Class dependent feature weighting and K-nearest neighbor classification

Elena Marchiori

Institute for Computing and Information Sciences, Radboud University Nijmegen, The Netherlands
elenam@cs.ru.nl

Abstract. Feature weighting in supervised learning concerns the development of methods for quantifying the capability of features to discriminate instances from different classes. A popular method for this task, called RELIEF, generates a feature weight vector from a given training set, one weight for each feature. This is achieved by maximizing in a greedy way the sample margin defined on the nearest neighbor classifier. The contribution from each class to the sample margin maximization defines a set of *class dependent feature weight* vectors, one for each class. This provides a tool to unravel interesting properties of features relevant to a single class of interest.
In this paper we analyze such class dependent feature weight vectors. For instance, we show that in a machine learning dataset describing instances of recurrence and non-recurrence events in breast cancer, the features have different relevance in the two types of events, with size of the tumor estimated to be highly relevant in the recurrence class but not in the non-recurrence one. Furthermore, results of experiments show that a high correlation between feature weights of one class and those generated by RELIEF corresponds to an easier classification task.
In general, results of this investigation indicate that class dependent feature weights are useful to unravel interesting properties of features with respect to a class of interest, and they provide information on the relative difficulty of classification tasks.

1 Introduction

Feature selection is a central problem in machine learning, because of its use in a wide range of real-life applications, such as in pattern recognition, text categorization and biological and biomedical data analysis [6]. Feature selection in supervised learning is motivated by the fact that using all available features may negatively affect generalization performance, especially in the presence of irrelevant or redundant features. Therefore, feature selection aims at making good predictions with few features.

Many feature selection algorithms have been proposed (see for instance the overviews in [6, 10]). In particular, feature weighting assigns real values (instead of zero or one) to features, describing their relevance to a learning problem. Among the existing feature weighting algorithms, RELIEF [7, 9, 8] is considered one of the most successful methods, due to its simplicity and effectiveness [1]. Interesting formalizations of RELIEF

have been proposed and used for developing new feature weighting algorithms [5, 11]. In particular it was shown that RELIEF can be interpreted as an online solution to a convex optimization problem, which maximizes a margin-based objective function, where the margin is defined on the 1-nearest neighbor (1-NN) classifier. This explains its good performance both when compared with filter methods, due to the performance feedback of a nonlinear classifier when searching for useful features, and when compared with wrapper methods, since it optimizes a convex problem, hence avoids any exhaustive or heuristic combinatorial search and can be implemented very efficiently [11].

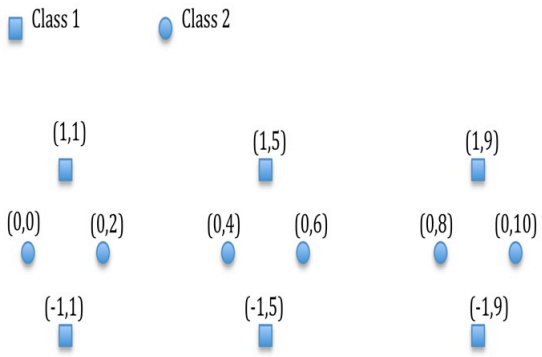

Fig. 1. Toy example.

This paper investigates a decomposition of RELIEF into class dependent feature weight terms, one for each class of the learning problem. Each class dependent term induces a weighted distance enlarging the sample margin of the corresponding class. This approach can be viewed as the supervised counterpart of clustering in subspaces spanned by different combinations of dimensions via local weightings of features (see, e.g., [4, 2]).

We show that complementary characteristics of a feature in different classes may yield different weight contributions which, when added, neutralize each other. This may prevent to detect the relevance of some features for a single class.

This situation is illustrated by the following toy example. RELIEF applied to the training data shown in Figure 1 assigns zero weight to all the features. However, if the class dependent feature weight terms are considered, the two features become relevant for class 1 and 2, respectively. This is shown in detail in Section 2.1.

The goal of this paper is to analyze class dependent feature weight vectors. First, we investigate whether class dependent feature weights provide useful information about the relative difficulty of classification tasks. Second, we investigate whether class dependent feature weights provide better information about the relevance of features than RELIEF for the underlying phenomenon under study.

Results of experiments show that a high correlation between feature weights of one class and those generated by RELIEF are associated to an easier classification task.

