Medical Image Analysis 34 (2016) 30–41

Contents lists available at ScienceDirect

Medical Image Analysis

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

Learning in data-limited multimodal scenarios: Scandent decision forests and tree-based features☆

Soheil Hor[a], Mehdi Moradi[b,*]

[a] University of British Columbia, Vancouver, BC, Canada
[b] IBM Almaden Research Center, San Jose, CA, USA

ARTICLE INFO

Article history:
Received 15 January 2016
Revised 25 April 2016
Accepted 28 July 2016
Available online 29 July 2016

Keywords:
Multimodal data analysis
Incomplete data analysis
Decision forest

ABSTRACT

Incomplete and inconsistent datasets often pose difficulties in multimodal studies. We introduce the concept of scandent decision trees to tackle these difficulties. Scandent trees are decision trees that optimally mimic the partitioning of the data determined by another decision tree, and crucially, use only a subset of the feature set. We show how scandent trees can be used to enhance the performance of decision forests trained on a small number of multimodal samples when we have access to larger datasets with vastly incomplete feature sets. Additionally, we introduce the concept of tree-based feature transforms in the decision forest paradigm. When combined with scandent trees, the tree-based feature transforms enable us to train a classifier on a rich multimodal dataset, and use it to classify samples with only a subset of features of the training data. Using this methodology, we build a model trained on MRI and PET images of the ADNI dataset, and then test it on cases with only MRI data. We show that this is significantly more effective in staging of cognitive impairments compared to a similar decision forest model trained and tested on MRI only, or one that uses other kinds of feature transform applied to the MRI data.

1. Introduction

In recent years there has been a surge of interest in multimodal data analysis. Different modalities provide researchers with complementary information about diseases and provide the means for more accurate detection and staging. This can be valuable in the case of progressive illnesses such as Alzheimer's disease and certain kinds of cancer. Simultaneous analysis of multiple modalities could also help us discover novel relations between different modalities, such as understanding the relationship of molecular changes caused by a disease and its imaging signature when both genetics and imaging data are available. Given these potential advantages, there has been a trend of merging different modalities in biomedical studies. For instance the Alzheimer's Disease Neuroimaging Initiative (ADNI), a six year $65 million study, has focused on using medical imaging modalities like magnetic resonance imaging (MRI) and positron emission tomography (PET) together with genetics and other clinical biomarkers for gaining better understanding of Alzheimer's Disease and its progression (http://adni.loni.usc.edu).

Acquiring multimodal data is generally more costly and time consuming than a single modality. As a result, multimodal datasets usually have valuable features, but a small set of samples with all features. This makes it difficult to build classifiers with large training data for highly multimodal protocols. For instance, in the case of the ADNI dataset, nearly half of the patients are missing the PET data. PET imaging is expensive and requires the use of radioactive tracers. As a result, a large number of patients only receive MRI scans, despite the fact that PET imaging provides unique brain functional information by quantification of the cerebral blood flow, metabolism, and receptor binding, which are not measured with MRI. This is a common scenario in dealing with multimodal data. A computational model that can be trained on both MRI and PET data (multimodal data), but be deployed in clinical settings where only MRI (single modal data) is available, is a valuable contribution in this area, provided that the model outperforms one that is solely trained on MRI data.

Another common scenario is the case of a new multimodal research protocol including at least one component which is only obtained in the course of the study itself. An example of this scenario

☆ For the Alzheimers Disease Neuroimaging Initiative: Part of the data used in preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how-to-apply/ADNI-Acknowledgement-List.pdf

* Corresponding author.
E-mail addresses: mmoradi@us.ibm.com, moradi2004@gmail.com (M. Moradi).

http://dx.doi.org/10.1016/j.media.2016.07.012
1361-8415/

is our study to understand the relationship of molecular signature of prostate cancer with the imaging signature of the disease obtained through multiparametric MRI (mpMRI). The hope is that the simultaneous analysis of the molecular and imaging data can provide clues towards building a reliable and affordable clinical staging test. Since prostate cancer is a multifocal disease with tumors at different stages in each foci, this study requires tissue samples for molecular analysis that are obtained from a specific area with known spatial registration to the MRI images. The steps taken to acquire this data are not part of the clinical routine and the pace of data acquisition is slow. On the contrary, we have access to hundreds of samples with only mpMRI data and known histology. In this scenario, we are building a computational model that would be studied on the mpMRI+genomics (multimodal) data. Here, we may benefit from a computational framework that can utilize the rather large single modality dataset during training, but be able to handle the multimodal data at the testing stage.

In this paper we present solutions, within the context of decision tree/forest paradigm of learning to address the problems posed in the two scenarios described above. Recent relevant work includes an investigation of the applications of imputation methods for dealing with missing values in the ADNI dataset (Campos et al., 2015). The results show that joining a multimodal dataset with a single modal dataset by imputation of the missing values improves the classification accuracy, compared to training a classifier on either the single modal or the available multimodal data. In our current work, we intend to go beyond the paradigm of imputation. This is due to the fact that multimodal studies do not necessarily hold the usual assumptions in imputation that only a small number of data points are missing at random. We intend to deal with situations where blocks of data are missing together and the missing values are not spread randomly.

One trend in dealing with block wise missing values in multimodal datasets is separately modeling different blocks of data and then joining the resulting models by using a merging classifier or an ensembling method. One of the most successful attempts in this field is applying multi-source learning techniques for dealing with block wise missing data in ADNI (Yuan et al., 2012; Xiang et al., 2013; 2014). The incomplete Multi-Source Feature learning method (iMSF) proposed by Yuan et al., models different blocks of data with similar feature sets as different tasks and learns a joint model by imposing a sparse learning regularisation on these tasks (Yuan et al., 2012). The authors also propose a different approach by using a model score completion scheme. This method is based on training independent classifiers on different blocks of data, and then using the prediction scores calculated by each classifier as a new presentation of the data that can then be imputed using conventional imputation techniques. A recent paper by Yu et al., proposes a new method based on Multi-task Linear Programming Discriminant (MLPD) analysis (Yu et al., 2014). This method formulates the problem as a multi-task learning scenario in a fashion similar to the iMSF method but does not constraint all of the tasks to share the same set of features, allowing joint learning of a more flexible model.

As a limitation to these studies, the training and testing datasets are assumed to have the same distribution and feature sets. Recently, Cheng et al., addressed this issue and proposed a method for multimodal data analysis based on multimodal manifold-regularized transfer learning method (Cheng et al., 2015). This method enables using data from different domains together with unlabeled data for multimodal classification. This work uses a feature transform based data fusion approach and includes a sparsity constraint in order to deal with the high dimensionality issue.

In this paper we address the same limitations reported in Cheng et al. (2015), but with different assumptions that fit our scenarios. We don't assume that there is unlabeled data available. We do assume that the feature set of the test data is a subset of the training data. For instance, in case of the ADNI dataset, we assume that the training dataset consists of a set of samples with both MRI and PET data (although incomplete) but the test sample only consists of MRI data. This scenario is aimed at enabling the use of multimodal datasets for training of a classifiers that requires only a subset of modalities for testing.

1.1. Scandent trees and tree-based feature transforms

An important issue in multimodal classification is the high dimensionality problem that poses difficulties in feature selection and classifier building. The majority of the methods in the literature use the multi-kernel SVM framework for multimodal classification and need to impose sparse conditions on the multimodal feature set in order to avoid over-fitting (Cheng et al., 2015; Jie et al., 2015; Zhang and Shen, 2012).

