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

Muscle Categorization Using PDF Estimation and Naive Bayes Classification

Tameem M. Adel, Benn E. Smith and Daniel W. Stashuk, *Member, IEEE*

Abstract— **The structure of motor unit potentials (MUPs) and their times of occurrence provide information about the motor units (MUs) that created them. As such, electromyographic (EMG) data can be used to categorize muscles as normal or suffering from a neuromuscular disease. Using pattern discovery (PD) allows clinicians to understand the rationale underlying a certain muscle characterization; i.e. it is transparent. Discretization is required in PD, which leads to some loss in accuracy. In this work, characterization techniques that are based on estimating probability density functions (PDFs) for each muscle category are implemented. Characterization probabilities of each motor unit potential train (MUPT) are obtained from these PDFs and then Bayes rule is used to aggregate the MUPT characterization probabilities to calculate muscle level probabilities. Even though this technique is not as transparent as PD, its accuracy is higher than the discrete PD. Ultimately, the goal is to use a technique that is based on both PDFs and PD and make it as transparent and as efficient as possible, but first it was necessary to thoroughly assess how accurate a fully continuous approach can be. Using Gaussian PDF estimation achieved improvements in muscle categorization accuracy over PD and further improvements resulted from using feature value histograms to choose more representative PDFs; for instance, using log-normal distribution to represent skewed histograms.**

I. Introduction

An electromyographic (EMG) signal is a voltage signal created by sampling the electric fields created by a contracting muscle. More specifically, the electrical activities of motor units recruited during muscle contraction summate to create an EMG signal [1]. In the literature, there are deterministic as well as probabilistic techniques that can be applied to extract as much information as possible from EMG signals. It is better to start with an abstraction and some definitions so that the reader can be more familiar with the literature used in this field. A muscle consists of muscle fibers. The muscle fibers of each muscle are grouped according to the specific α-motor neuron that innervates them. A single α-motor neuron and the muscle fibers it innervates are referred to as a motor unit (MU) [1], [2]. There are two forms of EMG; surface and intramuscular EMG. Accordingly, there are two types of electrodes used to detect EMG signals; surface or needle electrodes, respectively. With the electrode suitably positioned, the examined muscle is voluntarily activated and produces time varying voltage fields which are detected as EMG signals. The decision of which method to use depends on the objectives and scope of the investigation. Clinically detected intramuscular EMG signals are the focus of this paper.

T. M. Adel is at the University of Waterloo department of Systems Design Engineering. (e-mail: thesham@uwaterloo.ca).

D. W. Stashuk, is with the University of Waterloo department of Systems Design Engineering. (e-mail: Stashuk@uwaterloo.ca).

The clinical categorization of an examined muscle, acquired through EMG data, is used to support diagnosis and treatment of neuromuscular disorders. The objective is to build a system that automates diagnosis by categorizing an examined muscle. It is intuitive to assume that the MUs comprising a muscle being examined are used to form the basis upon which a muscle categorization is determined. In [3], Pino, Stashuk and Podnar proposed using a Naive Bayes classifier to aggregate decisions over MUs and categorize a muscle. They considered two muscle categories; normal and neuropathic. In this work, a Naive Bayes classifier that considers three categories (normal, neurogenic and myopathic) is applied. As the names suggest; a muscle that belongs to the normal category is a healthy muscle. A muscle belonging to the myopathic category suffers from a myopathic disorder, and correspondingly so for the neurogenic category. When there is a myopathic or neurogenic disorder in a muscle, the defining characteristics of its action potentials differ accordingly and this is why EMG signal analysis can be used in supporting diagnosis of these muscles. For example, action potentials of myopathic muscles usually have shorter durations than normal muscles while those belonging to neurogenic muscles usually have longer durations and much higher amplitude values.

The electrical activity of a single motor unit is represented by a motor unit potential train (MUPT). For each MUPT, there is a MUP template which is considered a representative of its respective MUPT [11]. MUPTs need to be characterized to help predict from what category of muscle they were detected. This can be done using several techniques; statistical or probabilistic; for example, a probabilistic approach accomplished by pattern discovery techniques is used in [4]. The methods proposed in this paper are probabilistic as well.

