Computers in Biology and Medicine 68 (2016) 27–36

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Unsupervised learning assisted robust prediction of bioluminescent proteins

Abhigyan Nath *, Karthikeyan Subbiah *

Department of Computer Science, Banaras Hindu University, Varanasi 221005, India

ARTICLE INFO

Article history:
Received 13 April 2015
Accepted 28 October 2015

Keywords:
Class imbalance
Training set diversity
Optimal class distribution
K-Means
SMOTE

ABSTRACT

Bioluminescence plays an important role in nature, for example, it is used for intracellular chemical signalling in bacteria. It is also used as a useful reagent for various analytical research methods ranging from cellular imaging to gene expression analysis. However, identification and annotation of bioluminescent proteins is a difficult task as they share poor sequence similarities among them. In this paper, we present a novel approach for within-class and between-class balancing as well as diversifying of a training dataset by effectively combining unsupervised *K*-Means algorithm with Synthetic Minority Oversampling Technique (SMOTE) in order to achieve the true performance of the prediction model. Further, we experimented by varying different levels of balancing ratio of positive data to negative data in the training dataset in order to probe for an optimal class distribution which produces the best prediction accuracy. The appropriately balanced and diversified training set resulted in near complete learning with greater generalization on the blind test datasets. The obtained results strongly justify the fact that optimal class distribution with a high degree of diversity is an essential factor to achieve near perfect learning. Using random forest as the weak learners in boosting and training it on the optimally balanced and diversified training dataset, we achieved an overall accuracy of 95.3% on a tenfold cross validation test, and an accuracy of 91.7%, sensitivity of 89. 3% and specificity of 91.8% on a holdout test set. It is quite possible that the general framework discussed in the current work can be successfully applied to other biological datasets to deal with imbalance and incomplete learning problems effectively.

1. Introduction

There are mainly two phenomena, bioluminescence and bioflourescence which are responsible for the emission of visible light from the living organisms. The mechanisms of these two processes are distinct as the former involves a chemical reaction, and the latter involves absorption of light from external sources and its emission after transformation. Bioluminescence is observed in both terrestrial and marine habitats. The chemical reaction, which is responsible for bioluminescence, generates very less heat and can be categorized into oxygen dependent (luciferin-luciferase system) and oxygen independent types (ex. Photoproteins). The colour of the emission is governed by the amino acid sequence, and by accessory proteins like yellow fluorescent proteins (YFP) and green fluorescent proteins (GFP) [1]. Diverse systems for bioluminescence exist in nature, for ex. in Dinoflagellates, specialized organelles known as Scintillons [2,3] exhibit bioluminescence. Bioluminescence plays an important role in bacterial intracellular chemical signalling and in symbiosis: a common example of which is shown by *Epryme scolopes* and *Vibrio fishcri* [4,5], in attracting for a mate and repelling the predators. The independent evolution of bioluminescence in different organisms has been discussed in Hastings et al. [1]. In some organisms, the usefulness of bioluminescence is still unknown.

In silico prediction of bioluminescent proteins (BLP) was first carried out by Kandaswamy et al. [6]. They developed Blprot, which is an SVM based method. Their prediction model was trained by using 544 amino acid physicochemical properties. The prediction of bioluminescent proteins was further improved by Zhao et al. (BLPre) [7] using evolutionary information in the form of PSSM (Position Specific Scoring Matrices) obtained from PSI-BLAST. Fan et al. [8] used a balanced dataset (equal number of positive and negative samples for training) with average chemical shift and modified pseudo amino acid composition for prediction of bioluminescent proteins. Recently, Huang [9] proposed a scoring card method (SCBM) for their prediction.

The imbalanced class ratios are often encountered in the protein family classification problems. This causes the over-representation of instances belonging to majority class and under representation of instances belonging to minority class in the

* Corresponding authors. Tel.: +91 9956015187, +91 9473967721.
E-mail addresses: abhigyannath01@gmail.com (A. Nath), karthinikita@gmail.com (K. Subbiah).

http://dx.doi.org/10.1016/j.compbiomed.2015.10.013
0010-4825/

training set. The machine learning models trained with the imbalance training dataset have classification bias towards majority class and behave like a majority class classifier. This issue of imbalanced dataset has not been given the required attention in the bioinformatics community as it deserves.

In the current prediction problem, bioluminescent proteins (BLPs) are the positive minority class (which is the class of interest) and the majority class consists of all the non-bioluminescent proteins (NBLPs) belonging to different other protein families. The negative class is naturally very large as compared to the number of BLPs. So, the bioluminescent prediction problem training dataset is one of the classic examples of imbalanced dataset. This imbalance in class distribution greatly affects the accuracy in predicting the positive class instances (as the prediction models tends to act as a majority class classifier) and it is also quite evident from the previous studies [6–9].

