International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Adaptive Rule-Based Classier using J48 For Extracting Big Biological Data

Dhanashree Patil and Prof. Dhanashree Kulkarni

Department of Computer Engineering D. Y. Patil college, Pune Savitribai Phule, Pune University Pune, 411041, Maharashtra, India.

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

Biological field of research have already produced big amounts of valuable biological data that is challenging to analyze due to its high dimensionality and complexity. With this huge amount of data there are several problems during classification of data such as: over fitting, noisy instances and class imbalance. This will affect both the accuracy and the efficiency of supervised learning methods. This paper proposes a data-adaptive rulebased classification system for biological big data classification that generates relevant rules by finding adaptive partitions. The proposed system is a rule based classifier, which is combination of random subspace and boosting approaches. To construct the classification rules without global optimization, system makes use of J48 Decision Tree and KNN algorithm. Random subspace is used to avoid over fitting problem, boosting approach is used for solving problem of noisy instances classification and finally J48 decision tree is deal with class imbalance problem. With J48 decision tree, rules are evaluated from training data and with KNN, misclassified instances are analyzed. The proposed approach will be compared with other rule-based and machine learning classifiers, and detailed results and discussion of the experiments are presented to demonstrate comparative analysis and the efficacy of the results. Results will prove the good prediction accuracy of classified DNA.

Keywords: *Biological data, J48 classifier, KNN classifier, misclassified instances.*

Introduction

The biologists are stepping up their reports to understand the biological processes that underlie disease pathways. This has resulted in a good of biological and clinical data from genomic sequences, DNA microarrays, and protein interactions, to biomedical images, disease pathways, and electronic health records. They are at position where the ability to generate biomedical data has greatly surpassed our ability to mine and analyze the 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 Three major issues are involved in this process collecting clinical and genomic data, retrieving relevant information from the data and extracting new knowledge from the information. Since last decade various life science study groups generated huge amount of clinical and genomic information from the Human Genome Project (HGP) and some of them are publicly available through online repositories. Routinely, the computational intelligence researchers have applied machine learning (ML)and data mining (DM) algorithms for illustrating the biological data. Typically the biological data are noisy, high dimensional space, small size of samples and some gene sequences have large variance, which results in the danger of obverting and low efficiency to classification. Biological data mining (BDM) is the process of deriving new knowledge (previously unknown) from the biological data. It represents major DM concepts, theories and applications in biological research. DM uses ML algorithms for identifying patterns and useful data from large data or databases. DM consist of two major functions such as classification (supervised learning) and clustering (unsupervised learning). In classification, the mining Classifiers predict the class value of a new/unseen instance after remarking the training data.

This paper proposed 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 optimization. The classifier considers random subspace approach to avoid overfiting, boosting approach for classifying noisy instances and ensemble

of decision trees to deal with class-imbalance problem. The classifier uses two popular classification techniques: J48 decision tree and k-nearest neighbor algorithms. J48 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 flavor. 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.

In this paper we study about the related work done, in section II, the proposed approach modules description, mathematical modeling, algorithm and experimental setup in section III and at final we provide a conclusion in section IV.

Review Of Literature

In this section discuss the literature review in detail about the recommendation system for online social network.

In this paper [1], Farid et al. introduce a new adaptive rule-based classifier for multiclass classification of biological data, where several problems of classifying biological data are addressed: over fitting, noisy instances and class-imbalance information. It is well known that rules are interesting way for illustrating data in a human interpretable way. The proposed rule-based classifier associate the random subspace and boosting approaches with ensemble of decision trees to construct a set of classification rules without involving global optimization.

This paper [2] proposed an in silico model composed of an in silico tumor growth model and a data verification technique using clustering. First, the extended in silico model upgraded their previous model for maspin dynamics and added cell ECM interactions, cell cell adhesion and cell movement constraints in presence of maspin. Our results suggest that maspin has influence on microenvironment constraint including cell ECM, cellcell adhesion and cell movements which show good agreement with the previous in vitro model hypotheses.

In this paper [3] a new framework for feature selection consisting of an ensemble of filters and classifiers is represented. Five filters, depend on different metrics, were employed. Each filter selects a different subset of features which is used to train and to test a specific classifier. The outputs of these five classifiers are combined by simple voting. In this study three well-known classifiers were employed for the classification task: C4.5, naive-Bayes and IB1. The rationale of the ensemble is to reduce the variability of the features selected by filters in different classification areas.

In this paper [4], Dr. Dewan Md. Farid and Prof. Dr. Chowdhury Mofizur Rahman proposed a new decision tree learning algorithm by assigning appropriate weights to each training instance in the training data that increases classification accuracy of the decision tree model. The main asset of this proposed technique is to set appropriate weights to training instances using nave Bayesian classifier before trying to construct the decision tree. In this approach the training instances are assigned to weight values based on the posterior probability. The training instances having less weight values are either noisy or posses unique characteristics compared to other training example.

