Aggregate Features in Multisample Classification Problems

Robert Varga, S. Marie Matheson, and Andrew Hamilton-Wright, *Member, IEEE*

***Abstract*—This paper evaluates the classification of multisample problems, such as electromyographic (EMG) data, by making aggregate features available to a per-sample classifier. It is found that the accuracy of this approach is superior to that of traditional methods such as majority vote for this problem. The classification improvements of this method, in conjunction with a confidence measure expressing the per-sample probability of classification failure (i.e., a hazard function) is described and measured. Results are expected to be of interest in clinical decision support system development.**

***Index Terms*—Bayes methods, decision support systems, machine learning, pattern analysis, statistical learning, supervised learning.**

I. Introduction

TYPICALLY, when a classifier is asked to combine information across multiple samples drawn from the same data source, the results are combined using a strategy such as majority vote [1]–[3]. The question then arises as to whether the sample-by-sample classification can be improved by means of incorporating some sort of information describing the full set of samples along with the per-sample values. One means of representing the data from a particular sample in such a set is to allow a classifier to consider each sample along with data describing an aggregate measure of all samples.

Problems of this sort arise in a number of milieux, however one that is of particular interest to the authors is in clinical characterization of disease state, based on quantitative electromyographic (QEMG) analysis. Here, one must ascertain the correct characterization of a muscle, in terms of its disease state, based on EMG signals that are produced by the structural groups of force production, called motor units (MUs).

An MU is the minimal control structure of a muscle, comprised of the set of muscle fibers coupled with, and controlled by, a single α-motor neuron. These MUs are interleaved with the other MUs in a muscle, and provide the force generation capability of a muscle while generating current observable as action potentials. These motor-unit potentials (MUPs) have a stable signature shape related to the fixed morphological structure of the terminal branches of the associated α-neuron. A signal composed of these potentials may be decomposed into per-MU activity [4], and converted into a table of QEMG measures describing the contribution of all observed MUs, (potentially among several contractions in a muscle study) which together are used as a source of diagnostic information [5]–[8] useful in a clinical decision support system (CDSS).

A CDSS is a particular application in the larger field of decision support systems that helps a clinical professional make a decision in a more complete, consistent, and informed fashion than they would be able to achieve without such a system. As clinical decisions are typically high in risk, transparency and explanatory ability is paramount; without these attributes, a CDSS will remain unused, even if diagnostic accuracy is improved [9]. Reviews of CDSS technologies are provided in [10], with an overview of some of the current issues provided in [11]. Recent work has seen the use of machine-learning-based classification systems in CDSS [12]–[15], however data aggregation has typically been achieved via summation of classification output information [14], [15], rather than the approach explored here.

The rest of this paper will be structured as follows. Section II will outline the data examined, as well as its preparation into cross-validation studies, the construction of the several additional feature sets examined, and the classifiers used in the experiments. Section III contains a description of the statistical examinations used to analyze the results. Results themselves will be presented in Section IV, followed by a discussion of the findings in Section V. The paper closes with a summary of the major points.

II. Methodology

Using Bayesian learning systems, we evaluate the efficacy of using additional feature sets (AFSs) on MUP data, where an overall muscular characterization is required based on the "study" of the problem, with multiple samples drawn from the same source. Some further exploration of these ideas using studies drawn from synthetically generated covaried data were also performed.

Values for an individual AFS are calculated by using a simple aggregation of all of the observed values for each feature within the study, and adding this result as a new feature to all samples, providing each sample information about the entire study. We inspect three simple aggregators in this initial examination of this idea: arithmetic mean, and maximum and minimum value.

Manuscript received November 23, 2013; revised February 16, 2014; and March 21, 2014; accepted March 24, 2014. Date of publication April 2, 2014; date of current version March 2, 2015.

R. Varga is with the School of Computer Science, University of Guelph, Guelph, ON N1G 2W1, Canada (e-mail: rob@robsplace.ca).

S. M. Matheson is with the Department of Mathematics and Computer Science, Mount Allison University, Sackville, NB E4L 1E2, Canada (e-mail: smmatheson@mta.ca).

A. Hamilton-Wright is with the School of Computer Science, University of Guelph, Guelph, ON N1G 2W1, Canada, and also with the Department of Mathematics and Computer Science, Mount Allison University, Sackville, NB E4L 1E2, Canada (e-mail: andrewhw@ieee.org).

Digital Object Identifier 10.1109/JBHI.2014.2314856

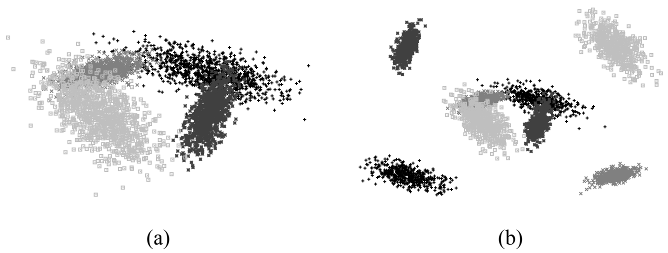


Fig. 1. Sample simulated data distributions.

A. Preparation of Evaluation Data

Evaluation was performed with two different types of data: MUP data based on QEMG studies, and synthetic data generated with specific distribution characteristics.

