Expert Systems With Applications 64 (2016) 305–316

Contents lists available at ScienceDirect

Expert Systems With Applications

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

An adaptive rule-based classifier for mining big biological data

Dewan Md. Farid [a,*], Mohammad Abdullah Al-Mamun [b], Bernard Manderick [a], Ann Nowe [a]

[a] Computational Modeling Lab, Department of Computer Science, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium
[b] Department of Population Medicine & Diagnostic Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY 14850, USA

ARTICLE INFO

Article history:
Received 2 January 2016
Revised 1 August 2016
Accepted 2 August 2016
Available online 3 August 2016

Keywords:
Brugada syndrome
Classification
Decision tree
Genomic data
Rule-based classifier

ABSTRACT

In this paper, we introduce a new adaptive rule-based classifier for multi-class classification of biological data, where several problems of classifying biological data are addressed: overfitting, noisy instances and class-imbalance data. It is well known that rules are interesting way for representing data in a human interpretable way. The proposed rule-based classifier combines the random subspace and boosting approaches with ensemble of decision trees to construct a set of classification rules without involving global optimisation. The classifier considers random subspace approach to avoid overfitting, boosting approach for classifying noisy instances and ensemble of decision trees to deal with class-imbalance problem. The classifier uses two popular classification techniques: decision tree and k-nearest-neighbor algorithms. Decision trees are used for evolving classification rules from the training data, while k-nearest-neighbor is used for analysing the misclassified instances and removing vagueness between the contradictory rules. It considers a series of *k* iterations to develop a set of classification rules from the training data and pays more attention to the misclassified instances in the next iteration by giving it a boosting flavour. This paper particularly focuses to come up with an optimal ensemble classifier that will help for improving the prediction accuracy of DNA variant identification and classification task. The performance of proposed classifier is tested with compared to well-approved existing machine learning and data mining algorithms on genomic data (148 Exome data sets) of Brugada syndrome and 10 real benchmark life sciences data sets from the UCI (University of California, Irvine) machine learning repository. The experimental results indicate that the proposed classifier has exemplary classification accuracy on different types of biological data. Overall, the proposed classifier offers good prediction accuracy to new DNA variants classification where noisy and misclassified variants are optimised to increase test performance.

1. Introduction

The emerging field of bioinformatics combines the challenging research areas of biology and informatics to develop different methods and tools for analysing biological data. The main challenge behind is to extract the relevant information from the large amount of clinical and genomic data, then transform it into useful knowledge (Taminau, 2012). Three major issues are involved in this process: (a) collecting clinical and genomic data, (b) retrieving relevant information from the data and (c) extracting new knowledge from the information. Since last decade various life science research groups generated huge amount of clinical and genomic data from the Human Genome Project (HGP) and some of them are publicly available through online repositories (Artimo et al., 2012; Berman et al., 2000; Lichman, 2013). Routinely, the computational intelligence researchers have applied machine learning (ML) and data mining (DM) algorithms for explicating the biological data (Latkowski & Osowski, 2015; Tzanis, Kavakiotis, & Vlahavas, 2011; Yang & Chen, 2015). Typically the biological data are noisy, high dimensional space (in thousands), small size of samples (in dozens) and some gene sequences have large variance (Alter et al., 2011), which results in the danger of overfitting and low efficiency to classification.

Biological data mining (BDM) is the process of extracting new knowledge (previously unknown) from the biological data. It presents extensive DM concepts, theories and applications in biological research. DM uses ML algorithms for discovering patterns and useful information from large data or databases (Farid et al., 2013; Han, Kamber, & Pei, 2011). DM has two major functions: (a) classification (supervised learning) and (b) clustering (unsupervised learning). In classification, the mining classifiers predict the

* Corresponding author. Fax: +32 26291879.
E-mail addresses: Dewan.Md.Farid@vub.ac.be (D.Md. Farid), Ma875@cornell.edu (M.A. Al-Mamun), bmanderi@vub.ac.be (B. Manderick), ann.nowe@vub.ac.be (A. Nowe).

http://dx.doi.org/10.1016/j.eswa.2016.08.008
0957-4174/

class value of a new/unseen instance after remarking the training data (Farid, Zhang, Rahman, Hossain, & Strachan, 2014; Nápoles, Grau, Bello, & Grau, 2014). The training instances are grouped into classes before mining the data. On the other-side, clustering (or segments) groups the instances into clusters based on the similarities among the instances on predefined features (Milone, Stegmayer, Kamenetzky, López, & Carrari, 2013). The instances within a cluster have more similarity in comparison to one another but are very dissimilar to instances in other clusters (Al-Mamun et al., 2016). Both classification and clustering methods play an important role to analyse biological data such as genomic/DNA microarray data classification and analysis (Hanczar & Nadif, 2011; 2012; Liew, Yan, & Yang, 2005; Lin, Liu, Chen, Chao, & Chen, 2006). But mining becomes more difficult when biological data has large number of features and small number of instances/variants (Gheyas & Smith, 2010; Hua, Tembe, & Dougherty, 2009).