In the current paper, by working within the decision tree/forest paradigm we benefit from its embedded way of dealing with high dimensional data through feature bagging (Breiman, 2001). Another motivation for the use of decision forest paradigm is that it provides the ability to morph the treatment of missing data within the framework of learning to maximize the classification performance. This area of work has seen significant contributions in recent years. These include the state of the art imputation methods embedded in the classification and regression tree (CART) algorithm (Steinberg and Colla, 2009; Quinlan, 2014) and in Random Forests (rfImpute) (Breiman, 2001).

A key element of our methodologies is the concept of scandent trees recently proposed in Hor and Moradi (2015). To our knowledge this is the first decision-forest-based method with an embedded way to deal with block wise missing data in multimodal datasets. In Hor and Moradi (2015), we only considered one scenario: a classifier that benefits from a large single modal dataset at the time of training, but is tested on multimodal data. This was motivated by our work in the area of prostate cancer staging.

Another key element within our work is the concept of tree-based feature maps. A disadvantage of decision forests compared with SVM is the lack of an embedded framework for kernel-based feature transformation in the case of forests. Using multi-kernel approaches, researchers have devised solutions for incorporation of various modalities in the SVM context.

Other related work includes the "auto-context" method introduced in Tu and Bai (2010) that provides a general interface to iteratively form feature transforms that can be interpreted as context features. However, similar to the other iterative methods, this method assumes the same feature set for both training and testing stages. Tree-based feature transforms have recently received some attention. For example, a recent work by Cao et al. (2015) uses stacked decision forests. This method is based on using the probability values estimated by trees in a random forest as a feature vector, and using this feature vector for training of an enhanced decision forest, potentially together with the original feature set. Inspired by the applications of multi-kernel SVMs in multimodal data analysis, we apply this concept of tree-based feature transforms for multimodal data analysis.

This manuscript reports two specific contributions: *First*, We report an improved version of our algorithm reported in Hor and Moradi (2015) for dealing with the missing data problem in the multimodal test scenario. We provide complementary results on a prostate cancer dataset and compare the scandent tree method to different state of the art methods for missing value imputation.

Second, entirely new to this work, we develop the idea of scandent tree-based feature transforms to solve the problem of missing data in the single modal testing scenario. This problem has many clinical applications in areas where expensive research protocols

meet the realities of clinical practice and high cost. Here, the assumption of a multimodal dataset with block wise missing values remains. However, there is no multimodal assumption about the test set. To solve this problem, we use the idea of tree-based feature transforms along with the scandent tree. This combination allows us to use tree-based feature transforms built on one modality to transform the features from a different modality. Using this approach, we use MRI and PET data in the ADNI dataset and train a classifier that only requires the MRI data for the prediction of different stages of Alzheimer's disease. We show that the inclusion of the PET data at the time of training results in an improved classification accuracy, even though the test cases are not subjected to PET imaging.

The structure of the remaining of the paper is as follows: In Section 2 an improved version of the scandent tree algorithm is presented. We then describe our new contribution for using the scandent tree model in the single modal classification task based on the concept of tree-based feature transforms. In Section 3, we introduce the datasets used in this work and the evaluation methods used in each scenario. Section 4 shows the experimental results on the prostate cancer dataset and the ADNI dataset, each targeting one of the two multimodal scenarios. Finally, Section 5 provides a detailed discussion on the results.

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 multi-modality dataset defined as $M = (m_1, m_2, \ldots, m_{N_m})$ which are described respectively by the multi-modal feature set F_m and the single-modal feature set F_s, where $F_s \subset F_m$. We do not set conditions on the feature or sample sizes but in practical scenarios, usually the multi-modality dataset has fewer samples ($N_m < N_s$). Also the single modal set is missing some of the more discriminative features. In this section we explain the proposed method for two target scenarios: First we aim to train a classifier using both S and M that can predict the outcome class C, for any test data described by F_m. Then we assume another scenario in which the classifier is trained to predict the outcome class C using the same two datasets, but the test data is described only by F_s. In other words, in the first scenario we make use of a single-modal dataset for optimization of a multi-modal decision forest. While in the second scenario we use the multi-modal dataset to improve the performance of a single-modal random forest.

2.1. First scenario: the multimodal classification task

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 the outcome probability at each leaf may not be accurate. The proposed method tries to reduce the 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. This tree can be one of the trees in a decision forest or an individual tree grown using any of the well known methods, such as C4.5 (Quinlan, 2014) and CART (Therneau et al., 2010b). The method used in this paper for growth of the support tree is based on the implementation of CART algorithm in the package "rpart" in R language (Therneau et al., 2010a).

Assuming that the tree is grown and optimized using the multimodal dataset M, there are two steps that might be the source of classification error in the tree: Division of sample space at inner branches, and majority voting at the leaves. The sample space division requires sufficient sample size at each division point which becomes an issue as the tree gets deeper. However, ensembling within the forest paradigm compensates for occasional incorrect divisions at inner branches, leaving majority voting at the leaves as the critical step to get a precise estimation of probability of the class label. This error can be compensated for by the scandent trees.

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. Let us define the subset of the samples of M in the i_{th} node as M_i and the feature used for sample space division at node i as f_i.

Intuitively, the idea of the scandent tree algorithm is to break the support tree into subtrees that partition the sample space either using only the missing modalities or only the shared modality. And then replace these sub-trees with a single modal local tree that divides the samples in a similar way. In order to reduce the number of consecutive estimations, it is critical that the single modal local trees estimate the largest subtrees in the support tree that hold these assumptions. We name the nodes that define the boundaries of such trees as 'link node's. Using mathematical notations a link node can be defined as follows:

For any arbitrary choice of node j, and its immediate parent node i, we define node j as a 'link node' if f_i belongs to a different feature set from f_j, or if node j is either the root node or a leaf. In other words, node j is a link node if and only if :

Node j is the root node,

or

Node j is a leaf node,

or

$f_j \in F_s$ and $f_i \notin F_s$,

or

$f_j \notin F_s$ and $f_i \in F_s$.

It can be seen in Fig. 1 that node $j2$ is a link node because it is based on a feature set different from it's immediate parent (node $i1$). For a similar reason nodes $i1$ and $j3$ are also link nodes while nodes $j1$, $i2$ and $j4$ are not. The other link nodes in this example are the root node (node R) and the leaf nodes (nodes $k1$ to $k8$).

We define the link nodes among the direct and indirect child nodes of node i that are found first in a Depth-First-Search (DFS) on the subtree rooting from node i as the set of nearest child link nodes of node i. Each link node and its nearest child link nodes can be used to address a subtree in the support tree that uses one and only one feature set for sample space partitioning. For instance in Fig. 1, the nearest child link nodes of node $i1$ are $k1$, $k2$ and $j2$.