Furthermore, we show that in a machine learning dataset describing instances of recurrence and non-recurrence events in breast cancer, the features have different relevance in the two types of events. In particular, the size of the tumor is estimated to be highly relevant in the recurrence class but not in the non-recurrence one.

In general, results of this investigation indicate that class dependent feature weights are useful to unravel interesting properties of features with respect to a class of interest, and can be used to analyze comparatively the difficulty of classification tasks.

2 Methods

We use the variant of RELIEF acting on all the training set instances [8] (see Algorithm 1 for binary classification problems).

Algorithm 1 RELIEF

Require: X, l: training data, each instance vector x has m features and class $l(x)$
Ensure: w: feature weight vector
  w = vector with m zeros
  for x in X **do**
    $w = w + \sum_{z \in \mathtt{KNN}(x,c), c \neq l(x)} |x - z| - \sum_{z \in \mathtt{KNN}(x,l(x))} |x - z|$;
  end for

X denotes a dataset of n instances, and x, z generic instances. Each instance $x = (x_1, \ldots, x_m)$ is a real-valued vector of dimension m, whose entries are here called features.

C denotes the set of class labels; $l : X \to C$ is the function mapping each instance x to its class label $l(x)$. Let c be a generic element of C, and X_c be the subset of X consisting of those instances having class label equal to c. Then $\mathtt{KNN}(x, c)$ is the set of K nearest neighbors (with respect to the Euclidean distance) of x computed using only the instances in X_c, excluded x.

For two instance vectors x and y, $|x-y|$ denotes the vector consisting of the absolute value of the difference of each pair of corresponding entries in x and y. For instance, if $x = (1, 2)$ and $y = (3, 4)$ the $|x - y| = (2, 2)$.

Given a training set of labeled instances, RELIEF generates the feature weight vector w of size equal to the number (m) of features. Each feature's weight in w is initialized to zero and updated iteratively by processing each instance vector x of X as follows. The K nearest neighbors of x from the opposite class are computed, and for each one of them, say z, the vector $|x - z|$ is added to w. The K nearest neighbors from the same class are computed, and for each one of them, say z, the vector $|x - z|$ is subtracted from w.

2.1 Class Dependent Feature Weighting

The feature weight vector computed by RELIEF can be re-written as $w = \sum_{c \in C} w_c$ where

$$w_c = \sum_{x \in X_c} \left\{ \sum_{z \in \mathrm{KNN}(x,c)} -|x - z| + \sum_{\substack{z \in \mathrm{KNN}(x,c') \\ c' \neq c}} |x - z| \right\}.$$

The term w_c can be viewed as a feature weight vector conditioned to class c: we call it *class dependent feature weight* vector.

For the toy example described in the Introduction (see Figure 1) the class dependent feature weight vectors for class 1 and 2 are $(-6, 6)$ and $(6, -6)$, respectively. Their sum is the vector $(0, 0)$. Therefore RELIEF assigns weight zero to all features. However, the two class dependent weight vectors estimate the second and first feature as 'relevant' for class 1 and 2, respectively.

This observation shows that, by averaging the weigths of features across different classes, information about the relevance of features with respect to a single class can be lost, while this information is visible if the weigths of features for each class are kept separated.

A class dependent feature weight w_c has a direct interpretation in terms of sample margin [5, 11]. In fact w_c can be viewed as the result of a greedy procedure for maximizing the sample margin *conditioned to class* c as described by the following objective function:

$$\theta_c = \sum_{\substack{x \ s.t. \\ l(x)=c}} \left(\sum_{c' \neq c} \sum_{z \in \mathrm{KNN}_w(x,c')} \|x - z\|_w - \right.$$

$$\left. \sum_{z \in \mathrm{KNN}_w(x,c)} \|x - z\|_w \right).$$

Here $\|x - z\|_w$ denotes the weighted Euclidean distance between x and z with respect to feature weights $w = w_1, \ldots, w_m$, $w_i \geq 0$, defined as $\left(\sum_{i=1}^{m} w_i (x_i - z_i)^2\right)^{-1/2}$. $\mathrm{KNN}_w(x)$ represents the list of K nearest neighbors of x in the training set computed using the weighted Euclidean distance $\| \cdot \|_w$.

Instead, the greedy procedure of RELIEF maximizes the sample margin over the entire training set, and in this way it considers the sum of the possibly competing objectives θ_c's.

