Comparison of Wrapper and Filter Feature Selection Algorithms on Human Activity Recognition

Jozsef Suto, Stefan Oniga
Department of Informatics Systems and Networks
University of Debrecen
Debrecen, Hungary
suto.jozsef@inf.unideb.hu, oniga.istvan@inf.unideb.hu

Petrica Pop Sitar
Technical University of Cluj-Napoca, Center Baia Mare
Department of Mathematics and Informatics
Baia Mare, Romania
petrica.pop@ubm.ro

***Abstract*—Feature selection is an increasingly important part of machine learning. The purpose of feature selection is dimension reduction in a large multi-dimensional data set and it can be the key step of successful knowledge discovery in those problems where the number of features is large. This research area has huge practical significance because it accelerates decisions and improves performance. The requirements of specific applications in different kinds of research areas have led to the development of new feature selection techniques with different properties. In the last few decades, several feature selection algorithms have been proposed with their particular advantages and disadvantages. Despite of the intensive research and the large amount of works, the different kinds of feature selection algorithms have not been tested yet in the human activity recognition problem. It was the main motivation of our work and this paper tries to fill this gap. Therefore, in this article we present a conceptually simple naïve Bayesian wrapper feature selection method and compare it with some widely used filter feature selection algorithms. The result of this work demonstrates that, the wrapper technique outperforms filter algorithms in this type of problem. In addition, this paper shows an example, when the classifier dependency of a wrapper method do not visible.**

Index Terms - artificial neural network; feature selection; human activity recognition; machine learning

I. Introduction

Machine learning is a multidisciplinary research area with countless application opportunities. The success of a machine learning application depends on many factors, one of them being the data quality. If the data contains irrelevant, redundant or noisy elements, then knowledge discovery will be difficult. In the last decade, the innovations in hardware technology highly increased the quantity and dimensionality of measurable information. In addition, new types of information sources have also evolved. Measured information has been becoming increasingly larger in both number of instances and number of features in many applications such as microarray data analysis, text classification, human activity recognition, image processing, etc., see for example [1-7]. The huge and high dimensional data may contain a large amount of irrelevant and redundant information which will reduce the performance of machine learning. Consequently, researchers and practitioners recognized that feature extraction and feature selection are essential steps not only in machine learning but also in other scientific areas [8].

Feature selection, also called feature reduction, is the process of choosing a subset of original features according to a well-defined evaluation criterion. It is a frequently used dimensionality reduction technique which removes irrelevant and redundant features. This approach has more useful effects for real applications because it accelerates the algorithms, improves the performance and simplifies the model. In contrast to other dimensionality reduction techniques like linear discriminant analysis (LDA) or principal component analysis (PCA) which are based on projection, feature selection do not alter the original representation of feature sets [9, 10]. According to the training data which may be tagged, untagged or partially tagged, there have been developed three categories of algorithms: supervised, unsupervised and semi-supervised feature selection where a tag refers to a given class. In addition, depending on the feature evaluation process, the feature selection algorithms belong to three different groups: filter, wrapper and embedded. A general introduction to feature selection can be found in [13]. Essentially, the purpose is to find the most appropriate hyperplane of the n-dimensional feature space in all cases, see [11, 12]. In this article we confine on the supervised category and particularly we are interested in filter and wrapper methods.

II. Related Work

Nowadays, a reach literature exists concerning the feature (or variable) selection. During the last years, there have been developed a large amount of filter algorithms. The proposed algorithms are based on different kinds of approaches such as statistical, information theory, rough set, etc. [3, 4, 17-22]. These algorithms are designed to serve different purposes in different kinds of models. For example, in [3] the authors proposed a special feature selection algorithm for background elimination of images, He et al. demonstrated the performance of the Laplacian score on Iris and PIE face data sets [20] while Bashir et al. developed a novel feature selection technique to gait energy images (GEI) [4]. In the past few years, some books and surveys have been published containing the most common feature selection methods and theirs application areas, see for example [23-25]. For instance, in [24] the features have been selected from EEG signal in order to increase prediction accuracy of sleep stage classification while the authors in [25] presented a detailed review about the application of feature selection techniques in bioinformatics. Moreover, some comparisons of feature selection methods exist in the literature where a few methods were tested on the same problem [26-28]. The authors of [26] compared nine filter algorithms on microarray data and examined the correlation between each other. In another study, the gait ratio and some correlation based approaches with different learning algorithms were used to select relevant features from a diabetic dataset [27]. In [28], five filter methods were evaluated in text categorization where the authors similarly noticed correlation between filter algorithms as in [26]. Two other approaches used the difference between categories and feature evaluation processes to the comparison. For example, in [29] the authors presented a joint study of supervised and unsupervised feature selection. However, despite of intensive research and a large amount of papers, feature selection methods have not been tested yet in the human activity recognition (HAR) machine learning problem. Nowadays, HAR is an intensive research field of machine learning where different features from the time, frequency and wavelet domains have been used, see for example [15, 16]. Although many articles were presented in this topic, some questions are unanswered yet. One of theme is the type of the most efficient feature selection algorithm which determines the best feature combination. Therefore, the goal of this work was to give an answer for this question.

