2015 IEEE Symposium Series on Computational Intelligence

Avoiding Bias in Classification Accuracy - a Case Study for Activity Recognition

Heli Koskimäki
Biomimetics and Intelligent Systems Group, University of Oulu
PO BOX 4500, 90014 University of Oulu
Email: heli.koskimaki@ee.oulu.fi

***Abstract*—The amount of studies on classification of human characteristics based on measured individual signals has increased rapidly. In wearable sensors based activity recognition a common policy is to report human independent recognition results using leave-one-person-out cross-validation scheme. This can be a suitable solution when feature or model parameter selection is not needed or it is done outside the validation scheme. Unfortunately, this is not always the reality. Thus in this article it is studied how the train-validate-test approach changes the recognition rates compared to basic leave-one-out cross-validation approach. Results of three different ways to perform the train-validate-test is presented: 1) single division to training and testing data, 2) 10-fold division to training and testing data, and 3) double leave-one-person-out cross-validation. In this article, it is shown that the best classifier or feature set selected based on the training and validation data using basic leave-one-out approach does not always perform best within independent testing data. Nevertheless, a larger bias to results can be achieved using single division or even 10-fold division into separate training and testing data. Thus it is stated that the double leave-one-person-out is the most robust version for reporting classification rates in future studies of activity recognition as well as other areas where human signals are used.**

I. Introduction

Wearable sensors based activity recognition is a research area where inertial measurement units based information is used to recognize human activities. The overall activity recognition process includes a data set collected from the activities wanted to be recognized, preprocessing, segmentation, feature extraction and selection, and classification [1]. Activity recognition is used in recognizing, for example, daily activities [2], [3], in sport sector [4], [5], [6] and in monitoring of assembly tasks [7], [8], [9]. To instruct and standardize the future studies an overall tutorial for activity recognition is published [1] and metrics to be deployed are presented [10]. Nevertheless, the usage of train-validate-test method is not instructed nor commonly used for reporting the recognition rates.

Although the use the separate test data is not a novel approach in machine learning [11] it is easy to ignore if there are no concrete studies about its impact in the area in question. Thus is this article, it is pointed out that the recommend and mainly used evaluation scheme based on leave-one-person-out cross-validation has some weaknesses. In theoretical perspective the question is; if in feature selection the same data is used for selecting the features and validating the features, as well as if in classifier training the model parameters are tuned based on the same data to be modeled, are the reported recognition rates biased. In this article, it is indeed shown that they are. For example, the leave-one-person-out cross-validation can be misleading when selecting the best features as well as the best classifier which is a major drawback, for example, when trying to introduce novel / improved classification processes. Nevertheless, the basic 70/30 division into separate training and testing data bias the results even more.

Moreover, in this article it is shown that one of the most common approach for feature selection using sequential forward selection (SFS) in person-wise cross validation, in practice, skews the activity recognition results even more. The initial assumption was that the skewness would be due the over-fitting and better classification accuracy for testing data would be achieved by selecting less features than the feature selection process implies. Nevertheless, it is the opposite. The highest classification accuracies with linear (LDA) and quadratic (QDA) discriminant analysis classifiers in testing data are achieved by selecting substantially more features than the SFS feature selection process implies.

II. Data sets

In this study three different activity recognition data sets are used; assembly data set, swimming data set and daily activities data set. With all the cases only 3D acceleration data collected from the right wrist was used to make the data sets comparable.

In assembling task data were collected from 10 persons assembling a wooden drawer containing 6 different activities: hammer use, screwdriver use, wrench use, tapping, leg adjusting with the right hand and attachment of small tips. Naturally the sequence also contained so-called null data, where none of the above-mentioned activities were performed. One assembly sequence lasted approximately 8 minutes, and these sequences were collected twice from each person. More information about the data set can be found in [12]. The swimming data consisted data from 19 swimmers of which 9 were professional and 10 were amateurs. The data were collected from three different swimming styles (free stroke, back stroke and breast stroke) as well as basic and flip turns based on skills of the swimmer. Thus, for example, the flip turn data was achieved only from professional swimmers. The data was collected at 50Hz but because of the higher recognition rate presented with lower frequency in [13] the data was sampled into 10Hz. The daily activity data set included data from 18 somewhat overlapping activities from 21 persons. The activities included walking, pushing lawnmower, biking, jogging, playing football (soccer), running, washing dishes, dusting, mopping, ironing,

