

# A Weighted Feature Selection Method for Instance-Based Classification

Gennady Agre<sup>1(✉)</sup> and Anton Dzhondzhorov<sup>2</sup>

<sup>1</sup> Institute of Information and Communication Technologies,  
Bulgarian Academy of Sciences, Sofia, Bulgaria  
agre@iinf.bas.bg

<sup>2</sup> Sofia University “St. Kliment Ohridski”, Sofia, Bulgaria  
anton.dzhondzhorov@gmail.com

**Abstract.** The paper presents a new method for selecting features that is suited for the instance-based classification. The selection is based on the ReliefF estimation of the quality of features in the orthogonal feature space obtained after PCA transformation, as well as on the interpretation of these weights as values proportional to the amount of explained concept changes. The user sets a threshold defining what percent of the whole concept variability the selected features should explain and only the first “stronger” features, which combine weights together exceed this threshold, are selected. During the classification phase the selected features are used along with their weights. The experiment results on 12 benchmark databases have shown the advantages of the proposed method in comparison with traditional ReliefF.

**Keywords:** Feature selection · Feature weighting · k-NN classification

## 1 Introduction

Feature selection problem has been widely investigated by the machine learning and data mining community. The main goal is to select the smallest feature subset given a certain generalization error, or alternatively to find the best feature subset that yields the minimum generalization error [19]. Feature selection methods are usually classified in three main groups: wrapper, filter, and embedded methods. Wrappers use a concrete classifier as a black box for assessing feature subsets. Although these techniques may achieve a good generalization, the computational cost of training the classifier a combinatorial number of times becomes prohibitive for high-dimensional datasets. The filter methods select some features without involving any classifier relying only on general characteristics of the training data. Therefore, they do not inherit any bias of a classifier. In embedded methods the learning part and the feature selection part can not be separated - the structure of the class of functions under consideration plays a crucial role. Although usually less computationally expensive than wrappers, embedded methods are still much slower than filter approaches, and the selected features are dependent on the learning machine. One of the most popular filter methods is ReliefF [10], which is based on evaluating the quality of the features. The present paper describes an approach for improving ReliefF as a feature selection method by its combination with PCA algorithm.

The structure of the paper is as follows: the next two sections briefly describe ReliefF and PCA algorithms. Section 4 presents the main idea of the proposed approach. The results of experiments testing the approach are shown in Sect. 5. Section 6 is devoted to discussion and related work and the final section is a conclusion.

## 2 Evaluation of Feature Quality by ReliefF

Relief [9] is considered as one of the most successful feature weighting algorithms. A key idea of it is to consider all features as independent ones and to estimate the relevance (quality) of a feature based on its ability to distinguish instances located near each other. In order to do this, the algorithm iteratively selects a random instance and then searches for its two nearest neighbours - a nearest hit (from the same class) and a nearest miss (from the different class). For each feature the estimation of its quality (weight) is calculated depending on the differences between the current instance and its nearest hit and miss along the corresponding attribute axis.

Sun and Li [21] have explained the effectiveness of Relief by showing that the algorithm is an online solution of a convex optimization problem, maximizing a margin-based objective function, where the margin is defined based on the nearest neighbour (1-NN) classifier. Therefore, compared with other filter methods, Relief usually performs better due to the performance feedback of a nonlinear classifier when searching for useful features. Compared with wrapper methods Relief avoids any exhaustive or heuristic combinatorial search by optimizing a convex problem and thus can be implemented very efficiently.

Igor Kononenko [10] proposed a more robust extension of Relief that was not limited to two class problems and could deal with incomplete and noisy data. Similarly to Relief, ReliefF randomly selects an instance  $E = \langle \{e_1, \dots, e_p\}, c_E \rangle$ , but then searches for  $k$  of its hits  $R^j = \langle \{r_1^j, \dots, r_p^j\}, c_E \rangle$ ,  $j = 1, \dots, k$  and misses  $N^j(c) = \langle \{n_1^j, \dots, n_p^j\}, c \rangle$ ,  $c \neq c_E, j = 1, \dots, k$  for each class  $c$  using metrics  $d(X, Y) = \left( \sum_{i=1}^p \delta^L(x_i, y_i) \right)^{\frac{1}{L}}$ ,  $L = 1, 2$ .

