A LATENT VARIABLE-BASED BAYESIAN REGRESSION TO ADDRESS RECORDING REPLICATIONS IN PARKINSON'S DISEASE

Pérez[*], *C. J., Naranjo*[*], *L., Martín*[*], *J., and Campos-Roca*[◇], *Y.*

[*]Department of Mathematics, University of Extremadura, Cáceres, Spain
[◇]Department of Computer and Communication Technologies, University of Extremadura, Cáceres, Spain

ABSTRACT

Subject-based approaches are proposed to automatically discriminate healthy people from those with Parkinson's Disease (PD) by using speech recordings. These approaches have been applied to one of the most used PD datasets, which contains repeated measurements in an imbalanced design. Most of the published methodologies applied to perform classification from this dataset fail to account for the dependent nature of the data. This fact artificially increases the sample size and leads to a diffuse criterion to define which subject is suffering from PD. The first proposed approach is based on data aggregation. This reduces the sample size, but defines a clear criterion to discriminate subjects. The second one handles repeated measurements by introducing latent variables in a Bayesian logistic regression framework. The proposed approaches are conceptually simple and easy to implement.

***Index Terms*—** Bayesian logistic regression, Data aggregation, Latent variable, Machine learning, Parkinson's disease, Voice features.

1. INTRODUCTION

Parkinson's disease (PD) is the second most common neurodegenerative disorder after Alzheimer's disease, affecting one in every 100 persons above the age of 65 years [1]. Depletion of dopaminergic nigrostriatal neurons gives rise to alterations in movement (tremor, rigidity, slow movements and/or unstable posture). Voice and speech, as dependent on movement of the articulators, are not spared. Non-dopaminergic changes can also affect language, cognition and mood, which can impact on communication [2].

Voice recordings have been used as a potential biomarker to diagnose some voice-related diseases. [3] provide a current view of automatic speech signal analysis for clinical diagnosis and assessment of speech disorders. In this context, [4] considered both linear and non-linear measures to discriminate healthy people from those with PD. Voice recordings have also been used to assess the progression of PD by relating voice characteristics to clinicians' ratings [5].

Telephone monitoring of PD has attracted interest as a potential mean of assessing this disorder. The current technologies allow the implementation of non-invasive and low cost procedures that make the assessment of PD easier both for patients and doctors. Purposeful-built devices have been developed to record various signals which can be associated with PD symptom severity [6]. Such technology also opens up the possibility of not just diagnosis and assessment, but also therapy being remotely performed [7].

Three steps are involved to discriminate healthy people from those with PD. Firstly, algorithms to extract features from the recordings must be used. Later, a suitable set of features must be selected. These two steps constitute the preprocessing stage. Finally, pattern recognition algorithms must be used to classify new individuals. The success of the classification highly depends on the good discriminatory properties of the selected features.

[4] presented one of the most used PD datasets consisting on 22 features extracted from 195 recordings of sustained /a/ phonations. These recordings belong to 32 people from both sexes, 24 of which were diagnosed with PD. Seven recordings were obtained from three subjects (S21, S27 and S35) and 6 from the others, leading to an imbalanced design. This dataset is available online at UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets/Parkinsons).

We have reviewed more than thirty different papers considering these data for feature selection and classification or only for classification. [8] compare their proposals with the ones from fifteen previously published papers that use these data. Very different overall accuracy rates were obtained depending on many factors, i.e., used features, reduction on features, classification methods and validation schemes. A common point among all the used approaches is that they are based on independent sample schemes instead of on repeated measurement frameworks. Each subject has six (or seven) related measures which are not independent. Treating the measures obtained from different recordings of the same subject as independent may produce wrong estimated parameters in

Thanks to M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig for distributing the Parkinson dataset. This research has been supported by *Ministerio de Economía y Competitividad*, Spain (Project MTM2011-28983-C03-02), *Gobierno de Extremadura*, Spain (Projects GRU10110 and GRU10126) and *European Union* (European Regional Development Funds).

many independence-based classifiers, which leads to wrong classifications. Besides, this treatment of the data artificially increases the sample size. Even more, incoherences are obtained when classifying subjects by using replicated samples, because it happens that some recordings of the same person can be classified as healthy and some as PD.

In this paper, subject-based approaches are considered to discriminate healthy people from those with PD. A discussion on the way the replications can be properly treated is provided. The idea of using latent variables in a Bayesian logistic regression model is used to provide a predictive model that can handle replications in an efficient way.