The subtree that roots from node $i1$ and divides $i1$ to nodes $k1$, $k2$ and $j2$ is a tree that only uses the missing modality for feature space division. A similar subtree would be the single modal

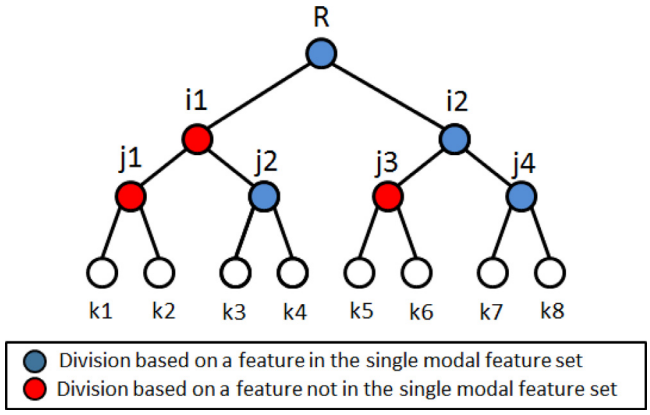

Fig. 1. The support tree.

subtree rooting from node R that only uses the shared modality to divide the sample between nearest child link nodes ($i1$, $j3$, $k7$ and $k8$). Assuming that for each division node i in the set of the link nodes of the support tree, there exists a set of nearest child link nodes $j_1, j_2, \ldots j_{ki}$. We define T_i as an optimum tree that can divide the set of multimodal samples at node i (M_i) to the set of multimodal samples at each child node (M_j) using the feature set F_s. The pseudo-code for forming such a tree is as follows:

For each link node i in the support tree,
{
  For each sample n in M_i and each node j in set of nearest child link nodes 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 local trees T_i for each node i that divide M_i to the child subsets M_j, using only the single modality features F_s. Here C' is a new categorical label-set defined for the corresponding local tree. For each sample in the parent node, the C' is assigned in a way that the samples belonging to a specific child node j are mapped to the same category within C'.

For each node i, if $f_i \in F_s$, then T_i is expected to divide M_i to the child subsets (M_j) with perfect accuracy. But if $f_i \notin F_s$, 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 link nodes as the support tree but only uses features of a single modality (F_s) for sample space division, we name this single modality tree, a scandent tree. Since T_i's are single modal trees, they can be used to predict the probability that each single modality sample s belongs to link 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 link node of $Node_j$, the term $p(s \in Node_j | s \in Node_i)$ is 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 other single modality classifiers. Because the scandent tree only has to predict the division boundaries for features that do not belong in F_s and other divisions will be perfectly accurate.

As an example, Fig. 2 shows that the subtree that divides node i to nodes $k1$, $k2$ and $j2$ in the support tree is replaced by a local tree T_i in the scandent tree that estimates the same sample partitioning.

Given the small multimodal sample size, the local trees could be prone to overfitting if only the few samples in the corresponding link nodes are used for training T_i's. As an improvement compared to the earlier version of this method, we now overcome this problem by using all of the available multimodal samples (M) for training of each local tree by running the whole multimodal training set through the corresponding sub-tree of the support tree. This will give each sample in the multimodal dataset a label from the set C'. This method adds more multimodal samples to the parent link node (M_i) and each child link node (M_j) which results in better estimation of T_i. We found that using this trick adds to the robustness of the scandent.

Leaf level inference: The standard method for leaf-level inference is majority voting. However, if there are a large number of single modality samples misplaced by the scandent tree, they might flood the original multimodal samples.

To tackle this problem we define the weights of each sample x in leaf i as:

$$w(x)_i = \begin{cases} 1/N, & x \in M_i \\ p(x \in Leaf_i)/N, & x \notin M_i \ \ \& \\ & p(x \in Leaf_i) > q \\ 0, & x \notin M_i \ \ \& \\ & p(x \in Leaf_i) < q \end{cases}$$

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 resulting in low precision of the probability estimation. This trade-off 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.

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, using out of the bag samples in case of a decision forest. Using non-uniform re-sampling instead of majority voting ensures that the single modal samples at the leaves are randomized. This randomization is critical because the single modal samples are not randomly selected in the scandent tree growth algorithm and without re-sampling, there is a possibility that many of the scandent trees in the resulting forest are not independent. This would violate one of the basic requirements of tree ensembling in a decision forest.

Although the proposed algorithm is explained only for one single modality dataset, the same method can be applied on different single modality datasets using the same support tree. As a result, the proposed framework can be used flexibly when different subsets of features are missing.

Implementation: For building the support trees, we randomly bagged 2/3 of the multimodal samples and randomly selected the square root of the dimension of the multimodal feature set as the feature bag. This bootstrapping and bagging phase is done separately for each of the outcome classes to ensure balanced class labels. Then the scandent trees are formed and for each leaf of each support tree in the forest the q parameter is optimized using the corresponding out of the bag samples.

After growing and optimizing each of the trees, the probability of outcome class C is calculated by averaging the corresponding

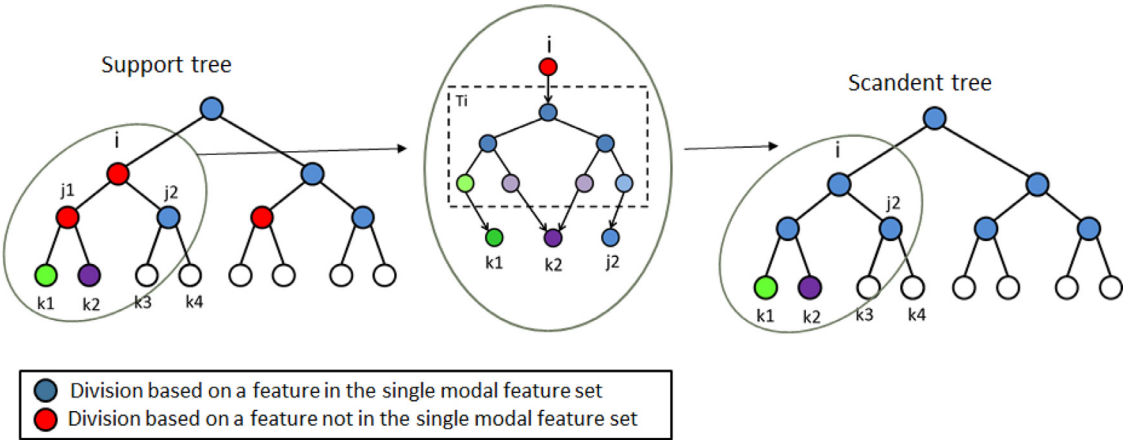

Fig. 2. The scandent tree.

probabilities of all trees in the forest. We use the R package "rpart" (Therneau et al., 2010a) both for growing each support tree and each of the local single modal trees (T_i's). 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 internal cross validation.

2.2. Second scenario: the single modal classification task

For the first scenario we assumed that the missing modalities only affect the training of the classifier and the test data was assumed to be complete. In this section, we describe a method that can use the scandent tree model for training a single-modal classifier that is able to predict the outcome class even when the test data is incomplete.

To obtain a forest that transfers the value of the multimodal dataset into a single modal environment, it is tempting to simply replace all the trees in a forest trained on the available multimodal training data with their corresponding scandent trees. However, this approach fails due to bias and the fact that many of the multimodal divisions of support trees might not be predictable by the single modal feature set.

Instead, we choose an approach inspired by the use of decision trees as feature maps. For this we start with growing a scandent forest similar to the method explained for the first scenario. However, instead of directly using the scandent trees, we use the set of local trees (T_i's) from all the scandent trees of a multimodal forest as tree-based "feature-transforms". Each T_i is a single modal tree which maps F_s to a new space defined by the corresponding $C^{'}$ set. This means that each T_i yields a categorical feature to describe each sample. Then we use the single modal dataset with the extended feature set, including the original and these tree-based features, to grow an improved single modal forest. Note that trees trained on single modal features can be directly used as categorical or continuous (similar to Cao et al. (2015)) feature transformers. In the current work, however, we use the scandent subtrees to link two inconsistent datasets.

