34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Multi-patient learning increases accuracy for Subthalamic nucleus identification in deep brain stimulation

Hernán Darío Vargas Cardona, Álvaro Ángel Orozco and Mauricio A. Álvarez

Abstract— **Establishing the exact position of basal ganglia is key in several brain surgeries, particularly in deep brain stimulation for patients suffering from Parkinson's disease. There have been recent attempts to introduce automatic systems with the ability to localize, with high accuracy, specific brain regions. These systems usually follow the classical supervised learning paradigm, in which training data from different patients are employed to construct a classifier that is patient-independent. In this paper, we show how by sharing information from different patients, it is possible to increase accuracy for targeting the Subthalamic Nucleus. We do this in the context of multi-task learning, where different but related tasks are used simultaneously to leverage the performance of a learning system. Results show that the multitask framework can outperform the traditional patient-independent scenario in two different real datasets.**

I. INTRODUCTION

Parkinson's Disease (PD) is a progressive degenerative condition of the Central Nervous System (CNS). It is caused by cell deterioration of a brain structure known as Substantia Nigra Reticulata-SNR, which leads to a decrease in dopamine levels. Patients with PD are usually subjected to drug treatment. In more advanced stages of the disease, it becomes apparent to proceed with a surgical treatment. Deep Brain Stimulation (DBS) is the most common surgical procedure for PD [1]. During DBS, the interpretation of physiological signals known as Microelectrode Recording (MER) signals is essential: the specialists analyze these recordings to locate specific target areas where a stimulating device should be implanted.

Identification of brain structures from processing MER signals has proved to be an excellent medical support for the correct localization of a target brain area and the respective insertion of neuroexcitatory devices. Previous works employed processing approaches based on temporal analysis of spikes [2], [3]. Another approach that has been used is the time-frequency analysis, where MER signals are transformed to different mathematical spaces. For example, the Short-Time Fourier Transform space (STFT) [4] or the Wavelet Transform space (WT) [5]. A more recent method includes a representation of the MER signals through adaptive filter banks (adaptive wavelets - AW) with lifting schemes [6].

To the best of our knowledge, all the methods used so far for basal ganglia identification follow the usual supervised learning paradigm. In a nutshell, each microelectrode recording is transformed to a feature space using some signal-processing representation (i.e. STFT, WT and AW). The feature vector thus obtained, $\mathbf{x}$, has an associated label t, assigned by the specialist. In practice, we usually have access to a set of feature vectors $\mathbf{X}$ and the corresponding set of labels $\mathbf{t}$. Based on a subset of $\mathbf{X}$ and $\mathbf{t}$, known as a *training set*, a learning algorithm is put to work, with the hope that the algorithm will exhibit an adequate generalization ability over a different subset of $\mathbf{X}$ and $\mathbf{t}$, known as a *validation set*. Learning algorithms that have been tested in basal ganglia identification problems include naive Bayes [6], hidden Markov models [7], [8], support vector machines [9], among others.

Different researchers have reported successful results when using the classical supervised learning method for targeting particular brain regions. In this paper, we look for improving the accuracy delivered by the classical supervised learning paradigm by including correlations between the patients that have undergone surgery. Our inspiration is the *multi-task learning* framework [10], that has been receiving increasing interest within the machine learning community in the last few years [11]. The idea behind multi-task learning is that by learning simultaneously different but related tasks, it is possible to increase the performance of a learning algorithm. The augmented performance is explained due to the transfer of information between tasks. In our context, we will assume that each patient undergoing DBS is a different task. We then attempt to increase the accuracy in targeting the Subthalamic Nucleus (STN) of that particular patient by using learning in multiple patients. In this setup, the multi-task learning becomes *multi-patient learning*.

