The Influence of Feature Selection Methods on Exercise Classification with Inertial Measurement Units

Martin A. O'Reilly†, *Member, IEEE*, William Johnston†, *Member, IEEE*, Cillian Buckley, Darragh Whelan, and Brian Caulfield, *Member, IEEE*

***Abstract*— Inertial measurement unit (IMU) based systems are becoming increasingly popular in the classification of human movement. While research in the area has established the utility of various machine learning classification methods, there is a paucity of evidence investigating the effect of feature selection on classification efficacy. The aim of this study was therefore to investigate the influence of feature selection methodology on the classification accuracy of human movement data. The efficacy of four commonly used feature selection and classification methods were compared using four IMU human movement data sets. Optimisation of classification and features selection methodologies resulted in an overall improvement in F1 score of between 1-8% for all four data sets. The findings from this study illustrate the need for researchers to consider the effect classification and feature selection methodologies may have on system efficacy.**

I. Introduction

Strength and conditioning (S&C), neuromuscular control and running mechanics training are essential components for the optimisation of athletic performance, injury rehabilitation, and the prevention of injury [1-3]. However, successful implementation of such training and rehabilitation programs is influenced by factors such as the individual's adherence, their ability to maintain adequate exercise technique, and the ability to objectively measure deficits and changes in performance. Currently, when assessing running and exercise form in a clinical setting, clinicians and S&C coaches are required to rely on subjective visual observation of technique, and traditional screening tools, such as the Functional Movement Screen [4]. In addition, when assessing neuromuscular control performance, practitioners have relied on surrogate measures, such as the Y Balance Test (YBT). While these yield macro-level information pertaining to balance performance, they fail to provide objective micro-level information pertaining to a person's biomechanical profile [5].

In recent times, wearable sensor and smart-phone technology has provided the means to address some of these challenges through the objective quantification and analysis of human movement, in both laboratory and clinical settings. Furthermore, research has demonstrated the potential of such inertial sensor based systems to classify different conditions (i.e. aberrant and acceptable exercise form) during tasks ranging from early stage orthopaedic rehabilitation exercises, to more complex compound exercises and dynamic balance assessment [5-9].

When researchers are investigating the role of an inertial measurement unit (IMU) based system for human movement analysis, there are several key statistics and machine learning methodological challenges that need to be addressed to ensure the development of an accurate and efficient system. The first challenge involves determining which machine learning statistical methods are most suitable, given the chosen set of features. This question has received plenty of attention in the literature with previous work often demonstrating the superior classification accuracy of the Random Forest Classifier when compared with alternative methods such as Naive-Bayes, Support Vector Machine, Radial-Basis, K-Nearest neighbour [10].

The second challenge involves selecting the most important features for classification. This is of interest as it allows for computational optimisation, facilitating the implementation of cheap and accessible classification systems with minimal processing power e.g. smart-phone applications. Additionally, restricting the number of variables reduces the risk of over-fitting when learning in high dimensions from few samples, thus resulting in better predictions when considering new samples [11]. Determining the most valuable features involved in differentiating two conditions can also facilitate the development of clinically relevant visualisations and metrics capable of informing clinical and training decisions.

The third challenge, which has received little attention in the literature, is how we can best combine the feature ranking/selection and machine learning classification techniques to obtain the highest degree of accuracy, sensitivity and specificity. Previous research in the field of genomics has demonstrated that the accuracy obtained from various classification methods can be strongly influenced by the feature selection method employed [11], however this has yet to be investigated in the field of human movement classification using IMUs.

As such, this study seeks to determine the influence of various feature selection methods on the accuracy of four commonly used machine learning classification methods when applied to IMU human movement data. Additionally, we aim to investigate the effect feature selection methods may have on classification accuracy. Finally, we aim to investigate if the number of selected top ranked features has an influence of classification accuracy.

