Cost-Sensitive Classification on Class-balanced Ensembles for Imbalanced Non-coding RNA Data

Bashier Elkarami
Department of Electrical and
Computer Engineering
University of Windsor
401 Sunset Avenue, Windsor, Ontario, Canada
elkaram@uwindsor.ca

Abed Alkhateeb, Luis Rueda
School of Computer Science
University of Windsor

401 Sunset Avenue, Windsor, Ontario, Canada
{alkhate, lrueda}@uwindsor.ca

***Abstract*—Many bioinformatics data sets have class-imbalanced data, where the number of samples in each class is not equal. Since most of data sets contain usual versus unusual cases, e.g. cancer versus normal or miRNAs versus other non-coding RNA, where the minority class with the least number of samples is the interesting class that contains the unusual cases. The learning models based on the standard classifiers, such as the support vector machine (SVM), random forest and k-NN are usually biased towards the majority class, which means that the classifier is most likely to predict the samples from the interesting class inaccurately. Thus, handling class-imbalanced data set has gained the researchers interests recently.**

A combination of proper feature selection, a cost-sensitive classifier and ensembling based on random forest method (BCE-CSC-RF) is proposed to handle the class-imbalanced data. Random class-balanced ensembles are built individually. Then, each ensemble is used as a training pool to classify the rest of out-bagged samples. Samples in each ensemble will be classified using class-sensitive classifier that incorporates random forest. The sample will be classified by selecting the most often class has been voted-for in all samples appearances in all the formed ensembles. A set of performance measurements including a geometric measurement suggests that the model can improve the classification of the minority class samples.

***Keywords*-Transcriptomatics, class-imbalanced data, non-coding RNA, mi-RNA, ensembles, cost-sensitive classifier, feature selections**

I. INTRODUCTION

A non-coding RNA (ncRNA) is an RNA molecule that is not translated into a protein. MicroRNAs (miRNAs) are small, highly conserved non-coding RNA molecules that regulates gene expression; miRNAs function via base-pairing with complementary sequences within mRNA molecules [1]. As a result, miRNAs may be involved in development of cancer or other diseases [2]. Since miRNAs are the minority group of ncRNAs, most ncRNAs data set are class-imbalanced, where miRNAs class is the minority class, this introduces the necessity of dealing with the class-imbalanced data set. The hairpin structure is a necessary feature for the computational classification of novel precursor miRNAs (pre-miRNAs). Pseudo hairpins are abundant genomic inverted repeats that can be potentially filtered to only discover novel species-specific pre-miRNAs [3], which yeilds the importance of the pre-miRNAs class over the pseudo hairpins class.

978-1-5090-0935-0/16/$31.00 ©2016 IEEE

Resampling techniques for handling imbalanced data have been proposed, including oversampling [4], undersampling [5], [6], and the combinations of these two method [7]. Oversampling can be approached by replicating samples randomly, or generating synthetic samples from the original ones. The first one does not add any new information which may lead to overfitting, while the latter generates unreal data that have no biological meaning. Undersampling techniques reduce the number of the majority class members in the training set [5], which may lead to bias in learning or loss of information in the majority class.

Feature selection is an important approach used in data mining applications. Feature selection can enhance the accuracy of the classifier by eliminating noisy features or unimportant features. It also helps in coping with the probability of over-fitting issues by reducing the data set dimensionally, which improves the overall efficiency and speed of the classifier. It can be divided into two main categories based on the underlying method of selection. Filter feature selection techniques [8], and wrappers techniques [9]. The first utilizes statistical measurements to build a decision based on the levels of relevance among features without taking the classifier into consideration. On the contrary, wrappers utilize classifiers to measure the scoring between the features. Wrapper techniques are usually greedy methods used to obtain a subset of features that interact well with each other. The classifier performs very well on the obtained data set. However, However the resulting subset may not be the optimal combination of features and the method is very costly.

Shakiba and Rueda proposed a method that transforms the non-coding RNA data set into a lower dimension one using linear dimensionality reduction (LDR) with explicit feature mapping, namely miLDR-EM [10] for pre-microRNA prediction from other sequences. Initially, all features are sent to a higher dimensional space explicitly after applying an exhaustive feature selection method that selects the optimum subset of k features in the higher dimensional

space. Then the LDR is applied on the chosen subset of features for classification. The implemented method is time consuming because of explicit mapping of the data to the higher dimensional space. In addition to an exhaustive feature selection strategy to find optimum subset of k features, which cannot guarantee the real optimum subset of features in the data set, that makes it works well for small value of k but not really effective for higher alues of k. The other problem of LDR is that usually LDR classifiers are affected by singular matrices. The authors stated that the problem does not exist for this particular data set; however, it may appear for other data sets. In the proposed model, correlation-based Feature Selection (CFS) is used to filter out noisy or redundant features and select the most informative observations. CFS uses a correlation based heuristic search strategy to evaluate the value of feature subsets when applied. CFS utilizes Pearsons correlation, where all feature vectors have been standardized to measure the correlation among the observations and the correlation between the selected subset of features and the classification.

