BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis
Volume 11(1)

Handling class imbalance problem in miRNA dataset associated with cancer

Ram Kothandan

Department of Biological Sciences, BITS PILANI K K Birla Goa Campus, Zuarinagar, Vasco Da Gama, India; Ram Kothandan – Email: mailram1986@gmail.com

Received January 09, 2015; Accepted January 23, 2015; Published January 30, 2015

Abstract:
MiRNAs are small (~22nt long) non-coding RNA sequences; binds to the complementarity target sites in 3' Untranslated Region (UTR) of mRNA sequences but not restricted to other mRNA regions *viz.,* 5' UTR and Coding sequences (CDS). Complementarity binding of miRNA to mRNA target sites either results in complete degradation of the mRNA itself or it may regulate the mRNA as an oncogene or as a tumor suppressor gene. However, the exact mechanism involved in identifying a miRNA to be associated with cancer is still unclear. Further, with the outburst in the number of miRNAs sequences recorded every year in miRBase, the gap is still widening mainly due to the laborious and economically unfavorable experimental procedures associated with the functional annotation. Motivated by the fact, we constructed a two-step support vector machine-based predictive model - miRSEQ and miRINT. However, the major pitfall during the construction of the model is the class imbalance problem. Hence, in order to overcome class imbalance problem, in the present study we empirically compare the effectiveness of two different methods *viz.,* Synthetic Minority Oversampling Technique (SMOTE) and cost-senstive learning method. Performance measures were evaluated in terms of Precision and Recall. Based on our result, it was observed that for miRNA dataset with high class imbalance utilized for predicting association of cancer, cost-sensitive method outperformed the oversampling method.

Keywords: Cost-sensitive, SMOTE, miRNA-mRNA interaction, Support Vector Machines.

Background

A dataset is imbalanced if the classification categories are not equally represented **[1]**. Class imbalance or skewed dataset mainly arises when most of the instances are labeled as one class (majority class), while very few are labeled as the other class (minority class). Traditional classifiers utilizing the entire training set for prediction are not suitable to deal with imbalanced dataset because they show bias towards the majority class due to over-prevalence. Particularly in case of disease related dataset (like ours) - miRNA dataset associated with cancer, the number of experimentally validated miRNAs are much higher than the number of miRNAs not associated with cancer. The main problem in training a classifier with high imbalanced dataset is that the minority class is often considered as noisy dataset and hence overlooked by the majority class.

Performance of the classifier constructed with a certain level of class imbalance is always unpredictable or deteriorating in many cases. Hence, to overcome the problem of class imbalance, machine learning algorithms generally utilize two methods *viz.,* resampling at the data level *i.e.* either oversampling the minority class e.g. Synthetic Minority Oversampling Method (SMOTE) **[2]** or under sampling the majority class e.g. Easy Ensemble and Balancing Cascade method **[3]**. Utilizing a resampling method is entirely a data driven process. On the other hand, class imbalance is ignored at the algorithm level by adjusting the cost of the classes to counter imbalance, adjusting the probabilistic estimates (in case of decision trees) and adjusting the decision threshold. In certain cases, both cost and resampling methods are used in combination, i.e. individual models are adjusted with these methods and combined as an ensemble to provide better performance **[4]**.