In real applications filter methods are frequently used for feature selection because they have some significant advantages [14]. Firstly, those methods are independently applicable with any types of machine learning techniques. Secondly, filter methods are faster than wrappers. However, wrapper methods are more efficient than feature ranking algorithms in more cases because they take into consideration the classifier hypothesis. This also means that, wrapper technique can handle feature dependencies. So, both types of feature selection methods have advantages and disadvantages. Unlike filter algorithms, the literature of wrapper techniques is smaller. One of the most famous articles about wrapper methods can be seen in [40]. In this paper, Kohavi and John give a detailed description about the strengths, weaknesses and search strategies of wrapper methods. In another study [30], the authors used the work of Kohavi and John and compared the correlation-based feature selection (CFS) algorithm with the wrapper models from [40]. Beyond the used classification algorithm, the operation of a wrapper method depends on more attributions such as search strategy, stop criterion, etc. However, the concrete description of the algorithms is missing (or deficient) from the articles as in [31]. It was the main motivation to use an own wrapper algorithm which effectively removes irrelevant features from the original feature set and tries to decrease the general disadvantages of wrapper techniques (computation cost, classifier dependency).

III. The Proposed Wrapper Algorith

The idea behind the proposed wrapper algorithm is based on the *Bayes'* rule. It gives back the *posterior* probability from the *prior* and *class-conditional* probabilities.

$$P(C_i \mid X_j) = \frac{P(X_j \mid C_i)\, P(C_i)}{P(X_j)} \tag{1}$$

In the above formula, $P(C_i)$ (prior probability) tells us how likely a sample belongs to the *i'th* class, $P(X_j)$ indicates how likely a sample will be in the *j'th* bin and $P(X_j|C_i)$ (conditional probability) is the probability of a sample is in *j* given that the class is *i*. In this context, bin is a short and closed subinterval of the whole interval. In our case, the samples are feature vectors (**X**) from each class with more features. Therefore, the class-conditional probability in the Bayes' rule can be written in the following form,

$$P(\mathbf{X}_j \mid C_i) = P(\mathbf{X}_j^1, \mathbf{X}_j^2, \ldots, \mathbf{X}_j^n \mid C_i) \tag{2}$$

where *n* is the size of the vector. If we suppose that the features are independent, (2) will be equal to the product of multiplying together each independent components.

$$\prod_{k=1}^{n} P(\mathbf{X}_j^k \mid C_i) \tag{3}$$

Since, $P(X_j)$ in formula (1) is just a normalizer factor, the posterior probability can be approximated by the following computation,

$$P(C_i \mid \mathbf{X}_j)^* = P(C_i) \prod_{k=1}^{n} P(\mathbf{X}_j^k \mid C_i) \tag{4}$$

The posterior probability assigns a new sample to one of the classes according to the following inequality,

$$P(C_i \mid \mathbf{X}_j)^* > P(C_j \mid \mathbf{X}_j)^* \quad \forall i \neq j \tag{5}$$

This inequality is the *maximum posteriori hypothesis* and it gives us the most likely class to each sample. The previously described idea is the base of the widely used *naïve Bayesian* classifier [32]. However, this idea is well applicable not only to classification but also to selection. Moreover, the naïve Bayesian classifier is faster than other complex algorithms such as artificial neural networks or support vector machines. Thus the computational time of such a wrapper algorithm is acceptable when the number of features is relatively small. This is the case in the HAR problem where the number of features is less than 100.