The risk of misclassifying the interesting or high risk class is higher than misclassifying the uninteresting normal state class. The main goal of most classifiers is to reduce the error rate or misclassification based on the assumption that all misclassifications costs are equal; however, this assumption is not completely correct. In medical research, A patient is classified as a negative class (healthy), while in fact is positive class (with cancer), this misclassification could cause death to the patient. On other hand, if the patient is classified as positive while the patient is negative, this misclassification will only cost the patient more medical tests. Moreover, the assumption of equal cost leads classifiers to be biased for the majority class in case of imbalance data because the minority class has a very small effect on the accuracy. These issues can be overcome by introducing different costs to different misclassification error by applying Cost- Sensitive learning [11].

Random forest is a meta classifier that utilizes a collection of decision tree classifiers. Each tree is built on randomly selected sub-samples of the data set, and each tree implements a majority voting technique in making classification decisions [12]. The Gini Index is used to select the variable on which to split the data at each node of the tree based on the relation between the variable and the classes. The splitting goes on until end nodes (leaves) of the tree are only one class. Random forest has some advantages such as its ability on running on large data sets efficiently, its ability of estimating missing data and processing unbalanced data, and its control over-fitting. In our study, we use random forest in conjunction with a cost- sensitive technique.

II. Materials and methods

A. Materials

The data set contains the two classes, pre-microRNAs from both pseudo hairpins and other non-coding RNAs. This data set is available as a supplementary material in [13]. More information about the two classes is given below.

1) The Minority Class: Known human pre-microRNAs: This data set includes 691 non-redundant human pre-microRNA sequences that fold into hairpin structures. Some of these sequences fold into multibranched loops at default parameters.

2) The Majority Class: Pseudo hairpins: The negative data set consists of 8,494 human pseudo hairpin sequences. These sequences were obtained originally from RefSeq genes [14].

B. Methods

After CFS feature selection is used to select a subset of k features, random class-balanced ensembles are created to be classified using a cost-sensitive classifier that uses random forest as an inner classifier, Figure 1 shows the schematic representation of the method.

1) Feature Selection: Feature selection is a preprocessing step in machine learning that is used to choose the most representative features of the data set. CFS is a method to select the most correlated k features with the class label vector. The basic assumption behind CFS is that good features are less correlated with each other and highly correlated with the classification. This relation between the average value of all feature-classification correlations $\overline{r_{zf}}$ and the average value of all feature-feature correlations $\overline{r_{ff}}$ can be formulated as follows:

$$r_{zc} = \frac{k\overline{r_{zf}}}{k + k\,(k-1)\,\overline{r_{ff}}}, \tag{1}$$

where r_{zc} is the correlation between the summed features and the classification, k is the number of features as criterion. CFS removes redundant features by not selecting them since they highly correlate with the selected subset of features.

2) Ensembling: The model creates N ensembles, each one has two bags of data, the first for training and the other for testing. The training data set is created by randomly selecting $2x$ samples, x from each of the majority and minority class. The remaining samples from each class are used as out of bag testing; x is determined by 80 percent of the samples in the minority class, while N is calculated as D/x, where D is the total number of samples in the whole data set. Figure 1 shows the bootstrapping of the ensembles; the created ensembles are the input to the classification process.

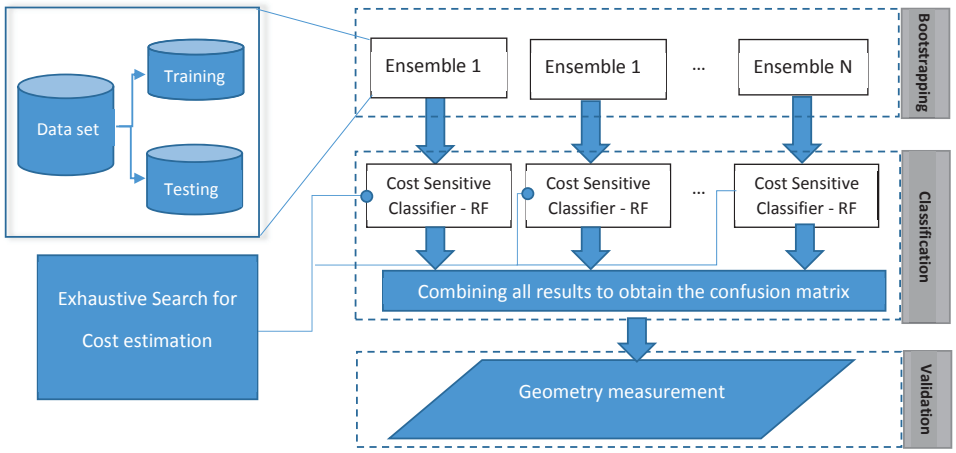

Fig. 1: Schematic representation of the proposed method

3) Cost-Sensitive Classifier: The goal of this type of learning is to minimize the risk by adding a factor or weight to the higher risk class to be misclassified. Equation 2 shows how the factor will be added to each class.