1) EMG Data: EMG from 21 muscles was produced using a simulation model [16], and decomposed using the DQEMG program [4], providing gold-standard data for nine muscles with *neuropathic involvement* (i.e., typical of a disease that attacks the nervous tissue controlling a muscle) nine with *myopathic involvement* (the class of diseases that attack the muscle fibers directly), and three with *no involvement* (i.e., healthy tissue). There are a total of 812 samples in these 21 studies, with the number of samples per study ranging between 18 and 53. Each sample has measures of the following features, described in [5]: amplitude, duration, phases, turns, area-to-amplitude ratio, size index, mean MU voltage.

2) Synthetic Data Distributions: In addition to the MUP data, further evaluation was performed based on data generated from four-class, four-feature synthetic distributions. These are partially overlapping normal distributions, arranged either unimodally or bimodally as indicated in Fig. 1, at various separations. These data are described fully in [17]. These synthetic data were used to validate the findings based on the MUP data described in this paper.

3) AFS preparation: Experimental treatments were constructed for each study, containing the original feature data, and additional columns containing the AFS data. Each AFS type (mean, min, max) was evaluated singly, in pairs, and as a triplet, for seven treatments with AFSs, and an eighth control treatment AFS(NONE) with no additional feature sets added.

These aggregator operators were chosen for their simplicity; arithmetic mean will provide a measure of central tendency of the study and min and max provide the range. As each individual sample in a study is classified, the presence of these AFS will allow comparison of the measured feature value against the distribution of values found in the study itself. This provides a mechanism by which a sample in a study can be identified as an outlier, information that has been noted to be important in correct QEMG classification [18]–[21].

The cost of this approach is clear, as each aggregator adds an extra feature for each original feature, greatly increasing the dimensionality of the search space, and therefore raising the issue of the "curse of dimensionality" [22], suggesting that significantly more training data may be required in order to successfully train the classifier. It may be noted, however, that as the aggregate values are based on other values within the grouped sample set, there is not independence between these values, implying a grouping within the data space and indicating that the "curse" may be somewhat ameliorated.

4) Discretization and Cross Validation: All data were quantized using maximum marginal entropy [23] in order that Bayesian event probabilities may be constructed on the data as quantized into ten bins.

Leave-one-out cross validation was used to better estimate classification accuracy, using each complete study as a single cross-validation set; this will ensure that all of the related AFS values from each study are grouped together into either testing or training datasets. While cross validation is known to produce an overestimate bias for accuracy [24], [25], as this is a comparative study based on identical data, the relative biases will be on average equal, and a relative comparison may meaningfully be made.

B. Classifiers Used

Several Bayesian classifiers were compared. PD/FIS*, a rule-based classier, and three Bayesian networks: naïve Bayes (NAÏVE-BN), tree-augmented naïve Bayes (TAN-BN), and an evolutionarily constructed Bayesian network (EVOLVED-BN).

1) PD/FIS:* This classifier [26] has previously been used with QEMG data [7], [27]–[30]. It functions by evaluating the frequency of occurrences of associations between values of the label and observed features in one or more of the input columns. By comparing these, using the adjusted residual [31], [32], one may detect associations that differ significantly from those expected by a model of random chance; these "patterns" are then used as rules for classification, weighted by their information content using the "occurrence/all" mechanism [33].

Classification is then the calculation of a weighted sum resulting in a set of assertions ($A_k \epsilon [-1 \ldots 1]$) representing a spectrum from total refutation (-1) to total support (1) for a given labeling. By comparing the A_k values for each label, one can identify τ, the highest value asserted for any label, and δ, the difference between the two highest assertion values. These provide access to internal measures of the amount of relative information used in calculating the class label, and have been used to construct a confidence measure $\mathcal{C}_{\mathrm{PD/FIS}*}$ for use in CDSS by observing how often the system is correct for similar τ and δ values.

The PD/FIS system by default will attempt to construct a labeling of "UNKNOWN" if $\tau \leq 0$, or if δ approaches 0. As this refusal-to-label behavior is not available in the other Bayesian systems used here, we will have suppressed this feature, labeling all input data regardless of the system confidence in the labeling; we therefore denote these results as PD/FIS*, to note the distinction in behavior relative to the original PD/FIS.

2) Bayesian networks: A Bayesian network (BN) is a directed acyclic graph-based representation of a probability distribution, using nodes to represent observable events, such as particular input values or class labels, and relations between events as arcs. Searching for an optimal graph based on training data is difficult, both due to the need to establish the degree of dependence between observed events, and the computational complexity of the search. We examine three common algorithms for obtaining a nonoptimal graph in a feasible manner.

Confidence may be measured for all Bayesian networks by examining the fraction of probabilistic support for the winning class: this fraction is then used as the confidence in the assigned classification, for $\mathcal{C}_{\text{NAÏVE}-\text{BN}}$, $\mathcal{C}_{\text{Tan}-\text{BN}}$, and $\mathcal{C}_{\text{Evolved}-\text{BN}}$.

3) Naïve Bayesian networks: Naïve Bayesian networks (Naïve-Bn) [34], [35], based on the assumption of complete independence between input values, are surprisingly effective classifiers, frequently outperforming more complex classifiers [36]. An important weakness of Naïve-Bn in CDSS design is that a failure to accurately reflect the probability distribution of the underlying data leads to a poor measure of decision confidence, and undermines transparency and understandability.

