Report about VOCs Dataset's Analysis based on randomForests Method

Zhang Huaizhong[1] Fred Hamprecht[2] Anton Amann[3]
1 Mathematics and Computer Science School,Nanjing Normal Uni., 210097 Nanjing, China
2 IWR, Heidelberg Uni., Germany
3 Universitatsklinik Anasthesie, Leopold-Franzens-Universitat, Austria
lotus_zhz@hotmail.com

Abstract

Volatile organic compounds (VOCs) play an important role in diagnosis and therapy of various diseases. We compare several main classifiers for data classification and point out the advantages of randomForests on supervising learning. So, in this project, we take the randomForests approach to analyze and appraise the VOCs data originally coming from the medical test. According to actual situation, combining the unsupervising and supervising methods, the important components and outliers are given. The evaluation for the classifying results has been acquired due to the cross-validation sampling methods.

1. The medical background

The medical experiments show the human breath contains a variety of endogenous volatile organic compounds (VOCs), the most abundant ones being acetone, methanol, ethanol, propanol and isoprene [1]. Now, more and more researches know about the origin and pathophysiological importance of these VOCs. The researchers can apply some devices, such as Proton-transfer-mass spectroscopy (PTR-MS) [2], to detect the components of VOCS online and rapid according to different patients. So, these online detections provide the evidences and valuable information to prove the connection between the VOCs and some diseases (as arrhythmia, cholesterol, diabetes, heart attack, high blood pressure, etc). For example, in a case study [1], breath isoprene reductions during lipid-lowering therapy (36%, be proportional to the whole components) were shown to correlate with cholesterol (32%) and LDL concentrations (35%) in blood ($p<0.001$) over a period of 15 days.Therefore, isoprene concentrations in human breath (measured by PTR-MS) might serve as an additional parameter to complement invasive tests for controlling lipid-lowering therapy and may be a useful parameter for lipid disorder screening. So, such VOCs might, in principle, be used for the screening, diagnosis and therapy control of various diseases [1]. The aim of our researching work is to provide the pattern analysis of VOCs dataset so that we can get some evidences for corresponding medical processing. The main results will include the important compounds analysis, the disease predicting and accuracy, the special situation detecting (outliers), and so on. The VOCs dataset comes from the University Clinic, Innsbruck and the compounds are instead of the variable names.

2. RandomForest method and the typical classifiers

2.1. The basic idea of statistical decision theory

The pattern analysis of VOCs dataset is based on the statistical decision theory. Nowadays, statistical learning plays a key role in many areas of science, finance and industry. How to learn from data is the main work in the field of statistics learning, data mining and artificial intelligence [4]. Our VOCs data analysis is mainly based on randomForests, one of the statistics decision methods. The randomForest method stands for the newest development of this field. The following is the two main aspects of data analysis: classification and clustering that is supervised and unsupervised learning methods. Our project is about the supervised leaning method, classification.

Bayes Rules is the foundation of statistical decision theory. There are some relevant terminologies and concepts. We can assume observations are independently and identically distributed (i.i.d) from an unknown multivariate distribution.

The class k prior, or proportion of objects of class k in the population, is denoted as $\Pi k=p(Y=k)$. Class k conditional density $pk(x)=p(x|Y=k)$. If we can know both Πk and $pk(x)$, we will solve the problem with Bayes rule. Namely, the Bayes rule predicts the class of an observation x by maximizing the posterior probability, argmaxk p(k|x), or minimizes the total risk under a symmetric loss function-Bayes risk.

Many classifiers can be viewed as of this general rule with particular parametric or non-parametric estimates of p(k|x) (the posterior probability). There are two general paradigms, one is direct function estimation approach, such as CART [5], and the other is density estimation approach, such as ML discriminant rule. RandomForest method belongs to the first class.

2.2. The main classifiers

Classifiers are built from past experience, i.e., from observations that are known to belong to certain classes. Such observations comprise the learning (training) set, L={(x1,y1),⋯,(xn,yn)}. So, classifier is applied to predict class for an observation x ∈ A(observation set) and get a estimating value =k. One important thing is that the random nature of the learning set L implies that the prediction is also random for a fixed value of x. The followings are some of the classifiers [5][7].

(1) Linear and quadratic discriminant analysis (LQDA)

LQDA methods were largely developed in the early times of statistics, but now we still study and use them in practical applications. It's main idea comes from Bayes rules or ML discriminant rules as the class features have Gaussian distributions. Thus, the basic idea can be explained as following: the predicted class for an observation x is the class with the closest mean vector, for a suitably defined distance function, using the Mahalanobis metric. LDDA is simple and easy to implement, and it has good performance in practice.

