A Clinical Application of Feature Selection: Quantitative Evaluation of the Locomotor Function

Luca Palmerini[1], Laura Rocchi[1], Sabato Mellone[1],
Franco Valzania[2], and Lorenzo Chiari[1]

[1] Biomedical Engineering Unit, DEIS, University of Bologna
Viale Risorgimento 2, 40136 Bologna, Italy
{luca.palmerini,l.rocchi,sabato.mellone,lorenzo.chiari}@unibo.it
[2] Department of Neuroscience, University of Modena and Reggio Emilia
via Pietro Giardini 1355, 41126 Baggiovara (MO), Italy
f.valzania@ausl.mo.it

Abstract. Evaluation of the locomotor function is important for several clinical applications (e.g. fall risk of the elderly, characterization of a disease with motor complications). We consider the Timed Up and Go test which is widely used to evaluate the locomotor function in Parkinson's Disease (PD). Twenty PD and twenty age-matched control subjects performed an instrumented version of the test, where wearable accelerometers were used to gather quantitative information. Several measures were extracted from the acceleration signals; the aim is to find, by means of a feature selection, the best set that can discriminate between healthy and PD subjects. A wrapper feature selection was implemented with an exhaustive search for subsets from 1 to 3 features. A nested leave-one-out cross validation (LOOCV) was implemented, to limit a possible selection bias. With the selected features a good accuracy is obtained (7.5% of misclassification rate) in the classification between PD and healthy subjects.

Keywords: Feature selection, Clinical, parkinson's disease, Accelerometer, Selection bias, Nested cross validation.

1 Introduction

Evaluation of the locomotor function is important for several clinical applications (e.g. fall risk of the elderly, characterization of a disease with motor complications). The Timed Up and Go (TUG) is a widely used clinical test to assess balance, mobility and fall risk in Parkinson's disease (PD). The traditional clinical outcome of this test is its duration, measured by a stopwatch. Since this single measure cannot provide insight on subtle differences in test performances, instrumented Timed Up and Go tests (iTUG) have been recently proposed [1], [2]. These studies demonstrated the potential of using inertial sensors to quantify TUG performance. As stated in [2], quantitative evaluation is especially important for early stages of PD when balance and gait problems are not clinically evident but may be detected by instrumented analysis. The aim of this study is to find, by means of a feature selection process, the

A. Fred et al. (Eds.): IC3K 2010, CCIS 272, pp. 151–157, 2013.

best set of quantitative measures that can allow an objective evaluation of gait function in PD and could be considered as possible early biomarkers of the disease. Feature selection has recently been used in the field of Parkinson's disease to quantify the performance of a PD subject [3]; in the mentioned study the quantitative data came from force/torque sensors.

2 Methods

We examined twenty early-mild PD subjects OFF medication (Hoehn & Yahr ≤ 3, 62±7 years old, 12 males and 8 females) and twenty healthy age-matched control subjects (CTRL, 64±6 years old, 7 males and 13 females). The OFF condition in PD subjects was obtained by a levodopa washout of at least 18 hours and a dopamine agonist washout of at least 36 hours. Subjects wore a tri-axial accelerometer, McRoberts© Dynaport Micromod, on the lower back at L5 level. They performed three TUG trials (single task, ST) and three TUG trials with a concurrent cognitive task (dual task, DT), which consisted in counting audibly backwards from 100 by 3s. The TUG trial consisted of rising from a chair, walking 7m at preferred speed, turning around, returning and sitting down again. A schematic representation of the task is shown in Fig. 1.