Specifically, each set of MUPTs represents the electrical activity of a sample of MUs from the muscle being examined. Each MUPT is represented by features of its MUP template, the stability of its MUP shapes and MU firing pattern statistics. The MUPTs are characterized by a set of conditional probabilities, one for each category, using either pattern discovery (PD) or PDF estimation. Bayes rule is then used to aggregate these MUPT characterization probabilities to obtain a set of muscle categorization probabilities. In the training dataset, MUPTs are represented by their respective numerical feature values. Out of all the features currently calculated, a feature set was chosen to be used for MUPT characterization. Pattern discovery (PD) is a technique that is built on quantizing feature values into events and using these events to create rules upon which classification is based. Even though there is some information lost due to the quantization step, PD has the added value of being more transparent [4]; the categorization of a muscle can be explained to clinicians more easily. When PDF estimation is

used for MUPT characterization, the training values of each feature and each category are used to estimate probability distributions and a Naive Bayes (NB) classifier is used to aggregate the resulting probabilities of each feature given a certain category to obtain the MUPT characterization probability given this category. In the testing phase, probability distributions formed in training are used along with values comprising the MUPTs of the muscle being examined to obtain MUPT characterization probabilities. Using PDFs does not require discretization as it is fully continuous. Therefore, there is no information lost as is the case when PD is used.

Ultimately, this paper evaluates the muscle categorization accuracies achieved when using a probabilistic approach that is built on either estimating PDFs to estimate MUPT characterizations or using PD. The main goal of this paper is to evaluate the difference in terms of accuracy between the best variations of the PDF estimation approach and the PD approach. Future work is to perform more investigation regarding how to enhance the accuracy of a hybrid technique that is built on both (continuous) PDF estimation and discrete PD, without losing transparency. Some other related issues, like feature selection and dependence relationships between features, are handled in this work as well.

II. METHOD

A. Muscle Categorization Using Bayesian aggregation

In [3], Pino et al. proposed using Bayesian aggregation to combine sets of MUPT conditional probabilities into a single set of probabilities, containing, for each muscle category, the probability of a muscle belonging to that category conditioned on the set of sampled MUPTs. As mentioned above, they used two categories; therefore the prior probabilities were 0.5 for each category. Equation (1) depicts the Bayesian probabilistic model concluded from Bayes theorem after assuming that all prior probabilities of the three categories are equal [3]:

$$P(muscle = y_k | \{MUP_1, MUP_2, \ldots, MUP_N\}) = \frac{\prod_i^N P(y_k|MUP_i)}{\sum_{j=1}^{3}(\prod_i^N P(y_j|MUP_i))} \quad (1)$$

Where *muscle* = y_k means that *muscle* belongs to category y_k and $\{MUP_1, MUP_2, \ldots, MUP_N\}$ is the set of N MUPs sampled from the muscle; MUP_i is the ith MUP of the set $\{MUP_1, MUP_2, \ldots, MUP_N\}$

Fig. 1 shows an example of how (1) works; it denotes the Bayesian aggregation when there are N MUPs and three categories as is the case with the methods proposed in this work. The pies on the left hand side show the characterization of each MUPT. The light sector denotes the probability that an MUPT is 'normal', the dark sector denotes 'myopathic' while the darker sector refers to the possibility it is 'neurogenic'. Using (1), the probabilities of MUPTs are aggregated and the categorization probability of the whole muscle is obtained. As depicted in Fig. 1, there is approximately a 53% probability that the muscle in question is normal, 38% it is myopathic and 9% it is neurogenic [3].

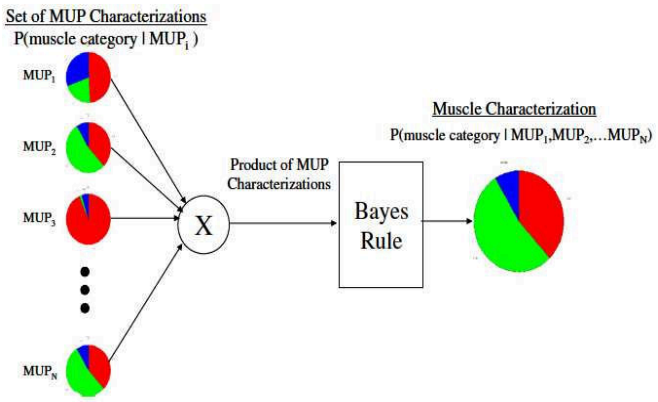

Figure 1. A NB classifier applied to aggregate probabilities of MUPTs [3]

