Ensemble Learning Based on Active Example Selection for Solving Imbalanced Data Problem in Biomedical Data

Min Su Lee[a], Sangyoon Oh[b], Byoung-Tak Zhang[a]
[a] School of Computer Science and Engineering, Seoul National University, Seoul, Korea
b Div. of Information and Computer Engineering, Ajou University, Suwon, Korea
mslee@bi.snu.ac.kr, syoh@ajou.ac.kr, btzhang@bi.snu.ac.kr

Abstract

The imbalanced data problem is popular in biomedical classification tasks. Since trained classifiers using imbalanced data are mostly derived from the majority class, their prediction performance is poor for the minority class. In this paper, we propose a novel ensemble learning method based on an active example selection algorithm to resolve the imbalanced data problem. To compensate a possible sub-optimal classifier, our proposed ensemble learning methods aggregates classifiers built by the active example selection algorithm. We implement this ensemble learning method based on the active example selection algorithm using incremental naïve Bayes classifiers. Our empirical results show that we greatly improve the performance of classification models trained by five real world imbalanced biomedical data. The proposed ensemble learning methods outperforms other approaches by 0.03~0.15 in terms of AUC which solve imbalanced data problem.

1. Introduction

The imbalanced data problem has been popular since machine learning techniques have applied in real worlds of internet, industry, scientific and business research [1, 2]. When we train a classifier from data, we call the training data is imbalanced if there is much less examples in one or more classes than others. It happens when class examples are rare inherently or it is very hard to collect data (e.g. biomedical data such as rare disease and abnormal prognosis or data which is obtained from expensive experiments). The most of machine learning algorithms train a classifier under the assumption that the numbers of training examples between classes are almost same. Thus, when we apply machine learning algorithms to imbalanced data, trained classifiers are mostly derived from the majority class. Also, we may miss or ignore essential patterns from the minority class. In this case, the prediction performance of a minority class is almost meaningless since the training for minority class has not been done. However, users are frequently interested more in the minority. Therefore, solving imbalanced data problem is very important to improve classification performance for training minority class patterns.

In Ref [3], we proposed an Active Example Selection (AES) method. AES is the method to build a classifier by starting from a small balanced subset of training data and training a classifier iteratively through adding useful examples into the current training set. Even though AES performs well for improving the imbalanced classification performances, AES has some cons too. Its computational cost is high because its model training step and example selection step are iterated. Also, its output classifiers can be different from each other depending on the initial training examples.

In this paper, we propose a novel ensemble learning (EAES) which is an extension of the active example selection algorithm to resolve the imbalanced data problem. We address AES's high computational cost from the iterative model training and example selection with applying an incremental learning algorithm. For the proposed EAES, we use incremental naïve Bayes algorithm as a base classifier of AES instead of iterative batch one. As a result, we make the training time of AES shorter than time of iterative batch learning algorithm.

Additionally, we build an ensemble model by connecting various classifiers from different initial training datasets to reduce the variance of classification errors of AES and to get a robust output classifier. By integrating the different predictions from individual classifiers, the ensemble model can increase

978-0-7695-3885-3/09 $26.00 © 2009 IEEE
DOI 10.1109/BIBM.2009.44

classification performance along with avoiding biased decisions.

We organized this paper as follows. In section 2, we present related works. We present our proposed ensemble learning method based on AES in detail in section 3. In section 4, we show our empirical experiments and discuss about the results. We conclude in section 5.

2. Related works

Recent research on the imbalanced data problem has focused on several major groups of techniques. The popular method to solve imbalanced data problem is balancing the number of training examples among classes by re-sampling examples. To balancing the number of training examples among classes, random under sampling (RUS) randomly discards examples of majority class while random over sampling (ROS) duplicates examples in a minority class. We can combine these two techniques to apply oversampling for minority class and under sampling for majority class respectively.

These random re-sampling techniques are easy to apply and improve the performance of classifiers by compensate imbalanced class distribution. However, they also produce unwanted effects such as overfitting or information loss through duplicating or deleting examples from training sets by the techniques. To overcome these imbalanced data problems of random re-sampling, several new techniques are introduced using intelligent approach (e.g. creating new examples for minority class which is inferred from existing examples and removing noise or duplicated examples from majority class [4, 5]). However, according to recent studies which compare performances of various re-sampling techniques, rather simple RUS or ROS generally produce better performance than new intelligent techniques mentioned above [6, 7].

