Scandent Tree: A Random Forest Learning Method for Incomplete Multimodal Datasets

Soheil Hor[1] and Mehdi Moradi[2,⋆]

[1] University of British Columbia, Vancouver, British Columbia, Canada
[2] IBM Almaden Research Center, San Jose, CA, USA
mmoradi@us.ibm.com

Abstract. We propose a solution for training random forests on incomplete multimodal datasets where many of the samples are non-randomly missing a large portion of the most discriminative features. For this goal, we present the novel concept of scandent trees. These are trees trained on the features common to all samples that mimic the feature space division structure of a support decision tree trained on all features. We use the forest resulting from ensembling these trees as a classification model. We evaluate the performance of our method for different multimodal sample sizes and single modal feature set sizes using a publicly available clinical dataset of heart disease patients and a prostate cancer dataset with MRI and gene expression modalities. The results show that the area under ROC curve of the proposed method is less sensitive to the multimodal dataset sample size, and that it outperforms the imputation methods especially when the ratio of multimodal data to all available data is small.

1 Introduction

In recent years there has been an interest in multimodality data analysis for disease detection. Ideally, multimodality methods should leverage the strengths of each modality and compensate for weaknesses. Another advantage of multimodality data analysis is discovering novel relations between different modalities. One example is finding the connection between genes related to Alzheimer's disease and related areas in functional MRI [1]. Acquiring multimodal data is, in general, more costly and time consuming than a single modality. As a result, multimodal datasets usually have valuable features, but small sample sizes. This makes it difficult to build classifiers, with large training data, for highly multimodal protocols. Multomodal data is also often high dimensional and pose difficulties in feature selection and classifier building. Ensemble classifiers such as random forest provide a solution for the large feature space in small datasets using feature bagging.

To tackle the issue of incomplete datasets, a variety of data imputation techniques exist. Some of these are non-parametric methods like hot deck imputation, KNN imputation or mean substitution. These methods ignore the possible correlations in data and could add bias. Model-based methods, on the other hand,

⋆ Corresponding author.

N. Navab et al. (Eds.): MICCAI 2015, Part I, LNCS 9349, pp. 694–701, 2015.
DOI: 10.1007/978-3-319-24553-9_85

assume a certain structure to the missing samples, like missing completely at random (MCAR) or missing not at random (MNAR). Examples of these methods include multiple imputation [2], maximum likelihood, stochastic regression [3], expectation maximization [3] and Bayesian methods [4]. While these methods could result in reduced bias, the assumption of specific pattern in the missing components may not be justified, especially in the case of small datasets with complex features. Therefore, a third approach to treating missing data has emerged that maximizes the performance of a classifier. An example is the imputation method proposed by Breiman in [5,6]. This method uses the proximity matrix of the random forest to iteratively predict the missing values in a way that maximizes the overall performance of the random forest and is designed to perform well even in MNAR conditions.

Our motivation in this area stems from the work on combining genomic biomarkers of prostate cancer with imaging biomarkers from multiparametric MRI (mpMRI) to enhance risk stratification. While imaging data is routinely acquired and archived from prostate cancer patients, there are very few patients with imaging data and spatially registered tissue specimens for genomic analysis. As a result, we have a relatively large number of data samples with only mpMRI data (which we call the single modality dataset in this work), and a small set of samples with both mpMRI and gene expression analysis from the same regions of interest (which we call multimodality data). While most of the imputation methods assume a small number of missing values (typically 10%-30% of the whole data), we are dealing with a situation where the multimodal samples only constitute around 10% of the data. While the patients recently recruited into the study provide multimodal data, we intend to find a solution to include the archival data with imaging only samples. In this work, we develop a solution to leverage a large single modality dataset to enhance the training of a classifier based on multimodal data. The proposed method is based on decision trees. However, we describe an entirely novel technique to link different feature sets and predict the class label using information from all of the datasets, multimodal and single modal. We use a large clinical benchmark dataset to show that our method outperforms the current state of the art in random forest imputation methods, particularly in the case of dataset with large missing ratio. We also report very promising preliminary results on our prostate cancer dataset.