B. MUPT Characterization using Pattern Discovery (PD)

Pattern discovery is an information theory based technique established on detecting significant patterns in the data and using these patterns in classification. PD was first introduced by Wong and Wang [5]-[8]. PD is applied on discrete data and, as a result, a quantization step is needed for each feature that has continuous data values, which is the case with most EMG features. The number of discretization bins can be identified according to the nature of the problem at hand and the dataset used. For instance, if the number of bins is three; low, medium and high for each feature, there might be some inaccuracy according to putting "very high" and "slightly high" values in the same bin. On the other hand, using five bins; very low, low, medium, high and very high would relatively solve such a problem but more training examples will be needed to keep the same number of patterns per bin. In the PD classification algorithm, the first step is to discover the "significant" patterns; patterns that are repeated more often than expected assuming a random occurrence [9]. Rules are composed of patterns in addition to the specific muscle category they represent. The order of a rule is equal to the number of features plus the muscle category (normal, myopathic or neurogenic in our work) to which the pattern belongs. For example, a high amplitude pattern in the neurogenic MUPTs is a 2nd order rule. Each rule has a weight of evidence (WOE) that denotes how much evidence the rule holds in support of a certain category [4]. For rule selection during testing, the highest order rule for each category is selected first. WOEs of selected rules are added to be normalized and this process continues until there are no more rules or all features have already been included in the previously selected rules [9]. In addition to the well known pros and cons of discretization, characterizations performed by PD are transparent, and this is very significant because the reason for not using a "black box" method like support vector machines or neural networks is their inability to provide clinicians with clear rationale for the classification process; something a clinical decision support system cannot do without. When PD is used, there is a decrease in accuracy due to discretization performed on the continuous MUPT feature values.

C. MUPT Characterization using a Naive Bayes classifier and PDF estimation

Each MUPT is represented by its MUP template, MUP shape stability and MU firing pattern feature values. A

feature set is selected first. Then a PDF is estimated for each feature given a category and the PDF parameters are estimated using maximum likelihood estimation (MLE). Histograms of each feature along with each category are used to choose estimated PDFs. A Naive Bayes (NB) classifier is utilized to aggregate the conditional feature probabilities and obtain each MUPT characterization probability given a certain category as in (2)

$$P(MUPT = y_k|\{F_1, F_2, \dots, F_N\}) = \frac{P(y_k)\prod_i^N P(F_i|y_k)}{\sum_{j=1}^{3}(\prod_i^N P(F_i|y_j))} \quad (2)$$

Where $MUPT = y_k$ means that *MUPT* belongs to category y_k and $\{F_1, F_2, \dots, F_N\}$ is the set of N features selected for characterization. $P(y_k)$ is the prior probability of the MUPT category y; for this work the prior probabilities of the three categories were assumed to be equal.

In all the proposed PDF estimation techniques, values of each feature in each category (label) are supposed to be drawn from a specific PDF. The accuracy of characterization depends on how close to the truth the estimated PDF and the estimated parameters are. Histograms show that most features do not follow a Gaussian distribution; and that assuming they are drawn from a log-normal distribution is closer to the truth. Characterization was performed using the Gaussian PDF assumption first and then the log-normal assumption to assess the difference between both. The PDF equations of the Gaussian and log-normal distributions are given in (3) and (4) respectively.

$$f_x(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (3)$$

$$f_x(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}} \quad (4)$$

Parameters μ and σ in (3) and (4) are estimated using maximum likelihood estimation (MLE) by calculating the mean and standard deviation of the values and log of the values, respectively, in the training data. Due to the fact that unknown parameters are being estimated, sample values obtained for each of them can contain errors; even with MLE searches to achieve minimum error. The larger the size of the training data, the more accurate the MLE parameter estimates, as the variance of estimated parameters gets closer to the Cramér–Rao bound. Using the selected feature values belonging to each MUP of the studied EMG signals, characterization probabilities of every MUP are calculated given each label. A NB classifier is used to aggregate characterization probabilities of the selected features to get the MUP characterization.

As the histogram of every feature can be displayed and characterization probabilities of each feature are calculated separately, there is no need to assume the same PDF for all the features. It is more precise to manually check histograms and choose the PDF that looks closer to the shape of the corresponding histogram. For example, after inspecting the "thickness" feature histogram, it can be concluded that its data are more likely to be drawn from a Gaussian PDF than a log-normal PDF. This proposed technique is referred to in the rest of this paper as the NB adaptive log-normal PDF estimation technique.