(2) Logistic discrimination method

This method arises from the principle of Bayes rule, and gets the model to estimate the class posterior probabilities with the multiple logit function. Logistic discrimination provides a more direct way of estimating posterior probabilities and is easier to generalize than classical linear discriminant analysis, e.g. neural networks.

(3) Nearest neighbor classifiers

They are based on a measure of distance between observations, such as the Euclidean distance. These classifiers were initially proposed as consistent non-parametric estimates of ML discriminant rules. The proportions of neighbors in each class are then used in place of the corresponding class conditional densities in the ML discriminant rule. There are several extensions of nearest neighbor rules, such as class priors, distance weights, feature selection, etc.

(4) Binary tree structured classifiers

These methods are constructed by repeated splits of subsets (nodes) of the measurement space into descendant subsets. Each terminal subset is assigned a class label and the resulting partition of the measurement space corresponds to the classifier. Different tree classifiers use different approaches to deal with the classification issues. In these models, CART is the typical classifier and is very popular [Breiman et al, 1984]. RandomForest classifier is based on the idea of CART [3].

(5) SVM classifier

The main idea is to maximize the margin, i.e., the sum of the distances from the hyperplane to the closest positive and negative correctly classified samples, while penalizing for the number of misclassifications, so that one can find the best hyperplane separating the two classes in the learning set (for binary classification, -1 vs. 1). The support vectors are those samples that determine the margin. According to the practical need, we can search for the hyperplane in the original space, linear SVMs, or in a higher-dimensional space, non-linear SVMs. SVMs are designed for binary outcomes. It can be generalized to multiclass problems by solving several binary problems simultaneously. SVM has good accuracy and is less prone to the overfitting problem.

2.3. RandomForest method and advantages

Random forests (Breiman, 1996, 1998) [3] are a combination of tree predictors such that each tree (which grows to maximum size and does not prune using CART methodology [5]) depends on the values of a random vector sampled independently with the same distribution for all trees in the forest. So, randomForests method is to grow an ensemble of trees and let them vote for the most popular class. In random forests methods, randomness is run through the procedure of algorithm. Firstly, in order to grow these ensemble trees, features are selected randomly that govern the growth of each tree in the ensemble. Secondly, at each node, the split is selected at random from among the best splits. Last, random training set, bagging is used in tandem with random feature selection. Each new training set is drawn, randomly, with replacement, from a bootstrap sample of the original training set. Then a tree is grown on the new training set using random feature selection. The trees grown are notpruned. Concretely, given a specific training set T, form bootstrap training sets Tk, construct classifiers h(x, Tk) and let these vote to form the bagged predictor. Then, for each y,x in the training set, aggregate the votes only over those classifiers for which Tk does not containing y, x (out of bag, OOB) [3].

Exclusion for classification, we can acquire the several important evidences from randomForest classifier.

(1) Test error rate can be estimated by built-in cross validation via the use of OOB samples.

(2) Each variable in OOB samples is randomly permuted and impact on prediction is measured, so the variables with high impact are deemed to be important.

(3) We can get the proximity matrix which measures how often a pair of points landed in the same terminal node, and it is useful for detecting outlier, clustering, missing value replacement, low dimensional projections.

Comparing with other classifiers, randomForests method has following advantages [6].

(4) It is fast algorithm (can be faster than growing/pruning a single tree) and easily parallelized.

(5) Because of taking an ensemble of unpruned trees, it can reduce variance and get low estimating bias, so it has good accuracy without over-fitting, comparable to SVM and Adaboost.

(6) It can handle high dimensional data without much problem.

(7) It can give internal unbiased estimate of test set error as trees are added to ensemble.

3 Analysis on VOCs dataset

The learning set includes 114 VOCs samples, and each sample has 65 VOC features. They come from 5 different classes, i.e., S2, W, L2, M, M_k

3.1. The basic classification results

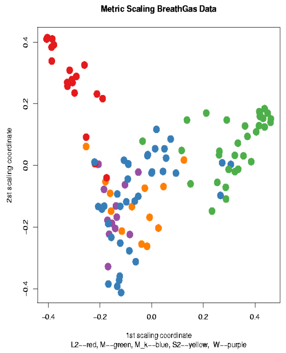
As in this heading, they should be Times 11-point boldface, initially capitalized, flush left, with one blank line before, and one after.

