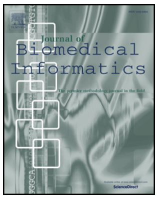
Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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



Mining features for biomedical data using clustering tree ensembles

Konstantinos Pliakos*, Celine Vens

KU Leuven, Campus KULAK, Department of Public Health and Primary Care, Faculty of Medicine, 8500 Kortrijk, Belgium

ARTICLE INFO

Keywords:
Tree-ensembles
Extremely randomized trees
Tree-embeddings
Biomedical data mining

ABSTRACT

The volume of biomedical data available to the machine learning community grows very rapidly. A rational question is how informative these data really are or how discriminant the features describing the data instances are. Several biomedical datasets suffer from lack of variance in the instance representation, or even worse, contain instances with identical features and different class labels. Indisputably, this directly affects the performance of machine learning algorithms, as well as the ability to interpret their results. In this article, we emphasize on the aforementioned problem and propose a target-informed feature induction method based on tree ensemble learning. The method brings more variance into the data representation, thereby potentially increasing predictive performance of a learner applied to the induced features. The contribution of this article is twofold. Firstly, a problem affecting the quality of biomedical data is highlighted, and secondly, a method to handle that problem is proposed. The efficiency of the presented approach is validated on multi-target prediction tasks. The obtained results indicate that the proposed approach is able to boost the discrimination between the data instances and increase the predictive performance.

1. Introduction

Significant progress has been made in biomedical data generation techniques and feature extraction methods. Microarrays, high-throughput sequencing, mass spectrometry and many others have been recognized as tools of great importance. The scientific community exploits these technological advances to generate more data in the field of biomedicine. However, a thoughtful question would be how informative these data truly are to make predictive inferences.

More precisely, among all the scientific tools in biomedical informatics, machine learning has proved to be of paramount importance, especially in predictive problems. It has provided contributions in a wide range of problems in biology and biomedicine [1,2]. Protein function prediction, interaction prediction between drugs and proteins, knowledge discovery related to biomarkers are only some of the various examples. In particular, in the framework of supervised learning several methods are provided to learn predictive models only from former observations of a system. In supervised learning, the instances are described by features (characteristics of each instance) and accompanied by targets. The goal is to learn a model on a training set of instances that can predict the target given the features [3]. This model is then used to predict the target of new unseen instances. If the target is numeric, the task is called regression. If the target is categorical (i.e., a class has to be predicted), the term classification is used. In some prediction problems, rather than a single target, a *set* of targets needs to be predicted. The mentioned applications of gene function prediction or interaction prediction are instances of this so-called multi-target prediction task [4]. Multi-target prediction is a generalization of multi-target regression and multi-target classification. Multi-label classification can be considered as an instance of multi-target classification with only two nominal values for each target [5,6]. These applications can be tackled by either transforming the multi-target problem into a set of single-target problems and applying a standard prediction algorithm, or by applying a specific multi-target prediction algorithm.

One of the most popular machine learning algorithms are the decision tree induction algorithms [7]. They have been applied extensively in biological and biomedical systems [2,8]. Decision tree algorithms have many advantages over other machine learning methods. They provide interpretability of the induced models, providing transparency and insights to scientists, leveraging this way knowledge discovery. Their models are also scalable, computationally efficient and accurate. The predictive performance of decision trees is specifically boosted when they are combined with ensemble methods [9], providing state-of-the-art results. Decision tree learning and its ensemble extension have been extended to the multi-target prediction setting (i.e., multi-label classification and multi-target regression) [4]. Multi-label classification has received more attention than multi-target regression [10].

* Corresponding author.
E-mail address: konstantinos.pliakos@kuleuven.be (K. Pliakos).

https://doi.org/10.1016/j.jbi.2018.07.012
Received 24 November 2017; Received in revised form 2 July 2018; Accepted 12 July 2018

1.1. Problem statement

In biomedicine and healthcare representing the data effectively is often a challenging task. In comparison to other fields, such as multimedia analysis or web mining, getting new features for an instance is a substantially more difficult task in biomedicine. In the case of a song or an image being under-represented one can easily get (extract) more features by collecting more information about the artist or applying an audio or image processing technique to the instance itself. However, this is not always possible in biomedicine. It requires significant time, expenditure, and the presence of human experts, as the features most of the times are the outcome of laboratory work. Furthermore, in medicine the data are usually produced through the process of patient care and thereby issues of privacy are included [1]. In many countries, hospitals are equipped with inefficient database systems that are mainly designed for billing purposes. In addition, medical data inevitably contain many missing values and many non-numeric features. Many traditional machine learning methods (e.g., SVM [11]) cannot be applied on such data without preprocessing. These peculiarities of the data in healthcare and biomedicine often lead to non-variant data representations and inevitably the harassment of the performance of machine learning methods [12]. To this end, methodologies to handle this uniqueness of biomedical data are needed.