2 Method

Let us assume that the training data consists of at least one single modality dataset defined as $S = (s_1, s_2, \ldots, s_{N_s})$ and at least one multimodality dataset defined as $M = (m_1, m_2, \ldots, m_{N_m})$ which is described by the multimodality feature set $F_m = (f_1, f_2, \ldots, f_{km})$. The aim is to train a classifier using both S and M that can predict the outcome class C, for any test data described by F_m. While we do not set conditions on feature or sample sizes, in practical scenarios, the multimodality dataset has fewer samples ($N_m < N_s$). In practice F_s is often a subset of F_m and is missing some of the more discriminative features.

As an advantage of having all the important features, trees formed by the multimodal dataset are expected to partition the feature space very effectively. But because of the low multimodal sample size, the estimation of outcome probability at each leaf may not be accurate. The idea of our method is to reduce prediction error at each leaf of the multimodal tree by using single modality samples that are likely to belong to the same leaf. In order to find these single modality samples, a feature space partitioning algorithm is needed that can simulate the feature space division of the target multimodal tree on the single modality dataset. The proposed method is to grow single modality trees that mimic the feature space division structure of the multimodal decision tree. Growing a tree that follows the structure of another tree from the root to the top brings analogy to the behaviour of "scandent"trees in nature that climb a stronger "support" tree. Considering this analogy, the proposed method can be divided into three basic steps: First, division of the sample space by a multimodal decision tree, called "the support tree". Second, forming the single modality trees that mimic the structure of the support tree, called "scandent trees". And third, leaf level inference of outcome label C, using the multimodal samples in each leaf and the single modal samples that are most likely to belong to the selected leaf.

Support Tree: The first step in the proposed method is growing a decision tree to predict the outcome class based on the multimodal dataset.

Scandent Trees: The second step is to form the scandent trees which enable the assignment of single modality samples to the leaves of the support tree. The process of feature space division in the support tree can be considered as grouping the multimodal data set (M) to different multimodal subsets at each node. Let us define the subset of the samples of the multimodal dataset (M) in the i_{th} node as M_i. The algorithm to form a scandent tree is as follows:

```
for each node $i$ in the support tree starting from the root node,
{
        for each sample $n$ in $M_i$ and each child node $j$ of node $i$
        {
                if $n \in M_j$, $C^{'}_{i,n} = j$
        }
        Grow $T_i$, as optimum tree that for each sample n in $M_i$,
        predicts $C^{'}_{i,n}$ using only $F_s$.
}
```

The above algorithm forms sub-trees T_i for each node i that divide M_i to the child datasets M_j using only the single modality features F_s. Let us assume that the sample space division at node i of the support tree is based on feature f. If $f \in F_s$, then T_i is expected to divide M_i to the child subsets (M_j) using only a single division node and with perfect accuracy. But if $f \notin F_s$, then T_i will be optimized to form the smallest tree that can divide the sample space in a similar manner to the support tree. Using T_i's for feature space division at each node, we can form a new tree that consists of the same division nodes as the support tree but only uses features of a single modality (F_s) for feature space division, we name this single modality tree, a scandent tree. Since the scandent tree is a single modality tree, it can be used to predict the probability that each single modality sample s belongs to each node j calculated by:

$$p(s \in Node_j) = p(s \in Node_j | s \in Node_i) p(s \in Node_i)$$

In which $Node_i$ is the parent node of $Node_j$. The term $p(s \in Node_j | s \in Node_i)$ in the above equation can be estimated by the corresponding sub-tree T_i and $p(s \in Node_i)$ is calculated by recursion. This method is expected to be generally more accurate than direct estimation of the leaves by any other single modality classifier. Because the scandent tree only has to predict the feature division for features that do not belong in F_s and other divisions will be perfectly accurate. Moreover, if two features are dependent over the whole sample space (unconditional dependence), they will also be predictable by each other over a sub-space of the sample space. But if the dependence is conditional, they cannot be universally predicted by each other. The scandent tree locally estimates each division and does not require a global dependence. As a result, it can predict the set of single modality samples that belong to each leaf of the support tree which may not be possible by any other single modality classifier.