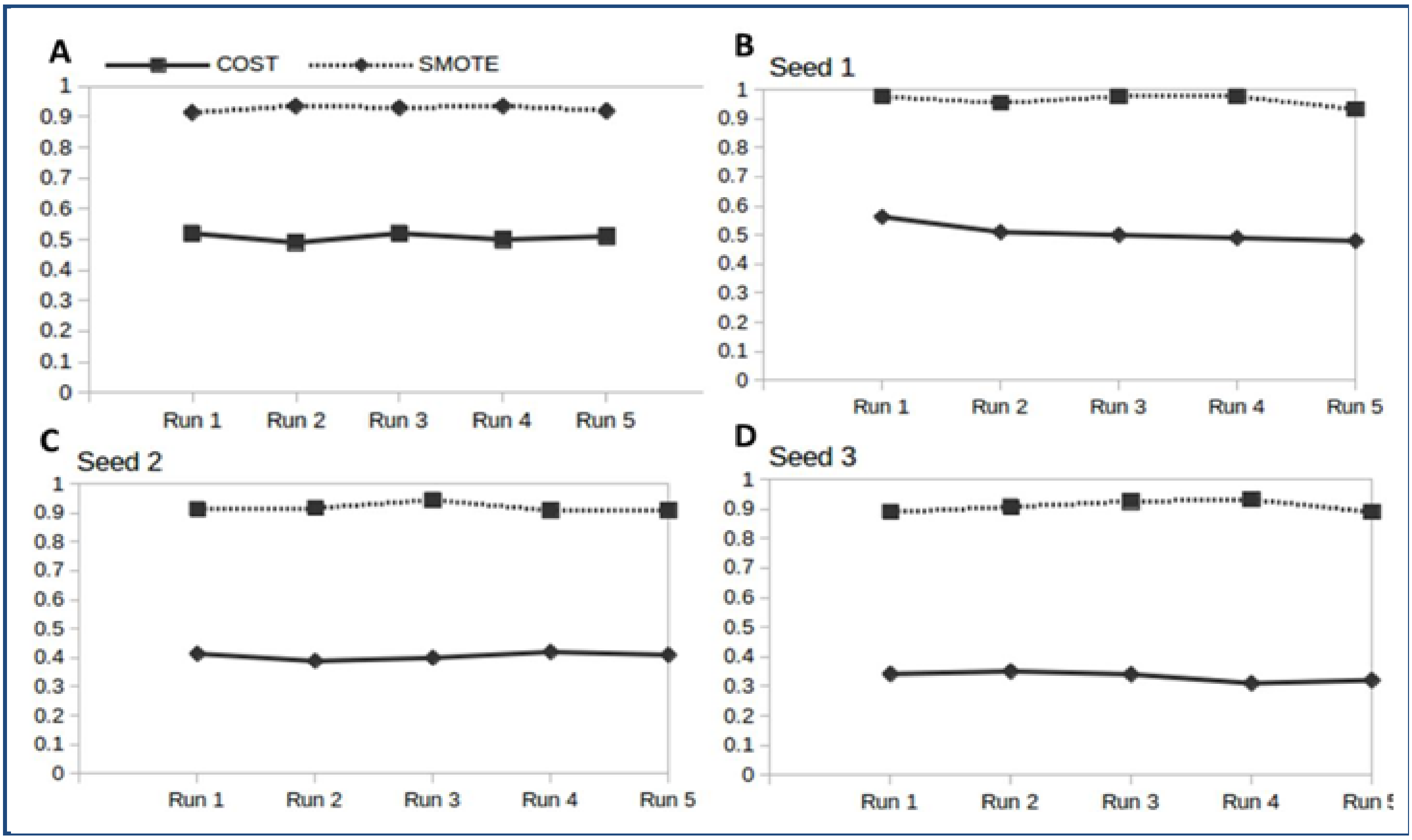

Figure 1: Comparison of SMOTE and cost-sensitive method to overcome class imbalance in the miRNA dataset associated with cancer: **A)** Comparison of SMOTE and cost-sensitive method with miRSEQ classifier trained with sequence based features only; **(B, C & D)** Comparison of SMOTE and cost-sensitive method with miRINT trained with miRNA-mRNA interaction based features. In both the classifier, SMOTE method tends to overfit the test dataset. SMOTE and cost-sensitive methods were tested with five independent test datasets (Run 1 to 5).

Generally in oversampling technique, class imbalance is overlooked by generating new instances with replacement from the minority class. But, generating similar instance at a specific region will overpopulate the minority class and results in bias during actual prediction **[5]**. Hence, in SMOTE, new synthetic samples are generated based on two parameters – the nearest neighbors (k) and the number of instance (n) required. In undersampling, multiple subset of majority class similar in size to the minority class is generated and trained. Since only a part of the dataset is utilized the computation cost and the time associated with this training is very less and efficient than the oversampling methods. However, undersampling methods ignore a large part of the training set making them vulnerable to miss many discriminative features present in them **[3].**

Most learning algorithms attempt to minimize the error rate in the classification by ignoring the difference between the types of misclassification errors. However, for real world problem this assumption wont hold true. Hence, to overcome the problem, cost-sensitive method is preferred generally over other class imbalance methods. Cost-sensitive method along with misclassification cost considers other cost like instance and attribute cost, active learning cost and computational cost. Among the cost, misclassification cost is more important in cost-sensitive learning and it can be either stationary (assigning a cost matrix) or dataset dependent. Thus, in the present study, we compare the effectiveness of two methods to overcome class imbalance in terms of precision and recall to construct an efficient classifier in predicting miRNAs associated with cancer.