Biomedical domain is our main focus of application of this paper. Here are some recent important studies about handling imbalance biomedical data problem: One of frequently used methods is dividing the original dataset into a balanced dataset and an imbalanced dataset using one for training and one for testing respectively. We can avoid imbalanced data problem by using a balanced dataset for training. The method is used to diagnose myocardial perfusion using cardiac SPECT (Single Proton Emission Computed Tomography) images and to predict polyadenylation signals in human sequences [8, 9].

For the imbalanced biomedical data, RUS techniques also can be applied easily. To discriminate deleterious nsSNPs from neutral nsSNPs with imbalanced training dataset, prediction performances are improved by applying RUS method combined with a decision tree algorithm [10]. As well, classifiers from the RUS method can be combined together into an ensemble machine (ERUS). An ensemble of under-sampled classifiers is constructed for predicting the activity of drug molecules based on structural characteristics of compounds and for predicting glycosylation sites in genomic sequences [11, 12].

3. Ensemble learning based on active example selection

In this section, we present our proposed EAES. Before we describe our method which is the main topic of this paper, we describe the AES to solve imbalanced data problem and the incremental naïve Bayes classifier which is a base learner of EAES.

3.1 Active example selection

Our AES is an active learning method to solve imbalanced data problem. AES starts with small number of examples which is balanced among classes and trains a classifier by adding useful examples incrementally. After the unselected examples are used as validation examples for the current classifier, AES evaluates the classification result of the validation examples. Then, useful examples which can make up the current classifiers will be added without considering the example ratio among the classes. These steps including model evaluation, example selection, and model update using selected examples are iterated until we get the output classifier. In this training process, all the examples in the dataset are utilized, but output classifier is trained using selected examples [3].

To select the useful examples, AES evaluates misclassified examples from validation data and ranks them according to their error degree. Then, it adds examples to training dataset by the rank. Let let $\mathbf{x}=(x_1, \dots , x_m)$ be a validation example represented by a attribute value vector, $y \in C$ be a target class of $\mathbf{x}$, and $\boldsymbol{\theta}$ be a parameter vector of the current classifier. Then, error degree $\varepsilon_p(\mathbf{x})$ can be calculated using following formula (1)

$$\varepsilon_p(\mathbf{x}) = \begin{cases} 0, & \text{if } y = \arg\max_{c \in C} P(c \mid \mathbf{x}, \boldsymbol{\theta}) \\ 1 - P(y \mid \mathbf{x}, \boldsymbol{\theta}), & \text{otherwise} \end{cases} \quad (1)$$

When AES selects useful examples, it counts only error degree, not counts imbalanced ratio degree among data. AES will terminate the training until it used up the validation examples or there is no validation error.

While we apply the procedure of AES repeatedly, the classifier evolves efficiently using small subset of training dataset. Even though it is not explicitly considering the imbalanced degree of given dataset, AES resolves the imbalanced data problem through selecting procedure of useful examples. Detailed description is presented in Ref [3].

3.2. Base learner: incremental naïve Bayes classifier

The active example selection (AES) can be applied as a wrapper learner of classification algorithms which outputs predicted class with confidence value. In this paper, we use incremental naïve Bayes classifier as a base learner of AES.

Naïve Bayes classifier is a simple probabilistic classifier based on Bayes' theorem. In particular, it assumes that the predictive attributes are conditionally independent given the class, and it hypothesizes that no hidden or latent attributes influence the prediction process [13]. These assumptions make classification algorithm efficient. Let c be the random variable denoting the class of an example and let $\mathbf{x}$ be an observed example. Further, let c represent a particular class label and let x_i represent the i-th attribute of $\mathbf{x}$. It selects the class label c* with the maximum probability which is calculated according to the following equation,

$$c^* = \arg\max_{c \in C}(P(c)\prod_j P(x_j \mid c)) \qquad (2)$$

Despite its naïve design and over-simplified assumptions, naïve Bayes classifier shows good performances in many complex real-world problems. Moreover, naïve Bayes classifier requires a small amount of training data for parameter estimation. Since independent attributes are assumed, only the variances of the attributes for each class need to be determined and not the entire covariance matrix. All the probabilities required for solving equation (2) can be computed from the training data in one step. As a result, it leads low computational cost and relatively low memory consumption.

