Relating Clinical and Neurophysiological Assessment of Spasticity by Machine Learning

Blaž Zupan[1], Dobrivoje S. Stokić[2], Marko Bohanec[1],
Michael M. Priebe[3,4], Arthur M. Sherwood[3]

[1] Dept. of Intelligent Systems, J. Stefan Institute, Ljubljana, Slovenia

[2] MMRC - Mississippi Methodist Rehabilitation Center, Jackson, MS, USA

[3] Dept. of Physical Medicine and Rehabilitation,
Baylor College of Medicine, Houston, TX, USA

[4] Spinal Cord Injury Service, Veteran Affairs Medical Center, Houston, TX, USA

Correspondence:

Blaž Zupan, Ph.D.
Department of Intelligent Systems
Jožef Stefan Institute
Jamova 39
SI-1000 Ljubljana
Slovenia

phone: +386 61 177 3880
fax: +386 61 125 1038
e-mail: blaz.zupan@ijs.si

Abstract

Spasticity following spinal cord injury (SCI) is most often assessed clinically using a five-point Ashworth Score (AS). A more objective assessment of altered motor control may be achieved by using a comprehensive protocol based on a surface electromyographic (sEMG) activity recorded from thigh and leg muscles. However, the relation between the clinical and neurophysiological assessments is still unknown. In this paper we employ three different classification methods to investigate this relationship. The experimental results indicate that, if the appropriate set of sEMG features is used, the neurophysiological assessment is related to clinical findings and can be used to predict the AS. A comprehensive sEMG assessment may be proven useful as an objective way of evaluating the effectiveness of various interventions and for follow-up of SCI patients.

Keywords

spasticity assessment
clinical assessment of spasticity
neurophysiological assessment of spasticity
Ashworth Score
classification
discriminant analysis
machine learning

1 Introduction

Severe trauma to the vertebral column, due to e.g. a motor vehicle accident, may result in spinal cord injury (SCI) which is characterized by various neurological impairments. Inability to perceive tactile and/or temperature sensation in the body parts below the level of SCI and to voluntarily activate the muscles innervated by the spinal cord segments distal to SCI are among the most prominent. These findings are closely accompanied by an increase in muscle tone, i.e. spasticity, which often represents a major patient's complain. Although different clinical interventions have been used for amelioration of spasticity, their efficacy is often the matter of debate. One of the main reasons is that an objective assessment of spasticity is still lacking. That is not a surprise since spasticity is a multidimensional problem encompassing various aspects of motor control. Secondly, most of the assessment scales currently used are subjective in nature and not sensitive enough to detect subtle differences noticeable sometimes by patients themselves.

Based on classical definition of spasticity ("velocity-depended increase in tonic stretch reflexes"), a qualitative five-point Ashworth Score (AS) [2] has been developed. Typically, by moving a limb or limb segment through the range of motion in a subjects attempting to relax, the examiner grades the resistance felt during such passive movements. In order to extend clinical evaluation of SCI subjects, a comprehensive neurophysiological protocol has been developed based on recordings of surface electromyographic (sEMG) activity from thigh and leg muscles. This approach may proven useful to get a more objective assessment of altered motor control, and thus spasticity, following SCI [9]. However, the relationship between the clinical and neurophysiological assessments is still unknown.

The purpose of this study was to find out whether there is a relationship between the AS, as a clinical, and the sEMG-based neurophysiological assessment of spasticity, and if so, which neurophysiological features are most useful in predicting the clinical findings. To determine the existence and the type of such a relationship this paper analyzes the data set from 98 SCI individuals. This data set includes the AS determined by a clinician, and a set of neurophysiological features based on sEMG. The data were analyzed by three different classification methods: machine learning methods that employ decision trees and k-nearest neighbors, and a statistical method of linear discriminant analysis. The paper evaluates the appropriateness of the three classification methods for this data set and investigates which neurophysiological features represent the best estimate of the AS.

The reminder of the paper is organized as follows. Section 2 describes the data set used. The classification methods used are overviewed in section 3. The experimental methodology is given in Section 4. Section 5 presents the experimental evaluation of the classification methods and the impact of the feature subset selection on the classification accuracy. The results of experiments are discussed in Section 6. Section 7 summarizes the experimental findings and possible future work.