†Joint lead authors

*Research supported in part by Science Foundation Ireland (SFI/12/RC/2289) and the Irish Research Council in an enterprise partnership with Shimmer (EPSPG/2013/574).

Martin O'Reilly, William Johnston, Cillian Buckley, Darragh Whelan and Brian Caulfield are with the Insight Centre for Data Analytics and the Department of Public Health, Physiotherapy and Sports Science, University College Dublin, Republic of Ireland (e-mail: martin.oreilly@insight-centre.org)

978-1-5090-6244-7/17/$31.00 ©2017 IEEE

II. METHODS

A. Data Sets

Four human movement IMU data sets were used in the classifier and feature selection methods comparison (Table 1). All IMU data sets consist of tri-axial accelerometer, gyroscope and magnetometer data collected from a single lumbar worn Shimmer3 IMU (Shimmer, Dublin, Ireland) (Figure 1).

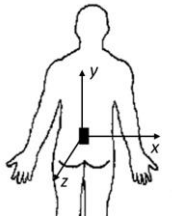

Figure 1: Illustrates the mounting location & orientation of the IMU for all four data sets.

The lunge and deadlift exercise data collection was completed as part of a larger study which included the collection of a range of IMU exercise data, requiring participants to complete repeated repetitions of the exercises with "acceptable" and "aberrant" form. While the detailed methodology for the deadlift data-set has yet to be published, it follows a similar structure to that outlined for the lunge data-set [7]. The YBT and running datasets involved the collection of "normal" and "aberrant" dynamic balance and running as influenced by fatigue. Detailed description of the methodologies for the collection of the two data set have previously been outlined in [5] and [10] respectively. The list of features used for all data sets replicates those described in [5] and [7].

Table 1: Description of the four IMU data sets. Each data-set is divided into two classes; "aberrant" (AB) and "acceptable" (AC). The number of participants (P), observations for each class and total (n) observations are provided.

Data Set	P	AC	AB	N
YBT	15	44	45	89
Running	21	292	292	584
Deadlift	80	796	1015	1811
Lunge	77	695	2334	3029

B. Feature Selection Methods

Four common feature selection methods were applied to each of the data sets described in Table 1. Each method was implemented using MATLAB 2016b (The MathWorks, Natwick, USA). In addition to the four described feature selection methods, variables were also assigned a random rank of importance. This, along with using all available variables for training and evaluating classifiers, added context to the results from utilising each feature selection method.

A paired t test (PTT) was initially applied to each data set. Features were ranked by their statistical significance according to P value in ascending order. The Wilcoxon sum-rank test (WSRT) was also applied to the variables in each dataset. Features were again ranked, in ascending order, by their statistical significance according to P value.

Random Forest recursive feature elimination (RF-RFE) was also employed to rank features in order of importance [12]. Initially benchmark accuracy was established through using all the computed features to train and test a random forests classifier. The process was to then eliminate each feature and measure how much the elimination decreases the accuracy of the model. For unimportant features, the permutation should have little to no effect on model accuracy, while holding-out important features should significantly decrease it. Following the recursive elimination of each feature, features were ranked based on their importance with the most important feature being that which caused the largest reduction in accuracy when permuted.

Finally, the feature rank value from the 3 feature selection methods (PTT, WSRT and RF-RFE) were aggregated by taking the mean rank. For instance, if the PTT deemed a variable the 10th most important, the WSRT deemed the same variable the 15th most important and the RF-RFE method deemed this variable the 20th most important, the ensemble-mean feature selection (EM-FS) method gave the variable a score of 15. All variables were then ranked in ascending order according to this mean score.

C. Data Analysis

To identify the best combination of feature selection method and classification algorithm for each dataset, the top 20-ranked variables (6.5% of all features) were taken from each variable ranking method and used to train and evaluate the four following classification algorithms: Random Forest (RF), support vector machines (SVM), k-nearest neighbours (KNN) and Naïve Bayes (NB). 400 trees were used for random forests classification in accordance with the previously published work on the same YBT and lunge data sets [5, 7]. The parameters of the SVM and KNN data sets were left as their default values as defined in MATLAB. Other parameters were also investigated to evaluate whether superior classification quality could be achieved.