2. BACKGROUND AND MOTIVATION

Building a predictive model with minimal bias is intended in this context, i.e., to maximize the generalization of the predictions so as to perform well with new samples. [4] used a Super Vector Machine (SVM) classifier and also performed an exhaustive search to select the optimal subset of features. They found that the combination of the features Harmonic-to-Noise ratio (HNR), Recurrence Period Density Entropy (RPDE), Detrended Fluctuation Analysis (DFA), and Pitch Period Entropy (PPE) obtained the best overall classification performance. Their best model provided a 91.4% accuracy rate based on a bootstrap resampling. However, this does not reflect a true prediction accuracy rate for future observations, since the model is trained and tested on the same sample [9] and the independence condition for the bootstrap resampling is not met [10].

Later, this dataset has been extensively used to analyze the global accuracy rates for many classifiers in different cross-validation schemes as one-leave-out, k-fold or stratified sampling (e.g., [8, 11]). Each subject has six (or seven) related measures which are not independent. Independence-based classification methods should not be used when data have been obtained by replicating recordings from the same subjects. This fact artificially increases the sample size. Even more, this leads to a diffuse criterion to decide when a subject should be classified as suffering from PD. For example, if the 195 22-dimensional vectors are used as training and testing datasets for a simple logistic regression, it is obtained that 4 out of 24 PD subjects and 4 out of 8 healthy subjects have different predictions in their own recordings. This means 25% of the subjects (50% of the healthy and 16.7% of the PD) have incoherence in any of their recording predictions. Note that in this case the global accuracy rate considering the recordings as independent would be 89.7% (72.9% for healthy and 95.2% for PD people).

Although applying cross-validation schemes is a usual practice in this context, neither the recordings in the training set nor in the testing set are independent. [12] noticed that the traditional cross-validation methods divide recordings from the same individual in the training set and testing set, creating an artificial situation untypical of a real testing scenario. They defined an adapted cross-validation method named one-leave-individual-out. All the recordings of one individual are used for the testing set whereas all the recordings from the remaining individuals are used for the training set. This is performed for all the subjects and the accuracy rates are averaged. Nevertheless, the underlying independence problem remains.

The next sections discuss and propose two subject-based approaches that solve this problem.

3. DATA AGGREGATION

[13] observed that these replicated data can not be treated with the traditional machine learning algorithms, since the data nature is dependent. They proposed to aggregate related data before learning by using some different functions as mean, minimum, maximum or a linear trend prediction. They compared their results to the ones obtained with the original dataset and obtained better global accuracy rates for 6 out of 14 classifiers. However, it must be taken into account that the sampling size was reduced from 195 to 32. Then, data aggregation is considered as a preprocessing step and, later, any machine learning algorithm based on independent samples may be used. This avoids the problem of defining which subject is healthy or not as it happens with the recordings.

There are many other functions that can be used to aggregate replicated data. For example, median is an interesting robust statistics for central tendency. Another option is to aggregate data by using the $\alpha-$trimmed mean, which is a hybrid of the mean and the median. The basic idea is to order all the elements and discard $\alpha/2 \cdot 100\%$ of the elements at the beginning and $\alpha/2 \cdot 100\%$ at the end, then calculate an average value using the remaining ones. Both median and $\alpha-$trimmed mean are less sensitive to extreme values than the arithmetic mean. Here, an experiment considering the mean (A1), the median (A2), and the $1/3-$trimmed mean (A3) as functions to aggregate data is performed. In this case, the $1/3-$trimmed mean discards the lowest and the largest measurement of each individual. Feature selection is not an objective in this paper, so the best four voice characteristics obtained in [4] are considered here, i.e., HNR, RPDE, DFA, and PPE.

Previously to the data aggregation the within-subjects variability is analyzed. Generalized linear mixed models for repeated measures are applied to the four voice characteristics. One between-subjects variable (status) and one within-subjects variable (repeated measurements) are considered. The interaction is not statistically significant for HNR, RPDE, DFA, and PPE (the p-values are 0.367, 0.230, 0.707, and 0.935, respectively). This suggests that the measurements across the replications are independent on the health status. Then, the main effects are analyzed. The main effect for the within-subjects variable is not statistically significant for the four feature variables (the p-values are 0.514, 0.255, 0.269, and 0.375, respectively). This means that the variability in the measurements within subjects is small for the four voice characteristics. Using these repeated measures within subjects as if they were independent leads to an artificially increased sampling size that would reduce the true variance with respect to an independent sample with the same sample size. The between-subjects effects indicate that Parkinson's patients provided higher means than healthy people for PPE and lower for HNR (the p-values are 0.029 and $2.76 \cdot 10^{-5}$, respectively). No significant differences were found for RPDE and DFA at a 0.05 level (the p-values are 0.057 and 0.177, respectively).