978-1-4799-7560-0/15 $31.00 © 2015 IEEE
DOI 10.1109/SSCI.2015.52

TABLE I: Data set descriptions

Data set	Amount of persons	Amount of activities	Length of window	Slide	Amount of windows
Assembly	10	6 + null	2s	0.5s	19 000
Swimming	19	4	6.4s	1.6s	13 500
Daily activities	21	18	4s	1s	90 000

vacuuming, ascending and descending stairs, sitting, working with computer, teaching at blackboard, reading book, reading magazines and packing/unpacking boxes. From every activity and from every person 4 minutes of data were collected.

III. Methods

To be able to classify the continuously measures acceleration signals all the data sets were divided into segments using the sliding window method. The window length and the overlapping were decided based on previous studies (see Table I).

Within each window specific features were calculated, although, to be able to test the feature selection step for activity recognition the amount of features was intentionally exploded in this study. Thus the features included previously used features for activity recognition (e.g. in [14]) as well as novel features called in this article *phase based features*. In practice, these features were calculated for every signal one signal at a time by searching crossings of the signal with a predefined limit *L*. Between these crossings *the biggest and smallest area* below and above of the limit as well as *the biggest and smallest amplitude* were calculated for the signal in question. In addition, area information at the corresponding sequence of the other two signals were selected as features as well as the corresponding amplitude of the signals (see Figure 1). In this article three limits were used (0, 1 and mean). These features were then divided into three sets where the first set basically consisted of statistical features (standard deviation, mean, minimum, maximum, percentiles, sequence periods, zero-crossings and mean-crossings for every channels and correlation information between the channels), second set added wavelet and FFT features and the third set the phase based features.

In this article, the best features were chosen using sequential forward selection (SFS) and minimum Redundancy Maximum Relevance Feature Selection (mRMR). In the SFS the the best features were selected one at a time using the classification accuracy of the model(s) in question as a selection criteria [15] . However, the selection was not stopped at local minimum but it was allowed to choose until "the best features" included all the features. On the other hand, with mRMR the feature selection was done model independently by selecting features having the highest correlation to the classification variable but locating far from each other [16]. Nevertheless, also when using mRMR the amount of features was selected using the model independent feature ranking with the model dependent maximum classification accuracy in leave-one-person-out cross-validation for each model separately.

The classifiers used in this study were the parametric linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), C4.5 tree classifier, k-nearest neighbor classifier (kNN) and support vector machines (SVM). The LDA and QDA

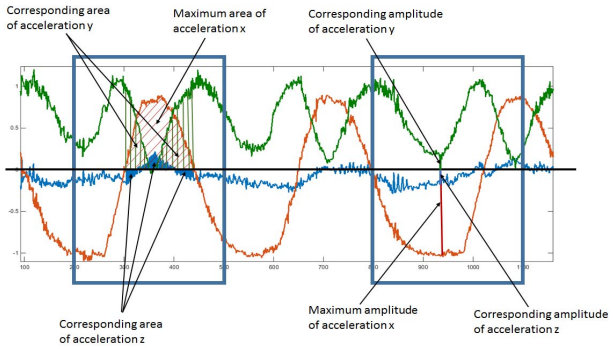


Fig. 1: The biggest area and maximum amplitude for acceleration signal x as well as the corresponding areas and amplitudes for acceleration signals y and z.

model the class-conditional densities parametrically as multivariate normals [17]. In practice, QDA separates classes using nonlinear decision boundaries while LDA uses linear decision boundaries. Both of the methods are fast to train, easy to implement and the memory requirements are small thus making them well-liked in practical applications and devices. The decision tree models are also computationally light models that partition the space banned by the input variables to maximize the score of class purity. Purity is achieved by ensuring that the majority of points in each cell of the partition belong to one class [11]. In C4.5 the partition is based in the difference in entropy [17]. On the other hand, kNN relies heavily on the training data requiring in most cases more memory and calculation capacity in the actual classification phase compared to the parametric methods. In practical applications targeted to small embedded devices the memory requirement makes it impossible to use but as an reference method it is appropriate due its capability to create multiform class boundaries. The basic idea of kNN classification is quite simple: a data point is classified into the class where most of its k nearest neighbors belong [18]. In this work, the features were standardized between $[0, 1]$ before calculating the distances. The SVM, on the other hand, relies on nonlinear mapping and transforms the original features to typically higher dimensional feature space. In this higher dimension a suitable hyperplane is sought to separate the classes [17]. In a sense of calculation the SVM is the most time consuming.