Methodology:

Dataset Preparation

Dataset preparation was carried out for positive and negative set individually. For training purpose, 239 experimentally validated miRNAs obtained from our previous work would serve as positive dataset **[6]**. For negative dataset, precautionary steps were undertaken to avoid randomness in the dataset, i.e. randomly generated and predicted dataset were completely avoided. Only experimentally validated 32 miRNAs obtained from TargetMiner were considered as negative dataset **[7]**. For evaluating the effectiveness of the two methods compared in the study, we constructed an independent test dataset not utilized in training purpose. A 10-fold cross-validation method is used as a standard method for revalidation during training **[8].**

Feature Extraction

A list of 60 features were extracted from experimentally validated miRNA sequences, miRNA-mRNA interaction data and thermodynamics of miRNA-mRNA binding as obtained from RNAhybrid **[6, 9, 10]**. We utilized Pairfinder, a perl script to parse the various features from the miRNA-mRNA hybridized structure **[6].** In this present study, a two-step classifier (*viz.,* miRSEQ and miRINT) was constructed. MiRSEQ preliminarily predicts the miRNA associated with cancer based on 26 sequence-based features; whereas miRINT utilized 34 miRNA-mRNA interaction-based features to confirm the association of miRNA with cancer.

Learning Algorithm

The choice of learning algorithm plays a critical role in overcoming class imbalance. In this present study, we employed Support Vector Machines (SVM) with Radial Basis Function (RBF) as kernel function for training the miRNA dataset **[11]**. In a binary classifier, SVM classifies two classes by constructing a hyperplane in three dimensional space separated by margins. We utilized LIBSVM package in Waikato Environment for Knowledge Analysis (WEKA) **[12]**. Random search method was employed to identify optimum algorithm parameter cost (c) and gamma (λ) rather than computationally expensive grid based method**.**

Both SMOTE and cost-sensitive method packages available within the WEKA environment were utilized to handle the class imbalance during the training process. For SMOTE, we considered the nearest neighbor to be five (k=5) and the percentage of instances generated (n) in each iteration to be 100. The number of iterations was limited till there is a shift in the class distribution. In a typical class imbalanced problem, cost-sensitive algorithms require a cost-matrix to represent costs for different misclassification types. The method tends to minimize the number of high cost error and then further generates a model with low misclassification cost. Misclassification cost can be assigned to both binary and multi-class classification problems. We constructed a 2x2 cost matrix for reweighing the data space. Cost for the correctly classified instances are assumed zero (*i.e.,* the cost associated with the True Positive (TP) and True Negative (TN) is zero) **[13]**. The main aim of utilizing cost-sensitive method is to construct a model with minimum misclassification cost and is given by the equation (1)

$$Cost = FNrate \times C(0,1) + FPrate \times C(1,0)$$

(Equation 1)

Where, C(0,1) and C(1,0) are the costs associated in prediction of False Positive (FP) and False Negative (FN) respectively.

Performance Evaluation

$$Precision = \frac{TP}{TP + FP} \qquad \text{(Equation 2)}$$

$$Recall = \frac{TP}{TP + FN} \qquad \text{(Equation 3)}$$

Results & Discussion:

The focus of the study is to obtain an efficient method for handling class imbalance in miRNA dataset associated with cancer. MiRSEQ and miRINT classifiers were constructed with both SMOTE and cost-sensitive method with SVM as the learning algorithm. Only experimentally validated miRNA were used for training purpose. Randomly generated, predicted miRNA sequences were neglected completely in order to avoid randomness in the dataset during the training process. Prior to training process dataset was normalized, since significant difference in the variance will dominate the RBF function and does not allow learning the dataset from other features. Utilizing mean value for missing attribute during the feature extraction was also avoided.

The performance of the constructed models were evaluated based on precision and recall. Usually in training machine learning algorithms, performance is evaluated using confusion matrix. However, for problems with high class imbalance, evaluating the performance of the classifier directly based on confusion matrix is not preferred. Alternatively, measures like precision and recall would reveal the actual predictive performance of the classifier. In disease related dataset, particularly miRNA dataset associated with cancer (like ours), precision would provide an exact measure of predictive performance of the constructed model since a single false prediction in disease related dataset would be catastrophic.

The predictive performance evaluated during the training process was marginally similar between the two methods being compared. However, when challenged with test dataset, cost-sensitive method performed better than the SMOTE. The underlying problem for poor predictive performance with SMOTE is due to overfitting (precision > 0.9 in all independent test runs are shown in **Table 1 See supplementary material).** One possible reason for overfitting with SMOTE is that the method centers more on the specific region in the feature space as the decision region for the minority class, than increasing the overall number of instances. Further, new instances are synthesized based on the number of the nearest neighbors chosen and also based on the number of new instances required per iteration. Thus SMOTE overpopulates a specific region rather than increasing the overall instances. Further, the classifier constructed with SMOTE method misclassified every instances as the minority class due to over-prevalence in the specific region during the independent test dataset prediction.