A cross-validation scheme is applied to the original and transformed data considering a stratified sampling to choose 75% for the training sample and 25% for the testing sample. Note that the three aggregated schemes learn from 18 PD and 6 healthy individuals, and are applied to 6 and 2, respectively. Averages from 100 iterations are obtained by using several classifiers with the default parameters of WEKA [14]. Note that each classifier could be tuned to their optimal (or near optimal) parameters and some changes on the results may be experienced. Table 1 shows the averaged accuracy rates.

Classifiers	A1	A2	A3	Original
Bayes Net (K2)	77.63	78.13	78.13	80.45
Bayes Net (TAN)	78.13	78.13	78.13	83.80
Naive Bayes	74.88	73.00	76.25	78.81
SVM-SMO	75.00	75.00	74.88	76.07
DTNB	77.50	77.88	77.88	83.96
One R	71.88	72.88	71.75	86.82
Zero R	75.00	75.00	75.00	75.39
PART	79.13	79.00	79.13	85.38
Decision Table	77.63	78.00	78.00	84.12
Decision Stump	78.13	78.50	78.50	82.19
J48	79.38	79.38	79.50	85.30
NBTREE	77.50	78.13	78.13	84.87
Random Forest	77.25	76.63	76.38	88.29
Simple Cart	79.50	81.25	82.00	85.17
Logistic	88.25	87.63	87.75	86.79
Simple Logistic	87.88	87.50	87.63	85.88
MultiPerceptron	85.88	85.63	84.50	87.21
RBF Network	73.75	74.25	73.38	84.68
IB1	70.38	72.38	72.63	87.08
IB5	77.88	78.00	79.50	85.83
Kstar	76.38	74.13	72.75	87.81
ClassClustering	62.38	62.13	62.00	66.43
ClassRegression	81.63	83.50	82.75	86.95
HyperPipes	75.63	77.38	80.88	79.61
FVI	75.25	77.00	71.38	61.80

Table 1. Global accuracy rates for aggregated and original datasets.

The three aggregated schemes provide similar accuracy rates because of the homogeneity of the replicated measurements. The results provided by logistic, simple logistic and multiple perceptron (neural networks) with the WEKA default parameters are remarkable. The results are good since models are learning from only 24 individuals (18 PD and 6 healthy) and tested on 8 ones (6 PD and 2 healthy). Note that the classifiers applied to the original data learn from 146 recordings and test on 49 ones, and different predictions may be associated to the same individual. Then, the sampling size has been artificially increased and the classifiers provide no logical interpretations, since they are based on the recordings and not on the subjects.

The information provided from replications can be used in a specific repeated measurement design. The next section proposes and applies an alternative modelling method to address repeated measurements in this context.

4. ALTERNATIVE MODELLING

A logistic regression model with latent variables is introduced to address repeated measurements in a classification context from a Bayesian viewpoint.

Let Y be the response variable, X_k, $k = 1, \ldots, K$, the measured variables (covariates), and Z_k, $k = 1, \ldots, K$, the (unknown) latent variables. The latent variables are used in the logistic regression as if they were observed, and then they are imputed through the relationship with the repeated measurements and the prior distribution, i.e.:

$$\begin{aligned}
\mathrm{Y}_i &\sim \mathrm{Bernoulli}(p_i), \\
\mathrm{logit}(p_i) &= \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \cdots + \beta_k Z_{ki}, \\
\mathrm{X}_{kij} &= Z_{ki} + \varepsilon_{kij}, \\
\varepsilon_{kij} &\sim \mathrm{Normal}(0, \sigma_k^2), \\
Z_{ki} &\sim \mathcal{F}_{Z_{ki}}, \\
\beta_k &\sim \mathcal{F}_{\beta_k}, \\
\sigma_k &\sim \mathcal{F}_{\sigma_k},
\end{aligned}$$

where $i = 1, \ldots, n$, denotes the subject, $j = 1, \ldots, T_i$, denotes the repeated measures for subject i, $k = 1, \ldots, K$, denotes the covariate, p_i is the proportion parameter, and ε_{kij} is the error parameter. $\mathcal{F}_{Z_{ki}}$, $\mathcal{F}_{\beta_k}$, and $\mathcal{F}_{\sigma_k}$ represent generic initial distributions for the latent variables, the regression parameters, and the variance parameter of the error, respectively.

In Bayesian methodology, the initial knowledge about the parameters (prior distribution) is combined with the model considering the observed data (likelihood) to provide the posterior distribution. The posterior distribution contains all the information about the model parameters. This methodology allows that the initial information from historical data or experts can be included in the model through the prior distribution. This can be a great advantage when information different from the current data is obtained. If no information is available, flat distributions can be used instead.