2 The experimental data

2.1 Subjects

Ninety-eight (2 female and 96 male) spinal cord injured subjects with lesions ranging from C2 to T12 neurological level (61 cervical and 37 thoracic lesions), sustained one month to 38 years prior to the examination (mean 7.9 $\pm$ 7.7 years) were recruited from the spinal cord injury service of a Veteran's Affairs Medical Center (VAMC). There were 39 ASIA Impairment Scale [1] (AIS) A's, 28 AIS B's, 18 AIS C and 12 AIS D subjects. The purpose of the study was explained to the subjects and informed consent was secured under a protocol approved by the local Institutional Review Board for Human Research. Subjects were relatively uniformly distributed in severity of spasticity (many were on antispastic medications). Their Ashworth

Figure 1: Schema of electrode placement

scores [2] ranged from 0 (flaccid) to 3 (marked resistance to passive movement), with two exceptions with a score of 4.

2.2 Data Collection

The overall purpose of the study was to compare clinical and neurophysiological measures of spasticity or altered motor control [13] after spinal cord injury. Data were collected in strict accordance with the BMCA (Brain Motor Control Assessment) protocol [9], beginning with 5 minutes relaxation followed by reinforcement maneuvers, voluntary maneuvers, passive maneuvers, tendon taps, clonus, application of vibration and finally plantar stimulation. For this paper, only the passive maneuvers are examined.

Subjects were placed in a comfortable, supine position and pairs of surface electrodes were placed bilaterally (Figure 1) centered over the long axis of the muscle bellies, 3 cm apart after preparing the skin to reduce electrode impedance below 5 kΩ measured at 30 Hz. Recordings were taken from 5 major muscle groups of each lower limb (total 10 channels) including the thigh muscles: quadriceps (q), adductors

(a), hamstrings (h), and leg muscles: tibialis anterior (ta) and triceps surae (ts) bilaterally. sEMG data were amplified 5000 times with a (-3 db) bandwidth of 40-600 Hz, using Grass 12A6 amplifiers (Grass Instruments, Quincy, MA). An event mark to denote timing of protocol commands was recorded along with the sEMG data.

After electrode placement, a physician carried out the clinical examination [6] incorporating standard scales for the level and severity of lesion, muscle strength and spasticity. To assess spasticity using the Ashworth Scale [2], he moved each lower limb through the range of motion in hip and knee joints in a single maneuver and graded the resistance felt. A single score to full maneuver (hip/knee flexion and extension) was given for each limb.

For the BMCA recording itself, each maneuver was repeated 3 times. Passive maneuvers for each side consisted of hip and knee flexion (first phase) then extension (second phase), followed by ankle dorsi- and plantar flexion. Each phase of each maneuver was maintained for a minimum of 5 seconds to provide time for the subject's responses to plateau.

2.3 Data Preprocessing

Clinical data were scored according to published clinical scales as previously described [6]. The sEMG data were analyzed from the EMG envelope calculated post-hoc using a root mean square (RMS) algorithm [3] (Figure 2). The sEMG recordings from relaxed healthy individuals are quiescent for the passive maneuvers described in this paper. Thus, an abnormal or spastic response is any non-zero activity resulting from passive maneuver attempts. Therefore, determination of abnormality could be based on a single number per channel per phase per maneuver. The average activity over 5 second window of each maneuver phase was corrected for baseline by subtraction of the average activity in the second immediately preceding

Figure 2: Surface EMG waveforms from passive limb movement.

the maneuver [10].

Figure 2 shows an example sEMG waveform with the computed envelope. Shown is the passive hip and knee flexion and extension of the right leg in a spastic spinal cord injured subject. The darker superimposed traces are the envelopes of the same data, computed as the root mean square (RMS) of the full-bandwidth data in such a manner as to reduce the effective sample rate of the envelope to 20 samples/second, from an original sample rate of 1800 samples/second.

