



Nonspecificity, strife and total uncertainty in supervised feature selection

Christoph Lohrmann*, Pasi Luukka

LUT University, School of Business & Management, Yliopistonkatu 34, 53850 Lappeenranta, Finland

ARTICLE INFO

Keywords:

Classification
Feature ranking
Filter method
Dimensionality reduction
Fuzzy logic
Machine learning

ABSTRACT

In this paper we propose three novel feature ranking methods for supervised feature selection in the context of classification which are based on possibility theory. All three methods – nonspecificity, strife and total uncertainty – are tested on eight artificial data sets and ten medical real-world data sets and benchmarked against ReliefF, the Fisher score, the fuzzy entropy and similarity (FES), the Fuzzy similarity and entropy (FSAE) filter, symmetrical uncertainty as well as using no feature selection. The feature ranking methods were applied following two approaches: (1) using a fixed threshold for the number of highest-ranking features selected and (2) using a hybrid feature selection approach with a classifier (k-nearest neighbor classifier, decision tree, similarity classifier, SVM) to select the optimal number of features to select. The results indicate that strife and the Fisher score are the two feature ranking methods that for both approaches are on average ranked the highest in terms of the test set accuracy on the real-world data sets. Besides that, for the hybrid approach, strife uses most of the time a considerably smaller number of features than nonspecificity and total uncertainty. In terms of stability, which was measured with the adjusted stability measure (ASM), the Fisher score and strife were among the most stable feature ranking methods in this study. Additionally, strife's feature subsets were diverse compared to those of the remaining feature selection methods, making it a good candidate to be included in a feature selection ensemble.

1. Introduction

In the era of information and big data where a 'wealth' of information can be obtained e.g. via the internet, the dimensionality of data and the handling of it has attracted a growing interest. At first glance, having more features (=variables) may appear advantageous for researchers and practitioners alike that aim to build models such as classification models based on high-dimensional data. However, such a situation also poses challenges, which are usually not covered in simple text book examples (Caruana and Freitag, 1994a). It is common sense that only those features that are relevant for a specific problem should be used to build a model so that it can identify the patterns in the data and generalize well to new observations to make accurate predictions. Since it may not be known in advance which features are actually relevant, there is an incentive to include numerous features into the training data for a model to avoid excluding any relevant feature (Almuallim and Dietterich, 1991). However, including a large number of features is often associated with the inclusion of irrelevant or redundant features, which can act as noise and may interfere with the remaining features and, hence, be detrimental to learning the actual generalizable patterns in the data (Caruana and Freitag, 1994a; Dessì and Pes, 2015). In addition to this potential negative impact on the model accuracy, including irrelevant or redundant features will increase the computational complexity by having high dimensional

data to model with Dessì and Pes (2015). Hence, researchers and practitioners that do not know which features are relevant may face a dilemma: if relevant features are excluded, the model may not be able to find the patterns in the data, whereas if irrelevant or redundant features are included, the model may overfit and not generalize well. Especially if a task is high-dimensional and complex, focusing on the relevant features becomes pivotal (Blum and Langley, 1997). Dimensionality reduction techniques aim to address this problem by reducing the number of features in the data while attempting to retain as much information contained in the data as possible. Dimensionality reduction techniques are commonly divided into feature extraction (e.g. principal component analysis) and feature selection (Liu and Motoda, 2001; Li et al., 2017). The focus in this research work is on feature selection, meaning the selection of a subset of the original features in the data that are deemed relevant according to some criterion (Liu and Setiono, 1996; Bins and Draper, 2001; Liu and Yu, 2005). Feature selection can be implemented following three forms: filter, wrapper, and embedded methods (Liu and Yu, 2005; Chandrashekar and Sahin, 2014). Filter methods are used for pre-processing the data and do not include a learning algorithm to rank features and/or select a feature subset (Blum and Langley, 1997; John et al., 1994). In contrast to that, wrapper and embedded methods use a learning algorithm for the feature subset selection. The former uses the performance (e.g. classification accuracy)

* Corresponding author.

E-mail address: christoph.lohrmann@lut.fi (C. Lohrmann).

as the evaluation criterion for the generated feature subsets (Liu and Yu, 2005; Saeys et al., 2007), whereas the later includes the feature selection as part of the training of the learning algorithm itself (Guyon and Elisseeff, 2003).

Filter methods can be differentiated into (1) those evaluating subsets of features (subset evaluation) and returning a feature subset directly and (2) feature ranking methods (individual evaluation) that rank features according to their relevance and require a decision on how many of the highest-ranking features to select in order to obtain a feature subset (Bolón-Canedo et al., 2013; Yu and Liu, 2004). The focus of this work will be on feature ranking methods for classification (supervised learning). Filter methods, especially feature ranking approaches, offer the advantage that they are often computationally inexpensive, fast and are independent of any classifier since they do not include a classifier to determine a feature ranking or subset (Liu and Setiono, 1996; Saeys et al., 2007). They represent a tradeoff since many feature ranking methods are univariate and, hence, evaluate features separately, which reduces the computational complexity but does not account for potential feature interactions (Kohavi and John, 1997).

Information theory is widely deployed in feature ranking methods, especially mutual information is frequently applied (Vergara and Estévez, 2014). Methods using Shannon's information entropy (Shannon, 1948) are common, but also fuzzy entropy based methods (Lohrmann et al., 2018; Luukka, 2011) can be found in the literature. In this paper, we approach this subject from a nonspecificity, strife and total uncertainty point of view from possibility theory (Klir and Yuan, 1995), which is an extension of our work on nonspecificity from information theory (Luukka and Lohrmann, 2019) and, to our knowledge, has not been done before. Moreover, in our proposed feature selection methods we connect similarity measures and ideal vectors of classes with nonspecificity, strife and total uncertainty for the first time. It is noteworthy that there is no single best feature selection method that performs superiorly with all learning algorithms and data sets (Bolón-Canedo and Alonso-Betanzos, 2018). This can be easily explained by the fact that the goodness of a feature subset is specific to the evaluation function used in the feature selection method (Liu and Yu, 2005; Dash and Liu, 2003, 1997) and that the performance (e.g. accuracy) also depends on how useful a feature subset is for the learning algorithm it is supplied to Kohavi and John (1997) and Caruana and Freitag (1994b). Thus, researchers often aim to find a feature selection method that works well for a specific data set and learning algorithm (Bolón-Canedo and Alonso-Betanzos, 2018) and the introduction of new methods can be useful for certain applications. Apart from that, a recent trend in feature selection is the implementation of ensemble feature selection. These ensembles combine multiple feature selection methods and aggregate their subsets or rankings to obtain more stable and, thus, reliable feature subsets (Bolón-Canedo and Alonso-Betanzos, 2018; Hilario and Kalousis, 2008; Bolón-Canedo et al., 2014a; Lazar et al., 2012; Yang et al., 2010). It is intuitive that using an ensemble with multiple methods instead of just applying a single method alone will only be useful when these methods are accurate but disagree at least in some cases (Opitz, 1999). Thus, in order to be of use for the learning algorithm, the different feature selection methods in the ensemble are supposed to provide diverse subsets of relevant features (Bolón-Canedo and Alonso-Betanzos, 2019). Therefore, introducing feature selection methods such as strife that generate subsets of relevant features which differ from those generated by existing methods, is desirable.

2. Mathematical background

2.1. Nonspecificity, strife and total uncertainty

Nonspecificity is a type of uncertainty which is connected to the sizes (cardinalities) of relevant sets of alternatives. It was first conceived in terms of classical set theory and was defined using the function

$$U(A) = c \times \log_b(\text{Card}(A)) \quad (1)$$

where $\text{Card}(A)$ denotes the cardinality of a finite nonempty set A , whereas b and c are positive constants. If c is equal to 1 and b is set to 2, uncertainty is measured in bits and the corresponding function is called a Hartley function (Hartley, 1928). This uncertainty assumes that larger sets, so sets with a larger cardinality, are less specific than smaller sets, with the most specific set being a set containing only a single element (Klir and Yuan, 1995). A generalization of the Hartley function to measure nonspecificity was introduced in the context of possibility theory, termed the U-uncertainty (Klir and Yuan, 1995). Within possibility theory, ordered possibility distributions are used that represent fuzzy sets and are denoted by $r = (r_1, r_2, \dots, r_n)$ where $1 = r_1 \geq r_2 \geq \dots \geq r_n$ so that each element in r is larger than or equal to its successor (Klir and Yuan, 1995). The U-uncertainty of r is then calculated as

$$U(r) = \sum_{i=2}^n (r_i - r_{i+1}) \times \log_2 i \quad (2)$$

where n is the number of elements in the set. The summation runs until $i = n$ and in this case, according to convention, r_{n+1} is set to 0 (Klir and Yuan, 1995), highlighting that no other outcome than those considered is possible. This formula calculates the differences between successive possibility values and weights them using $\log_2 i$ which is a monotonically increasing weight. Thus, the differences between possibility values that are later in the sequence of the ordered possibility distribution have a higher weight. In case the differences between the ordered possibility values are large, the U-uncertainty will take a large value as well. In particular, the smallest possibility values in the distribution, which could be considered the tail of the ordered possibility distribution, have the largest weights and, hence, a larger contribution to the U-uncertainty. Also, since nonspecificity focuses on the sizes of the set (cardinalities), larger sets tend to have larger U-uncertainty values given that more values are summed, and the weights are monotonically increasing. The minimum of the U-uncertainty is reached with a set of a single tuple so that $U(r) = 0$, indicating the absence of any uncertainty.

Strife is another measure from possibility theory that relies on ordered possibility distributions. In contrast to nonspecificity, which measures to what extent a set of alternatives is not specific, strife expresses conflict or disagreement among alternatives (Klir and Yuan, 1995). It can be regarded as a generalization of the Shannon Entropy and can be expressed with the help of the possibilistic nonspecificity (U-uncertainty) $U(r)$ as Klir and Yuan (1995):

$$S(r) = U(r) - \sum_{i=2}^n (r_i - r_{i+1}) \times \log_2 \sum_{j=1}^i r_j \quad (3)$$

This can be rewritten as

$$S(r) = \sum_{i=2}^n (r_i - r_{i+1}) \times \log_2 \frac{i}{\sum_{j=1}^i r_j} \quad (4)$$

It is apparent that $S(r)$ in this compact written form also utilizes the differences between the ordered possibility values to determine the uncertainty in a fuzzy set. However, the weight of the differences is different since the factor $\log_2 i$ as used for $U(r)$ is replaced by the $\log_2$ of the quotient of i to the sum of the possibility values.

The last measure to discuss in this context is total uncertainty, which is a combination of the two types of uncertainty: nonspecificity and strife (Klir and Yuan, 1995). The total uncertainty NS can be obtained by the addition of the nonspecificity measure and the strife measure as $NS(r) = U(r) + S(r)$. In possibility theory, this measure can be expressed as

$$NS(r) = \sum_{i=2}^n (r_i - r_{i+1}) \times \log_2 \frac{i^2}{\sum_{j=1}^i r_j} \quad (5)$$

where the difference in the equation for total uncertainty $NS(r)$ to strife $S(r)$ lies in the square of i in the weights for the differences in the possibility values. As with nonspecificity and with strife, the range

of values total uncertainty can take lies within $[0, \log_2 |X|]$, where $|X|$ denotes the cardinality of set X (Klir and Yuan, 1995). It is noteworthy that the weights for nonspecificity and total uncertainty tend to be considerably higher than those of strife (see Fig. A.1 in the appendix), also indicating that nonspecificity as the first component of total uncertainty has a larger impact on total uncertainty than its second component strife. For all three approaches the weights are monotonically increasing but nonspecificity and total uncertainty appear to assign relatively low weights to the first possibility values in the ordered possibility distribution, thus placing a stronger emphasis on the middle and tail of the possibility distribution. In contrast, the weights for strife appear to increase close to linear and do not dedicate such relatively low weights to the initial possibility values. Given that the weights associated with the difference in the possibility values constitute the only difference between these approaches, this is a distinguishing characteristic of strife. Finally, the weights of nonspecificity are the only ones that are independent of the actual possibility values since they only contain the index of the possibility values but not the possibility values themselves.

All three measures can be used to compare different ordered possibility distributions and determine the level of uncertainty they represent and order them accordingly to make conclusions on which distribution contains a higher level of uncertainty. This idea will be used in the following in the context of supervised feature selection.

2.2. Similarity

In data handling, many procedures begin with a mutual comparison of the observations in the data. In the simple case, the comparison encompasses obtaining a relation $\mu_R: \mathbb{R} \times \mathbb{R} \rightarrow [0, 1]$. Initially, a mapping of the elements into the unit interval is required and, subsequently, similarity measures conduct a mapping $s: [0, 1] \times [0, 1] \rightarrow [0, 1]$. Several different similarity measures exist (Bandemer and Näther, 1992), but in this work the similarity introduced in Lukasiewicz (1970) is deployed. Hence, the similarity of two elements can be defined as

$$s(x, y) = 1 - |x - y| \quad (6)$$

Where $x, y \in [0, 1]$. This can be extended by following the generalized Łukasiewicz structure as Luukka et al. (2001):

$$s(x, y) = (1 - |x^p - y^p|)^{\frac{1}{p}} \quad (7)$$

where parameter $p \in \mathbb{R}^+$ can be used to adjust the strength of the similarity for the data at hand. The similarity value $s(x, y)$ is within $[0, 1]$ where a value close to '0' indicates that x and y are very dissimilar, a value close to '1' that x and y are very similar and a value of close to '0.5' indicates ambivalence, meaning that it is not clear whether x is more similar or more dissimilar to y .

3. Feature selection using nonspecificity, strife and total uncertainty

Let us assume a data set X , which contains n observations and has D features (= variables). Each observation belongs to one out of N classes ($i = 1, 2, \dots, N$). The number of observations contained in a class i is denoted by n_i .

First Step: Normalize each feature d in the data set separately into the unit interval so that $X^D \rightarrow [0, 1]^D$, where D denotes the number of features in the data set.

Second Step: For each class i in the data set compute a so-called ideal vector that represents that class. The ideal vector for a class i is denoted by v_i , where $v_i = (v_{i,1}, v_{i,2}, \dots, v_{i,D})$ contains for each feature in the class a scalar value that represents the values that the feature takes within that class. In this work the arithmetic mean is used for the calculation of the ideal vector using only observations contained in the training data.

$$v_{i,d} = \left(\frac{1}{n_i} \sum_{x \in X_i} x_d \right), d = 1, 2, \dots, D, i = 1, 2, \dots, N \quad (8)$$

Where $v_{i,d}$ denotes the arithmetic mean of the d th feature for observations belonging to class i . Other functions can also be used to calculate the ideal vector elements, such as the generalized mean, the Bonferroni mean (Bonferroni, 1950) or ordered weighted average (OWA) operators (Yager, 1988).

Third Step: For each feature of each observation, the similarity to the ideal vector element for that feature in each class is calculated. For an observation x_j the similarity of its d th feature to the d th element of the ideal vector of class i can be calculated as

$$s(x_{j,d}, v_{i,d}) = \left(1 - |x_{j,d}^p - v_{i,d}^p| \right)^{\frac{1}{p}} \quad (9)$$

The simple intuition of the formula being that if the value for the d th feature for observation x_j is close to the value of this feature in the ideal vector v_i of class i , then these values are very similar. In other words, the observation is similar to class i in terms of the d th feature. In contrast to that, if these two values differ considerably, then the similarity value will be low, indicating that they are dissimilar. This calculation is conducted for each observation $j = 1, 2, \dots, n$ for each feature $d = 1, 2, \dots, D$ for all classes $i = 1, 2, \dots, N$.

Fourth Step: This step concerns the calculation of the nonspecificity, strife or total uncertainty measure for a feature d and a class i . Hence this step exclusively considers observations of each class and not all observations for the calculation of the uncertainty measure. For the ordered possibility, the ordered similarity value distribution for each class is used that includes the similarities calculated in the third step. The ordered similarity distribution for class i and feature d can be denoted as $s_{i,d} = (s_{1,i,d}, s_{2,i,d}, \dots, s_{n_i,i,d})$ where n_i denotes the number of observations in class i and $s_{1,i,d} \geq s_{2,i,d} \geq \dots \geq s_{n_i,i,d}$ highlights the ordering. It should be remarked that the first number in the subscript of the ordered similarity values refers to the position in the ordered sequence and not to the original number of the observation.

