Proceedings of the Federated Conference on Computer Science and Information Systems pp. 281–285
DOI: 10.15439/2016F244
ACSIS, Vol. 8. ISSN 2300-5963

The new method of the selection of features for the k-NN classifier in the arteriovenous fistula state estimation

Marcin Grochowina
University of Rzeszów
al. Rejtana 16, 35-310 Rzeszów, Poland
Email: gromar@ur.edu.pl

Lucyna Leniowska
University of Rzeszów
al. Rejtana 16, 35-310 Rzeszów, Poland
Email: lleniow@ur.edu.pl

***Abstract*—In this paper the application of a new method of features selection was presented. Its effects were compared with several other methods of features selection. The study were performed using a data set containing samples of the sound signal emitted by the arteriovenous fistula. The aim was to create a solution with multiclass classification based on the k-NN classifier family allowing for effective and credible assessment of the state of arterial-venous fistula.**

I. Introduction

EACH classification process is based on the set of features delivered to the classifier on the basis of which a decision is taken and the result obtained. Proper selection of a set of features significantly improves the quality of the classification process.

The approach ensuring the best quality is to test all possible non-empty subsets of the input set. Unfortunately, the number of non-empty subsets of n-element is $2^n - 1$, which implies the possibility of a full review of the subsets only for the n that does not exceed a dozen elements. Full analysis for larger values of n is too time-consuming. It is therefore the use of quasi-optimal methods, which is determinative of a subset of the features of possible high efficiency of classification.

K-NN is the most popular minimum distance classifier. It assigns the unknown sample to the class most often representing its neighborhood[14][15]. There are many variants of this method. They differ among themselves, inter alia, by methods of calculating the distance and the method of voting that determines the result.

The most common variation of k-NN is weighted k-NN , in which weight of the neighbor of samples $\boldsymbol{x}$ depends on its distance from the $\boldsymbol{x}$[13]. An interesting solution is the Diplomatic Nearest Neighbors (k-DN) [12], which seeks k neighbors of each class separately, and then selects the class for which the average distance from the found neighbors to the tested sample is the smallest.

Due to its flexibility, simplicity and the possibility of use in tasks of classification and regression k-NN is popular despite its flaws: it requires storage in the memory the whole training set and high demand for computing power, especially for large training sets.

II. Data set

In the studies the data set consisting of sounds emitted by the arteriovenous fistula was used. The studies to date [1][2] show that the character of the sound emitted by the blood flowing within the fistula differs depending on the condition of the fistula.

The research data set was collected from 19 patients with radiocephalic fistula. Aquisition of the material consisted in recording the sound of the blood flowing through the arteriovenous fistula. Material was collected using a dedicated head equipped with an electret microphone CZ034 manufactured by Ringford, with a sensitivity of -42dB (0dB=1V/Pa, 1kHz), ie. 8mV/Pa and an interval signal/noise ratio greater than 60dB. To register a signal, an integrated sound card was used as part of the RV730 Radeon 4000 manufactured by AMD as well as dedicated software running under the Linux operating system. Sampling frequency was set at 8 kHz.

Numerical processing of data was performed using WEKA 3.7.13 package running with the JRE Oracle Java 1.8. The calculations were performed on a computer with Intel Core 2 T6570 2.1GHz under the Linux operating system. During the measurements the algorithms time requirement only a single core processor was used.

Fistulas were rated as effective, however, to differing degrees. Eight groups representing a fistula with varying degrees of stenosis were extracted. A total of 1190 samples was collected.

The groups were lettered with labels ***a-h*** , wherein the group ***a*** were fistulas in the best condition and in the group ***h*** in the worst condition. With the collected data set 23 features were extracted; 6 in the time and 17 in the frequency domain. Features in the time domain named t0, t4, y0, y4, p0 and p4 describe the timing, amplitude and shape of the signal envelope within a single period of the rhythm of the heart. Features in the frequency domain named f1-f17 describe the density of the frequency spectrum of the recorded signal at specific intervals from the scope of 20-600Hz.