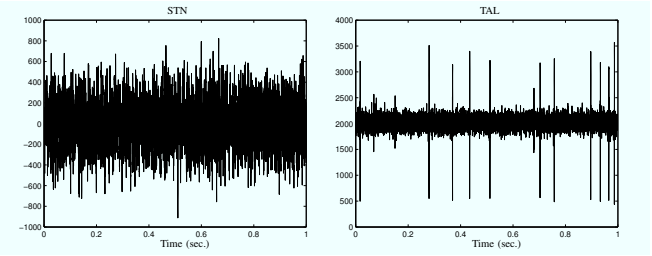
Several algorithms for multi-task learning have been proposed in the machine learning literature. In this work, we employ the multi-task Gaussian processes framework that exhibits state of the art performance in multi-task problems. To obtain the input vectors $\mathbf{X}$ we use adaptive wavelets. We show how the multi-patient learning framework improves accuracy when compared to the usual patient-independent setup, in two different datasets.

II. MATERIALS AND METHODS

A. Databases

A first database comes from Universidad Tecnológica de Pereira (DB-UTP). It contains recordings of surgical procedures in four patients with Parkinson's disease. Microelectrode recordings were obtained using the ISIS MER

H. D. Vargas, A. A. Orozco and M. A. Álvarez are with the Department of Electrical Engineering, Faculty of Engineering, Universidad Tecnológica de Pereira, Pereira, Colombia. {hernan.vargas, aaog, malvarez}@utp.edu.co



(a) MER from STN (b) MER from TAL

Fig. 1. Samples of Microelectrode Recordings from the DB-UTP database. On the left, a microelectrode recording from the Subthalamic Nucleus. On the right, a microelectrode recording from Thalamus. Microelectrode recordings from Zone Incerta and Substantia Nigra Reticulata follow a similar spiky shape.

system (Inomed Medical GmbH).[1] MER signals were labeled by neurophysiology and neurosurgery specialists from the Institute of Parkinson and Epilepsy of the Eje Cafetero, located in the city of Pereira, Colombia. In total, there are 400 recordings of 1 second of duration, sampled at 25 KHz with 16-bit resolution. The signals come from four patients and we only consider two classes: 200 recordings belong to the Subthalamic Nucleus and 200 recordings belong to other brain regions (Thalamus-TAL, Zone Incerta-ZI, Substantia Nigra Reticulata-SNR). Samples from STN and TAL are shown in figure 1.

A second database comes from Universidad Politécnica de Valencia (DB-UPV). Surgeries were carried out in the General University Hospital of Valencia, Spain, and labeled by specialists in neurophysiology and electrophysiology. The equipment used for data acquisition was the LeadPointTM Medtronic (Medtronics Functional Diagnostics).[2] Each signal is 1 sec. long, sampled at 24 KHz. In total, there are 240 recordings coming from four patients: 120 recordings belong to STN and 120 come from other brain regions.

B. Feature Extraction with Wavelets

After preprocessing each MER signal by removing its artifacts and by normalizing it, we use a dual adaptive scheme to decompose the original signal into two levels. The signal is partitioned on windows of 80 msecs with an overlap of 50%. From the approximation coefficients obtained from each window, we calculate the normalized average, the absolute maximum, the kurtosis and the energy, obtaining 8 features ($\mathbf{x} \in \mathbb{R}^8$) per MER signal. The reader is referred to [12] for a detailed description of the above feature extraction method.

C. Learning algorithms

We use several standard learning algorithms for classification in the patient-independent context, this is, when no correlation among patients is taken into account. For multi-patient learning, we use different alternatives of multiple-output Gaussian processes.

[1] http://www.inomed.com

[2] http://www.medtronic.com/

1) Standard classifiers: We test different parametric and non-parametric classifiers. Within the parametric family, we use the Naive Bayes classifier with a shared covariance matrix among classes, also known as the linear discriminant classifier (LDC) and the Naive Bayes classifier with a different covariance matrix per class, also known as the quadratic discriminant classifier (QDC). Within the non-parametric family, we use the K-nearest neighbors (KNN) algorithm with $K = 1$ and $K = 3$ (KNN1 and KNN3, respectively); a support vector machine with a radial basis kernel (SVM); a Gaussian process regressor used as a classifier (GPR) and a Gaussian process classifier (GPC). The theory behind each of the above classifiers is well known. The interested reader is referred to [13].[3]

2) Multi-output Gaussian Processes: Since this a relatively new topic in the machine learning literature, we spend a couple of lines here to describe the different multiple output Gaussian processes methods employed in the experimental section. A detailed description of several alternatives can be found at [11].