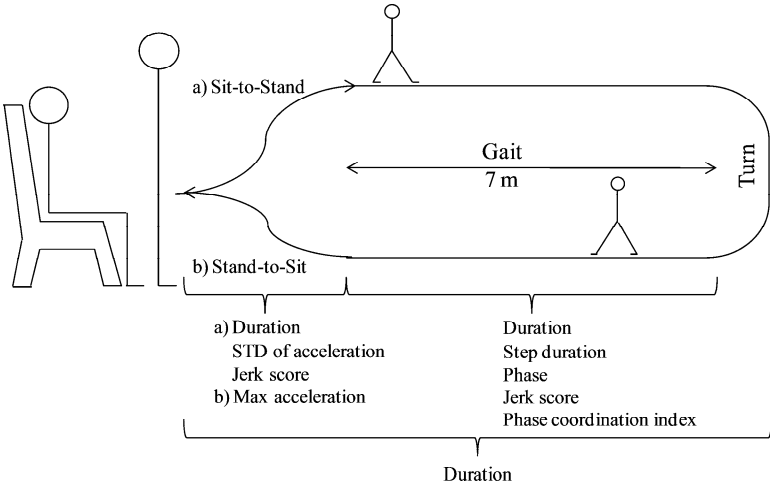

Fig. 1. Timed Up and Go Test and extracted measures

It has to be noted that a modified version of the TUG was used in this study, with a Gait section of 7m (instead of 3) to get more accurate measures of the various steps.

Several temporal (including total duration of the test), coordination and smoothness measures were extracted from the acceleration signals in different sections of the TUG. In Fig. 1 the main measures are reported.

Considering the Gait section, each stride (from one heel strike to the consecutive heel strike of the same leg) defines one gait cycle. The phase is determined by the

ratio between the duration of the first step of the gait cycle and the entire duration of the gait cycle: a factor of 360 is used to transform the variable into degrees (360 degrees would correspond to the entire gait cycle) [4]. Among the other measures, phase coordination index measures the symmetry of gait [4] and jerk score (for both Sit-to-Stand and Gait sections) can be seen as an index of movement smoothness.

In the Gait section, jerk score and step duration were computed for each step; for the following analysis their averages across all the steps were considered, together with measures of variability between different steps (standard deviation, STD, and coefficient of variation, CV). Similarly, phase was computed for each gait cycle but only its average and variability measures were considered.

Jerk score (for both Sit-to-stand and Gait sections), Root Mean Square (RMS), and max value of acceleration, were computed along two orthogonal axes of the accelerometer: the first aligned with the direction of gait progression and coincident with the biomechanical anteroposterior (AP) axis of the body; the second in the left/right direction and coincident with the biomechanical mediolateral (ML) axis of the body.

For each measure, both in ST and in DT, we computed the mean value across the three repeated trials for the following analyses.

2.1 Feature Selection

The total number of measures (features) is higher (56, 28 for ST and 28 for DT) than the available samples (40 subjects). Therefore feature selection is necessary to avoid overfitting and to improve the performance of the classifiers. To select, from all the available features, the subset which has the best discriminative ability, a "wrapper" feature selection [6] was implemented: the objective function was the predictive accuracy of a given classifier on the training set. We used the following classifiers: linear and quadratic discriminant analysis (LDA and QDA, respectively), Mahalanobis classifier (MC), logistic regression (LR), K-nearest neighbours (KNN, K=1) and linear support vector machines (SVM). An exhaustive search among subsets of cardinality from one to three was implemented; the limit of three was chosen to permit a clinical interpretation of the result (it would be difficult to associate too many features with different aspects of the disease). Subsets of different cardinalities were considered separately.

The adopted procedure is similar to the one proposed by [3] where an exhaustive search of subsets of three features was performed. Still, in the present study, feature selection bias was also considered.

Since feature selection is part of the tuning design of the classifier, it needs to be performed on the training set, in order to avoid a possible bias (selection bias) in the final evaluation of the accuracy of the classifier [5]. The most common solution to this problem is to use a nested cross validation procedure [6]: the internal feature selection step is repeated for each training set resulting from the external cross validation. In this study, because of the small sample size (40), a leave-one-out cross validation (LOOCV) was implemented both for the feature selection steps and for the final evaluation of the classifier.

As it can be seen in Fig. 2, the external cross validation used for estimation of the accuracy of the classifier ($LOOCV_{ext}$) splits the dataset in 40 different training and testing sets (TR_i,TS_i $1 \leq i \leq 40$); for each TR_i, a different feature selection step was performed (FS_i, $1 \leq i \leq 40$). The objective function (predictive accuracy) of each feature selection was evaluated by an internal LOOCV ($LOOCV_{int}$). After each FS_i, a list of optimal subsets of features was generated: there was generally more than one subset with the same highest $LOOCV_{int}$ accuracy (more than one optimal subset). In the nested procedure TS_i should be classified from the classifier built with a single subset chosen by FS_i; in this study, since more than one optimal subset was found, it was not possible to make a unique choice. Moreover different FS_i's led to different lists of optimal subsets. So we decided to extract the subset which was selected as optimal more frequently over all the FS_i's (overall optimal subset, see Fig. 2). The number of times a certain subset was selected as optimal (selection times) can be seen as an index of how that subset is robust to changes in the training set, and therefore to selection bias. Eventually, the accuracy of the classifier (misclassification rate, MR) was computed by $LOOCV_{ext}$ for the overall optimal subset (see Fig. 2).