978-83-60810-90-3/$25.00

III. METHODS

In this study five methods of feature selection were tested. Each of them belongs to a different category of methods (Figure 1).

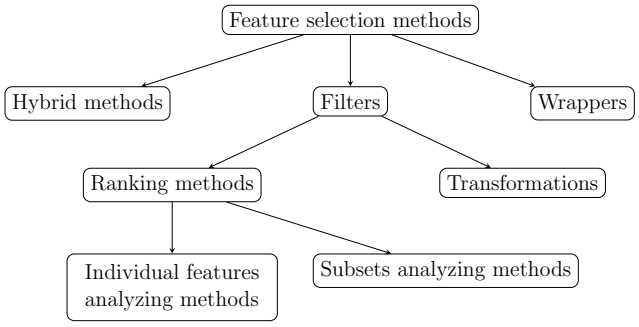


Fig. 1: Feature selection methods

The first four are commonly known and available in the WEKA package. The fifth is an own method developed proprietarily to the needs of this particular task.

The methods used are:

- *Correlation* - builds ranking of features evaluating the characteristics of each of them individually. The rate criterion is the absolute value of the correlation of coefficient feature with the class. The higher the correlation, the higher the position of feature in the ranking.
- *SVMeval* - evaluates the worth of an attribute by using an SVM classifier. Attributes are ranked by the square of the weight assigned by the linear SVM classifier.
- *PCA* - performs a linear transformation of the features into another space in which features included in the new set are mutually uncorrelated and sorted with respect to the amount of input information in the classification process.
- *Forward search* - wrapper method building the set of features starting from one and gradually adding these features that provide the best quality of classification. This method is based on a classifier to be used in the target solution - in this case the k-NN.
- *Joined pairs* - a method developed by the author. It creates a ranking of features based on the ability of pairs of features for classification. In the first stage, a collection of all possible two-element subsets of features is formed. Then, basing on each subset of features a classifier is constructed and evaluated. As a result, each two-element subset is assigned a numerical value that indicates the quality of classifier built on the basis of this subset. Finally, the ranking of features is created. Features are added into in order indicated by quality of classifiers built in the previous step. The principle of operation of the method is shown in algorithm1.

The method has been tested using four selected data sets available from the UCI Machine Learning Repository[16] – glass, vote, segment challenge and wine quality. In each of the cases a rapid convergence of the level of quality classifications

Algorithm 1: Joined pairs

```
input : tf: table of features
output: fr: features ranking
// variable: pair of features
def pof: structure:
    featureA
    featureB
    quality
for each possible pairs of features from tf do
    add new pair to pof
    pof.quality ←
     classifierQuality(pof.featureA,
     pof.featureB)
Sort(pof) by pof.quality, ascending
for each pof do
    if pof.featureA ∉ fr then
        add pof.featureA to fr
    if pof.featureB ∉ fr then
        add pof.featureB to fr
return fr
```

to the maximum value was obtained, indicating that the joined pairs method works properly Figure 2 shows the graphs indicating the level of quality of classification described by the F-measure as a function of features number taken into account during the classification process. Number of features included was increased by adding features one by one, in the order indicated by the ranking produced by joined pairs algorithm.

In the study, k-NN classifier with distance weighing was used. For the distance measure the Manhattan metric was used:

$$d(X,Y) = \sum_{i=1}^{N} |X_i - Y_i|, \tag{1}$$

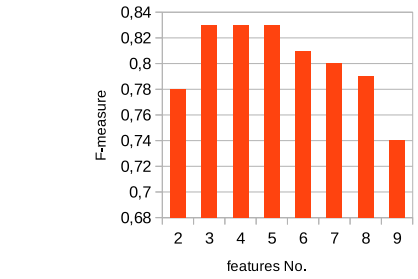
where X and Y are the points in N-dimensional space of features and d is a distance betwen these points. The tested element was assigned to a class on the basis of the vote. The weight of the vote of the $i-th$ neighbor was distance weighed according to the formula:

$$w(i) = \frac{1}{d(i) + 0.0001}. \tag{2}$$

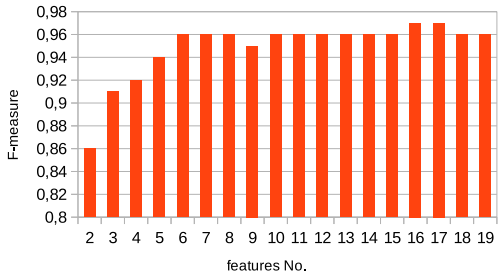
Value of 0.0001 in the denominator is added to the distance in order to avoid division by zero when the distance is equal to zero[6].

Quality rating of classification was based on the F-measure[1] indicator. The indicator can be betwen 0 and 1 and the quality of classification is the higher the F-measure value is closer to 1. The test method was 10-fold cross-validation.

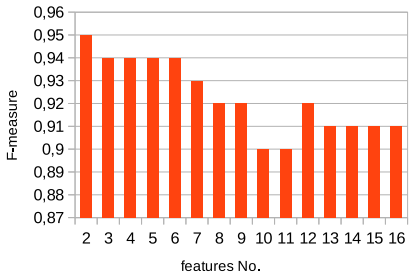
[1]F-score, F1-score



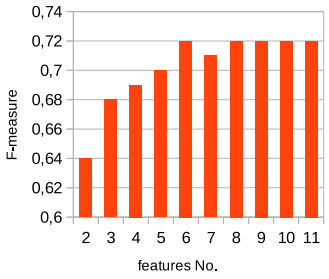
(a) glass



(b) segment challenge



(c) votes



(d) wine quality

Fig. 2: The quality of the classification as a function of the number of features for the selected data sets

IV. Results and discussion

The quality of feature sets obtained using each method was evaluated by construction of the k-NN classifier and assessment of its quality. For each set of the features, 23 subsets of features were generated, containing from 1 to 23 features. In the next subsets the features were included in the order indicated by the ranking. For each subset of features, 15 classifiers differ by a n parameter were generated. Parameter n was varied from 1 to 15. Summary of rankings of features for each method are schown in table.I.

TABLE I: Features ranking

no.	Correlation	SVMeval	PCA	Forward search	Joined pairs
1	f14	f11	v1	f3	-
2	f15	f5	v2	f13	f3,f13
3	f8	f14	v3	f11	f11
4	f16	f16	v4	y4	f1
5	f7	f13	v5	f10	f12
6	f6	f9	v6	f4	f4
7	f13	f15	v7	f1	f9
8	f9	f3	v8	f14	f2
9	f5	f8	v9	f15	f7
10	f4	f12	v10	f16	f5
11	f10	f6	v11	f9	f10
12	f12	f7	v12	f8	f8
13	f11	f10	v13	t4	y4
14	f3	f1	v14	f7	t4
15	f1	f2	v15	f12	f14
16	f2	f4	v16	fm	f15
17	fm	fm	v17	f2	fm
18	t1	y4	v18	f5	y0
19	y0	t4	v19	t1	t1
20	p4	y0	v20	p4	f16
21	t4	t1	v21	y0	f6
22	y4	p1	v22	f6	p1
23	p1	p4	v23	p1	p4

Graphical comparison of results of calculations for the classification was presented in figure 3.

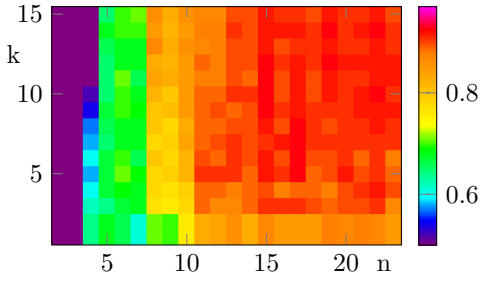
The worst result was achieved by the correlation method with its F-measure not exceeding 0.93. Not much better were SVMeval and PCA methods for which F-measure reached a value of 0.94. All the above methods have achieved the maximum quality for $n \geq 15$.