Another interesting aspect of the algorithm is that it is easy to implement in an incremental fashion because only counters are used. Naïve Bayes classifier builds a table for each attribute. The table reflects the distribution on the training data of the attribute-values over the classes. Incremental naïve Bayes classifier is initialized with zero training examples. Then it can learn incrementally using one example at a time by updating the tables. The trained incremental naïve Bayes classifier can be utilized by calculating the class membership probabilities for the given test example based on the tables.

The AES works well with iterative naïve Bayes classifier because it has small number of parameters to be tuned and spends short training time. In addition, incremental learning algorithms [14] are very suitable for incorporating with iterative procedure of AES.

3.3 An ensemble learning based on active example selection

AES resolves the imbalanced data problem nicely by iteratively selecting useful examples and update a current classifier. An output classifier is resulted from initial training examples which are just a small part of entire training data. However, since used examples cover a part of sample space, slight changes to the training data may easily lead to changes to the output model. Therefore, we propose ensemble learning based on active example selection (EAES) to improve classification performance and make it more robust. Ensemble learning methods combine multiple models and use them as a committee for decision making. Ensemble learning method increase memory and computational cost. Nevertheless, it increases prediction performance over a single model in many cases, because it reduces the variance of prediction errors and avoids biased decisions [15].

Our EAES builds an ensemble of component classifiers learned with different composition of training data which is derived from AES. Since each component classifiers are trained with different compositions of initial training examples, diverse subset of original training data is used for training each component classifier. Each of the component classifiers covers different part of sample space. Thus, resulting ensemble model can improve generalized prediction performance. For making final decision, our EAES use weighted voting policy: using the weighted sum of classification results with prediction probability distribution and weight derived from classifier's training performance.

EAES algorithm can be pseudo-coded in Figure 1 and overview of EAES is depicted in Figure 2. As depicted, EAES trains N component classifiers based on AES using randomly selected initial training data. In this study, we use incremental naïve Bayes classifier as a base learner of AES. After that, EAES binds N component classifiers from N different training subset to make a final classification result. EAES calculates a prediction result using weighted voting.

Let $\mathbf{x}$ be a test example, and $\boldsymbol{\theta}_i$ (i=1,...,N) be a parameter vector of the i-th component classifier from AES. To get the target class of test example $\mathbf{x}$, we calculate it as follows:

$$f(\mathbf{x}) = \arg\max_{c \in C} \sum_{i=1}^{N} \alpha_i P_i(c \mid \mathbf{x}, \boldsymbol{\theta}_i) \qquad (3)$$

where, α_i is calculated based on training error rate ε_i of the i-th component classifier of the form:

$$\alpha_i = \frac{\exp(-\varepsilon_i)}{\sum_{i=1}^{N} \exp(-\varepsilon_i)} \qquad (4)$$

Our EAES not only address the imbalanced data problem using AES, but also achieve more competitive performance by combining AES with ensemble learning method.

```
Model generation
  For each of $N$ iterations:
    Perform AES
      Initialize a component classifier using a predefined classification algorithm
      Randomly select initial training examples per class.
      Until meet termination conditions
        Update classifier using selected examples
        Validate current classifier
        Select useful examples
Classification
  For each of the $N$ models:
    Assign weight to each model using its training error
    Predict a class label of an example using weighted voting
  Return the class label with highest voting score.
```

