Neurocomputing 000 (2017) 1–12

Contents lists available at ScienceDirect

Neurocomputing

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

The role of pertinently diversified and balanced training as well as testing data sets in achieving the true performance of classifiers in predicting the antifreeze proteins

Abhigyan Nath*, Karthikeyan Subbiah*

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

ARTICLE INFO

Article history:
Received 4 July 2015
Revised 21 June 2017
Accepted 4 July 2017
Available online xxx

Communicated by Zidong Wang

Keywords:
Antifreeze proteins
Imbalance data set
Incomplete learning
K-means clustering
Representative training set
Physicochemical-n-grams

ABSTRACT

Antifreeze proteins (AFPs) are those proteins, which inhibit the ice nucleation process and thereby enabling certain organisms to survive under sub-zero temperature habitats. AFPs are supposed to be evolved from different types of protein families to perform the unique function of antifreeze activity and turn out to be the classical example of convergent evolution. The common sequence similarity search methods have failed to predict putative AFPs due to poor sequence and structural similarity that exists among the different sub-types of AFP. The machine learning techniques are the viable alternative approaches to predict putative AFPs. In this paper, we have discussed about the criteria (like apposite feature selection, balanced data sets and complete learning) that are needed to be taken into account for successful application of machine learning methods and implemented these criteria by using a clustering procedure in order to achieve the true performance of the learning algorithms. Diversified and representative training and testing data sets are very crucial for perfect learning as well as true testing of machine learning based prediction methods for two reasons: first is that a training dataset that lacks definable subset of input patterns makes prediction of patterns belonging to this subset either difficult or unfeasible (thus resulting in incomplete learning) and secondly a testing data set that lacks definable subset of input patterns does not tell about whether this subset of patterns can be correctly predicted by the classifier or not (thus resulting in incomplete testing). Moreover, balanced training and testing data sets are equally important for achieving the true (robust) performance of classifiers because a well-balanced training set eliminates bias of the classifier toward particular class/sub-class due to over-representation or under-representation of input patterns belonging to those classes/sub-classes. We have used *K*-means clustering algorithm for creating the diversified and balanced training as well as testing data sets, to overcome the shortcoming of random splitting, which cannot guarantee representative training and testing sets. The current clustering based optimal splitting criteria proved to be better than random splitting for creating training and testing set in terms of superior generalization and robust evaluation.

1. Introduction

The sub-zero cold environmental habitats are prevalent on our planet earth from deep seas to permafrost. The living organisms in these habitats have adapted to thrive in extreme cold environmental conditions. The water in the liquid state is essential for the survival of living organisms. If the surrounding temperature goes below the freezing point of water (0°C) then the ice formation may begin which will be lethal to cell's functionalities and therefore threaten the very life of the organism itself. In nature ice anti-nucleation is facilitated with the assistance of antifreeze proteins [1]. Antifreeze proteins (AFPs) are those proteins which inhibit the ice nucleation process by lowering the freezing point of the intracellular fluids [2] and thereby enabling a wide range of organisms such as fishes, insects, plants, algae, bacteria, etc., to survive under sub-zero temperatures [3–6]. AFPs are also known as ice-structuring proteins and it was first discovered in the blood plasma of fishes by DeVries et al. [7,8]. These proteins are supposed to be evolved from different types of protein families to perform the unique function of antifreeze activity and turn out to be the classical example of convergent evolution.

An interesting example of diverse AFPs can be seen in fishes, in which they are classified into five different sub-types, namely: I, II, III, IV and AFGP. All these sub-types have no significant sequence and structural similarity among themselves and interestingly show homology to different protein families. In spite of all AFPs having

* Corresponding author.
E-mail addresses: abhigyannath01@gmail.com (A. Nath), karthinikita@gmail.com (K. Subbiah).

http://dx.doi.org/10.1016/j.neucom.2017.07.004
0925-2312/

Table 1
A case of imbalanced dataset and its effect on accuracy.

AFP class	0 (TP)	10(FN)	0.0% (Accuracy)
Non-AFP class	0(FP)	90(TN)	100% (Accuracy)

the same functionality, till date, no conserved ice binding motif has been discovered. Besides AFPs pose a problem for their In silico identification using common similarity search tools such as BLAST [9], PSI-BLAST [10] etc., due to the absence of any significant sequence similarity among them. These facts create the study on AFP- ice binding mechanism, extremely difficult and prompted few researchers to speculate few hypotheses for convergent evolution of antifreeze proteins in fishes [11,12]. As a result, the classifiers based on machine learning have become a viable alternative to predict AFPs.