IV. Study

In this study three different ways to perform the train-validate-test approach were tested (see Figure 2). Results of these approaches are presented in their own subsections while the overall discussion is left to Section V. In every subsection the results of the basic leave-one-person-out cross-validation for the training data are also presented as a comparison because the selected approach also reflected to the cross-validation results. Thus when it is stated that the results are presented for training/validation data it means that the results are shown as an average of person-wise validation data within the training data in question.

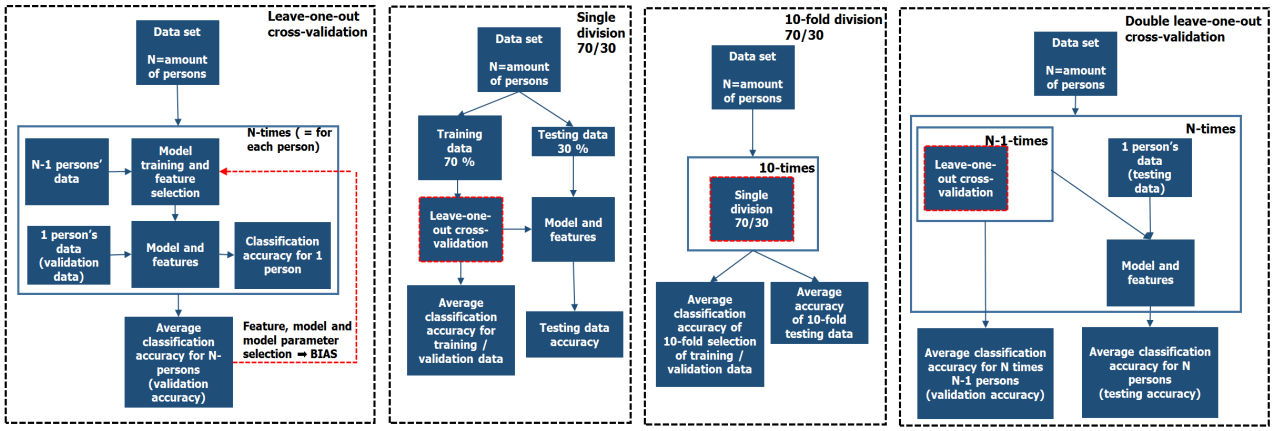


Fig. 2: Principles of basic leave-one-person-out cross-validation and three ways for train-validate-test approach: single division, 10-fold division and double leave-one-person-out cross-validation. The bias in basic leave-one-person-out cross-validation is due the back propagation marked with slashed red arrow. If there are no back propagation the basic scheme is adequate.

A. Single division into training and testing data (70/30)

At this subsection, the data sets were randomly divided into separate training (70 % of data) and testing sets (30 % of data) before the models were trained or suitable features were selected based on leave-one-person-out cross-validation procedure within the training data. However, the separation for training and testing data was done person-wise in a sense that 30 percent of the persons were selected into testing. This means that the test data consisted of data from more than a single person.

While the mRMR is a classifier dependent, at least when considering the order of features selected, it was fast to adapt with 5 different classifiers. On the other hand, while the SFS is computationally more challenging than mRMR the SFS feature selection was only studied with LDA and QDA classifiers.The results for this study can be seen from Table II. In the table it is shown the recognition rates themselves for training/validation and testing data but also the number of features providing the highest recognition rate within training/validation data. With SFS also the best overall accuracy is presented in the table.