The weakness of lots of feature selection algorithms is the assumption that features are independent, which is not true in most cases. However, this weakness has been avoided in the proposed feature selection algorithm which can be seen on Fig. 1. It uses *forward selection* strategy with the following two stop conditions: $n \le i$ and $accuracy_{i+1} \le accuracy_i$, where i indicates iterations and *n* is the number of predetermined features. The algorithm requires the *n* and a training set (**F**) and returns the indexes of selected features (**Sf**). For simplicity, suppose that **F** is a three dimensional data block (features, samples, classes) where the colon refers to each element of a dimension (as in Matlab). At first, our proposed algorithm normalizes the training set. It means that, the features are divided by the highest feature value plus one so the normalized values will be in the [-1, 1) interval. The second step consists in calculating the probabilities $P(C_i)$, given by the number of samples in class *Ci* divided by the total number of samples. Thereafter, we have to determine the class-conditional probabilities from the whole training set. It results a *Pxc* three dimensional matrix (features, bins, classes) which indicates how likely a feature in the *i'th* class belongs into the *j'th* bin. Fig. 2 shows a pseudocode to the *Pxc* matrix calculation. It divides [-1, 1) interval into bins and counts how many times a feature belongs into a bin. Finally, it turns count into probability in the last loop. In the following step, the algorithm determines a score to each feature by the *bayesianScore* function and stores the index of the selected in **Sf** which has the greatest score. The pseudocode to *bayesianScore* can be seen on Fig. 3. It determines a bin for each feature in a sample and selects that class which satisfies inequality (5) according to (4). The *calculate_rate* function at the end of the pseudocode compares the predicted and correct class labels and returns the recognition rate in percent. The remaining part of the feature selection algorithm selects following features while the stop condition does not met. In the embedded loops, each unselected feature will be concatenated with the selected feature set one by one and that feature will be added to the set which maximize the summarized score of selected features.

IV. EFFICIENCY INVESTIGATION

The appearance of data mining was a milestone of modern biomedical applications. An interesting and rapidly expanding part of this area is the human activity recognition. In this type of problem, we want to determine the activity of people from the information which comes from one or more accelerometer-based data collector devices. Generally, the sensors are placed to different parts of the body and provide information about the functional ability and lifestyle of an observed person [34]. However, it is not clear yet, which feature combination is the most appropriate for classification. In order to create a comprehensive comparison between feature selection methods, we have to collect all of the well-known methods from the literature. Fortunately, Zhao et al. created a generally applicable repository to feature selection research [33]. It was designed to collect the most common algorithms that have been developed in this topic. The repository suggests implementations (in Matlab) of the algorithms which were used in this investigation. For efficiency analysis some feed-forward artificial neural networks (ANNs) were used as classifier because previous studies have shown that ANN is well applicable in the case of HAR [6, 35]. Finding the right architecture of an ANN for a specific purpose is a time-consuming task because it requires lots of simulations. Our ANN architecture design is based on the work of Oniga and Suto [35], where the authors demonstrated that a simple feed-forward ANN with only one hidden and one output layers is enough to HAR. Therefore, in this study, some feed-forward ANNs were generated with the same architecture as in their work. Thus in the used networks, the number of neurons on the input, hidden and output layers are equal to *n*, *3n*, *number of classes*, respectively.

```
Require: n, F
Ensure: Sf
  F ← normalize(F)
  Pc ← calculatePc(F)
  Pxc ← calculatePxc(F)
  for f : 1 to features do
    Fs(f) ← bayesScore(F(f,:,:), Pxc(f,:,:), Pc)
  end for
  Sf(1) ← maxIndex(Fs)
  for i : 1 to n do
    for f : 1 to features do
      if f was not selected yet then
        for sf : 1 to i do
          F_temp(sf,:,:) ← F(Sf(sf),:,:)
          Pxc_temp(sf,:,:) ← Pxc(Sf(sf),:,:)
        end for
        F_temp(i+1,:,:) ← F(f,:,:)
        Pxc_temp(i+1,:,:) ← Pxc(f,:,:)
        Fs(f) ← bayesScore(F_temp, Pxc_temp, Pc)
      else
        Fs(f) ← 0
      end if
    end for
    findex ← maxIndex(Fs)
    recognition_rate ← max(Fs)
    Sf(i+1) ← findex
    if accurecy_{i+1} ≤ accurecy_i then
      break loop
    end if
  end for
```