4) Tree augmented Naïve Bayesian networks: Tree augmented naïve Bayesian networks (Tan-Bn) [37], [38] attempt to exploit the strengths of Naïve-Bn classifiers by relaxing the independence assumption, allowing the feature nodes in a network to form a fully dependent tree, creating systems that can outperform Naïve-Bn systems [39].

5) Evolutionary algorithms: Further utilizing randomized search, an evolutionary algorithm can be used to construct the network, by using tournament selection randomization to select networks for merging, and by swapping arcs, and finally pruning by the use of a Markov blanket based on the class node as described in [40].

C. Computational Environment

All algorithms discussed were run on SHARCNET, a portion of the Canadian academic supercomputing network. Each training run was given a maximum runtime of 7 days. Any run that did not complete in that time was terminated, and not included in the results below.

III. Analysis

A. Classification Accuracy

Classification accuracies are compared on a sample-by-sample basis, to examine the degree to which each individual classification may be improved by the addition of an AFS.

The total number of correct samples classified in each cross-validation run was tallied for each dataset. The Durkalski formulation of the McNemar test [41] was then applied to the counts of correct and incorrect classifications produced, repeated over each of the cross-validation sets. The McNemar test focuses on the cells on the secondary diagonal, which captures occasions when only one classifier is correct observing the difference between the marginal proportions of correctness of each system. Using this method, one can calculate the differences between the marginal proportions of correct classification of each system (Δ_{McN}) and establish a 95% confidence interval on the improvements made by choosing one classifier over another across all data examined.

In order to determine whether the addition to the data of one or more AFS columns made a significant improvement to classification accuracy for a given classifier, the per-sample classification results were evaluated by applying the nonparametric Kruskal–Wallis [42] one-way analysis of variance to the classifications made. Separate results were constructed for each dataset investigated (MUP, unimodal covariate, and bimodal covariate). Note that the standard parametric ANOVA is inappropriate in this case, as the measured accuracies do not follow a Normal distribution, nor are the variances near equal. Minitab 15 (Minitab, Inc., State College, PA, USA) was used to calculate all Kruskal–Wallis (K–W) results.

B. Classification Confidence

Due to our interest in using this system as a CDSS, we measure not only classification accuracy, but additionally classification confidence: the estimated probability that the classifier has arrived at the correct labeling. This may be based on some measure of the classifier's internal state based on a particular input data value, and thus may be seen as analogous to the hazard function modeling the risk of failure in a physical system. As an aid to transparency, confidence is suggested in the design of clinical decision support systems regularly [1], [33], [43]–[48], and is incorporated in some form in now-available rule-based, general purpose clinical decision support construction tools [49], [50].

By incorporating the probabilistically based (and therefore $[0\ldots 1]$ bounded) $\mathcal{C}_{(\text{System})}$ values described above with an indicator of correctness, we can obtain a measure of the error in confidence, $\mathcal{E}^{\text{Confidence}}_{(\text{System})}$, as follows:

$$\mathcal{E}^{\text{Confidence}}_{(\text{System})} = \begin{cases} 1 - \mathcal{C}_{(\text{System})} & \text{if correct, and} \\ \mathcal{C}_{(\text{System})} & \text{if incorrect.} \end{cases}$$

A *confidence error* close to 0 implies that either the system is very confident in its correct classification or very uncertain of its incorrect classification. This reflects the situation that as the risk of error rises, the confidence decreases, and therefore if an error is likely to be made, this fact can be accurately indicated to a human decision maker.

We will calculate changes in confidence error both on individual samples, and also on a per-study level, by calculating a mean confidence for the study. This is expected to be useful in a CDSS context, as it is the study-level decision, and its related confidence, that will be of most interest to a decision maker. This will be referred to as "study-level confidence error." K–W statistics were computed to identify significant changes, as with classification accuracy.

IV. Results

We shall explain the MUP results found in detail, and then summarize the relationship between these results and the synthetic distributions.

McNemar results are shown in Fig. 2, where one can see that the difference in observed marginal proportion for the Naïve-Bn versus Tan-Bn systems is defined by a confidence interval $\Delta_{\text{McN}} = (0.105, 0.109)$. All classifier names have been ordered so that the first mentioned classifier outperforms the second, to avoid confusion due to unnecessary sign changes. A lower

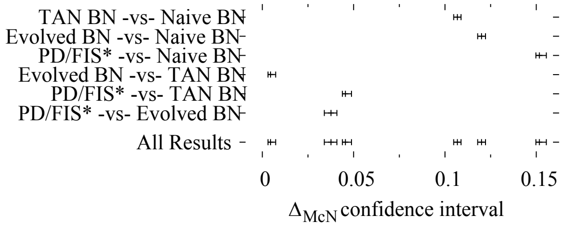


Fig. 2. Results from pairwise McNemar tests of the systems.

number for Δ_{McN} indicates more similarity in performance; zero is identical performance. In all instances, $p < 10^{-5}$ for Δ_{McN}. Note the groupings in performance in which all NAÏVE-BN system comparisons are similar (all values > 0.1) while in all other comparisons $\Delta_{\mathrm{McN}} < 0.05$.