Figure 1. Pseudo-code for EAES

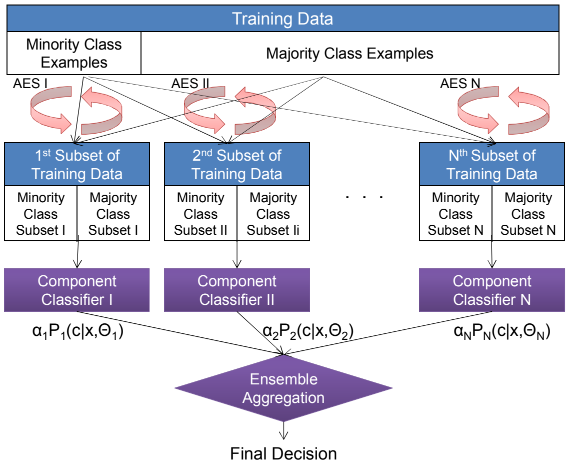

Figure 2. Ensemble learning based on active example selection

4. Experiments and evaluations

In this section, we present empirical results which show the performance of EAES. Since we already presented the architecture, characteristics, and performance on imbalanced data of AES in Ref [3], we focus on analyzing the classification performance of EAES with imbalanced biomedical data in this section.

4.1. Experimental datasets

We perform empirical experiments using five real-world biomedical benchmark datasets: hepatitis clinical data (Hepatitis), voice data of Parkinson's disease (Parkinson), diabetes clinical data (Diabetes), image data of prognostic breast cancer (WPBC), and image data of cardiac disease (SPECT) from UCI machine learning repository [16]. We consider binary classification problems in this study and the overview of the datasets is given in Table 1. An interesting target class is called the positive class and a normal class is called the negative class. As a data preprocessing step continues, a range of numeric attributes in the dataset is discretized into nominal attributes for naïve Bayes classifier.

Table 1. Overview of datasets

Dataset	# of Examples	# of Features	Class Distribution	Imb. Ratio
Hepatitis	158	19	Positive (terminal): 32 Negative (survival):123	1:3.84
Parkinson	194	22	Negative (healthy): 47 Positive (disease): 147	1:3.13
Diabetes	768	8	Positive (diabetes): 268 Negative (healthy): 500	1:1.87
WPBC	198	33	Positive (recur): 47 Negative (non-recur): 151	1:3.12
SPECT	267	43	Negative (normal): 55 Positive (abnormal): 212	1:3.85

4.2. Experiments

To investigate proposed AES and EAES performances, we conducted experiments to compare 1) naïve Bayes classifier algorithm (NB), 2) RUS, 3) AES, 4) ERUS and 5) EAES. Since AES produces a subset of training set, we compare the performance of AES with that of RUS which produces a randomly selected subset of training set. We choose RUS because it shows generally good performance than new intelligent approaches [6]. Since EAES is an ensemble method based on AES, we compare the performance of EAES with that of ERUS which is an ensemble method based on RUS. ERUS is chosen because it was used to solve many biomedical imbalanced data problem [11, 12]. RUS and ERUS are incorporated with naïve Bayes classifier algorithm. To make training time shorter, AES and EAES are incorporated with incremental version of naïve Bayes classifier.

For all data, the parameters for AES procedures are set as follows: the number of initial training example per class is 1 and the incremental example size is 2. In ensemble learning (i.e. ERUS and EAES), the number of component classifiers is set to 15.

To evaluate the performance of classification methods, AUC (Area Under the ROC Curve), overall accuracy, and true positive rates are calculated. When dataset is highly skewed and the overall accuracy tends to be overwhelmed by the prediction power for the majority class, the performance comparison of overall accuracy is very much misleading. For this reason, we used the AUC which give balanced evaluation by incorporating measures of both positive and negative classes with equal weights. In the imbalanced data problem, the AUC have been widely used as a performance evaluation measure. We also use the true positive rate (TPR) as an evaluation measures which represent the classification performances per class. The true positive rates are computed by the ratio of correctly predicted examples of a class among all available examples of the class during the test.

To estimate general performances of AES and EAES, for each combination of 5 datasets and 5 learning strategies, 10-fold cross validation were executed. The performances of total runs for each combination are averaged with standard deviation. The results are shown from Table 2 to Table 4.