Fig. 1. The proposed feature selection algorithm

```
Require: F
Ensure: Pxc
  dx ← 2/bins
  for c : 1 to classes do
    for f : 1 to features do
      for s : 1 to samples do
        for b : 0 to bins - 1 do
          if (-1 + b*dx) ≤ F(f,s,c) and (-1 + (b+1)*dx) >
          F(f,s,c) then
            Pxc(f,b+1,c) ← Pxc(f,b+1,c) + 1
          end if
        end for
      end for
      for b : 1 to bins do
        Pxc(f,b,c) ← Pxc(f,b,c)/(features * samples)
      end for
    end for
  end for
```

Fig. 2. Pxc matrix calculation

```
Require: F, Pc, Pxc
Ensure: recognition_rate
  for c : 1 to classes do
    for s : 1 to samples do
      for f : 1 to features do
        for b : 0 to bins - 1 do
          if (-1 + b*dx) ≤ F(f,s,c) and (-1 + (b+1)*dx) >
          F(f,s,c) then
            bin(f) ← b + 1
          end if
        end for
      end for
      for c1 : 1 to classes do
        Pcx(c1) ← 1
        for f : 1 to features do
          Pcx(c1) ← Pcx(c1) * Pxc(f, bin(f), c1)
        end for
        Pcx(c1) ← Pcx(c1) * Pc(c1)
      end for
      cindex ← maxIndex(Pcx)
      predicted_class(c, s) ← cindex
    end for
  end for
  recognition_rate ← calculate_rate(predicted_class)
```

Fig. 3. The bayesScore function

The aim of the test was to determine the most appropriate feature combination which gives the best recognition rate with the fewest features. In our experiments we used a public data source which was collected at University of California, Berkeley. The data was acquired from 20 volunteers who executed more than ten activities [36]. Our tests are focused on eight main activities like standing, sitting, laying, walking, jogging, jumping and climbing stairs (up, down). For comparison two subjects were selected from the database, the youngest and oldest persons.

At first, the raw data was split into windows and the features were extracted from all of them. The most important features from the literature were collected into Table I [37, 38]. Most of them have been obtained from the time and frequency domains. Wavelet features were omitted because time-frequency features are more useful, see [16]. One part of the features is one dimensional while the other part is multidimensional according to the used three dimensional accelerometer sensor (*x*, *y*, *z*). Therefore, three dimensional features were partitioned into one dimensional. Thus, 50 one dimensional features were generated. Thereafter, the features were determined by different feature selection methods and they were the input of the networks. The most relevant feature selection methods are the following,

- Correlation Feature Selection (CFS)
- Chi Square
- Fast Correlation-Based Filter (FCBF)
- Fisher Score
- Information Gain
- Kruskal-Wallis
- Minimum Redundancy Maximum Relevance (mRMR)
- T-test

During the test, the number of features was increased by one from 1 to 6 in every step and the performance of ANN was recorded. Selected features can be seen in Table II and Table III which was sorted in selection order. According to the number of features and observed subjects, Table IV and Table V contain the measured recognition rates in percentage. The elements within the tables are the best recognition rates of 30 training processes and the values in bold in the last two tables indicate the best recognition rates provided by the methods.

TABLE I. USED FEATURES

Category	Feature	Abbreviation
Time	Mean	M
	Variance	V
	Mean absolute deviation	MAD
	Root mean square	RMS
	Zero crossing rate	ZCR
	Interquartile range	IQR
	75'th percentile	PE
	Kurtosis	KS
	Signal magnitude area	SMA
	Min-max	MM
Frequency	Spectral energy	SE
	Spectral entropy	E
	Spectral centroid	SC
	Principal frequency	PF
Other	Correlation between axis	CORR
	First autoregressive coefficient	AR1
	Second autoregressive coefficient	AR2
	Tilt angle	TA

TABLE II. SELECTED FEATURES FOR SUBJECT 1

Method	Features
Wrapper	RMS$_y$, MM$_x$, CORR$_x$, PE$_z$, SMA, KS$_z$
CFS	M$_x$, M$_z$, V$_x$, RMS$_x$, RMS$_y$, RMS$_z$
Chi Square	SMA, PE$_y$, RMS$_y$, RMS$_x$, PE$_z$, PE$_x$
FCBF	PE$_x$, PE$_y$, M$_z$, M$_x$, PE$_z$, V$_y$
Fisher score	SE$_z$, TA, RMS$_x$, E$_y$, M$_x$, MM$_x$
Information gain	KS$_z$, AR1$_y$, AR2$_z$, AR1$_z$, KS$_y$, CORR$_y$
Kruskal-Wallis	KS$_x$, KS$_z$, KS$_y$, PE$_z$, M$_x$, M$_y$
mRMR	SMA, TA, AR2$_x$, AR1$_x$, AR2$_y$, PF$_z$
T-test	M$_x$, TA, SE$_z$, SE$_x$, RMS$_z$, RMS$_x$