If the NAÏVE-BN system is disregarded, the Δ_{McN} values for the remaining pairs are much smaller, falling within 0.05 of each other. It is clear then that the NAÏVE-BN systems perform much more poorly than any other system, which are competitive in their ability to classify MUP data correctly.

A. Independent MUP Sample Classification Accuracy

Analysis using K–W found that a significant improvement in per-sample classification accuracy was observed when any AFS was added to the system ($p < 0.0005$), when considering each dataset separately, or as a pool.

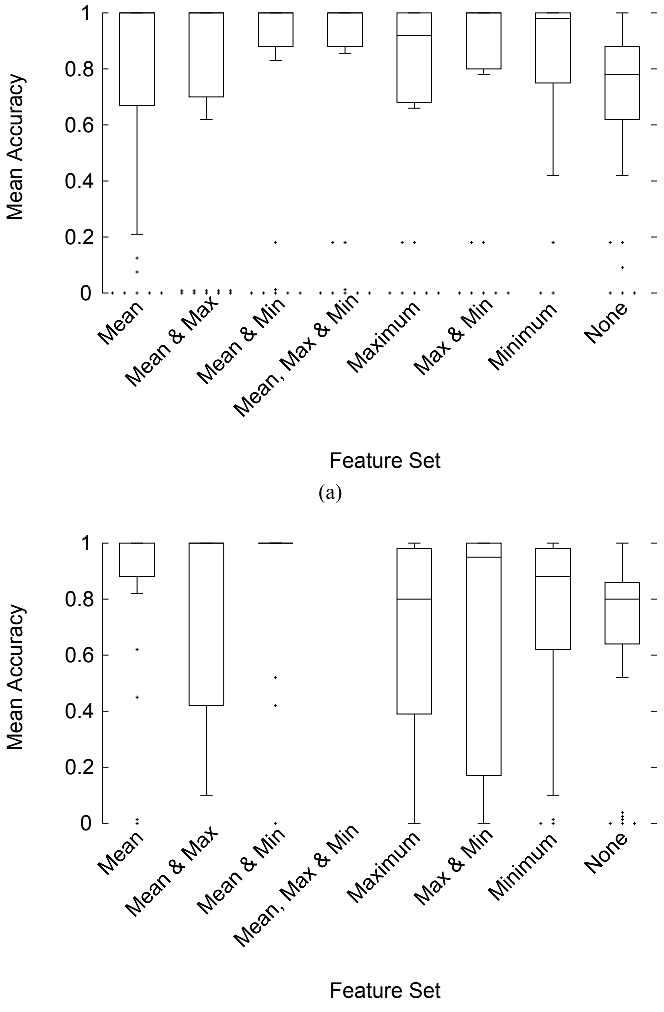
Additionally, the effect of adding any AFS was so strong that it was not possible to distinguish between the various AFS measures for NAÏVE-BN or TAN-BN. Both the EVOLVED-BN and PD/FIS* systems were observed to have a number of very low accuracies, as indicated by the boxplots for the MUP tests presented in Fig. 3, however the addition of an AFS still significantly improved the performance. In all tests $p < 0.0005$.

B. Independent MUP Sample Confidence Error

Both NAÏVE-BN and TAN-BN significantly reduced their confidence error based on the addition of an AFS, with $p < 0.0005$, again as measured independently on any of the datasets examined. Conversely, for the EVOLVED-BN case, there were no significant effects noted due to the addition of an AFS. In the case of PD/FIS*, AFS(MEAN) showed improvement in confidence error with $p < 0.001$; AFS(MAX) and AFS(MIN) however showed no such improvement; results for PD/FIS* on MUP data are shown in Fig. 4.

C. Study-Level MUP Confidence Error

The study-level confidence error drops to near zero for NAÏVE-BN and TAN-BN strategies, with $p < 0.0005$. In contrast, for PD/FIS* the result is particularly sensitive to the AFS used, and those involving the arithmetic mean are the values with the best result (in each dataset examined). This is shown in Table I which provides for each AFS the median value for the measured confidence error associated with that AFS, as well as the mean of the rankings computed on all values corresponding to this AFS within the ordered sequence of all results. N refers to the number of successful runs in this set (time limited, as indicated in Section II-C). For the EVOLVED-BN system, there is no significant effect ($p = 0.594$).



(a)

(b)

Fig. 3. Independent MUP sample classification accuracies. (a) EVOLVED-BN Systems. (b) PD/FIS* Systems.

D. Synthetic Covaried Data Results

These results are largely corroborated by the analysis of the covaried synthetic data; the rankings and relative strengths of the AFS classifiers remain the same when examined on covaried synthetic data with similar amounts of distribution overlap as the MUP data results. As the distributions become more separate, the effect is weakened as the overall classification accuracy rises toward unity; similarly as the distributions are moved together, the effect weakens as overall classification accuracy moves toward random chance.

E. Overall Results Summary

Table II displays the overall results found, showing, for each classifier, the AFS with the most significant effect for each test, or *no effect* if no significant effect was observed.

