), validation data ($D_{validation}$), and independent test data (D_{test}) from one of 100 runs with Diabetes dataset. The AESNB learning is terminated at 105 -th iteration step, and the validation AUC converges into 1. In the early part of incremental training, the training, validation and test curves are oscillated. However, after mid-part, they are increased steadily. Even though the learned classifiers based on AESNB seem to be overfitted to validation datasets, as we can see in Fig. 2, the classifiers are not overfitted to total training datasets.

The iterative procedure of AESNB learning can be terminated when the validation dataset is exhausted or there is no misclassified example in validation dataset. However, all classifiers derived from 10 different runs of 10-fold cross-validation with 5 datasets are terminated in case with the absence of misclassified examples. Data efficiency of AESNB training can be demonstrated with Fig. 3 and Table VI. Fig. 3 shows training data distribution for each AESNB iteration step. The right-most bar indicates total Diabetes data distribution. Even though only one-third of training examples are used to train AESNB classifier, the classifier shows outstanding prediction performances as we can show in Table II ~ Table V. Table VI indicates the average number of used examples for training with AESNB. Data

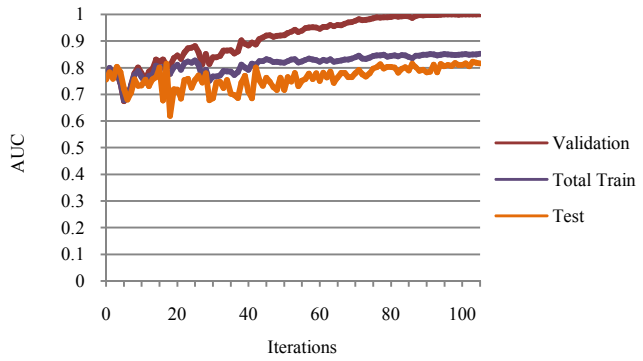

Figure 2. Training, validation and test curves in Diabetes dataset

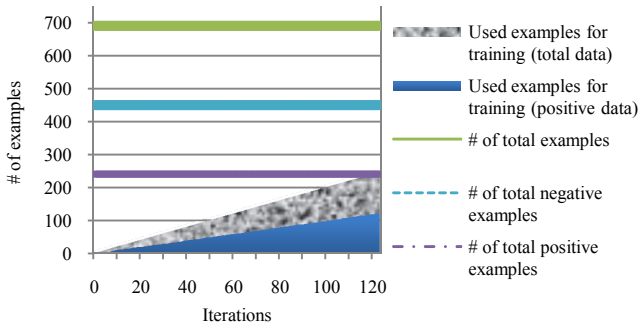

Figure 3. Incremental patterns of trianing data

TABLE VI. THE NUMBER OF USED EXAMPLES FOR TRAINING WITH AESNB

Dataset	Total training examples (imbalance ratio)	Used examples for training (imbalance ratio)	% of used examples for training
Diabetes	Pos 241.2 Neg 450.0 (1:1.9)	Pos 116.7 Neg 117.1 (1:1.0)	33.8%
WDBC	Pos 190.8 Neg 321.3 (1:3.1)	Pos 23.9 Neg 17.6 (1.36:1)	8.1%
Parkinson's disease	Neg 42.3 Pos 132.3 (1:3.1)	Neg 11.6 Pos 48.1 (1:1.4)	58.2%
Colon	Neg 19.0 Pos 36.0 (1:1.8)	Neg 13.5 Pos 19.0 (1:1.4)	58.2%
Promoter	Pos 1655.1 Neg 4687.2 (1:2.8)	Pos 176.8 Neg 233.8 (1:1.3)	6.5%

efficiency percentage of AESNB range from 6.5% to 58.2%. By the active example selection process, AESNB effectly trims less important or redundant training examples.

The high false negative rate of a class is an indicator of a severe imbalanced data problem. To characterize AESNB operation for handling the imbalanced data problem, we perform a correlation analysis between the number of added examples and false negative rates on an iterative AESNB learning process. The false negative rate indicates the proportion of positive examples that were erroneously classified as negative. By investigating the correlation between them for each class, we can check whether the active example selection incrementally expands training examples reflecting the weakness of the current estimation or not. In the Diabetes, WDBC, Parkinson's disease, and Colon dataset, although a range of the number of added examples is from 0 to 2 and a range of the false negative rate is from 0 to 100, the Pearson's correlation coefficients ranges from 0.20 to 0.57. In the Promoter dataset, the range of added example is from 0 to 50, and the Pearson's correlation coefficients ranges from 0.70 to 0.78. The positive correlations indicate that for each iterative learning step, the AESNB procedure identifies the weakness of the current classifier and adds critical examples which have high error score to the current classifier regardless of an original class distribution.