3.1.1. RF classifying situation headings. According to Breiman's suggestion, we select the number of variables randomly sampled as candidates at each split, i.e., the parameter mtry=8, and set the number of trees, i.e., ntree=5000. We find the misclassification error rate is about 23~27% in an RF run. Because of the randomness, the result of each prediction is different from the previous. The following is the confusion matrix for one run.From the table, S2, W classes are very confusing and most of the samples can't be distinguished. So, the estimate error is mainly caused from these two classes. But, L2, M, M_k are classified clearly, and there is a little misclassifying error. In section 3.4, we will give these 3 classes for classification and cross-validation.

	PREDICTION					
OBSERVATION		L	M	M k	S	W
	L2	1	0	0	0	0
	M	0	32	3	2	0
	M k	0	1	31	0	0
	S2	0	8	0	5	0
	W	0	12	1	0	0

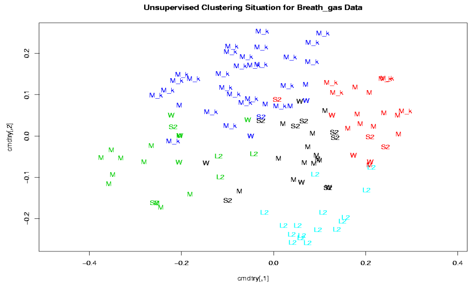
Table 1

3.1.2. Metric scaling graph. The proximity matrix is the by-product from the randomForests. The proximities between samples can provide the neighboring situation of samples. The graph 1 is drawed from the two scaling coordinates according to supervised learning. The 2-dimensional plots of the first scaling coordinate vs. the second often gives useful information about the data. We can directly view the relation among the different classes. As to the graph, L2, M, M_k samples are relatively concentrated, but S2, W samples are scattered around. So, in the following analysis, we will discard the S2, W samples, and pay more attention to the L2, M, M_k samples in order to get better accuracy.



Graph 1

3.1.3. Clustering situation-unsupervising learning results. We use the unsupervised learning method [10] to cluster the raw VOCs data set so that it can provide some useful information to further analysis. Labels indicate true class, colors indicate class membership according to unsupervised clustering. From the graph, L2, M_k are clustered very well, but the other classes are mixing up.



Graph 2

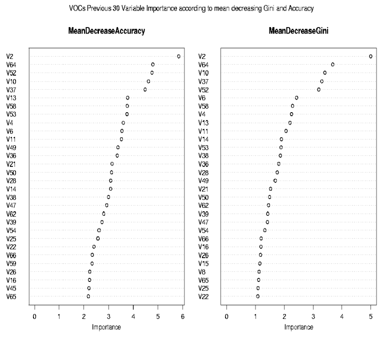
3.2. The most important VOCs components

In randomForests, by computing the variable's magin, i.e., the proportion of votes for its true class minus the maximum of the proportion of votes for each of the other classes at the end of run, we can get the variable's importance. Then, the value of importance of the mth variable is the average margin across all samples when the mth variable is randomly permuted.

In VOCs samples, the components' importances (the 30 most important components) are presented in graph 3.

The five most important components are V2,V64,V10,V37,V52, corresponding to the features 1, 63, 9, 36, 51 because V1 is the label in the training data set. The most important component in VOCs is component 1.

The next graph 4 shows the range of each component's Gini index value during 20s runs and substantiates the above analysing result.



Graph 3

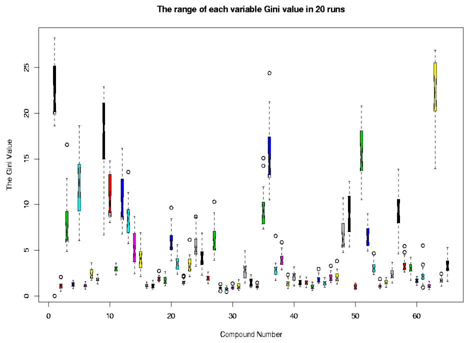
3.3. The special samples (outliers)

In randomForests method, outliers are defined as samples having small proximities to all other samples. The outlyingness is defined only with respect to other data in the same class as the given sample because some classes may be more spread out than others. So, we can find the outliers through proximity matrix. The graph 5 is the outlyingness measure of 114 samples.