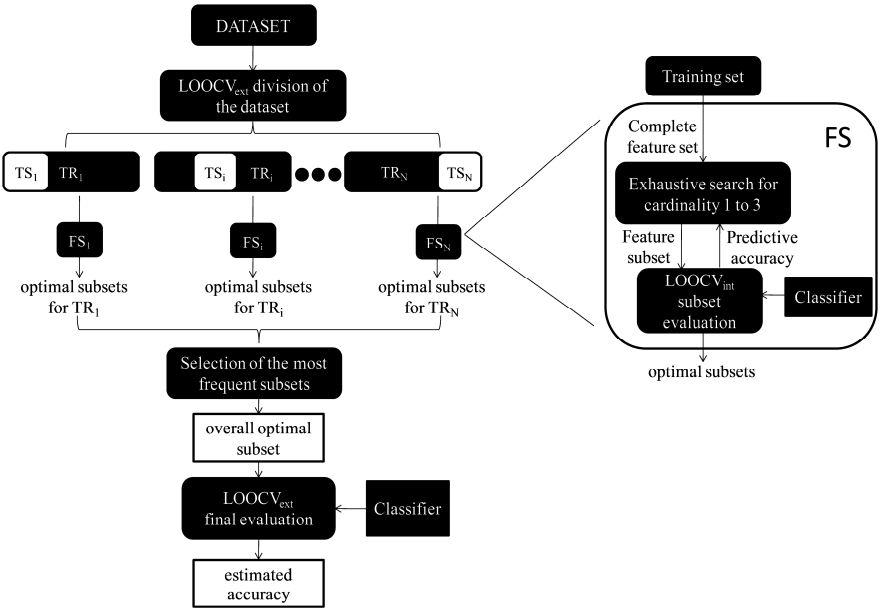

Fig. 2. Feature selection procedure

3 Results and Discussion

In Table 1 the results of the feature selection procedure for subsets of 3 measures are reported; the estimated accuracy is presented together with the *selection times* (the number of times a subset was selected as optimal among the 40 different feature selection procedures). Subsets of 3 measures were preferred since subsets of lower

cardinality led to higher misclassification rates. It can be seen that a good misclassification rate could be achieved (7.5%-10%) by all the classifiers. As discussed in section 2, estimates of misclassification rates of subsets with higher *selection times* should be considered as more reliable, regarding selection bias, with respect to estimates with lower *selection times*.

The best subset from this point of view is the subset selected by the KNN classifier which is exclusively made of measures related to the jerk score in different sections of the TUG. It can then be seen that the smoothness of the movement during Gait and Sit-to-Stand is very important in discriminating between control and PD subjects.

Table 1. Results of the feature selection procedure

Class.	Overall optimal subsets	Task	Selection times /40	MR
LDA	RMS of AP acceleration during Sit-to-Stand	single task	32	7.5%
	Max AP acceleration during Stand-to-Sit	dual task		
	STD of the phase during Gait	dual task		
QDA	RMS of ML acceleration during Sit-to-Stand	single task	25	7.5%
	Max AP acceleration during Stand-to-Sit	single task		
	CV of the step duration during Gait	dual task		
LR	Jerk score of AP acceleration during Sit-to-Stand	single task	28	7.5%
	Jerk score of ML acceleration during Gait	single task		
	STD of the step duration during Gait	dual task		
KNN	Jerk score of AP acceleration during Sit-to-Stand	single task	36	7.5%
	Jerk score of AP acceleration during Gait	dual task		
	CV of the jerk score of ML acceleration during Gait	dual task		
MC	Jerk score of AP acceleration during Sit-to-Stand	single task	32	10%
	Jerk score of ML acceleration during Gait	single task		
	max AP acceleration during Stand-to-Sit	dual task		
SVM	Jerk score of ML acceleration during Sit-to-Stand	single task	25	7.5%
	CV of the jerk score of ML acceleration during Gait	single task		
	max AP acceleration during Stand-to-Sit	dual task		

Considering the overall optimal subsets from all the classifiers, the procedure always selected a measure related with the sit-to-stand and one or two measures related with the gait phase. In four subsets there is also a measure extracted during Stand-to-Sit. It should also be remarked that every subset presented in Table 1 is made of both single and dual task related measures.