The best was the Forward search method, which reached a maximum value of F-measure equal to 0.97 for $n = 9$. In addition, a large area, stretching from $n \in \langle 8 - 18 \rangle$ and $k \in \langle 5 - 15 \rangle$, for which $F - measure \geq 0.95$ provides a good stability of the solution. Comparable in quality but far superior in the minimum amount of features was Joined pairs method. The maximum value specified by $F - measure = 0.96$ was achieved for $n = 6$ and $k = 12$.

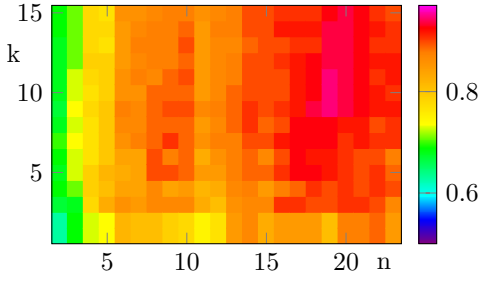
A tabular summary of the F-measure for selected values of k was presented in Table.II.

The chart shows that the *Joined Pairs* method attains the best F-measure using the smallest set of features. However, an increase in the feature count causes quality loss, which is regained only for $n = 15$ and $n = 16$. The Forward Search method achieved a stable maximum for n=9. Other schemes generated feature sets that were best for high values of n, yet none reached the quality level of *Joined Pairs* or *Forward Search*.

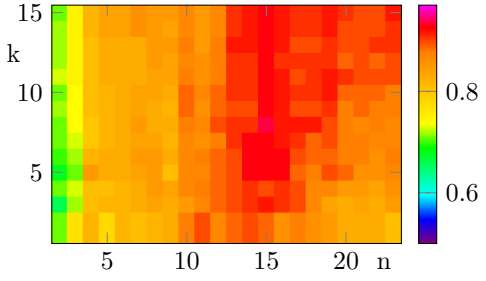
The *PCA* method allows the use of non-empirical methods for selecting the amount of features (eg. the igenvalues criterion), therefore evaluation time assumed zero. Evaluation time for *Forward search* method is zero because the evaluation of set is made up to date during the construction of the rankings.