$$R\,(i\backslash x) = \Sigma P(i|x)C(i|j), \tag{2}$$

Which means that minimizing the cost of predicting the sample x to class i while it belongs to j, where C is the Cost matrix. If we have two classes, the cost matrix will be as in Table I, which also shows the estimated cost for the confusion matrix elements (TP, TN, FP, FN). The costs have been found using an exhausting search method for the best cost for each ensemble.C(0,1) represents the cost of classifying a positive sample as negative (pseudoHairpin as pre-miRNA), while C(1,0) represents the cost of classifying a negative sample as positive (pre-miRNA as pseudoHairpin).

	Actual positive	negative
Predicted positive	C(0,0) = TP	C(0,1) = FP
Predicted negative	C(1,0)= FN	C(1,1) = TN

TABLE I: The assigned cost matrix for the confusion matrix

4) Random Forest Classifier: We choose random forest to classify each randomly created class-balanced ensemble due to its strength with random classification and the ability to average the overall of classification performance. 150 random trees have been generated on each ensemble to vote for the out of bag test samples. Random forest trees are planted using Gini ratio as a ranker. T is the number of trees which is 150 in our experiment. M is the randomly chosen variable each time which approximately equals half the number of features for each subset.

5) Geometric Measurement Score: As mentioned earlier, the class imbalance problem occurs when there are different numbers of samples in the positive and negative classes, which leads to poor performance of the classifier with respect to the minority class [15], the number of positive samples in comparison to the negative samples is small, with a ratio of about 1:13. In this case, standard classifiers such as SVMs have tendency to classify well the largest class, while ignoring the smallest class.

In classifiers in which the numbers of samples in different classes are imbalanced, the performance of the classifier cannot be assessed accurately based on the percentage of test samples that are correctly classified. This is because, for example, when the number of samples in the negative class heavily outnumber the samples from the positive class and the classifier always classifies samples as negative, the accuracy is high, although the classifier is useless. Hence, other indicators are required for analysis of the classification performance. In [15], it is suggested that the geometric mean Gm, can be a good indicator:

$$Gm = \sqrt{SE \times SP} \tag{3}$$

Where SE is the sensitivity and SP is the specificity. The Geometric measurement reflects the way in which the classifier handles the minority class.

III. Results and Discussions

The number of randomly selected samples for each class x was set to 400, N the number of ensembles, was set to 15 ensembles to ensure coverage of all samples. The number of planted trees in random forest was set to 150 and the cost matrix for cost-sensitive classifier was set to:

$$C\begin{Bmatrix} 0 & 7 \\ 9 & 0 \end{Bmatrix}$$

For different k features that are selected by CFS, the method created the class-balanced ensembles from only the selected feature vectors for all data. The specificity, sensitivity, and Gm for each k, suggests that the best result is $k = 7$ selected features. The results of the $7-$features model as in Figure2 shows different values for specificity, sensitivity and Gm for

TABLE II: performance comparisons between miLDR-EM and BCE-CSC-RF.

	miLDR-EM			BCE-CSC-RF		
No. of Features	SE	SP	Gm	SE	SP	Gm
3	85.53%	97.51%	91.32%	90.13%	91.42%	90.80%
5	84.23%	96.16%	90.00%	89.56%	92.18%	90.85%
7	86.54%	96.91%	91.58%	90.85%	92.62%	91.72%
10	87.12%	93.18%	90.10%	90.60%	92.70%	91.64%

the different created ensembles due to the randomness property of the model. The randomness of the model comes from selecting the samples randomly for each ensemble, in addition to the randomness nature of the random forest. The combined results for the overall model shows that the sensitivity equals 92.62, while specificity equals 90.85, In comparison to miLDR-EM results, the best miLDR-EM result is also for the $7-$features model, where the sensitivity equals 86.54 and specificity equals 96.91, which indicates that miLDR-EM performs less efficient on the positive class ẗhe minority classs̈ince the sensitivity measures the positive rate. The Gm for the proposed method equals 91.72% while it is 91.58% for the miLDR-EM method.

