ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Data normalization and supervised learning to assess the condition of patients with multiple sclerosis based on gait analysis

Samir Azrour, Sébastien Piérard, Pierre Geurts, and Marc Van Droogenbroeck

University of Liège - Department of Electrical Engineering and Computer Science, Sart-Tilman B28, 4000 Liège - Belgium

Abstract. Gait impairment is considered as an important feature of disability in multiple sclerosis but its evaluation in the clinical routine remains limited. In this paper, we assess, by means of supervised learning, the condition of patients with multiple sclerosis based on their gait descriptors obtained with a gait analysis system. As the morphological characteristics of individuals influence their gait while being in first approximation independent of the disease level, an original strategy of data normalization with respect to these characteristics is described and applied beforehand in order to obtain more reliable predictions. In addition, we explain how we address the problem of missing data which is a common issue in the field of clinical evaluation. Results show that, based on machine learning combined to the proposed data handling techniques, we can predict a score highly correlated with the condition of patients.

1 Introduction

Gait impairment is considered as an important feature of disability in multiple sclerosis (MS) [1]. However, the measurement of gait characteristics by neurologists is usually limited to the use of a stopwatch. The problem is that stopwatches realize a very incomplete evaluation of the gait. The effect of new drugs or physical therapies is therefore hardly noticeable. The *GAit Measuring System* (*GAIMS*) [2] was created to address this lack by providing a wider range of measures that allow the definition of several relevant gait descriptors (inter-feet distance, step length, double support duration, ...).

The main score used by neurologists to assess the condition of MS patients is the *Expanded Disability Status Scale* (*EDSS*) [3]. It ranges from 0 (healthy persons) to 10 (death due to MS) in steps of 0.5. In this paper, we focus on predicting the condition of MS patients (*i.e.* their *EDSS*) based on their gait descriptors. As the relation between the gait descriptors and the *EDSS* is unknown and complex, we rely on machine learning techniques for the prediction.

In machine learning problems, it often happens that the data used to predict some output are influenced by external variables independent of the predicted output. This external influence can be seen as noise on the data and may affect the quality of predictions if we ignore it. In our applicative domain, we are confronted to this issue because the morphological characteristics of individuals (expressed in the space $\mathcal{C} = height \times weight \times gender \times age \times shoe\,size$) influence the gait while being, in first approximation, independent of the disease

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

level. This could lead to misleading results if, for instance, there is an inhomogeneous selection bias with respect to these morphological characteristics among the different classes in the dataset. Thus, a normalization with respect to these morphological characteristics is performed in order to obtain more reliable predictions. In addition to this normalization step, we had to develop an appropriate original technique to face the problem of missing data which is a quite common imperfection of the datasets in the biomedical domain.

The paper is organized as follows. Section 2 details the normalization of the gait descriptors with respect to the morphological characteristics. Section 3 explains our methodology for predicting the condition of patients (*EDSS*), shows the results and discusses them. In Section 4, we conclude the paper and give some prospects for improvement.

2 Normalization of the gait descriptors with respect to the morphological characteristics

2.1 Goal of the normalization

Gait analysis is a difficult task because the gait can be affected or modified by many factors including physiological factors (neurological, heart or metabolic diseases), psychological and emotional factors, dual task [4], morphological characteristics (height, weight, gender, and age), medications, alcohol intake, etc. To the contrary of the other parameters, the morphological characteristics are easy to determine precisely and are known most of the time in the context of clinical studies. The influence of height, gender, and age on gait has clearly been demonstrated [5, 6, 7]. Even if, to the best of our knowledge, there is no study showing the precise effect of weight on gait, we expect that overweight persons will walk differently than underweight persons. We will also consider the shoe size as it is likely to influence the gait especially for walking tests performed in tandem gait which is a gait where the toes of one foot touch the heel of the other foot at each step.

These morphological characteristics introduce biases in the gait descriptors which can lead to unreliable predictions. For instance, suppose that we want to discriminate healthy persons from MS patients and that the population of MS patients is significantly older than the healthy population. In this case, the learning algorithm fitted on these data may systematically classify a new old person as suffering from MS which is not acceptable. It is therefore important to remove the effect of the morphological characteristics on the gait descriptors before learning the relation between MS and these gait descriptors.

2.2 Normalization procedure

To normalize the gait descriptors with respect to the morphological characteristics, we start by defining two functions for each gait descriptor. The first one, denoted $\mu(\cdot)$, takes as input a person p whose morphological characteristics are

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

$c(p) \in \mathcal{C}$ and returns the expected value of the gait descriptor in a healthy population having the same morphological characteristics as p. If G_i is the random variable associated to the ith gait descriptor g_i, we can write this function as follows:

$$\mu_i(p) = \mathbb{E}\left[G_i \,|\, \mathcal{C} = c(p)\right], \tag{1}$$

where $\mathbb{E}$ denotes the mathematical expectation.