The nonspecificity w.r.t a class i and a feature d using ordered similarity values of class i can be obtained with the help of Eq. (2) as follows:

$$u_{1,i,d} = \sum_{k=2}^{n_i-1} (s_{k,i,d} - s_{k+1,i,d}) \times \log_2 k \quad (10)$$

Where n_i denotes the number of observations in class i , and $s_{k,i,d}$ and $s_{k+1,i,d}$ two successive similarity values (meaning it is the closest smaller or equal similarity value to $s_{k,i,d}$) of the ordered similarity distribution for observations in class i . In addition, the subscript 1 in $u_{1,i,d}$ highlights that this nonspecificity is specific to the class. It is noteworthy that the calculation only includes the summation until $n_i - 1$, so that the last value in the ordered similarity distribution is $s_{n_i,i,d}$ and not $s_{n_i+1,i,d}$. In other words, it does not follow the convention mentioned in Section 2.1 that the last value considered is zero at the end of the ordered sequence (referred to as r_{n+1} previously). The reason for this is simple: it is conceptually intuitive for an ordered possibility distribution to assume that the next value outside the set takes a value of zero, meaning that any value not specified in the possibility distribution is not obtainable. However, the same logic does not apply for ordered similarity values. In addition, following this convention can increase the value of the uncertainty measure manifold just based on the difference between the last actually measured smallest similarity and a hypothetical similarity of zero — as also practical examples conducted by the authors of this article have indicated.

A similar calculation as presented in Eq. (10) is required in case the strife measure or the total uncertainty measure is used. The strife measure w.r.t a class i and a feature d using ordered similarity values of class i can be computed as

$$u_{1,i,d} = \sum_{k=2}^{n_i-1} (s_{k,i,d} - s_{k+1,i,d}) \times \log_2 \frac{k}{\sum_{l=1}^k s_{l,i,d}} \quad (11)$$

And for total uncertainty in the same context as

$$u_{1,i,d} = \sum_{k=2}^{n_i-1} (s_{k,i,d} - s_{k+1,i,d}) \times \log_2 \frac{k^2}{\sum_{l=1}^k s_{l,i,d}} \quad (12)$$

Depending on the selected uncertainty measure, either Eq. (10) for nonspecificity, Eq. (11) for strife or Eq. (12) for total uncertainty should be used. This step is repeated for each of the N classes and each of the D features.

Fifth Step: This step concerns the calculation of the nonspecificity, strife or total uncertainty measure for a feature d for all observations in the data. This is clearly distinct from the fourth step where for each measure only the observations within a specific class were incorporated into the calculation. Hence, this step concerns the ordered similarity distribution of the similarity values of all observations for feature d with respect to the ideal vector of a class i . The formula for nonspecificity, strife and total uncertainty can be represented as

$$u_{2,i,d} = \sum_{k=2}^{n-1} (s_{k,i,d} - s_{k+1,i,d}) \times \log_2 k \quad (13)$$

$$u_{2,i,d} = \sum_{k=2}^{n-1} (s_{k,i,d} - s_{k+1,i,d}) \times \log_2 \frac{k}{\sum_{l=1}^k s_{l,i,d}} \quad (14)$$

$$u_{2,i,d} = \sum_{k=2}^{n-1} (s_{k,i,d} - s_{k+1,i,d}) \times \log_2 \frac{k^2}{\sum_{l=1}^k s_{l,i,d}} \quad (15)$$

Where the summation of the weighted differences in the similarity values is implemented over all observations in the data and not, as in Eq. (10)–(12), only over the observations in a specific class. The subscript 2 in $u_{2,i,d}$ highlights that this nonspecificity covers the similarity values of all observations in the data.

It is obvious, that the choice of uncertainty measure (nonspecificity, strife, total uncertainty) in the fifth step is supposed to correspond to the selection of the uncertainty measure in the fourth step. This step is repeated for each of the N classes and each of the D features.

Sixth Step: This step contains the calculation of the information transmission value T_d for the d th feature. This step incorporates the nonspecificity, strife or total uncertainty values calculated in the previous fourth and fifth step. In particular, the information transmission for a feature is measured as the difference between the uncertainty of the feature values of all observations to each class's ideal vector element and the uncertainty of the feature values for observations in a class to the corresponding class's ideal vector element. This can be formulated as follows

$$T_d = \sum_{i=1}^N u_{2,i,d} - u_{1,i,d} \quad (16)$$

where $u_{2,i,d}$ denotes the uncertainty for the d th feature in the entire data set using the ideal vector of class i and $u_{1,i,d}$ the uncertainty for the d th feature only within class i . The information transmission T_d is then the sum of the differences in uncertainty between the feature in the entire data set and that in each of the N classes. This step is repeated for each of the D features and results in D feature-specific information transmission values. A high information transmission value for a feature is essentially indicative of the ideal vector element for a class being considerably better at representing the observations in that class than at representing all observations.

Seventh Step: The features can be ranked according to their feature-specific information transmission value in descending order. A feature is relevant for classification if the knowledge of the class label reduces the uncertainty for the values the feature takes. This is represented by a large difference between the uncertainty in the entire data compared to the uncertainty in each of the classes, which means a high information transmission value obtained from Eq. (16).

Finally, a certain number of the highest-ranking features, meaning with the largest information transmission values, needs to be selected.

There are two common forms how feature ranking methods are used to generate feature subsets. The first is selecting a fixed subset of the highest-ranking features using a threshold value for the feature importance or selecting a user-specified number $k < D$ of the highest-ranking features (Bolón-Canedo et al., 2013; Das, 2001). In the second approach, the feature ranking is followed by a wrapper approach where the number of highest-ranking features to select is decided using cross-validation (Liu and Yu, 2005; Duch, 2006), which can be referred to as a hybrid feature selection approach since it combines two feature selection methods (feature ranking and a wrapper) sequentially (Bolón-Canedo et al., 2014b).

Both approaches to using a feature ranking will result in selecting a feature subset and discarding the remaining features to reduce the dimensionality of the data set for the subsequent learning algorithm.

4. Data and cross-validation procedure

There are overall 18 data sets used as part of this study: eight artificial data sets and ten real-world data sets. The details for each data set are depicted in Table 1 (the majority class(es) for each data set is highlighted in bold).

Five artificial cases are based on three artificial examples and two extensions of them. These first three examples are two-dimensional and displayed in Fig. 1.

The first example contains two classes, which each have a single decision region and two independent features. The second example is an extension of the first example with two decision regions for each class and a different level of variance of the data points for the point clusters. The third example is a three-class problem with a single decision region per class and different covariance matrices and sample sizes. For details on the distribution and specifications, please see Table A.1 in the Appendix. A detailed description how nonspecificity, strife and total uncertainty work on these three artificial cases is provided in the appendix.

The fourth artificial example is an extension of the second artificial example with 100 additional random uniform features, whereas the fifth artificial example is an extension of the third artificial example with 100 additional random uniform features. The remaining three artificial data sets are the three MONK's problems data sets (Thrun, 1992) from the UCI Machine Learning Repository (Dua and Graff, 2019). In contrast to the previous five data sets, the features in the MONK's problems are discrete, integer-valued and the target concept is based on logical rules. There are seven features in the data set (a_1 to a_7) and for MONK-1 the target concept is ($a_1 = a_2$) or ($a_5 = 1$), for MONK-2 exactly two out of the first six features need to be 1, and for MONK-3 either ($a_5 = 3$ & $a_4 = 1$) or ($a_5 \neq 4$ & $a_2 \neq 3$) with 5% of class noise (Thrun, 1992).

The real-world data sets that are utilized in this paper are also taken from the UCI Machine Learning Repository (Dua and Graff, 2019) and from the GitHub repository (Primekangkang, 2019). The authors selected 10 medical data sets that differ considerably in the number of observations, features, classes as well as the ratio of observations per feature. So-called microarray data sets which are commonly found in the medical context (especially gene selection) are characterized by a small number of observations and a considerably larger number of features. These characteristics make them challenging data sets for learning that often require some form of dimensionality reduction, which is the reason they were selected for this study. In contrast, the remaining medical data sets possess more observations and are considerably lower-dimensional, increasing the variety and characteristics of data sets used in this study. The six data sets obtained from the UCI Machine learning repository are: the cryotherapy data (Khozeimeh et al., 2018), the diabetic retinopathy Debrecen data (Antal and Hajdu, 2014), the cardiocardiography data as a ten-class and as a three-class problem (Marques de Sá et al., 2010), the chronic kidney disease data (Soundarapandian and Rubini, 2015) and the dermatology data

Table 1
Summary of the data sets

Artificial Data	Features	Samples	Classes	1	2	3	4	5	6	7	8	9	10
Artificial Case 1	2	1200	2	50 %	50 %								
Artificial Case 2	2	1800	2	50 %	50 %								
Artificial Case 3	2	1500	3	20 %	40 %	40 %							
Artificial Case 4	102	1800	2	50 %	50 %								
Artificial Case 5	102	1500	3	20 %	40 %	40 %							
Monk 1	7	124	2	50 %	50 %								
Monk 2	7	169	2	62 %	38 %								
Monk 3	7	122	2	51 %	49 %								
Real-world Data	Features	Samples	Classes	1	2	3	4	5	6	7	8	9	10
Cryotherapy	6	90	2	47 %	53 %								
Diabetic Retinopathy	19	1151	2	47 %	53 %								
Cardiotocography-1	22	2126	10	18 %	27 %	2 %	4 %	3 %	16 %	12 %	5 %	3 %	9 %
Cardiotocography-2	22	2126	3	78 %	14 %	8 %							
Chronic Kidney	24	156	2	74 %	26 %								
Dermatology	34	358	6	31 %	17 %	20 %	13 %	13 %	6 %				
SRBCT	2308	83	4	35 %	30 %	13 %	22 %						
Leukemia1	5327	72	3	53 %	13 %	35 %							
DLBCL	5469	77	2	75 %	25 %								
Prostate Cancer	10509	102	2	51 %	49 %								

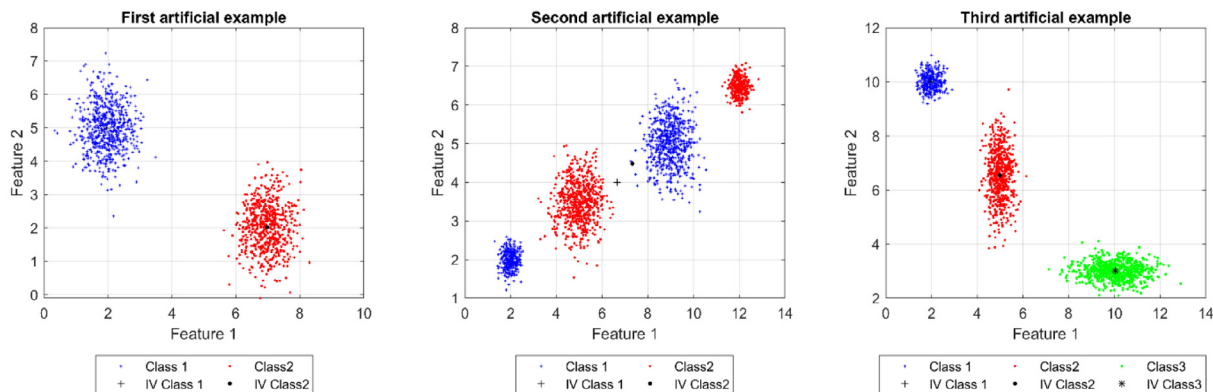


Fig. 1. First three artificial examples.

set (Ilter and Guvenir, 1998). The four data sets obtained from GitHub contain: the small-round-blue-cell tumor (SRBCT), the leukemia1, the diffuse large B-cell lymphoma (DLBCL), and the prostate cancer data set.

In terms of features, the smallest real-world data set is with only six features the cryotherapy data set, whereas the largest data set is the prostate cancer data with more than 10'000 features. The number of observations also differ considerably between the data sets with the leukemia1 data set encompassing only 72 observations with a considerable number of features, while the cardiotocography data sets contain more than 2000 observations with only 22 features. This also highlights that this selection of data sets contains data with a high number of observations to the number of features as well as data such as the prostate cancer data that have considerably more features than observations — which makes the model fitting on this type of data more susceptible to overfitting (curse of dimensionality). The number

of classes range from two, such as for the diabetic retinopathy data set, up to ten classes for one of the cardiotocography data sets. With respect to the class balance, meaning how equally the observations are distributed among classes, the prostate cancer data set reflects the best balance of all data sets with a division that is close to 50–50, whereas the cardiotocography data sets belong to the most imbalanced ones with the minority classes being up to 10 times smaller than the largest class in the respective data set. Overall, the varying characteristics of these data sets support the investigation whether the feature ranking methods suggested in this paper can perform well on different types of data sets.

4.1. Cross-validation

Several authors pointed out that using the entire data set for feature selection and, subsequently, deploying a cross-validation procedure for

the classifier training leads to a bias with respect to the accuracy on the test set (= 1 - generalization error). This bias, which is the difference between the true generalization error and the estimated generalization error, is often referred to as ‘feature selection bias’ or ‘selection bias’ (Miller, 2002; Singhi and Liu, 2006). The reason for a too optimistic estimate of the generalization error is the fact that the selection of the feature subset includes data that was later used for the calculation of the error on the supposedly independent test set (Ambroise and McLachlan, 2002). This bias may be large for data sets with a small set of observations and a large number of features, as is the case for microarray data for gene analysis (Singhi and Liu, 2006; Krawczuk and Łukaszuk, 2016).

Hence, Ambroise and McLachlan (2002) conclude that the test set should not be part of the feature selection process. Such external cross-validation, meaning that the test set is independent of the feature selection process, is required to obtain unbiased results on the test set (Ambroise and McLachlan, 2002; Krawczuk and Łukaszuk, 2016). However, it is pivotal to stress that a single external cross-validation may not be sufficient for additional parameter tuning/ optimization of a classifier together with feature selection. Similar to Wood et al. (2007), this study will deploy a two-level external cross-validation to address the ‘optimization bias’, which is sometimes also termed ‘parameter selection bias’ (Varma and Simon, 2006), as well as the ‘selection bias’.

The first external cross-validation is implemented using 10-fold cross-validation, as recommended by Ambroise and McLachlan (2002), and the second external cross-validation (termed “nested CV”) is based on a simple holdout data split (80% training set, 20% validation set). Each of the two cross-validations serve a different purpose. The outer external CV is used to set aside an independent test set to measure the generalization error of the final model with data that has not been used for feature selection, parameter selection and model training. The second external CV is deployed for model selection, meaning feature selection (and parameter selection) for the learning algorithm (e.g. classifier).

In this study, we use the two-level CV and follow two different approaches to conduct feature selection based on a feature ranking: (1) using a fixed number of the highest-ranking features and (2) using a hybrid approach where the feature ranking is inputted to a wrapper method to cross-validate the number of highest-ranking features to select.

For the first approach, the simple rules applied in the study of Bolón-Canedo et al. (2013) are used which state that depending on the number of features D , the following percentage of highest-ranking features can be selected: if $D < 10$ then 75%, if $10 \leq D < 75$ then 40%, if $75 \leq D < 100$ then 10%, and if $100 \leq D$ then 3%. The only exceptions in our study are the five artificial cases for which only one feature is relevant and, thus, only the highest-ranking feature is selected. Thus, following this first approach, for each data set the number of highest-ranking features selected is fixed and is the same for all filter methods used. The two-level cross-validation used with a fixed number of highest-ranking features is illustrated in Fig. 2, the first and second data split are highlighted in blue and green, and the step in which the feature subset is selected is highlighted in gray.