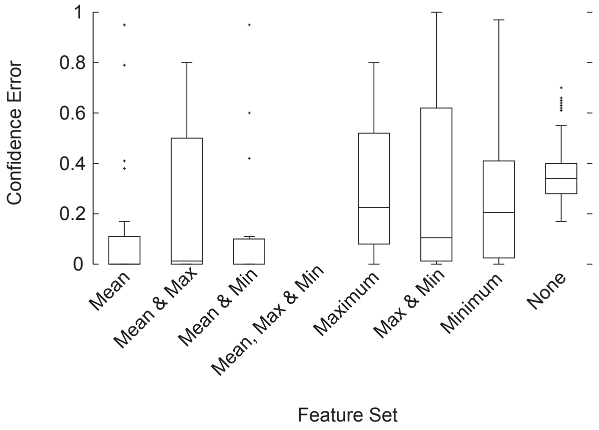


Fig. 4. Independent sample confidence error for the PD/FIS* systems.

TABLE I
KRUSKAL–WALLIS TESTS ON PD/FIS* CONFIDENCE ERROR

System Group	N	Median	Mean Rank	z
AFS(MEAN)	30	0.00	50.5	-4.98
AFS(MEAN+MAX)	10	0.01	74.8	-1.31
AFS(MEAN+MIN)	18	0.00	50.9	-3.70
AFS(MAX)	41	0.18	113.1	2.00
AFS(MAX+MIN)	18	0.11	100.9	0.27
AFS(MIN)	35	0.13	103.6	0.71
AFS(NONE)	42	0.28	134.7	4.85

H = 57.05, DF = 6, $p < 0.0005$

TABLE II
SUMMARY OF RESULTS

System Name	Per-Sample Classification Results		Overall Confidence Error
	Classification Accuracy	Confidence Error	
NAÏVE-BN	AFS(MEAN)	AFS(MEAN)	AFS(MEAN)
TAN-BN	AFS(MIN)	AFS(MIN)	AFS(MIN)
EVOLVED-BN	AFS(MEAN+MIN)	*no effect*	*no effect*
PD/FIS*	AFS(MEAN+MIN)	AFS(MEAN)	AFS(MEAN)

V. DISCUSSION

The NAÏVE-BN systems are shown to perform poorly compared to all of the other systems, likely because of feature dependence. The TAN-BN and EVOLVED-BN systems are shown to perform similarly to each other, but not as well as the PD/FIS* -based systems. There are a couple of different possible reasons that the PD/FIS*-based systems perform better than the BN systems. It is possible that the BN structures imposed or found do not truly represent the distribution of the training data. For example, the features in the networks are definitely not independent, and so the assumptions that the NAÏVE-BN systems make do not hold. Another potential explanation is that the BN systems are unable to apply Occam's Razor and find simple, generalized patterns. The PD/FIS*-based systems only keep patterns in the rule-base that are found as significant, whereas the BN systems keep full observed probability distributions for dependencies from the training data, and apply them when classifying.

When the systems are trained with MUP data, with various AFSs added, different strengths show in the results. The BN systems do not appear to have difficulty in learning problems with multifeature AFSs added to the original data, whereas the PD/FIS*-based systems have increasing difficulty as the size of the AFSs increase.

These results imply that, while the BN systems perform worse than the PD/FIS*-based systems, they can be trained with more complex datasets in a reasonable period of time. If the addition of AFSs to the original MUP data shows a significant increase in classification accuracy, then BN systems may have the advantage because they are able to handle the larger datasets. This result is not limited to simple data with AFSs added, but to any data that has a large number of features.

It is also interesting to note that, for most of the metrics discussed, the EVOLVED-BN systems behave inconsistently with the addition of AFSs as any of the other systems. Independent sample classification accuracy does improve with the systems using AFS(MEAN+MIN), but there is no statistically significant effect on sample classification confidence error observed from the addition of any of the AFSs to the original MUP data.

As EVOLVED-BN systems are trained, there is no requirement that every feature node is included in the systems. This means that they are the only BN learning approach investigated in this paper that performs feature selection. There are two different potential benefits of adding AFSs to data: there will be more information, and there will be better information. Since the EVOLVED-BN systems are more likely than the other systems to disregard features inherently, they lose the potential benefit of there being more information. However, the fact that there is better information does allow the EVOLVED-BN systems to classify more accurately on a per-sample basis. However, these improvements are not enough to translate to better overall study classification accuracy.

A. Sample Classification

All systems have at least one AFS that tends to cause an improved independent sample classification accuracy, with a confidence of 99.9%. Additionally, but with the exception of the EVOLVED-BN systems, all systems have at least one AFS that tends to cause a decreased independent sample confidence error, with a confidence of 99.9%.

The TAN-BN(MIN) systems appear to discriminate against samples that the TAN-BN(NONE) systems have difficulty in classifying correctly. In other words, for samples classified with low accuracy but high confidence by TAN-BN(NONE), the TAN-BN(MIN) system could correctly represent a low confidence for these samples. The same effect occurs with the addition of several different AFSs to the PD/FIS* systems, with confidences of at least 90%. This means that the additional information provided by the AFSs gives the systems enough information to recognize samples that may be more difficult to classify. This can potentially be used to flag problematic classifications, and alert a clinician of these, rather than make an error; we hope to address this question more fully in future work.

