Paper: Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks

Reviewer: Md. Omar Shahariar

Date: 20-12-2020

Abstract: In this paper, two independent hybrid mining algorithms have been proposed to improve the classification accuracy rates of decision tree (DT) and naïve Bayes (NB) classifiers for the classification of multi-class problems. Both DT and NB classifiers are useful, efficient and commonly used for solving classification problems in data mining. Since the presence of noisy contradictory instances in the training set may cause the generated decision tree suffers from overfitting and its accuracy may decrease, in first proposed hybrid DT algorithm, used naïve Bayes (NB) classifier to remove the noisy troublesome instances from the training set before the DT induction. Moreover, it is extremely computationally expensive for a NB classifier to compute class conditional independence for a dataset with high dimensional attributes. Thus, in the second proposed hybrid NB classifier, DT induction to select a comparatively more important subset of attributes for the production of naïve assumption of class conditional independence. The experimental results indicate that the proposed methods have produced impressive results in the classification of real life challenging multi-class problems. They are also able to automatically extract the most valuable training datasets and identify the most effective attributes for the description of instances from noisy complex training databases with large dimensions of attributes.

Introduction: During the past decade, a sufficient number of data mining algorithms have been proposed by the computational intelligence researchers for solving real world classification and clustering problems. classification is a data mining function that describes and distinguishes data classes or concepts. The goal of classification is to accurately predict class labels of instances whose attribute values are known, but class values are unknown. Clustering is the task of grouping a set of instances in such a way that instances within a cluster have high similarities in comparison to one another, but are very dissimilar to instances in other clusters. It analyzes instances without consulting a known class label. This paper presents two independent hybrid algorithms for scaling up the classification accuracy of decision tree (DT) and naïve Bayes (NB) classifiers in multi-class classification problems. DT is a classification tool commonly used in data mining tasks such as ID3 (Quinlan, 1986), ID4 (Utgoff, 1989), ID5 (Utgoff, 1988), C4.5 (Quinlan, 1993), C5.0 (Bujlow, Riaz, & Pedersen, 2012), and CART (Breiman, Friedman, Stone, & Olshen, 1984). The goal of DT is to create a model that predicts the value of a target class for an unseen test instance based on several input features. In this paper, two hybrid algorithms have been proposed respectively for a DT classifier and a NB classifier for multi-class classification tasks. The first proposed hybrid DT algorithm finds the troublesome instances in the training data using a NB classifier and removes these instances

from the training set before constructing the learning tree for decision making. Otherwise, DT may suffer from overfitting due to the presence of such noisy instances and its accuracy may decrease. Moreover, it is also noted that to compute class conditional independence using a NB classifier is extremely computationally expensive for a dataset with many attributes. Second proposed hybrid NB algorithm finds the most crucial subset of attributes using a DT induction. The weights of the selected attributes by DT are also calculated. Then only these most important attributes selected by DT with their corresponding weights are employed for the calculation of the naïve assumption of class conditional independence. We evaluate the performances of the proposed hybrid algorithms against those of existing DT and NB classifiers using the classification accuracy, precision, sensitivity– specificity analysis, and 10-fold cross validation on 10 real benchmark datasets from UCI (University of California, Irvine) machine learning repository (Frank & Asuncion, 2010). The experimental results prove that the proposed methods have produced very promising results in the classification of real world challenging multi-class problems. These methods also allow us to automatically extract the most representative high quality training datasets and identify the most important attributes for the characterization of instances from a large amount of noisy training data with high dimensional attributes.

Reason Accepting this paper:

Figs. 1 and 2 respectively show the comparison of classification accuracy rates between basic NB and the proposed NB classifiers and between the C4.5 DT and the proposed DT classifier for each dataset with 10-fold cross validation on a selected data. Fig. 3 also shows the comparison of classification accuracy rates of all classifiers on 10 datasets with 10-fold cross validation. Overall, Algorithm 1 is able to automatically remove noisy instances from training datasets for DT generation to avoid overfitting. It thus possesses more robustness and generalization capabilities. Algorithm 2 is capable of identifying the most discriminative subset of attributes for classification. The evaluation results prove the efficiency of the proposed DT and NB algorithms (Algorithm 1 and 2) for the classification of challenging real benchmark datasets. They respectively outperform the traditional C4.5 DT and NB classifiers in all the test cases (see Figs. 2 and 1).