The feature ranking methods are applied to the (internal) training set and a fixed number of the highest-ranking features is selected for each method. The learning algorithm (here: classification algorithm) is trained on the same (internal) training data multiple times — each time with a different parameter value. Subsequently, the best parameter(s) is selected based on the highest validation performance (model selection). The (external) training data is used to train the learning algorithm with the previously obtained feature subsets and the corresponding best model parameter(s), and the final model obtained after training is then tested on the independent test set (model evaluation) to estimate the generalization error.

For the second approach, the number of highest-ranking features is selected using the validation data set. In particular, the feature ranking

is used as a starting point and the lowest-ranking feature is removed from the set of features in each iteration of the wrapper method (= backward elimination). Finally, the highest classification accuracy of the learning algorithm (in the wrapper) on the validation data set is used to select the best number of the highest-ranking features. The cross-validation procedure for this hybrid approach is presented in Fig. 3 and the steps in which the feature subset is selected are highlighted in gray.

The essential difference of this second approach to the first one is that the validation data is not only deployed to determine the optimal parameters (for a learning algorithm on a feature subset of fixed size) but simultaneously to select the best number of highest-ranking features for the subset as well. After the optimal combination of the feature subset and parameter set is determined based on the highest validation set accuracy in the nested CV, this combination is utilized to train the model on the training set of the outer external CV and test the generalization performance on the independent test set (external CV). The result obtained from each loop of the outer external CV can then be averaged over all 10 folds to obtain an estimate of the generalization error of the feature selection process together with the specific training algorithm used.

The classifiers applied in this study are a k-nearest neighbor (KNN) classifier (Cover and Hart, 1967), a binary decision tree (Breiman et al., 1984), a similarity classifier (Luukka et al., 2001), and a support vector machine (SVM) (Cortes and Vapnik, 1995). Due to the computational complexity of using two-level external cross-validation and training the classification models for each feature subset size for each fold in case of the hybrid feature selection approach, only a linear SVM was used and the similarity classifier is deployed solely with standard parameters ($p = 1, m = 1$). For the k-nearest neighbor classifier the number of neighbors k and for the decision tree the minimum leaf size are tuneable parameters for the parameter selection in this study. The average classification accuracies accomplished on the test set (external CV) for each feature selection method are compared in terms of the significance of the results against using no feature selection by using a one-sided Welch’s test.

4.2. Stability of a feature selection algorithm

Using an external cross-validation procedure for feature selection can lead to different feature rankings and subsets since different partitions of the same data set are used in each iteration to determine the feature subsets (Ambroise and McLachlan, 2002; Krawczuk and Łukaszuk, 2016). However, it is clearly desirable that a feature selection method gives the same or a comparable feature subset when applying it to different partitions of the same data set. Hence, the stability of a feature selection algorithm, meaning its ability to generate similar feature subsets from large samples of the same data set, is a crucial aspect of feature selection (Krawczuk and Łukaszuk, 2016). Similar to Krawczuk and Łukaszuk (2016), this study uses the adjusted stability measure (ASM) conceptualized by Lustgarten et al. (2009) to evaluate the stability of a feature selection algorithm based on the feature subsets obtained from the nested external cross-validation. The ASM can be calculated as follows (Lustgarten et al., 2009):

$$ASM = \frac{2}{c * (c - 1)} \sum_{i=1}^{c-1} \sum_{j=i+1}^c S_A(f_{s_i}, f_{s_j}) \quad (17)$$

Where c denotes the number of feature subsets (here equal to the number of folds) and S_A the similarity measure between two feature subsets f_{s_i} and f_{s_j} . The double sum is used to sum up the similarity measure for all combinations of feature subsets, not considering permutations of them (e.g. feature subset one is compared with feature subset two, but, subsequently, feature subset two does not need to be compared to feature subset one again). The multiplication of the sum with the term $\frac{2}{c * (c - 1)}$ averages the summed similarity measures with the number of similarity measures in the sum. For instance, if $c = 10$, meaning there

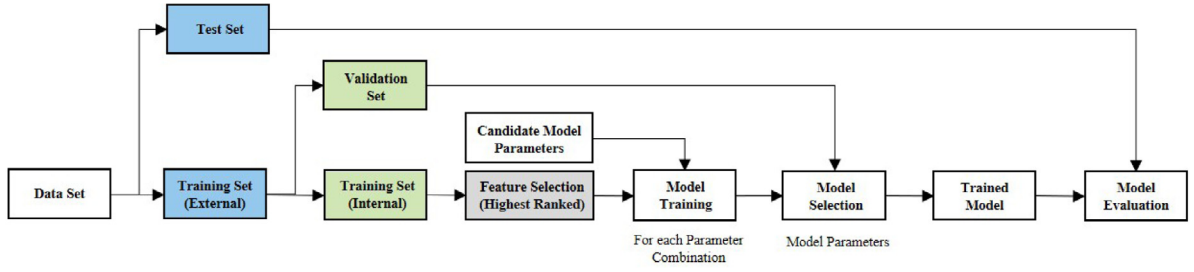


Fig. 2. Two-level cross-validation for parameter selection with a fixed feature subset of the highest-ranking features.

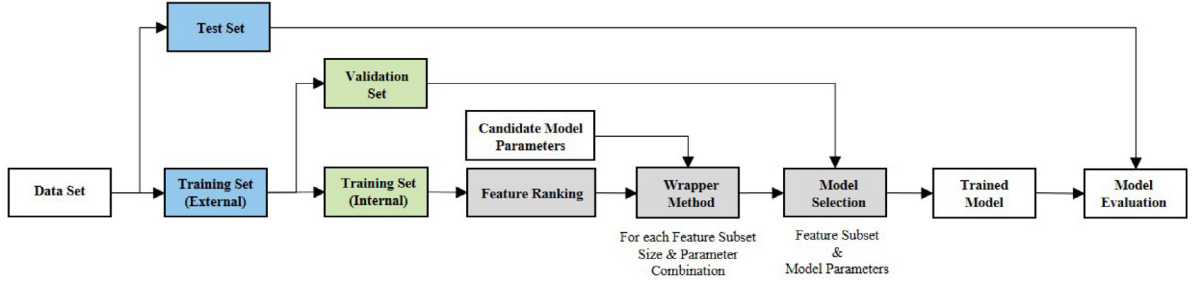


Fig. 3. Two-level cross-validation for feature subset and parameter selection using a hybrid feature selection approach.

are 10 feature subsets, technically $c * c$ feature subset comparisons are possible if every subset is compared with every subset. However, if the comparisons of each feature subset with itself is excluded, this number reduces to $c * (c - 1)$. If permutations are excluded, only $\frac{c*(c-1)}{2}$ combinations remain, which is the exact number the sum of the similarity measures is divided by. Hence, the ASM is an average of the similarity between all combinations of the feature subsets.

The similarity measure S_A for two feature subsets fs_i and fs_j , which originated in the idea by Kuncheva (2007) to account for the effect of chance, is calculated as

$$S_A(fs_i, fs_j) = \frac{r - \frac{k_i * k_j}{D}}{\min(k_i, k_j) - \max(0, k_i + k_j - D)} \quad (18)$$

In this formula r is the cardinality of the intersection of the two feature sets fs_i and fs_j , meaning how many features they have in common. The quotient subtracted from r in the numerator represents the expected value for r , meaning the number of features two subsets fs_i and fs_j with cardinality k_i and k_j for a data set with D features are expected to have in common (Lustgarten et al., 2009). Hence, the numerator is the difference between the actual number of features two subsets have in common minus the number of features these two subsets are expected to have in common. The denominator represents the range of values these two subsets can have in common. In particular, the first term represents the largest number of features the two subsets can have in common, whereas the second term, that is subtracted from it, embodies the smallest possible number of mutual features in these two subsets. The similarity measure S_A can take values within the range $(-1, 1]$ (Krawczuk and Łukaszuk, 2016) and since ASM is an average of all similarity values, it has the same range. For further details, please see Lustgarten et al. (2009).

4.3. Benchmark feature selection methods

The classification performance of the three proposed supervised filter methods are compared to five other supervised filter methods

as well as to no feature selection, meaning to simply use the entire feature set for classification. The five feature selection methods are the well-known Relief algorithm, the fuzzy entropy and similarity (FES) filter (Luukka, 2011), the fuzzy similarity and entropy (FSAE) filter, the Fisher score, and symmetrical uncertainty.

ReliefF is an extension of Relief, a supervised filter for feature selection introduced by Kira and Rendell (1992a), that also works on incomplete and noisy data (Kononenko et al., 1997). It determines the relevance of a feature by using the k Near-hit and Near-miss observations, meaning the closest observations from the same and other classes to each considered observation (Kononenko et al., 1997; Kira and Rendell, 1992b). In particular, the weights for a feature are initialized at zero and are adjusted based on the average difference of particular observations to the corresponding k Near-hit and Near-miss instances (Robnik-Šikonja and Kononenko, 2003):

$$W_{d,new} = W_{d,old} - \sum_{j=1}^k \frac{diff(d, x_i, h_j)}{m * k} + \sum_{C \neq class(x_i)}^{n-1} \left[\frac{P(C)}{1 - P(class(x_i))} \sum_{j=1}^k \frac{diff(d, x_i, m_j(C))}{m * k} \right] \quad (19)$$

Where W_d denotes the weight for feature d , m is the number of observations considered, $P(class(x_i))$ denotes the probability of an observation to be in the class that also x_i belongs to, whereas $P(C)$ is the probability of each different class C . In addition, $diff(d, x_i, h_j)$ denotes the difference between observation x_i and the Near-hit h_j in terms of the d th feature, and $diff(d, x_i, m_j(C))$ represents the difference between observation x_i and the Near-miss $m_j(C)$ in terms of the d th feature from a different class C . For numerical features, this difference is calculated as

$$diff(d, x_i, x_j) = \frac{|x_{i,d} - x_{j,d}|}{\max(x_d) - \min(x_d)} \quad (20)$$

Whereas for categorical features, it can be determined as

$$\text{diff}(d, x_i, x_j) = \begin{cases} 0 & \text{if } x_{i,d} = x_{j,d} \\ 1 & \text{otherwise} \end{cases} \quad (21)$$

The *diff* function is also used to find the Near-hit and Near-miss instances, by summing up the *diff* values for each instance over all features (Manhattan distance) (Robnik-Šikonja and Kononenko, 2003). The weight $W_d \in [-1, 1]$ that is obtained for each feature represents the feature relevance, so that the features can be ranked according to this weight in descending order and a number of user specified features can be selected (Kira and Rendell, 1992b; Souza et al., 2006). In this work, ReliefF will be deployed with 10 Near-hits and Near-misses, which is a common choice (Kononenko et al., 1997; Kononenko, 1994). In addition, this study uses the modification introduced by Robnik-Šikonja & Kononenko (Robnik-Šikonja and Kononenko, 2003) with 70 Near-hit and Near-miss instances and a new decay parameter $\sigma = 20$.

Fuzzy entropy and similarity (FES), was in previous works referred to as “Feature selection by Luukka (2011)” (Lohrmann et al., 2018), is a filter method that uses similarity and fuzzy entropy measures to determine the relevance for each feature. The approach first calculates for each class i an ideal vector v_i using, for instance, a generalized mean (Luukka, 2011). Subsequently, for each observation the similarity of the observation’s feature value with each class ideal vector element is determined. For the value of feature d for observation j , denoted $x_{j,d}$, the similarity with the corresponding element of the ideal vector of class i , denoted $v_{i,d}$, can be calculated as

$$S(x_{j,d}, v_{i,d}) = \sqrt[p]{1 - |x_{j,d}^p - v_{i,d}^p|} \quad (22)$$

Where p comes from the generalized Łukasiewicz structure (Luukka et al., 2001). Subsequently, to determine the summed entropy value for each feature, the similarity values for each feature are summed over all observations n and all classes N .

$$H_d = \sum_{i=1}^N \sum_{j=1}^n H(S(x_{j,d}, v_{i,d})) \quad (23)$$

The features can be ranked according to the summed entropy values in ascending order so, for instance, a user-specified number of features with the lowest entropy values can be selected.

Fuzzy similarity and entropy (FSAE) feature selection is a modification of the filter feature selection approach FES introduced by Luukka (2011). As in FES, ideal vectors are calculated for each class and the similarity is measured for each feature for each observation with each corresponding class ideal vector element. In contrast to Eq. (23), the similarity values for a given feature in the FSAE are only summed over all observations but not over all classes (Lohrmann et al., 2018):

$$H_{i,d} = \sum_{j=1}^n H(S(x_{j,d}, v_{i,d})) \quad (24)$$

In addition to this feature- and class-specific summed entropy $H_{i,d}$, a scaling factor is calculated in order to account for the difference between ideal vector elements of different classes.

$$SF_{i,d} = 1 - \frac{\left(\sum_{i \neq j} |v_{i,d} - v_{j,d}|^t \right)^{\frac{1}{t}}}{N - 1} \quad (25)$$

The summed entropy values are then multiplied with the corresponding feature- and class-specific scaling factor and summed over all N classes:

$$SE_d = \sum_{i=1}^N (H_{i,d} * SF_{i,d}) \quad (26)$$

As with the entropy values obtained by the approach of Luukka, the scaled entropy values for the FSAE are ordered in ascending order and,

for instance, a user-specified number of features with the lowest scaled entropy values is selected.

Symmetrical uncertainty is another measure that is rooted in information theory, that was introduced by Press et al. (1988) and essentially embodies a normalized version of mutual information (Sosa-Cabrera et al., 2019). For a categorical random feature x that can take I possible values $\{x_1, x_2, \dots, x_I\}$, the entropy as a measure of uncertainty can be expressed as

$$H(x) = - \sum_i^I P(x_i) \log_2(P(x_i)) \quad (27)$$

Where x_i represents one specific value the feature x can take and $P(x_i)$ the prior probability that x takes the value x_i . In addition to the entropy for a single feature x , the conditional entropy of a feature x given another feature y with possible values $\{y_1, y_2, \dots, y_J\}$ can be expressed as

$$H(x|y) = - \sum_j^J \left(P(y_j) \sum_i^I P(x_i|y_j) \log_2(P(x_i|y_j)) \right) \quad (28)$$

Where $P(x_i|y_j)$ is the posterior probability of observing value x_i for feature x given that feature y has a value of y_j . Combining the entropy of x with the conditional entropy of x given y allows to measure the reduction in uncertainty of the values feature x takes if the value of y is known (Sosa-Cabrera et al., 2019). This quantity is referred to as the mutual information and is calculated as

$$MI(x|y) = H(x) - H(x|y) \quad (29)$$

The symmetrical uncertainty is the mutual information normalized by the entropy values of feature x and feature y in order not to be biased towards features with many categories (Sosa-Cabrera et al., 2019).

$$SU(x, y) = 2 \left(\frac{MI(x|y)}{H(x) + H(y)} \right) \quad (30)$$

The symmetrical uncertainty $\epsilon [0, 1]$ represents the association between two features. In case the symmetrical uncertainty is used to quantify the uncertainty between the dependent variable (= class labels) and each of the features, these values can be ordered in descending order and, for instance, a user-specified number of the features with the highest symmetrical uncertainty values can be selected.

Fisher Score is a similarity-based filter method for supervised feature selection (Duda et al., 2012). It is based on the distance between the class means of a feature to the mean of the feature, standardized by the observation-weighted standard deviation for that feature in the classes. It can be defined as:

$$FisherScore_d = \frac{\sum_{i=1}^N n_i (\bar{x}_{i,d} - \bar{x}_d)^2}{\sum_{i=1}^N n_i \sigma_{i,d}} \quad (31)$$

Where d is a feature, $\bar{x}_{i,d}$ is the mean of feature d in class i , $\bar{x}_d$ is the mean of feature d , $\sigma_{i,d}$ is the standard deviation of feature d in class i , N is the number of classes and n_i is the number of observations in class i . The Fisher score is calculated for each feature and then ordered in descending order. Subsequently, for instance, a user-specified number of features with the largest Fisher score can be selected for the feature subset (Li et al., 2017).