3 Applications

The goals of our investigation are twofold. First, we investigate whether class dependent feature weights provide useful information about the difficulty of the underlying classification task. Second, we investigate whether class dependent feature weights provide better information than RELIEF does about the relevance of features in relation to the underlying phenomenon under study.

To this aim we consider a number of life sciences datasets available at the UCI Machine Learning repository (see http://archive.ics.uci.edu/ml/datasets.html). Their characteristics are given in Table 1.

Table 1. Datasets used in the experiments. CL = number of classes, TR = training set, TE = test set, FS = number of features, Cl.Inst. = number of instances in each class.

DATASET	FS	TR	CL.INST.	TE	CL.INST.
B.CANCER	9	200	140-60	77	56-21
DIABETES	8	468	300-168	300	200-100
HEART	13	170	93-77	100	57-43
SPLICE	60	1000	525-475	2175	1123-1052
THYROID	5	140	97-43	75	53-22
BREAST-W	9	546	353-193	137	91-46
BUPA	6	276	119-157	69	26-43
PIMA	8	615	398-217	153	102-51

Examples of class dependent and RELIEF weights of the considered problems are shown in Figure 4. The plots in this figure show that in some cases the two classes are in strong disagreement with respect to the way they estimate the relevance of different features.

For instance, on the B.Cancer data the feature rankings induced by the two class dependent weight vectors are rather different. A similar situation can be observed for the Bupa data, where the resulting rankings are the reverse of each other. These observations show that also on real-life problems the contributions of the classes to the feature weight vector computed by RELIEF may be rather different.

3.1 Class dependent weights and relative difficulty of the classification task

We want to assess whether the *diversity* between the class dependent feature weights and the weights computed by RELIEF and the *hardness* of classification tasks are linked with each other.

Table 2. Results for datasets used in the experiments. Acc = average test accuracy of K-NN classifier (with K=5) over 10 runs, Ave Corr1 = average of the correlations between RELIEF weights and class 1 dependent weights, Ave Corr2 = average of the correlations between RELIEF weights and class 2 dependent weights. Max Corr = average over 10 runs of the maximum between Corr 1 and Corr 2. Standard deviation (std) over 10 runs is reported between brackets.

DATASET	ACC (STD)	CORR1 (STD)	CORR2 (STD)	MAX CORR (STD)
B.CANCER	70.12 (5.44)	-0.35 (15.75)	64.43 (10.79)	64.43 (10.79)
DIABETES	70.65 (4.57)	-5.73 (24.38)	86.24 (4.85)	86.24 (4.85)
HEART	80.00 (3.39)	36.33 (16.64)	63.88 (16.37)	67.41 (10.38)
SPLICE	72.89 (1.72)	43.50 (13.65)	91.64 (3.37)	91.64 (3.37)
THYROID	91.86 (3.84)	42.90 (24.65)	89.03 (3.67)	89.03 (3.67)
BREAST-W	97.95 (1.32)	21.93 (19.29)	97.91 (1.01)	97.91 (1.01)
BUPA	65.07 (6.12)	61.29 (34.41)	-13.47 (38.37)	67.53 (23.31)
PIMA	72.73 (1.08)	24.45 (35.18)	89.53 (6.61)	89.53 (6.61)

To this end we perform 10 runs on each dataset. In each run a partition of the dataset into training and test set is used, where the size of training and test set reported in the UCI ML repository are employed (see Table 1). On each partition we compute the class dependent weight vectors and the RELIEF weights using the training data, and apply the K-NN classifier (with K=5) to the test data.

We measure hardness using the mean test accuracy, and diversity using the mean of the maximum linear correlations between class dependent feature weights and the RELIEF ones (see Table 2, columns 'Corr1', 'Corr2' , and 'Max Corr').

Then we use the Spearman's correlation to assess how well the relationship between these two variables can be described using a monotonic function. A positive Spearman correlation coefficient corresponds to an increasing monotonic trend between the two variables.

We compute the Spearman's correlation coeffient ρ with right tail (that is, with alternative hypothesis 'correlation is greater than zero' against which to compute p-values for testing the hypothesis of no correlation) between the variables 'hardness' and 'diversity' as described by the vectors generated by computing their values on each dataset. The resulting hardness and diversity vectors are the columns 'Acc' and 'Max Corr' of Table 2. Computation of the coefficient yields a positive correlation $\rho = 0.5952$ with p-value 0.06.