For the purpose of this paper, the RMS values obtained from two phases of each maneuver were averaged across different muscles yielding a single number per muscle per maneuver. The sEMG data were thereby reduced to a set of numbers (features), each of which represented the average response of the muscle (q, a, h, ta, ts) to each maneuver (hip/knee, ankle). Each leg (total of 196 observations) was therefore represented by a single AS and by 10 sEMG-based neurophysiological features (5 muscles x 2 maneuvers). The initial set of features was later extended with 16 additional features after averaging original ones (see Tables 1 to 3). For example, the feature AVE(ta_{hk}, ts_{hk}) represents an average sEMG of two leg muscles during hip/knee maneuver.

Based on the AS, two classes were formed: AS $<$ 2 and AS $\geq$ 2, a margin which divides a slight from a more marked increase in muscle tone (56.7% vs. 43.3% of all instances, respectively). The complete data set used in the experiments thus included 196 data instances belonging to one of the two classes and described with 26 features.

3 Classification Methods

In the paper we experimentally evaluate the appropriateness of three different classification methods: top-down induction of decision trees (TDIDT), k-nearest neighbor (k-NN), and linear discriminant analysis. Given a data set, a TDIDT machine learning algorithm develops a classifier in the form of a decision tree [8]. A decision tree is a representation of a decision procedure for determining the class of a given instance [11]. Each node of the tree specifies either a class name or a specific test that partitions the space of instances at the node according to the possible outcomes of the test. Each subset of the partition corresponds to a classification subproblem for that subspace of instances, which is solved by a subtree. Leaf nodes contain class names. An instance is classified by finding a corresponding path from the root of the tree to one of its leaves.

Starting with the whole data set, a TDIDT algorithm first develops a condition, which is typically based on a single feature that best discriminates between the classes. Various measures are used for this purpose, such as information measure, GINI index of diversity, and gain-ratio measure [5]. The condition splits the data set to two or more data sets. The decision-tree development process is continued recursively on the new data sets until the resulting sets sufficiently well represent a single class. Most commonly, the process is stopped either when all the instances in a data subset belong to a single class, either the subset is empty, or the data subset cannot be further split using the available features.

The result of a TDIDT algorithm is a decision tree that is both a representation of the underlying data and can be used to classify new, previously unseen instances. In this work, we use a well-known implementation of TDIDT called C4.5 [7].

k-NN is an instance-based classifier that, given an instance to classify, searches the data set for k instances that are the closest to that instance according to a suitable distance measure. The instance is then classified to the most frequent class among its neighbors. We use a D. Wettschereck's implementation of k-NN called NGE [12], the Euclidean distance measure, and $k = 10$.

Linear discriminant analysis [4] is a statistical method that for a two-class problem aims to derive a hyper-plane in the form

$$a_0 + a_1 x_1 + \ldots a_n x_n = 0$$

where a_i are the coefficients and x_i are the features used in the data set. The coefficients are derived from the data set so that the hyper-plane best discriminates between the two classes. We use a maximum likelihood linear discriminant method, where the hyper-plane dividing two classes is drawn so as to bisect the line joining the centers of those classes, and the direction of the hyper-plane is determined by shape of the clusters of instances of the same class. The equation of the hyper-plane in this case is

$$\mathbf{x}^T S^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)^T S^{-1}(\mu_1 - \mu_2) = 0$$

where μ_i denotes the population mean for class i, and S the covariance matrix.

To assess the classification accuracy of the methods, we used a stratified 10-fold cross-validation [4]. This divides the data set to 10 sets of approximately equal size and equal distribution of classes. In each experiment, a single set is used for testing the classifier that has been developed from the remaining nine sets. The classification accuracy of each method is then assessed as an average of 10 experiments. The same training and testing data sets were used for all classification methods.

4 Experiments

The experiments were performed in the following steps:

1. given a data set of 196 instances and 26 features, select a subset of features to be used,
2. for each of the three methods, assess the classification accuracy,
3. where possible, interpret the resulting classifier.

Twenty different feature subsets were investigated. As for a relationship between the AS and the maneuver type, four possibilities were explored: sEMG obtained from hip/knee maneuver only, ankle maneuver only, hip/knee and ankle maneuvers combined, and an sEMG averaged across hip/knee and ankle maneuvers (Table 1, column labels). Similarly, activity from five recorded muscles was tested independently (q, a, h, ta, ts) or in four various combinations that include an average of only thigh muscles, an average of only leg muscles, an average of all muscles, or an average of thigh and leg muscles, respectively.