4.3. Results and discussion

From the experiment results, we find some interesting issues to be discussed. First, we argue that our EAES and AES settles imbalanced data problem and achieve superior classification performance against RUS and ERUS. The improvement in AUC by 0.04~0.15 implies that our EAES effectively deals with the imbalanced data problem (Table 2). The AUC of EAES is higher than that of AES. It indicates that proposed EAES reduces the possibility of distorting the data distribution which is caused by training a model using a subset of total data. Also, the improvement in accuracy by 3.3~14.6% implies that our EAES upgrade general performance of output classifier by employing several classifiers as a decision committee (Table 3).

Second, our empirical study shows that real imbalanced data problem is not an imbalanced class distribution but an imbalanced prediction performance. In terms of true positive rates, the imbalanced class distribution does not always induce the imbalanced classification performance. As we can see in Table 4, in Parkinson and SPECT dataset, the true positive rates of majority class are lower than that of minority class.

Table 2. Comparison of AUC

Dataset	NB	RUS	AES	ERUS	EAES
Hepatitis	0.86±0.08	0.88±0.07	0.92±0.03	0.89±0.07	**0.94**±0.03
Parkinson	0.85±0.12	0.86±0.11	0.91±0.07	0.86±0.11	**0.92**±0.09
Diabetes	0.81±0.04	0.82±0.04	**0.84**±0.04	0.82±0.03	**0.84**±0.03
WPBC	0.69±0.13	0.66±0.15	0.79±0.07	0.67±0.10	**0.84**±0.14
SPECT	0.86±0.08	0.85±0.07	**0.91**±0.05	0.86±0.08	0.90±0.04

Table 3. Comparison of accuracy (%)

Dataset	NB	RUS	AES	ERUS	EAES
Hepatitis	83.7±9.2	81.8±8.7	**93.5**±8.1	81.1±9.2	92.8±9.0
Parkinson	68.7±10.0	68.7±10.4	76.8±9.2	69.2±10.1	**81.4**±9.1
Diabetes	75.1±4.0	74.5±2.3	77.9±4.3	74.3±3.1	**78.4**±4.3
WPBC	67.2±7.9	61.7±11.9	73.7±8.7	61.6±7.6	**81.8**±7.8
SPECT	68.9±5.9	66.0±6.4	77.2±5.6	66.3±4.7	**79.8**±6.6

Table 4. Comparison of true positive rate per class (%)

Dataset	Class	NB	RUS	AES	ERUS	EAES
Hepatitis	Pos	67.5±27.3	77.5±27.2	**87.5**±16.3	77.5±22.2	**87.5**±16.3
	Neg	87.7±10.6	82.9±9.6	**95.1**±8.8	82.1±10.2	94.3±11.0
Parkinson	Neg	86.5±19.2	89.0±19.1	**95.5**±9.6	91.0±19.1	93.5±10.6
	Pos	62.5±11.7	61.9±12.5	70.6±11.5	61.9±12.9	**77.4**±9.9
Diabetes	Pos	60.1±9.8	68.6±4.7	68.7±7.1	68.3±5.5	**70.2**±8.1
	Neg	**83.2**±4.2	77.6±2.5	82.8±3.9	77.6±4.2	82.8±3.7
WPBC	Pos	47.0±17.2	53.0±17.2	57.5±18.7	56.0±16.3	**62.0**±20.3
	Neg	73.4±9.6	64.1±16.5	78.8±8.1	63.5±7.5	**88.0**±8.2
SPECT	Neg	87.7±11.7	93.0±9.1	96.7±7.0	93.0±9.1	**100**±0
	Pos	64.2±8.3	59.0±8.6	72.2±7.0	59.5±5.7	**74.6**±8.1

The learning pattern of abnormal class is often harder than that of normal ones. More examples are needed to capture the patterns of the class, if the training examples of a class are derived from several groups. In those cases, AES selects more examples of complicated class regardless of balancing the training data distribution [3]. By adding useful examples into current classifier, AES effectively covers up weak points of the current existing classifier and improves prediction performance of each class. Moreover, we can achieve better prediction performance in almost every case by combining ensemble learning with AES.

Third, by adopting incremental learning algorithm (in this paper, we use incremental version of naïve Bayes classifier), AES selects new examples by active selection strategy and update the current model with the selected examples without training all the examples repeatedly. As a result, we make training time of AES shorter than time of iterative batch learning algorithm. Overall computational cost of EAES is strictly reduced by using incremental learning algorithm, because EAES includes several component classifiers which are trained based on iterative AES procedure. Due to space limitation, details of the results are omitted.

