Identifying Parkinson's Patients: A Functional Gradient Boosting Approach

Devendra Singh Dhami[1], **Ameet Soni**[2], **David Page**[3], and **Sriraam Natarajan**[1]

[1]Indiana University Bloomington

[2]Swarthmore College

[3]University of Wisconsin-Madison

Abstract

Parkinson's, a progressive neural disorder, is difficult to identify due to the hidden nature of the symptoms associated. We present a machine learning approach that uses a definite set of features obtained from the Parkinsons Progression Markers Initiative(PPMI) study as input and classifies them into one of two classes: PD(Parkinson's disease) and HC(Healthy Control). As far as we know this is the first work in applying machine learning algorithms for classifying patients with Parkinson's disease with the involvement of domain expert during the feature selection process. We evaluate our approach on 1194 patients acquired from Parkinsons Progression Markers Initiative and show that it achieves a state-of-the-art performance with minimal feature engineering.

1 Introduction

We consider the problem of predicting the incidence of Parkison's disease, a progressive neural disorder. Specifically, we consider the data collected as part of the Parkisons Progression Marker Initiative (PPMI) and aim to predict if a subject has Parkinson's based on clinical data – particularly, motor assessments (motor functions) and non-motor assessments (neurobehavioral and neuropsychological tests). One of the most important challenges for this task is that there seem to be no real strong indicator that explains the progression clearly [1].

Our hypothesis, which we verify empirically is that instead of considering a small set of strongly influencing risk factors, it might be more effective to consider a large set of weakly influencing risk factors. To this effect, we adapt the recently successful gradient-boosting algorithm [2] for this task. We exploit the use of a domain expert in identifying the right set of features and consider learning from the longitudinal data. Unlike standard methods that require projecting the data to a feature vector format (using what are called aggregation or propositionalization methods), our proposed approach models the data faithfully using time as a parameter of a logical representation. We evaluate our proposed approach on ≈ 1200

patients from the PPMI study and demonstrate the effectiveness and efficiency of the proposed approach.

We make the following key contributions: we consider the challenging task of predicting Parkinson's from 37 different features. These features were chosen by interacting with a domain expert. The key advantage of this approach is that it models the underlying data faithfully by utilizing logic based framework. Evaluation on the PPMI dataset shows that our approach is superior to standard classification aproaches.

2 Background

Parkinsons Progression Markers Initiative(PPMI) is an observational study with the main aim of identifying features or biomarkers that impact Parkinsons disease progression [1]. The collected data can be divided broadly into four distinct categories: Imaging data, Clinical data, Biospecimens and Subject demographic data. We focus primarily on the clinical data which mainly consists of motor assessments and non-motor assessments. Since Parkinson's mainly affect the motor system (i.e. the part of the nervous system associated with movement) and the initial symptoms are mostly movement related, using motor assessment data seems a natural approach. A subset of the clinical data is selected based on the expert input after which a set of 37 features are obtained which can be broadly defined in these categories: (1) Motor-UPDRS (consists of 34 features) [4], (2) Montreal Cognitive Assessment (MoCA), a non-motor assessment that contains the MoCA score of a patient for a single visit and the difference in MoCA scores of the patient from the last visit, (3) total UPDRS score.

MoCA consists of series of questions assessing various parameters of a subject such as the visual capability, capacity of recognizing objects, the attention span and memorizing words to name a few. Each of the questions are scored with the total score being 30. A subject with score of 26 is considered to be normal. Unified Parkinson Disease Rating Scale (UPDRS) is a rating scale used for determining the extent of Parkinson's disease progression in a subject and each assessment ranges from 0 to 4, with 0 being normal behavior and 4 representing severe abnormal behavior with respect to the assessment. Motor-UPDRS refers to the value of motor assessments in the UPDRS scale. Total UPDRS score refers to the sum of all the motor-UPDRS features.