Likewise, the second function, denoted $\sigma(\cdot)$, also takes a person p as input and returns the expected variance of the gait descriptor g_i in a healthy population having the same morphological characteristics as p:

$$\sigma_i^2(p) = \mathbb{E}\left[(G_i - \mu_i(p))^2 \,|\, \mathcal{C} = c(p)\right]. \tag{2}$$

Using (1) and (2), the normalization procedure consists to replace each gait descriptor g_i of a person p (healthy and suffering from MS) in the dataset by:

$$g_i^{normalized}(p) = \frac{g_i(p) - \mu_i(p)}{\sigma_i(p)}. \tag{3}$$

2.3 Estimators of the functions $\mu(\cdot)$ and $\sigma(\cdot)$

It is not possible to know $\mu(\cdot)$ and $\sigma(\cdot)$ perfectly, because of the finiteness of our dataset. Thus, we can only use estimators $\hat{\mu}(\cdot)$ and $\hat{\sigma}(\cdot)$ of the real functions which are obtained based solely on the healthy population of our dataset. These two estimators are determined for each gait descriptor independently.

To determine the first estimator $\hat{\mu}(\cdot)$ for a given gait descriptor, we use a least squares linear regression to model the relationship between the morphological characteristics and the values of the gait descriptor. The second estimator $\hat{\sigma}(\cdot)$ is calculated with a similar procedure, except that the gait descriptors $g_i(p)$ are replaced by $(g_i(p) - \mu_i(p))^2$. Due to the small size of our healthy dataset (71 persons), we chose a model with a low complexity to fit the relation between morphological characteristics and gait descriptors in order to avoid overfitting.

3 Predicting the condition of patients based on normalized gait descriptors

3.1 Input and normalization

The gait descriptors are derived from 12 different walking tests that can be performed by the persons whose gait is analyzed by *GAIMS* (71 patients and 71 healthy persons). These tests include different walking modes (preferred pace, as fast as possible, and tandem gait) and different distances (7.62, 20, 100, and 500 m) [2]. 27 gait descriptors are derived for the short walking tests (7.62 m) and 135 are derived for the long tests (> 7.62 m). All these gait descriptors are normalized with respect to the morphological characteristics following the procedure explained in Section 2 before applying the following techniques.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

3.2 Missing data and imbalanced dataset

One difficulty originates from the absence of some walking tests in the database. This problem, which occurs mainly for patients, is due either to a lack of time (time constraint of the clinical routine) or because the patient's inability to perform the test. Therefore, some samples in the dataset can have a high percentage of missing data. To handle this problem of missing data, tree-based techniques like surrogate splits (used in the CART algorithm [8]) are not adapted to our case, because the missing values are not punctual but appear in block and can be numerous. On the other hand, other techniques like imputation [9] can lead to misleading predictions for the samples with a high percentage of missing data, because the imputed –and therefore more uncertain– data may hide the known data. Hence, as the missing values are grouped by walking tests, we chose to decompose the dataset in 12 parts, one for each test. A classifier predicting the *EDSS* (each *EDSS* level in the learning set is considered as a class) is created for each sub-dataset independently, and the final prediction is obtained by taking the arithmetic mean of the intermediate predictions. In this way, we can simply remove the missing walking tests in each sub-datasets and there is no more problem of missing data.

Another problem arises from the imbalance of our dataset, i.e. the very non uniform coverage of the *EDSS* scale in the training set. To limit the bias on the final predictions, it is important to use techniques to compensate for this problem. A review of the existing techniques solving the multiple classes imbalanced problem has been proposed by Fernández *et al.* [10]. Following their conclusions, we used in our experiments the combination of the 'one versus one' approach with an oversampling method based on the SMOTE algorithm [11].

3.3 Algorithm

The further results are obtained with the procedure explained above associated with the learning algorithm AdaBoost (based on decision trees) which is the algorithm that yielded the higher Pearson's correlation coefficient ρ (compared to SVM and ExtRaTrees [12]) between estimated *EDSS* and real *EDSS*.

A linear transformation is performed on the *EDSS* predictions in order to compensate for the bias originating from the absence of an extrapolation mechanism in decision trees. Indeed, the estimated *EDSS* are within the range of the real *EDSS* in our learning set. Hence, the *EDSS* of healthy persons will tend to be overestimated whereas the *EDSS* of the most severely diseased patients will tend to be underestimated. It should be noted that, to avoid two levels of leave-one-person-out, we have determined this correction from all pairs of real and estimated *EDSS*. Even if this may lead to a slightly optimistic RMS error prediction, such a linear correction does not change ρ.