5 Results

The results of 20 experiments are presented in Tables 1 to 3. It is clear that the C4.5 and k-NN classification algorithms outperform the discriminant analysis, which for only a single experiment (row 2, third column in Table 3) reaches the margin of 56.7% - the a priori probability of the majority class. Based on these observations, it can be concluded that the discriminant analysis is not an appropriate classification method for the problem domain and sets of features investigated in this paper. Thus, further analyses focused on results of the C4.5 and k-NN classification algorithms.

	hip/knee	ankle	hip/knee, ankle	AVE(hip/knee, ankle)
q,a,h,ta,ts	63.7 ±11.7	60.8 ±7.5	63.7 ±11.7	67.4 ±10.3
AVE(q,a,h)	**74.0 ±8.9**	61.3 ±9.7	**74.0 ±8.9**	73.0 ±10.7
AVE(ta,ts)	60.7 ±11.2	56.2 ±11.7	59.1 ±10.3	66.4 ±10.4
AVE(q,a,h,ta,ts)	70.5 ±7.5	65.4 ±10.3	70.5 ±7.5	60.8 ±11.2
AVE(q,a,h),AVE(ta,ts)	**73.5 ±9.0**	60.8 ±10.2	72.4 ±7.3	73.0 ±10.7

Table 1: Classification accuracies of C4.5

	hip/knee	ankle	hip/knee, ankle	AVE(hip/knee, ankle)
q,a,h,ta,ts	68.0 ±7.7	60.9 ±11.4	70.0 ±10.9	68.9 ±8.7
AVE(q,a,h)	72.0 ±10.5	62.3 ±5.6	**74.6 ±11.8**	72.0 ±10.8
AVE(ta,ts)	67.0 ±11.9	57.7 ±8.6	68.5 ±12.2	62.9 ±11.6
AVE(q,a,h,ta,ts)	69.5 ±9.3	63.3 ±7.3	70.5 ±6.5	64.4 ±8.6
AVE(q,a,h),AVE(ta,ts)	**75.6 ±7.4**	62.3 ±9.5	72.5 ±10.8	67.3 ±8.2

Table 2: Classification accuracies of k-NN

	hip/knee	ankle	hip/knee, ankle	AVE(hip/knee, ankle)
q,a,h,ta,ts	55.5 ±8.0	42.7 ±14.9	53.4 ±9.0	50.7 ±10.0
AVE(q,a,h)	53.1 ±7.1	50.5 ±13.9	58.0 ±7.6	53.6 ±8.6
AVE(ta,ts)	50.1 ±10.0	45.2 ±13.1	44.1 ±16.3	51.5 ±10.9
AVE(q,a,h,ta,ts)	53.1 ±4.7	50.4 ±10.5	53.4 ±12.2	53.6 ±8.1
AVE(q,a,h),AVE(ta,ts)	54.1 ±9.4	53.0 ±11.0	49.9 ±9.4	53.0 ±10.4

Table 3: Classification accuracies of linear discriminant analysis

The k-NN performs slightly better than C4.5 for most experiments, but the differences are not significant at .05 level. For easier comparison, the classification accuracies higher than 73.0% in Tables 1 and 2 are printed in bold. These are in fact the experiments where the classification accuracy is significantly better (.05 significance level) than in the remaining experiments using C4.5 and k-NN, respectively. In can be seen that there are two such feature subsets for k-NN and three feature subsets for C4.5. Interestingly, two subsets yielding the best classification accuracy are shared between k-NN and C4.5. Thus, a common set of features is found consistently superior using two different algorithms.