This effect, however, is limited to the two aforementioned systems—the other systems either do not have trouble classifying, or do not differentiate between samples that the AFS(NONE) systems have difficulty with.

Some of the systems just simply have problems classifying systems with AFSs added to the original MUP data. As AFSs are added, and more complex datasets are thus created, it becomes less likely that patterns seen in the training data will appear in the testing data. When a pattern never seen by a BN during training appears in testing, a classification cannot be made because the probability assigned to that pattern is 0, causing the probability of any class being correct as 0. The PD/FIS* systems are designed so that they find patterns in the training data that appear more often than random chance should allow. When patterns that have never been seen before appear in testing data, these two systems simply do not use those patterns as part of the classification process. However, as there are more patterns available in the training data, it is likely that more significant patterns will be found. Not all of these patterns are necessarily useful—some of them may even be misleading. When a classification is made, rules that would otherwise have been integral in classifying the sample correctly may be weighted too low due to the large number of other rules that are also triggered.

In other words, the AFSs analyzed in this paper add information about how individual samples relate to the studies in which they are a part of. With each additional aggregator in an AFS, the number of features added to the original dataset increases, which provides more and more information about how exactly each individual sample relates to the whole. This, however, also results in a larger number of possible relationships between the individual samples and the studies as a whole because of the larger total feature space. In turn, this requires more data to cover the space—this is the well-known "curse of dimensionality." Due to this increased number of patterns, the number of apparently significant patterns also increases causing truly significant and useful patterns to get lost in less useful patterns.

The most obvious solution to this dilemma is to add more training data. As training data are added, it becomes more probable that significant relationships between samples and the studies they are part of are found. Also, more data will decrease the statistical variability in significance when finding patterns in the data for the PD/FIS* systems, allowing them to be more confident that the found statistically significant patterns are truly statistically significant. However, this poses two more problems.

First, it is still not a guarantee that all significant patterns will be found. This problem is not unique to systems using datasets that have AFSs—the job of a learning algorithm is to find general patterns in training data that it can use to classify new samples. If any system is given patterns that it does not recognize from the training data, then it will be unable to classify. This would imply that the system was either unable to find general patterns in the data, or there are no general patterns in the data. If the system is unable to find general patterns, then different systems need to be investigated. However, if no general patterns exist in the data, then it may become a very difficult classification problem that needs further study.

The second problem is that more data may not be available. In a clinical setting, data can be hard to acquire. It could involve getting a large number of people to participate in potentially obtrusive tests. Data could potentially be generated by using computer simulations of human functions, but this poses the risk of introducing biases if the simulations do not mimic the human functions in their entirety. Ultimately, real data are preferred for training real systems to classify real problems, and so acquiring more data just may not be possible.

B. Study Classification

Section IV-C and Table II show that there exists at least one AFS for every group of systems, excluding the EVOLVED-BN systems, where overall confidence error is lower than the AFS(NONE) systems, with a confidence of 99.9%. This is not surprising, since the same systems tend to have increased independent sample classification accuracies, and decreased independent sample confidence errors, when certain AFSs are added. It is not surprising that the EVOLVED-BN systems do not show a decreased confidence error, either, as the addition of AFSs to the original MUP data did not improve independent sample confidence error.

C. Transparency and Confidence

The addition of an AFS shows a significant improvement in classification accuracy, confidence, and in confidence error. This significant effect indicates that use of an AFS in decision support systems has multiple benefits, both in terms of system accuracy and in terms of system transparency.

VI. CONCLUSION

While the inclusion of an AFS may provide a significant classification benefit for multisample problems, different classifiers are sensitive to different AFS choices, as seen in Table II. The use of the AFS(MEAN) is generally found to be advantageous. This may relate to the fact that MUP data have been noted to have a great deal of information in its outliers [18]–[21], indicating that other choices of AFS that take this into account, such as a trimmed mean, or quartile measure, may be of interest.

The addition of AFSs to the original MUP data increases independent sample classification accuracy, but that does not translate into an increased study classification accuracy. When trained and tested with AFS(MEAN), the PD/FIS* systems show a reduced confidence error. When AFS(MAX) is used to train and test the PD/FIS* systems, confidence error remains unchanged.

REFERENCES

[1] C. Christodoulou and C. Pattichis, "Medical diagnostic systems using ensembles of neural SOFM classifiers," in *Proc. 6th IEEE Int. Conf. Electron., Circuits Syst.*, 1999, vol. 1, pp. 121–124.

[2] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. New York, NY, USA: Wiley, 2004.

[3] L. Rokach, "Ensemble-based classifiers," *Artif. Intell. Rev.*, vol. 33, no. 1/2, pp. 1–39, Feb. 2010.

[4] D. W. Stashuk, "Decomposition and quantitative analysis of clinical electromyographic signals," *Med. Eng. Phys.*, vol. 21, no. 6/7, pp. 389–404, Feb. 1999.

[5] D. W. Stashuk and W. F. Brown, "Quantitative electromyography," in *Neuromuscular Function and Disease*, vol. 1, W. F. Brown, C. F. Bolton, and M. J. Aminoff, Eds. Philadelphia, PA, USA: Saunders, 2002, pp. 311–348.