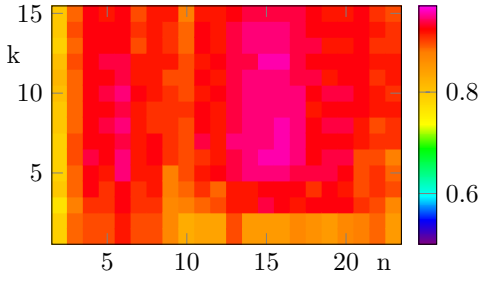
(a) Correlation



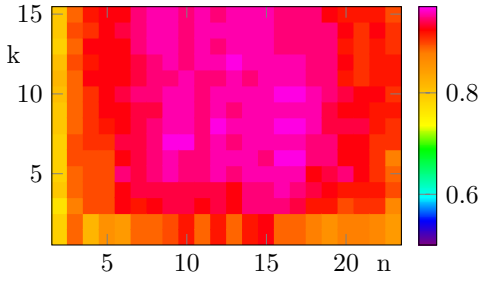
(b) SVMeval



(c) PCA



(d) Joined pairs



(e) Forward search

Fig. 3: F-measure as a function of k and n

TABLE II: F-measure

k=	12	10	6	7	12
n	Correlation	SVMeval	PCA	Forward search	Joined pairs
2	0,14	0,68	0,71	0,79	0,81
3	0,39	0,73	0,74	0,9	0,9
4	0,49	0,77	0,8	0,91	0,93
5	0,65	0,79	0,82	0,91	0,94
6	0,68	0,86	0,83	0,94	0,96
7	0,68	0,86	0,85	0,94	0,92
8	0,83	0,89	0,83	0,95	0,92
9	0,82	0,89	0,83	0,97	0,92
10	0,85	0,89	0,88	0,97	0,9
11	0,89	0,85	0,87	0,95	0,92
12	0,88	0,87	0,9	0,96	0,93
13	0,9	0,89	0,91	0,95	0,94
14	0,9	0,9	0,91	0,96	0,94
15	0,93	0,9	0,94	0,96	0,96
16	0,91	0,9	0,92	0,96	0,96
17	0,92	0,92	0,92	0,96	0,94
18	0,91	0,93	0,92	0,95	0,93
19	0,93	0,95	0,9	0,95	0,93
20	0,92	0,94	0,88	0,93	0,93
21	0,92	0,93	0,87	0,93	0,93
22	0,92	0,92	0,88	0,91	0,92
23	0,92	0,91	0,87	0,91	0,92

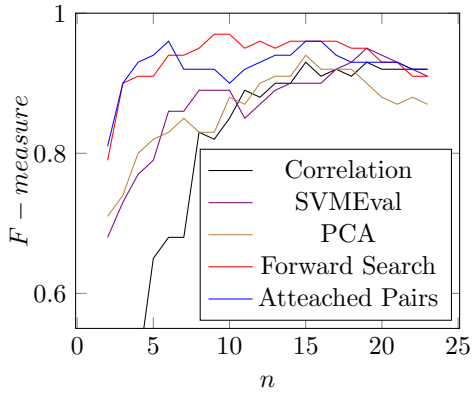


Fig. 4: F-measure as a function of n for optimal k

Table III summarizes the results for all the feature selection methods. Its first part presents the time requirements for each schema, including the time needed to generate the ranking and the time necessary to assess the quality of respective sets of features. The filtering methods (*Correlation* and *SVMeval*), unsurprisingly used the smallest amount of time. Similarly, the time requirements of the *PCA* method were negligible. The running time of both *Forward Search* and *Joined Pairs* was significantly slower. The *Forward Search* method used 276 classifiers: 23 one-feature classifiers, 22 two-feature classifiers, 21 classifiers that used three features, and so on. This was the main source of higher time requirements. The *Joined Pairs* method analyzed all the pairs of features and thus built in total 253 classifiers. Due to lower complexity of the classifiers, this approach needed less time that the *Forward Search* method. The quality of the ranking for *Correlation*, *SVMeval* and *Joined Pairs* methods was based on generation and quality assessment for all the feature sets for n from 2 to 23 and each k from 1 to 15. Since the only variable components of

TABLE III: Summary of results

	Correlation	SVMeval	PCA	Forward search	Joined pairs
rankings construction time	00:00:01	00:00:10	00:00:01	02:48:35	01:40:27
sets evaluation time	00:17:03	00:17:03	0	0	00:17:03
total time	00:17:04	00:17:13	00:00:01	02:48:35	01:57:30
optimal n	15	19	15	9	6
optimal k	12	10	8	7	12
F-measure	0,93	0,95	0,94	0,97	0,96

the ranking process were the fearure sets, the running time was the same for all of them. The *PCA* method allowed for non-empirical ways of choosing the size of feature sets (for example, the Kaiser criterion or scree plots).

As this algorithms are not computationally-heavy, their time-requirements were assumed to be zero. The running time of quality assessment for the *Forward Search* method was assumed to be zero as well, because the method does the necessary calculations online, while generating the ranking. The second part of Table 4 presents the optimal values for k and n with the respective F-measure.

All methods of feature selection achieved similar quality indicators of the constructed models.

V. Conclusion

It is possible to notice the general principle that computing power consumption feature selection algorithm translates into the quality of the obtained subsets of features. Undemanding methods of filter group indicated subsets of more features than other methods. The *Joined pairs* algorithm gives good results in the classification task.