Leaf Level Inference: The standard method for leaf-level inference is majority voting. However, if there are single modality samples misplaced by the scandent tree, they may flood the true observations. The proposed method for weighted majority voting is to re-sample from each leaf i and calculate the probability of outcome C by non-uniform bootstrapping. The bootstrap probability of each sample x in leaf i is defined by:

$$p(x)_{bootstrap,leaf_i} = \begin{cases} 1/\mathrm{N}, & x \in M_i \\ p(x \in Leaf_i)/N, & x \notin M_i \quad \& \quad p(x \in Leaf_i) > q \\ 0, & x \notin M_i \quad \& \quad p(x \in Leaf_i) < q \end{cases}$$

In which q is the selected minimum threshold for the probability that a single modality sample belongs to the selected leaf i, and N is the total number of samples in leaf i (single modal and multimodal). As q value increases, the probability that a misplaced sample is used in the leaf-level inference is reduced. This may increase the accuracy of the majority voting but increasing q will also reduce the number of single modality samples at each leaf which decreases the accuracy of the probability estimation. This tradeoff is more evident at the two ends of the spectrum, for $q = 1$ the tree will be the same as the support tree which suffers from low sample size at the leaves. For $q = 0$ all the single modality samples will be used for inference at each leaf and the feature space division of the support tree will have no direct effect on the inference. The optimization of the q parameter for each leaf is essential for optimal performance of the resulting tree. This can be done by cross validation over the multimodal dataset. Because M_i is smaller at deeper nodes, as the support tree gets deeper and develops more division points, it gets harder for the scandent tree to accurately follow the tree structure. Moreover, a higher number of consequent divisions by T_i's leads to accumulative errors. So forming smaller trees and ensembling those in form of a random forest may lead to a more accurate classifier.

Implementation: In building the support trees, we bagged 2/3 of bootstrapped samples and the square root of the dimension of the multimodal feature set. The out of bag samples are used for optimization of the q parameter at each leaf. After growing and optimizing each of the trees, the probability of outcome class C is calculated by averaging the corresponding probabilities of all trees in the forest. We use the R package "rpart" [7] to grow the support tree. This package uses internal cross validation to form the optimal tree. But for the purpose of controlling the bias-variance of the resulting forest, the depth of support tree is limited by controlling the minimum of samples needed for each division. The depth of T_i's in each scandent tree is optimized by cross validation.

Evaluation: We evaluate the performance of the proposed method using a dataset available from University of California at Irvine (UCI) Machine Learning Repository [8]. We also report preliminary results on a prostate cancer multimodal dataset. The datasets are summarized in Table 1.

Heart disease data: This set consists of data from two different studies reported in [9]. One set (data from the Hungarian Institute of Cardiology) is missing two out of 14 features. We use this as the single modal dataset in our experiments. In real world problems, such as our prostate cancer study, the single modal dataset is missing some of the most discriminative features. To simulate this condition we used a classical random forest feature ranking approach. We study the effect of decreasing the number of features in the single modality dataset on the overall performance by sweeping from 12 to two features, always removing the most top-ranking ones. The multimodal dataset in this experiment was the Cleveland dataset consisting of 303 samples. 100 samples were randomly separated and used as test data. The remaining samples were used as the multimodal data for training the support trees. We experimented with scenarios that included 10% to 90% of this data in training of the support trees.

Table 1. Evaluation Datasets

Datasets	Heart Disease		Prostate Cancer	
	Cleveland	Hungarian	MRI and Genetic	MRI only
Sample Size	303	294	27	400
Feature size	14	12	43	4

Prostate cancer data: We also test our method on a dataset that is a perfect example of the target scenario, a small multimodal prostate cancer dataset ($N_m = 27$) accompanied by a relatively large single modal dataset ($N_s = 400$). The single modality dataset consists of four multiparametric MRI features from dynamic contrast enhanced (DCE) MRI and diffusion MRI on a 3 Tesla scanner.