In the early eighties, Vianelli proposed a generalized version of the log-normal distribution [10]. This version has an additional shape parameter *r*, and the original log-normal distribution represents this generalized version with $r = 2$ [10]. The PDF of the 3-parameter generalized log-normal, also known as the exponential power distribution is given in (5) [10].

$$f_x(x; \mu, \sigma_r, r) = \frac{1}{2xr^{(1/r)}\sigma_r\Gamma(1+1/r)} e^{-\frac{|\ln x-\mu|^r}{r\sigma_r^r}} \quad (5)$$

The parameter μ is the mean as in (3), while $\sigma_r = [|\log x - \mu|^r]^{1/r}$ and $r > 0$ is another shape parameter. $\Gamma(1+1/r)$ is the value of the Gamma function for $(1+1/r)$

III. RESULTS

Data used were acquired under IRB approval and sanitized of any personal identifying information and then processed using decomposition-based quantitative electromyography (DQEMG). Data are gathered in a database that is used in both training and testing. This database contains information from 371 muscles in four groups, as per their location in the body; forearm, arm, lower leg and upper leg. As the forearm group contains only three myopathic muscles, it was excluded from these experiments as three muscles are considered very few for one category and do not provide enough information for classification. There are three muscle categories; normal, myopathic or neurogenic. In the database, there are entries for 9822 MUPTs extracted from EMG signals acquired during 1408 contractions. There are currently five groups of features calculated following signal decomposition that describe each MUPT [11]. One group contains MUP template size features like area and amplitude, another group contains MUP template shape features. The third group contains features that describe MUP template complexity like fiber count while the fourth has features describing the stability of the MUPs in the MUPT. The fifth group includes features that describe the MU firing pattern. The MUP stability and MU firing pattern features represent the whole MUPT, not only the MUP template [11].

Apart from the sixth row, the set of features used to obtain the results displayed in all of the other rows of Table I, is referred to as Feature Set 1. This feature set consists of turns, area, thickness & Ajiggle [12]. It achieved higher accuracy than all other feature sets consisting of 1 to 5 features with NB adaptive Log-normal, apart from the feature set used to obtain the results in the sixth row, as will be clarified later. Leave one out cross validation was used to obtain all the accuracy values shown in Table I.

The first two rows of Table I show the accuracy results when using PD to characterize the MUPTs. Classification accuracy using PD with 5 quantization bins to discretize the continuous feature values is clearly better than when only 3 bins are used.

Histograms of most features are positively skewed; which supports the assumption that such features are drawn from a log-normal distribution. Nonetheless, a trial was still

implemented using the Gaussian PDF assumption, just to compare it with the other techniques. The third row of Table I shows the results of applying the NB Gaussian technique. The more plausible approach of assuming log-normal PDFs was then applied and the results are displayed in the fourth row of Table I. The log-normal distribution estimation technique has a higher average accuracy, which is in line with the information obtained from the histograms. The fifth row of the Table I exhibits the results of using the NB adaptive log-normal PDF estimation technique. Looking at the histograms of the features in Feature Set 1, the "thickness" feature was assumed to be drawn from a Gaussian distribution rather than log-normal. The average accuracy of the adaptive log-normal technique is 2.1% higher than the rigid log-normal technique. The seventh row displays the accuracy results when using the generalized log-normal estimation technique. The average accuracy is 76.7; it is 1.4% higher than the adaptive log-normal technique. The value of r was selected to be 2.25 empirically as this is the value that led to the highest average accuracy (the range of values of r from 1.7 to 2.4 led to average accuracy values higher than r values outside this range). For all the continuous PDF methods (rows 3-8), classification error was higher for myopathic muscles than for those with a neurogenic disorder.

NB classifiers assume independence between features. As this is not always true in reality, it was intriguing to check what happens if the dependence relationships between the features are taken into consideration. Correlation values were used to express this dependence. The last row of Table I shows the results of using Feature Set 1 and the NB adaptive Log-normal technique without assuming independence between features. The average performance did not improve. Taking into account the flexibility induced by the independence assumption and the fact that it is faster, more straightforward and does not require the storing of feature correlation values for each category, it is clear that assuming independence is valid and is favored. The independence assumption was applied to get the accuracy results for all the techniques introduced in this work.