It is challenging to use traditional data mining method to deal with real-time data stream classifications. Existing mining classifiers need to be updated frequently to adapt to the changes in data streams. To address this issue, this paper [5] developed an adaptive ensemble approach for classification and novel class detection in concept drifting data streams. The proposed technique uses traditional mining classifiers and updates the ensemble model automatically so that it represents the most recent concepts in data streams. For novel class detection we consider the idea that data points belonging to the same class should be closer to each other and should be far apart from the data points belonging to other classes. If a data point is well separated from the existing data clusters, it is identified as a novel class example.

In this paper [6], Farid et al. developed two independent hybrid mining algorithms to upgrade the classification accuracy rates of decision tree (DT) and nave Bayes (NB) classifiers for the classification of multi-class issue. in data mining, both DT and NB classifiers are useful, efficient and commonly used for solving classification problems. Since the presence of noisy contradictory instances in the training set may cause the generated decision tree suffers from over fitting and its accuracy may decrease, in our first proposed hybrid DT algorithm, they employ a nave Bayes (NB) classifier to remove the noisy troublesome instances from the training set before the DT induction.

In this paper [7], Gehrke et al. developed a unifying framework called Rain Forest for classification tree construction that separates the scalability aspects of algorithms for constructing a tree from the central features that determine the quality of the tree. The generic algorithm is easy to instantiate with specific split selection methods from the literature (including C4.5, CART, CHAID, FACT, ID3 and extensions, SLIQ, SPRINT and QUEST).

System Architecture / System Overview

A. Proposed System Overview

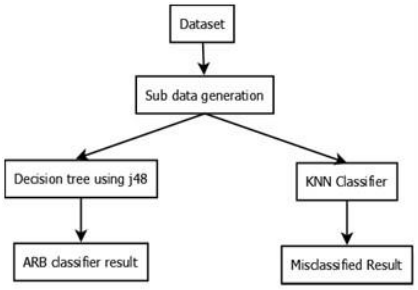

Fig. 1. Proposed System Architecture

Techniques used to implement this system:

1) Input Dataset:

This system take cancer, diabeties and iries dataset as a input in with attributes to classify the data and generate adaptive rule-based classifier.

2) Sub-data Generation:

In this module this biological data deivided in sub-data generation where arff files of each dataset is generated as an output.

3) J48 Decision Tree:

In this module decision tree is implementing to generate adaptive rules. The key disadvantage of DTs is that without proper pruning (or limiting tree growth), trees tend to overfit the training data. the base paper used C4.5 method to extract the classification rules from training data. Each rule is generated for each leaf node of the tree. Each path in tree from the root to a leaf corresponds with a rule. In contribution we used J48 classifier to overcome the C4.5 disadvantages. J48 is an extension of ID3. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. It save both time and memory and provide accurate classification result.

4) KNN Classification:

In this module KNN classifier is used to get missclassifier result. kNN is a simple classifier, which uses the distance measurement techniques that widely used in pattern recognition. The main disadvantage of the kNN classifier is that it is a lazy learner, i.e. it does not learn anything from the training data and simply uses the training data itself for classification. In this paper, we have used kNN classifier to check the class labels for the misclassified instances and removing vagueness between the contradictory rules. adaptive rule-based (ARB) Classifier: The proposed ARB classifier combines the random subspace and boosting approaches with ensemble of DTs to construct a set of classification rules. Random subspace method (or attribute bagging) is an ensemble classifier that consists of several classifiers each operating in a subspace of the original feature space, and outputs the class based on the outputs of these individual classifiers.

5) Misclassified Classifier:

Most of the probability based ML algorithms like DT and NB classifier suffer from the overfitting problem, because of the redundant instances in training data. To check the classes of misclassified instances we used the kNN classifier with feature selection and weighting approach. We applied DT induction for feature selection and weighting approach. To analyse the misclassified instances, firstly we built a tree, DT from the misclassified instances.

B. Algorithm

Algorithm 1: J48 Algorithm

J48 classifier is the classification algorithm used for detecting the novel and multi novel class. For the problem to the classification the methodology of decision tree is used. For modeling the classification process tree is build. While the tree is generated it is connected with each column of the database and results in classification for that column.

1) Input: Training data
2) Output: Decision Tree
3) DSTBUILD (∗DS)
4) {
5) DT = φ;
6) DT=Generate root node and label with splitting attribute;
7) DT=Add arch to root node for each splitting predicate and label;
8) DS=By applying split predicate to DS database is created;
9) If stopping point reached for this path, then;
10) DTr = generate leaf node and label with the appropriate class;
11) DTr = DSTBUILD(∗DS);
12) Else
13) DTr = DSTBUILD (DS);
14) DT = add DTr to arc;
15) }

The J48 classifier for establish the tree does not require any code. While constructing a tree, J48 rejects the missing qualities i.e. the quality for those things can be anticipated focused around which is the thought about characteristics qualities for the other record.

System Analysis

A. Experimental Setup

The system is built using Java framework on Windows platform. The Net beans IDE is used as a development tool. The system doesn't require any specific hardware to run; any standard machine is capable of running the application.

B. Expected Result

In this section discussed the experimental result of the proposed system.

In table 1 shows the accuracy of proposed and existing system for the training dataset.