It is crucial for the performance of machine learning models that the features that describe the data instances are enriched with valuable information and are able to discriminate the data instances. However, there are many examples of broadly used benchmark biomedical datasets suffering from this phenomenon of lacking variance. An extreme version of the aforementioned phenomenon is the existence of biomedical datasets that contain identical feature vectors for different data instances. Some examples from the field of gene function prediction (e.g., [13,14]) or interaction prediction (e.g., [15]) are shown in Table 1 [16,17]. This is irrational as there are instances (e.g., genes), that have different targets (e.g., functions) but identical feature vectors.

In the *pheno* dataset for example, 32% of the instances have a unique feature vector. As a unique feature vector we denote a feature vector that exists only once in the dataset after removing the replicates (i.e., instances that are described by exactly the same feature vectors). This dataset contains an instance for every gene in the *S. cerevisiae* organism, and the instance corresponds to a mutant, i.e. an organism where the corresponding gene has been altered. There are 69 features corresponding to different growth medium (e.g., caffeine, sorbitol, benomyl, …), on which growth of the mutant is recorded. The values of the features correspond to the observed sensitivity or resistance of the mutant: no effect, less growth or better growth compared to the wild type, or no data for the growth medium. The large number of replicates (i.e., instances that are described by exactly the same feature vectors) can be explained by two observations: (1) the data is very sparse, i.e. for many mutants few growth media have been tested, and (2) in many cases the tested growth media has no effect on the mutant. The *church* dataset also contains many instances with identical feature values. It consists of 27 mostly real-valued features. Interestingly, the replicate issue is not limited to datasets with few features. For example, *hom* contains homology information encoded by 47034 binary features, *struc* contains 19628 binary features related to predicted secondary structure of the protein. Moreover, apart from genes the aforementioned problem is also existing in interaction networks. In Table 1, an example of a drug-protein interaction network (DPI) [18] is presented where there are some drugs or proteins having the same feature vector. In drug-protein interaction networks connections are formed between drugs and proteins when the drug targets the protein. Both interaction parts are described by their own set of features, for example in the current dataset each drug is described by the presence or absence of 660 chemical substances and each protein by the presence or absence of 876 PFAM domains.

Table 1
Datasets, the number of features, instances and unique feature vectors.

Context	Dataset	|features|	|instances|	|unique feature vectors|
Gene fct. prediction	church	27	3755	2352
(*S. cerevisiae*)	pheno	69	1592	514
	hom	47034	3854	3646
	seq	478	3919	3913
	struc	19628	3838	3785
(*A. thaliana*)	scop	2003	9843	9415
	struc	19628	11763	11689
Interaction prediction	drugs	660	1862	1779
(DPI network)	proteins	876	1554	**683**

Lack of variance in the data representation can heavily impair the performance of machine learning algorithms. Indicatively, by applying an 1 nearest neighbor (1-NN) classifier to a training dataset (including the specific query instance), one expects to get 100% accuracy. However, in case there are instances in that dataset with exactly the same feature representation, it is not guaranteed that an instance is mapped onto itself. As an example we could refer to [16] where ML-KNN [19] with $K = 1$ was applied on pheno dataset from Table 1 and it yielded a precision of only 51.59%. Moreover, instances with identical feature vectors will end up into the same leaf in case of decision trees or their ensemble extension. Although this is what is expected, it causes a performance issue if those instances (i.e., instances in the same leaf) have completely different labels. The algorithm is supposed to separate instances with very different labels into different leaves. In Fig. 1, the distances between the feature vectors representing the 1592 data instances of pheno and the distances between the corresponding label vectors are displayed. More specifically, distances were computed between the feature vectors and between the label vectors. Each cell (i, j) of the matrix was assigned a color based on the distance between the feature (label) vectors that correspond to instances i and j. White corresponds to distance equal to zero and black to distance equal to one. It is shown that instances represented by identical feature vectors (distance equal to zero) are associated with different labels.

The issue of replicate feature vectors is an extreme example of the more general phenomenon of lacking variance in biomedical data. Although it is difficult to provide any algorithmic solution to the existence of replicate feature vectors, the scientific community should be aware of its presence.

1.2. Related work

Whereas numerous studies have focused on feature selection [20–22], the more difficult task of feature construction or induction has received less attention. The main goal of feature construction is to augment the feature space by creating or inferring additional features [23].

In [24], a feature construction method specifically focused on multi-label classification was presented. A distinct feature set was assigned to every label, increasing this way the performance of a classifier trained for that specific label. The label-specific features were generated by first clustering the positive and negative instances (separately) of the label. Next, the distances of each instance to the obtained cluster centroids were calculated. Vens and Costa [25] proposed a feature induction technique based on random forest [9]. A metric transformation was proposed, mapping the identity of the tests performed in each node of a decision tree to a feature indicator. Next, the final representation is yielded by concatenating the features associated with the trees in the forest and encoding them with hashing. The method was demonstrated on binary and multi-label classification tasks. A similar work involving a set of random clustering forests was proposed in [26,27]. The application was focused on the construction of visual vocabularies. More specifically, randomized trees were used to generate the new feature