Computers in Biology and Medicine 73 (2016) 38–46

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Can-CSC-GBE: Developing Cost-sensitive Classifier with Gentleboost Ensemble for breast cancer classification using protein amino acids and imbalanced data

Safdar Ali, Abdul Majid*, Syed Gibran Javed, Mohsin Sattar

Department of Computer & Information Sciences, Pakistan Institute of Engineering & Applied Sciences (PIEAS), Nilore 45650, Islamabad, Pakistan

ARTICLE INFO

Article history:
Received 22 October 2015
Received in revised form
31 March 2016
Accepted 2 April 2016

Keywords:
Breast cancer
Tissue protein
Imbalanced data
Cost-sensitive classifier
GentleBoost ensemble

ABSTRACT

Early prediction of breast cancer is important for effective treatment and survival. We developed an effective Cost-Sensitive Classifier with GentleBoost Ensemble (Can-CSC-GBE) for the classification of breast cancer using protein amino acid features. In this work, first, discriminant information of the protein sequences related to breast tissue is extracted. Then, the physicochemical properties hydrophobicity and hydrophilicity of amino acids are employed to generate molecule descriptors in different feature spaces. For comparison, we obtained results by combining Cost-Sensitive learning with conventional ensemble of AdaBoostM1 and Bagging. The proposed Can-CSC-GBE system has effectively reduced the misclassification costs and thereby improved the overall classification performance. Our novel approach has highlighted promising results as compared to the state-of-the-art ensemble approaches.

1. Introduction

Worldwide, cancer is the second most fatal disease and breast cancer is the second main cause of cancer related deaths. The International Agency for Research on Cancer has reported for 2012 that 1,677,000 women were diagnosed with breast cancer and 577,000 women died with this disease [1]. In Pakistan, due to lack of treatment facilities approximately 17,552 women would die with breast cancer in 2015. Like other diseases, however, breast cancer can be successfully treated if diagnosed in the early stages [2–4]. For the early classification of breast cancer, development of an effective decision support system is a critical task.

Several conventional, statistical, and machine learning (ML) techniques are used for the detection/prediction of breast cancer. Mammography/imaging-diagnosis is one of the methods used for the detection of breast cancer; however, it has considerable variation in interpreting the results of graphs. For this reason, in recent times, many other statistical and ML approaches have been proposed for the prediction of cancers.

The fields of sequencing of human genome and proteins are well established. The growth of proteome data increased rapidly. For example, various international projects such as The Human Proteome Project and The Human Genome Project has been generated huge amount of data. This data is being used in various medical applications of diagnostic, prognostic, therapeutic, and preventive. The knowledge extracted from these projects increasingly promises the potential for future widespread adoption in personalized medicine and diagnostics. As the development in human genome and proteins has increased rapidly, so in recent times, researchers and practitioners would use protein data for clinic/research. In the view of fact, the ML based approaches (such as the proposed approach) will have very bright practical applications and benefits.

The sequencing of human genome and proteins are well established that are now utilized to find new cancer related molecular descriptors and biomarkers [5]. These markers are employed to develop improved decision support systems, since the mutated proteins are associated with breast cancer. Therefore, features extracted from such protein disorders would be used for cancer classification, treatment, and drug discovery. Recently, protein sequences are employed for the classification of ovarian cancer [6], lung cancer [7], colon cancer, and breast cancer [8]. Xin et al. used sequence-based features for the prediction of DNA binding residues in protein sequences [9].

As cancer is a genetic disease, investigating various kinds of mutations in genetic material such as DNA, RNA, and proteins, can reveal important underpinnings of carcinogenesis and cancer proliferation. In this paper, for classification of breast cancer, we have utilized the discriminant information of mutated protein

* Corresponding author.
E-mail addresses: safdarali_11@pieas.edu.pk (S. Ali), abdulmajiid@pieas.edu.pk (A. Majid), gibranjaved_11@pieas.edu.pk (S.G. Javed), mohsin_14@pieas.edu.pk (M. Sattar).

http://dx.doi.org/10.1016/j.compbiomed.2016.04.002
0010-4825/© 2016 Elsevier Ltd. All rights reserved.

molecules with physicochemical properties hydrophobicity (H_b) and hydrophilicity (H_p) of amino acids. These physicochemical properties exhibit excellent discriminant capability in different feature spaces for amino acid sequences classification [10].

Usually, in cancerous data, number of cancer and non-cancer patients is inherently imbalanced i.e., the particular diagnosis class is not easily achievable. Thus, the decision boundary of conventional classifier is biased towards majority class. To address this problem, different techniques are suggested; either processing the input data or use the cost sensitive learning (CSL). In this study, we have employed CSL technique, which minimized the misclassification costs.

The over-sampling technique is used to put more novel data examples to the rare class to balance dataset. The new data examples are produced using synthetic techniques or by duplicating the data examples of the minority class. Synthetic minority oversampling technique (SMOTE) adds new synthetic samples to the minority class by randomly interpolating pairs of the closest neighbors [11]. Usually, SMOTE engages in making copies of samples and consequently lead to overfitting [12]. For small medical dataset, Mega-Trend diffusion method adds new synthetic samples to the minority class by employing membership function rather than normal distribution to compute the possibility values of synthetic [13]. On the other hand, under-sampling can discard potentially useful medical and biological information of the majority data class that could be important for the induction process. For example, given imbalance ratio of 100:5, in order to get a close match for the minority class, it might be undesirable to throw away 95% of majority class instances. Therefore, to avoid the risk of deleting useful information form majority data class and to prevent overfitting in case of over-sampling, we preferred to use cost-sensitive learning (CSL) approach.