All feature selection methods and calculations were implemented with MATLAB™ - software (version 2020a). The code for the symmetrical uncertainty was obtained from the “Feature Selection Library” from the MathWorks file exchange (Roffo et al., 2017). The code files for the feature ranking methods proposed in this study and for the artificial cases are freely available at the GitHub repository (Lohrmann and Luukka, 2020).

5. Experimental results and analysis

The results of the filter feature selection algorithms discussed in the previous section and obtained using a k-nearest neighbor (KNN) classifier, a decision tree for classification, a similarity classifier and a linear support vector machine (SVM) are depicted below.

The results with each classifier using the feature ranking with a fixed number of features selected are presented in Tables 2 to 5 (the best results per classifier are highlighted in gray; ‘.’, ‘*’, ‘***’, and ‘****’, refer to the 10%, 5%, 1% and 0.1% significance level corresponding to Welch’s test of a filter method versus using no feature selection, respectively). It is apparent that for the first simple artificial example all feature selection methods consistently find the single most relevant feature in the data set. For the KNN classifier (Table 2) the second

artificial case and its extension, the fourth artificial case, can only consistently produce perfect accuracies using the ReliefF setups and symmetrical uncertainty. However, it is noteworthy that strife, non-specificity and total uncertainty only offer a slightly worse performance whereas the Fisher score, FSAE and FES provide a roughly 10%-points lower accuracy. In contrast to that, for the third artificial case and its extension, the fifth artificial case, the three approaches proposed in this paper (strife, nonspecificity and total uncertainty) as well as the Fisher score and symmetrical uncertainty consistently rank the most relevant feature first. It is worth mentioning that the KNN classifier as a distance-based classifier is very susceptible to many additional irrelevant features as can be seen by the deterioration of the classification accuracy without feature selection in the fourth and fifth artificial example case. On the MONK’s data sets the performance of

Table 2

Average test set accuracy and standard deviation for the KNN classifier with fixed feature subset (Filter).

KNN (Filter)	ReliefF (10)	ReliefF (70,20)	FES	FSAE	Fisher Score	Nonspecificity	Strife	TU	SU	No FS
Artificial Case 1	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
Artificial Case 2	100 ± 0	100 ± 0	90.7 ± 1.9 ***	90.7 ± 1.9 ***	90.7 ± 1.9 ***	98.4 ± 3.2 .	97.4 ± 4 *	98.4 ± 3.2 .	100 ± 0	100 ± 0
Artificial Case 3	99.7 ± 0.4 *	99.7 ± 0.4 *	99.9 ± 0.2	99.6 ± 0.4 **	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
Artificial Case 4	100 ± 0 ***	100 ± 0 ***	90.3 ± 2 ***	90.3 ± 2 ***	90.3 ± 2 ***	98.6 ± 4.2 ***	97.7 ± 4.8 ***	98.6 ± 4.2 ***	100 ± 0 ***	48.6 ± 3.1
Artificial Case 5	99.5 ± 0.4 ***	99.5 ± 0.4 ***	99.9 ± 0.3 ***	99.5 ± 0.4 ***	100 ± 0 ***	100 ± 0 ***	100 ± 0 ***	100 ± 0 ***	100 ± 0 ***	68.3 ± 3.9
Monk 1	92.8 ± 6.7 ***	91.2 ± 10.2 ***	91.3 ± 10.7 ***	91.3 ± 10.7 ***	70.1 ± 23.1	91.3 ± 10.7 ***	69.1 ± 13.9	91.3 ± 10.7 ***	91.3 ± 10.7 ***	67.6 ± 13
Monk 2	61.5 ± 8.9 *	62.7 ± 8.1 *	56.2 ± 5.6 **	56.2 ± 5.6 **	62.8 ± 7.2 *	56.2 ± 5.6 **	61 ± 8.3 *	56.2 ± 5.6 **	56.2 ± 5.6 **	71.1 ± 11.8
Monk 3	83.7 ± 12.8	85.3 ± 11.5	89.5 ± 8 *	89.5 ± 8 *	81.9 ± 17.4	89.5 ± 8 *	49 ± 11.1 ***	89.5 ± 8 *	89.5 ± 8 *	82.9 ± 8.3
Cryotherapy	85.6 ± 7.1	84.4 ± 12.4	86.7 ± 9.7	85.6 ± 11.2	85.6 ± 13.2	87.8 ± 7.8	84.4 ± 10.2	87.8 ± 7.8	86.7 ± 12	88.9 ± 8.6
Retinopathy	64.8 ± 5	66 ± 4.2 *	54.5 ± 3.3 ***	54.5 ± 3.3 ***	67.8 ± 4.2 **	59.1 ± 4.9 *	63 ± 4.7	58.6 ± 5 *	67.6 ± 6.4 *	62.6 ± 3.1
CTG-1	78.6 ± 2.6	78.5 ± 2.6	61.7 ± 3.7 ***	63.6 ± 2.6 ***	75.6 ± 2.8 *	74.1 ± 1.8 ***	79.4 ± 3.3	74.1 ± 1.8 ***	74.3 ± 3.8 **	77.9 ± 1.9
CTG-2	89.7 ± 1.7 .	90.1 ± 1.7	87.5 ± 2.1 ***	87.9 ± 2.1 **	92.5 ± 1.6 *	86.7 ± 2.4 ***	92.1 ± 1.6 .	86.4 ± 2 ***	89.9 ± 2.5	90.9 ± 1.8
Kidney	99.4 ± 1.9	99.4 ± 1.9	98.1 ± 2.9	98.8 ± 2.5	97.5 ± 4.2	96.2 ± 4.2	98.1 ± 2.9	96.2 ± 4.2	99.4 ± 1.9	96.8 ± 5.8
Dermatology	81.5 ± 4.6 ***	81.5 ± 4.6 ***	82.1 ± 4.5 ***	85.5 ± 3.2 ***	84.4 ± 2.2 ***	84.3 ± 3.6 ***	85.5 ± 3.2 ***	84.3 ± 3.6 ***	87.2 ± 5.2 ***	96.9 ± 3.6
SRBCT	100 ± 0 ***	100 ± 0 ***	76.4 ± 15 .	89.3 ± 8.3	100 ± 0 ***	91.8 ± 10.2 .	98.8 ± 3.8 ***	93.1 ± 9.4 *	91.5 ± 7.9 .	85.4 ± 9.1
Leukemia1	93 ± 7 **	94.5 ± 6.8 **	62 ± 14.9 **	81.8 ± 10.9	94.5 ± 6.8 **	88.8 ± 8.5 .	95.9 ± 6.3 **	91.6 ± 6.9 *	94.5 ± 6.8 **	82.1 ± 10.1
DLBCL	96.3 ± 8 *	92.3 ± 10.1	75.2 ± 7.1 **	77.9 ± 8.3 *	93.8 ± 8.4 .	78 ± 12 .	94.8 ± 8.5 .	79.6 ± 11.2 .	89.8 ± 13.5	87.3 ± 12.5
Prostate	89.2 ± 8.3	87.2 ± 6.2	66.4 ± 9.6 **	60.9 ± 9.2 ***	90.1 ± 6.3	75.5 ± 8.1 *	89.1 ± 5.4	77.4 ± 10	86.2 ± 10.3	84.2 ± 12.9

Table 3

Average test set accuracy and standard deviation for the decision tree classifier with fixed feature subset (Filter).

DT (Filter)	ReliefF (10)	ReliefF (70,20)	FES	FSAE	Fisher Score	Nonspecificity	Strife	TU	SU	No FS
Artificial Case 1	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
Artificial Case 2	100 ± 0	100 ± 0	90.8 ± 2.2 ***	90.8 ± 2.2 ***	90.8 ± 2.2 ***	98.7 ± 2.6 .	97.7 ± 3.6 *	98.7 ± 2.6 .	100 ± 0	100 ± 0
Artificial Case 3	99.5 ± 0.8 *	99.5 ± 0.8 *	99.8 ± 0.6	99.3 ± 0.7 **	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
Artificial Case 4	100 ± 0	100 ± 0	90.4 ± 2.3 ***	90.4 ± 2.3 ***	90.4 ± 2.3 ***	98.6 ± 4.2	97.7 ± 4.8 .	98.6 ± 4.2	100 ± 0	100 ± 0
Artificial Case 5	99.5 ± 0.6 *	99.5 ± 0.6 *	99.9 ± 0.2	99.5 ± 0.6 *	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
Monk 1	81.6 ± 11	82.4 ± 12	79 ± 11.2	79 ± 11.2	68.6 ± 16.4 *	79 ± 11.2	67.4 ± 13.9 **	79 ± 11.2	79 ± 11.2	82.4 ± 11.5
Monk 2	61.5 ± 11	62.1 ± 12.5	54.3 ± 10.1	53.2 ± 8.7 .	59.2 ± 8.7	54.3 ± 10.1	66.9 ± 10.9	54.3 ± 10.1	53.8 ± 10.3	60.3 ± 14
Monk 3	93.5 ± 6.2	93.5 ± 6.2	93.5 ± 6.2	93.5 ± 6.2	93.5 ± 6.2	93.5 ± 6.2	54 ± 10.1 ***	93.5 ± 6.2	93.5 ± 6.2	93.5 ± 6.2
Cryotherapy	85.6 ± 10	88.9 ± 7	91.1 ± 6.7	88.9 ± 9.9	90 ± 6	90 ± 10.5	91.1 ± 6.7	90 ± 10.5	90 ± 6	88.9 ± 7
Retinopathy	66 ± 4.7 **	66 ± 4.5 ***	55.3 ± 4.1 *	55.3 ± 4.1 *	66.2 ± 5.3 **	58.2 ± 4.2	66.7 ± 4.6 ***	59.3 ± 3.4	62.7 ± 5.9 .	59.3 ± 3.6
CTG-1	77.1 ± 2.1 ***	78.2 ± 2 **	61 ± 4.2 ***	64.1 ± 3.2 ***	74.5 ± 2 ***	75.1 ± 2.2 ***	78.4 ± 2.2 **	75 ± 2.4 ***	72.9 ± 2.9 ***	81.3 ± 2.5
CTG-2	91.6 ± 1.4 *	91.9 ± 1.7 .	87.6 ± 1.6 ***	87.2 ± 2.1 ***	92.5 ± 1.7	86.8 ± 1.1 ***	92.3 ± 1	86.7 ± 1.1 ***	91.1 ± 1.7 *	92.9 ± 1.6
Kidney	99.4 ± 1.9	99.4 ± 1.9	98.8 ± 2.5	99.4 ± 1.9	99.4 ± 1.9	98.1 ± 2.9	98.1 ± 2.9	98.1 ± 2.9	99.4 ± 1.9	99.4 ± 1.9
Dermatology	82.2 ± 5.2 ***	82.2 ± 5.2 ***	81.5 ± 2.8 ***	84.4 ± 2.2 ***	83.8 ± 2.4 ***	83.8 ± 3.7 ***	78.2 ± 5.6 ***	83.8 ± 3.7 ***	85.8 ± 4.7 ***	94.2 ± 2.6
SRBCT	80.7 ± 14.9	79.6 ± 14.7	77.6 ± 17.2	79.9 ± 15.6	82.1 ± 11.1	71.1 ± 15.8 .	85.4 ± 13.3	81 ± 14.4	85.7 ± 8.9	82.1 ± 12.5
Leukemia1	90.4 ± 8.5 *	90.4 ± 8.5 *	73.9 ± 12.4	79.1 ± 13.1	89.1 ± 7.8 *	79.1 ± 9.5	88.9 ± 12 *	78.9 ± 13	88.9 ± 10.2 *	77.7 ± 13.1
DLBCL	83.2 ± 10.1	87.3 ± 7.9	66.6 ± 10.7 ***	66.6 ± 10.7 ***	90.7 ± 8.8	69.1 ± 16.6 **	88.2 ± 9.4	74.1 ± 14.7 *	89.6 ± 8.3	88.2 ± 9.4
Prostate	80.5 ± 9.8	79.5 ± 10.3	66.6 ± 6.6 **	65.6 ± 11.2 **	75.5 ± 9.2	77.5 ± 11.8	79.5 ± 12.1	74.5 ± 14.3	79.5 ± 11.8	80.5 ± 12.5

Table 4

Average test set accuracy and standard deviation for the similarity classifier with fixed feature subset (Filter).

Sim (Filter)	ReliefF (10)	ReliefF (70,20)	FES	FSAE	Fisher Score	Nonspecificity	Strife	TU	SU	No FS
Artificial Case 1	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
Artificial Case 2	33.3 ± 0.3 ***	33.3 ± 0.3 ***	39.6 ± 0.4 ***	39.6 ± 0.4 ***	39.6 ± 0.4 ***	34.6 ± 2.5 *	35.3 ± 3	34.6 ± 2.5 *	33.3 ± 0.3 ***	36.8 ± 0.3
Artificial Case 3	99 ± 1 **	99 ± 1 **	99.7 ± 0.6	98 ± 0.2 ***	99.9 ± 0	99.9 ± 0	99.9 ± 0	99.9 ± 0	99.9 ± 0	99.9 ± 0
Artificial Case 4	33.3 ± 0.5 ***	33.3 ± 0.5 ***	39.6 ± 0.5 ***	39.6 ± 0.5 ***	39.6 ± 0.5 ***	33.9 ± 1.9 ***	34.5 ± 2.6 ***	33.9 ± 1.9 ***	33.3 ± 0.5 ***	46.2 ± 0.7
Artificial Case 5	98 ± 0.1 ***	98 ± 0.1 ***	99.4 ± 0.9	98 ± 0.1 ***	99.9 ± 0 ***	99.9 ± 0 ***	99.9 ± 0 ***	99.9 ± 0 ***	99.9 ± 0 ***	99.8 ± 0.1
Monk 1	68.8 ± 3.1	68.9 ± 2.5	70.1 ± 2.6	70.2 ± 2.5	68.2 ± 2.3	70.1 ± 2.6	66.1 ± 1.6 **	70.1 ± 2.6	70.2 ± 2.5	69.3 ± 2.2
Monk 2	54.8 ± 1.9 ***	55.8 ± 2.3 **	57.8 ± 1.6	57.8 ± 1.6	57.5 ± 2.6	57.8 ± 1.6	56.3 ± 2 **	57.8 ± 1.6	57.8 ± 1.6	58.8 ± 2.1
Monk 3	78.6 ± 2.7	78.6 ± 2.1	78.3 ± 2.3	78.3 ± 2.3	78.2 ± 2.1	78.3 ± 2.3	65.5 ± 8.4 ***	78.3 ± 2.3	78.3 ± 2.3	78.1 ± 1.7
Cryotherapy	80.1 ± 1.3	80.6 ± 1.6	84.6 ± 1.9 ***	80.1 ± 1.4	80.5 ± 1.6	84.8 ± 2.1 ***	85.1 ± 2.6 ***	84.8 ± 2.1 ***	80.7 ± 1.5	80.2 ± 1.5
Retinopathy	57.7 ± 0.6 *	57.6 ± 0.6 *	55.1 ± 0.4 ***	55.1 ± 0.4 ***	57.6 ± 0.6	60.6 ± 1.9 ***	57.6 ± 0.6	61.3 ± 1.1 ***	59.1 ± 1 ***	57.1 ± 0.5
CTG-1	52.4 ± 0.7 ***	51.9 ± 0.8 ***	42.5 ± 4.2 ***	48 ± 1.8 ***	54.2 ± 0.4 **	53.3 ± 0.3	52.4 ± 1.6 *	53.3 ± 0.3	43.7 ± 4.5 ***	53.6 ± 0.5
CTG-2	77.2 ± 2.4 ***	75.5 ± 1.5 ***	61.1 ± 0.2 ***	63.7 ± 0.5 ***	78.6 ± 1.3 ***	65.2 ± 5.3 **	74.5 ± 0.7 ***	64.5 ± 5.7 **	78.4 ± 0.8 ***	71.6 ± 0.3
Kidney	97 ± 0.8 ***	96.6 ± 0.8 ***	92 ± 2.4 **	93.2 ± 1.2 *	97.4 ± 0.6 ***	83.3 ± 1.5 ***	98.6 ± 0.3 ***	83.4 ± 1.7 ***	94.2 ± 0.7	94.2 ± 0.7
Dermatology	80 ± 0.6 ***	80 ± 0.6 ***	77.4 ± 4.1 ***	76.3 ± 0.7 ***	81.3 ± 0.5 ***	79.8 ± 1.4 ***	85.8 ± 1.4 ***	79.8 ± 1.4 ***	82.9 ± 0.8 ***	96.2 ± 0.5
SRBCT	99.9 ± 0.4 ***	100 ± 0 ***	58.1 ± 4.2 ***	79 ± 6.4 ***	100 ± 0 ***	79.9 ± 3.5 ***	98.9 ± 1.2 ***	82.3 ± 2.7 ***	95.5 ± 3	96 ± 1.6
Leukemia1	97.4 ± 1.2 **	97.4 ± 1.2 **	50.8 ± 6.1 ***	82.1 ± 3.6 ***	97.1 ± 1.3 **	93.8 ± 1.2	97.2 ± 1.2 **	94.3 ± 1.2	94.9 ± 1.8	94.4 ± 2.3
DLBCL	96 ± 1.9 ***	94.4 ± 2.4 ***	65.7 ± 13.4 **	66.9 ± 12.8 *	90.5 ± 2.4 ***	70.9 ± 7.9 **	79.1 ± 1.1	72.2 ± 5.9 **	93.8 ± 1.4 ***	78.1 ± 2.4
Prostate	91.2 ± 1.7 ***	89.8 ± 1.3 ***	65.9 ± 4.2 **	67.1 ± 4.9 **	84.6 ± 3.9 ***	65.6 ± 1.4 ***	79.8 ± 5.1 ***	65.7 ± 1.4 ***	83.6 ± 4.7 ***	61.7 ± 1.2