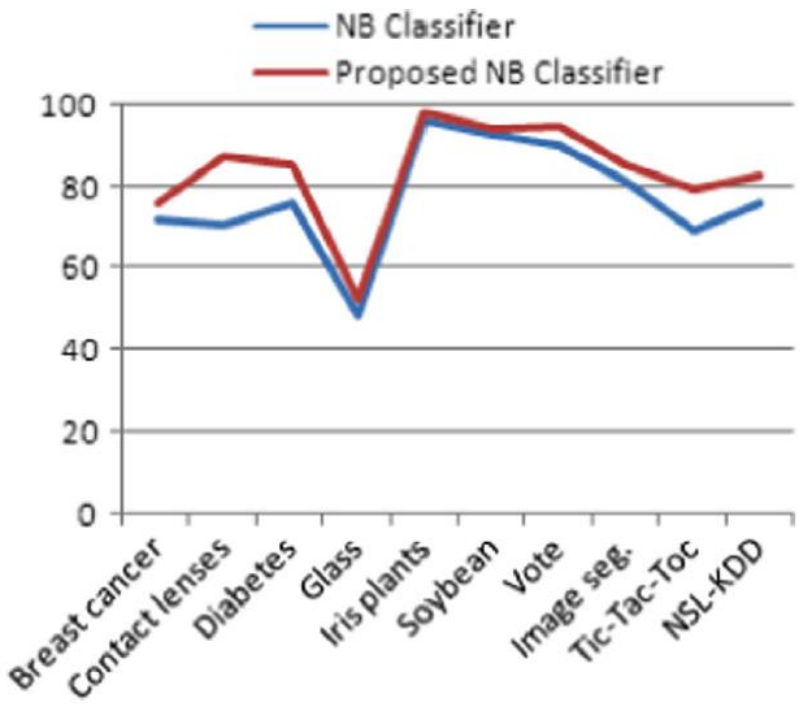

Fig. 1. The comparison of classification accuracy rates between the basic NB and the proposed hybrid NB classifiers on 10 datasets with 10-fold cross validation.

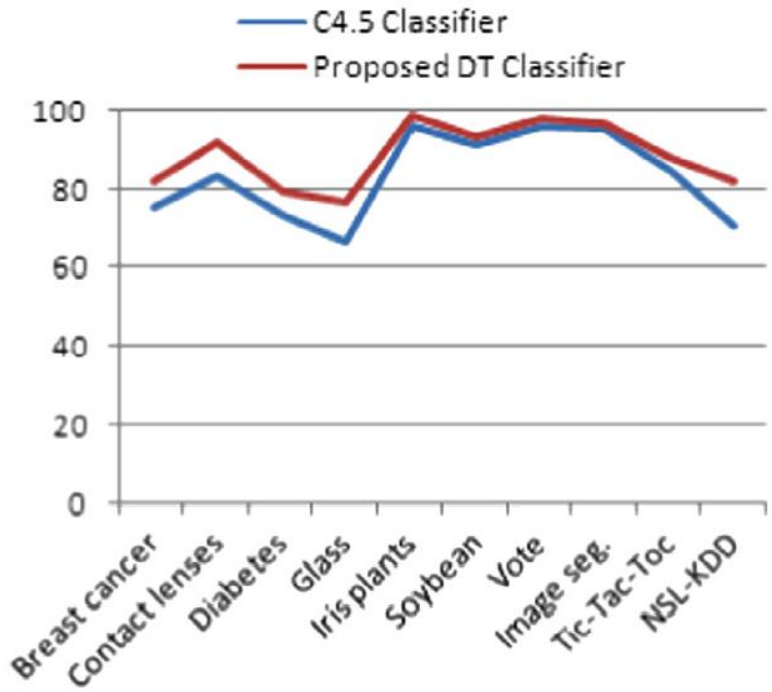

Fig. 2. The comparison of classification accuracy rates between the C4.5 DT and theproposed DT classifiers on 10 datasets with 10-fold cross validation

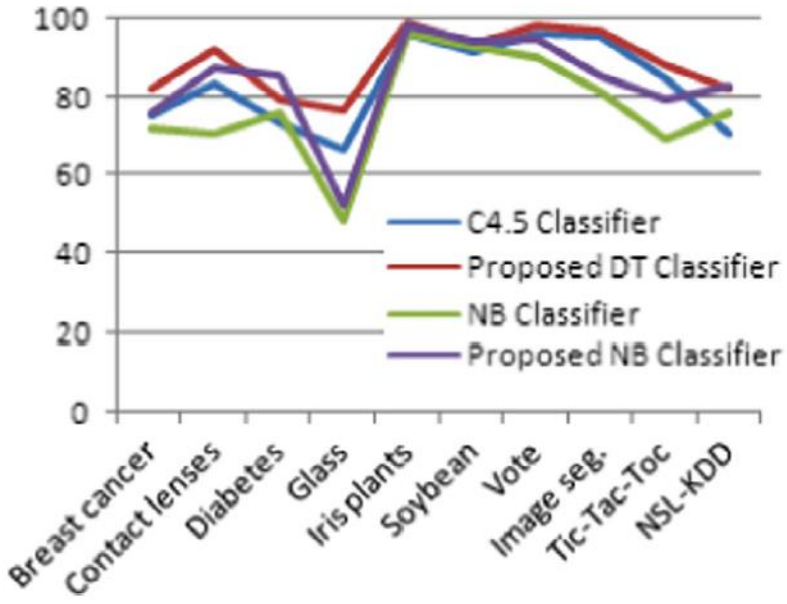

Fig. 3. The comparison of classification accuracy rates of all classifiers on each dataset with 10-fold cross validation.

Conclusions:

In this paper, they have proposed two independent hybrid algorithms for DT and NB classifiers. The proposed methods improved the classification accuracy rates of both DT and NB classifier in multi-class classification tasks. The first proposed hybrid DT algorithm used a NB classifier to remove the noisy troublesome instances from the training set before the DT induction, while the second proposed hybrid NB classifier used a DT induction to select a subset of attributes for the production of naïve assumption of class conditional independence. The performances of the proposed algorithms were tested against those of the traditional DT and NB classifiers using the classification accuracy, precision, sensitivity specificity analysis, and 10-fold cross validation on 10 real bench mark datasets from UCI machine learning repository. The experimental results showed that the proposed methods have produced impressive results for the classification of real life challenging multi-class problems. In future work, other classification algorithms, such as naïve Bayes tree (NBTree), genetic algorithms, rough set approaches and fuzzy logic, will be used to deal with real-time multi-class classification tasks under dynamic feature sets.