Talanta 189 (2018) 71–78

Contents lists available at ScienceDirect

Talanta

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

Uncharted forest: A technique for exploratory data analysis

Casey Kneale, Steven D. Brown*

Department of Chemistry and Biochemistry, University of Delaware, 163 The Green, Newark, DE 19716, USA

ARTICLE INFO

Keywords:
Exploratory data analysis
Random forest
Provenance
Clustering

ABSTRACT

Exploratory data analysis is crucial for developing and understanding classification models from high-dimensional datasets. We explore the utility of a new unsupervised tree ensemble called uncharted forest for visualizing class associations, sample-sample associations, class heterogeneity, and uninformative classes for provenance studies. The uncharted forest algorithm can be used to partition data using random selections of variables and metrics based on statistical spread. After each tree is grown, a tally of the samples that arrive at every terminal node is maintained. Those tallies are stored in single sample association matrix and a likelihood measure for each sample being partitioned with one another can be made. That matrix may be readily viewed as a heat map, and the probabilities can be quantified via new metrics that account for class or cluster membership. We display the advantages and limitations of using this technique by applying it to two classification datasets and three two provenance study datasets. Two of the metrics presented in this paper are also compared with widely used metrics from two algorithms that have variance-based clustering mechanisms.

1. Introduction

Chemometric classification methods are often used to discriminate high-dimensional chemical signatures of unknown samples to determine their most likely class label [1]. These methods have proven to be valuable for the fields of archaeometry and forensics where the origin, or provenance, of a manufactured item is often of interest [1–6] because the chemical data, especially those obtained from neutron activation analysis (NAA) [7,8], tend to be highly multivariate. The high dimensionality of data poses a challenge for understanding sample relationships because it cannot be easily visualized or interpreted directly. While, these methods can succeed at classifying samples, they do so by including information from the class labels into the model, and they provide little insight about trends or patterns found in the data with the absence of labels. Information pertaining to which samples are most similar to one another between classes, or which are different from the rest of the samples within a given source class, without the influence of label discrimination is not available from classification methods.

Exploratory methods are often used to reveal trends and other patterns hidden in data without the use of class labels [1]. Exploratory data analysis (EDA) is a form of data analysis which encompasses a number of visual exploration methods. Success in an EDA study depends on the creativity of the analyst as much as on the technique. Although there is no strict definition of EDA, it has been stated that, “Exploratory data analysis isolates patterns and features of the data and reveals these forcefully to the analyst” [9], and also that “Exploratory data analysis' is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe to be there” [10]. The coupling of EDA methods with preprocessing and/or feature selection can aid an analyst in the discovery of patterns, or the lack thereof, in data. The information gained from EDA may be used with feature selection, preprocessing and modeling methods to iteratively improve a data analysis pipeline.

The two predominant types of EDA focus on dimension reduction and clustering. These methods have been extensively used for the EDA of archaeometric data [3–5,11] because high dimensional data are not readily visualized. Dimension reduction techniques such as principal component analysis (PCA), discriminant analysis, or exploratory projection pursuit are used to change the basis of the data to one based on lower-dimensional projections [1]. Unfortunately, these methods do not offer a guarantee that the resulting 2-D or 3-D projections will offer meaningful information about class relationships or sample-class relationships, due to the presence of class overlap or class heterogeneity [1]. Unsupervised classification as implemented in clustering algorithms suffers from the opposite problem clustering methods readily create groupings of high-dimensional data but tend to offer little information about the cluster assignments themselves. Clustering methods have been known to incorrectly associate archeological samples depending on the method used and on the chemistry of the samples [11]. Information about sample or class relationships may be inferred by employing many techniques, but the resulting information is often hidden, and typically key relationships in the data cannot be seen by an analyst [11].

* Corresponding author.
E-mail address: sdb@udel.edu (S.D. Brown).

https://doi.org/10.1016/j.talanta.2018.06.061
Received 21 April 2018; Received in revised form 18 June 2018; Accepted 19 June 2018
Available online 20 June 2018
0039-9140/

This work reports a new approach, which we call the uncharted forest, to the visualization and measuring of relationships within and between classes of data. This approach has elements of a clustering algorithm in that it groups similar samples with one another, but is also similar to dimension reduction methods in that it outputs a single heat map which can be interpreted to reveal information about the samples. Uncharted forest analysis uses a partitioning method that is related to the sample partitioning approach used in decision trees but, it does not use class labels like most tree methods do [12]. Instead, the uncharted forest analysis explores how samples relate to one another under the context of univariate variance partitions. Although the method is unsupervised, we show that when the results are overlaid with external class labels, the method can be used to investigate sample relationships as they pertain to class labels. We demonstrate that this technique can be used as a tool for exploratory data analysis to visualize class or cluster associations, sample-sample associations, class heterogeneity, and uninformative classes. The utility of uncharted forest analysis is demonstrated on two classification datasets and two provenance datasets. Additionally, two empirical clustering metrics are compared with two of the metrics obtained by uncharted forest analysis on another provenance dataset.