Table 5

Average test set accuracy and standard deviation for the support vector machine with fixed feature subset (Filter)

SVM (Filter)	ReliefF (10)	ReliefF (70,20)	FES	FSAE	Fisher Score	Nonspecificity	Strife	TU	SU	No FS
Artificial Case 1	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
Artificial Case 2	33.3 ± 3 ***	33.3 ± 3 ***	39.6 ± 3.5 ***	39.6 ± 3.5 ***	39.6 ± 3.5 ***	34.3 ± 4.1 ***	34.9 ± 3.4 ***	34.3 ± 4.1 ***	33.3 ± 3 ***	58.7 ± 3.1
Artificial Case 3	86.1 ± 6.3 ***	86.1 ± 6.3 ***	81.3 ± 4 ***	91.9 ± 2.3 ***	80 ± 0 ***	80 ± 0 ***	80 ± 0 ***	80 ± 0 ***	80 ± 0 ***	97.8 ± 1.3
Artificial Case 4	33.3 ± 4.7 ***	33.3 ± 4.7 ***	39.6 ± 4.6 ***	39.6 ± 4.6 ***	39.6 ± 4.6 ***	34.3 ± 5.1 ***	35.1 ± 5 ***	34.3 ± 5.1 ***	33.3 ± 4.7 ***	55.8 ± 3.1
Artificial Case 5	91.7 ± 2.2 **	91.7 ± 2.2 **	83.5 ± 5.5 ***	91.7 ± 2.2 **	80 ± 0 ***	80 ± 0 ***	80 ± 0 ***	80 ± 0 ***	80 ± 0 ***	94.8 ± 1.7
Monk 1	66.1 ± 14.9	67.6 ± 15.1	69.2 ± 13.3	69.2 ± 13.3	66.1 ± 11.6	69.2 ± 13.3	66 ± 14.1	69.2 ± 13.3	69.2 ± 13.3	64.4 ± 13.5
Monk 2	62.1 ± 2.8	62.1 ± 2.8	62.1 ± 2.8	62.1 ± 2.8	62.1 ± 2.8	62.1 ± 2.8	62.1 ± 2.8	62.1 ± 2.8	62.1 ± 2.8	62.1 ± 2.8
Monk 3	78.1 ± 11.5	77.2 ± 12	76.4 ± 10.6	76.4 ± 10.6	78.7 ± 8.3	76.4 ± 10.6	54.9 ± 12.6 ***	76.4 ± 10.6	76.4 ± 10.6	81.3 ± 9.4
Cryotherapy	85.6 ± 10	83.3 ± 11.4	84.4 ± 11.3	84.4 ± 11.3	83.3 ± 11.4	86.7 ± 9.7	87.8 ± 9.2	86.7 ± 9.7	84.4 ± 12.4	85.6 ± 10
Retinopathy	65.7 ± 5.7	65.1 ± 5.6	53.4 ± 0.4 ***	53.4 ± 0.4 ***	64.6 ± 5.2	60 ± 6.1 **	65.3 ± 4.5	60.3 ± 5.8 **	61.1 ± 5.9 **	67.7 ± 4.4
CTG-1	66.3 ± 4.4 *	65.5 ± 4.5 **	46.9 ± 4.9 ***	48.8 ± 1.8 ***	64.5 ± 2.7 ***	63.2 ± 1.5 ***	64.7 ± 3.4 ***	63.1 ± 1.4 ***	59 ± 2.7 ***	70.9 ± 3.8
CTG-2	86.9 ± 1.3 *	86.8 ± 1.3 **	80.9 ± 1.4 ***	81.2 ± 1.3 ***	87.3 ± 1.8	81.5 ± 1.7 ***	87 ± 2.1 *	81.5 ± 1.7 ***	84.1 ± 1.3 ***	88.5 ± 1.7
Kidney	99.4 ± 1.9	99.4 ± 1.9	98.1 ± 4	100 ± 0	99.4 ± 1.9	96.8 ± 4.3	98.7 ± 2.6	97.5 ± 4.2	100 ± 0	99.4 ± 1.9
Dermatology	84.1 ± 3 ***	84.1 ± 3 ***	82.9 ± 3.8 ***	85.5 ± 2.5 ***	86 ± 1.8 ***	85.2 ± 3.1 ***	81.6 ± 6.2 ***	85.2 ± 3.1 ***	88 ± 3.4 ***	98.3 ± 1.8
SRBCT	100 ± 0	100 ± 0	86.8 ± 11	96.5 ± 7.3	100 ± 0	95.3 ± 9.5	100 ± 0	95.3 ± 7.6 *	97.6 ± 4.7	100 ± 0
Leukemia1	95.9 ± 6.3	95.9 ± 6.3	84.6 ± 7.8 ***	94.5 ± 6.8	94.5 ± 6.8	95.9 ± 6.3	95.9 ± 6.3	95.9 ± 6.3	95.9 ± 6.3	97.1 ± 5.7
DLBCL	97.5 ± 7.5	96.1 ± 6	76.8 ± 4.6 ***	75.4 ± 3.7 ***	94.8 ± 8.5	81.1 ± 10 ***	94.6 ± 6.6	81.1 ± 10 ***	95 ± 8.3	98.8 ± 3.8
Prostate	94.1 ± 4.8 *	95.1 ± 4.9 *	63.9 ± 10.8 ***	61.9 ± 12.2 ***	92.2 ± 5.7	88.2 ± 7.3	92.1 ± 4	90.2 ± 7.5	91.2 ± 5	90.1 ± 4.5

nonspecificity and total uncertainty is often considerably better than that of strife, which seems to have difficulty with the unusual, logical target concepts. The performance on the MONK 1 data set is still comparable with the Fisher score, but the target concept for the MONK 3 data set is clearly not captured using strife. On the real-world data sets, the KNN classifier appears to perform best with the Fisher score, resulting for four out of the ten data sets in the best test accuracy. Strife is often showing a comparable performance to the Fisher score and is for the CTG-1 (79.4%) data set and the Leukemia1 (95.9%) data set the best performing feature ranking method. With exception of the Cryotherapy data set, strife is more accurate on all real-world data sets than nonspecificity and total uncertainty. Overall, feature selection appears to be beneficial for the accuracy of the KNN classifier. For eight out of ten data sets the performance with at least one feature selection

method is higher than the one using no feature selection, and in five of these cases the result is significantly higher than that of no feature selection.

The results using the decision tree (Table 3) are largely comparable to those obtained using the KNN classifier, especially on the artificial example cases. Both strife and FES achieve the best test accuracy of 91.1% of any feature selection method and classifier combination on the cryotherapy data set. The two ReliefF setups seem to perform very well with the decision tree, resulting for three data sets in the highest test set accuracies and for the remaining data sets in very competitive results compared to the remaining filter methods. It is noteworthy that in six out of the ten real-world data sets, at least one feature selection method resulted in a higher average test set accuracy and in

Table 6

Diversity of feature ranking methods on all real-world data sets.

ASM (All)	ReliefF (10)	ReliefF (70,20)	FES	FSAE	Fisher Score	Nonspecificity	Strife	TU	SU
ReliefF (10)	0.8	0.6	-0.1	0.0	0.4	0.0	0.3	0.0	0.2
ReliefF (70,20)	0.6	0.8	-0.1	0.0	0.5	0.0	0.3	0.0	0.2
FES	-0.1	-0.1	0.8	0.6	0.0	0.4	-0.1	0.4	0.1
FSAE	0.0	0.0	0.6	0.8	0.1	0.4	-0.1	0.4	0.1
Fisher Score	0.4	0.5	0.0	0.1	0.8	0.1	0.3	0.1	0.2
Nonspecificity	0.0	0.0	0.4	0.4	0.1	0.8	0.0	0.7	0.1
Strife	0.3	0.3	-0.1	-0.1	0.3	0.0	0.8	0.0	0.1
TU	0.0	0.0	0.4	0.4	0.1	0.7	0.0	0.8	0.1
SU	0.2	0.2	0.1	0.1	0.2	0.1	0.1	0.1	0.8

Table 7

Average test set accuracy and standard deviation for the KNN classifier using a hybrid approach.

KNN (Hybrid)	ReliefF (10)	ReliefF (70,20)	FES	FSAE	Fisher Score	Nonspecificity	Strife	TU	SU	No FS
Artificial Case 1	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
Artificial Case 2	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
Artificial Case 3	100 ± 0	100 ± 0	100 ± 0	99.9 ± 0.3	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
Artificial Case 4	100 ± 0 ***	100 ± 0 ***	91.2 ± 2.3 ***	91 ± 2.5 ***	100 ± 0 ***	100 ± 0 ***	100 ± 0 ***	100 ± 0 ***	100 ± 0 ***	51.9 ± 3.1
Artificial Case 5	99.7 ± 0.7 ***	99.7 ± 0.7 ***	99.8 ± 0.6 ***	99.7 ± 0.7 ***	100 ± 0 ***	100 ± 0 ***	100 ± 0 ***	100 ± 0 ***	100 ± 0 ***	68.4 ± 2.7
Monk 1	92.8 ± 5.4 ***	92.8 ± 5.4 ***	83.2 ± 8.4 ***	92.5 ± 5.4 ***	86.1 ± 11.3 ***	82.8 ± 8 ***	70 ± 6.2	82.8 ± 8 ***	84.6 ± 8.1 ***	67 ± 4.3
Monk 2	73.1 ± 6.9	73.2 ± 6.3	73.2 ± 7.2	73.2 ± 7.2	74.8 ± 5.2	73.2 ± 7.2	74.5 ± 5.9	73.2 ± 7.2	75.3 ± 4.3	76.4 ± 3.7
Monk 3	94.9 ± 3.5 ***	94.9 ± 3.5 ***	88 ± 2.8 ***	94.9 ± 3.5 ***	96.7 ± 1.7 ***	88 ± 2.8 ***	80.4 ± 2.7	88 ± 2.8 ***	88 ± 2.8 ***	81.6 ± 0.5
Cryotherapy	86.7 ± 10.9	86.7 ± 10.9	87.8 ± 10.5	84.4 ± 11.3	86.7 ± 10.9	88.9 ± 11.1	86.7 ± 8.3	88.9 ± 11.1	91.1 ± 8.3	87.8 ± 10.5
Retinopathy	65.9 ± 5.4	66.8 ± 4.7 *	63.4 ± 3.3	63.0 ± 3.1	68.0 ± 4.2 **	61.6 ± 5.4	66.3 ± 5.2	61.6 ± 5.4	67.7 ± 3.4 **	63.2 ± 3.9
CTG-1	78.3 ± 3.5	77.9 ± 3.4	80.5 ± 2.6 **	79.6 ± 3.2	80.5 ± 2.9 *	79.4 ± 2.1 *	80.4 ± 2.7 *	79.2 ± 2.1	77.1 ± 3.8	77.6 ± 2.3
CTG-2	90.9 ± 3.1	89.9 ± 2.4	91.6 ± 1.6	91.5 ± 1.8	92.5 ± 1.8	91.7 ± 1.7	92.1 ± 1.7	91.7 ± 1.6	90.5 ± 1.8	91.1 ± 2.6
Kidney	98.8 ± 2.5	99.4 ± 1.9 *	97.5 ± 4.1	97.5 ± 4.1	100 ± 0 **	96.1 ± 4.2	98.0 ± 3.0	96.1 ± 4.2	97.5 ± 4.1	96.8 ± 3.2
Dermatology	95.0 ± 3.0	95.0 ± 3.0	96.7 ± 2.1	98.3 ± 1.8	97.5 ± 3.4	97.2 ± 2.8	97.5 ± 1.9	97.2 ± 2.8	95.5 ± 4.2	97.2 ± 3.3
SRBCT	95.1 ± 8.2 *	96.2 ± 8.0 *	84.3 ± 14.3	90.4 ± 10.5	100 ± 0 **	97.8 ± 4.4 **	94.0 ± 7.8	96.7 ± 5.1 *	96.4 ± 5.5 *	87.1 ± 10.7
Leukemia1	91.4 ± 9.5	92.9 ± 9.6	77.5 ± 9.8 *	80.5 ± 13.0	94.5 ± 6.8 *	87.3 ± 14.9	95.9 ± 6.3 *	88.9 ± 1	90.2 ± 11.1	85.9 ± 11.1
DLBCL	89.8 ± 12.3	89.6 ± 12.3	75.4 ± 9.8 **	78.9 ± 14.	92.3 ± 11.5	84.6 ± 13.8	83.2 ± 12.8	84.6 ± 11.3	90.7 ± 10.5	88.4 ± 11
Prostate	90.2 ± 7.5	94.2 ± 6.3 *	75.4 ± 16.6	75.4 ± 16.7	89.1 ± 11.4	89.2 ± 5.4	88.2 ± 11.5	90.1 ± 7.7	91.2 ± 6.7	85.2 ± 10.2

Table 8

Average test set accuracy and standard deviation for the decision tree classifier using a hybrid approach.

