FISHER LINEAR DISCRIMINANT CONSTRUCTION FOR DISTRIBUTIONS OF DISEASED DATA

Andrew Hamilton-Wright,† Linda McLean‡
†Mount Allison University, New Brunswick; University of Guelph, Ontario
‡Queen's University, Kingston, Ontario

INTRODUCTION

When constructing a clinical diagnostic tool, one frequently turns to statistical methods in order to discern "diseased" data from that drawn from "normative" subjects. Particular care must be taken when using statistical methods in order to ensure that the underlying assumptions of the technique are valid in the area of application of interest.

Ideally, when using a two-outcome test, a "positive" test outcome is uniquely associated with instances of actual disease, and a "negative" outcome will be similarly associated with normative data. It is important to note the underlying assumption that the samples defining each class in a two-class discernment problem all represent the same disease state. This assumption cannot truly be made in most cases of diagnostic disease data. Consider the case when a patient begins to exhibit symptoms of a disease. Some fraction of the cells under study in a diseased subject will show signs of the disease involvement, while other cells may be completely healthy. Any measurement of these healthy cells will not show any symptoms, and should therefore be identified as negative by the diagnostic test. Stated simply: a data set acquired from a diseased patient will contain data representative of both the disease and of completely normal cells.

The application area of interest to the authors is quantitative electromyography (QEMG). QEMG data is an important tool that is beginning to be used in electrodiagnostic medicine (Brown, Bolton and Aminoff, 2002; Doherty and Stashuk, 2003; Preston and Shapiro, 2005). As the mechanism underlying the observation of electromyographic data is the proper function of individual muscle and nerve cells, the proportional involvement of individual cells will have a profound effect on whether or not the disease is detected.

In order to evaluate muscle performance, a series of motor unit potentials (MUPs) may be acquired; these form the smallest functional units of the muscle. If MUPs are acquired from a truly diseased patient, it is therefore expected that only a subset of these MUPs will be produced by diseased tissue. In turn, this implies that the set of MUPs observed in a contraction or study acquired from such a diseased individual will contain MUPs that should truly have a "positive" test outcome as well as "true negative" MUP samples.

This may cause a significant problem when attempting to use a statistical method to obtain a robust and accurate method to separate normative from diseased data. When considering Fisher's (1936) Linear Discriminant, it is apparent from the standard formulation (Duda, Hart and Stork, 2001, pp. 117-121) that good estimates of the mean and covariance of all data are required. If the composition of the diseased data includes a significant number of "negative-outcome" data points, then any estimation of the mean (or covariance) based on a the collection of all "disease" points will be severely biased.

METHODS

We propose a novel technique to calculate the vector forming the Fisher Linear Discriminant between two classes of data. In order to establish this vector, we rely on the fact that normative data are well understood and easy to acquire.

First, we obtain a mean and covariance for the normative (negative-outcome) data set, indicated by $\vec{\mu}_N$ and $\vec{\Sigma}_N$, respectively, and use these to calculate the Mahalanolbis distance (Duda *et al*, 2001, pp. 35), relative to the distribution of negative-outcome training points:

$$r_i = \sqrt{(\vec{x}_i - \vec{\mu}_N)^t \vec{\Sigma}_N (\vec{x}_i - \vec{\mu}_N)} \qquad (1)$$

Equation (1) provides a distance in units of standard deviation of any point $\vec{x}_i$ relative to the class distribution mean $\vec{\mu}_N$. Such a measure can be seen to produce the familiar *z*-scores used in conjunction with a table describing a Gaussian probability function. The Mahalanolbis distance describes the relative probability of association with the distribution described by $\vec{\mu}_N$ and $\vec{\Sigma}_N$, providing a weighting to indicate the likelihood of any point being associated

with this distribution, versus the likelihood of being an outlier, possibly associated with another class (in this case, the desired positive-outcome disease class).

The distribution of negative-outcome points will fill a region of some *m*-dimensional space. The points belonging to the "diseased" samples can be visualized as forming the tips of arrows emanating from this region. The volume around this region in which the majority of the arrows "point" can then be seen as the likely volume or site where those points lie that should truly be labelled with a positive outcome. If we preferentially count only the arrows that extend out of the normative distribution (by using the Mahalanolbis distance), then it becomes possible to distinguish between the desired positive and negative outcome points found in the study of a patient from the diseased set.