2. Theory

To motivate the use of an unsupervised tree ensemble, a brief review of supervised trees and bagged tree ensembles as used in classification is provided here. A complete review of these methods can be found elsewhere [12,13].

2.1. Supervised classification trees

Before describing the mechanisms by which supervised decision trees are used to partition data, a survey of the vocabulary common to these methods is presented. Classification trees are a type of supervised classifier, which means that the assignment of new data to a class label requires that each sample the tree was trained on also has a label. The aim of developing a classification tree is the establishment of a set of sequential rules that can be applied to label a sample based on its features. Supervised decision trees are collections of many binary decisions, where each decision is made at one of three locations: roots, branches, or terminal nodes/leaves. These three locations are displayed in their hierarchical ordering in Fig. 1. A root is the first decision made in the tree. Branch nodes indicate later, non-terminal decisions. Terminal nodes indicate where a branch has been terminated and where final decisions that assign class labels are made.

Every sample from the training set is partitioned at a branch or node based on whether a selected variable for the sample has a value that is greater than or less than a specified value. The combination of a variable and value that partitions a set of samples is often referred to as a decision boundary. Decision boundaries are obtained by exhaustively searching each variable and finding the threshold value which affords the highest gain [12]. Here, gain is a metric-dependent measure of how well a decision separates the available samples according to their class labels. Metrics that can be used to minimize class label impurities at terminal nodes, such as the Gini impurity, informational entropy, and classification accuracy are commonly used to train supervised classification trees by finding a set of decision boundaries [12].

For example, if the samples available at a branch node are better separated into their respective classes by application of some decision, two new branches are made at that node. If not, the partitioning terminates, the node becomes a terminal node, and the decision tree ceases growth using those samples. In this way, each observation is sorted from the root of a tree through branches and eventually to terminal nodes, based on whether creating a new decision boundary better separates samples by known class labels [12].

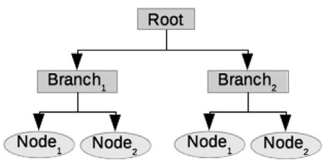

Fig. 1. A block diagram showing the relationships between roots, branches and terminal nodes via arrows.

One advantage to the use of decision trees is that the classifier that results can be easily understood by examining the decision boundaries that partition the data and the relation of these boundaries to class labels. However, the predictive performance of classification trees on new samples is generally poor relative to other models because the partitions are sensitive to noise in the variables [14]. Tree-based classifiers also tend to overfit the data, and some sort of branch removal, or pruning step, is usually employed to reduce the complexity of the tree [12].

2.2. Random forest classification

The random forest classifier is a multiclass classification technique that utilizes the results from an ensemble, or collection of supervised decision trees [13] to assign class labels. In this technique, a user-specified number of supervised decision trees is created, each overfit to a selection of bootstrapped samples and random selections of variables. Random forests make use of the complexity of many, biased decision trees to make robust classification assignments by using the average class designation assigned by every tree to predict overall class membership for new data. This approach to ensemble model building is referred to as bootstrap aggregating, or bagging [15]. The average of many, complex trees tends to perform well in practice [13].

Random forest classification has some distinct advantages relative to many other classifiers. It is robust to outliers because sample selection at branches is bootstrapped, and the effect of a few outlying samples does not affect its cost function significantly [13]. Random forest models are also robust to noise in variables because many classifiers, each with randomly selected variables, are used to decide on class memberships. Another notable advantage to random forest classification is that there is no requirement for linearity in class boundaries. Random forest models, unlike methods such as linear discriminant analysis or support vector machines, are not based on the hypothesis that classes are linearly separable [16,17]. Random forest classification also requires relatively few hyperparameters and works without the need for significant tuning of the classifier on many kinds of data [14].

2.3. Uncharted forest tree

The concept of unsupervised decision tree modeling is relatively new to chemometric applications. Only a few approaches using unsupervised trees have been reported. An approach reported by Khudanpur et al. is based on an algorithm similar to the one reported here, but their approach uses the Kullback-Leibler distance to find sets of unsupervised partitions that are optimal with respect to an information theoretic measure [18]. Other methods for developing unsupervised decision trees are based on the use of an assumed class label for each observation; the trees are trained in a manner that is similar to that used for training a random forest classifier [19,20].

Our approach to unsupervised decision trees, the uncharted forest tree algorithm, focuses on intuitive concepts from statistics or machine learning rather than on information theory, to allow for easy interpretation. The trees can be created without the need for class labels because the gain function that is optimized in the construction of each tree relies only on information in the data matrix, not on the labels. This algorithm does not utilize the usual supervised metrics for optimization such as the Gini importance or entropy [13]. Instead, the tree hierarchies used in uncharted forest are constructed from decision boundaries that reduce measures of spread in a given variable. The reduction of spread is common in pattern recognition and chemometrics; it is the