DT (Hybrid)	ReliefF (10)	ReliefF (70,20)	FES	FSAE	Fisher Score	Nonspecificity	Strife	TU	SU	No FS
Artificial Case 1	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
Artificial Case 2	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
Artificial Case 3	99.9 ± 0.2	99.9 ± 0.2	99.9 ± 0.2	99.7 ± 0.4 *	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
Artificial Case 4	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
Artificial Case 5	99.8 ± 0.4	99.8 ± 0.4	99.9 ± 0.4	99.8 ± 0.4	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
Monk 1	83.3 ± 0	83.3 ± 0	89.2 ± 5.9 **	85.3 ± 5.6	84.2 ± 6.1	89.7 ± 5.6 **	81.5 ± 3.8	89.7 ± 5.6 **	90.2 ± 4.4 ***	82.5 ± 2.5
Monk 2	64.3 ± 2.9	64 ± 2.2	65.3 ± 2.9	65.4 ± 2.9	65.9 ± 3	65.4 ± 2.9	63.5 ± 2.6	65.4 ± 2.9	65.5 ± 3.5	65.5 ± 3.5
Monk 3	97.2 ± 0	97.2 ± 0	97.2 ± 0	97.2 ± 0	97.2 ± 0	97.2 ± 0	97.2 ± 0	97.2 ± 0	97.2 ± 0	97.2 ± 0
Cryotherapy	88.9 ± 7.0	88.9 ± 7.0	85.6 ± 14.9	83.3 ± 13.4	90 ± 6.0	90 ± 10.5	91.1 ± 9.7	90 ± 10.5	87.8 ± 7.8	86.7 ± 9.7
Retinopathy	66.0 ± 2.7 **	67.6 ± 3.8 **	61.6 ± 6.3	60.9 ± 6.7	66.1 ± 4.3 **	60.5 ± 4.9	66.7 ± 3 **	60.5 ± 4.9	61.6 ± 6.1	59.9 ± 5.6
CTG-1	80.9 ± 3.4	80.8 ± 3.3	81.0 ± 2.6	81.2 ± 2.6	81.5 ± 2.6	82.2 ± 3	82.1 ± 2.4	82.2 ± 3	80.5 ± 3.8	81.5 ± 2.2
CTG-2	92.2 ± 1.5	92.2 ± 2.0	92.6 ± 1.9	92.4 ± 2	92.4 ± 2.5	93 ± 1.4	92.7 ± 1	93 ± 1.4	91.9 ± 2.1	92.1 ± 2.4
Kidney	98.7 ± 2.5	99.4 ± 1.9	98.7 ± 2.5	98.7 ± 2.5	99.4 ± 1.9	98.1 ± 4.1	98.7 ± 2.6	98.1 ± 4.1	98.8 ± 2.5	99.4 ± 1.9
Dermatology	95.5 ± 3.1	95.3 ± 3.7	95.0 ± 4.4	95.0 ± 4.4	94.7 ± 2.9	94.4 ± 4.1	96.1 ± 2.8	94.4 ± 4.1	94.7 ± 2.9	94.1 ± 3.8
SRBCT	83.6 ± 12.6	84.7 ± 11.5	73.3 ± 10.9	73.7 ± 18.2	93.9 ± 8.3 ***	77.2 ± 8.1	79.9 ± 14.6	72.8 ± 15.5	84.3 ± 7.9	78.5 ± 10.3
Leukemia1	84.5 ± 10.0	85.9 ± 9.0	87.7 ± 9.5	79.1 ± 14.6	91.4 ± 11.4	86.1 ± 11.9	89.1 ± 7.7	86.1 ± 10.8	84.8 ± 7.2	87.5 ± 9.5
DLBCL	88.2 ± 9.4	88.2 ± 9.4	77.9 ± 18.1	73.9 ± 18.0 *	86.8 ± 10.7	77.5 ± 21.5	81.4 ± 16.5	87.5 ± 9.7	74.1 ± 12.8 *	88.2 ± 13.5
Prostate	85.2 ± 6.8	85.2 ± 6.9	71.5 ± 16.0 *	81.2 ± 10.6	86.2 ± 6.7	83.5 ± 8.6	86.4 ± 9	82.5 ± 8.3	87.1 ± 11	86.1 ± 10.2

Table 9

Average test set accuracy and standard deviation for the similarity classifier using a hybrid approach.

Sim (Hybrid)	ReliefF (10)	ReliefF (70,20)	FES	FSAE	Fisher Score	Nonspecificity	Strife	TU	SU	No FS
Artificial Case 1	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
Artificial Case 2	36.8 ± 0.3	36.8 ± 0.3	39.6 ± 0.4 ***	39.6 ± 0.4 ***	39.6 ± 0.4 ***	36.8 ± 0.3	37 ± 1	36.8 ± 0.3	36.8 ± 0.3	36.8 ± 0.3
Artificial Case 3	99.9 ± 0	99.9 ± 0	99.9 ± 0	99.9 ± 0	99.9 ± 0	99.9 ± 0	99.9 ± 0	99.9 ± 0	99.9 ± 0	99.9 ± 0
Artificial Case 4	46.5 ± 2 *	46.8 ± 0.8 ***	52.9 ± 1 ***	52 ± 1.5 ***	46.6 ± 1.2 **	46.9 ± 1 **	45.7 ± 4.6 .	46.6 ± 1.3 **	45.8 ± 2 **	47.9 ± 0.4
Artificial Case 5	99.9 ± 0 ***	99.9 ± 0 ***	99.9 ± 0 ***	99.9 ± 0 ***	99.9 ± 0 ***	99.9 ± 0 ***	99.9 ± 0 ***	99.9 ± 0 ***	99.9 ± 0.1 ***	99.6 ± 0.1
Monk 1	67.1 ± 1.9 **	67.2 ± 2 **	65.6 ± 5.3 *	67.6 ± 1.9 **	67.7 ± 1.9 **	67.5 ± 2.1 *	66.8 ± 1.3 ***	67.5 ± 2.1 *	67.4 ± 4.1 .	69.4 ± 0
Monk 2	55 ± 1.5 ***	55.1 ± 1.6 ***	56.4 ± 1.6 **	57.3 ± 1.1 *	58.2 ± 0.5 .	57 ± 1 **	55.1 ± 2 ***	57 ± 1 **	58.4 ± 0.8 .	58 ± 0
Monk 3	77.7 ± 0.8 ***	78.1 ± 0.8 ***	74.6 ± 4.1	77.5 ± 2.7 .	78.2 ± 0.8 ***	74.6 ± 4.1	73.2 ± 3.1 **	74.6 ± 4.1	75.9 ± 3.1	76.2 ± 0
Cryotherapy	81.5 ± 1.9 *	80.3 ± 3.8	82.2 ± 4.9 .	82.8 ± 2.1 **	82.2 ± 1.8 **	83.2 ± 4.3 *	83 ± 2 **	83.2 ± 4.3 *	79.4 ± 3.3	79.8 ± 2.2
Retinopathy	59.3 ± 0.6 ***	59.2 ± 0.6 ***	58.4 ± 1.7 *	58.4 ± 2.2 .	59.2 ± 0.5 ***	60.2 ± 2.2 ***	59.2 ± 0.4 ***	60.2 ± 2.2 ***	59.1 ± 2.7 *	57.1 ± 0.5
CTG-1	51.5 ± 2.7 *	52.9 ± 2.5	56 ± 1.1 ***	58.4 ± 1.5 ***	56.4 ± 2.1 **	59.7 ± 1.3 ***	53.5 ± 2.2	59.7 ± 1.3 ***	54.5 ± 1 *	53.7 ± 0.3
CTG-2	75.3 ± 1.9 ***	73.9 ± 1.5 ***	74.6 ± 0.3 ***	75.9 ± 1.8 ***	80.6 ± 1.4 ***	81.2 ± 0.4 ***	78.2 ± 1.2 ***	81.2 ± 0.4 ***	79.5 ± 1.4 ***	71.7 ± 0.4
Kidney	96.2 ± 1.4 ***	96 ± 1.2 ***	94.8 ± 0.6 **	95.2 ± 0.9 **	97.4 ± 1 ***	92.2 ± 3.8 .	97.7 ± 1.1 ***	91.2 ± 4.6 *	95.1 ± 0.7 **	94.2 ± 0.4
Dermatology	96 ± 0.6	95.7 ± 0.8 .	94.4 ± 2.1	94.3 ± 2.1	96.5 ± 0.7	94.7 ± 1.4	95.3 ± 0.9	94.7 ± 1.4	94 ± 2.2	96.2 ± 0.3
SRBCT	95.1 ± 6.8	95.6 ± 6.6	91 ± 9 .	96.1 ± 2.5	97.3 ± 2.2 .	87 ± 7.5 **	97.3 ± 2.6	87 ± 6.8	93.3 ± 7.2	96 ± 1.9
Leukemia1	94.8 ± 1.6	93.7 ± 3.9	94.8 ± 6.4	89.8 ± 9.9 .	92.4 ± 8.9	95.2 ± 4.8	93.2 ± 2.9	96.1 ± 3.3	91.2 ± 5.3 .	94.4 ± 2.5
DLBCL	86.6 ± 8.5 **	87.6 ± 8.6 **	65.8 ± 10.4 **	65.5 ± 10.2 **	89.6 ± 3.9 ***	71.2 ± 2.6 ***	86.7 ± 4.1 ***	71.7 ± 2.6 ***	77.3 ± 10.8	78.2 ± 2.2
Prostate	91.9 ± 1.8 ***	91.7 ± 1.7 ***	62.5 ± 5.4	62.6 ± 5.1	90.6 ± 1 ***	71.1 ± 7.5 **	92.3 ± 1.7 ***	73.6 ± 7.8 ***	90.7 ± 4.6 ***	61.8 ± 1.8

Table 10

Average test set accuracy and standard deviation for the support vector machine using a hybrid approach.

SVM (Hybrid)	ReliefF (10)	ReliefF (70,20)	FES	FSAE	Fisher Score	Nonspecificity	Strife	TU	SU	No FS
Artificial Case 1	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
Artificial Case 2	58.9 ± 4.7	58.9 ± 4.7	58.9 ± 4.7	58.9 ± 4.7	58.9 ± 4.7	58.9 ± 4.7	58.9 ± 4.7	58.9 ± 4.7	58.9 ± 4.7	58.9 ± 4.7
Artificial Case 3	97.9 ± 0.9	97.9 ± 0.9	97.9 ± 0.9	97.9 ± 0.9	97.9 ± 0.9	97.9 ± 0.9	97.9 ± 0.9	97.9 ± 0.9	97.9 ± 0.9	97.9 ± 0.9
Artificial Case 4	56.7 ± 5	57.6 ± 4.6	54.4 ± 3.9	54.2 ± 4.9	58.4 ± 4.2	58.9 ± 3.6 .	58.6 ± 2.9	58.6 ± 4	58.3 ± 5.1	56.6 ± 4
Artificial Case 5	97.7 ± 1.2 ***	97.7 ± 1.4 ***	97.5 ± 1.2 ***	97.6 ± 1.5 ***	97.8 ± 1.4 ***	97.7 ± 1.5 ***	97.5 ± 1.6 ***	97.5 ± 1.8 ***	97.6 ± 1 ***	94.5 ± 1.5
Monk 1	67.7 ± 1.4 **	67.5 ± 1.3 **	67.3 ± 0.4 ***	67.5 ± 1.3 **	67 ± 0.8 **	67.3 ± 0.4 ***	66.4 ± 0.2 **	67.3 ± 0.4 ***	62.6 ± 8.3	66.2 ± 0
Monk 2	67.1 ± 0	67.1 ± 0	67.1 ± 0	67.1 ± 0	67.1 ± 0	67.1 ± 0	67.1 ± 0	67.1 ± 0	67.1 ± 0	67.1 ± 0
Monk 3	80.5 ± 5.6	80.5 ± 5.6	75.2 ± 9.2 *	78.8 ± 7.5 .	79.9 ± 5.4	75.2 ± 9.2 *	78 ± 7.4 .	75.2 ± 9.2 *	77.2 ± 8.7 .	82.2 ± 0
Cryotherapy	84.4 ± 7.4	85.6 ± 7.1	91.1 ± 6.7	90 ± 6	86.7 ± 6.7	87.8 ± 7.8	87.8 ± 9.2	87.8 ± 7.8	82.2 ± 8.9 .	88.9 ± 9.9
Retinopathy	67.6 ± 3.5	67.4 ± 4	68.2 ± 4.1	67.9 ± 3.8	68.4 ± 3.8	67.4 ± 4.4	67.2 ± 5.1	67.4 ± 4.4	67.7 ± 4.5	68.1 ± 4.5
CTG-1	69.9 ± 2.6	69.4 ± 2.5	70.7 ± 2.8	69.5 ± 2.4	69.1 ± 3.2	71.1 ± 2.1	70.5 ± 2.1	71.1 ± 2.1	69.2 ± 2.4	70.2 ± 2.6
CTG-2	88.4 ± 2.1	88.4 ± 1.9	88.5 ± 2	88.9 ± 2.3	88 ± 2.1	88.9 ± 1.7	88.9 ± 1.9	88.9 ± 1.7	87.4 ± 3.1	88.7 ± 1.9
Kidney	97.5 ± 4.2	97.5 ± 4.2	98.1 ± 2.9	97.5 ± 4.2	98.7 ± 2.6	97.4 ± 4.4	99.4 ± 1.9	97.4 ± 4.4	97.5 ± 4.2	99.4 ± 1.9
Dermatology	97.2 ± 2.2	97.2 ± 1.8 .	97.5 ± 2	98.1 ± 1.3	97.8 ± 1.7	96.7 ± 2.1	97.2 ± 1.8	96.7 ± 2.1	96.4 ± 1.8	98.3 ± 1.9
SRBCT	95.1 ± 6 *	97.6 ± 4.7 .	90.4 ± 8.9 **	100 ± 0	95.3 ± 5.8 *	98.9 ± 3.3	95 ± 11.5	98.9 ± 3.3	97.5 ± 5 .	100 ± 0
Leukemia1	97.3 ± 5.4	97.3 ± 5.4	85 ± 10.7 **	86.1 ± 12.8 *	94.6 ± 6.6	87.3 ± 10 *	94.5 ± 9.3	92.9 ± 7.1	94.6 ± 6.6	96.1 ± 6
DLBCL	91.3 ± 9.8 *	92.5 ± 10 .	89.6 ± 9.5 *	88.4 ± 10.5 *	90 ± 9.4 *	92.1 ± 9.1 .	96.3 ± 5.7	93.4 ± 9.2	92.5 ± 10 .	97.5 ± 5
Prostate	93.1 ± 7.8	96.1 ± 4.8 .	88.3 ± 9.3	87.4 ± 12.8	96.1 ± 4.8 .	92.2 ± 6	93.1 ± 6.4	89.2 ± 8.3	91.1 ± 5.4	92.2 ± 7.4

two out of these cases the test accuracy was even significantly higher than that without feature selection. This is of interest since decision trees are embedded feature selection methods that select features to build the decision tree (forward selection) (Blum and Langley, 1997; Duch, 2006). However, the results highlight that using feature selection prior to the decision tree can still in some cases improve the accuracy significantly.

The results for the similarity classifier (Table 4) are mixed and show very low test set accuracies for the CTG-1 and CTG-2 data sets with all feature ranking methods. However, at the same type, this classifier leads to high test accuracies on the microarray data sets (SRBCT, Leukemia1, DLBCL and Prostate), indicating that this simple classifier performs very well on this types of data sets, especially using ReliefF

for feature ranking. The Fisher score and strife provide very high test accuracies for most data sets and lead to the best test set accuracies for three and two of the real-world data sets, respectively. ReliefF appears to perform particularly well on the microarray data sets and provides in up to three data sets the highest test set accuracy. With exception of the dermatology data set, the test accuracies for most feature ranking methods exceed that of using no feature selection. For nine out of the ten real-world data sets at least one feature ranking method improves the test accuracy significantly. As in the KNN classifier, the similarity classifier includes all features in the classification model, making it susceptible to irrelevant and noisy features. Thus, the fact that feature selection in most cases improves the accuracy of this classifier is intuitive.

Table 11
Average rank of classifier performance, feature number and ASM on the real-world data sets.

Approach	Filter		Hybrid						
	Avg. Acc	ASM (All)	Avg. Acc.	Avg. Feat.	ASM KNN	ASM DT	ASM Sim	ASM SVM	ASM Overall
Relieff (10)	4.0	5.45	5.6	3.2	4.7	5.45	4.2	3.7	4.6
Relieff (70,20)	4.4	5	5.3	3.1	4.65	5.15	4.6	2.7	3.9
FES	8.6	5.25	6.6	8.2	6.05	5.6	6	6.8	6.5
FSAE	7.8	2.85	6.7	6.5	5	4.8	5.3	5.8	5
Fisher Score	3.9	3.55	3.8	2.7	3	3.5	2.5	2.7	2.5
Nonspecificity	6.9	4.7	5.2	6.5	5.3	4.95	7.25	5.95	6.2
Strife	4.0	5.75	4.5	3.5	4.6	4.2	3	5	3.8
TU	6.6	5.25	5.1	6.2	5.5	5.05	6.95	6.15	6.4
SU	4.4	7.2	6.4	5.2	6.2	6.3	5.2	6.2	6.1
No FS	4.4	10	5.8	10	10	10	10	10	10

The support vector machine (SVM) with linear kernel (Table 5) performs poorly on most artificial data sets due to its simple kernel and the fact that it can neither handle multiple decision regions for each class (artificial case two and four) nor use the actually most relevant feature to construct two hyperplanes to divide the feature space perfectly according to the classes. However, the SVM performs extraordinarily well on the SRBCT, DLBCL and Prostate data sets for many of the feature ranking methods including Relieff, Fisher score, strife and symmetrical uncertainty. Relieff (70,20) and strife both result for two data sets in the highest test accuracies and strife's performance is in most cases comparable with that of the Fisher score. Overall, the linear SVM's test set accuracies are in most cases (seven out of ten data sets) the highest when using no feature selection. Out of the three cases in which using feature selection leads to higher test accuracies than using no feature selection, only two times the test set accuracy is significantly higher (Prostate data set with Relieff).