These measures improve the discrimination power between CTRL and PD with respect to the traditional TUG duration (the best misclassification rate that can be

obtained by using this single measure with the reported classifiers, in ST or in DT, is 35%), which interestingly was not selected in any of the overall optimal subsets. Moreover TUG duration alone was not significantly different between the two groups (as in [1] and [2]) and therefore it could not discriminate between CTRL and early-mild PD. Instead, considering various quantitative measures related to different parts of the TUG (see Table 1), allowed us to obtain good accuracy in the classification of PD subjects.

This accuracy would not have been obtained without feature selection; considering all the features altogether, the number of features is higher than the number of samples. In this case LDA, QDA and MC cannot be used because it is not possible to estimate the covariance matrix; similarly, in LR the model is over-parameterized and some coefficients of the logistic model are not identifiable. So the only classifiers that can be used without feature selection are KNN and SVM which, using all the features, have a MR of 52% and 20%, respectively; this reflects the importance of performing feature selection in this kind of datasets.

Furthermore it has to be noted that, even if our relatively small sample size limits the power of our data mining perspective, a nested cross validation was applied to limit the possible feature selection bias. Since it was not possible to follow the typical nested procedure (because several different combinations of features were selected as optimal), a value was derived which can be seen as an index of the reliability of the estimation of the misclassification rate.

4 Conclusions

The main result achieved by this work is that a set of few quantitative measures, derived from a clinical test for locomotor evaluation, can discriminate with a good accuracy between early-mild PD and CTRL subjects.

Further experiments should be made on new subjects to have an independent data set and validate these findings; in particular, the selected optimal measures could be tested on PD subjects in an earlier stage of their disease in order to check if they could also be used as early biomarkers of PD. On the other hand it should be investigated whether the presented measures remain valid and maintain their superiority over TUG duration for later stages of the disease. In fact, even if the presented subsets are optimal for classifying early-mild PD, there is no guarantee that they would be optimal to monitor the disease progression or to detect changes in gait patterns after a particular medical treatment; in this context, the next step will be a follow-up of the study with the same subjects.

Another future goal will be to assess if and how the TUG carried out under DT can add discriminative power with respect to the ST alone (as suggested by this study), since this would have important implications on the experimental design.

Acknowledgements. The authors wish to thank Luca Codeluppi, MD, and Valentina Fioravanti, MD, from the Department of Neuroscience, University of Modena and Reggio Emilia, Modena, Italy, for clinical supervision and assistance in data.

References

1. Weiss, A., Herman, T., Plotnik, M., Brozgol, M., et al.: Can an accelerometer enhance the utility of the Timed Up & Go Test when evaluating patients with Parkinson's disease? Med. Eng. & Phys. 32(2), 119–125 (2010)
2. Zampieri, C., Salarian, A., Carlson-Kuhta, P., Aminian, K., et al.: The instrumented timed up and go test: potential outcome measure for disease modifying therapies in Parkinson's disease. J. of Neurol., Neurosurg. & Psychiatry 81(2), 171–176 (2009)
3. Brewer, B.R., Pradhan, S., Carvell, G., Delitto, A.: Feature selection for classification based on fine motor signs of Parkinson's disease. In: 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (2009)
4. Plotnik, M., Giladi, N., Hausdorff, J.M.: A new measure for quantifying the bilateral coordination of human gait: effects of aging and Parkinson's disease. Exp. Brain Res. 181(4), 561–570 (2007)
5. Simon, R., Radmacher, M.D., Dobbin, K., McShane, L.M.: J. Natl. Cancer Inst. 95(1), 14–18 (2003)
6. Kohavi, R., John, G.H.: Wrappers for Feature Subset Selection. Art. Intel. 97(1-2), 273–324 (1997)