V. DISCUSSION AND CONCLUSION

Our empirical results raise many interesting issues to be discussed. First, the proposed AESNB shows competitive prediction performance against random sampling method and cost-sensitive learning strategy, even though it uses only subset of the entire training dataset.

Second, we learn that the final imbalance ratio which is usually decreased in other resampling methods, is sometimes increased or even reversed because AESNB try to achieve better prediction performance. We can explain it by comparing previous methods with our AESNB method in used training dataset distribution. Most methods to solve the imbalanced learning problems are based on balancing the training data distribution. However, the proposed AESNB method does not consider the balance of class distribution but only consider the prediction performance. This also shows that there are positive correlations between false negative rates and added examples. By adding critical examples of given validation dataset, AESNB method effectively tries to cover up weak points of the currently existing classifier.

Table IV shows that the prediction performance of the minority class is not always lower than the majority class' in terms of true positive rates. The imbalanced class distribution does not always induce the imbalanced classification performance. Table VI depicts changes for the imbalance ratios between original training datasets and used training examples in the AESNB method. In the WDBC dataset, an imbalance direction between the total dataset and used training examples is reversed. In the Parkinson's disease dataset and the Colon dataset, AESNB method selects more positive examples than the negatives. In the two datasets, the general performances of majority class's true positive rates are lower than the minority class. It is depicted in Table IV. The learning pattern of abnormal class is often harder than that of normal class. To learn the distinguishing patterns of the positive class, it seems that the more number of the positive training examples is sometimes needed than the negative examples. If the training examples of negative (i.e. less-interesting) class are derived from several groups, to capture the characteristics of negative class, more negative examples are needed. Hence AESNB procedure selects more number of the negative examples than the positive examples. The negative examples of the Promoter dataset are artificially generated from various biologically meaningful regions in DNA sequences. Thus the more number of negative training examples are used than positive examples to learn the discriminative patterns between two classes.

Our study shows that real imbalanced data problem is not an imbalanced class distribution but an imbalanced prediction performance. Thus, the most important factor to solve the imbalanced data problem is not providing more

balanced dataset but selecting task-relevant examples for improving the prediction performance.

In this paper, we proposed an active example selection with naïve Bayes classifier to solve the imbalanced learning problem. Unlike previous resampling techniques, our proposed method is not using uniform class distribution. Rather, the proposed method selects and adds critical examples regardless of original class distribution for reflecting the weakness of the current classifier. By focusing on the informative examples, the AESNB improves imbalanced prediction performances with the subset of total training dataset, and it also achieves competitive prediction performances.

Our empirical results by using five real-world biomedical datasets help us to conclude that the AESNB performs better than other popular resampling or cost-sensitive learning methods in dealing with the imbalanced learning problem.

Our method can be used to select discriminative or representative examples of some classes. In addition, the AESNB can be applied for an effective active learning for the huge amount of dataset.

ACKNOWLEDGMENT

This work was supported in part by the Ministry of Education and Human Resources Development (BK21-IT Program) and the Ministry of Science and Technology Foundation (National Research Laboratory Program). The authors thank to Dr. Sun Kim for insightful discussions.

REFERENCES

[1] N. V. Chawla, N. Japkowicz, and A. Kolcz, "Editorial: special issue on learning from imbalanced data sets", *SIGKDD Explorations*, vol. 6, no. 1, June 2004, pp.1-6..

[2] G. M. Weiss, "Mining with rarity: a unifying framework", *SIGKDD Explorations*, vol.6, no. 1, June 2004, pp.1-6.

[3] G. -Z. Li, H. -H. Meng, W. -C. Lu, J. Y. Yang, and M. Q. Yang, "Asymmetric bagging and feature selection for activities prediction of drug molecules", *BMC Bioinformatics*, vol. 9(suppl. 6), Aug. 2007, article no. S7.

[4] C. Caragea, J. Sinapov, A. Silvescu, D. Dobbs, and V. Honavar, "Glycosylation site prediction using ensembles of support vector machine classifiers", *BMC Bioinformatics*, vol. 8, Nov. 2007, article no. 438,.