The calculation of feature weight updates averages the contribution of all the hits and all the misses ( $m$  is the number of iterations):

$$w_i \leftarrow w_i - \sum_{j=1}^k \frac{\delta(e_i, r_i^j)}{k \cdot m} + \frac{1}{m \cdot k} \sum_{c \neq c_E} \frac{P(c)}{1 - P(c_E)} \sum_{j=1}^k \delta(e_i, n_i^j(c))$$

The weights  $w_i$  calculated by ReliefF are varied in interval  $[-1, 1]$ , as features with the weight equal to  $-1$  are evaluated as absolutely irrelevant, and those with the weight of  $1$  - as absolutely relevant.

### 3 Selection of Features by PCA

Principal Component Analysis (PCA) is one of the most frequently used feature selection methods, which is based on extracting the basis axes (principle components) on which the data shows the highest variability [8]. The first principle component is in the direction of maximum variance of the given data. The remaining ones are mutually orthogonal and are ordered in a way of maximizing the remaining variance. PCA can be considered as a rotation of the original coordinate axes to a new set of axes that are aligned with the variability of the data - the total variability remains the same but the new features are now uncorrelated.

The dimensionality reduction (from  $n$  to  $k$  features,  $k < n$ ) is done by selecting only a part of all principle components - those orthonormal axes that have the largest associated eigenvalues of the covariance matrix. The user sets the threshold  $h$  defining a part of the whole variability of data that should be preserved and first  $k$  new features, which eigenvalues in common exceed the threshold, are selected.

An optimization of the classical PCA was proposed by Turk and Pentland [18]. It is based on the fact that when the number of instances  $N$  is less than the dimension of the space  $n$ , that only  $N-1$  rather than  $n$  meaningful eigenvectors will exist (the rest eigenvectors will have associated eigenvalues equal to zero). Having a normalized data matrix  $\mathbf{D}$ , the proposed optimization procedure allows calculating eigenvectors of original covariance matrix  $\mathbf{S} = \mathbf{D}\mathbf{D}^T$  based on eigenvectors of matrix  $\mathbf{L} = \mathbf{D}^T\mathbf{D}$ , which has the dimensions  $N \times N$ . This variant of PCA is very popular for solving tasks related to image representation, recognition and retrieval, when the number of images is significantly less than the number of features (pixels) describing an image.

### 4 Our Approach

Being very effective method for feature weighting, ReliefF has several drawbacks. One of them is that the algorithm assigns high relevance scores to all discriminative features, even if some of them are severely correlated [13]. As a consequence, redundant features might not be removed, when ReliefF is used for feature selection [21].

Although ReliefF has been initially developed as an algorithm for evaluating the quality of features, it has been commonly used as a feature subset selection method that is applied as a preprocessing step before the model is learnt [9]. In order to do that, a threshold  $h$  is introduced and only features, which weights are above this threshold, are selected. The selection of a proper value for the threshold is not an easy task and the dependence of ReliefF from this parameter has been mentioned as one of its shortcomings as a feature selection algorithm [6]. Several methods for selecting the threshold have been proposed - for example, Kira and Rendell [9] suggested the following bounds for this value:  $0 < h \leq \frac{1}{\sqrt{\alpha m}}$ , where  $\alpha$  is the probability of accepting an irrelevant feature as relevant and  $m$  is the number of iterations used. However, as it is mentioned in [13], “the upper bound for  $h$  is very loose and in practice much smaller values can be used”.

In its turn, PCA is traditionally used as a feature selection algorithm mainly in an unsupervised context – in most cases only first  $k$  uncorrelated attributes with total amount of eigenvalues above the user defined threshold  $h$  are selected. The main drawback of PCA, applied to the classification task, is that it does not take into account the class information of the available data. As it was mentioned in [4], the first few principal components would only be useful in those cases where the intra-class and inter-class variations have the same dominant directions or inter-class variations are clearly larger than intra-class variations. Otherwise, PCA will lead to a partial (or even complete) loss of discriminatory information. This statement was empirically confirmed for  $k$ -nearest neighbours (k-NN), decision trees (C4.5) and Naive Bayes classifiers tested on a set of benchmark databases [12].

Our approach is an attempt to explore the best features of the mentioned above methods for compensating their deficiencies. First, in order to improve the quality of ReliefF as *a feature weighting method* we apply it to a set of uncorrelated features found by PCA transformation decreasing in such a way the weights of redundant features and removing duplicating ones.