With the assembly data set the difference between training/validation and testing data is shown very drastically. When considering the classification accuracies in testing data the difference can be almost 20 percentage units less than with training/validation data. At this point, it has to be remembered that this is not due an anomaly caused by a single person but the testing data included in this case data from three different persons. With swimming, however, a rise in accuracy can be seen. Nevertheless, the table also shows how the classifier and feature set giving the best accuracy within training data do not necessarily produce the best classification within separate testing data. For example, with daily activities the best recognition rate 74.0 % is achieved using all the features with LDA while in testing the highest accuracy is produced by using QDA with feature sets 1 and 2. On the other hand, the table also shows that the generalization problem is not highly depended on the classifier used.

Within the two different feature selection methods the SFS seems to outperform the mRMR if only the validation accuracy is studied. In swimming, however, the testing data accuracy is higher with mRMR. This is at least the case when comparing just the features selected by using the maximum recognition rate of the training/validation data. An another interesting aspect seen in Table II is the amount of the features chosen with different data sets and models. It was mentioned that mRMR selects the features model independently, but as it can be seen, the amount of features corresponding to highest recognition rate changes remarkably based on the classifier. For example, with swimming data and using all the features the amount of chosen features with LDA is 224 while with SVM only 4 features are used. Also, interesting is to see that the most simplest classifiers LDA and QDA produces the best recognition rates with every data set. On the other hand, when shifting the interest towards the best possible recognition rates within testing data by selecting more and more features using the order decided by SFS, an another interesting aspect is noticed. The initial assumption was that the SFS would suffer from overfitting and better classification accuracy for testing data would be achieved by selecting less features than the feature selection process implies. Nevertheless, it is the opposite. The highest classification accuracies for testing data are achieved by selecting substantially more features than the feature selection process implies. In fact with LDA results, this is true for almost all the activities and within all the feature sets. Only exception to this is found when using the assembly data and using the so called basic features.

B. 10-fold division into training and testing data (70/30)

While the results seemed quite drastic with a single separation into training and testing data the results presented in this subsection were averaged using 10 different runs of 70/30 division. While the calculation burden made it troublesome to run comprehensive tests with separate classifiers and feature selection methods within the 10-fold train-validate-test selection approach (e.g. in SVM the calculation with one selection took weeks in calculation) at this point the most of the comparison was left out. Thus, the classification accuracies are calculated only by using LDA and kNN classifiers,

TABLE II: Recognition rates for 5 different classifiers, 3 feature sets and mRMR or SFS feature selection algorithm.

Dataset	Classifier	Feature set	mRMR feature selection		SFS feature selection		
			Validation accuracy (number of features)	Testing accuracy selected based on training data	Validation accuracy (number of features)	Testing accuracy selected based on training data	Best validation accuracy (number of features)
Assembly	LDA	1	80.1 (25)	61.5	83.3 (16)	60.7	66.0 (11)
		1 & 2	83.0 (84)	66.3	85.8 (19)	65.4	68.0 (73)
		1 & 2 & 3	82.9 (126)	66.6	86.4 (71)	66.1	68.9 (181)
	QDA	1	81.9 (19)	64.3	85.1 (22)	70.6	**72.8 (43)**
		1 & 2	**83.5 (23)**	**68.2**	86.3 (32)	70.0	72.0 (15)
		1 & 2 & 3	83.3 (20)	66.2	**87.4 (40)**	**72.0**	**72.8 (68)**
	C4.5	1	59.9 (19)	50.2			
		1 & 2	63.7 (45)	56.6			
		1 & 2 & 3	65.5 (91)	55.2			
	kNN	1	77.0 (45)	63.9			
		1 & 2	76.6 (86)	62.3			
		1 & 2 & 3	75.8 (129)	61.9			
	SVM	1	62.4 (7)	51.3			
		1 & 2	65.5 (6)	49.0			
		1 & 2 & 3	64.8 (8)	48.9			
Swimming	LDA	1	86.1 (51)	88.9	87.3 (15)	88.0	89.0 (48)
		1 & 2	**89.8 (86)**	91.0	92.0 (46)	**90.4**	91.2 (67)
		1 & 2 & 3	88.6 (224)	**91.5**	92.5 (41)	90.0	**91.9 (231)**
	QDA	1	80.4 (21)	85.7	86.1 (16)	85.6	86.6 (5)
		1 & 2	81.9 (5)	87.1	89.3 (19)	87.9	88.3
		1 & 2 & 3	87.4 (22)	87.2	**92.7 (28)**	89.4	89.5 (49)
	C4.5	1	79.6 (20)	84.5			
		1 & 2	79.5 (35)	80.9			
		1 & 2 & 3	80.7 (15)	82.8			
	kNN	1	85.4 (49)	89.5			
		1 & 2	86.3 (14)	87.1			
		1 & 2 & 3	87.0 (34)	87.9			
	SVM	1	80.3 (7)	76.5			
		1 & 2	81.5 (5)	86.2			
		1 & 2 & 3	85.5 (4)	85.0			
Daily	LDA	1	66.9 (50)	64.0	67.5 (41)	63.7	64.2 (48)
		1 & 2	71.9 (88)	68.2	72.6 (54)	67.8	68.3 (83)
		1 & 2 & 3	**74 (253)**	69.5	75.1 (119)	67.9	69.5 (241)
	QDA	1	64.6 (51)	62.5	70.9 (16)	66.8	66.8 (16)
		1 & 2	71.4 (88)	**70.2**	75.7 (24)	**72.4**	**72.4 (27)**
		1 & 2 & 3	73.4 (252)	69.0	**75.9 (29)**	72	72.3 (26)
	C4.5	1	64.7 (33)	62.6			
		1 & 2	67.7 (81)	65.9			
		1 & 2 & 3	67.7 (197)	67.2			
	kNN	1	72.0 (42)	68.4			
		1 & 2	72.0 (59)	69.0			
		1 & 2 & 3	72.6 (78)	68.8			
	SVM	1	70.1 (15)	66.1			
		1 & 2	70.6 (11)	68.9			
		1 & 2 & 3	70.6 (18)	67.2			