This approach uses the relationship between the covariates and the latent variables jointly with the prior distributions to achieve posterior estimations of the latent variables. The approach considers the latent variables as missing data and provides imputations from the distribution conditioned on the observed variables and the parameters. The estimations from the latent variables are used to estimate the regression parameters in the logistic model. The variability among the repeated measures is taken into account by following this procedure.

This general method is applied to the problem in hand by considering the four features HNR (X_1), RPDE (X_2), DFA (X_3), PPE (X_4), and the status (Y) as response variable (0 healthy and 1 PD). In this case, no historical or expert information has been obtained independently of the data. Then, the following prior distributions are considered:

$$\begin{aligned} Z_{ki} &\sim \mathrm{Normal}(0, 100), \\ \beta_k &\sim \mathrm{Normal}(0, 100), \\ \sigma_k &\sim \mathrm{Unif}(0, 10). \end{aligned}$$

The posterior estimates are obtained by using Markov Chain Monte Carlo (MCMC) methods [15]. WinBUGS software has been used to implement the MCMC simulations [16]. This procedure has been implemented with the same specifications described in the previous section for the cross-validation scheme with 100 iterations. The averaged results are presented in Table 2. The following notation is followed: TP (True Positive), TN (True Negative), FP (False Positive), FN (False Negative), and $n = 32$.

		Mean	SD
Accuracy rate	(TN+TP)/n	0.804	0.102
Sensitivity	TP/(TP+FN)	0.965	0.090
Specificity	TN/(TN+FP)	0.320	0.314
Precision	TP/(TP+FP)	0.815	0.080

Table 2. Accuracy rate and other indicators for the approach with four covariates.

This approach is providing an averaged accuracy rate of 0.804, which is better than 20 (out of 25) classifiers analyzed by aggregating data in the same cross-validation scheme. Note that there is a low specificity (0.320) with a high standard deviation (0.314). This is due to the fact that, for each iteration, there is only 6 healthy subjects in the training set and 2 in the testing test. The results obtained in Table 2 can be improved by considering interactions between the covariates. The fact that the observed variables are related suggests that the obtained results could be improved by considering second-order interactions related to the highly correlated covariates HNR, RPDE, and PPE. In this case, the following modification is performed:

$$\begin{aligned} \mathrm{logit}(p_i) &= \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \beta_3 Z_{3i} + \beta_4 Z_{4i} + \\ &\quad \beta_5 Z_{1i} Z_{2i} + \beta_6 Z_{1i} Z_{4i} + \beta_7 Z_{2i} Z_{4i}. \end{aligned}$$

Following the same cross-validation scheme used through the paper, the results are presented in Table 3. The WinBUGS code is presented in the Appendix.

		Mean	SD
Accuracy rate	(TN+TP)/n	0.904	0.087
Sensitivity	TP/(TP+FN)	0.972	0.079
Specificity	TN/(TN+FP)	0.700	0.326
Precision	TP/(TP+FP)	0.916	0.089

Table 3. Accuracy rate and other indicators for the model with four covariates and three second order interactions.

By considering interactions, the accuracy rate has increased to 0.904, which is a very high rate for an experiment with this small sample size. Sensitivity maintains at the same level, whereas precision substantially increases. The specificity is dramatically increased from 0.320 to 0.700. Its standard deviation keeps at the same level as in the model without interaction, but for a much higher mean estimation. Then, the model with interactions is able to classify the healthy people better, which is a problem in this dataset, due to the reduced number of healthy people.

5. CONCLUSION

Many experimental data are collected in repeated measurement statistical designs. When replications on the same subjects are used, conventional machine learning methods are not appropriate since observations are no longer independent. Then, alternative methods as the ones proposed in this paper must be used to address repeated measurements.

The proposed approaches are conceptually simple and easy to implement. However, there is space for improving. The Bayesian approach can be extended by a generalized linear model allowing different types of link functions that may provide better fittings. Also, implementing the Metropolis-Hasting and Gibbs sampling algorithms by developing the distributions necessary to generate the Markov chains is interesting. This would give more control to the user than simply using WinBUGS code. Although the computational cost is not high with WinBUGS, it may be reduced.

The success of the classification also highly depends on the good discriminatory properties of the selected features, so integrating specific feature selection algorithms is important. The approach could also be extended to allow ordinal response data, which could be useful to classify PD patients by severity levels with the Hoehn and Yahr's scale.