Our goal is to estimate the conditional distribution – $P(par|\mathbf{x})$ where $\mathbf{x}$ represents the set of motor and non-motor assessments (i.e., features) and par denotes the incidence of Parkinson's disease for the particular patient. One could apply any machine learning algorithm for learning this distribution. We focus on the gradient-boosting technique which has had success in medical applications [6]. Gradient-boosting is a gradient-ascent technique performed on the functional space. For probabilistic models, gradient-boosting represents the conditional distributions using a functional representation, for instance a *sigmoid* function. Then the gradients are calculated w.r.t to this function. For instance, one could represent $P(par|\mathbf{x}) = \frac{e^{\psi(y|\mathbf{x})}}{1+e^{\psi(y|\mathbf{x})}}$. Friedman [2] suggested that instead of obtaining the gradients w.r.t the global function ψ, one could obtain the gradient for each example $\langle \mathbf{x}_i, y_i \rangle$,

where y denotes the target i.e., presence of Parkinson's. The key idea is that these gradients are good approximations of the overall gradients because the direction of these gradients point to the same direction as that of the true gradient. This functional gradient of likelihood w.r.t. ψ for each example i is $\frac{\partial log(P(y_i|\mathbf{x}_i))}{\partial \psi(P(y_i|\mathbf{x}_i))}=I(y_i=1) - P((y_i=1|\mathbf{x}_i))$ where I is the indicator function (equal to 1 if $y_i = 1$ and equal to 0 if $y_i = 0$). The gradient is the difference between the true label and the current prediction and is positive for positive examples and negative for negative examples. In simpler terms, the negative examples are pushed towards 0 and the positive examples are pushed to 1 resulting in a well defined decision boundary. The ψ function is represented by relational regression trees (RRT) [3] which uses the relational features as input to the trees that can be easily guided by the domain expert who could provide preference information.

3 Proposed Approach

Since the aim of the PPMI study is to determine the most predictive features for the disease progression, the involvement of a domain expert during the feature selection process becomes beneficial. Our specific aim is to predict the incidence of Parkinson's in a patient. Since the data is longitudinal in nature, it is important that we model time faithfully. We view time as a special type of relation and hence we create features in the predicate logic format as *feature name(patient id, time, feature value)*. This allows the number of days since the start of the study (time) to be an argument of the predicate.

The first step of the process starts with the correlation analysis of the raw data obtained. The raw clinical data consists of 81 features. A correlation matrix is constructed with each entry representing the correlation coefficient (we use the *Pearson correlation coefficient*) between each variable and the others. The 50 features with low mutual correlation and high correlation to the class label are selected. The second step consists of the expert evaluating the obtained features giving us a further pruned set of 37 features and thus the final data to be used for the classification task. The key reason for considering a relational representation is two fold: First, relational models allow learning using the natural representation of the data and second, data is longitudinal i.e. a patient has multiple entries in the data. Propositional classifiers have limitations in learning such data (require aggregators).

The learner is provided with the training data which it uses to learn a relational regression tree. The last step is the prediction phase where the learned model can be queried to return the probability of the target being true given the evidence. Since all the evidence is observed, inference requires simply querying all the relational regression trees, summing up their regression values and returning the posterior estimates i.e. the probability that the given test example belongs to the positive class.

Interpretability

One key limitation of the proposed approach is the interpretability of the final model. While each boosted tree in itself is interpretable, given that they are not learned independently of each other other makes the model difficult to interpret. To make the model comprehensible, we take an approximate approach that we call the Craven approach [5] which was originally

developed for making neural networks interpretable. The key idea is to relabel the training data based on the boosted model that we have learned and then train an overfitted tree to this labeled data. The intuition is that this new large tree will represent the decisions made by the original set of trees due to its performance on the training data. Recall that our original training data consists of Boolean labels (Parkinson's vs negative). But the relabeled data consists of regression values that are being learned in the new tree. Hence, the resulting tree is closer to the original learned model as we show in our experiments.

4 Experiments

In our empirical evaluations, we aim to explicitly ask the following questions:

> **Q1:** How effective is the feature selection with expert in predicting Parkinson's?

> **Q2:** Given the longitudinal nature of the data, is our method more effective than using standard classifiers in this prediction task?

Table 1 shows the result of learning from 50 features obtained after correlation and 37 features after the expert advice. This helps us in answering **Q1** affirmatively. Across all scoring metrics, expertly selected features outperforms models built using the larger feature set.

We compare our method, BoostPark, to three propositional classifiers: Logistic Regression, Gradient-Boosting and Support Vector Machines. The propositional data is aggregated using three aggregator functions: min, max and mean over time. Our data consists of records for 1194 patients, with 378 positive examples (i.e. Parkinson's patients) and 816 negative examples. Regression trees are learned on the given data which form the training model. We perform 10-fold cross validation and present the results.

Since we aggregate the propositional data using 3 aggregators, the best performing aggregator for all the propositional classifiers is selected and compared to our methods BoostPark10 and BoostPark20 as shown in table 2. Our methods perform considerably better than the propositional counterparts in terms of AUC-ROC and AUC-PR and performs equally well in terms of accuracy. This helps answer **Q2** positively. Our method is more effective than the standard classifiers in this prediction task.