Second, in order to use ReliefF as *a feature selection method*, we apply a new method for selecting features, which is inspired by the interpretation of weights calculated by ReliefF as a portion of explained concept changes [13]. In such a way, the sum of the weights can be seen as an approximation to the value of concept variation – a measure of problem difficulty based on the nearest neighbor paradigm [13]. Similar to the approach used by PCA for selecting feature subsets in an unsupervised context, we select only the first  $k$  features (i.e. features with the biggest nonnegative ReliefF's weights) with *total amount of weights* above the user defined threshold  $h$ . Such an approach may be seen as a unified method for solving feature selection task based on evaluation of quality of features both in unsupervised and supervised context – in the first case we set a desired portion of explained total variability of the data, while in the second – the variability of the data is changed by the variability of concept to be learnt.

Lastly, the third aspect of our approach concerns the use of features selected by the proposed combination of algorithms in the context of the instance-based classification. As it has been shown in [20], ReliefF estimation of feature quality can be successfully used as weights in a distance metrics used for the instance-based classification:

$$d(X, Y) = \left( \sum_{i=1}^n w_i * \delta^L(x_i, y_i) \right)^{\frac{1}{L}}, w_i > 0$$

That is why, we use all features selected by our method *together with their weights*, when a classification task is going to be solved by an instance-based algorithm.

## 5 Experiments and Results

In order to test our ideas we have selected 12 databases described by numerical attributes (features) which number is varied from 4 to 16500 (see Table 1) – 9 benchmark bases are from UCI Machine Learning Repository (UCI)<sup>1</sup>, 2 – from Kent Ridge Bio-medical Dataset (KR)<sup>2</sup> and one is our own database [15].

**Table 1.** Databases used in the experiments.

Database	Source	Atts	Examples	Classes	Missing attributes	Default accuracy
Diabetes (DB)	UCI	8	768	2	No	0.651
Breast Cancer Wisconsin (BCW)	UCI	10	699	2	Yes	0.655
Glass (GL)	UCI	9	214	6	No	0.327
Wine (WN)	UCI	13	178	3	No	0.399
Iris (IR)	UCI	4	150	3	No	0.333
Vehicle Silhouettes (VS)	UCI	18	846	4	No	0.258
Liver Disorders (LD)	UCI	6	345	2	No	0.560
Sonar (SN)	UCI	60	208	2	No	0.534
Banknote (BN)	UCI	4	1372	2	No	0.555
Faces (FC)	Custom	16500	102	2	No	0.559
Lung Cancer (LC)	KR	12533	181	2	No	0.829
Ovarian Cancer (OC)	KR	15154	253	2	No	0.640

All databases were used for evaluating a classifier by means of a hold-out cross-validation schema – for each database 70 % of randomly selected examples were used for training the classifier and the rest 30 % - for testing it. The experiments were repeated 70 times and the results were averaged. A traditional 5-NN algorithm with Euclidian distance was used as a basic classifier. The same algorithm was used for evaluating the quality of different feature weighting schemas but with the weighted Euclidian distance. The differences in classification accuracy of the 5-NN classifiers used different algorithms for calculating feature weights were evaluated by Student t-paired test with 95 % significance level.

### 5.1 Evaluation of Feature Weighing Schemas

The following algorithms for calculating feature weights were evaluated:

<sup>1</sup> <https://archive.ics.uci.edu/ml/datasets.html>.

<sup>2</sup> <http://datam.i2r.a-star.edu.sg/datasets/krbd/>.

- *PCA*: the transformation was applied to each training set and the calculated eigenvectors were used for transforming the corresponding testing set. Since PCA does not change the distance between examples, the classification accuracy of an instance-based classifier is not changed when it is applied to the PCA transformed data. That is why, in order to evaluate the quality of PCA as a *feature weighting algorithm*, we used the calculated eigenvalues as feature weights in classification.
- *ReliefF*: the feature weights were calculated by applying ReliefF algorithm to each training set. In our experiments we used 10 nearest neighbours and up to 200 iterations as the ReliefF parameters.
- *PCA+ReliefF*: each training set was initially transformed by application of PCA and the resulted dataset was then used for calculating feature weights by means of ReliefF.

The experiment results are shown in Table 2. Classification accuracy values are shown in bold (meaning statistically significant better performance than the basic (5-NN) classifier), in italics (for worse performance) or in regular fonts for cases with no statistically significant differences. The “Total” row summarises this information.