A general method for multiple output Gaussian processes employs convolution integrals of latent functions $\{u_q(\mathbf{x})\}_{q=1}^{Q}$ with smoothing kernels $\{G_d(\mathbf{x}-\mathbf{z})\}_{d=1}^{D}$, to describe D outputs or tasks $\{f_d(\mathbf{x})\}_{d=1}^{D}$,

$$f_d(\mathbf{x}) = \sum_{q=1}^{Q} \int G_d(\mathbf{x}-\mathbf{z}) u_q(\mathbf{z}) d\mathbf{z}.$$

Assuming that the latent functions are independent Gaussian processes with covariance functions $k_q(\mathbf{x},\mathbf{x}')$, the outputs $f_d(\mathbf{x})$ form a joint Gaussian process with covariance function $k_{d,d'}(\mathbf{x},\mathbf{x}')$ with $d,d' = 1,\ldots,D$,

$$k_{d,d'}(\mathbf{x},\mathbf{x}') = \sum_{q=1}^{Q} \int \int G_d(\mathbf{x}-\mathbf{z}) G_{d'}(\mathbf{x}'-\mathbf{z}') k_q(\mathbf{z},\mathbf{z}') d\mathbf{z} d\mathbf{z}'.$$

We call this covariance the Convolved Multiple Output Covariance or CMOC.

The linear model of coregionalization (LMC) is a particular case of the above covariance, one for which $G_d(\mathbf{x}-\mathbf{z}) = a_d \delta(\mathbf{x}-\mathbf{z})$, being $\delta(\mathbf{x})$ the Dirac delta function. The covariance $k_{d,d'}(\mathbf{x},\mathbf{x}')$ reduces then to

$$k_{d,d'}(\mathbf{x},\mathbf{x}') = \sum_{q=1}^{Q} a_d a_{d'} k_q(\mathbf{x},\mathbf{x}').$$

The number that results from the product $a_d a_{d'}$ is a measure of the correlation between the two tasks $f_d(\cdot)$ and $f_{d'}(\cdot)$. A further simplification of the above function, $k_{d,d'}(\mathbf{x},\mathbf{x}')$, can be obtained assuming that some of the terms $k_q(\mathbf{x},\mathbf{x}')$ are the same (notice, however, that the latent functions $u_q(\mathbf{x})$ are still considered to be orthogonal). This model receives the name of the intrinsic coregionalization model (ICM).

[3] The parametric classifiers, KNN1, KNN3 and the SVM are implemented using the PRTOOLS toolbox obtained from http://www.prtools.org/. GPR is implemented using the Gaussian Process Toolbox from http://staffwww.dcs.shef.ac.uk/people/N.Lawrence/gp/. GPC is implemented using the Gaussian Process Toolbox from http://www.gaussianprocess.org/gpml/code/matlab/doc/.

Once any of the CMOC, LMC or ICM covariances has been selected, the traditional technology of Gaussian processes for single outputs, as explained for example in [14], can be employed for doing multi-task regression or multi-task classification.

In the case of multi-task regression, the usual inference method is based on maximum likelihood, whereas for the multi-task classification, inference methods include the Laplace approximation, Expectation-Propagation (EP), among others.