In terms of stability, the adjusted stability measure (ASM) indicates that all feature ranking methods tend to lead to comparably stable feature subsets with adjusted stabilities of 0.5 to 0.7 (see Table A.3 in the appendix). It is noteworthy that symmetrical uncertainty shows the lowest stabilities on most of the microarray data sets which have a large number of features (in the thousands) and only a small number of observations.

In this study, the ASM is also used as a diversity measure by comparing the stability of subsets generated by different feature selection methods in the same iterations of the cross-validation. If two feature selection methods generate considerably different feature subsets, the ASM will attribute a low score to them, indicating that the selected features are not similar and, thus, diverse. The average ASM scores comparing the feature selection methods with each other are displayed in Table 6.

The results indicate that nonspecificity and total uncertainty lead to highly similar feature subsets. This is likely on account of the similar behavior of their weights (see Fig. A.1 in the appendix) and indicates that using both algorithms may be of limited use in the context of a feature selection ensemble due to a lack of diversity. The same appears to hold true for the two Relieff setups and the FSAE compared to FES. The average Spearman rank correlations calculated over the entire feature rankings (not just the considerably smaller feature subsets) supports this finding (see Table A.4 in the appendix). In addition, both the average Spearman rank correlations and average ASM scores indicate that the feature subsets obtained from symmetrical uncertainty, strife and the Fisher score are not similar to those of the remaining feature ranking methods, indicating that they are diverse and also may be suitable additions to feature selection ensembles to add diversity.

The results using the hybrid feature selection approach where the feature rankings are used as input to a wrapper to determine the best number of highest-ranking features to select for the feature subset are presented in Tables 7 to 10. The results for the KNN classifier using the hybrid approach are displayed in Table 7 and show that in contrast to the previous filter results, for the second and fourth artificial cases perfect test set accuracies can also be obtained using strife, nonspecificity and total uncertainty. Thus, using a hybrid approach with the classifier avoided removing the most relevant feature in all cases. As with the

feature ranking with a fixed threshold, the results obtained using the Fisher score are the best for most data sets using the KNN classifier. Strife generally results in competitive results to the Fisher score with exception of the SRBCT and DLBCL data sets for which the Fisher score clearly outperforms all other feature ranking methods. For the KNN classifier, strife leads to the highest classification accuracy for the Leukemia1 data set, which is significantly higher than using no feature selection, and outperforms most approaches, including symmetrical uncertainty and the Relieff setups. For all ten real-world data sets at least one approach is outperforming no feature selection in terms of the test set accuracy, and for seven of these the test accuracy was improved significantly. In terms of stability and feature subset size, there is no method that is consistently more stable or consistently generates smaller average feature subsets. Anyhow, the Fisher score seems to be overall the most stable method with the KNN classifier (see Table A.5).

The results for the decision tree classifier are displayed in Table 8. Once again, for the second and fourth artificial cases using the hybrid approach avoids for most approaches, including strife, nonspecificity, total uncertainty as well as the Fisher score the removal of the most relevant feature. For the decision tree the three proposed feature ranking methods all work well, and each method leads for two data sets to the highest average test accuracy. It is noteworthy that strife leads on the cryotherapy data set with the decision tree to the highest accuracy of 91.11% of all classifiers, a result only achieved otherwise with symmetrical uncertainty and the KNN classifier as well as FES with the SVM. Compared to the results of the decision tree using a fixed threshold, now in eight instead of six cases the best test accuracy on the real-world data sets is achieved with a feature

¹ Since the ideal vector element for each class is not representative of the feature values of observations belonging to the opposing class, there is a steep drop in the ordered similarity values for all observations, which translates into a peak for the weighted strife values (at around 600 observations). This in turn contributes to a higher strife value, which is indicative of a considerably higher uncertainty within all data than within each class.

selection algorithm but the result is again only significantly better than no feature selection for two data sets (SRBCT, Retinopathy). It is remarkable that all feature ranking methods seem to perform considerably worse on the SRBCT data set when used in combination with the decision tree classifier instead of the KNN, similarity or SVM classifier. With the decision tree classifier, strife and the Fisher score appear to generate on average the smallest and most stable feature subsets (see Table A.6).

With the similarity classifier (Table 9) the Fisher score is performing on average the best for only two data sets but is displaying competitive results for most of the remaining data sets. It is worth mentioning that for most data sets the similarity classifier with the best performing filter methods is significantly or highly significantly more accurate than using no feature selection. However, this may be since for many data sets the test accuracy using this classifier without feature selection is considerably lower than using any of the other classification algorithms in this study. Nonspecificity and total uncertainty give the highest test set accuracies on four of the real-world data sets, but strife is competitive in performance in most of these cases. In contrast, for those data sets that strife is performing best, these two feature ranking methods perform considerably worse, being 5 to 20%-points lower than strife's performance. In terms of the average feature subset size and the stability of the feature subsets, the Fisher score performs best in conjunction with the similarity classifier (see Table A.7). Besides that, strife leads for most data sets to competitive feature subset stabilities.

The results for the SVM classifier are presented in Table 10. For this classifier, there appears to be no filter method that tends to work consistently better than the other feature selection methods on all the data sets. Strife is for most data sets competitive to all remaining feature selection methods, including nonspecificity and total uncertainty, and in some cases clearly better than these two other proposed methods (Kidney, Leukemia, DLBCL). For six out of the ten real-world data sets at least one feature selection approach outperforms using no feature selection but not in even a single case this higher test accuracy is significantly larger. The Fisher score appears to generate the most stable feature subsets for this classifier (with exception of a very low ASM for the CTG-1 data set), but ReliefF and strife also show good results in terms of stability and average feature subset size (see Table A.8).

In terms of diversity, the results for the hybrid approach are similar to those presented in Table 6 for the feature subsets of fixed size. Even though the feature subsets generated by the hybrid approach are not of fixed size and, thus, are classifier dependent, the finding that nonspecificity and total uncertainty, the two ReliefF setups, as well as FSAE and FES are pairs of feature selection methods that generate similar subsets still holds true.

The average rank of the performance and stability on the real-world data sets is illustrated in Table 11 (Top 3 methods for each category are highlighted in gray). For the filter version where the feature ranking is used together with a fixed threshold on the number of features, the Fisher score, strife and ReliefF (70, 20) are the three methods that rank the highest with almost equal average ranks. In terms of stability, the Fisher score is the only of these three that also ranks within the top three most stable methods, with ReliefF and strife being clearly lower ranking in terms of stability when a fixed subset is used.

² The partial overlap of the first feature is reflected in the ordered similarity values. Some values for each class are closer to the opposing class ideal vector element than some of the own class's observations. In general, the larger the overlap of the feature values for different classes, the more the ordered similarity values will also be mixed (= not clearly differentiated by class).

³ The drop in the ordered similarity values for both classes shows that there is a set of observations that is much less well represented by the class representative than the other observations in that class. Hence, it is indicative of two distinct decision regions for each of the classes for that feature.

When using the hybrid approach where a classifier is used to determine the number of highest-ranking features to retain, again strife and the Fisher score are among the top three, whereas ReliefF is only ranked fifth and sixth.

The high average rank for the test accuracies obtained with the Fisher score and strife can be based on the fact that both methods compare the feature values of all observations to those feature values observed within the classes. The former does so by comparing the overall average feature value with the average value within a class, whereas strife compares the uncertainty of a feature value overall to the uncertainty of the feature value within the classes. Nonspecificity and total uncertainty indeed function in a similar manner, but the weights they use are considerably different. In particular, the weights for high similarity values, which belong to observations close to the class representatives (ideal vector) are comparably low, placing a larger emphasis on the middle and tail of the ordered similarity distribution. Since this is the main difference to strife, and since nonspecificity and total uncertainty often provide lower test accuracies, it can be concluded that placing comparably low weights on observations that are well represented by the class representative is the reason for the difference in the test accuracies. This would imply that considering observations that represent a class well in terms of the feature regarded are important for determining whether a feature is relevant. However, it is noteworthy that for the discrete MONK data sets that have a challenging logical target concepts, the opposite was true and nonspecificity and total uncertainty were clearly more accurate than strife. This stands in contrast to the results on the numerical artificial cases and real-world data sets for which strife overall outperforms nonspecificity and total uncertainty.

In terms of the average number of features selected, ReliefF together with the Fisher score form the top three whereas strife is only ranked fourth. It is noteworthy that the average rank of the feature subset size of strife is considerably better than that of nonspecificity and total uncertainty. In terms of stability for the hybrid approach, the top three results appear to be classifier-dependent, with only the Fisher score always being among the top three. With exception of the SVM, this holds true for strife with all classifiers as well. The ReliefF setups are also in several cases among the top three but seem on average not to work very well with the decision tree (rank six and seven).

6. Conclusions

In this paper, three feature ranking methods for supervised feature selection in the context of classification were proposed that are grounded in possibility theory. These methods, namely nonspecificity, strife and total uncertainty, and their implementation for feature ranking were depicted in detail. All three proposed methods were deployed on eighteen data sets, which included eight artificial data sets and ten medical real-world data sets. These data sets were selected so that they have a wide variety of characteristics. In particular, the real-world data sets possess between tens to thousands of observations, from few to thousands of features, from two to ten classes and strongly varying ratios of observations per feature. Two approaches were followed in this study: the first using the feature ranking with a fixed threshold for the number of highest-ranking features to select and the second following a hybrid approach by deploying a wrapper method on the feature ranking to select the number of features. Both approaches were included in a two-level external cross-validation procedure that avoids the selection bias and the parameter optimization bias. In order to benchmark and evaluate the results obtained by the three proposed methods, five other filter methods for feature selection (two ReliefF setups, Fisher score, FES, FSAE, and symmetrical uncertainty) as well as simply retaining the complete feature set (no feature selection), were included in the comparison of methods.

The classifiers included in this study were a KNN classifier, a binary decision tree, a similarity classifier as well as a linear SVM. In more general terms, the results showed that feature selection in many cases can

improve the classification accuracy, in a few cases even significantly, or lead to a comparable test accuracy with less features. The benefit appeared to be the largest for distance-based classifiers that include all features for their assignment of a class label, but improvements were also possible in some cases for the decision tree, which is an embedded feature selection method. In addition, the results on the ten real-world data sets demonstrated that strife and the Fisher score appear to be the feature selection methods that overall performed the best for both the filter and hybrid approach, with the two ReliefF setups and nonspecificity and total uncertainty in several cases also offering very good results. Besides that, strife uses most of the time a considerably smaller number of features than nonspecificity and total uncertainty while showing a comparable or better performance, indicating that it is likely the most capable of the three new proposed methods overall. This especially holds true in terms of numerical data, whereas the challenging logical target concepts of the artificial MONKs data sets were better addressed using nonspecificity and total uncertainty. In terms of stability, which was measured with the adjusted stability measure (ASM), the Fisher score and strife were the two most stable methods using the hybrid approach whereas the results using a fixed threshold were mixed. Apart from that, the proposed feature ranking method strife, and to some degree nonspecificity (or total uncertainty), shows low rank correlations and high diversity with the remaining feature ranking methods in this study.

A limitation of this study is that only selected classifiers were used, and that parameter tuning was only conducted for a single parameter for KNN and the decision tree. The reason for this course of action is the high computational complexity of using two-fold external cross-validation and testing all feature subset sizes by training classification models on each of them for the hybrid feature selection approach. Furthermore, the proposed filter methods such as strife were introduced as univariate filter methods that do not take into account conditional dependency between features for conducting the feature ranking. Different approaches to extend the univariate form to a multivariate one (e.g. using nearest neighbors as it is done for ReliefF) while keeping the increase in computational complexity at a reasonable level, can be analyzed in the future as well.

Funding

Finnish Strategic Research Council, grant number 313396/MFG40 Manufacturing 4.0.

Code availability

MATLAB code files for strife, nonspecificity and total uncertainty as well as the artificial examples are available on GitHub (<https://github.com/christophLUT/Nonspecificity-strife-and-total-uncertainty-in-supervised-feature-selection>).

CRediT authorship contribution statement

Christoph Lohrmann: Conceptualization, Methodology, Investigation, Validation, Software, Visualization, Writing – original draft. **Pasi Luukka:** Conceptualization, Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

⁴ The second class's ideal vector element leads to the lowest uncertainty for all observations (compared to the ideal vector elements for the first and third class) since the class is located closest to the center of the feature's range. Anyhow, for all three classes it is apparent that the uncertainty for all observations is considerably higher than for each class separately.

Availability of data and material

All real-world data sets are freely available online on GitHub and the UCI Repository for Machine Learning (sources are provided in the manuscript).

Appendix

Feature Ranking on the three artificial examples using nonspecificity, strife and total uncertainty

To illustrate the logic underlying the three proposed feature ranking methods, the feature ranking and scores for the nonspecificity, strife and total uncertainty filter on the three simple artificial examples (no data split) are displayed in Table A.2 (the highest-ranking feature and score is highlighted in gray). It is apparent that all three approaches rank the features correctly and according to expectation. Moreover, it is noteworthy that nonspecificity and total uncertainty are not only comparable in magnitude, but also take very similar values, whereas strife has values that are much lower in magnitude. This is based on the fact that the weights for nonspecificity are in general much higher than those of strife and, thus, for total uncertainty as well, which is the combination of nonspecificity and strife (Fig. A.1).

For the first artificial example, for the first feature both classes are well represented by their respective ideal vector element, whereas the opposing class is not well represented by it. Hence, the similarity values for observations of a class to their ideal vector element are high and of similar magnitude, whereas the similarity of observations to the opposing class's ideal vector element is considerably smaller and varies more (see Fig. A.2). For feature two the classes partially overlap and have a higher variation, leading to similarity values of observations to the own ideal vector that are on average not as high as for feature one. However, to the opposing ideal vector element the similarity values of observations are comparably higher than they were for feature one. Due to the overlap, some observations of each class are more similar to the opposing class's ideal vector element than to their own class's ideal vector element (see Fig. A.3). Due to that, for feature two the uncertainty within each class is higher, whereas the uncertainty for all observations is smaller than for feature one, leading overall to a lower difference in these two uncertainties. This holds true for all three approaches and results in the first feature being ranked higher.

For the second artificial example, the situation is similar, with the main difference being that the ideal vector elements for each class do not represent that class well in either dimension. It is apparent that the uncertainty within each class is higher than previously given that the observations in each class are on average less similar to the class ideal vector element and the similarity varies more (see Fig. A.4). For the first feature, the uncertainty within each of the classes is higher than for the second feature since the separate clusters (= groupings) in each class are further from one another. However, also the uncertainty for all observations is higher since the data is quite dispersed for the first feature in general. For the second feature the uncertainty within the class is comparably lower, but even more so is the uncertainty for all observations since there is some overlap between the pairs of clusters (see Fig. A.5). Since the first feature corresponds to a higher difference between the uncertainty overall versus the uncertainty within the class than the second feature, the first feature is ranked higher. This example illustrates that it is not always necessary for the proposed feature ranking methods that the ideal vector element represents the feature values a class takes well, as long as it represents the observations for that class better than it represents all observations.