When comparing the first two columns of Tables 1 and 2, respectively, it can be seen that sEMG features derived from the hip/knee maneuver are better predictor of the AS classes than those from the ankle maneuver for any muscle combination tested. Interestingly, for C4.5, when hip/knee and ankle maneuvers were taken together but considered separately (third column), the classification accuracy, compared to the hip/knee maneuver only (first column), was either spoiled (rows 3 and 5) or exactly the same (rows 1,2 and 4). The respective decision trees, constructed for the experiments where the classification accuracy was equal, yielded identical results, i.e. none of the features derived from the ankle maneuver of the hip/knee-ankle data set was used. Another interesting finding is that by averaging the muscle activity across the two maneuvers (fourth column) the classification accuracy slightly improved in three experiments or remained the same (rows 1 and 3 for C4.5 and rows 1 and 2 for k-NN). In majority of cases, however, the performance of two classifiers dropped.

As linear discriminant analysis performs poorly and k-NN only implicitly builds a classification model, we have further investigated only the classification trees obtained by C4.5 for the three best cases which are marked bold in Table 1. For those two that use the average activity of thigh muscles only (row 2) the decision tree constructed using the whole data set was the same and consisted of one internal node only (single condition). For the case where both thigh and leg muscles are used (row

5, first column), the decision tree was more complex. However, the most accurate decision trees were composed of the same average thigh muscle activity located in the root of the tree thus indicating that this feature is the most predictive one.

6 Discussion

This study revealed an existence of relationship between the sEMG and the AS for patients who sustained SCI. This in fact confirmed that neurophysiological assessment of spasticity based on carefully selected sEMG features is related to clinical assessment using the AS.

Of the three classification methods tested, k-nearest neighbor and top-down induction of decision trees by C4.5 outperformed the classifier by the linear discriminant. The performance of k-NN and C4.5 was comparable in most cases and there were no significant differences. Because k-NN does not explicitly develop a classification model, and because the decision trees of C4.5 were for the best cases relatively simple and transparent, the C4.5 may be an algorithm of choice among the three methods investigated.

After the relationship was established two questions emerged. First, which maneuver(s), and second, which sEMG feature(s) are the best predictors of the AS classes. The average sEMG activity recorded during hip/knee maneuver appeared to be an excellent predictor whether the patient's AS was less than 2 or 2 and higher. Conversely, ankle maneuver seemed to be a poor predictor of the AS classes. The features obtained by combining the two yielded similar results regardless of whether sEMG from the two maneuvers was averaged or considered independently. This was a consistent finding using either C4.5 or k-NN. Developed decision trees reveal, however, that even in those cases the sEMG features obtained from the ankle maneuver play a minor role. Thus, it can be concluded that neurophysiological features derived from hip/knee maneuver are the best predictor of the AS classes.

As for the second question, the sEMG features derived from the thigh muscles only (row 2) yielded the highest prediction accuracy for all maneuvers, excluding the ankle maneuver, using C4.5. The same was true for k-NN except in one case (column 1). Furthermore, three of the five cases that showed significantly better prediction accuracy included the average of thigh muscles only. The remaining two cases, in addition to the average of thigh muscles, included also the average of leg muscles (row 5). When the decision tree for the latter was developed, it consisted of only one node represented by the average of thigh muscle activity. Therefore, even in those cases the most predictive was the sEMG recorded from the thigh muscles. Prediction based on the sEMG of the leg muscles (row 3) was consistently the worst. Thus, it appears that the features based on the thigh muscles' sEMG are the best predictor of the AS classes.

In order of their predictive accuracy, some of the best decision trees were also the simplest ones: they used the average sEMG of the thigh muscles and included hip/knee maneuver. If, for example, the average sEMG during hip/knee maneuver was larger than 3.6, the subject's AS was likely 2 or greater. Our results, therefore, show that in order to make a comfortable prediction in which of the two AS classes SCI patients belong to, the hip/knee maneuver and the respective sEMG features of the thigh muscles need to be selected. This may not be a surprise since multiple pairs of electrodes placed over the thigh picked up the activity from several large muscle groups that act upon the hip and/or knee joints. Also, the AS in this study represented a subjective impression of the entire limb resistance felt during passive movements occurring mostly in hip and knee joints, rather than resistance of the muscle groups acting upon individual joints. In other words, the examiner himself graded the "average" resistance of the limb that seems to be related to the average sEMG activity of the thigh muscles. Neither ankle maneuver nor the muscles acting upon the ankle (ta,ts) were good predictors of the AS. Thus, the classification methods used in this study clearly pinpointed the key neurophysiological features that may have been anticipated, should the AS be related to sEMG.