3.4 Results and discussion

Figure 1 illustrates the results obtained for the prediction of the *EDSS*. The bisector of the first quadrant represents the locus of the points where the predicted

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

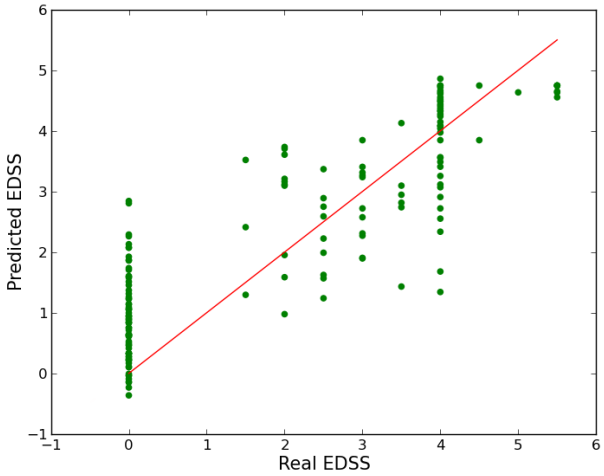

Figure 1: *EDSS* predictions obtained with our method.

EDSS would be equal to the real *EDSS*. The plotted points are the predictions of the *EDSS* obtained by *leave-one-person-out*, which means that the *EDSS* of each person in our dataset is predicted using only the data related to the other persons, after correction. As can be seen on the figure, the estimated *EDSS* is well correlated with the real *EDSS*. We obtained $\rho = 0.86$, and a RMS error of 1.11.

An interesting result is that all the predictions of the *EDSS* lower than 0.97 correspond to healthy persons. Similarly, all the predictions of the *EDSS* above 4.15 correspond to persons with an *EDSS* larger than or equal to 4. Moreover, we can see that almost all the persons with an *EDSS* ranging from 1 to 4 have their predictions in the same range.

It should be noted that the *EDSS* is a global score [3] assessing the condition of MS patients and is not only based on gait but also depends on other factors such as the vision, the cognition, or the upper extremity function. The results obtained are interesting in the sense that it might be possible to predict the *EDSS* based solely on the gait. A similar observation was made by Cao *et al.* from posturographic data [13]. The results could possibly be further improved with a larger and more balanced dataset. Indeed, the actual dataset is rather small and the problem of imbalance worsens this situation.

4 Conclusion

In this paper, we present a methodology to predict the condition of patients with multiple sclerosis based on their gait descriptors obtained with a gait analysis system. First, we explain the normalization of the gait descriptors with respect

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

to the morphological characteristics which influence the gait while being independent, in first approximation, of the disease level. This can lead to unreliable predictions if we do not compensate for their effect. In addition, to avoid the effect of missing data and the bias due to the imbalanced dataset, appropriate methods are used. The results show that we are able to predict a score from gait descriptors which is highly correlated with the condition of patients suffering from multiple sclerosis.

Further work will include the acquisition of more data in order to obtain a larger and more balanced dataset, and the identification of the most relevant gait descriptors and walking tests to predict the condition of patients.

Acknowledgments. We are grateful to the volunteers who have accepted to be recorded, the university hospital of Liège, and the staff of neurology for their help. We also thank Rémy Phan-Ba and Amaury Giet for their contribution to the project. Samir Azrour has a research fellowship of the Belgian National Fund for Scientific Research (F.R.S.-FNRS).

References

[1] C. Heesen, J. Böhm, C. Reich, J. Kasper, M. Goebel, and S. Gold. Patient perception of bodily functions in multiple sclerosis: gait and visual function are the most valuable. *Multiple Sclerosis*, 14:988–991, 2008.

[2] S. Piérard, R. Phan-Ba, V. Delvaux, P. Maquet, and M. Van Droogenbroeck. GAIMS: a powerful gait analysis system satisfying the constraints of clinical routine. *Multiple Sclerosis Journal*, 19(S1):359, October 2013. Proceedings of ECTRIMS/RIMS 2013 (Copenhagen, Denmark), P800.

[3] J. Kurtzke. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology*, 33(11):1444–1452, November 1983.

[4] E. Lamberg and L. Muratori. Cell phones change the way we walk. *Gait & Posture*, 35(4):688–690, April 2012.

[5] R. Bohannon. Comfortable and maximum walking speed of adults aged 20-79 years: reference values and determinants. *Age Ageing*, 26(1):15–19, January 1997.

[6] S. Cho, J. Park, and O. Kwon. Gender differences in three dimensional gait analysis data from 98 healthy korean adults. *Clinical Biomechanics*, 19(2):145–152, 2004.

[7] M. Murray, A. Drought, and R. Kory. Walking patterns of normal men. *Journal of Bone Joint Surgery*, 46a(2):335–360, 1964.

[8] L. Breiman, J. Friedman, R. Olsen, and C. Stone. *Classification and Regression Trees*. Wadsworth International (California), 1984.

[9] A. Donders, G. Heijden, T. Stijnen, and K. Moons. Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59(10):1087–1091, 2006.

[10] A. Fernández, V. López, M. Galar, M. Jesús, and F. Herrera. Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *Knowledge-based systems*, 42:97–110, 2013.

[11] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer. SMOTE: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[12] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, April 2006.

[13] H. Cao, L. Peyrodie, S. Boudet, F. Cavillon, O. Agnani, P. Hautecoeur, and C. Donzé. Expanded disability status scale (EDSS) estimation in multiple sclerosis from posturographic data. *Gait & Posture*, 37(2):242–245, February 2013.