This method has a few advantages compared to the conventional method for forming a single modal decision forest or directly using the scandent trees as a new set of trees in a single modal decision forest. First, because at each split of each tree in the single modal forest, the tree growth algorithm searches for the best division feature among both the original single modal features and the new features generated by the local trees (T_is), the resulting tree is expected to be more accurate than both the scandent tree and the tree grown using only the original single modal features. Second, although the T_is are formed by a small multimodal dataset, the feature selection criteria (Gini impurity or information gain) is calculated based on the large single modality dataset. In other words, the single modal forest uses the features inspired by the multimodal forest, but it is completely randomized and optimized based on the larger single modal dataset.

Implementation: The first step is to grow a multimodal forest and the related scandent trees using the method explained in the previous sections. Then the local trees (T_is) are extracted from each tree and each T_i is used as a feature generator for single modal dataset. Given that each T_i is a single modal classifier, it can assign labels relative to the local class labels ($C^{'}$) to each single modal sample. The resulting labels are used as new categorical features which can be calculated for any test data using the corresponding T_i. We then use a conventional decision forest growth method similar to what was explained in the previous section to grow a forest using this set of new features together with the original single modal feature set.

It should be mentioned that because the local trees are trained using the small multimodal dataset, many of the generated features might not be useful for the single modal decision forest. Considering the large number of local trees in a random forest, this can flood the original single modal features.So we filter the new features by a conventional feature selection algorithm, namely based on the feature importance measure in a decision forest. We apply feature bagging separately to the set of the original single modal features and the new features, and then merge them together to form the feature bag used for each single modal tree.

Table 1
Evaluation Datasets.

Parameter	Prostate cancer	ADNI
Number of Multimodal features	44	9
Multimodal sample size	27	218
Number of Single modal features	5	7
Single modal sample size	428	508

3. Evaluation

We report results on a prostate cancer multimodal dataset and an Alzheimer's disease dataset. A summary of the datasets used in this paper can be seen in Table 1.

3.1. Prostate cancer data

This consists of a small genomics+MRI prostate cancer dataset ($N_m = 27$) accompanied by a relatively large MRI only dataset ($N_s = 428$). The single modal dataset consists of five multi-parametric MRI features from dynamic contrast enhanced (DCE) MRI and diffusion MRI on a 3 Tesla scanner. We used the apparent diffusion coefficient (ADC) and fractional anisotropy (FA) 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 (Haq et al., 2015; Moradi et al., 2012).

This data is from patients undergoing radical prostatectomy at Vancouver General Hospital and has been collected with informed consent, and with the approval of the Research Ethics Board of the Vancouver General Hospital. Imaging is performed a week before the surgery. After the surgery, the prostate specimens were processed with wholemount cuts that matched the slices in the MRI scans. A cutting device and the procedure described in Drew et al. (2010) ensured that the cuts matched the MRI slices. An experienced pathologist outlined the area of the tumor/normal from wholemount histopathology slides.

The tissue samples were then obtained by needle biopsy from the corresponding formalin-fixed paraffin-embedded (FFPE) tissue blocks and RNA was extracted and purified from these samples. An elaborate specimen cutting and registration mechanism described in previous work (Haq et al., 2015; Drew et al., 2010) ensures that the MRI feature are calculated and averaged over the same region of interest of the prostate gland that is sampled for gene expression analysis. The expression level of 39 genes that form the most recent consensus on the genetic signature of prostate cancer for patients with European ancestry as reported and maintained by National Institutes of Health (Date last modified 02/20/2015) were used as features (Appendix A).

We have 27 samples with gene expression data and registered imaging data (14 normal, 13 cancer) from 21 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 samples, with all the single modality data samples and features used for forming the scandent trees.

3.2. Alzheimer's disease data

We test the proposed single modal classification method on a dataset from Alzheimer's Disease Neuro-imaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Dr. Michael W. Weiner. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimers Disease (AD) (For up-to-date information, see www.adni-info.org). The ADNI study is an example of a multimodal scenario in which a large portion of samples are missing one of the modalities. In this paper we take the samples that come from patients with both MRI and PET scan as multimodal dataset ($N_m = 218$) accompanied by a relatively large single modal dataset ($N_s = 508$) consisting of patients with only MRI data. This includes the MRI data from the 218 multimodal samples.

The single modality dataset consists of MRI volume measurements of six ROIs in the human brain (ventricles, hippocampus, whole-brain, entorhinal, fusiform and mid-temporal) and intracranial volume (ICV) in mm^3. The multimodal feature set consists of the same MRI features together with two additional PET scan features, FluoroDeoxyGlucose (FDG) measurement and AV45 uptake measurement. The outcome labels include cognitively normal patients (NL), patients with confirmed dementia (AD) and patients with mild cognitive impairment (MCI). The MCI group can be divided into progressive (pMCI) that eventually converts to dementia and stable (sMCI). In this paper we assume a maximum of 36 month conversion time for the MCI class to be considered pMCI.

The distribution of different outcome classes in the two datasets is as follows: for the normal class we have 178 samples in the single modal dataset versus only 18 samples in the multimodal dataset, for the dementia class we have 108 single modal samples versus 29 multimodal samples, for the sMCI class we have 126 single modal versus 144 multimodal samples and for the pMCI class we have 96 single modal samples versus 27 multimodal samples. In other words, the multimodal dataset is much smaller than the single modal dataset, and it also does not have the same distribution of outcome classes. This makes the data fusion between the two datasets extremely difficult with traditional approaches such as imputation. We examine the performance of the proposed method by reporting AUC for three classification scenarios: NL versus pMCI, sMCI versus AD. and sMCI versus pMCI.

3.3. Baseline methods used for comparison

A natural choice for a baseline imputation method, in the multimodal test scenario of prostate cancer, is the state of the art imputation method embedded in decision forests. In our results, this method is referred to as rfImpute. This iterative imputation approach starts with a median imputation of the whole dataset and then grows a random forest using the imputed dataset. In the next step the estimations of each missing value are updated by using the proximity matrix of the resulting random forest as weights in a voting scheme. This process is iterated using the new imputed values until a stable estimation is achieved. Another imputation method worth investigating is the state of the art imputation method of C5.0 trees. This method uses an algorithm similar to the proposed method in the sense that it assigns fractional weights (probabilities of belonging to a certain node). However, the fractional weights are chosen separately for each node and are only based on the proportion of samples in each parent node that end up in the corresponding child nodes. Moreover, the inference method is based on a simple weighted-voting scheme. The complete list of comparison methods include two data-discarding methods where we simply drop one or the other dataset (the single modal forest and the multimodal forest), two forest-based imputation methods (C5.0 forest and rfImpute) and two other general purpose imputation methods, namely replacing the missing values with zero, and replacing with the weighted average value of the K nearest neighbors (KNN), K = 10 in our work (Ashab et al., 2014).

In the single modal scenario, our proposed method uses the available multimodal training data to find a helpful tree-based feature transform to be used on the single modal dataset. Since we do not have the assumption of similar feature sets in training and testing in this scenario, imputation does not provide a fair comparison. Our method is a tree-based feature transform that uses scandent trees. Therefore, we compare it with two other similar methods that involve a transformation, but not the scandent trees. These include (1) PC-forest: using principal components (PCs) of the single modal features used along with the original single modal features for training an enhanced-forest baseline, and (2) Single modal feature transform forest: uses tree-based feature transforms generated using only single modal features. This is similar to stacking forests. However, we use categorical, as opposed to continuous probability features unlike (Cao et al., 2015). (3) Simple single modal forest trained and tested on single modal data.