If we consider the list of test accuracies and the list of maximum correlations (between class dependent feature weights and the RELIEF ones) generated in the 10 runs instead of their mean, then a smaller yet more significant correlation is obtained, $\rho = 0.4155$ with p-value 0.0001.

Furthermore, if we consider the test classification achieved using only the top 5 features selected using the feature ranking induced by the weights generated by RELIEF, then the correlation signal between hardness and diversity becomes highly significant. Indeed, the Spearman's correlation coefficients become $\rho = 0.8333$ (p-value 0.0154) and $\rho = 0.6476$ (p-value 0) when the mean (over the 10 runs) results and the results in all the runs are considered, respectively.

The above results indicate that class dependent feature weights provide information about the relative difficulty of the learning tasks (as measured by the test error).

3.2 Class dependent weights and feature relevance

In this section we analyze the class dependent feature weights and the RELIEF ones in the context of breast cancer data analysis, using two breast cancer datasets B.Cancer and the Breast-W.

The B.Cancer Data This dataset was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. It consists of instances from recurrence (class 2) and non-recurrence (class 1) breast cancer patients. The instances consist of the following features:

1. age: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99,
2. menopause: lt40, ge40, premeno,

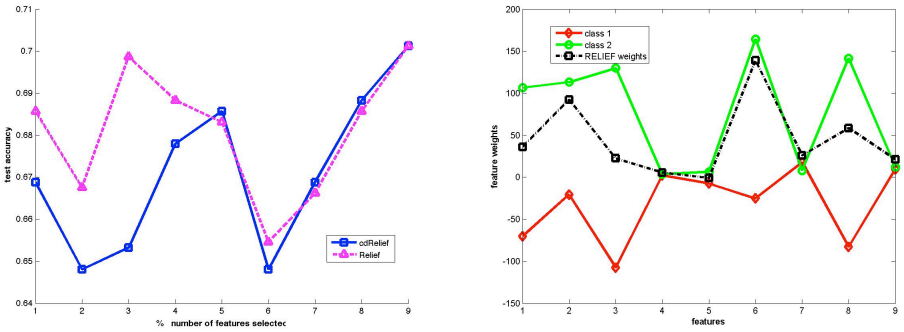

Fig. 2. Average accuracy of the K-NN classifier (K=5) for different number of selected features (left) and example of feature weight vectors (right) on dataset B.Cancer.

3. tumor-size: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59,
4. inv-nodes: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39,
5. node-caps: yes, no,
6. deg-malig: 1, 2, 3,
7. breast: left, right,
8. breast-quad: left-up, left-low, right-up, right-low, central,
9. irradiat: yes, no.

Figure 2 (right side) shows the weights of the features according to RELIEF, and the class dependent ones. The two classes assign rather different values to some features. In particular, class 2 'recurrence events' assigns high values to feature 'tumor-size' (feature 3), 'deg-malig' (6), and breast-quad (8), while class 1 'non-recurrence-events' considers them not relevant (assigns negative weights).

Thus when the class dependent relevance values are added by RELIEF, feature 'tumor-size' (3) becomes not very relevant for this learning task, less relevant than feature 1 ('age') and 2 ('menopause'). However, 'tumor-size' is recognized as relevant for the 'recurrence events' class. Indeed, research in breast cancer has shown tumor size to be strongly associated to the its recurrence hence to survival (see e.g. [3]).

Figure 2 (left side) shows the average test accuracy (over 10 runs) obtained by the K-NN classifier (with K=5) when varying the number of selected features. The features are selected using the ranking induced either by the weights computed by RELIEF or by the class dependent weights with highest variance ('cdRelief' classifier in Figure 2). Results show that test accuracy decreases when using only one class to estimate the relevance of features. This is expected since the feature ranking is biased towards one class and the two classes have different feature rankings.

Wisconsin Breast Cancer Dataset The Wisconsin Breast Cancer dataset (Original), donated from Dr. WIlliam H. Wolberg (University of Wisconsin Hospitals Madison, Wisconsin, USA), is publicly available at the UCI Machine Learning Repository. The dataset contains malignant (cancerous, class 2) and benign (non-cancerous, class 1) instances. The instances consist of the following features:

1. Clump Thickness: 1 - 10,
2. Uniformity of Cell Size: 1 - 10,
3. Uniformity of Cell Shape: 1 - 10,
4. Marginal Adhesion: 1 - 10,
5. Single Epithelial Cell Size: 1 - 10,
6. Bare Nuclei: 1 - 10,
7. Bland Chromatin: 1 - 10,
8. Normal Nucleoli: 1 - 10,
9. Mitoses: 1 - 10.