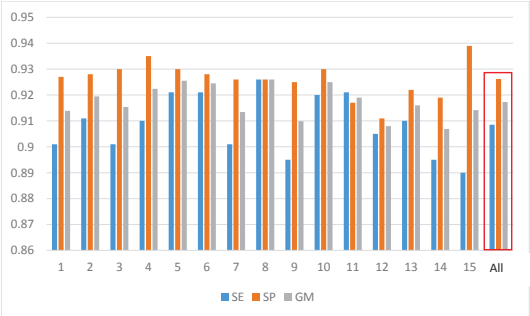

Fig. 2: Measurements for the 7-features model, $x-$axis represents number of ensemble, $y-$axis represents the percentage value for SE,SP and Gm measurements

As stated earlier, the performance of miLDR-EM in the smaller number of selected features models, $3-$features model for example, is slightly better than the proposed BCE-CSC-RF. However, the performance of BCE-CSC-RF is better than miLDR-EM in the higher number of selected features models. for example for the $10-$features model, Gm measurement is 90.10% for miLDR-EM where it is up to 91.64% for BCE-CSC-RF. But the real improvement in the assumption of the risk, where we assume that the true positive rate is much important than false negative rate. In other words, the misclassification of the minority class is higher risk than the misclassification of the majority class. The results in Table II suggest that BCE-CSC-RF highly outperformed miLDR-EM in SE measurement, which is more important since it represent the high biological informative class human pre-microRNAs than the Pseudo hairpins class.

IV. Conclusion and Future Works

Based on the fact of the higher risk of the misclassification of the pre-miRNAs samples as a pseudo hairpins.BCE-CSC-RF is a cost-sensitive classifier that is proposed based on randomly created class-balanced ensembles, The method handles the class-imbalanced data where the minority class is much important than the majority class which the standard machine learning classifiers usually fail. CFS is used as a preprocessing to filter out the noise features and select the most informative features. The results suggest that BCE-CSC-RF has better geometric performance than miLDR-EM and improves the true positive rate than the true negative rate.

Many different feature selection methods can be tested to improve the selected subset of features. collaboration with biologists to evaluate the selected subset of features is highly recomended to gain more biological information from the model. The model also can be applied on different class-imbalanced data where the minority class is higher importance than the majority class, for example, cancer versus normal data set. grid search to find better model's setting can help to improve the overall performance.

References

[1] D. P. Bartel, "Micrornas: target recognition and regulatory functions," *Cell*, vol. 136, no. 2, pp. 215–233, 2009.

[2] Y. S. Lee and A. Dutta, "The tumor suppressor microrna let-7 represses the hmga2 oncogene," *Genes & development*, vol. 21, no. 9, pp. 1025–1030, 2007.

[3] K. L. S. Ng and S. K. Mishra, "De novo svm classification of precursor microrna from genomic pseudo hairpins using global and intrinsic folding measures," *Bioinformatics*, vol. 23, no. 11, pp. 1321–1330, 2007.

[4] H. Guo and H. L. Viktor, "Learning from imbalanced data sets with boosting and data generation: the databoost-im approach," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 30–39, 2004.

[5] M. Kubat, S. Matwin *et al.*, "Addressing the curse of imbalanced training sets: one-sided selection," in *ICML*, vol. 97. Nashville, USA, 1997, pp. 179–186.

[6] D. R. Wilson and T. R. Martinez, "Reduction techniques for instance-based learning algorithms," *Machine learning*, vol. 38, no. 3, pp. 257–286, 2000.

[7] X.-M. Zhao, X. Li, L. Chen, and K. Aihara, "Protein classification with imbalanced data," *Proteins: Structure, function, and bioinformatics*, vol. 70, no. 4, pp. 1125–1132, 2008.

[8] H. Liu, R. Setiono *et al.*, "A probabilistic approach to feature selection-a filter solution," in *ICML*, vol. 96. Citeseer, 1996, pp. 319–327.

[9] R. Kohavi and D. Sommerfield, "Feature subset selection using the wrapper method: Overfitting and dynamic search space topology." in *KDD*, 1995, pp. 192–197.

[10] N. Shakiba and L. Rueda, "Microrna identification using linear dimensionality reduction with explicit feature mapping," in *BMC proceedings*, vol. 7, no. Suppl 7. BioMed Central Ltd, 2013, p. S8.

[11] T. Wang, Z. Qin, S. Zhang, and C. Zhang, "Cost-sensitive classification with inadequate labeled data," *Information Systems*, vol. 37, no. 5, pp. 508–516, 2012.

[12] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[13] R. Batuwita and V. Palade, "micropred: effective classification of pre-mirnas for human mirna gene prediction," *Bioinformatics*, vol. 25, no. 8, pp. 989–995, 2009.

[14] K. D. Pruitt and D. R. Maglott, "Refseq and locuslink: Ncbi gene-centered resources," *Nucleic acids research*, vol. 29, no. 1, pp. 137–140, 2001.

[15] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *Machine Learning: ECML 2004*. Springer, 2004, pp. 39–50.