In this paper we use multi-task Gaussian process regression with the CMOC and LMC covariances, for classification purposes. This practice is sometimes known as least-square classification. We refer to the multi-task GP with CMOC as MC and to the multi-task GP with LMC covariance as ML. We also use the ICM covariance in a multi-task Gaussian process classifier as introduced in [15], and refer to this method as MI.[4]

D. Validation

To test the statistical significance of our results, we follow the procedure proposed for model selection in [16]. We split each dataset in a training set and a validation set. We train the different methods using the training set and then we measure the accuracy over the validation set. We repeat this procedure 50 times with a different training set and validation set per repetition. To study if there are differences that are statistically significant among the classifiers, we apply first a Lilliefors test for normality over the 50 repetitions of each classifier. If the null hypothesis for normality is rejected, we perform a Kruskal-Wallis test to compare average performances among the classifiers. If null hypothesis for equal medians is rejected, we perform a multiple comparison test using Tukey-Kramer to study further which classifiers are different. All the significance levels are measured at 5%.

Two different types of experiments are performed. In the first type of experiment, we test the performance of the different classifiers using 50% of the datapoints from each patient for training, and then validate the performance over the other 50% of datapoints per patient. The experiment is performed over both databases. We refer to this type of experiment as E1. The second type of experiment is performed only on DB-UTP. The idea here is to test the generalization ability of the method when few datapoints per patient are used in the training phase. In detail, we use 50% of the datapoints for three patients and only 10% of the datapoints for the fourth patient. We then report the accuracy over the remaining 90% datapoints for the fourth patient for validation purposes. We will refer to this experiment as E2.

III. Experimental Results

Figures 2 and 3 show accuracy results for E1, this is, when the same amount of datapoints per patient is used for the training stage and the validation stage.

[4]We implement MC and ML using the MULTIGP Toolbox retrieved from `http://staffwww.dcs.shef.ac.uk/people/N.Lawrence/multigp/`. We implement MI using software available at `http://homepages.inf.ed.ac.uk/gsanguin/software.html`.

Figure 2 shows the mean accuracy performances for different classifiers applied to the database DB-UTP. We also include in the figure two standard deviations away from the mean performance. Notice that the methods employing multi-patient learning (MI, MC, and ML) exhibit better performance than methods disregarding correlations between patients (KNN1, KNN3, LDC, QDC, SVM, GPR and GPC). This increased performance is further tested using the hypothesis tests described in section II-D. The null hypothesis of equal means between the group of multi-patient learning algorithms and the group of patient-independent algorithms is rejected. According to the same analysis, the difference in performances between MI, MC and ML is not statistically significant.

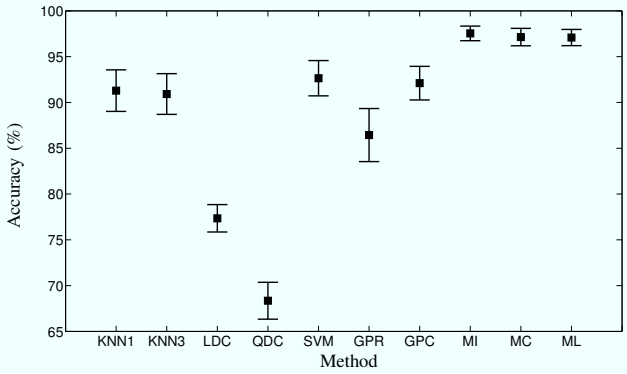


Fig. 2. Mean accuracy and two standard deviations for different classifiers applied to DB-UTP for experiment E1. KNNX stands for K-nearest neighbors, where X is either 1 or 3. L(Q)DC stands for linear(quadratic) discriminant classifier. SVM stands for support vector machine. GPR stands for Gaussian Process Regressor. GPC stands for Gaussian Process Classifier. MI represents a multi-patient GP classifier with ICM covariance. MC represents a multi-patient GP regressor with CMOC. ML represents a multi-patient GP regressor with LMC covariance.

Figure 3 shows the accuracy performances for different classifiers applied to the database DB-UPV. The mean accuracy performances for the multi-patient algorithms (MI, MC and ML) is superior to the mean accuracy performances of the standard classifiers.

