Overview of Random Forest : Effective Ensemble Method in Modern Data Mining

S.P. Rahayu, [1] A. Embong, [2]
[1] Department of Statistics, Faculty of Mathematics and Natural Sciences ,
Insititut Teknologi Sepuluh Nopember
Kampus Keputih Sukolilo Surabaya Indonesia
[2] Faculty of Computer Science & Information Technology,
Universiti Malaysia Pahang,
Karung Berkunci 12, 25000 Kuantan, Pahang

Abstract — Random Forest (RF) is one of the most effective ensemble methods that has been proposed in recent years. This ensemble method has been proven to improve model performance in data mining task. In this paper, we describe the basic algorithm, characteristics, empirical comparison result and applications in real problem of RF.

INTRODUCTION

Many traditional machine learning in data mining methods generate a single model such as tree method or neural network. Ensemble data mining methods, also known as Commite Methods, generate multiple model. Basic goal when designing an ensemble is to increase the power of multiple models to achieve better model performance (low bias and variance) than any of the single models could on their own [16]. Model performance demonstrated by ensembles data mining methods is usually comparable or even better that can be achieved by the best single sophisticated models in many application.

Various ensembles data mining method have been proposed in recent years. The most widely used ensemble methods are boosting and bootstrap aggregating (bagging). Boosting is based on sample re-weighting but bagging uses bootstrapping [8]. Bootstrapping is a statistical technique for replicating a data set with a distribution approximating the original. A bootstrap data set is formed by drawing samples randomly from the original data set with replacement. RF uses bagging to form an ensemble of classification and regresion tree. Bootstrap samples are drawn to construct multiple trees and each tree is grown with a randomized subset of predictors, hence the name "random" forest. The trees are grown to maximum size without pruning and aggregation is by averaging or majority voting the trees [1], [2].

BASIC ALGORITHM

The RF algorithm (for both classification and regression) is as follows: [2],[4]

(1) Draw *ntree,* the number of bootstrap sample or the number of trees to grow , from original data
(2) For each of bootstrap sample, grow unpruned classification or regression tree, with the following modification : at each node, rather than choosing the best split among all predictors, randomly sample *mtry* of the predictor and choose the best split from among those variables. (Bagging can be thought as special case of RF, obtained when *mtry* = p, the number of predictors)
(3) Predict new data by aggregating the predictions of the *ntree*. (i.e. majorite votes for classification, average for regression)

CHARACTERISTICS

RF has several chacteristics that make it effective : [2], [9], [12], [14]

a) Can be used when there are many more variables than observations
b) Runs efficiently on high dimensional data set
c) Can be used in supervised learning task (classification and regression), unsupervised learning (clustering) and give useful views of data.
d) Can handle a mixture of categorical and continous predictor variables
e) Incorporated interactions among predictor variables
f) Can be used both for two-class and multi-class for classification tasks
g) Can balance error in unbalanced class population data set
h) It does not over fit in regression task
i) It produces highly accurate results in classification task
j) Relatively robust to outlier and noise
k) Can give useful internal ('out of bag') estimates of error, strength and variable importance

(generates information about the relation between the variables and the classification)

l) There is little need to fine–tune parameters to achieve effective performance

m) Generated forest can be saved for future use on other data

EMPIRICAL COMPARISON RESULT

Multiple recent empirical studies demonstrate RF to be competitive with many other machine learning algorithm : [2], [9], [12], [17]

a. RF predictive performance is as good as boosting and sometimes better (especially on noisy data sets).
b. RF is faster than many other ensembles, bagging and boosting in particular.
c. RF comparable in accuracy to Support Vector Machine, Neural Network (NN) and kernel Nearest Neighbour (k-NN).

APPLICATION OF RANDOM FOREST

RF method has a proven record of many successful application domain in data mining tasks. Two examples are classification & regression and clustering tasks:
Examples of application in classification and regression tasks are :

a. Bioinformatics
Uriarte and Andrez (2006) investegated the use of RF for classification of microarray data (including multiclass problems) and propose a new method of gene selection in classification problems based on RF [9].

b. Cheminformatics
Svetnik et al (2003) investegated the use of RF for predicting a compound's quantitative or categorical biological activity based on quantative description of the compund's molecular structure [11].

c. Neuroscience
Oh et al (2003) describe briefly an application of RF to a neuronal ensemble data set. RF was used to make predictions about wether single trials in the task were associated with correct or error responses [6].