1.1. Previous works

The selection of an apposite combination of input features plays a significant role in the performance of the prediction methods based on the machine learning algorithms. Researchers used different combinations of input features for creating the input vectors in their AFP prediction methods: (i) The amino acid composition along with PSSM profiles as evolutionary information were used as an input feature vector by Zhao et al. [13], (ii) The physicochemical properties of amino acid chains were used as an input feature vector by Kandaswamy et al. [14], (iii) n-peptides are used as one of the effective input features by Yu and Lu [15] and (iv) n-peptides along with physicochemical properties were used as input vector by Wen et al. [16] for further improving the prediction of AFPs. A discriminative model using AFP structural information about ordered surface carbon atoms was developed by Doxey et al. [17], in which the importance of physicochemical pattern was emphasized. Statistically significant avoidance and preference of amino acids and its property groups in the sequences of different fish AFP sub-types was also reported by Nath et al. [18].

1.2. Motivation

The discrimination of AFPs from proteins of other protein families presents an example of the class imbalance problem. A class imbalance happens when the members of a particular class label outnumber members of other classes by a good margin. The imbalanced class ratios are often encountered in the protein family classification problems. In the current classification task, AFPs are the positive minority class (which is the class of interest) and the majority class consists of all the non-AFPs belonging to protein families other than AFPs. Naturally the members belonging to the AFP class are much undersized as compared to the negative class non-AFPs.

Under the influence of heavy imbalance, the accuracy cannot be a representative of the true performance of the classifier. This imbalance in class distribution affects the accuracy in predicting the positive class instances to a great extent. For instance, imagine there are 90 patterns for the non-AFP class and 10 patterns for the AFP class (Table 1), the classifier can easily get biased toward the majority non-AFP class. If the trained model classifies all the patterns as non-AFPs, the overall accuracy will be 90% and this gives a false indication of the classifier performance. The classifier has a 100% recognition rate for the non-AFP class, but the trained model fails to perform optimally for the minority AFP class. Classifying all the patterns into the majority class still maintains high accuracy.

The issue of imbalanced data set was frequently addressed by researchers and they had proposed various methods to deal with it. For example: (i) Weighted under-sampling was used by Anand and Suganthan [19], (ii) Further Anand et al. [20] had successfully implemented the framework of class-wise feature selection by using class-wise optimized genes and probability estimates for multiclass cancer classification, (iii) Estabrooks et al. [21] had used a multiple re-sampling method to deal with imbalance data sets and (iv) Chawla et al. [22] had proposed SMOTE to deal with imbalance data sets. Besides, a few more researchers [23–27], had also used different techniques to overcome the problems of imbalanced dataset. In these papers, researchers had discussed about the impact of imbalanced data set and how it can drastically affect the results of the classifiers.

Previously it has also been pointed out that the training of machine learning algorithms with a natural distribution of instances belonging to different classes may not be optimal [28]. The issue of imbalanced data (both within and between class imbalances) has not been given the required consideration by the Bioinformatics community as it deserves. The random splitting of data into disjoint training and testing set is a common approach for measuring the performance of machine learning based prediction/classification methods. Keeping more diversified training data gives a better learning and consequently the better generalization and similarly more diverse testing data gives a true estimate for the classification error. Ideally, both training and testing sets should be a reasonable approximation of the entire input space and also free from bias. True performance evaluation metrics can be obtained from the classifiers if they are tested on the true distribution of all cases in the population. The choice of how particular training and testing sets are created strongly influences the estimated performance evaluation metrics. Data as well as a suitable choice of classifier performance evaluation metrics are equally important. Ideally the testing set should contain all the diverse patterns from both the majority and minority classes. If the testing set is small and biased, it may lead to imprecise performance evaluation metrics. If the classifier is trained on a subset of the available data, then the trained model will on average, perform worse than a classifier on diversified training set. Random partitioning into training and testing sets certainly reduces the size of the training set, which consequently reduces the model's ability to learn all the diverse patterns present in the entire input space.

Apart from between class imbalance that happens due to differences in the number of instances belonging to the majority and minority classes, there is also within class imbalance that occurs due to differences in the number of common patterns and rare patterns within each class. Also, these rare patterns constitute the small disjuncts, which are small regions in the input space having a very small number of instances [25]. Random splitting does not guarantee the inclusion of all common as well as rare patterns from both majority and minority classes in the training and testing set, i.e. it can't guarantee a representative training and testing set.

An unrepresentative training set may lead to poor generalization for the trained models [29] and to get the performance evaluation metrics to be meaningful, the training as well as the testing sets should be representative of the true distribution of the entire input space. Data splitting plays an important role in determining the true performance of the classifier. The current work addresses of how to optimally split the data into training and testing sets for assessing the true performance of the classifier. The representative testing set gives an unbiased estimate of classifier performance evaluation metrics as it has a representation of both common and rare patterns belonging to different class labels. The optimal split criterion will be helpful in those situations in which the performance of a classifier will be assessed on a hold out test set. If the data are less in training set, the learning algorithm may not generalize well on the holdout testing data. The optimal split criteria facilitates in creating a representative training set without

Please cite this article as: A. Nath, K. Subbiah, The role of pertinently diversified and balanced training as well as testing data sets in achieving the true performance of classifiers in predicting the antifreeze proteins, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2017.07.004