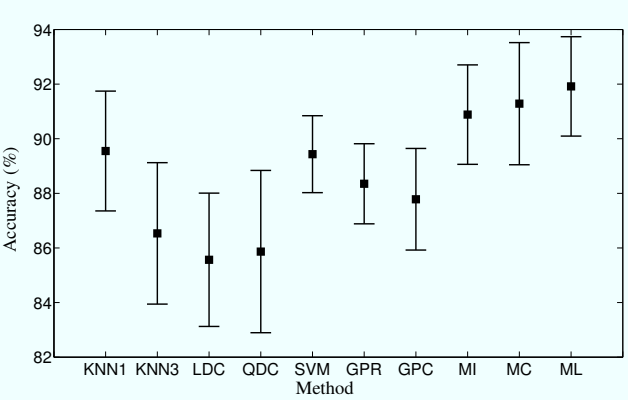


Fig. 3. Mean accuracy and two standard deviations for different classifiers applied to DB-UPV for experiment E1. The specification for each classifier is given in section II-C.1 or in the caption of figure 2.

Based on the multiple comparison test, we conclude that

MI and MC are not statistically significant when compared to KNN1. However, we can reject the null hypothesis of equal means between ML and KNN1. Also, the mean performances between MI and SVM are not statistically significant. Nevertheless, the post test analysis rejects the null hypothesis of equal means between MC and SVM and between ML and SVM. The post test analysis also rejects the null hypothesis of equal means between the multi-patient learning algorithms, and the other standard classifiers (KNN3, LDC, QDC, GPR and GPC).

Figure 4 shows the mean accuracy performance and two standard deviations for E2 over database DB-UTP. Recall from section II-D that in E2, we use 50% of the datapoints for three patients and only 10% data points for the other patient, for the training stage. For this particular experiment, we use 50% of the datapoints available for patient 2, patient 3 and patient 4, and 10% of the datapoints available for patient 1, for training. In Figure 4, we report the performance over the remaining 90% datapoints for patient 1.

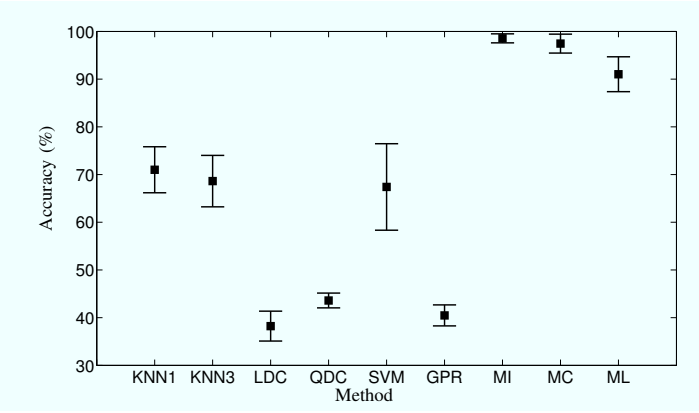


Fig. 4. Mean accuracy and two standard deviations for different classifiers applied to DB-UTP for experiment E2. The specification for each classifier is given in section II-C.1 or in the caption of figure 2.

It can be seen that the multi-patient learning algorithms clearly outperform the standard learning algorithms. The null hypothesis tests further confirm this result.

IV. Discussion and Conclusions

We have seen in the experimental section that methods using multi-patient learning increase the accuracy over standard learning techniques. Experiment 2 (see figure 4) is particularly illustrative of the way in which multi-patient learning can leverage the performance of the learning system, when only a small number of datapoints are available for a particular patient.

There is perhaps an even deeper insight about this experiment. To obtain the results appearing in figure 4, we trained the different methods using 50% of the datapoints per patients 2, 3 and 4, and 10% of the datapoints per patient 1. We also compared performances when using 50% of the datapoints per patients 1, 3 and 4, and 10% of the datapoints for patient 2. The mean accuracies computed over the 90% remaining points for patient 2, were not statistically significant. A closer look at the feature space let us realized that patient 1 is negatively correlated with patients 2, 3 and 4. Patient 2 is positively correlated with patients 3 and 4. These experiments indicate that multi-patient learning is robust to scenarios where few negatively correlated samples are included in the training phase.