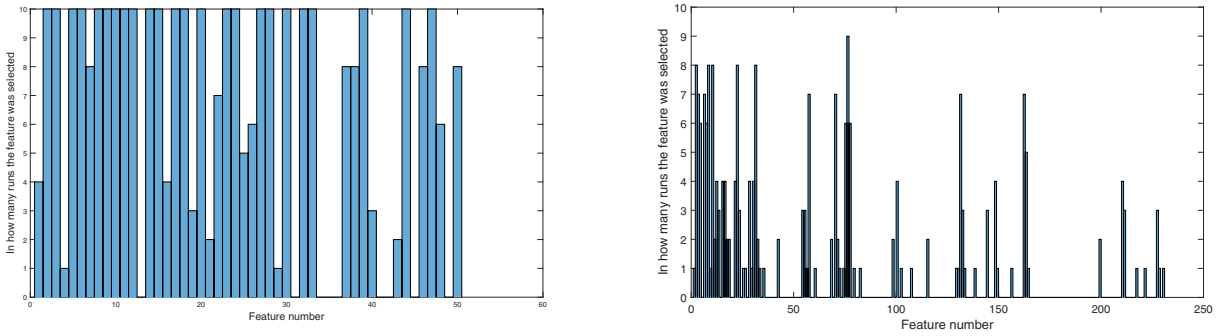
mRMR feature selection and the feature set 1 while the results are shown as an average of 10 different runs for both to training/validation data as well as to testing data for three of the data sets (Table III). To get comparable results the same selections were used with both classifiers.

TABLE III: Average recognition rates with 10-fold train-validate-test selection with LDA and kNN classifiers

Data set	Validation accuracy with LDA	Testing accuracy with LDA	Validation accuracy with kNN	Testing accuracy with kNN
Assembly	**77.5 ±1.8**	**72.1 ± 6.5**	74.7 ± 1.9	69.0 ± 3.6
Swimming	**86.0 ± 1.0**	88.9 ± 2.9	84.6 ± 1.5	**89.0 ± 3.0**
Daily activities	65.8 ± 0.8	65.8 ± 1.8	**70.1 ± 1.1**	**70.3 ± 1.4**

Nevertheless, also these simple results show that the best classification accuracy in training/validation data is not the one leading to the highest accuracy within the separate testing data. From the Table III all the possible behaviors of testing data can be seen. With daily activities the recognition rates correspond well and with swimming data set the recognition rates are even higher within the separate testing data but with assembly there is a drop of five percentage units in accuracy. Moreover, with the assembly data the variation between separate runs of train-validate-test selection is remarkably high.