Although there have been important advances in the diagnosis and progression of Parkinson's disease by using speech signals, there is a scientific challenge to develop reliable procedures that can be included into medical protocols in neurological units. Also, telemonitoring is a current challenge.

6. APPENDIX

The WinBUGS code is presented here for the approach considering interactions, i.e.:

```
model{
  for(i in 1:n){
    Y[i] ~ dbern(p[i])
    logit(p[i]) <- B[8] + B[1]*Z1[i] +
    B[2]*Z2[i] + B[3]*Z3[i] + B[4]*Z4[i] +
    B[5]*Z1[i]*Z2[i] + B[6]*Z1[i]*Z4[i] +
    B[7]*Z2[i]*Z4[i]
    for(j in 1:T[i]){
      X1[i,j] ~ dnorm(Z1[i],Tw[1])
      X2[i,j] ~ dnorm(Z2[i],Tw[2])
      X3[i,j] ~ dnorm(Z3[i],Tw[3])
      X4[i,j] ~ dnorm(Z4[i],Tw[4])
    }
    Z1[i] ~ dnorm(0,0.01)
    Z2[i] ~ dnorm(0,0.01)
    Z3[i] ~ dnorm(0,0.01)
    Z4[i] ~ dnorm(0,0.01)
  }
  for(h in 1:8){
    B[h] ~ dnorm(0,0.001)
  }
  for(k in 1:4){
    Tw[k] <- pow(Sw[k],-2)
    Sw[k] ~ dunif(0,10)
  }
}
```

REFERENCES

[1] M. C. de Rijk, L. J. Launer, K. Berger, M. M. Breteler, J. F. Dartigues, M. Baldereschi, L. Fratiglioni, A. Lobo, J. Martinez-Lage, C. Trenkwalder, and A. Hofman, “Prevalence of parkinsons disease in europe: a collaborative study of population-based cohorts,” *Neurology*, vol. 54, no. 11, pp. S21–S23, 2000.

[2] N. Miller, “Communication changes in Parkinson’s disease,” *Rev. Logopedia, Foniatría y Audiología*, vol. 29, no. 1, pp. 37–46, 2009.

[3] L. Baghai-Ravary and S. W. Beet, *Automatic Speech Signal Analysis for Clinical Diagnosis and Assessment of Speech Disorders*, Springer Briefs in Electrical and Computer Engineering - Speech Tecnologies. Springer, 2013.

[4] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, “Suitability of dysphonia measurements for telemonitoring of Parkinson’s disease,” *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 4, pp. 1015–1022, 2009.

[5] A. Tsanas, M. A. Little, P. E. McSharry, and L.O. Ramig, “Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson’s disease symptom severity,” *The Royal Society Interface*, vol. 8, no. 59, pp. 842–855, 2011.

[6] A. Tsanas, M. A. Little, P. E. McSharry, and L.O. Ramige, “Using the cellular mobile telephone network to remotely monitor Parkinson’s disease symptom severity,” *IEEE Transactions on Biomedical Engineering (submitted)*, 2013.

[7] S-C. Yin, R. Rose, O. Saz, and E. Lleida, “A study of pronunciation verification in a speech therapy application,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2009, pp. 4509–4612.

[8] M. Hariharan, K. Polat, and R. Sindhu, “A new hybrid intelligent system for accurate detection of Parkinson’s disease,” *Computer Methods and Programs in Biomedicine*, In press.

[9] B. Shahbaba and R. Neal, “Nonlinear models using dirichlet process mixtures,” *Journal of Machine Learning Research*, vol. 10, pp. 1829–1850, 2009.

[10] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman and Hall/CRC, 1993.

[11] D. Gil and J. Magnus, “Diagnosing Parkinson by using artificial neural networks and support vector machines,” *Global Journal of Computer Science and Technology*, vol. 9, no. 4, pp. 63–71, 2009.

[12] C. O. Sakar and O. Kursun, “Telediagnosis of Parkinson’s disease using measurements of dysphonia,” *Journal of Medical Systems*, vol. 34, pp. 591–599, 2010.

[13] T. Silva and I. Dutra, “T-SPPA trended statistical preprocessing algorithm,” in *The International Conference on Digital Information Processing and Communications*, J. Platos V. Snasel and E. El-Qawasmeh, Eds. 2011, vol. I, pp. 118–131, Springer-Verlag.

[14] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten, “The WEKA data mining software: an update,” *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[15] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in practice*, Chapman and Hall, 1996.

[16] I. Ntzoufras, *Bayesian Modeling Using WinBUGS*, Wiley Series in Computational Statistics. Wiley, 2011.