Each classifier was evaluated using leave-one-subject-out-cross-validation (LOSOCV) as is common practice in the field of human activity recognition and exercise classification with IMUs [13, 14]. Classification quality, for each combination of feature selection method and classification algorithm, was assessed using the below formulae (1-4). F1 score was chosen as the most important metric for comparing each approach as the 'accuracy' score may be skewed by imbalances in the training/test data sets (Table 1).

$$1.\ Accuracy = \frac{TP + FN}{TP + FP + TN + FN}$$

$$2.\ Sensitivity = \frac{TP}{TP + FN}$$

$$3.\ Specificity = \frac{TN}{TN + FP}$$

$$4.\ F1\ Score = 2\left(\frac{Sensitivity\ x\ Specificity}{Sensitivity + Specificity}\right)$$

Following the identification of the strongest combination of classifier type and feature selection method for each dataset Cohen's kappa coefficient was computed to ensure correct classification had not been due to chance. We then aimed to identify how classification could be further improved. The effect of incrementing the number of top-ranked features to include in training and testing classifiers was investigated. Classification quality was again determined using LOSOCV and formulae 1-4. The optimal number of top-ranked features to include for classification was determined for the YBT, running, lunge and deadlift data sets.

III. Results

Table 2 shows the F1-scores achieved using LOSOCV for each classifier type, and each feature selection method used to select the top 20 features for each data-set. When the WRST feature selection method is utilised to select the top 20 features, the running dataset demonstrated an 8% improvement in F1-score. Cohen's kappa for this combination was 0.26. Previously published lunge results using all features and a random forests classifier were exceeded by 6% when using RF-RFE feature selection and an SVM classifier. Cohen's kappa in this circumstance was 0.31. The deadlift data set had the least improvement in F1-score, with only the RF-RFE feature selection method providing an improved F1-score when compared to using all features. Cohen's kappa for this combination was 0.33. Cohen's kappa for the YBT classifier with WSRT was 0.28.

Figure 2 shows the F1-score achieved when using different proportions of the WRST and RF-RFE top-ranked train and evaluate a random forests classifier on the YBT data-set. The maximum F1-Score achieved with the WSRT was 69.62% with 11% of top-ranked features. The maximum with RF-RFE was 70.95% with 60% of top ranked features. The horizontal black line demonstrates the F1-score when using all features for RF training and evaluation (62%).

IV. Discussion

This paper aimed to investigate the effect different combinations of feature selection and classification methods have on classification quality in 4 IMU data sets pertaining to exercise. When considering the 4 classification methods, the RF and SVM methods consistently achieved the highest F1-scores. In contrast, no single feature selection method consistently outperformed the others. Rather, the best feature selection method varied considerably depending on the data set under study. Notably, the WRST, RF-RFE and EM-FS methods considerably outperformed the PTT method when using RF classification algorithm. Our results differ vastly to hose presented by Haury et al [11] who demonstrated the PTT as the optimal feature selection method for four genomics data sets. As such, it may be suggested that developers of IMU based classification systems thoroughly investigate which feature selection method is most appropriate for their specific application.