[6] D. W. Stashuk and T. J. Doherty, "The normal motor unit action potential," in *Neuromuscular Function and Disease*, vol. 1, W. F. Brown, C. F. Bolton, and M. J. Aminoff, Eds. Philadelphia, PA, USA: Saunders, 2002.

[7] D. W. Stashuk, L. Pino, A. Hamilton Wright, T. Doherty, and S. Boe, *Interpretation of QEMG data," presented at the General Meet. Amer. Assoc. Neuromuscular Electrodiagnostic Med.*. Phoenix, AZ, USA, 2007.

[8] J. V. Basmajian and C. J. De Luca, *Muscles Alive: Their Functions Revealed by Electromyography*, 5th ed. Baltimore, MD, USA: William & Wilkins, 1985.

[9] R. L. Teach and E. H. Shortliffe, "An analysis of physician attitudes regarding computer-based clinical consultation systems," *Comput. Biomed. Res.*, vol. 14, no. 6, pp. 542–558, Dec. 1981.

[10] A. X. Garg, N. K. J. Adhikari, H. McDonald, M. P. Rosas-Arellano, P. J. Devereaux, J. Beyene, J. Sam, and R. B. Haynes, "Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: A systematic review," *J. Amer. Med. Assoc.*, vol. 293, no. 10, pp. 1223–1238, 2005.

[11] M. Greenberg and M. S. Ridgely, "Clinical decision support and malpractice risk," *J. Amer. Med. Assoc.*, vol. 306, no. 1, pp. 90–91, 2011.

[12] A. Dupuy and R. M. Simon, "Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting," *J. Nature Cancer Inst.*, vol. 99, no. 2, pp. 147–157, Nov. 2007.

[13] Y. Wang, Y. Fan, P. Bhatt, and C. Davatzikos. (2010, May). High-dimensional pattern regression using machine learning: From medical images to continuous clinical variables. NeuroImage[Online]. 50(4), pp. 1519-1535. Available:http://www.sciencedirect.com/science/article/pii/S1053811909013810

[14] H. Lu, H.-L. Eng, C. Guan, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularized common spatial pattern with aggregation for EEG classification in small-sample setting," *IEEE Trans. Bio-Med. Eng.*, vol. 57, no. 12, pp. 2936–2946, Dec. 2010.

[15] G. Pfeiffer, "The diagnostic power of motor unit potential analysis: An objective Bayesian approach," *Muscle Nerve*, vol. 22, no. 5, pp. 584–591, 1999.

[16] A. Hamilton-Wright and D. W. Stashuk, "Physiologically based simulation of clinical EMG signals," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 2, pp. 171–183, Feb. 2005.

[17] A. Hamilton-Wright, D. W. Stashuk, and H. R. Tizhoosh, "Fuzzy classification using pattern discovery," *IEEE Trans. Fuzzy Syst.*, vol. 15, no. 5, pp. 772–783, Oct. 2007.

[18] E. V. Stålberg, C. Bischoff, and B. Falck, "Outliers, a way to detect abnormality in EMG," *Muscle Nerve*, vol. 17, no. 4, pp. 392–399, 1994.

[19] S. Podner, "Comparison of different outlier criteria in quantitative anal sphincter electromyography," *Clin. Neurophysiol.*, vol. 116, no. 8, pp. 1840–1845, 2005.

[20] M. Sonoo, M. Kobayashi, N. Kokubun, T. Imai, Y. Arimura, S. Kuwabara, and T. Komori, "How to determine the percentile value in the outlier analysis of quantitative electromyography," *Muscle Nerve*, vol. 46, no. 4, pp. 637–637, 2012.

[21] G. L. Sheean, "A self-referential outlier detection method for quantitative motor unit action potential analysis," *Med. Hypotheses*, vol. 78, pp. 430–431, 2012.

[22] R. Bellman, "Computational aspects of dynamic programming," in *Adaptive Control Processes: A Guided Tour*. Princeton, NJ, USA: Princeton Univ. Press, p. 94, 1961.

[23] D. V. Gokhale, "On joint and conditional entropies," *Entropy*, vol. 1, no. 2, pp. 21–24, 1999.

[24] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1924, Oct. 1998.

[25] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th Int Joint Conf. Artif. Intell.*, 1995, pp. 1137–1143.

[26] A. Hamilton-Wright and D. W. Stashuk, "Statistically based pattern discovery techniques for biological data analysis," in *Applications of Computational Intelligence in Biology* (Studies in Computational Intelligence), vol. 122, T. G. Smolinski, M. G. Milanova, and A.-E. Hassanien, Eds. Berlin, Germany: Springer-Verlag, 2008, ch. 1, pp. 3–31.

[27] A. Hamilton-Wright and D. W. Stashuk, "Clinical characterization of electromyographic data using computational tools," in *Proc. Symp. Comput. Intell. Bioinf. Comput. Biol.*, Toronto, ON, Canada, Sep. 2006, pp. 1–7.

[28] A. Hamilton Wright and D. W. Stashuk, "Clinical decision support by fuzzy logic analysis of quantitative electromyographic data," presented at the 16th Int. Soc. Electromyography Kinesiology Congr., Torino, Italy, 2006.