We used the apparent diffusion coefficient (ADC) from diffusion MRI, and three pharmacokinetic parameters from DCE MRI: volume transfer constant, k^{trans}, fractional volume of extravascular extracellular space, v_e, and fractional plasma volume v_p [10,11].

For the 27 multimodal samples, besides the four described imaging features, biopsy tissue samples with known pathologic state (cancer or normal determined by a histopathologist) were also available. RNA was extracted and purified [12]. The expression level of 39 genes that form the most recent consensus on the genetic signature of prostate cancer for patients with European ancestry were used as features. This signature is reported and maintained by National Institute of Health [13]. We have 27 samples with genomic analysis and registered imaging data (14 normal, 13 cancer) from 19 patients. The evaluation of the proposed method on this small dataset was carried out in a leave-one-out scheme. Each time, the support trees were trained using 26 multimodal samples, with all the 400 single modality data samples used for forming the scandent trees.

3 Results and Discussion

Heart Disease Dataset: Figure 1 shows the AUC of the proposed method and the rfImpute method for different multimodal sample sizes. Each box in this figure shows AUC values for different single modal feature set sizes and a fixed multimodal dataset sample size. The expected upward trend in AUC *vs.* multimodal sample size is evident and it can be seen that the proposed method outperforms the rfImpute method especially in smaller samples sizes. For example, when only 14 multimodal samples are available, the rfImpute method results in a mean AUC of 0.90 whereas the proposed method delivers an AUC of 0.94. As the number of multimodal samples increases to 112, the performances increase for rfImpute and scandent tree to 0.96 and 0.97, respectively. In other words, the scandent tree approach has a clear advantage when the dataset with multimodal data is significantly smaller.

Figure 2 shows the AUC of the proposed method and the rfImpute method for different single modality feature set sizes. Each box shows changes of AUC for different sample sizes at a fixed feature set in the single modality data. Smaller variances of the boxes for the proposed method, especially in smaller feature set sizes, show that the proposed method is on average less sensitive to the multimodal sample size especially when the single modality dataset has a large number of missing features. For example, at feature vector size of 2 for the single modality dataset, the performance of rfImpute varies from 0.88-0.98, whereas scandent tree shows a performance range of 0.93-0.98. This stable behavior is due to the unique ability of the scandent trees to predict division points for missing features that only conditionally depend on the available features.

Prostate Cancer Dataset: In leave one out validation, the proposed method resulted in an AUC of 0.95 for the prostate cancer data. The rfImpute approach resulted in an AUC of 0.8. The difference was statistically significant (p <0.02). This dataset is an example of the worst case scenario of missing data: a large

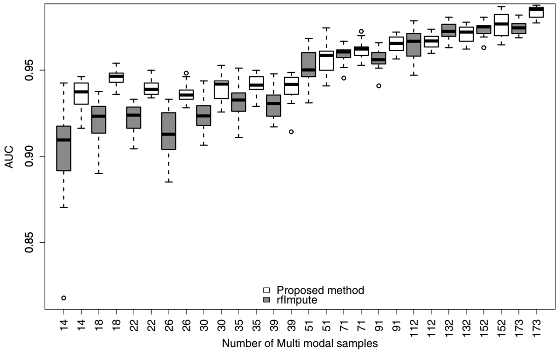

Fig. 1. AUC vs multimodal sample size for heart disease dataset (each box shows AUC values for different single modal feature sets)