We would still like to validate multi-patient learning in more extreme scenarios, for example, when we use database DB-UTP for training the learning system, and DB-UPV for testing it.

Acknowledgments

The authors of this paper would like to thank Rubén Darío Pinzón for his comments. We also thank the MD Hans Carmona Villada and the Institute of Epilepsy and Parkinson of Eje Cafetero, Colombia, who helped in organizing the database DB-UTP. We also acknowledge Dr. Enrique Guijarro, for providing us with the DB-UPV database. This research has been developed under the project: *"Desarrollo de un sistema automático de mapeo cerebral y monitoreo intraoperatorio cortical y profundo: aplicación neurocirugía"*, financed by Colciencias with code 111045426008.

References

[1] A.L. Benabid, *Deep brain stimulation for Parkinsons disease*. Curr Opin Neurobiol 13: 696706, 2003.

[2] H. Chan, T. Wu, S. Lee, M. Lin, S. He, P. Chao and Y. Sai, *Unsupervised wavelet-based spike sorting with dynamic codebook searching and replenishment*, Neurocomputing, vol. 73, pp. 1513-1527, 2010.

[3] R.A. Santiago, J. MacNames, K. Burchiel and George G. Lendaris *Developments in understanding neuronal spike trains and functional specializations in brain regions*, Neural Networks, 16:601-07, 2003

[4] P. Novak, S. Daniluk, S. Elias and J. Nazzaro, *Detection of the subthalamic nucleus in microelectrographic recordings in parkinson disease using the high frequency (500 hz) neuronal background*, Neurosurgey, no. 106, pp. 175-179, 2007.

[5] P. Gemmar, O. Gronz, T. Henrichs and F. Hertel, *Advanced methods for target navigation using microelectrode recordings in stereotactic neurosurgery for deep brain stimulation*, in Proc. of the CBMS 2008, pp. 99-104, 2008.

[6] R. Pinzon, A. Orozco, G. Castellanos and H. Carmona, *Towards High Accuracy Classification of MER Signals for Target Localization in Parkinson's Disease*, in Proc. of the IEEE-EMBC, pp 4040-43, 2010.

[7] A. Tahgva, *Hidden Semi-Markov Models in the Computerized Decoding of Microelectrode Recording Data for Deep Brain Stimulator Placement*, World Neurosurgery, pp 758-764, 2011.

[8] A. Orozco, M. Alvarez, E. Guijarro and G. Castellanos, *Identification of Spike Sources using Proximity Analysis through Hidden Markov Models*, in Proc. of the IEEE-EMBC 2006, pp 5555-5558, 2006.

[9] P. Guillen, F. Martinez, R. Sanchez, M. Argez, and L. Velsquez, *Characterization of Subcortical Structures during Deep Brain Stimulation utilizing Support Vector Machines*, in Proc. of the IEEE-EMBC 2011, pp 7949-7952, 2011.

[10] R. Caruana, *Multitask Learning*, Machine Learning, 28:41-75, 1997.

[11] M. A. Alvarez, L. Rosasco, N. D. Lawrence *Kernels for vector-valued functions: a review*, Available at `http://arxiv.org/pdf/1106.6251v1`, 2011.

[12] E. Giraldo, G. Castellanos and A.A. Orozco, *Feature Extraction for MER Signals Using Adaptive Filter Banks*, in Electronics, Robotics and Automotive Mechanics Conference, pp 582-585, 2008.

[13] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[14] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.

[15] G. Skolidis and G. Sanguinetti, *Bayesian Multi-task Classification with Gaussian Process Priors*, IEEE TNN, pp 2011 - 2021, 2011.

[16] J. Pizarro, E. Guerrero and P. L. Galindo, *Multiple comparison procedures applied to model selection*, Neurocomputing, 48: 155-173, 2002.