**Table 2.** Classification accuracy when the calculated weights are used during classification

Database	5-NN	PCA	ReliefF	PCA+ReliefF
DB	0.734 $\pm$ 0.025	<i>0.728 <math>\pm</math> 0.024</i>	0.734 $\pm$ 0.026	0.733 $\pm$ 0.026
BCW	0.968 $\pm$ 0.013	<i>0.965 <math>\pm</math> 0.011</i>	<i>0.966 <math>\pm</math> 0.012</i>	<i>0.965 <math>\pm</math> 0.012</i>
GL	0.626 $\pm$ 0.050	<b>0.645 <math>\pm</math> 0.053</b>	<b>0.651 <math>\pm</math> 0.055</b>	<b>0.646 <math>\pm</math> 0.048</b>
WN	0.960 $\pm$ 0.024	<b>0.967 <math>\pm</math> 0.023</b>	0.962 $\pm$ 0.024	<b>0.977 <math>\pm</math> 0.018</b>
IS	0.961 $\pm$ 0.027	0.959 $\pm$ 0.029	0.959 $\pm$ 0.030	0.962 $\pm$ 0.028
VS	0.686 $\pm$ 0.022	<i>0.561 <math>\pm</math> 0.028</i>	<b>0.693 <math>\pm</math> 0.025</b>	<b>0.706 <math>\pm</math> 0.025</b>
LD	0.607 $\pm$ 0.037	<i>0.593 <math>\pm</math> 0.037</i>	0.613 $\pm$ 0.037	<b>0.644 <math>\pm</math> 0.043</b>
SN	0.797 $\pm$ 0.053	0.786 $\pm$ 0.048	0.791 $\pm$ 0.054	0.799 $\pm$ 0.053
BN	0.998 $\pm$ 0.002	<i>0.994 <math>\pm</math> 0.004</i>	<b>0.999 <math>\pm</math> 0.001</b>	<b>0.999 <math>\pm</math> 0.001</b>
FC	0.822 $\pm$ 0.055	<i>0.783 <math>\pm</math> 0.074</i>	<i>0.811 <math>\pm</math> 0.055</i>	0.821 $\pm$ 0.061
LC	0.925 $\pm$ 0.034	<b>0.978 <math>\pm</math> 0.026</b>	<b>0.938 <math>\pm</math> 0.034</b>	<b>0.963 <math>\pm</math> 0.034</b>
OC	0.915 $\pm$ 0.034	<i>0.832 <math>\pm</math> 0.042</i>	<b>0.959 <math>\pm</math> 0.023</b>	<b>0.939 <math>\pm</math> 0.028</b>
Total		3+, 7-, 2=	5+, 2-, 5=	7+, 1-, 4=

As it may be expected, PCA has been shown as a weak feature weighting algorithm in the classification context; however, it still lead to a significantly better accuracy on 3 databases and achieved a very high result on the Lung Cancer database.

The experiments have confirmed the commonly acknowledged statement that ReliefF is a rather strong feature weighting algorithm – in our case it has 5 statistically significant wins against 2 statistically significant loses and 5 statistically equal results in comparison with the unweighted variant of 5-NN algorithm.

The proposed combination of PCA and ReliefF has provided the best results – 7 statistically significant wins against only 1 statistically significant lose and 4 statistically equal results. The evaluation of differences in behaviour of pure ReliefF and ReliefF applied after PCA is shown in Table 3.

**Table 3.** Classification accuracy of PCA+ReliefF against ReliefF.

Database	ReliefF	PCA+ReliefF
DB	$0.734 \pm 0.026$	$0.733 \pm 0.026$
BCW	$0.966 \pm 0.012$	$0.965 \pm 0.012$
GL	$0.651 \pm 0.055$	$0.646 \pm 0.048$
WN	$0.962 \pm 0.024$	<b><math>0.977 \pm 0.018</math></b>
IS	$0.959 \pm 0.030$	$0.962 \pm 0.028$
VS	$0.693 \pm 0.025$	<b><math>0.706 \pm 0.025</math></b>
LD	$0.613 \pm 0.037$	<b><math>0.644 \pm 0.043</math></b>
SN	$0.791 \pm 0.054$	$0.799 \pm 0.053$
BN	$0.999 \pm 0.001$	$0.999 \pm 0.001$
FC	$0.811 \pm 0.055$	$0.821 \pm 0.061$
LC	$0.938 \pm 0.034$	<b><math>0.963 \pm 0.034</math></b>
OC	$0.959 \pm 0.023$	$0.939 \pm 0.028$
Total		4+, 1-, 7=

It can be seen, that in most cases the application of PCA before ReliefF has really improved the quality of ReliefF as a feature weighting algorithm.