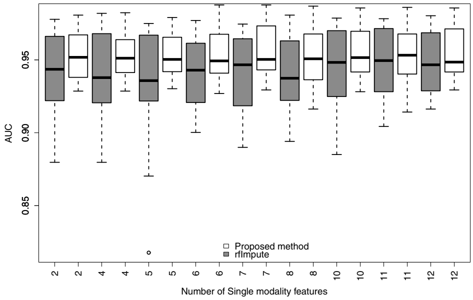

Fig. 2. AUC vs single modal feature set size for heart disease dataset (each box shows AUC values for different multimodal sample sizes)

non-random portion of the data is missing the potentially more powerful genomic features resulting in a very small multimodal dataset. At the same time, the number of features on the single modality (imaging) side is small. As a result, the power of the proposed method in comparison with rfImpute is on full display. It is also important to understand that the scandent tree is providing a platform to incorporate the genomic data, despite the very limited number of samples. In the absence of such methodology, if one uses only the imaging features with an optimized random forest, the AUC is 0.74.

4 Conclusion

In this paper we addressed the problem of incomplete multimodal datasets in random forest learning algorithms in a scenario where many of the samples are non-randomly missing a large portion of the most discriminative features. We introduce the novel concept of scandent trees. The results show that the proposed method outperforms the embedded missing value imputation method of random forests introduced in [5], particularly in smaller samples sizes. The method is in general less sensitive to the number of missing features and the multimodal sample size. We showed that the proposed method enables the integration of a small genomic plus imaging dataset, with a relatively large imaging dataset.

In this paper we used a single modal dataset to improve the accuracy of leaf-level inference in multimodal trees. Future work will address the possibility of using single modality data at the test stage.

Acknowledgment. Funding from Canadian Institutes of Health Research (CIHR). Prostate imaging and genomic data were obtained at Vancouver General Hospital and UBC Hospital with approval from Clinical Research Ethics Board and informed patient consent. The authors would like to acknowledge Drs. Peter Black, Larry Goldenberg, Piotr Kozlowski, Jennifer Locke, Silvia Chang, Edward C. Jones, Ladan Fazli, all from UBC; Dr. Elai Davicioni, Christine Buerki, Heesun Shin, and Zaid Haddad from GenomeDx Biosciences Inc.

References

1. Liu, J., Calhoun, V.D.: A review of multivariate analyses in imaging genetics. Frontiers in Neuroinformatics 8, 29 (2014)
2. Rubin, D.B.: Multiple imputation for nonresponse in surveys, vol. 81. John Wiley & Sons (2004)
3. Gold, M.S., Bentler, P.M.: Treatments of missing data: A Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization. Structural Equation Modeling 7(3), 319–355 (2000)
4. Kong, A., Liu, J.S., Wong, W.H.: Sequential imputations and bayesian missing data problems. Journal of the American Statistical Association 89(425), 278–288 (1994)
5. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
6. Liaw, A., Wiener, M.: Classification and regression by randomforest. R News 2(3), 18–22 (2002)
7. Therneau, T.M., Atkinson, B., Ripley, B.: rpart: Recursive partitioning. R package version 3.1-46. Ported to R by Brian Ripley 3 (2010)
8. Lichman, M.: UCI machine learning repository (2013)
9. Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.J., Sandhu, S., Guppy, K.H., Lee, S., Froelicher, V.: International application of a new probability algorithm for the diagnosis of coronary artery disease. The American Journal of Cardiology 64(5), 304–310 (1989)
10. Haq, N.F., Kozlowski, P., Jones, E.C., Chang, S.D., Goldenberg, S.L., Moradi, M.: A data-driven approach to prostate cancer detection from dynamic contrast enhanced MRI. Computerized Medical Imaging and Graphics 41, 37–45 (2015)
11. Moradi, M., Salcudean, S.E., Chang, S.D., Jones, E.C., Buchan, N., Casey, R.G., Goldenberg, S.L., Kozlowski, P.: Multiparametric MRI maps for detection and grading of dominant prostate tumors. Journal of Magnetic Resonance Imaging 35(6), 1403–1413 (2012)
12. Erho, N., et al.: Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. PloS One 8(6), e66855 (2013)
13. National Institutes of Health: National cancer institute: PDQ genetics of prostate cancer (Date last modified February 20, 2015)