This paper presents an adaptive rule-based (ARB) classifier for classifying multi-class biological/genomic data to improve the prediction accuracy of DNA variants classification task. Where it uses two efficient and effective supervised learning algorithms: decision tree (DT) and k-nearest-neighbor (kNN) method. DTs are used for evolving a set of classification rules from the training data, while kNN is used for analysing the misclassified instances and removing ambiguity between the contradictory rules. It is suggested that rules make it easy to deal with complex classification problems. Rule-based classifier has various advantages: (a) highly expressive as DT, (b) easy to interpret, (c) easy to generate, (d) can classify new instances rapidly, and (e) performance comparable to DT. Also, new rules can be added to existing rules without disturbing ones already in there and rules can be executed in any order (Han et al., 2011). Usually, there are two characteristics of the rule-based classifier: (a) classifier contains mutually exclusive rules if the rules are independent of each other, where every instance is covered by at most one rule, and (b) classifier exhaustive coverage if it accounts for every possible combination of feature values, where each instance is covered by at least one rule (Witten, Frank, & Hall, 2011). A rule-based classifier makes use of a set of *IF-THEN* rules for classification. The *IF* part of the rule is called rule antecedent or precondition. The *THEN* part of the rule is called rule consequent. The antecedent part of condition consists of one or more feature tests and these tests are logically *ANDed*. The consequent part consists of the class prediction.

The proposed ARB classifier combines the random subspace and boosting approaches with ensemble of DTs to construct a set of classification rules. We consider random subspace approach to avoid overfitting, boosting approach for classifying noisy instances and ensemble of DTs to deal with class-imbalance problem. It considers a series of k iterations to generate classification rules. A rule set is generated using a DT with random subspace from the training data in each iteration. Each rule is generated for each leaf node of the tree. Each path in the tree from root to a leaf corresponds with a rule. The rules are extracted from the training data using the C4.5 (DT induction) algorithm. The ARB classifier mainly concentrates to the misclassified training instances in the next iteration and conclusively finds the instances those are difficult to classify. An instance's weight reflects how difficult it is to classify. The weights of instances are adjusted according to how they are classified in each iteration. If an instance is correctly classified, then its weight is decreased, otherwise if misclassified, then its weight is increased. The ARB classifier produces good classification rules without any need for global optimisation. We have tested the performance of proposed ARB classifier on 148 Exome data sets (Brugada syndrome variants classification) and 10 real benchmark life sciences data sets from the UCI (University of California, Irvine) machine learning repository (Lichman, 2013). The experimental analysis of proposed classifier has compared with very popular and strong classification algorithms: (a) RainForest tree, (b) DT (C4.5), (c) naïve Bayes (NB) classifier, and (d) kNN classifier. The DT, NB and kNN classifiers have been using for mining biological data in last few years. We have chosen these classifiers because they also result in transparent models that can are human interpretable. Other models like random forests usually result in better performance, but are poor from a transparency point of view.

The rest of the paper is organised as follows. Section 2 presents the knowledge discovery process from big biological data. Section 3 introduces DT and kNN algorithms. Section 4 presents the proposed ARB classifier in details. Section 5 presents the algorithm for analysing misclassified instances. Section 6 provides experimental results on genomic and benchmark life sciences data sets. Finally, Section 7 concludes the findings and proposed directions for future work.

2. Mining big biological data

2.1. Mining big data

Mining big data is the process of extracting knowledge to uncover large hidden information from the massive amount of complex data or databases (Al-Jarrah, Yoo, Muhaidat, Karagiannidis, & Taha, 2015; Assunção, Calheiros, Bianchi, Netto, & Buyya, 2015). The data in big data comes in different forms including two-dimensional tables, images, documents and complex records from multiple sources (Kambatla, Kollias, Kumar, & Grama, 2014). It must support search, retrieval and analysis (Barbierato, Gribaudo, & Iacono, 2014; Singh & Reddy, 2014). The three V's define big data: *Volume* (the quantity of data), *Variety* (the category of data) and *Velocity* (the speed of data in and out). It might suggest throwing a few more V's into the mix: *Vision* (having a purpose/plan), *Verification* (ensuring that the data conforms to a set of specifications) and *Validation* (checking that its purpose is fulfilled) (Jin, Wah, Cheng, & Wang, 2015).

Regarding data and their uses, there is a big difference between big data and business intelligence (BI). BI provides historical, current and predictive views of business operations, which includes reporting, online analytical processing, business performance management, competitive intelligence, benchmarking and predictive analytics. But in big data, the data is processed employing advanced ML and DM algorithms to extract the meaningful information/knowledge from the data (Chen & Zhang, 2014; Najafabadi et al., 2015). In big data mining applications, very large training sets of millions of instances are common. Most often the training data will not fit in memory. The efficiency of existing ML and DM algorithms, such as DT, NB and kNN, has been well established for relatively small data sets (Gehrke, Ramakrishnan, & Ganti, 2000). These algorithms become inefficient due to swapping of the training instances in and out of main and cache memories. So, more scalable approaches are required to increase the capability of handling training data that are too large to fit in memory.

2.2. Classifying big biological data

In this 'Omics' era, life science generates the genomic, transcriptomic, epigenomic, proteomic, metabolomics data scaling from terabyte (TB) to petabyte (PB), even exabyte (EB) (López, Aguilar, Alonso, & Moreno, 2012). It is plausible to mention that these biological data have similar properties 3V's (volume, variety and velocity) like big data has. But these biological data are highly heterogeneous than other contemporal big data like Google, WeChat and Ali Baba, where 3V features showcase causal relationships among certain elements like genes, proteins, and pathways in molecular level (Toga & Dinov, 2015). Knowledge discovery from biological data is a DM application. The process of