TABLE III. SELECTED FEATURES FOR SUBJECT 2

Method	Features
Wrapper	RMS$_y$, M$_y$, PE$_x$, MM$_x$, KS$_y$, M$_x$
CFS	M$_x$, M$_z$, RMS$_x$, RMS$_y$, RMS$_z$, IQR$_y$
Chi Square	PE$_x$, PE$_y$, SE$_x$, RMS$_y$, RMS$_x$, M$_x$
FCBF	RMS$_y$, PE$_x$, M$_y$, PE$_y$, M$_x$, RMS$_z$
Fisher score	SE$_z$, RMS$_x$, M$_x$, TA, PE$_x$, M$_z$

Information gain	PE_y, PE_x, SE_x, RMS_x, RMS_y, M_x
Kruskal-Wallis	M_y, KS_z, KS_x, TA, $CORR_x$, M_x
mRMR	PE_y, TA, PF_z, PF_y, PF_x, $AR2_z$
T-test	RMS_x, SE_x, PE_x, M_x, E_y, $CORR_y$

TABLE IV. MEASURED RECOGNITION RATES FOR SUBJECT 1

Method	**n = 1**	**n = 2**	**n = 3**	**n = 4**	**n = 5**	**n = 6**
Wrapper	**50.31**	**81.77**	**88.36**	**94.11**	**96.53**	**98.61**
CFS	50.00	76.38	88.35	90.10	95.14	98.13
Chi Square	50.00	81.42	59.91	60.07	92.82	94.79
FCBF	49.30	78.29	86.45	92.92	95.31	98.01
Fisher score	42.70	50.00	70.83	72.91	76.90	84.21
Information gain	27.25	38.71	49.33	74.13	83.16	90.45
Kruskal-Wallis	43.92	53.29	63.19	87.67	92.53	95.48
mRMR	50.00	79.51	86.45	93.05	95.65	97.22
T-test	50.00	80.83	38.67	49.69	45.03	53.87

TABLE V. MEASURED RECOGNITION RATES FOR SUBJECT 2

Method	**n = 1**	**n = 2**	**n = 3**	**n = 4**	**n = 5**	**n = 6**
Wrapper	**52.26**	**86.11**	**96.35**	**98.95**	**99.65**	**100**
CFS	52.24	75.34	90.97	98.43	99.30	99.65
Chi Square	47.04	82.98	75.69	77.77	68.40	64.53
FCBF	**52.26**	77.43	**96.35**	98.91	99.47	99.68
Fisher score	49.47	63.89	73.09	78.29	83.51	83.33
Information gain	52.23	84.20	84.95	86.28	88.02	89.41
Kruskal-Wallis	50.00	72.91	83.68	92.71	97.39	99.13
mRMR	52.23	82.63	90.27	94.61	97.04	97.91
T-test	48.26	44.09	41.32	50.52	51.90	43.40

V. DISCUSSION

The efficiency examination in the previous section demonstrates some interesting results. Firstly, from Table II and Table III, we can observe that there are not two identical feature sets. It clearly indicates the difference between algorithms. In addition, the selected features for subject 1 and subject 2 are just partially overlapping. This result can be explained by the variation of movement patterns according to the age. Secondly, the selected features by the proposed wrapper method provides better recognition rates in all cases as Table IV and Table V show. In addition, only the wrapper algorithm could reach 100% recognition rate for subject 2. We can notice some other advantages of the wrapper method. One of theme is the linearity. In both cases (subject 1 and 2), the improvements are continuous and linear in contrast to Chi Square, Fisher score and T-test. Moreover, an important advantage of the wrapper method is the concrete feature amount determination. It means that, the method gives back as much features as necessary. Finally, the investigation demonstrated that the proposed method is efficient not only with the naïve Bayesian classifier but also with other types of machine learning algorithms (in this case, with artificial neural network).