With respect to the problem of evaluation of the arteriovenous fistula it can be concluded that the results are very good. Each of these methods has allowed to obtain a very high quality classification. It is suspected that, such optimistic results may be the effect of insufficient amount of analyzed data. Vectors describing individual patients form in the a feature space the easily separated clusters.

Verification of the results should be made on unrelated set of test data and having regard to a greater number of patients and samples.

Therefore, it would be appropriate to extend the scope of the study, increasing the set of input data.

References

[1] Marcin Grochowina, Lucyna Leniowska and Piotr Dulkiewicz, "Application of Artificial Neural Networks for the Diagnosis of the Condition of the Arterio-venous Fistula on the Basis of Acoustic Signals," *Brain Informatics and Health,* Springer, 2014, pp. 400–411.

[2] Marcin Grochowina and Lucyna Leniowska, "Comparison of SVM and k-NN classifiers in the estimation of the state of the arteriovenous fistula problem," *Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on,* IEEE, 2015, pp. 249–254.

[3] Zbigniew Suraj, Neamat El Gayar and Pawel Delimata, "A rough set approach to multiple classifier systems," *Fundamenta Informaticae,* IOS Press, 2006, pp. 393–406.

[4] Mikkel Grama , Jens Tranholm Olesena , Hans Christian Riisa , Maiuri Selvaratnama and Michalina Urbaniaka, "Stenosis detection algorithm for screening of arteriovenous fistulae," *15th Nordic-Baltic Conference on Biomedical Engineering and Medical Physics (NBC 2011),* Springer, 2011, pp. 241–244.

[5] Fan, Rong-En and Chen, Pai-Hsuen and Lin, Chih-Jen, "Working set selection using second order information for training support vector machines," *The Journal of Machine Learning Research vol.6,* JMLR. org, 2005, pp. 1989–1918.

[6] "WEKA documentation," http://www.cs.waikato.ac.nz/ml/weka/documentation.html

[7] Remco R. Bouckaert, Eibe Frank, Mark Hall, Richard Kirkby, Peter Reutemann, Alex Seewald, David Scuse, "WEKA Manual," University of Waikato, 2013.

[8] Tadeusz Morzy, "Eksploracja danych - metody i algorytmy," PWN, 2013.

[9] Dymitr Ruta "Robust Method of Sparse Feature Selection for Multi-Label Classification with Naive Bayes," Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on, IEEE, 2014, pp. 375–380. http://dx.doi.org/10.15439/2014F502

[10] Zdravevski, Eftim and Lameski, Petre and Kulakov, Andrea and Gjorgjevikj, Dejan "Feature selection and allocation to diverse subsets for multi-label learning problems with large datasets," Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on, IEEE, 2014, pp. 387–394. http://dx.doi.org/10.15439/2014F500

[11] Daniel T. Larose, "Data mining methods and models," John Wiley & Sons, Inc, 2006.

[12] Sierra, B., Larrañaga, P., Inza, I. "K Diplomatic Nearest Neighbour: giving equal chance to all existing classes," Journal of Artificial Intelligence Research, 2000

[13] Dudani, S. A. "The distance-weighted k-nearest neighbor rule," IEEE Tran-sactions on Systems, Man, and Cybernetics, Vol. SMC-6, No. 4, 1976, pp. 325–327

[14] Fix, E., Hodges Jr., J. L. "Discriminatory analysis — nonparametric discrimination: Consistency properties," Project 21-49-004, Report No. 4, USAF School of Aviation Medicine, Randolph Field, TX, USA, 1951, pp. 261–279.

[15] Fix, E., Hodges Jr., J. L. "Discriminatory analysis — nonparametric discrimination: Small sample performance," Project 21-49-004, Report No. 11, USAF School of Aviation Medicine, Randolph Field, TX, USA, 1952, pp. 280–322.

[16] Lichman, M. "UCI Machine Learning Repository," http://archive.ics.uci.edu/ml Irvine, CA: University of California, School of Information and Computer Science 2013