5 Conclusion

Identifying important features responsible for the progression in Parkinson's disease in a patient remains a compelling challenge. We use a human domain expert to guide our method with identifying a relatively large set of influencing risk factors. We then present a learning method that can consider this large set of weak influences in learning a probabilistic model. We evaluated our results on the PPMI data and demonstrated that our learning approach outperforms standard machine learning classifiers. Since Parkinson's is a progressive disease, developing robust temporal models for this task remains an interesting challenge. Extending our learning algorithm to handle hybrid temporal models will allow for modeling the continuous data more faithfully. Finally, scaling up the learning algorithm to learn from a

broader set of population data rather than be limited by a single study remains an interesting open problem.

References

1. Marek, Kenneth, et al. The parkinson progression marker initiative (PPMI). Progress in neurobiology. 2011; 95(4)

2. Friedman, Jerome H. Greedy function approximation: a gradient boosting machine. Annals of statistics. 2001

3. Blockeel, Hendrik, Raedt, Luc De. Top-down induction of first-order logical decision trees. Artificial intelligence. 1998; 101(1)

4. Goetz, Christopher G., et al. Movement Disorder Societysponsored revision of the Unified Parkinson's Disease Rating Scale (MDSUPDRS): Process, format, and clinimetric testing plan. Movement Disorders. 2007; 22(1)

5. Craven, Mark W., Shavlik, Jude W. Extracting tree-structured representations of trained networks. Advances in neural information processing systems. 1996

6. Natarajan, Sriraam, et al. Early Prediction of Coronary Artery Calcification Levels Using Machine Learning. IAAI. 2013
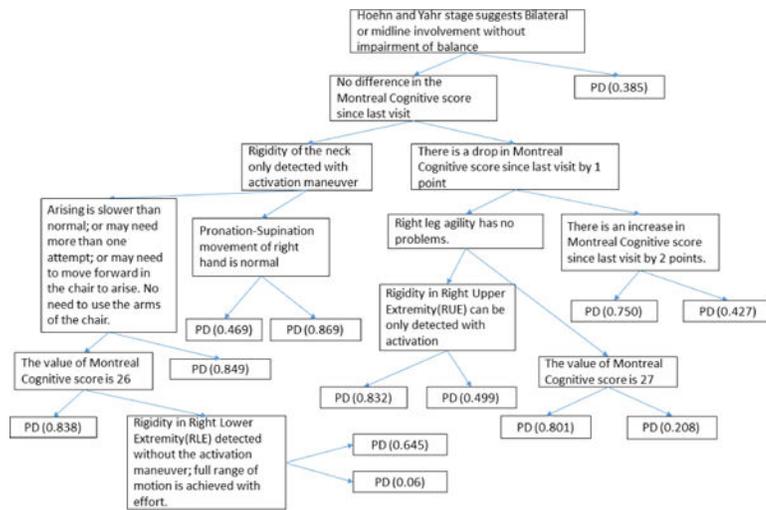
Fig. 1.
Combined tree learnt with BoostPark10.

Author Manuscript

Author Manuscript

Table 1

Results for BoostPark with and without expert advice

Classifier	With Expert				Without Expert			
	Accuracy	AUC-ROC	AUC-PR	F1 score	Accuracy	AUC-ROC	AUC-PR	F1 score
BoostPark10	0.889	0.973	0.937	0.808	0.854	0.932	0.9	0.797
BoostPark20	0.901	0.977	0.947	0.851	0.881	0.94	0.87	0.832

Table 2

Classifier results. Only the best classifiers among the aggregators are shown.

Classifier	Accuracy	AUC-ROC	AUC-PR	F1 score
BoostPark10 (10 relational regression trees)	0.889	**0.973**	**0.937**	0.808
BoostPark20 (20 relational regression trees)	0.901	**0.977**	**0.947**	0.851
Gradient Boosting(aggregator = mean)	**0.920**	0.914	0.904	**0.885**
Logistic Regression(aggregator = max)	0.903	0.896	0.884	0.862
Support Vector Machine(kernel = linear, aggregator = max)	0.897	0.892	0.877	0.855
Support Vector Machine(kernel = polynomial, aggregator = max)	0.895	0.883	0.875	0.848
Support Vector Machine(kernel = RBF, aggregator = mean)	0.757	0.806	0.784	0.734