Table I Accuracy Comparison For Training Dataset

System	Accuracy
Accuracy of proposed system for training data	90%
Accuracy of existing system for training data	83%

Following figure 2 shows the comparison of proposed system and existing system on the basis of their accuracy for training dataset. From the graph it shows that accuracy of the proposed system is more than the accuracy of the existing system.

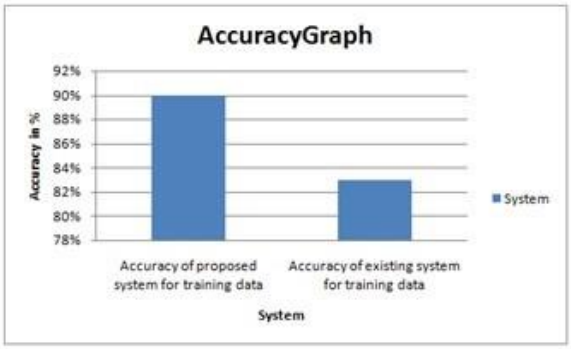

Fig. 2. Accuracy Graph for Training Dataset

In table 2 shows the accuracy of proposed and existing system for the testing dataset.

Table II Accuracy Comparison For Testing Dataset

System	Accuracy
Accuracy of proposed system for testing data	60%
Accuracy of existing system for testing data	55%

Following figure 3 shows the comparison of proposed system and existing system on the basis of their accuracy for testing dataset. From the graph it shows that accuracy of the proposed system is more than the accuracy of the existing system.

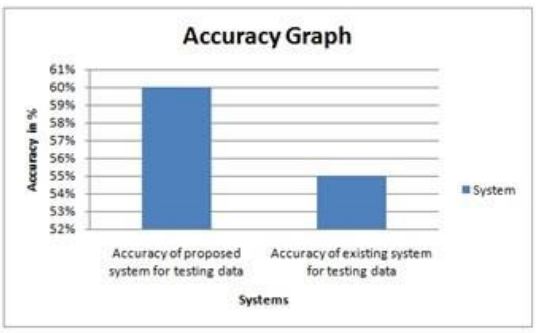

Fig. 3. Accuracy Graph for Testing Dataset

Conclusion

The proposed system is a rule based classifier, which is combination of random subspace and boosting approaches. To construct the classification rules without global optimization, system makes use of J48 Decision Tree and KNN algorithm. Random subspace is used to avoid over fitting problem, boosting approach is used for solving problem of noisy instances classification and finally J48 decision tree is deal with class imbalance problem. With J48 decision tree, rules are evaluated from training data and with KNN, misclassified instances are analyzed. The proposed approach will be compared with other rule-based and machine learning classifiers, and detailed results and discussion of the experiments are presented to demonstrate comparative analysis and the efficacy of the results. Results will prove the good prediction accuracy of classified DNA. Also save time and memory and enhanced the performance of an adaptive rule-based classifier method.

Acknowledgment

The authors would like to thank the researchers as well as publishers for making their resources available and teachers for their guidance. We are thankful to the authorities of Savitribai Phule University, Pune and concern members of cPGCON2017 conference, organized by, for their constant guidelines and support. We are also thankful to the reviewer for their valuable suggestions.

References

[1]. Dewan Md. Farid, Mohammad Abdullah Al-Mamun, Bernard Manderick, Ann Nowe, "An adaptive rule-based classifier for mining big biological data", Expert Systems With Applications 64 (2016) 305-316.
[2]. Al-Mamun, M. A., Farid, D. M., Ravenhill, L., Hossain, M. A. , Fall, C., Bass, R.(2016). An in silico model to demonstrate the effects of maspin on cancer cell dynamics. Journal of Theoretical Biology, 388 , 37-49.
[3]. Alter, M. D., Kharkar, R., Ramsey, K. E., Craig, D. W., Melmed, R. D., Grebe, T. A., et al. (2011). Autism and increased paternal age related changes in global levels of gene expression regulation. PLOS ONE, 6 (2), 1-10.
[4]. Farid, D. M., Rahman, C. M. (2013). Assigning weights to training instances increases classification accuracy. International Journal of Data Mining & Knowledge Management Process, 3 (1), 13-25.
[5]. Farid, D. M., Zhang, L., Hossain, A., Rahman, C. M., Strachan, R., Sexton, G., et al. (2013). An adaptive ensemble classifier for mining concept drifting data streams. Expert Systems with Applications, 40 (15), 5895-5906.
[6]. Farid, D. M., Zhang, L., Rahman, C. M., Hossain, M., & Strachan, R. (2014). Hybrid decision tree and nave bayes classifiers for multi-class classification tasks. Expert Systems with Applications, 41 (4), 1937-1946.
[7]. Gehrke, J., Ramakrishnan, R., & Ganti, V. (20 0 0). Rainforest - a framework for fast decision tree construction of large datasets. Data Mining and Knowledge Discovery, 4 (2-3), 127-162.
[8]. Jin, X., Wah, B. W., Cheng, X., & Wang, Y. (2015). Significance and challenges of big data research. Big Data Research, 2 (2), 59-64.