On the other hand, cost-sensitive method seamlessly performed better than SMOTE because it considers misclassification cost based on the dataset utilized in the training (precision 0.52 for miRSEQ and average precision 0.4 in all seed based models for miRINT) (**Table 1)**. From (**Figure 1),** it is evident that SMOTE method tends to overfit the dataset in both miRSEQ and miRINT classifier, whereas cost-sensitive showed significantly a steady performance in all test runs. Further, in order to boost the performance of classifier with SMOTE method, we reduced the number of instances generated per iteration. This will avoid over populating the minority class in a specific region. However, it was observed that there was no significant improvement in the performance measurement. For miRINT, the dataset was segregated based on the number of seed region formed in the hybridized structure. Similar to miRSEQ performance, the SMOTE method did not show much improvement in terms of precision, rather they tend to overfit (precision > 0.9) the dataset and thus left no room for further improvement.

Conclusion:

The work presented in this paper gives an empirical comparison of two methods to overcome class imbalance (*viz.,* SMOTE and cost-sensitive method) in prediction of miRNA associated with cancer. Among the two methods compared the SMOTE handles class imbalance at the data level and cost-sensitive method at the algorithm level. Handling class imbalance at the data level for disease related prediction (like ours) would induce several synthesized instances. Even though, oversampling method provide a good performance measure at the training step, when challenged with independent test datasets the performance of the classifier deteriorated completely. To further support the hypothesis, the prediction obtained from classifier constructed show overfitting of the test dataset.

On the other hand, cost-sensitve method provided a steady performance measure in each of the independent runs and thus acts as an effective method in handling class imbalance in miRNA dataset. The performance of cost-sensitive method can be further enhanced by utilizing appropriate feature selection method like Recursive Feature Elimination method (RFE) prior to the training process. Prioritizing most discriminative features would increase the performance of the classifier with cost-sensitive method. Further, utilizing different learning algorithm along with cost-sensitive method would boost the performance significantly and such a work is under progress in our group. Thus, we conclude that for prediction of miRNA associated with cancer with high class imbalance in dataset, cost-sensitive method performs better than the oversampling method.

Acknowledgment:
Ram Kothandan was supported by a scholarship from Council of Scientific Research and Industrial Research, Govt. of India.

References:

[1] Ding J *et al. BMC Bioinformatics* 2010 **14:** 11 [PMID: 21172046]
[2] Lertampaiporn S *et al. Nucleic Acids Res.* 2013 **41:** e21 [PMID: 23012261]
[3] Liu XY *et al. IEEE Trans Syst Man Cybern B Cybern*. 2009 **39:** 539 [PMID: 19095540]
[4] Yin HL & Leong TY, *Stud Health Technol Inform*. 2010 **160:** 856 [PMID: 20841807]
[5] Hao M *et al. Anal Chim Acta.* 2014 **806:** 117 [PMID: 24331047]
[6] Kothandan R & Biswas S, *Bioinformation* 2013 **9:** 524 [PMID: 23861569]
[7] Bandyopadhyay S & Mitra R, *Bioinformatics* 2009 **25:** 2625 [PMID: 19692556]
[8] Gamzon ER *et al. Plos One.* 2010 **5:** e13534 [PMID: 20975837]
[9] Batuwida R & Palade V, *Bioinformatics* 2009 **25:** 989 [PMID: 19233894]
[10] Sharma S & Biswas S, *Bioinformation* 2011 **6:** 364 [PMID: 21814397]
[11] Vapnik VN, *IEEE Trans Neural Netw*. 1999 **10:** 988 [PMID: 18252602].
[12] www.cs-waikato.ac.nz/ml/weka
[13] Zidelmal Z *et al. Comput Methods Programs Biomed.* 2013 **111**: 570 [PMID: 23849928]

Edited by P Kangueane
Citation: **Kothandan,** Bioinformation 11(1): 006-010 (2015)

Supplementary material:

Table 1: Comparison of SMOTE and cost-sensitive method in terms of Precision and Recall. Only average value of five independent runs are tabulated. For miRINT, miRNA-mRNA hybrid structures were segregated into seed 1, seed 2 and seed 3 models based on the number of seed region formed in their structures and trained individually.

MiRSEQ		**Precision**	**Recall**
Cost-sensitive		0.52	0.521
SMOTE		0.927	0.9345
MiRINT	**Number of Seeds**	**Precision**	**Recall**
Cost-Sensitive	**Seed 1**	0.562	0.426
	Seed 2	0.414	0.644
	Seed 3	0.341	0.584
SMOTE	**Seed 1**	0.9627	0.9042
	Seed 2	0.9181	0.931
	Seed 3	0.908	0.9141

* Models with Precision > 0.9 misclassified all instances as minority class in SMOTE