Finally, the performance of RUS and ERUS is not good comparing to our proposed methods. They sometimes show slightly improved AUC by at most 0.03 (Table 2) and no improvement in the classification accuracy (Table 3). In terms of true positive rates of RUS and ERUS, the classification performance on majority class is rather decreased by -5.2~-9.9% (Table 4). We presume that the performance degradation on majority class may be caused by the information loss from randomly discarding majority class examples.

5. Conclusion

Examples in the imbalanced data may exist redundantly or some of them are less useful. Our AES [3] solves the imbalanced data problem by iteratively collecting the useful training examples from entire training data as well as ignoring redundant or less-useful examples to select. By paying no attention to redundant or less-useful examples and learning a classifier using informative examples, AES can effectively make up the performance degradation caused by the imbalanced data problem. However, the composition of selected training examples may make variations of the resulting model.

In this paper, we propose an ensemble learning method based on AES called EAES to avoid biased decisions. In addition, we adopt incremental version of naïve Bayes classifier algorithm to speed up iterative AES procedure. Empirical results from five real-world biomedical datasets shows that our EAES and AES perform better than RUS and ERUS in dealing with the imbalanced data problem and improve prediction performance strictly.

We expect that our EAES and AES can be applied to other real world data mining applications where we suffer from the imbalanced data problem. Also, our EAES can be used to identify discriminative or representative examples of some classes via investigating selected training examples which are commonly appeared among various AES runs.

Acknowledgement

This work was supported in part by the MEST under the BK21-IT Program, by the IITA through the MARS project (IITA-2009-A1100-0901-1639), by the KRF Grant funded by the Korean Government (MOEHRD) (KRF-2008-314-D00377) and by the MKE under the ITRC support program supervised by the NIPA (NIPA-2009-C1090-0902-0003). The ICT at Seoul National University provided research facilities.

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, pp.1-6, June 2004.

[2] G. M. Weiss, “Mining with rarity: a unifying framework”, *SIGKDD Explorations*, vol.6, no. 1, pp.7-19, June 2004.

[3] M. S. Lee, J.-K. Rhee, B.-H. Kim, and B.-T. Zhang, “AESNB: active example selection with naïve Bayes classifier for learning from imbalanced biomedical data”, *In Proc. of the IEEE International Conference on Bioinformatics and Bioengineering*, pp.15-21, 2009.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique”, *Journal of Artificial Intelligence Research*, 16:321-357, 2002

[5] 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*, pp.179-186, 1997.

[6] J. V. Hulse, T. M. Khoshgoftaar, and A. Napolitano, “Experimental perspectives on learning from imbalanced data”, *In Proc. of the 24th Int. Conf. on Machine Learning*, pp.935-942, 2007.

[7] 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, pp.20-29, June 2004.

[8] 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.

[9] H. Liu, H. Han, J. Li, and L. Wong, “An in-silico method for prediction of polyadenylation signals in human sequences”, *In Proc. 14th Int. Conf. on Genome Informatics*, vol. 14, Dec. 2003, pp.84-93.

[10] R. J. Dobson, P. B. Munroe, M. J. Caulfield and M. AS Saqi, “Predicting Deleterious nsSNPs: an analysis of sequence and structural attributes”, *BMC Bioinformatics*, 7:217-225, 2006.

[11] 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.

[12] 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,.

[13] W. L. Buntine. “Operations for learning with graphical models”, *Journal of Artificial Intelligence Research*, 2:159-225, 1994.

[14] C. Giraud-Carrier, “A note on the utility of incremental learning”, *AI Communications*, 13 (4). ISSN 0921-7126, pp. 215–223. December 2000.

[15] E. Bauer, and R. Kohavi, “An empirical comparison of voting classification 37 algorithms: bagging, boosting, and variants”, *Machine Learning*, 36(1-2), pp.105-139. 1999.

[16] A. Asuncion, & D. J. Newman, (2007). UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html].