By considering the diseased data as a set of vectors projecting from the normative distribution, it is clear that further information is available by constructing a "neighbourhood" measure of the type found in Bezdek's (1981) fuzzy *c*-means or the Kohonen (1989) Self-Organizing Map. This information will provide a means of calculating an overall direction. This will, in turn, determine whether or not the vectors point towards the diseased state or are simply outliers of the normative distribution. This is achieved through the inter-vector angle

$$\theta_\delta(i,j) = \arccos\left(\frac{\vec{x}_i \cdot \vec{x}_j}{\|\vec{x}_i\| \|\vec{x}_j\|}\right), \quad (2)$$

constructed from the vectors and their norms and relating it to the underlying probability distribution function.

Once constructed, the value $\theta_\delta(i,j)$ is used as the distance in a neighbourhood function to determine the overall importance of a given vector.

This neighbourhood function value, ω_i, is calculated for every vector i in the list of all N_{Diseased} vectors acquired from diseased patients using

$$\omega_i = \sum_{\substack{j=1 \\ j \neq i}}^{N_{\text{Diseased}}} \frac{e^{-[\theta_\delta(i,j)]^2/2}}{\sqrt{2\pi}}. \quad (3)$$

Equation (3) therefore provides a neighbourhood weighting that will accentuate vectors that are close together over vectors that are distant from each other. By combining these ω_i values with the Mahalanolbis distance values (r_i) and the original vectors ($\vec{x}_i$) from the diseased data class, we can calculate an inter-mean vector

$$\vec{w}'_{N\to D} = \sum_{i=1}^{N_{\text{Diseased}}} \omega_i r_i \vec{x}_i. \quad (4)$$

The units of equation (4) are meaningless (and quite large), so the vector reported is converted to unit length by scaling it by its norm:

$$\vec{w}_{N\to D} = \frac{\vec{w}'_{N\to D}}{\|\vec{w}'_{N\to D}\|}. \quad (5)$$

RESULTS

The data set used for the analysis presented here contains contraction data obtained from 11 patients exhibiting pain characteristic of a particular repetitive strain injury (forming the diseased data set), as well as from 39 subjects (the normative data set) who were asymptomatic for repetitive strain. All subjects provided informed consent prior to participation.

Concentric needle (micro) and surface (macro) EMG data were collected from the *extensor carpi radialis brevis* muscle using a 32-gauge concentric needle electrode and silver/silver chloride surface electrodes located on the skin overlaying the needle acquisition site. Subjects performed repeated isometric contractions ranging between 5 and 20% of maximum voluntary contraction force. All acquisition was performed using Comperio™ clinical EMG amplifiers. Acquisition settings used were as reported in Calder, Stashuk and McLean (2006): 10Hz-10kHz bandpass for micro data sampled at 31250 samples/second, and for macro a bandpass of 5Hz-5kHz at 3125 samples/second. Data acquired were then decomposed using the DQEMG algorithm of Stashuk (1999). The QEMG features produced by this program were then used as the input data for the algorithm described here.

The names of the features collected along with the values describing the resulting vector are shown in Table I. The value for each feature is shown as a

mean over 50 leave-one-out jackknife trials, along with the standard deviation calculated over the set of trials.

In order to provide a comparative result, a naïve evaluation of the Fisher linear discriminant was performed as described in Duda *et al* (2001, pp. 117-121), using the normative and diseased class distributions directly (*i.e.*; without consideration for possible normative values in the diseased class distribution). The production of results using this method was unsuccessful due to the presence of an inverse operation in the calculation of the Fisher vector; the disease data acquired for this experiment becomes singular in the Fisher decomposition, and therefore the textbook calculation of the Fisher vector cannot be obtained.

Using the new method presented here, it is therefore possible to calculate a result unavailable from the traditional Fisher methodology. As can be seen in Table I, the vector values obtained are very stable, with a standard deviation in all cases being several orders of magnitude lower than the mean. Evaluating the classification performance of the discriminant on separate testing data gives 62% (std. dev. ±0.490) correct classification, with a specificity of 0.909 (±0.302) and with a sensitivity of 0.538 (±0.505).