It should be noted that in order to guarantee a fair comparison between the proposed methods and the baseline classifiers,

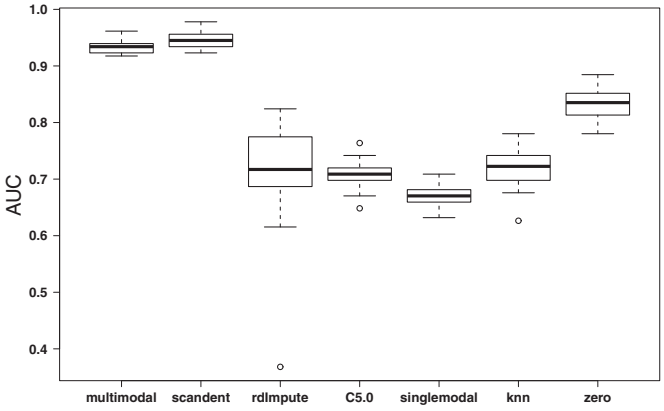

Fig. 3. AUC for multimodal classification task, prostate cancer dataset.

the support decision forest classifier and the single modal forests used in all the methods are designed to be similar.

4. Results

4.1. Multimodal classification task

Fig. 3 shows the AUC obtained on this data, for detection of prostate cancer, for several experiments, namely from left to right the bars show the distribution of AUC areas for (1) a multimodal decision forest that simply ignores the existence of archival imaging data, (2) our proposed scandent tree approach to use the archival data to improve the performance of a forest trained and testes on multimodal data, (3) the standard rfImpute method applied at the forest level to include the single modal data in training, (4) the standard C5.0 method applied at trees level, (5) training and testing a tree using only the single modal features of the multimodal set, (6) KNN imputation, and (7) zeroing of the missing feature values.

It can be seen that the multimodal forest is performing significantly better than the single modal forest even though the sample size of the single modal dataset is significantly larger than the multimodal dataset. This suggests that the missing modality, in this case the genetic features, is far more discriminative than the shared modality, MRI. The imputation methods outperform a single modal forest, but they fail to outperform the multimodal forest. This shows that even the state of the art imputation methods may misguide the decision forest when a large portion of data is missing, to the extent that a simple imputation method like zero replacement outperforms the state of the art imputation approaches.

In case of the proposed method, scandent forest, the significant advantage over a single modal forest, and each of the imputation methods is evident. Moreover, the proposed method does not introduce bias into the prediction like the other imputation methods and as a result, it outperforms both the multimodal forest and the single modal forest. However, because the shared modality is significantly less discriminative than the missing modality, the improvement in performance is small (mean AUC of 94% for the scandent forest and 93% AUC for the multimodal forest), although it is statistically significant ($p < 0.01$).

4.2. Single modal classification task

In the single modal scenario, we used experiments on the ADNI dataset to evaluate our proposed method of using scandent trees to extract tree-based feature transforms, in comparison with other approaches to enhance the single modal forest that do not use scandent trees. The performance of all these methods is evaluated for three classification tasks: discrimination of normal samples from progressive MCI (NL vs. pMCI), discrimination of stable MCI from progressive MCI (sMCI vs. pMCI) and stable MCI from dementia (sMCI vs. AD).

5-fold cross-validated ROC curves of the baseline single modal forest, PC forest, single modal feature transform forest, and the scandent tree multimodal feature transform forest for NL vs. pMCI classification task are shown in Fig. 4.

As it can be seen in Table 2, the feature transform forests significantly outperform the baseline single modal forest and the PC forest. The difference between the baseline and the feature transform methods is statistically significant ($p = 0.01$) for the single modal transformed features and ($p = 0.002$) for multimodal feature transforms. However, the improvement in the performance achieved by the PC-based features is not statistically significant (p-value $= 0.92$). The multimodal feature transforms are more effective compared to the single modal feature transforms. This difference is significant ($p = 0.04$).

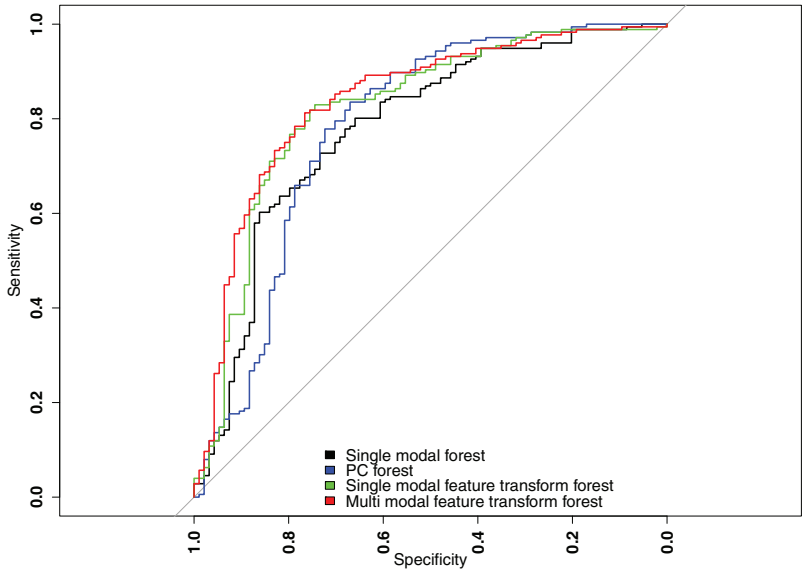

Fig. 4. ROC curve for NL vs. progressive MCI classification, single modal classification task, ADNI dataset.

Table 2
Accuracy (Acc), Sensitivity (Sens), Specificity (Spec) and Area under ROC curve (AUC) of the proposed methods and the baseline forest for the NL vs. pMCI single modal classification task, ADNI dataset.

	Acc	*Sens*	*Spec*	*AUC*
Single modal forest	0.744	0.663	0.791	0.779
PC forest	0.774	0.878	0.747	0.781
Single modal feature transform	0.781	0.691	0.844	0.819
Multimodal feature transform	0.788	0.747	0.805	0.837

Table 3
Accuracy (Acc), Sensitivity (Sens), Specificity (Spec) and Area under ROC curve (AUC) of the proposed methods and the baseline forest for the sMCI vs. AD single modal classification task, ADNI dataset.

	Acc	*Sens*	*Spec*	*AUC*
Single modal forest	0.731	0.824	0.699	0.814
PC forest	0.752	0.758	0.748	0.836
Single modal feature transform	0.782	0.734	0.863	0.868
Multimodal feature transform	0.795	0.737	0.897	0.892

Another classification problem worth investigating is discrimination of samples with stable MCI from dementia cases using the MRI feature set. Fig. 5 and Table 3 show ROC curves and performance measures of the enhanced and baseline forests for this classification task.

It can be seen that similar to the NL vs. pMCI task, the forests enhanced by the new feature sets are outperforming the baseline single modal forest. The improvement observed in the PC forest is more significant than the previous task but it still can not be considered statistically significant (p-value $= 0.08$). On the other hand the proposed tree-based feature transform methods significantly outperform the baseline methods with p-values of 0.0001 and 3.698e−07 for the single and multimodal feature transforms, respectively, and the multimodal feature transforms are more effective than single modal feature transforms ($p = 0.0003$).