This study shows that neurophysiological assessment can supplement clinical examination in attempt to assess spasticity after SCI. More importantly, neurophysiological assessment, due to its sensitivity, has the potential to reveal subclinical features that may remain undetected otherwise. Also, it provides a mean to objectively follow-up SCI patients and to evaluate the effectiveness of various clinical interventions aiming to modify spasticity or other aspects of altered motor control after SCI.

7 Conclusion

To conclude, for C4.5 and k-NN, the classification accuracy in most cases was well above the prior probability of the most probable class (56.70%). The results thus indicate that neurophysiological assessment based on sEMG is related to the AS and can be used to predict the clinical findings. The results also show that prediction accuracy can be improved by carefully selecting neurophysiological features. Future work is needed to demonstrate usefulness of the sEMG-based comprehensive assessment of spasticity that may provide a more objective assessment of altered motor control after SCI. The future work should also take the advantage of similar approaches to aid the analysis of clinical and neurophysiological data.

So far, the protocol described has been used for research purposes only. In the past, its widespread clinical use was limited by high technical requirements on one side, and by insufficient evidence of its clinical usefulness on the other. This and several other studies were undertaken to justify its applicability in clinical settings. Our current results suggest that the comprehensive sEMG-based assessment of spasticity should be a complementary to the already existing clinical scales. Although the specificity of the method is yet to be proven, its higher sensitivity can help to reveal subclinical features, frequently left undetected on clinical examination, thus providing a more objective assessment of altered motor control after SCI.

So far, the protocol described has been used for research purposes only. In the past, its widespread clinical use was limited by challenging technical requirements on the one hand, and by insufficient evidence of its clinical usefulness on the other. This and several other studies were undertaken to justify its applicability in clinical settings. Our current results suggest that the comprehensive sEMG-based assessment of spasticity should be viewed as complementary to existing clinical scales. Although the specificity of the method is yet to be proven, its higher sensitivity can help to reveal subclinical features, frequently left undetected on clinical examination, thus providing a more objective assessment of altered motor control after SCI. Furthermore, the high degree of reproducibility demonstrates its suitability for serial studies in the same subject.

References

[1] American Spinal Injury Association. *International Standard for Neurological and Functional Classification of Spinal Cord Injury.* ASIA, Chicago, IL, 1992.

[2] B. Ashworth. Preliminary trial of carisoprodol in multiple sclerosis. *Practitioner*, 192:540–542, 1964.

[3] J. V. Basmajian and C. J. De Luca. *Muscles Alive: Their Functions Revealed by Electromyography.* Williams and Wilkins, Baltimore, fifth edition edition, 1985.

[4] D. Michie, D. J. Spiegelhalter, and C. C. Taylor, editors. *Machine learning, neural and statistical classification.* Ellis Horwood, 1994.

[5] J. Mingers. An empirical evaluation of selection measures for decision-tree induction. *Machine Learning*, 3:319–342, 1989.

[6] M. M. Priebe, A. M. Sherwood, N. F. Kharas, and J. Markowski. Clinical assessment of spasticity in spinal cord injury: A multi-dimensional problem. *Arch Phys Med Rehab*, 77:713–716, 1996.

[7] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

[8] R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[9] A. M. Sherwood, W. B. McKay, and M. R. Dimitrijević. Motor control after spinal cord injury: Assessment using surface emg. *Muscle & Nerve*, 19:966–79, 1996.

[10] A. M. Sherwood, M. M. Priebe, and D. E. Graves. Consistency of multichannel surface emg recordings: Application in spinal cord injured subjects. *J Electrophysiol Kinesiol*, 7:97–111, 1997.

[11] P. E. Utgoff. Incremental induction of decision trees. *Machine Learning*, 4(2):161–186, 1989.

[12] D. Wettschereck. A hybrid nearest-neighbor and nearest-hyperrectangle algorithm. In *Proc. 7th European Conference on Machine Learning*, pages 323–335, 1994.

[13] R. R. Young. Treatment of spastic paresis (editorial). *New England Journal of Medicine*, 320:1553–1555, 1989.