d. Biometrics (Ecology)
Prasad et al (2006) evaluated the use of RF predictive vegetaion mapping under current and future climate scenarios according to he Canadian Climate Centre global circulation model [3].

e. Modern Physics
Kiesling and Zimmerman (2004) described application of RF in high-energy and astro-physics analysis [7].

f. Multisource Remote Sensing and Geograpic
Gislason et al (2006) investegated the use of RF for land cover classification and compared the accuracy of RF to other better–known ensemble methods [8].

g. Landscape epidemiology
Furlanello et al (2003) studied the use of the RF predictor and developing a spatial model of the probability of tick presence, given enviromental biotic and abiotic input variabel [13] .

h. Economy
Figini (2006) proposed a non parametric approach on based on Random Survival Forest in the field of credit risk measurement and compared its performance with a standard logit model [15].

i. Medical
Ishwaran et al (2004) proposed Relative Risk Forest, a novel method that combined RF methodology with survival trees grown using Poisson likelihoods, as a predictor of mortality from heart disease [5].

In clustering task, Shi et al (2005) describe Random Forest Clustering for tumor profiling based on tissue micro array data.
The random forest clustering procedure is carried out as follows. The random forrest dissimilariy (one major input of a clustering analysis is the dissimilarity measure) is use to represent each observation (patient) as a point in two-dimensional space with the aid of multidimensional scalling. The distance between the points are used in partitioning around medoids clustering. The number of clusters is chosen by using the partitioning around medoids silhouette plots and inspecting corresponding multidimensional scaling plots [10].

CONCLUSION

In this paper, we have described briefly about the use of Random Forest, one of the most effective ensemble data mining method. Multiple recent empirical studies have demonstrated Random Forest to be competitive in model performance on the level boosting, Support Vector Machine, NN and k-NN. Beside that, Random Forest method has a proven record of many succesful application domain in data mining tasks.

REFFERENCES

[1] L. Breiman, "Bagging Predictor," Machine Learning, vol. 24, pp. 123-140, 1996.
[2] L. Breiman, "Random Forest," Machine Learning, vol. 45, pp. 5-32, 2001.

[3] A.M. Prasad et al, “Newer Calssification and Regression Tree Techniques : Bagging and Random Forest for Ecological Prediction,” Ecosystem, vol.9, pp.181-199, 2006.
[4] A. Liaw and M. Wiener,” Classification and Regression by Random Forest,” The Newsletter of the R Project, vol. 2, pp. 18-22, December 2002.
[5] H. Ishwaran et al, ”Relative Risk Forest for Exercise Heart Rate RECOVERY AS A Predictor of Mortality,” Journal of the American Statistical Association, vol. 99, pp.591-599, September 2004
[6] J.Oh, “Estimating Neuronal Variable Importance with Random Forest,” IEEE, pp. 33-34, 2003.
[7] Kiesling and J. Zimmerman,”Statistical Learning Method in High- Energy- and Astrophysics,” IEEE, pp. 325-329, 2004.
[8] P.O. Gislason et al,”Random Forest for Land Cover Classification,” Pattern Recognition Letter, vol 27, pp. 294-300, 2006.
[9] R.D.-Uriarte and S.A. Andres,”Gene Selection and Classification of Microarray Data using Random Forest,” BMC Bioinformatics, vol. 7, pp.1-13, January 2006.
[10] T. Shi et al, ”Tumor Classification by Tissue Microarray Profiling: Random Forest Clustering applied to renal cell carcinoma,” Modern Patology, vol.18, pp. 547-557, 29 October 2004.
[11] V. Svetnik et al, ”Random Forest: A Classification and Regression Tool for Compound Classification and Qsar Modelling,” J. Chem. Inf. Comput. Sci., vol. 43, pp. 1947-1958, 2003.
[12] Y. Truong et al, “Learning a Complex Metabolomics Dataset using Random Forest and Support Vector Machine”, presented at the KDD’04 Conference, Washington, USA, August 22-25, 2004.
[13] C. Furlanello et al,”GIS and the Random Forest Predictor: Integration in R for Tick-Borne Disease Risk assessment,” in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, 2003, pp. 1~10.
[14] J. Chia et al,”Derivation of a Perennial vegetaion Density map for The Australian Continent,” CSIRO Mathematical and Information Science, Floreat, Australia, 2006.
[15] D.F.S.Figini, ”Random Survival Forest models for SME Credit Risk Measurement,” 2006
[16] N.C. Oza,”Ensemble Data Mining Methods,” NASA Ames Research Center, USA.
[17] T. Hastie et al, *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer-Verlag, 2001.