## 5.2 Evaluation of Feature Selection Schemas

The results of the experiments for evaluating the ability of the mentioned above algorithms to select relevant features are presented in Table 4 (the first column presents the accuracy of 5-NN algorithm that uses the full set of features). *All selected features are used by 5-NN algorithm without taking into account their weights.*

**Table 4.** Classification accuracy when the calculated weights are used only for feature selection ( $h = 0.95$ )

Database	5-NN	PCA	ReliefF	PCA+ReliefF
DB	$0.734 \pm 0.025$	<b><math>0.738 \pm 0.026</math></b>	$0.732 \pm 0.026$	$0.734 \pm 0.025$
BCW	$0.968 \pm 0.013$	$0.969 \pm 0.011$	$0.966 \pm 0.013$	$0.968 \pm 0.013$
GL	$0.626 \pm 0.050$	$0.625 \pm 0.051$	<b><math>0.657 \pm 0.052</math></b>	$0.627 \pm 0.048$
WN	$0.960 \pm 0.024$	<b><math>0.969 \pm 0.021</math></b>	$0.959 \pm 0.026$	<b><math>0.966 \pm 0.023</math></b>
IS	$0.961 \pm 0.027$	$0.950 \pm 0.028$	$0.961 \pm 0.027$	$0.961 \pm 0.029$
VS	$0.686 \pm 0.022$	$0.607 \pm 0.025$	$0.683 \pm 0.020$	$0.688 \pm 0.022$
LD	$0.607 \pm 0.037$	$0.575 \pm 0.040$	$0.609 \pm 0.035$	$0.613 \pm 0.041$
SN	$0.797 \pm 0.053$	$0.800 \pm 0.050$	$0.796 \pm 0.049$	$0.801 \pm 0.052$
BN	$0.998 \pm 0.002$	$0.973 \pm 0.008$	$0.999 \pm 0.002$	$0.999 \pm 0.002$
FC	$0.822 \pm 0.055$	$0.818 \pm 0.056$	$0.811 \pm 0.055$	$0.814 \pm 0.057$
LC	$0.925 \pm 0.034$	$0.898 \pm 0.039$	$0.912 \pm 0.045$	$0.920 \pm 0.048$
OC	$0.915 \pm 0.034$	$0.912 \pm 0.034$	<b><math>0.931 \pm 0.030</math></b>	$0.914 \pm 0.034$
Total		2+, 6-, 4=	2+, 4-, 6=	1+, 0-, 10=

As it is expected, the behaviour of PCA as a feature selection algorithm in the supervised context still remains unsatisfactory, even though it is better than when it has been used for feature weighting.

A significant degradation can be observed in the behaviour of ReliefF. It should be mentioned that even when *all* features are used for weighted instance-based classification, in practice ReliefF *removes* highly irrelevant features, i.e. operates as (partially) feature selection algorithm. Such irrelevant features are those with the weights less or equal to zero. In our case, in addition to these irrelevant features ReliefF has been forced to remove slightly relevant or redundant attributes as well. Since the algorithm does not take into account the possible correlation between the features, it tends to underestimate less important (or redundant) features. As a result, the ordering of features by its importance created by ReliefF has occurred to be not very precise, which leads to removing some features that play important role for classification.

The mentioned above explanation is confirmed by the results shown in the last column of the table – the application of PCA eliminates the existing correlation between the features and allows ReliefF to evaluate the importance of the transformed features in a more correct way. As it can be seen, the accuracy of 5-NN algorithm running on the selected subset of features is statistically the same or even higher (for Wine database) than the accuracy of the same algorithm exploiting the whole set of features.

However, the comparison of results from Tables 2 and 4 has shown that as a whole, the classification accuracy of 5-NN algorithm used the combination of PCA and ReliefF for feature subset selection is *less* then the accuracy of the same algorithm that used the same combination for feature weighting. A possible explanation of this fact is that even after removing some irrelevant features the correct weighting of the rest features remains a very important factor for k-NN based classification. In order to prove this assumption we have conducted experiments in which feature weighting is combined with feature selection – during the classification phase all selected features are used with their weights calculated by the corresponding feature weighting algorithm at the pre-processing phase. The results of these experiments are shown in Table 5.