*Table 2: F1- scores for each combination of feature selection & classifier methods. Highlighted in bold & marked with an * is the highest F1 score achieved for each data-set. ** Indicates support vector machine could not converge.*

Classifier	Dataset	All Features	Random	PTT	WSRT	RF-RFE	EM-FS
RF 400 trees	YBT	0.62	0.56	0.55	**0.64***	0.62	**0.64***
	Running	0.67	0.24	0.64	**0.75***	0.66	0.67
	Deadlift	0.64	0.53	0.64	0.61	**0.65***	0.63
	Lunge	0.64	0.61	0.66	0.67	0.67	0.67
SVM	YBT	**	0.43	0.34	0.52	0.62	0.60
	Running	**	0.27	0.56	0.64	0.41	0.42
	Deadlift	**	0.36	0.56	0.32	0.42	0.35
	Lunge	**	0.61	0.66	0.66	**0.70***	0.67
KNN	YBT	0.52	0.50	0.55	0.56	0.51	0.45
	Running	0.48	0.48	0.59	0.51	0.62	0.60
	Deadlift	0.51	0.51	0.52	0.53	0.57	0.55
	Lunge	0.54	0.55	0.57	0.56	0.59	0.54
NB	YBT	0.49	0.16	0.47	0.33	0.46	0.55
	Running	0.58	0.50	0.59	0.54	0.62	0.63
	Deadlift	0.60	0.53	0.59	0.50	0.59	0.50
	Lunge	0.58	0.50	0.55	0.55	0.44	0.57

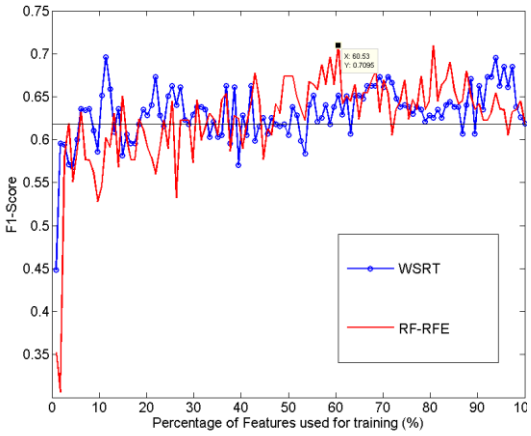

Figure 2: F1-score versus percentage of top-ranked features used for training and evaluating RF classifier on the YBT data set. The horizontal black line demonstrates the F1-score when using all features for RF training and evaluation.

In addition to identifying the optimal feature selection and classification method for specific IMU classification applications, our results demonstrated that finding the optimal number of top-ranked features to train and evaluate classifiers can further improve system accuracies. Figure 2 shows the F1-score achieved using the RF classifier with 400 trees when trained with incrementing subsets of top-ranked features from the YBT data set. Determining the optimum number of features allowed for an F1-score improvement of 8% compared to using all features. These findings suggest that depending on the data set size and type, incorporation of additional features that do not directly aid classification, may have a detrimental effect on overall classification accuracy. Therefore, we recommend that developers of IMU based classification systems also identify the optimal number of top-ranked features to include in training classifiers to ensure best accuracy. In addition, when determining the total number of features to include in a classification system, researchers should factor in the accuracy and efficiency required for specific applications.

There are several contextual factors to this investigation. Firstly, this study only investigated four IMU data sets pertaining to exercise. Therefore, comprehensive conclusions cannot yet be drawn on how best to approach feature selection on other IMU data sets. The data sets used, whilst variable in size, do not include any exceptionally large data sets. Therefore, we cannot make any inferences regarding the benefit or necessity of applying various feature selection methods to massive IMU data sets. It should also be noted that only a limited number of feature selection and classification methods were used in this study. Therefore, for the data sets under study, we cannot be certain we have identified the optimal feature selection-classifier combination. However, we have demonstrated that comparing such combinations can have beneficial outcomes on classification accuracy, sensitivity, specificity, F1-score and Cohen's kappa. This is in addition to the inherent benefits feature selection has for computational efficiency and algorithm interpretability.

In conclusion, this study has demonstrated the interesting interaction classification and feature selection methods, the number and combination of features selected and the data set type have on classifier training. It is imperative that researchers consider the need for individualised classification and feature selection methods, specific to the application, to ensure the development of accurate and efficient systems. Optimising these relationships in the design of IMU based systems can allow for considerably improved system efficacy, and improved understanding of each dataset and the associated variables of importance.