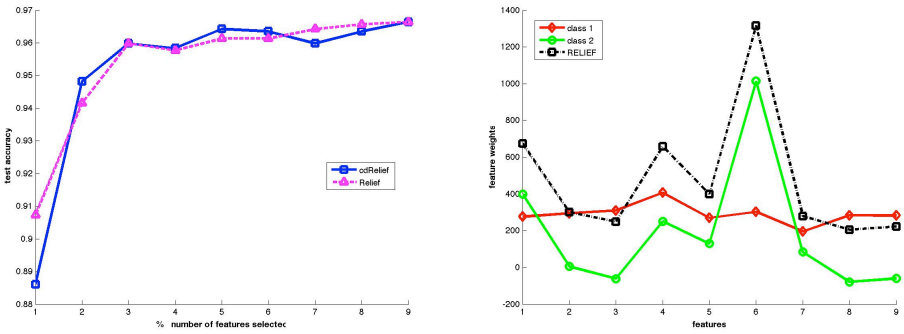

Fig. 3. Average accuracy of the K-NN classifier (K=5) for different number of selected features (left) and example of feature weight vectors (right) on dataset Breast-W.

On this dataset the class dependent feature weights for class 2 (malignant cancer) are in strong accordance with those of RELIEF (see the right plot in Figure 3 for an example of weights generated by a run). Therefore the relevance as estimated by RELIEF reflects the importance of the features for malignant cancer.

Moreover the average test classification performance is not significantly affected when using the feature ranking generated by class 2 (see the performance of 'cdRelief' in the left plot in Figure 3). Indeed, on this problem both the average accuracy (97.95) and the maximum Pearson correlation, that is, that between RELIEF and class 2 weights (97.91) are high.

4 Conclusions

We investigated a decomposition of RELIEF into class dependent feature weight vectors, where each vector describes the relevance of features conditioned to one class. The results of experiments indicated the usefulness of this decomposition for unraveling relevant features for a single class and for providing information about the difficulty of the considered learning task.

In general, results indicated that using only one class to estimate the relevance of features is not beneficial for the classification performance. This is expected, and shows that feature relevance for classification is different than feature relevance for a single class. The latter type of relevance is important when the goal is to unravel useful information about the phenomenon under study, like in the example about the recurrence of breast cancer events we discussed in this paper.

The contributions of this work provide initial insights and results about the advantages of using feature relevance in a way that depends on the single classes. Future work includes the development of a theoretical and methodological framework in order to better understand, use and combine class dependent feature weights. For instance, on the methodological side, an interesting problem for future research is to investigate whether the use of multi-objective optimization for feature weighting could be employed to improve also the classification performance, where the objectives are the θ_c's (that is, the sample margin conditioned to the classes c in C).

References

1. T.G. Dietterich. Machine-learning research: Four current directions. *The AI Magazine*, 18(4):97–136, 1998.
2. C. Domeniconi, D. Gunopulos andS. Ma, B. Yan, M. Al-Razgan, and D. Papadopoulos. Locally adaptive metrics for clustering high dimensional data. *Data Min. Knowl. Discov.*, 14(1):63–97, 2007.
3. E.B. Elkin, C. Hudis, C.B. Begg, and D. Schrag. The effect of changes in tumor size on breast carcinoma survival in the u.s.: 1975-1999. *Cancer*, 104(6):1149–57, 2005.
4. J.H. Friedman and J.J. Meulman. Clustering objects on subsets of attributes. *Journal of the Royal Statistical Society*, 6:815–849, 2004.
5. R. Gilad-Bachrach, A. Navot, and N. Tishby. Margin based feature selection - theory and algorithms. In *ICML*, 2004.
6. I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.
7. K. Kira and L.A. Rendell. The feature selection problem: Traditional methods and a new algorithm. In *National Conference on Artificial Intelligence*, pages 129–134, 1992.
8. R. Kohavi and G. John. Wrappers for feature subset selection. *Artificial Intelligence journal*, 97(1-2):273–324, 1997.
9. I. Kononenko. Estimating attributes: Analysis and extensions of RELIEF. In Springer, editor, *ECML*, volume LNCS 784, pages 171–182, 1994.
10. H. Liu and H. Motoda. *Computational Methods of Feature Selection (Chapman & Hall/Crc Data Mining and Knowledge Discovery Series)*. Chapman & Hall/CRC, 2007.
11. Y. Sun. Iterative RELIEF for feature weighting: Algorithms, theories, and applications. *IEEE TPAMI*, 29(6):1035–1051, 2007.