**Table 5.** Classification accuracy when the calculated weights are used both for feature selection and during classification ( $h = 0.95$ )

Database	5-NN	PCA	ReliefF	PCA+ReliefF
DB	$0.734 \pm 0.025$	$0.731 \pm 0.022$	$0.733 \pm 0.025$	$0.733 \pm 0.025$
BCW	$0.968 \pm 0.013$	<b><math>0.965 \pm 0.011</math></b>	$0.965 \pm 0.012$	$0.965 \pm 0.012$
GL	$0.626 \pm 0.050$	<b><math>0.643 \pm 0.052</math></b>	<b><math>0.656 \pm 0.055</math></b>	<b><math>0.646 \pm 0.048</math></b>
WN	$0.960 \pm 0.024$	<b><math>0.968 \pm 0.023</math></b>	$0.961 \pm 0.024$	<b><math>0.978 \pm 0.018</math></b>
IS	$0.961 \pm 0.027$	$0.956 \pm 0.031$	$0.959 \pm 0.029$	$0.961 \pm 0.028$
VS	$0.686 \pm 0.022$	$0.549 \pm 0.030$	$0.690 \pm 0.024$	<b><math>0.704 \pm 0.025</math></b>
LD	$0.607 \pm 0.037$	$0.584 \pm 0.042$	$0.615 \pm 0.038$	<b><math>0.642 \pm 0.041</math></b>
SN	$0.797 \pm 0.053$	$0.786 \pm 0.048$	$0.793 \pm 0.052$	$0.798 \pm 0.051$
BN	$0.998 \pm 0.002$	$0.962 \pm 0.010$	<b><math>0.999 \pm 0.001</math></b>	<b><math>0.999 \pm 0.001</math></b>
FC	$0.822 \pm 0.055$	$0.784 \pm 0.065$	$0.811 \pm 0.054$	$0.819 \pm 0.062$
LC	$0.925 \pm 0.034$	<b><math>0.977 \pm 0.026</math></b>	<b><math>0.939 \pm 0.033</math></b>	<b><math>0.965 \pm 0.033</math></b>
OC	$0.915 \pm 0.034$	$0.832 \pm 0.042$	<b><math>0.959 \pm 0.024</math></b>	<b><math>0.940 \pm 0.028</math></b>
Total		3+, 6-, 3=	4+, 1-, 7=	8+, 1-, 4=



As one can see, in the context of the instance-based classification the best results have been achieved by the combination of ReliefF feature weighting method applied to PCA transformed databases with the proposed by us schema for feature selection.

### 5.3 Dimensionality Reduction

The last question that should be discussed is the dimensionality reduction that has been achieved by the proposed method. The number of removed features is shown in Table 6.

**Table 6.** Contribution of different algorithms to dimensionality reduction.

DB	All Atts	PCA Weighting	ReliefF Weighting	PCA +ReliefF Weighting	PCA Selection (h = 0.95)	ReliefF Selection (h = 0.95)	PCA+ReliefF Selection (h = 0.95)
DB	8	0	0	0	1	1	0
BCW	10	0	0	0	3	1	2
GL	9	0	0	0	3	1	1
WN	13	0	0	0	3	1	3
IS	4	0	0	0	2	0	1
VS	18	0	0	0	12	2	2
LD	6	0	1	1	1	1	1
SN	60	0	3	14	34	15	26
BN	4	0	0	0	1	0	0
FC	16500	16429	1799	16467	16460	5888	16475
LC	12533	12407	3130	12450	12433	6265	12469
OC	15154	14977	123	15075	15131	3277	15097

It should be mentioned that our implementation of PCA includes the optimization proposed in [18], which is very efficient in cases, when the number of features ( $n$ ) is significantly greater than the number of examples ( $N$ ). In such cases PCA behaves as a feature selection algorithm that preserves only  $N$  relevant (most important) attributes. The contribution of such implementation of PCA to the reduction of the final feature subset is shown in the table column named ‘PCA Weighting’. The number of features evaluated by ReliefF as highly irrelevant (i.e. with non-positive values of feature weights) is shown in the column ‘ReliefF Weighting’. The next column displays the number of features evaluated as highly irrelevant in cases when ReliefF has been applied after PCA transformation. The last three columns show the number of features that have been removed by the corresponding algorithm when the threshold  $h$  has been set.

The results show that when a database is described by a relatively small number of features (the first 9 databases in the table), our algorithm has succeeded to remove, in average, 10.8 % of them without compromising or even significantly raising in most cases (5+, 1–, 3=) the classification accuracy of 5-NN algorithm. The main contribution to this reduction belongs to ReliefF algorithm, which has evaluated (in average)

4.5 % of the features as highly irrelevant and 6.3 % of them – as weakly irrelevant or redundant.