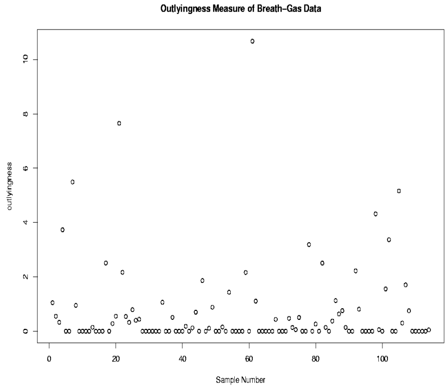
Generally, if 2 is used as a cutoff point, there are 13 samples that are to be suspected as outlying samples.

There are 2 samples from S2 class, 3 from class, 0 from L2, 4 from M class, and 4 from M_k class.

If a value above 4 is reason to suspect the sample of being outlying, then there are 5 samples, 7, 21, 61, 98, 105, as outliers. They are 1 from S2, 1 from W, 0 from L2, 1 from M, 2 from M_k respectively.



Graph 4



Graph 5

3.4. Evaluate the classifying performance of subset L2, M, M_k

In the previous analysis, we find the L2, M, M_k classes having good performance in classifying procedure. Now, we give the classifying situation and cross-validation (CV) result about these 3 classes data set. We use the bootstrap CV method, that is, the test samples comes from the original data. 10 samples are selected randomly for testing in each run. We have run the classifying procedure for 20 times totally.

Graph 6 shows the CV total error, each class error range.

From the graph, randomForests brings about 11~16 percent OOB error, and each class gets the coresponding error value. The classificatione error is pretty much the same and each of them contributes to the OOB error almost equally.

3.5. Evaluate the classifying performance of subset S2, M, W

In the VOCs data, the S2, M, W members are badly confusing. Now, we give the cv result of randomForests about these subdata.

In the graph 7, we can know the general OOB error is about 40 percent. It is quite high. But, the M class classifying error is much lower than the other two. It is impossible to distinguish the classes S2, W. The situation is the same as the mixture of five classes

4. Conclusion

Summing up, we can get some conclusions about VOCs with randomForests classifying method.

(1) To classify the data with RF, the error is about 23%. The main part of error comes from S2, W. The classes L2, M, M_k can be classified quite well.

(2) In the all components of VOCs, the components 1, 63, 9,36,51 are the most important components.

(3) The samples 7, 21, 61, 98, 105 are most suspected to be as outliers according to the analysis. In general, the most number of outlier comes from M, M_k, W.

(4) If we discard S2, W from VOCs, we can get a good classifying performance about the subset of L2, M, M_k. The OOB error is about 11~16 percent.

(5) It is impossible to classify the S2, W classes.

10. References

[1] A. Amann, G. Poupart, et al, "Applications of breath gas analysis in medicine". Breath Gas Analysis in Medicine.

[2] J. Rieder, P. Lirk, et al, "Analysis of volatile organic compounds: possible applications in metabolic disorders and cancer screening", the middle european journal of medicine, p181-185, 2001.

[3] L. Breiman, RandomForests, January 2001.

[4] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning, Springer-Verlag, 2001

[5] R. O. Duda, P. E. Hart, D. G. Stork, Pattern Classification, second edition, 2001, John Wiley & Sons, Inc.

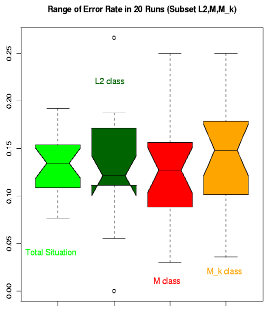
[6] L. Breiman, RFtools—for predicting and understanding data, Interface Workshop, April 2004.

[7] S. Dudoit, R. Gentleman, Classification in microarray experiments, Bioconductor short course, 2003.

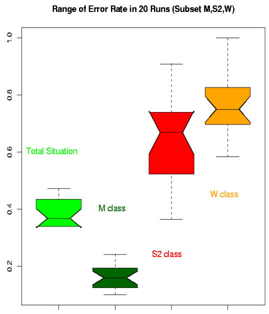
[8] K. S.Remlinger, Introduction and application of random forest on high throughput screening data from drug discovery, http://www4.ncsu.edu/~ksremlin

[9] A. Liaw, The randomForest Package, http://stat-www.berkeley.edu/users/breiman/randomforests

[10] S. Tao, S. Horvath, Unsupervised learning with Random Forest predictors, Nov 2004.



Graph 6



Graph 7