When we use any machine learning algorithm to build a prediction model, the major motive is to maximize the generalization ability of the model. This insures that the trained predictive model will yield good prediction accuracy on the future unseen data. Ideally the training dataset that is presented to the learning algorithm should be properly diversified by covering the representatives from the entire input instance space to achieve the maximum possible generalization ability. If the training data are composed of a large number of very similar instances, it may get biased towards those instances. This notion holds true in the cases of both between-class (inter-class) and within-class (intra-class) instances. So the diversification of the training set is essential to gain enhanced generalization. Both between-class imbalance and within-class imbalance have a negative influence on the performance of machine learning algorithms [10].

In the present study, we have created a diversified and balanced training dataset by using unsupervised *K*-Means clustering algorithm (to deal with the within class imbalance where each class contains subgroups of similar instances of varying numbers) and then using SMOTE [11] (to selectively amplify the representative minority class sequences for balancing the between-class imbalance). The boosted random forest algorithm which has performed considerably better than the other machine learning algorithms was used to create our prediction model.

As the next part of this study, we have investigated the effect on prediction performances by varying the balancing ratio from ideal ratio (that is 1:1) to the original imbalance ratio. Analyzing the experimental results has revealed that the best prediction performance can be achieved at an optimal balancing ratio rather than at ideal balancing ratio. It was found that another performance factor (diversity) gets affected at the ideal balancing ratio of 1:1. This has motivated us to probe for optimal class distribution which is required to achieve superior accuracy (provides the best trade-off between inter-class balancing ratio and the diversity). The optimal class distribution is seldom explored in bioinformatics.

Finally individual features are ranked using the Relieff feature ranking algorithm and investigated the performance of the classifier by varying the number of features starting from 5 most discriminating features up to 40 (according to their rank) and recorded the calculated performance evaluation metrics obtained for RARF. The prediction performance increases with the increasing number of features (according to their ranks). This has authenticated the presence of large diversity among BLPs and there is a need for finding the optimal class distribution in order to achieve the best prediction performance. The superiority of the proposed framework as compared to random sampling is also discussed.

2. Materials and methods

2.1. Dataset

We used the dataset of Kandaswamy et al. [6] which consists of 441 positive class sequences (bioluminescent proteins) having less than 40% sequence identity and 18202 negative class sequences (non-bioluminescent protein NBLP) having more than 40% sequence identity. The redundant sequences in the dataset may result in bias and overestimation of model evaluation parameters. So we have used CD-HIT [12] to reduce the redundancy by removing sequences having more than 40% sequence identity, which resulted in 13,446 negative sequences. The final dataset consisted of approximately 1:30 positive to negative instances ratio. The data imbalance is intrinsically present in most of the protein family classification problems and affects the accuracy of predicting the members of a particular protein family. So the datasets are needed to be appropriately balanced to achieve the true performance of the classifiers.

2.2. Sequence based input features

The input vectors were created by extracting the following three types of features from every protein sequence.

(i) **Amino acid composition**: We used the percentage composition of amino acid residues (aa) as one of the feature vectors. This feature was selected on the assumption that there are some specific avoidances and preferences of certain amino acids in the formation of a protein family to perform a common functionality, which resulted in distinguishable frequency compositions (f_{res}).

$$f_{res} = \frac{N_{res,i}}{N_{total\_res,i}} \times 100 \quad (1)$$

where

$\boldsymbol{res}$ stands for one of the 20 different amino acid residues
$\boldsymbol{f_{res}}$ denotes the amino acid percentage frequency of the specific residue in *i*th Sequence.
$\boldsymbol{N_{res,i}}$ denotes the total count of amino acid of the specific type in the *i*th sequence.
$\boldsymbol{N_{total\_res,i}}$ denotes the total count of all residues in the *i*th sequence (i.e. sequence length).

(ii) **Amino acid property group composition:** The percentage frequency counts of amino acid property groups were used as the second component in the feature vector. The different amino acid property groups [13] that are selected for this study are given in Table 1. This is a refinement over amino acid frequency composition where specific property group count is computed instead of the individual amino acid count.

$$f_{pg} = \frac{N_{pg,i}}{N_{total\_res,i}} \times 100 \quad (2)$$

where

$\boldsymbol{pg}$ denotes one of the 11 different amino acid property groups
$\boldsymbol{f_{pg}}$ denotes the percentage frequency of the specific amino acid property group in the *i*th sequence.
$\boldsymbol{N_{pg,i}}$ denotes the total count of the specific amino acid property group in the *i*th sequence.
$\boldsymbol{N_{total\_res,i}}$ denotes the total count of all residues in the *i*th sequence.