The third classification task which separates sMCI from pMCI cases is potentially the most clinically relevant model. The ROC curves and performance measures for this task can be seen in Fig. 6 and Table 4.

The trends remain the same: the tree-based feature transforms outperform a simple single modal forest with $p = 0.01$ and and $p = 0.0002$ for the single modal (MRI-based) and multimodal (MRI+PET) feature transforms, respectively. It can also be seen that the PC-based features fail to enhance the baseline forest to a statistically significant level ($p = 0.672$). Similar to the previous experiments, the multimodal feature transforms yield a larger AUC than single modal feature transforms with $p = 0.01$.

Table 4
Accuracy (Acc), Sensitivity (Sens), Specificity (Spec) and Area under ROC curve (AUC) of the proposed methods and the baseline forest for the sMCI vs. pMCI single modal classification task, ADNI dataset.

	Acc	*Sens*	*Spec*	*AUC*
Single modal forest	0.743	0.713	0.744	0.810
PC forest	0.757	0.769	0.746	0.815
Single modal feature transform	0.777	0.819	0.750	0.848
Multimodal feature transform	0.815	0.831	0.803	0.872

5. Discussion

5.1. Scandent tree: limitations and future work

The prostate cancer dataset is an example of the worst case scenario of missing data: a large 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. It is, therefore, revealing that even in this situation, the use of scandent tree methodology provides a clear advantage against the traditional approaches to deal with a situation like this, such as simply ignoring one or the other set, or imputation approaches.

There are two limitations to our work with prostate cancer data. First, as our experiments show the missing modality (gene expression) is far more discriminative than the shared modality (MRI). This makes it extremely difficult for the proposed method to model the relationships between the modalities and effectively merge the two datasets. Second, the small number of features in the shared modality (MRI) makes the feature-bagging in the support tree unbalanced between the modalities. As a result, many of the support trees are completely grown based on the missing modality (gene expression) and scandent trees have to follow the

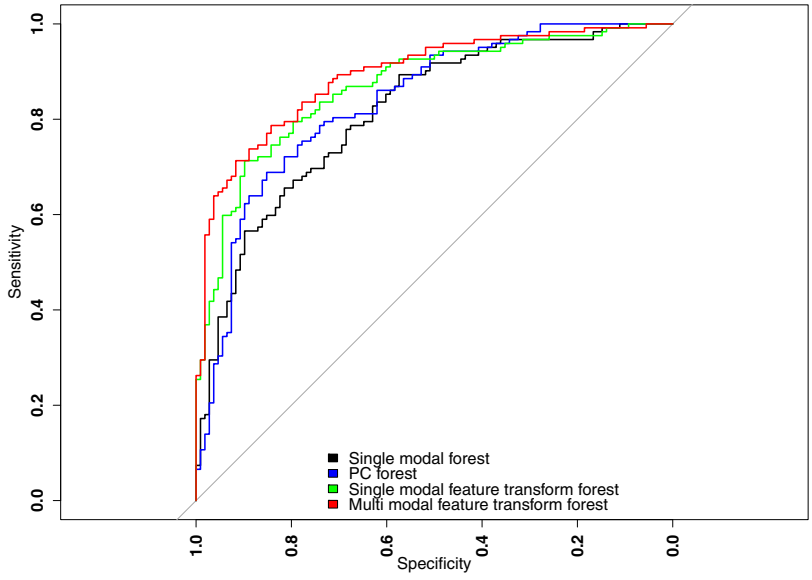

Fig. 5. ROC curve for stable MCI vs. AD classification, single modal classification task, ADNI dataset.

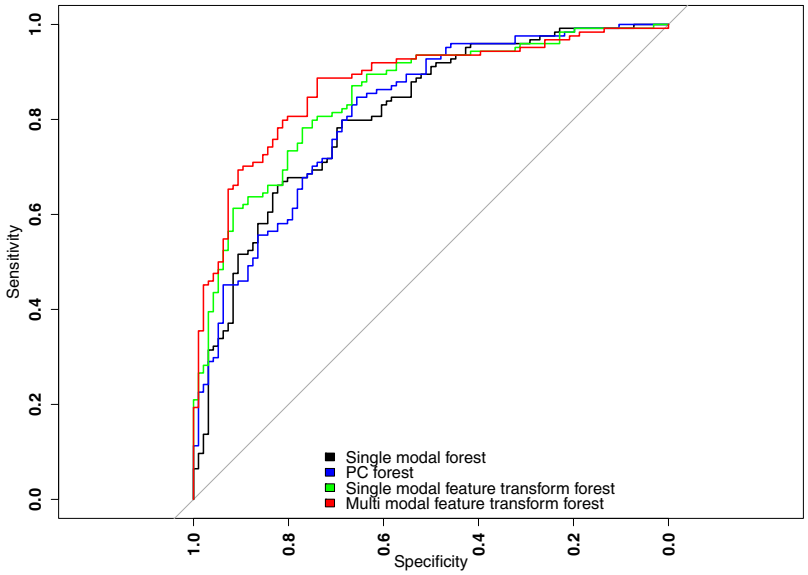

Fig. 6. ROC curve for stable MCI vs. progressive MCI classification, single modal classification task, ADNI dataset.

structure of a whole support tree. This together with small sample size of the multimodal dataset can cause over-fitting.

Considering these limitation of the prostate cancer study, one might question the value of training a complex model like a scandent tree using such a small dataset with unbalanced feature sets. Further tree-level investigations of the proposed model show that even in the extreme case of the prostate cancer dataset, 56.3% of the trees of the scandent forest were observed to outperform the corresponding support trees. We observed that because of the low sample size of the multimodal prostate cancer dataset, the inference algorithm is unable to optimize many of the leaves of the scandent forest and as a result most of the leaves of a scandent forest are not different from the support forest. This makes it hard to show the effect of the local trees separately from the inference method using the overall performance of a scandent forest.

However, on the leaves that have enough out of the bag samples needed for optimization, our observations show that the effect of local trees on the performance of the inference algorithm is evident. We also used a leaf-level error-rate calculated based on out of bag samples as a measure to compare the local trees method with a genetic-unaware partitioning of the samples. The results show that the inference method achieves lower error rates using the local trees and the reduction in error rates is statistically significant.

Given the limitations of the prostate cancer study, a more revealing test of the performance of the solution proposed for the multimodal scenario can be achieved in the study of benchmark datasets. One such study was presented in our MICCAI 2015 work where we examined the performance of the scandent tree method for different mutlimodal sample sizes and different feature sets using a heart disease benchmark dataset publicly available from the University of California Irvine (UCI) database (Lichman, 2013).

In comparison with the state of the art imputation method for decision forests (rfImpute), we observed that in larger multimodal sample sizes or when only a small number of features were missing from the single modal dataset, both of the methods perform very well in handling the missing values for multimodal classification. However, in smaller multimodal datasets or when a large portion of features are missing from the single modal samples, the scandent tree method showed significantly better performance in comparison with the rfImpute method. Another observation was that for a fixed sample size, the scandent tree method is less sensitive to the number of missing features, especially in smaller multimodal sample sizes. This advantage was also evident from the results on the prostate cancer dataset.

We can envision two future improvements to the implementation of the scandent tree method.