[29] A. Hamilton-Wright and D. W. Stashuk, "Fuzzy rule based decision making for electromyographic characterization," in *Proc. 11th Int. Conf. Inf. Process. Manage. Uncertainty in Knowledge-Based Syst.*, 2006, pp. 1168–1173.

[30] A. Hamilton-Wright, L. McLean, D. W. Stashuk, and K. M. Calder, "Bayesian aggregation versus majority vote in the characterization of nonspecific arm pain based on quantitative needle electromyography," *J. NeuroEng. Rehabil.*, vol. 7, no. 8, Feb. 2010.

[31] S. J. Haberman, *Analysis of Qualitative Data* (Springer Series in Statistics). Toronto, ON, Canada:: Academic, 1979, vol. 1, pp. 78–79, 82–83.

[32] S. J. Haberman. (1973, Mar.). The analysis of residuals in cross-classified tables. *Biometrics* [Online]. *29(1)*, pp. 205–220. Available:http://www.jstor.org/stable/2529686

[33] A. Hamilton-Wright, D. W. Stashuk, and L. Pino, "Internal measures of reliability in 'pattern discovery' based fuzzy inference," in *Proc. 11th Int. Conf. Inform. Process. Manage. Uncertainty in Knowledge-Based Syst*, Le Cordeliers, Paris, France, 2006, pp. 340–347.

[34] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York, NY, USA: Wiley, 1973, p. 37.

[35] S. Marsland, *Machine Learning: An Algorithmic Perspective* (Machine Learning & Pattern Recognition Series).. Boca Raton, FL, USA: CRC Press, 2009, pp. 171–173.

[36] I. Rish, "An empirical study of the naive Bayes classifier," in *Proc. IJCAI-01 Workshop Empir. Methods Artif. Intell.*, 2001.

[37] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. Inf. Theory.*, vol. 14, no. 3, pp. 462–467, May 1968.

[38] S. Lee and K. C. Lee, "Context-prediction performance by a dynamic Bayesian network: Emphasis on location prediction in ubiquitous decision support environment," *Expert Syst. Appl.*, vol. 39, pp. 4908–4914, 2012.

[39] N. Friedman and M. Goldszmidt, "Bayesian network classifiers," *Mach. Learn.*, vol. 29, pp. 131–161, 1997.

[40] R. Varga, S. M. Matheson, and A. Hamilton-Wright, "Evidence aggregation for diagnosis: Bayesian and fuzzy strategies," presented at the Int. Joint Conf. North Amer. Fuzzy Information Processing Society Biannual Conf., Cincinnati, OH, USA, Jun. 2009.

[41] V. L. Durkalski, Y. Y. Palesch, S. R. Lipsitz, and P. F. Rust, "Analysis of clustered matched-pair data," *Statist. Med.*, vol. 22, no. 15, pp. 2417–2428, Jul. 2003.

[42] W. H. Kruskal and A. W. Wallis, "Use of ranks in one-criterion variance analysis," *J. Amer. Statist. Assoc.*, vol. 47, no. 260, pp. 583–621, Dec. 1952.

[43] P. Bonissone and W. Cheetham, "Fuzzy case-based reasoning for decision making," in *Proc. 10th IEEE Int. Conf. Fuzzy Syst.*, Dec. 2001, vol. 3, pp. 995–998.

[44] J. DeLeo and J. Dayhoff, "Medical applications of neural networks: Measures of certainty and statistical tradeoffs," in *Proc. Int. Joint Conf. Neural Netw.*, 2001, vol. 4, pp. 3009–3014.

[45] P. Fortier, S. Jagannathan, H. Michel, N. Dluhy, and E. Oneill, "Development of a hand-held real-time decision support aid for critical care nursing," in *Proc. 36th Annu. Hawaii Int. Conf. Syst. Sci.*, Jan. 2003.

[46] N. Tóth and B. Pataki, "On classification confidence and ranking using decision trees," in *Proc. 11th Int. Conf. Intell. Eng. Syst.*, Jun. 2007, pp. 133–138.

[47] T. Wetter, "Lessons learnt from bringing knowledge-based decision support into routine use," *Artif. Intell. Med.*, vol. 24, no. 3, pp. 195–203, Mar. 2002.

[48] C. P. Friedman, G. G. Gatti, T. M. Franz, G. C. Murphy, F. M. Wolf, P. S. Heckerling, P. L. Fine, T. M. Miller, and A. S. Elstein, "Do physicians known when their diagnoses are correct?" *J. General Internal Med.*, vol. 20, no. 4, pp. 334–339, Apr. 2005.

[49] A. Minutolo, M. Esposito, and G. De Pietro, "A pattern-based knowledge editing system for building clinical decision support systems," *Knowl.-Based Syst.*, vol. 35, pp. 120–131, 2012.

[50] M. Samwald, K. Fehre, J. de Bruin, and K.-P. Adlassnig, "The Arden Syntax standard for clinical decision support: Experiences and directions," *J. Biomed. Informat.*, vol. 45, pp. 711–718, 2012.

[51] *Proc. 11th Int. Conf. Inf. Processing and Manage. Uncertainty Knowl.-Based Syst.*, Le Cordeliers, Paris, France, 2006.

Authors' photographs and biographies not available at the time of publication.