In the last three databases, in which the number of attributes is significantly greater than the number examples, our algorithm has eliminated, in average, 99.7 % of the features evaluating 99.6 % of them as highly irrelevant and only 0.1 % - as redundant. The main contribution to this dimensionality reduction is due to the role of PCA implementation used. However, setting threshold  $h$  to 95 % of the total explained concept variability has forced ReliefF algorithm to evaluate 25.2 % of features remaining after PCA transformation as redundant. In the same time, the average classification accuracy of 5-NN algorithm using, in average, only 0.3 % of all features has significantly raised (2+, 0–, 1=).

All mentioned above have proved that the proposed method can be successfully used for dimensionality reduction without compromising the accuracy of instance-based classification.

## 6 Discussion and Related Work

The question that should be discussed is the scalability of the proposed approach. The first aspect of this problem is related to types of features that can be processed. ReliefF does not have any problems with processing nominal features by changing, for example, Euclidian distance with Manhattan distance [13]. Although the computation of principle components in PCA explores the apparatus of linear algebra, the algorithm can be easily adapted to work with nominal features by their binarization. [7, 17]. Since several binarization methods exist, we have not purposefully included any databases with nominal features into our experiments in order to exclude possible influence of such methods to the final quality of the proposed approach. However, the binarization allows applying our approach to databases with nominal features as well.

The other aspect is the dependence of the approach from the number of features and instances. For a database contained  $N$  instances described by  $n$  features, the complexity of Relief using  $m$  iterations is  $O(mnN)$  – the same is valid for ReliefF as the most complex operation is finding  $k$  nearest neighbours for an instance. However, many different techniques for fast and efficient k-NN search have been developed [3], which can be applied in ReliefF implementation as well. Fast and scalable implementations exist also for PCA transformation (see e.g. [11]), so the proposed combination of PCA and ReliefF is also scalable.

Another question concerns the classification accuracy of k-NN algorithms that use the proposed approach for data pre-processing. In such a context our method could be considered as a wrapper feature selection method and the typical cross-validation approach can be applied as for selecting optimal  $k$ , as for selecting the proper value of threshold  $h$  optimizing the accuracy of the k-NN classifier.

The similar approach for feature selection based on ReliefF and PCA was proposed in [22] in the context of underwater sound classification. The authors also used PCA for removing correlation between the features and then ReliefF – for evaluating the feature quality. However, the selection of the features was done in the traditional manner (by comparison of each feature weight against a threshold), which led to unconvincing

results. Moreover, the weights of the selected features were not used for classification. The approach was tested only on a single dataset of a small dimension (39 features) and was compared only with PCA without presenting any information about statistical significance of the results.

Considering our approach as a method for adapting PCA to the classification task, it can be related to [17] and works on class-dependent PCA methods (see e.g. [14]).

Considering the proposed method as an approach for improving ReliefF algorithm, it can be related to such works as [5, 21]. The first work proposes a so-called Orthogonal Relief, which is a combination of sequential forward selection procedure, the Gram-Schmidt orthogonalization procedure and Relief. The algorithm was tested only on 4 databases with 2 classes. The second work proposes a variant of Relief algorithm called WACSA, where correlations among features are taken into account to adjust the final feature subset. The algorithm was tested on five artificial well-known databases from the UCI repository.

Our future plans include more intensive testing of the proposed approach on a more diverse set of databases and comparing it with other state-of-the-art methods for feature selection.

## 7 Conclusion

The paper presents a new method for feature selection that is suited for the instance-based classification. The selection is based on the ReliefF estimation of the quality of attributes in the orthogonal attribute space obtained after PCA transformation, as well as on the interpretation of these weights as values proportional to the amount of explained concept changes. Only the first “strong” features, which combined ReliefF weights exceed the user defined threshold defining the desired percent of the whole concept variability the selected features should explain, are chosen. During the classification phase the selected features are used along with their weights calculated by ReliefF. The results of intensive experiments on 12 datasets have proved that the proposed method can be successfully used for dimensionality reduction without compromising or even raising the accuracy of instance-based classification.