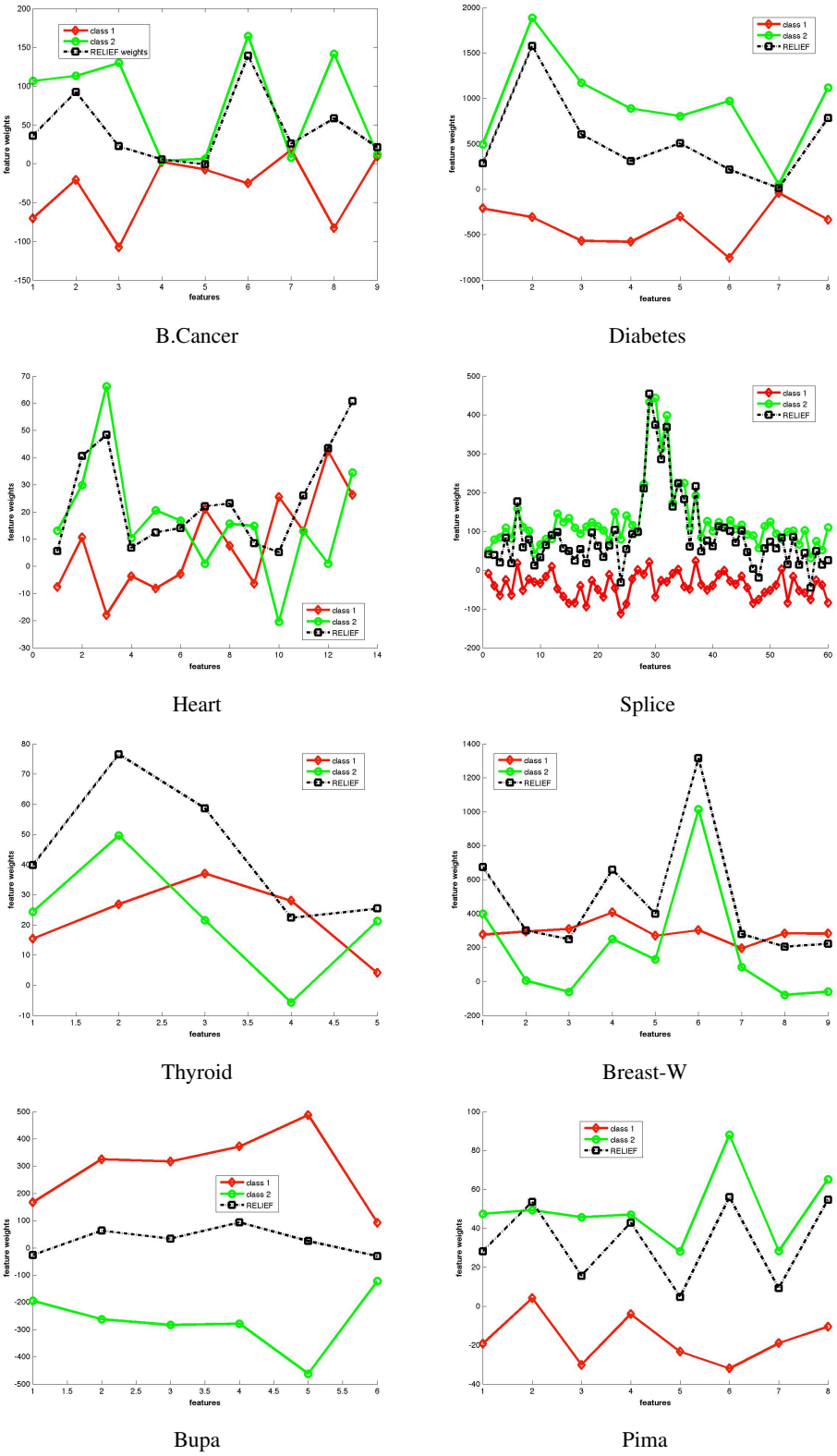

Fig. 4. Examples of class dependent and RELIEF weights of the considered datasets.