The first improvement would be optimization of the proposed multimodal classifier for computational efficiency. There are two computationally expensive steps in forming a scandent forest, forming the scandent trees and the inference. Forming the scandent tree is relatively complex in comparison to a conventional tree. Moreover, its computational cost is highly dependent on the relationships between feature sets. However, the computational cost of training a scandent tree can be neglected in comparison to the inference method. The current proposed inference method is based on the optimization of the scandent trees in a leaf by leaf manner. This is one of the main reasons for better performance of the scandent tree method in smaller sample size. However, the computational load of the inference step can be considerable in case of large multimodal datasets. This depends on many parameters besides the number of leaves of the tree. For instance, the number of bootstrapped re-samples performed at each leaf and the algorithm used for finding the optimum sample selection threshold, q. Considering all these parameters, a valuable direction for future work would be to redesign the proposed method for better computational efficiency with emphasis on designing of the inference method. Note that our proposed multimodal method keeps the testing cost the same as the baseline support tree. In contrast to the proposed multimodal classifier, the scandent forests used for single modal classification do not necessarily need the inference step. As a result, we do not see a need for redesigning the proposed single modal classifier. However, because the feature transforms are used in both training and testing steps, a more computationally efficient implementation of the feature transforms can be valuable.

The second area for continued work is improving the baseline trees. Because we needed full control over each division of each tree in the forest, we could not use the off-the-shelf decision forest packages available in R. Therefore, the support forest which is the base of the scandent tree method is our in-house implementation and can be improved.

Table 5
Comparison of the proposed single modal method with the state of the art for sMCI vs. pMCI prediction, ADNI dataset.

Method	Sample size	Modalities	Performance			
			Acc	*Sens*	*Spec*	*AUC*
Proposed method[a]	122	MRI	0.815	0.831	0.803	0.872
Proposed method[b]	357	MRI	0.759	0.688	0.774	0.737
(Cheng et al., 2015)	99	MRI, PET, CSF	0.801	0.853	0.733	0.852
(Suk et al., 2014)	204	MRI, PET	0.759	0.48	0.952	0.746
(Campos et al., 2015)	397	MRI, PET, CSF	0.732	0.655	0.767	0.786
(Eskildsen et al., 2013)	388	MRI	0.754	0.705	0.776	0.82
(Wee et al., 2013)	200	MRI	0.751	–	–	0.84
(Young et al., 2013)	143	MRI, PET, CSF, APOE	0.741	0.787	0.656	0.795
(Zhang and Shen, 2012)	91	MRI,PET,CSF	0.739	0.686	0.736	0.797
(Coupé et al., 2012)	405	MRI	0.71	0.7	0.72	–
(Westman et al., 2012)	162	MRI, CSF	0.685	0.741	0.63	0.76

[a] MCI:SMCI+EMCI
[b] MCI:SMCI+EMCI+LMCI

5.2. Scandent tree feature transforms

5.2.1. General discussions

We examined the robustness of the proposed method for different feature selection algorithms on the ADNI dataset, namely by using importance measure in a decision forest, in a C5.0 tree and p-value of a Pearson's Chi-squared test. We observed that the proposed method is robust to the choice of the importance measure but is sensitive to the size of the feature generating forests and the number of selected features. This suggests that the statistical independence between the features plays a more important role comparing to the choice of the feature selection technique. A valuable direction for feature work would be to design a method that in addition to ranking the features based on their discrimination power, eliminates the statistically dependent features.

One of the interesting results on the ADNI dataset is the advantage of the multimodal feature transform forest over the single modal feature transform forest. Considering the fact that multimodal feature transforms are optimized for mimicking the structure of the support tree, it might seem odd that they can be more useful than the single modal feature transforms optimized directly based on the outcome label. This can be justified by the fact that the divisions formed by the missing modality might guide the local tree to form feature transforms that could not be easily observable by a conventional single modal tree growth algorithms. These new feature-sets belong to the same optimization space of a single modal tree. Howevee, because each single modal tree is optimized based on step-wise sample space partitioning, the single modal forest might not easily converge to the features generated by the scandent forest.

Another interesting point of discussion is the kind of relationships that can be modeled by the scandent tree model. Naturally, statistical dependence between the features is assumed as a strict requirement for any relationship model to work. Although it cannot be guaranteed that the proposed method can handle any type of statistical dependence, it can be shown that in case of simple relationships like correlation between features, the local trees can be as effective as the other models used in state of the art imputation methods. However in case of modalities that do not have a trivial relationship, we believe that using the local trees might have some advantages. Two well-known examples of modalities that measure different quantities and do not have a trivial relationship in general are the MRI and PET modalities in the ADNI study.

For example, PET and MRI values are not trivially related. Although we know that the MRI values may not be predictable by PET values (and vice versa), we are hoping that they are not statistically independent. The local trees in each scandent tree avoid the prediction of the exact values and instead translate the knowledge of the missing modality into questions of the form "given that a patient belongs to a partition of the PET feature space, what is the probability that the patient belongs to a partition in the MRI feature space?". For instance, "if a patient is similar to another patient in the PET space, is it likely that these patients are also similar in the MRI space?". We are trying to show that sometimes, the answer to this question is sufficient for a more accurate classification and we do not need to predict exact values of a modality by the other one.

5.2.2. Comparison with other work on ADNI

We investigated the performance of the proposed method for single modal classification on the ADNI dataset. In this study we focused on the problem of leveraging the multimodal set of samples with both MRI and PET features for designing a single modal classifier that only needs MRI for classification. The solution presented here relies on the new concept of tree-based feature transforms. We showed that in all clinically relevant questions related to the ADNI dataset, the use of a the scandent trees as a means to define tree-based feature transforms based on both PET and MRI data, and using them along with the original MRI features for training and testing single modal data, results in an improved performance in comparison with methods that rely only on MRI features.

The block wise missing value problem is a well-known issue of the ADNI dataset and it is addressed in many papers in literature. However, none of them has the same goal and assumptions as our study. For instance, this paper focuses on improving the performance of decision forests with the assumption that a decision forest is the classifier of choice for a given multimodal dataset. However, most of the studies on the ADNI dataset use other classifiers like multi-kernel SVM for multimodal classification. As a result, it is difficult to compare our results with the available literature as any such comparison will be mostly informed by the choice of classification paradigm.

One other issue that makes the comparison difficult is the different feature sets and sample sizes impacted by patient selection criteria. A simple example is the different assumptions on the conversion time for MCI to AD for differentiating progressive versus stable MCI. In our study, we assumed a 36 month conversion time for progressive MCI cases and used the summarized set of features extracted by adnimerge R package as our feature set. This package is accessible from the ADNI website (https://adni.loni.usc.edu).

With all these differences and limitations in mind, we have gathered a list of comparable methods with performance measures reported in the literature in Table 5. These are all on the sMCI vs

pMCI classification task. We report the performance of the proposed method for two common choices of the MCI class, one consisting of MCI casses labeled as SMCI or EMCI and the other one consisting of MCI classes labeled as SMCI, EMCI or LMCI. As it can be seen, the proposed method matches or surpasses the performance of the state of the art, even in cases where multimodal data is available for all cases.