Previously, Wang et al. [14] constructed five-year breast prognosis models by combining Logistic Regression (LR) and Decision Tree (DT) with cost-sensitive classifier (CSC) technique, Bagging, and Boosting. Their proposed CSC ensemble models showed improved performance than the original models. They reported accuracy up to 91.30%. Liu et al. have applied under-sampling approach with DT to deal with imbalanced problem for breast cancer survivability and reported 86.52% survival rate of patients [15]. Zhang et al. utilized gene expression profiles for the prediction of breast cancer by employing LR, Support Vector Machine (SVM), AdaBoost, LogitBoost and Random Forest (RF) [16]. They achieved maximum value of Area Under the receiver operating characteristic Curve (AUC) measure of 88.6% and 89.9% for SVM and RF models, respectively. Delen et al. used surveillance and epidemiology results for prediction of breast cancer [17]. They employed three classification models of DT, LR, and Artificial Neural Network (ANN) based learning approaches. They obtained the highest value AUC of 84.9% and 76.9% for LR and DT models, respectively. In another study, Khalilia et al. developed prediction models from highly imbalanced data using SVM, Bagging, Boosting and RF [18]. They demonstrated that, in terms of AUC measure, RF model (91.2%) outperformed SVM (90.6%), Bagging (90.5%), and Boosting (88.9%).

It is a challenging task for an individual learner to develop an improved prediction models for breast cancer [4,19]. Therefore, Boosting and Bagging based ensemble systems were developed for cancer dataset. These ensemble systems were constructed by a set of trained classifiers with the same learning classifier. They attempt to enhance the performance by iteratively retraining the base classifiers with a subset of most informative data and then combining their predictions with novel examples. These systems have limited performance due to small number of samples and class imbalance. The main novelty in this study is the development of CSL based GentleBoost ensembles (Can-CSC-GBE) using physicochemical properties H_b and H_p of amino acids as molecule descriptors in different feature spaces. To the best of our knowledge, previously, this aspect has not been explored for imbalanced data in the context of breast cancer classification.

In the proposed study, first, molecule descriptors are generated in four different feature spaces of (i) Amino Acid Composition (AAC) of dimensions 20, (ii) Split Amino Acid Composition (SAAC) of dimensions 60, (iii) Pseudo Amino Acid Composition-Series (PseAAC-S) of dimensions 40, and (iv) Pseudo Amino Acid Composition-Parallel (PseAAC-P) of dimensions 60. The CSL technique is then employed in feature spaces to reduce the misclassification costs. In the next step, we employed GentleBoost ensemble to construct Can-CSC-GBE system (GBE_{FS}^{CSC}) for different feature spaces (FS) using 10-fold jackknife technique. For comparison purpose, ensemble system of AdaBoostM1, and Bagging are implemented using CSC technique to develop $\mathrm{AdaM1}_{FS}^{CSC}$ and Bag_{FS}^{CSC} models. The experimental results demonstrate that $\mathrm{GBE}_{PseAAC-S}^{CSC}$ model using PseAAC-S feature space is superior to individual, $\mathrm{AdaM1}_{FS}^{CSC}$, and Bag_{FS}^{CSC} models.

2. Material and methods

Framework of the proposed CSC based GentleBoost ensemble cancer classification is shown in Fig. 1. The proposed system consists of three main modules: the feature space, the CSC development, and the ensemble development. The proposed system is assessed using two datasets of protein amino acid sequences for cancer/non-cancer (C/NC) and breast/non-breast cancer (B/NBC). These datasets are borrowed from [8]. These datasets consist of 1056 protein sequences. First C/NC dataset is composed of 865 non-cancer and 191 cancerous protein sequences, whereas the second B/NBC dataset is containing 865 non-cancer and 122 breast-cancer related protein sequences. The next subsection describes the feature space generation of protein primary sequences.

2.1. Feature generation

Proper input representation of protein primary sequences make easier for a classifier to recognize underlying regularities in the sequences. The native twenty amino acids in a protein sequence are usually illustrated by set of single letter codes of English letters. A protein of length L_r is formally represented as an ordered sequence $p = (a_1, a_2, \ldots, a_L)$ with elements a_i, from the finite set ={A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}, where A stand for Alanine, C stand for Cysteine, E stand for Glutamic acid, etc. The raw information given by the protein sequence is customarily restructured for prediction such that each representation of protein sequences is suitable for a feature space. For cancer prediction, we used correlation based discriminant feature extraction strategies of AAC, SAAC, PseAAC-S, and PseAAC-P.

In AAC feature space (20-dimensions), a vector of the relative frequencies of the 20 native amino acids represents each protein in its sequence as:

$$f_i = \frac{n_i}{L_r} \qquad (i = 1, 2, \ldots, 20) \tag{1}$$

where f_i represents the occurrence frequency of the i-th native amino acid in the protein, n_i is the number of the i-th native amino acid in sequence. Then, the AAC feature vector is expressed as:

$$\mathbf{x}_{AAC} = \left[f_1, \ f_2, \ \ldots, \ f_{20}\right]^T \tag{2}$$

In SAAC feature space generation, the given protein sequence is split into three dissimilar sections, named the N-terminal, the Internal segments and the C-terminal [20,21]. The amino acid