Table I: Features Examined and Vector Found

Feature Name	*Fisher Vector*	*Units*
Micro Features		
Amplitude	$0.594 \pm 5.0e^{-3}$	μV
Duration	$0.013 \pm 1.1e^{-4}$	ms
Phases	$3.4e^{-3} \pm 3.1e^{-5}$	
Turns	$4.4e^{-3} \pm 3.9e^{-5}$	
AAR	$1.9e^{-3} \pm 1.8e^{-5}$	ms
Macro Features		
Amplitude	$0.119 \pm 1.0e^{-3}$	μV
Neg. Peak Area	$0.717 \pm 4.9e^{-4}$	μV•ms
Neg. Peak Amplitude	$0.068 \pm 5.6e^{-4}$	μV
Neg. Peak Duration	$0.034 \pm 3.2e^{-4}$	ms
Timing Features		
IPI Mean	$0.091 \pm 8.9e^{-4}$	ms
IPI Std. Dev.	$0.014 \pm 1.6e^{-4}$	
IPI Covariance	$2.1e^{-4} \pm 2.8e^{-6}$	
Inter-Discharge Rate	$0.075 \pm 7.9e^{-4}$	pps
Firing Rate	$0.020 \pm 2.0e^{-4}$	pps
Mean Con. Diff	$3.4e^{-4} \pm 7.0e^{-6}$	pps
# MUPs	$0.313 \pm 4.4e^{-3}$	
Mean MU Voltage	$0.011 \pm 8.3e^{-5}$	μV

DISCUSSION

The data in Table I support several interesting observations, most notably that the contribution to the discrimination between classes is not uniform among the acquired features. Note the relatively high value associated with Negative Peak Area; this measure has been shown to be informative in identifying fatigued versus non-fatigued muscle (Calder *et al*, 2006), presumably resulting from slowed muscle fibre conduction velocities, and a similar scenario would be expected here. Similarly, given that Phases and Turns are not frequently observed in most neuromuscular disorders, it is not surprising that the value associated with both of these features is more than 100 times lesser in magnitude than that of Negative Peak Area. It is clear that this classification decision can be made with a reduced number of features and that low numbers in Table I indicate which features may be more easily omitted than those with higher values.

As the Fisher vector provides a one-dimensional projection into a linear space (Duda *et al*, 2001, pp. 117), the degree of information present in the vector weights will themselves be linearly related allowing the relative importance of any feature to be directly read from the table.

CONCLUSIONS

Several important conclusions can be drawn from the presented results:

the magnitude of a component in an inter-mean vector may be useful separating class distributions and can be constructed using a neighbourhood algorithm. The resulting vector is analogous to the inter-mean vector of the Fisher Linear Discriminant, and may therefore be used to differentiate between points belonging to two different classes.

such an inter-mean vector may be constructed in cases where the traditional approach fails due to singularities in the covariance matrix of the data.

the resulting one-dimensional projection created by the inter-mean vector will be useful in assessing the relative contribution of each feature to the inter-class discernment problem. Features described with lower weights in the modified Fisher vector are less important for making this type of decision. The relative importance of these numbers may be described in a linear fashion.

Future work will include the assessment of the Fisher vector constructed here along with an inter-class decision threshold. The resulting data will allow classifications to be made, providing a linear-space

classifier for diseased MUP data, or other data of this form.

REFERENCES

[1] Bezdek, J.C. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum.

[2] Brown, W.F., Bolton, C.F. and Aminoff, M.J. (2002). Neuromuscular Function and Disease. Philadelphia: W.B.Saunders.

[3] Calder, K.M., Stashuk, D.W. and McLean, L. (2006). Physiological characteristics of motor units in the brachioradialis muscle across fatiguing low-level isometric contractions. Journal of Electromyography and Kinesiology, (in press), PubMed ID: 17113787.

[4] Doherty, T.J. and Stashuk, D.W. (2003). Decomposition-based quantitative electromyography: methods and initial normative data in five muscles. Muscle and Nerve, 2(28), 204-211.

[5] Duda, R.O., Hart, P.E. and Stork, D.G. (2001). Pattern Classification. 2nd ed. Wiley.

[6] Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7(2), 179-188.

[7] Kohonen, T. (1989). Self-Organization and Associative Memory, 3rd ed. Berlin: Springer.

[8] Preston, D.C. and Shapiro, B.E. (2005). Electromyography and Clinical Neuromuscular Disorders. Elsevier.

[9] Stashuk, D.W. (1991). Decomposition and quantitative analysis of clinical electromyographic signals. Medical Engineering and Physics, 21(6), 389-404.