6. Conclusion

We propose the novel concept of scandent trees for enriching a multimodal classifier with large training dataset from a subset of modalities. The results show that the proposed method for multimodal classification outperforms the embedded missing value imputation method of decision forests introduced in Breiman (2001) and other state of the art imputation methods, particularly in smaller samples sizes and when a large portion of features are missing. We showed that the proposed method enables the integration of a small genomic plus imaging dataset, with a relatively large imaging dataset. We also descibe a novel learning method for training on multiple modalities and testing on one modality. To this end, we introduced the concept of tree-based feature transforms. We showed that using this approach, we can efficiently transfer the discriminative power of PET imaging into the training phase of building a model that would only use the MRI data at the testing phase.

Acknowledgments

Funding from Canadian Institutes of Health Research (CIHR, Operating Grant Priority Announcement, IC1-134055) and Natural Sciences and Engineering Research Council of Canada (Discovery Grant, RGPIN 435597-13) is acknowledged. 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/VGH; Dr. Elai Davicioni, Christine Buerki, Heesun Shin, and Zaid Haddad from GenomeDx Biosciences Inc.

ADNI Disclosure: Data collection and sharing for parts of this work was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimers Association; Alzheimers Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Table A.6
The Genes used in prostate cancer study.

Probe ID	Gene name	Probe ID	Gene name
2376037	MDM4	3947604	BIK
4008427	NUDT11	3956433	CHEK2
2436826	KCNN3	2887633	BOD1
2562343	GGCX	3968303	SHROOM2
3128411	EBF2	2920619	ARMC2
3761737	ZNF652	3286921	08-Mar
3754797	HNF1B	2934521	SLC22A3
2852766	AMACR	2852742	AMACR
2731257	AFM	2652027	CLDN11
2736322	PDLIM5	2949901	NOTCH4
3127978	NKX3-1	3043264	JAZF1
2484970	EHBP1	3349660	HTR3B
2845829	TERT	3359180	TH
3739668	VPS53	3739679	VPS53
2738146	TET2	3014159	LMTK2
2536531	FARP2	3338060	MYEOV
3839538	KLK3	3049522	TNS3
2417390	CTBP2	2469157	GRHL1
3311417	CTBP2	2636483	SIDT1
3413787	TUBA1C		

Appendix A. Prostate cancer dataset

The 39 genes used as biomarkers for the prostate cancer study are listed in Table A.6.

References

Ashab, H.A., Kozlowski, P., Goldenberg, S.L., Moradi, M., 2014. Solutions for missing parameters in computer-aided diagnosis with multiparametric imaging data. In: Machine Learning in Medical Imaging. Springer, pp. 289–296.

Breiman, L., 2001. Random forests. Machine learning 45 (1), 5–32.

Campos, S., Pizarro, L., Valle, C., Gray, K.R., Rueckert, D., Allende, H., 2015. Evaluating imputation techniques for missing data in adni: a patient classification study. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Springer, pp. 3–10.

Cao, Y., Wang, H., Moradi, M., Prasanna, P., Syeda-Mahmood, T.F., 2015. Fracture detection in x-ray images through stacked random forests feature fusion. In: Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on. IEEE, pp. 801–805.

Cheng, B., Liu, M., Suk, H.-I., Shen, D., Zhang, D., 2015. Multimodal manifold-regularized transfer learning for MCI conversion prediction. Brain Imag. Behav. 1–14.

Coupé, P., Eskildsen, S.F., Manjón, J.V., Fonov, V.S., Pruessner, J.C., Allard, M., Collins, D.L., 2012. Scoring by nonlocal image patch estimator for early detection of Alzheimer's disease. NeuroImage 1 (1), 141–152.

Drew, B., Jones, E.C., Reinsberg, S., et al., 2010. Device for sectioning prostatectomy specimens to facilitate comparison between histology and in vivo MRI. J. Magnetic Resonance Imag. 32, 992–996.

Eskildsen, S.F., Coupé, P., García-Lorenzo, D., Fonov, V., Pruessner, J.C., Collins, D.L., 2013. Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. Neuroimage 65, 511–521.

Haq, N.F., Kozlowski, P., Jones, E.C., Chang, S.D., Goldenberg, S.L., Moradi, M., 2015. A data-driven approach to prostate cancer detection from dynamic contrast enhanced MRI. Comput. Med. Imag. Graph. 41, 37–45.

Hor, S., Moradi, M., 2015. Scandent tree: A random forest learning method for incomplete multimodal datasets. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015. Springer, pp. 694–701.

Jie, B., Zhang, D., Cheng, B., Shen, D., 2015. Manifold regularized multitask feature learning for multimodality disease classification. Human Brain Map. 36 (2), 489–507.

Lichman, M., 2013. UCI machine learning repository.

Moradi, M., Salcudean, S.E., Chang, S.D., Jones, E.C., Buchan, N., Casey, R.G., Goldenberg, S.L., Kozlowski, P., 2012. Multiparametric MRI maps for detection and grading of dominant prostate tumors. J. Magnetic Resonance Imag. 35 (6), 1403–1413.

National Institutes of Health, National cancer institute: PDQ genetics of prostate cancer, Date last modified 02/20/2015.

Quinlan, J.R., 2014. C4. 5: Programs for Machine Learning. Elsevier.

Steinberg, D., Colla, P., 2009. Cart: classification and regression trees. The Top Ten Algorithms in Data Mining 9, 179.

Suk, H.-I., Lee, S.-W., Shen, D., et al., 2014. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. NeuroImage 101, 569–582.

Therneau, T.M., Atkinson, B., Ripley, B., 2010a. rpart: Recursive partitioning. R package version 3.1-46. Ported to R by Brian Ripley. 3.

Therneau, T.M., Atkinson, B., Ripley, B., et al., 2010b. rpart: Recursive partitioning. R Package Version 3, 1–46.

Tu, Z., Bai, X., 2010. Auto-context and its application to high-level vision tasks and 3d brain image segmentation. Pattern Analysis and Machine Intelligence, IEEE Transactions on 32 (10), 1744–1757.

Wee, C.-Y., Yap, P.-T., Shen, D., 2013. Prediction of Alzheimer's disease and mild cognitive impairment using cortical morphological patterns. Human Brain Map. 34 (12), 3411–3425.

Westman, E., Muehlboeck, J.-S., Simmons, A., 2012. Combining mri and csf measures for classification of Alzheimer's disease and prediction of mild cognitive impairment conversion. Neuroimage 62 (1), 229–238.

Xiang, S., Yuan, L., Fan, W., Wang, Y., Thompson, P.M., Ye, J., 2013. Multi-source learning with block-wise missing data for Alzheimer's disease prediction. In: Proceedings of the 19th ACM SIGKDD International conference on Knowledge Discovery and Data Mining. ACM, pp. 185–193.

Xiang, S., Yuan, L., Fan, W., Wang, Y., Thompson, P.M., Ye, J., 2014. Bi-level multi--source learning for heterogeneous block-wise missing data. NeuroImage 102, 192–206.

Young, J., Modat, M., Cardoso, M.J., Mendelson, A., Cash, D., Ourselin, S., 2013. Accurate multimodal probabilistic prediction of conversion to alzheimer's disease in patients with mild cognitive impairment. NeuroImage 2, 735–745.

Yu, G., Liu, Y., Thung, K.-H., Shen, D., 2014. Multi-task linear programming discriminant analysis for the identification of progressive MCI individuals. PLOS One 9, e96458.

Yuan, L., Wang, Y., Thompson, P.M., Narayan, V.A., Ye, J., 2012. Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. Neuroimage 61 (3), 622–632.

Zhang, D., Shen, D., 2012. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. Neuroimage 59 (2), 895–907.