VI. CONCLUSION

In this paper we compared a naïve Bayesian based wrapper and more well-known filter feature selection algorithms on the HAR machine learning problem. The paper contains the algorithm of the wrapper method and gives a detailed description about it. In order to assess the reliability of the comparison, public and independent data and algorithm sources were used. WARD 1.0 was the data source while the applied feature selection algorithms were derived from an open source repository. The WARD database and the feature selection repository were created in University of California at Berkeley and Arizona State University, respectively. At the beginning of investigation, the most important features in HAR were listed. Features were extracted from two data blocks which were acquired from the youngest and oldest participants of WARD. The algorithms selected 6 features and they were the input of neural networks in increasing order. The efficiency of an algorithm depended on the recognition rate of the networks and it were stored in Table IV and V. As the experimental results indicate, in this type of problem, the proposed wrapper method outperforms each filter algorithms in the case of both subjects. In addition, the results clearly show that the selected feature set by the wrapper method is efficiently usable with other types of machine learning algorithms.

REFERENCES

[1] G. C. Cawley, N. L. C. Talbot, “Gene selection in cancer classification using sparse logistic regression with bayesian regularization”, Bioinformatics, vol. 22(19), pp. 2384-2355, 2006.

[2] C. P. Lee, Y. Leu, “A novel hybrid feature selection method for microarray data analysis”, Applied Soft Computing, vol. 11(1), pp. 208-2013, 2011.

[3] T. Parag, A. Elgammal, A. Mittal, “A framework for feature selection for background subtraction”, Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, USA, pp. 1916-1923, 2006.

[4] K. Bashir, T. Xiang, S. Gong, “Feature selection on gait energy image for human identification”, Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, USA, pp. 985-988, 2008.

[5] N. Bankman, Handbook of Medical Image Processing and Analysis, Elsevier, USA, 2009, pp. 50-120.

[6] S. Oniga, J. Suto, “Human activity recognition using neural networks”, Proceedings of 15th Carpathian Control Conference, Velke Karlovice, Czech Republic, pp. 403-406, 2014.

[7] A. Abbasi, H. Chen, A. Salem, “Sentiment analysis in multiple languages: feature selection for opinion classification in web forums”, ACM Transactions on Information Systems, vol. 26(3), pp. 1-34, 2008.

[8] U. Stanczyk, L. C. Jain, Feature Selection for Data and Pattern Recognition, Springer, USA, 2015, pp. 1-76.

[9] C. H. Cheng, P. S. P. Wang, Handbook of Pattern Recognition and Computer Vision, 3th ed., World Scientific, Singapore, 2005, pp. 40-180.

[10] S. T. Bow, Pattern Recognition and Image Preprocessing, 2th ed., Marcel Dekkere, USA, 2002, pp. 90-130.

[11] N. M. Murty, D. V. S. Devi, Introduction to Pattern Recognition and Machine Learning, World Scientific, Singapore, 2015, pp. 110-150.

[12] A. R. Webb, Statistical Pattern Recognition, 2th ed., John Wiley & Sons, UK, 2002, pp. 73-126.

[13] I. Guyon, A. Elisseeff, “An introduction to variable and feature selection”, Journal of Machine Learning Research, vol. 3(1), pp. 1157-1182, 2003.

[14] H. Liu, H. Motoda, R. Setiono, Z. Zhao, “Feature selection: an ever evolving frontier in data mining”, Proceedings of the 4th Workshop on Feature Selection in Data Mining, Hyderabad, India, pp. 4-13, 2010.

[15] D. O. Lara, M. A. Labrador, “A survey on human activity recognition using wearable sensors”, IEEE Communication Survey & Tutorials, vol. 15(1), pp. 1192-1209, 2013

[16] J. S. Preece, J. Y. Goulermas, L. P. J. Kenney, D. Howard, “A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data”, IEEE Transactions on Biomedical Engineering, vol. 56(3), pp. 871-879, 2009.

[17] C. Bea, W. C. Yeh, Y. Y. Chung, S. L. Liu, “Feature selection with intelligent dynamic swarm and rough set”, Expert Systems with Applications, vol. 37(1), pp. 7026-7032, 2010.

[18] T. M. Cover, J. A. Thomas, Elements of Information Theory, 2th ed., John Wiley & Sons, USA, 2006, pp. 26-69.