While the results implied that the classification accuracies of testing data do not necessarily correlate with the results achieved using the leave-one-person-out cross-validation an interest raised about the features selected in these different runs. To test the variation in feature selection a test was done using the assembly data and selecting the best features within two different cases: using mRMR, LDA and feature set 1, and using mRMR, QDA and all the features (although the positive definiteness ruled some of the features out). The features selected are shown in Figure 3. With the LDA case the variation does seem quite rational; in the 10 different runs the amount of features selected was between 25 and 39 and it can be seen that 24 same features were selected in every run. Thus, it can be assumed that there are 24 person-independent and highly informative features. Nevertheless, with QDA the amount of features selected between 10 separate runs changed between 10 and 44 and none of the features was selected in



(a) Using LDA and feature set 1

(b) Using QDA and all the features

Fig. 3: Histograms for features selected during 10 different runs using the assembly data.

every run. Moreover, there is only one feature selected even in 9 runs. Thus, in this case the features selected cannot be assumed to be person independent.

C. Double leave-one-person-out cross-validation

In double (or nested) leave-one-person-out cross-validation the bias is avoided by adding an outer loop into cross-validation. Data from one person at a time is chosen as separate testing data while the rest N-1 persons data is left for basic leave-one-person-out cross-validation. This approach is computationally the most challenging especially when amount of persons in data set increases but when reporting the results it is model, model parameter and feature selection independent.

The results for this train-validate-test is presented in Table IV. With this time the features chosen were the basic statistical features added with fourier and wavelet features (feature sets 1 & 2) while the best features were selected using mRMR feature selection. The classifiers used were LDA and knn. From the results it can be seen that the difference of training/validation data and separate testing data is this time smaller or even non-existing. Nevertheless, a drop in recognition rate is noticed within the assembly data with both methods and within swimming results achieved by knn.

TABLE IV: Average recognition rates with double leave-one-out cross-validation train-validate-test selection with LDA and kNN classifiers (k=5)

Data set	Validation accuracy with LDA	Testing accuracy with LDA	Validation accuracy with kNN	Testing accuracy with kNN
Assembly	**80.6**	**80.2**	75.2	74.5
Swimming	**89.9**	**89.9**	85.5	84.4
Daily activities	**71.5**	**71.4**	70.9	70.7

Although, the differences in classification accuracies between double and basic leave-one-person-out cross-validation were not large, an interest raised if it would have in the model parameter selection. For this purpose, the value of k for knn was tested using the same scenario as above. This test was done with assembly and swimming data sets. With this test and by using assembly data the effect is shown (Table V). The value of k selected using back propagation in leave-one-out cross-validation is not the one leading to best results in double version.

TABLE V: Recognition rates of knn classifier with k = 5 and k = 9.

Data set	Validation accuracy k=5	Testing accuracy k=5	Validation accuracy k=9	Testing accuracy k=9
Assembly	**75.2**	74.5	74.9	**74.8**
Swimming	**85.5**	84.4	**86.0**	**84.9**

V. Discussion

As the results showed the classification accuracy of the training/validation data does not always correlate with the accuracy in the testing data in activity recognition studies. It was shown that the selection of best classifier or the best features based on training data can be quite misleading. This can be a major drawback when developing new methods or introducing new features into activity recognition. One reason for the variation in results is the fact that every person is individual.

Based on the results of this article, it is also obvious that the selection of only a single, random testing data can be misleading. The results show in someway the best and the worst cases with separate testing data. To be able to have more weight on the results of the separate testing data also the train-validate-test procedure should be repeated several times, for example, applying at least the 10-fold testing data selection or by using the double leave-one-person-out cross-validation. It seems that from these the later would be more recommendable while it does not need any randomization.

Beside the remarks of train-validate-test procedure usage this article also contributed to feature extraction and selection parts of activity recognition. The novel features presented gave a way to study how the amount of features used effect the recognition rates. It was shown that in the most of the cases with LDA classifier the more the features the better. For example, with mRMR and all the features the amount of features selected moved between 126 and 253 which is

shockingly large amount. Moreover, it is not reasonable to select the best features using mRMR given impact values while the impact of features to the recognition rates is highly dependent of the classifier. On the other hand, it was noted that the SFS does not select too much features but in practice too few.