Feature Set 1 contains features from four different groups of features (review feature groups described earlier), which leads to a more comprehensive description of the MUPTs classified and that was confirmed by applying an exhaustive feature selection technique for all (1 to 5 features) feature sets. The technique is wrapper-based as classification accuracy is the criterion used to judge the quality of a feature set. Using NB adaptive Log-normal to test all the (1 to 5) feature combinations (sets), the only competitive feature set to set 1 was Feature Set 2 consisting of turns, amplitude, thickness & Ajiggle. Feature Set 2 led to a better accuracy result for both the lower leg and upper leg muscles while the NB adaptive log-normal techniques using Feature Set 1 for the arm muscles had better accuracy than Feature Set 2 as shown in the sixth row of Table I.

TABLE I. ACCURACY RESULTS OF ALL THE TECHNIQUES

Technique	Muscle group			
	Lower leg	*Upper leg*	*Arm*	*Average*
PD (using 3 bins)	54.3%	64.3%	61.7%	60.1%
PD (using 5 bins)[a]	73.0%	69.2%	62.9%	68.4%
NB Gaussian PDF	74.2%	69.3%	75.1%	72.9%
NB log-normal PDF	76.0%	71.5%	72%	73.2%
NB adaptive log-normal PDF	78.8%	71.0%	75.9%	75.3%
NB adaptive log-normal with features in Feature Set 2	80.2%	72.2%	74.1%	75.5%
NB generalized log-normal PDF (r = 2.25)	**82.3%**	**70.4%**	**77.2%**	**76.7%**
NB adaptive log-normal (assuming features dependence)	78.2%	71.4%	75.6%	75.1%

a. All rows (methods) except the sixth use Feature Set 1 (defined earlier)

IV. CONCLUSION

Even though muscle categorization techniques, like other clinical decision support systems, should be transparent, it is still important to check how accurate a classification system would be if transparency was not an issue.

As things stand, using a fully continuous approach is more accurate than using a fully discrete and transparent PD approach. This work quantified the difference in terms of accuracy between both approaches. Now that the accuracy values of the PDF estimation technique are identified, using a hybrid approach should follow. Theoretically, there should be a decrease in accuracy as it is not going to be a fully continuous approach but the aim is to improve the accuracy of this hybrid approach without losing transparency.

REFERENCES

[1] H. Parsaei, D. W. Stashuk, S. Rasheed, C. Farkas, and A. Hamilton-Wright, "Intramuscular EMG Signal Decomposition," Crit Rev Biomed Eng, vol. 38, no. 5, pp. 435–465, 2010.

[2] C. Farkas, A. Hamilton-Wright, H. Parsaei, and D. W. Stashuk, "A review of clinical quantitative electromyography," Crit Rev Biomed Eng, vol. 38, no. 5, pp. 467–485, 2010.

[3] L. Pino, D. W. Stashuk, and S. Podnar, "Bayesian characterization of external anal sphincter muscles using quantitative electromyography," Clin Neurophysiol, vol. 119, pp. 2266–2273, 2008.

[4] L. Pino, D. W. Stashuk, S. G. Boe, and T. J. Doherty, "Motor unit potential characterization using 'pattern discovery'," Med Eng Phys, vol. 30, pp. 563–573, 2008.

[5] A. Wong, and Y. Wang, "Pattern discovery: a data driven approach to decision support," IEEE Trans Syst Man Cybern Part C: Appl Rev, vol. 33, pp. 114–124, 2003.

[6] A. Wong and Y. Wang, "High-order pattern discovery from discrete valued data," IEEE Trans Knowled Data Eng vol. 9, pp. 877–893, 1997.

[7] A. Wong and Y. Wang "Discovery of high order patterns," IEEE Trans Syst Man Cybern, vol. 2, pp. 1142–1147, 1995.

[8] Y. Wang, "High-order pattern discovery and analysis of discrete-valued data sets," Ph.D. Thesis, Univ. of Waterloo, Waterloo, 1997.

[9] L. Pino, "Neuromuscular Clinical Decision Support using Motor Unit Potentials Characterized by 'Pattern Discovery'," Ph.D. Thesis, Sys. Design Eng., Univ. of Waterloo, Waterloo, 2008.

[10] S. Vianelli, "The family of normal and lognormal distributions of order r," Metron, vol. 41, pp. 3-10, 1983.

[11] D. Stashuk, "Decomposition and quantitative analysis of clinical electromyographic signals," Med Eng Phys, vol. 21, pp. 389–404, 1999.

[12] E. Stalberg and M. Sonoo, "Assessment of variability in the shape of the motor unit action potential, the 'jiggle' at consecutive discharges", Muscle Nerve, vol. 17 pp. 1135-1144, 1994.