[19] G. C. Cawley, N. L. C. Talbot, M. Girolami, “Sparse multinomial logistic regression via bayesian L1 regularisation”, Proceedings of the 23th Conference on Advances in Neural Information Processing Systems, Montreal, Canada, pp. 209-216, 2007.

[20] X. He, D. Cai, P. Niyogi, “Laplacian score for feature selection”, Proceedings of the 18th Conference on Advances in Neural Information Processing Systems, Montreal, Canada, pp. 507-514, 2005.

[21] H. Peng, F. Long, C. Ding, “Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27(8), pp. 1226-1238, 2005.

[22] P. A. Estevez, M. Tesmer, C. A. Perez, J. M. Zurada, “Normalized mutual information feature selection”, IEEE Transactions on Neural Networks, vol. 20(2), pp. 189-201, 2009.

[23] H. Liu, H. Motoda, Computational Methods of Feature Selection, Taylor & Francis Group, USA, 2008, pp. 85-134.

[24] B. San, M. Peker, A. Cavusoglu, F. W. Celebi, “A conparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms”, Journal of Medical Systems, vol. 38(18), pp. 2507-2517, 2014.

[25] Y. Saeys, I. Inza, P. Larranaga, “A review of feature selection techniques in bioinformatics”, Bioinformatics, vol. 23(19), pp. 2507-2517, 2009.

[26] J. V. Hulse, T. M. Khoshgoftaar, A. Napolitano, R. Wald, “Feature selection with high-dimensional imbalanced data”, Proceedings of the IEEE International Conference on Data Mining Workshops, Miami, USA, pp. 507-514, 2009.

[27] A. G. Karegowda, A. S. Majunath, M. A. Jayaram, “Comparative study of attribute selection using gain ration and correlation based feature selection, International Journal of Information Technology and Knowledge Management, vol. 2(2), pp. 271-277, 2010.

[28] Y. Yang, J. O. Pedersen, “A comparative study on feature selection in text categorization”, Proceedings of the 14thInternational Conference on Machine Learning, Nashville, USA, pp. 412-420, 1997.

[29] Z. Zhao, H. Liu, “Spectral feature selection for supervised and unsupervised learning”, Proceedings of the 24th International Conference on Machine Learning, Corvallis, USA, pp. 1151-1157, 2007.

[30] M. A. Hall, M. A. Smith, “Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper”, Proceedings of the Florida Artificial Intelligence Symposium, Florida, USA, pp. 235-239, 1999.

[31] M. A. Jayaram, A. G. Karegowda, A. S. Manjunath, “Feature subset selection problem using wrapper approach in supervised learning”, International Journal of Computer Applications, vol. 1(7), pp. 13-17, 2010.

[32] S. Marsland, Machine Learning - An Algorithmic Perspective, 2th ed., CRC Press, USA, 2015, pp. 27-31.

[33] Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, H. Liu, Advancing Feature Selection Research – ASU Feature Selection Repository, Technical Report, Arizona State University, 2010.

[34] J. Suto, S. Oniga, A. Buchman, “Real time human activity monitoring, Annales Mathematicae et Informaticae, vol. 44(1), pp. 187-196, 2015.

[35] S. Oniga, J. Suto, “Activity recognition in adaptive assistive systems using artificial neural networks”, Elektronika ir Elektrotechnika, vol. 22(1), pp. 68-72, 2016.

[36] A. Y. Yang, L. Jafari, S. S. Sastry, R. Bajcsy, “Distributed recognition of human actions using wearable motion sensor networks”, Journal of Ambient Intelligence and Smart Environment, vol. 1(1), pp. 1-5, 2009.

[37] M. Shoaib, S. Bosch, O. D. Incel, H. Scholten, P. J. M. Havinga, “A survey of online activity recognition using mobile phones”, Sensors, vol. 15(1), pp. 2059-2085, 2015.

[38] M. Shoaib, S. Bosch, O. D. Incel, H. Scholten, P. J. M. Havinga, “A survey of online activity recognition using mobile phones”, Sensors, vol. 15(1), pp. 2059-2085, 2015.

[39] M. Young, The Technical Writer’s Handbook, Mill Valley, CA: University Science, 1989, pp. 76-142.

[40] R. Kohavi, G. H. John, “Wrappers for feature subset selection”, Artificial Intelligence, vol. 97(1), pp. 273-324, 1997.