[5] R. J. Dobson, P. B. Munroe, M. J. Caufield, and M. AS. Saqi, "Predicting deleterious nsSNPs: an analysis of sequence and structural attributes", *BMC Bioinformatics*, vol. 7, Apr. 2006, article no. 217.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority oversampling technique", *Journal of Artificial Intelligence Research*, vol. 16, 2002, pp. 321-357

[7] H. Han, W. Y. Wang, and B. H. Mao, "Borderline SMOTE: a new oversampling method in imbalanced data sets learning", In Proc. of the *Int. Conf. on Intelligent Computing, Lecture Notes in Computer Science*, vol. 3644, 2005, pp.878-887.

[8] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one sided selection", In Proc. *of the 14th Int. Conf. on Machine Learning*, 1997, pp.179-186.

[9] R. Barandela, R. M. Valdovinos, J. S. Sanchez, and F. J. Ferri, "The imbalanced training sample problem: under or oversampling?", *In Joint IAPR Int. Workshops on Structural, Syntactic and Statistical Pattern Recognition, Lecture Notes in Computer Science*, vol. 3138, 2004, pp.806-814.

[10] J. V. Hulse, T. M. Khoshgoftaar, and A. Napolitano, "Experimental perspectives on learning from imbalanced data", In Proc. of the *24th international conference on machine learning*, 2007, pp.935-942.

[11] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data", *SIGKDD Explorations*, vol. 6, no.1, , June 2004, pp.20-29.

[12] P.Domingos, "Metacost: a general method for making classifiers cost-sensitive", In Proc. of *Int. Conf. on Knowledge Discovery and Data Mining*, 1999, pp.155-164.

[13] N. Japkowicz, and S. Stephen, "The class imbalance problem: a systematic study" *Intelligent Data Analysis*, vol. 6, no. 5, 2002, pp.203-231.

[14] L. A. Kurgan, K. J. Cios, R. Tadeusiewicz, M. Ogiela, and L. Goodenday, "Knowledge discovery approach to automated cardiac SPECT diagnosis", *Artificial Intelligence in Medicine*, vol. 23, no. 2, Oct. 2001, pp.149-169.

[15] H. Liu, H. Han, J. Li, and L. Wong, "An in-silico method for prediction of polyadenylation signals in human sequences", in Proc. *14th International Conference on Genome Informatics*, vol. 14, Dec. 2003, pp.84-93.

[16] S. Kim, "Protein β-turn prediction using nearest-neighbor method", *Bioinformatics*, vol. 20, no. 1, Jan. 2004, pp. 40-44.

[17] A. P. Sales, G. D. Tomaras, and T. B. Kepler, "Improving peptide-MHC class I binding prediction for unbalanced datasets", *BMC Bioinformatics*, vol. 9, no.1, Sep. 2008, article no. 385.

[18] K. Y. Lee, D. W. Kim, D. K. Na, K. H. Lee, and D. Lee, "PLPD: reliable protein localization prediction from imbalanced and overlapped datasets", *Nucleic Acids Research*, vol. 34, no. 27, , Sep. 2006, pp.4655-4666.

[19] B. -T. Zhang, "Accelerated learning by active example selection", *International Journal of Neural Systems*, vol. 5, no.1, 1994, pp.67-75.

[20] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus", In Proc. of the *Symposium on Computer Applications and Medical Care*, 1988, pp. 261-265.

[21] O. L. Mangasarian, W. N. Street and W. H. Wolberg. "Breast cancer diagnosis and prognosis via linear programming", *Operations Research*, vol. 43, no. 4, July-August 1995, pp. 570-577.

[22] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. E. Costello, I. M. Moroz, "Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection", *BioMedical Engineering OnLine* vol. 6, no. 23, June 2007.

[23] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J.Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", *Proc. Natl. Acad. Sci*, USA, vol. 96 June 1999, pp. 6745-6750.

[24] B. H. Kim, S. B. Park, and B. -T. Zhang, "PromSearch: a hybrid approach to human core-promoter prediction", In Proc. of *the Int. Conf. on Intelligent Data Engineering and Automated Learning*, *Lecture Notes in Computer Science*, vol. 3177, Aug. 2004, pp.125-131.

[25] T. Fawcett, "ROC Graphs: Notes and Practical Considerations for Researchers", *Technical report*, Palo Alto, USA: HP Laboratories, March 2004.