V. References

[1] N. van der Horst, D. W. Smits, J. Petersen, E. A. Goedhart, and F. J. Backx, "The preventive effect of the nordic hamstring exercise on hamstring injuries in amateur soccer players: a randomized controlled trial," *Am J Sports Med,* vol. 43, no. 6, pp. 1316-23, Jun, 2015.

[2] H. J. Hale, Olmsted-Kramer LC, "The Effect of a 4-Week Comprehensive Rehabilitation Program on Postural Control and Lower Extremity Function in Individuals With Chronic Ankle Instability," *Journal of Orthopaedic & Sports Physical Therapy,* vol. 37, no. 6, pp. 303-311, 2007/06/01, 2007.

[3] A. R. Diebal, R. Gregory, C. Alitz, and J. P. Gerber, "Forefoot running improves pain and disability associated with chronic exertional compartment syndrome," *Am J Sports Med,* vol. 40, no. 5, pp. 1060-7, May, 2012.

[4] G. Cook, L. Burton, and B. Hoogenboom, "Pre-participation screening: the use of fundamental movements as an assessment of function–part 1," *North American journal of sports physical therapy: NAJSPT,* vol. 1, no. 2, pp. 62, 2006.

[5] W. Johnston, M. O'Reilly, K. Dolan, N. Reid, G. F. Coughlan, and B. Caulfield, "Objective Classification of Dynamic Balance Using a Single Wearable Sensor," in Proceedings of the 4th International Congress on Sport Sciences Research and Technology Support, Porto, Portugal, 2016, pp. 15-24.

[6] D. Whelan, M. O'Reilly, T. Ward, E. Delahunt, and B. Caulfield, "Evaluating Performance of the Single Leg Squat Exercise with a Single Inertial Measurement Unit," in REHAB '15, Lisbon, Portugal, 2015, pp. 144-147.

[7] D. Whelan, M. O'Reilly, T. Ward, E. Delahunt, and B. Caulfield, "Evaluating Performance of the Lunge Exercise with Multiple and Individual Inertial Measurement Units," in Pervasive Health, Cancun, Mexico, 2016.

[8] M. O'Reilly, D. Whelan, C. Chanialidis, N. Friel, E. Delahunt, T. Ward, and B. Caulfield, "Evaluating squat performance with a single inertial measurement unit," in 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Boston, MA, 2015, pp. 1-6.

[9] D. Whelan, M. O. Reilly, B. Huang, O. Giggins, T. Kechadi, and B. Caulfield, "Leveraging IMU data for accurate exercise performance classification and musculoskeletal injury risk screening." pp. 659-662.

[10] C. Buckley, M. O'Reilly, D. Whelan, A. Valley Farrell, L. Clark, V. Longo, M. D. Gilchrist, and B. Caulfield, "Binary Classification of Running Fatigue using a Single Inertial Measurement Unit," in Conf. on Wearable and Implantable Body Sensor Networks, Eindhoven, Netherlands, 2017.

[11] A.-C. Haury, P. Gestraud, and J.-P. Vert, "The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures," *PLOS ONE,* vol. 6, no. 12, pp. e28210, 2011.

[12] a. Liaw, and M. Wiener, "Classification and Regression by randomForest," *R news,* vol. 2, no. December, pp. 18-22, 2002.

[13] S. J. Preece, J. Y. Goulermas, L. P. J. Kenney, D. Howard, K. Meijer, and R. Crompton, "Activity identification using body-mounted sensors--a review of classification techniques," *Physiological measurement,* vol. 30, no. 4, pp. R1-R33, 2009.

[14] O. D. Lara, and M. a. Labrador, "A Survey on Human Activity Recognition using Wearable Sensors," *IEEE Communications Surveys & Tutorials,* vol. 15, no. 3, pp. 1192-1209, 2013.