Classifying Very High-Dimensional Data with Random Forests Built from Small Subspaces

Baoxun Xu, Harbin Institute of Technology Shenzhen Graduate School, China
Joshua Zhexue Huang, Shenzhen Institutes of Advanced Technology and Chinese Academy of Sciences, China
Graham Williams, Shenzhen Institutes of Advanced Technology, and Chinese Academy of Sciences, China
Qiang Wang, Harbin Institute of Technology Shenzhen Graduate School, China
Yunming Ye, Harbin Institute of Technology Shenzhen Graduate School, China

ABSTRACT
The selection of feature subspaces for growing decision trees is a key step in building random forest models. However, the common approach using randomly sampling a few features in the subspace is not suitable for high dimensional data consisting of thousands of features, because such data often contains many features which are uninformative to classification, and the random sampling often doesn’t include informative features in the selected subspaces. Consequently, classification performance of the random forest model is significantly affected. In this paper, the authors propose an improved random forest method which uses a novel feature weighting method for subspace selection and therefore enhances classification performance over high-dimensional data. A series of experiments on 9 real life high dimensional datasets demonstrated that using a subspace size of $\left\lceil \log_2(M) + 1 \right\rceil$ features where $M$ is the total number of features in the dataset, our random forest model significantly outperforms existing random forest models.

Keywords: Classification, Decision Tree, High-Dimensional Data, Random Forests, Random Subspace

1. INTRODUCTION
One of the major current challenges to many classical classification algorithms is dealing with large high dimensional data. Examples include text data, microarray data and digital images which often have thousands of features and hundreds of thousands or millions of objects. Such very high dimensional data has two special characteristics that affect the performance of classification algorithms. One is that different
classes of objects are present in subspaces of the data dimensions. For example, in text data, documents relating to sport are categorized by the key words describing sport, while documents relating to music are represented by the key words describing music. The other characteristic is that a large number of dimensional features are uninformative to the class feature. That is, many features are only weakly correlated to the class feature, if at all and have a low power in predicting object classes (Saxena & Wang, 2010).

The random forest (Breiman, 2001) algorithm is a popular classification method used to build ensemble models of decision trees from subspaces of data. Experimental results have shown that random forest models can achieve high accuracy in classifying high dimensional data (Banfield et al., 2007). Interest in random forests has grown in many domains where high dimensional data is prominent, including domains such as bioinformatics (Pang et al., 2006; Diaz-Uriarte & De Andres, 2006; Chen & Liu, 2005; Bureau et al., 2005), medical data mining (Ward et al., 2006) and image classification (Bosch, Zisserman, & Muoz, 2007).

Several methods have been proposed to build random forest models from subspaces of data (Breiman, 2001; Ho, 1995, 1998; Dietterich, 2000). Among them, Breiman’s method (Breiman, 2001) has been popular due to its good performance compared to other methods (Banfield, Hall, Bowyer, & Kegelmeyer, 2007). Breiman uses a simple random sampling from all the available features to select subspaces when growing unpruned trees within the random forest model. Breiman suggested selecting \( \lceil \log_2(M) + 1 \rceil \) features in a subspace, where \( M \) is the total of independent features in data. This works well for data with certain dimensions (e.g., less than 100 features) but is not suitable for very high dimensional data consisting of thousands of features. In contrast, for very high dimensional data, Breiman’s subspace size of \( \lceil \log_2(M) + 1 \rceil \) is too small. Such data are dominated by uninformative features which have very low predictive power with respect to the target classification. Using a simple random sampling results in informative features not being included in subspaces (Amaratunga, Cabrera, & Lee, 2008). As a result, weak trees are created and classification performance of the random forest is significantly affected. To increase the chance of selecting informative features in subspaces, the subspace size has to be enlarged extensively. However, this increases the computational requirements of the algorithm and increases the likelihood of the resulting trees being correlated. Correlated trees reduce the classification performance of a random forest model (Breiman, 2001).

To address this problem Amaratunga (Amaratunga, Cabrera, & Lee, 2008) proposed a feature weighted method for subspace sampling. The weight of a feature is computed with respect to the correlation between the feature and the class. The weights are treated as the probability with which a feature is selected for inclusion in a subspace. Using this feature weighted method to sample subspaces there is a high chance that informative features are selected when growing trees for a random forest model. This method might be compared to the method of Adaboost (Freund & Schapire, 1996; Qiu, Wang, & Bi, 2008) which selects training samples according to the sample weights computed from the result of the previous classification. Such a method increases the probability for selecting informative features for inclusion in each subspace. This results in an increase in the average strength of the trees making up the random forest model, and thus the generalization error bound is reduced. Consequently, classification performance of the random forest model is increased.

Amaratunga’s method is only valid for two-class problems, using the t-test of variance analysis to calculate the feature weights. In this paper, we propose a feature weighting method for subspace selection to solve multi-class problems. Instead of the t-test we calculate the chi-square statistic or information gain ratio as the feature weights. Both measures capture the correlation between a feature and the class for multi-class problems. The larger the weight, the more informative the feature to the classification. We normalize the set of weights to
The Challenges of Data Cleansing with Data Warehouses
www.igi-global.com/chapter/the-challenges-of-data-cleansing-with-data-warehouses/157685?camid=4v1a

Influence of Domain and Model Properties on the Reliability Estimates’ Performance
www.igi-global.com/article/influence-domain-model-properties-reliability/37405?camid=4v1a