## References

1. Bins, J., Draper, B.: Feature selection from huge feature sets. In: Proceedings of the Eighth IEEE International Conference on Computer Vision, vol. 2, pp. 159–165 (2001)
2. Chang, C.-C.: Generalized iterative RELIEF for supervised distance metric learning. *Pattern Recogn.* **43**(8), 2971–2981 (2010)
3. Dhanabal, S., Chandramathi, S.: A review of various k-nearest neighbor query processing techniques. *Intern. J. Comput. Appl.* **31**(7), 14–22 (2011)
4. Diamataras, K.I., Kung, S.J.: *Principal Component Neural Networks. Theory and Applications.* Wiley, New York (1996)

5. Florez-lopez, R.: Reviewing RELIEF and its extensions: a new approach for estimating attributes considering high-correlated features. In: Proceedings of IEEE International Conference on Data Mining, Maebashi, Japan, pp. 605–608 (2002)
6. Freitag, D., Caruana, R.: Greedy attribute selection. In: Proceedings of Eleven International Conference on Machine Learning, pp. 28–36 (1994)
7. Hall, M.A.: Correlation-based feature selection of discrete and numeric class machine learning. In: Proceedings of International Conference on Machine Learning (ICML-2000), San Francisco, CA, pp. 359–366. Morgan Kaufmann, San Francisco (2000)
8. Jolliffe, I.T.: Principal Component Analysis. Springer, New York (1986)
9. Kira, K., Rendell, L.A.: The feature selection problem: traditional methods and a new algorithm. In: Proceedings of AAAI 1992, San Jose, USA, pp. 129–134 (1992)
10. Kononenko, I.: Estimating attributes: analysis and extensions of RELIEF. In: Proceedings of European Conference on Machine Learning, Catania, Italy, vol. 182, pp. 171–182 (1994)
11. Ordóñez, C., Mohanam, N., García-Alvarado, C.: PCA for large data sets with parallel data summarization. *Distrib. Parallel Databases* **32**(3), 377–403 (2014)
12. Pechenizkiy, M.: The impact of feature extraction on the performance of a classifier: kNN, Naïve Bayes and C4.5. In: Kégl, B., Lee, H.-H. (eds.) *Canadian AI 2005. LNCS (LNAI)*, vol. 3501, pp. 268–279. Springer, Heidelberg (2005)
13. Robnik-Sikonja, M., Kononenko, I.: Theoretical and empirical analysis of ReliefF and RReliefF. *Mach. Learn. J.* **53**, 23–69 (2003)
14. Sharma, A., Paliwala, K., Onwubolu, G.: Class-dependent PCA, MDC and LDA: a combined classifier for pattern classification. *Pattern Recogn.* **39**, 1215–1229 (2006)
15. Strandjević, B., Agre, G.: On impact of PCA for solving classification tasks defined on facial images. *Intern. J. Reason. Based Intell. Syst.* **6**(3/4), 85–92 (2014)
16. Sun, Y., Li, J.: Iterative RELIEF for feature weighting: algorithms, theories, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(6), 1035–1051 (2007)
17. Tsymbal, A., Puuronen, S., Pechenizkiy, M., Baumgarten, M., Patterson, D.W.: Eigenvector-based feature extraction for classification. In: Proceedings of FLAIRS Conference, pp. 354–358 (2002)
18. Turk, M., Pentland, A.: Eigenfaces for recognition. *J. Cogn. Neurosci.* **3**(1), 71–86 (1991)
19. Vergara, J., Estevez, P.: A review of feature selection methods based on mutual information. *Neural Comput. Appl.* **24**, 175–186 (2014)
20. Wettschereck, D., Aha, D.W., Mohri, T.: A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artif. Intell. Rev.* **11**, 273–314 (1997)
21. Yang, J., Li, Y.-P.: Orthogonal relief algorithm for feature selection. In: Huang, D.-S., Li, K., Irwin, G.W. (eds.) *ICIC 2006. LNCS*, vol. 4113, pp. 227–234. Springer, Heidelberg (2006)
22. Zeng, X., Wang, Q., Zhang, C., Cai, H.: Feature selection based on ReliefF and PCA for underwater sound classification. In: Proceedings of the 3rd International Conference on Computer Science and Network Technology (ICCSNT), Dalian, pp. 442–445 (2013)

Artificial Intelligence: Methodology, Systems, and  
Applications

17th International Conference, AIMS 2016, Varna,  
Bulgaria, September 7-10, 2016, Proceedings

Dichev, C.; Agre, G. (Eds.)

2016, XII, 370 p. 84 illus., Softcover

ISBN: 978-3-319-44747-6