While the classification accuracy was not the main criterion in this article, it is not assumed that better results could not be achieved for the individual problems. For example, the optimization of the methods was not taken into true consideration. Nevertheless, it does not undermine the outcomes of this study - the missing of separate testing data can bias the activity recognition results or results in any area where signals measured from humans are classified.

VI. Conclusion

In this article the bias of classification accuracy caused by the back propagation step in leave-one-person-out cross-validation was studied. It was shown that the bias do not just effect to the classification accuracy itself but the selection of the optimal features as well as classifier and its parameters can be effected. This similar effect can be anticipated also other areas where signals measured from humans are used. Thus extra consideration should be put when reporting new results.

How to pick the best unbiased model and features based on the presented results will be the next step of the study.

References

[1] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Comput. Surv.*, vol. 46, no. 3, pp. 33:1–33:33, Jan. 2014. [Online]. Available: http://doi.acm.org/10.1145/2499621

[2] O. Banos, M. Damas, H. Pomares, A. Prieto, and I. Rojas, "Daily living activity recognition based on statistical feature quality group selection," *Expert Systems with Applications*, vol. 39, no. 9, pp. 8013–8021, 2012.

[3] M. Zhang and A. A. Sawchuk, "Human daily activity recognition with sparse representation using wearable sensors," *Biomedical and Health Informatics, IEEE Journal of*, vol. 17, no. 3, pp. 553–560, May 2013.

[4] K. Chang, M. Chen, and J. Canny, "Tracking free-weight exercises," *UbiComp 2007: Ubiquitous Computing*, pp. 19–37, 2007.

[5] H. Koskimäki and P. Siirtola, "Recognizing gym exercises using acceleration data from wearable sensors," in *Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on*. IEEE, 2014, pp. 321–328.

[6] P. Siirtola, H. Koskimäki, V. Huikari, P. Laurinen, and J. Röning, "Improving the classification accuracy of streaming data using sax similarity features," *Pattern Recognition Letters*, vol. 32, no. 13, pp. 1659–1668, 2011.

[7] T. Stiefmeier, D. Roggen, G. Tröster, G. Ogris, and P. Lukowicz, "Wearable activity tracking in car manufacturing," *IEEE Pervasive Computing*, vol. 7, no. 2, pp. 42–50, 2008.

[8] H. Koskimäki, V. Huikari, P. Siirtola, P. Laurinen, and J. Röning, "Activity recognition using a wrist-worn inertial measurement unit: a case study for industrial assembly lines," *The 17th Mediterranean Conference on Control and Automation*, pp. 401–405, 2009.

[9] H. Koskimäki, V. Huikari, P. Siirtola, and J. Röning, "Behavior modeling in industrial assembly lines using a wrist-worn inertial measurement unit," *Journal of Ambient Intelligence and Humanized Computing*, vol. 4, no. 2, pp. 187–194, 2013.

[10] J. A. Ward, P. Lukowicz, and H. W. Gellersen, "Performance metrics for activity recognition," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 1, pp. 6:1–6:23, 2011.

[11] D. J. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*. MIT Press, 2001.

[12] V. Huikari, H. Koskimäki, P. Siirtola, and J. Röning, "User-independent activity recognition for industrial assembly lines-feature vs. instance selection," in *5th International Conference on Pervasive Computing and Applications*. IEEE, 2010, pp. 307–312.

[13] P. Siirtola, P. Laurinen, J. Röning, and H. Kinnunen, "Efficient accelerometer-based swimming exercise tracking," in *Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on*. IEEE, 2011, pp. 156–161.

[14] O. Lara and M. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 1192–1209, 2012.

[15] P. A. Devijver and J. Kittler, *Pattern recognition: A statistical approach*. Prentice-Hall London, 1982, vol. 761.

[16] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 8, pp. 1226–1238, Aug 2005.

[17] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[18] E. Fix and J. L. Hodges Jr., "Discriminatory analysis - nonparametric discrimination: Consistency properties," *Technical Report 4, U.S. Air Force, School of Aviation Medicine, Randolph Field, TX*, 1951.