The third artificial example is most similar to the first artificial example in that regard that each class has a single grouping of observations. For the first feature which does not display any overlap, the ideal vector element of each class represents the class well. The sole exception is class three, which displays considerable variation for feature one, which leads to a higher uncertainty for that feature in that

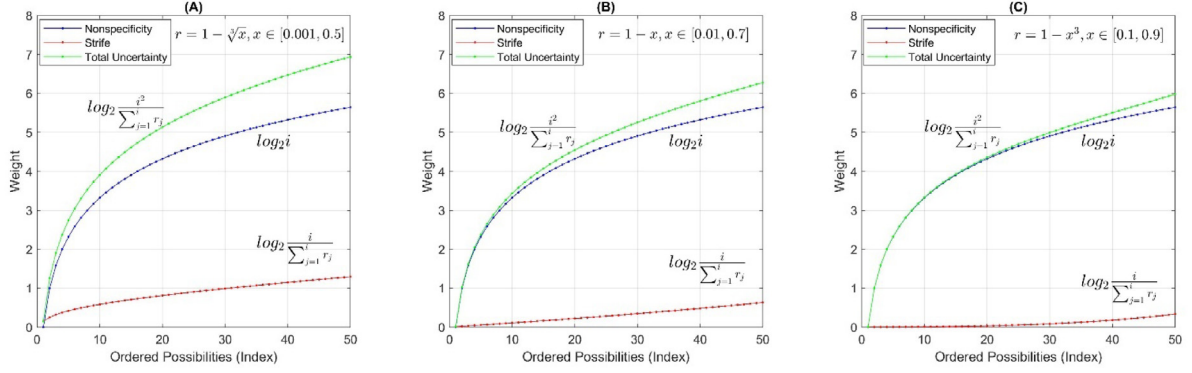


Fig. A.1. Weights of nonspecificity, strife and total uncertainty for (A) decreasing, (B) constant, (C) increasing differences in the possibility distribution.

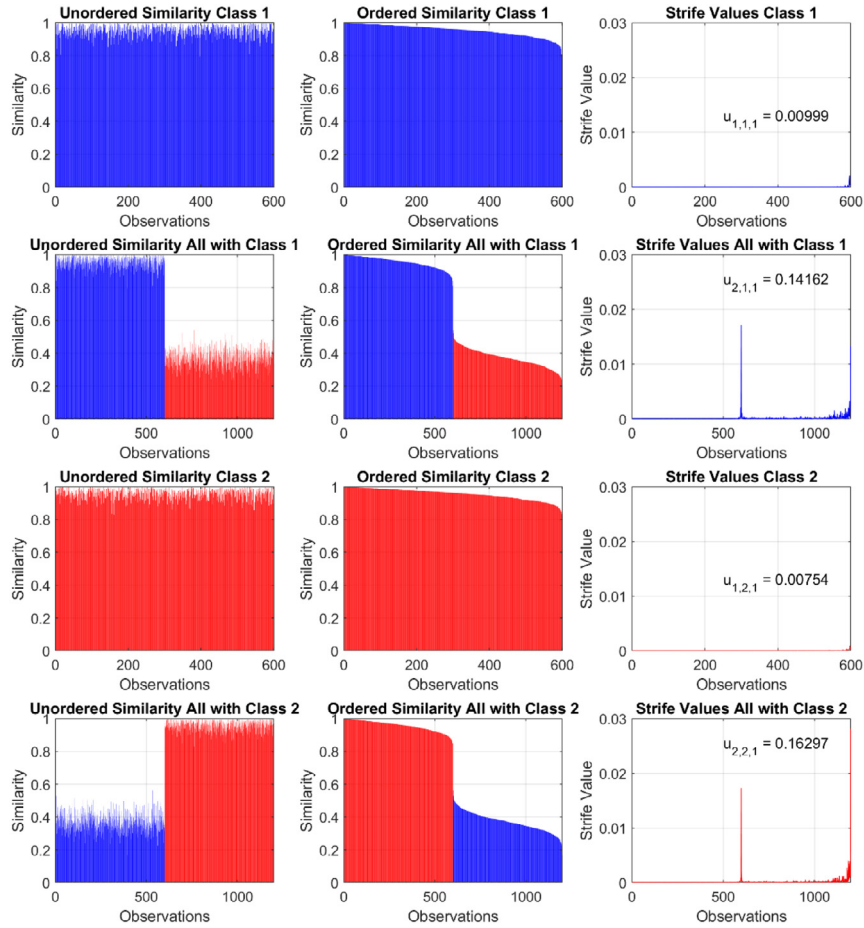
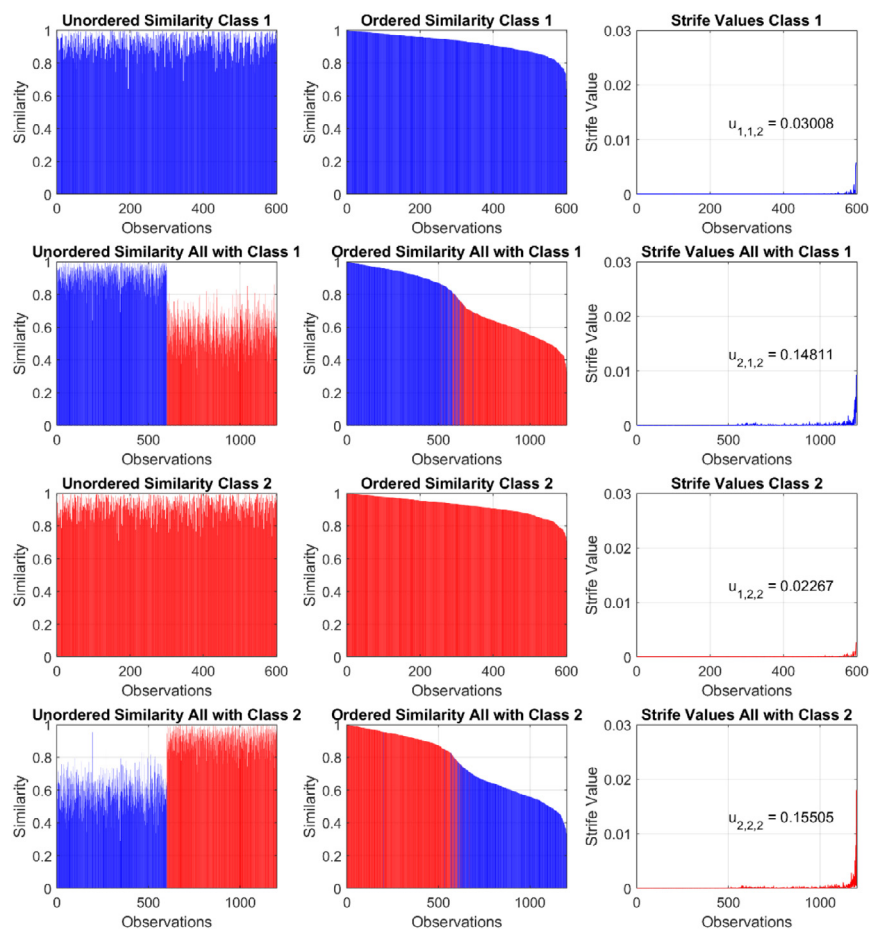


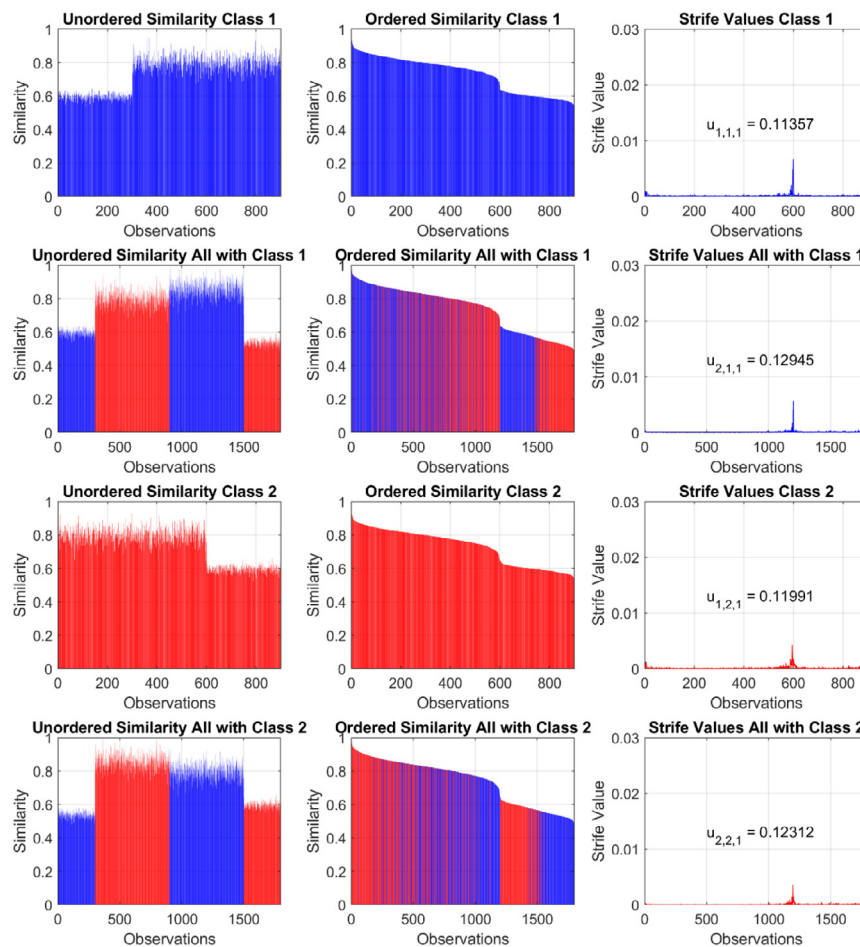
Fig. A.2. First artificial example strife components for feature 1.¹

class. Since the variation in the class is much smaller than for the entire data, especially for class one and class two, the difference in these uncertainties is quite high (see Fig. A.6). For the second feature, the variation for class three is considerably lower, leading to a markedly lower uncertainty within the class but also a higher uncertainty for all observations since these are not well represented by class three's ideal vector element (see Fig. A.7). However, the larger difference in the uncertainties for class three for the second feature is overcompensated by the fact that the uncertainty within class two is considerably higher

than for feature one, leading to higher uncertainty within the class and, thus, a lower difference in the uncertainties for all observations to those within the class. Hence, overall, all three feature ranking methods conclude that the first feature should be ranked higher than the second one.

Overall, the three artificial examples highlighted how the three proposed feature ranking methods work and that they can accomplish intuitive feature rankings.

Fig. A.3. First artificial example strife components for feature 2.²

Fig. A.4. Second artificial example strife components for feature 1.³

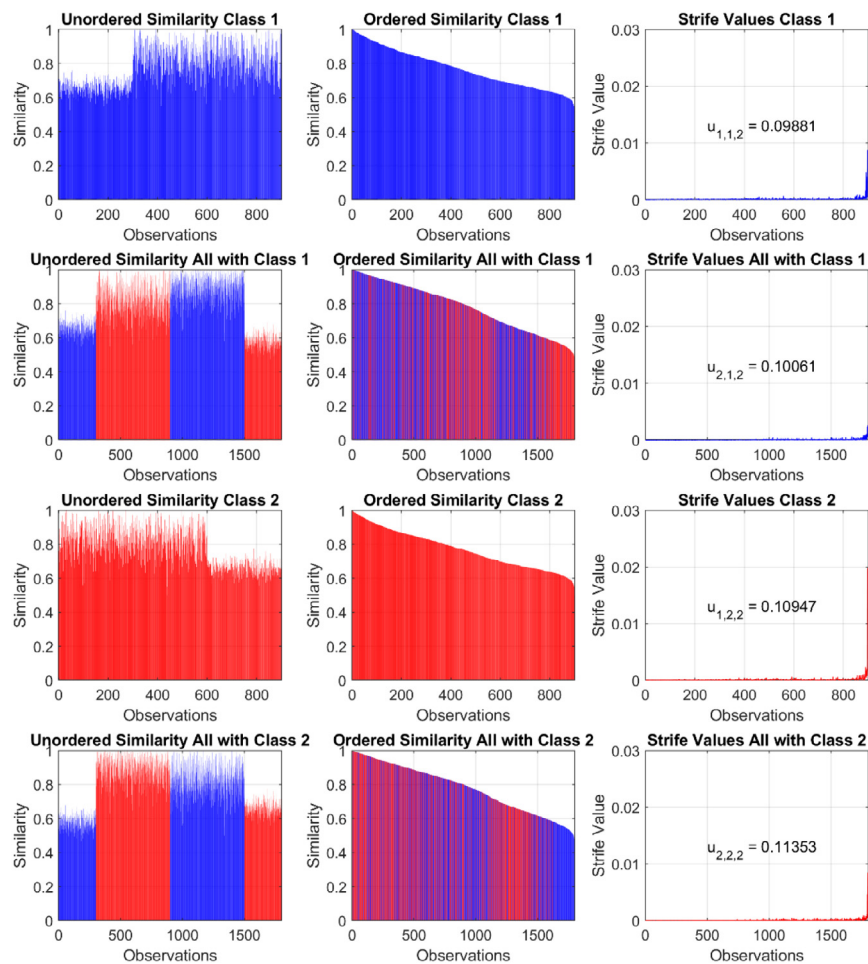
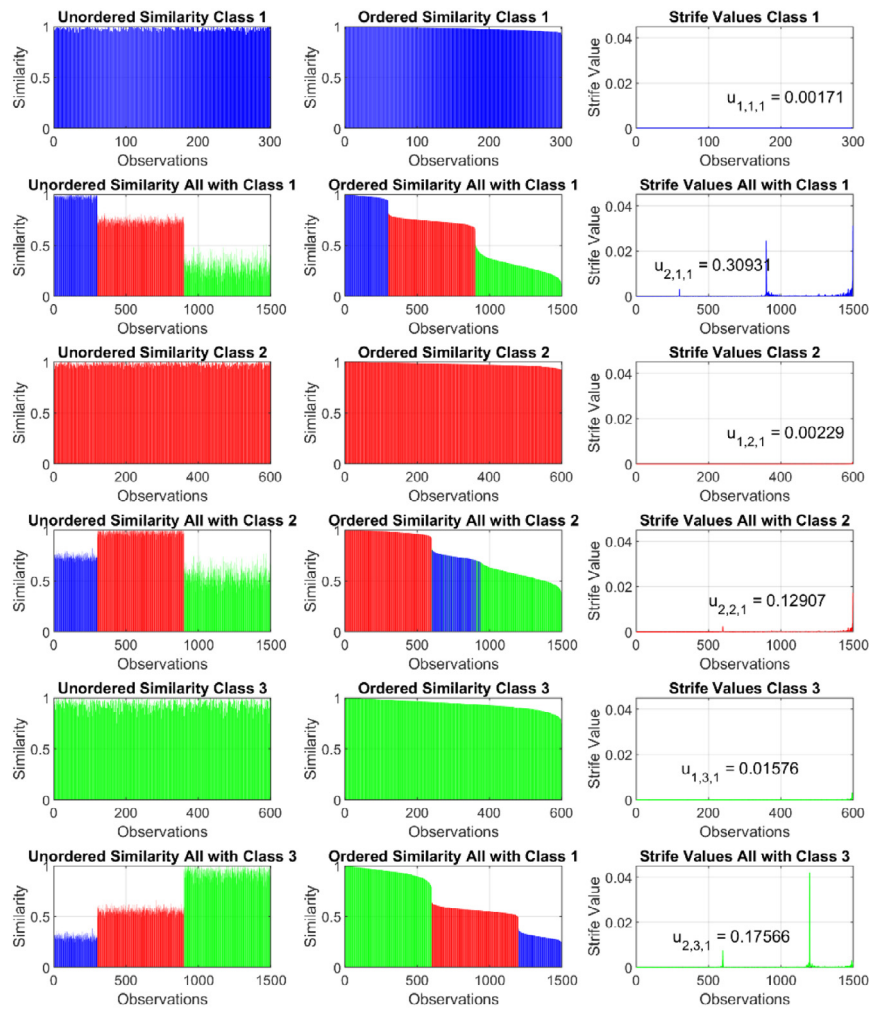


Fig. A.5. Second artificial example strife components for feature 2.

Fig. A.6. Third artificial example strife components for feature 1.⁴

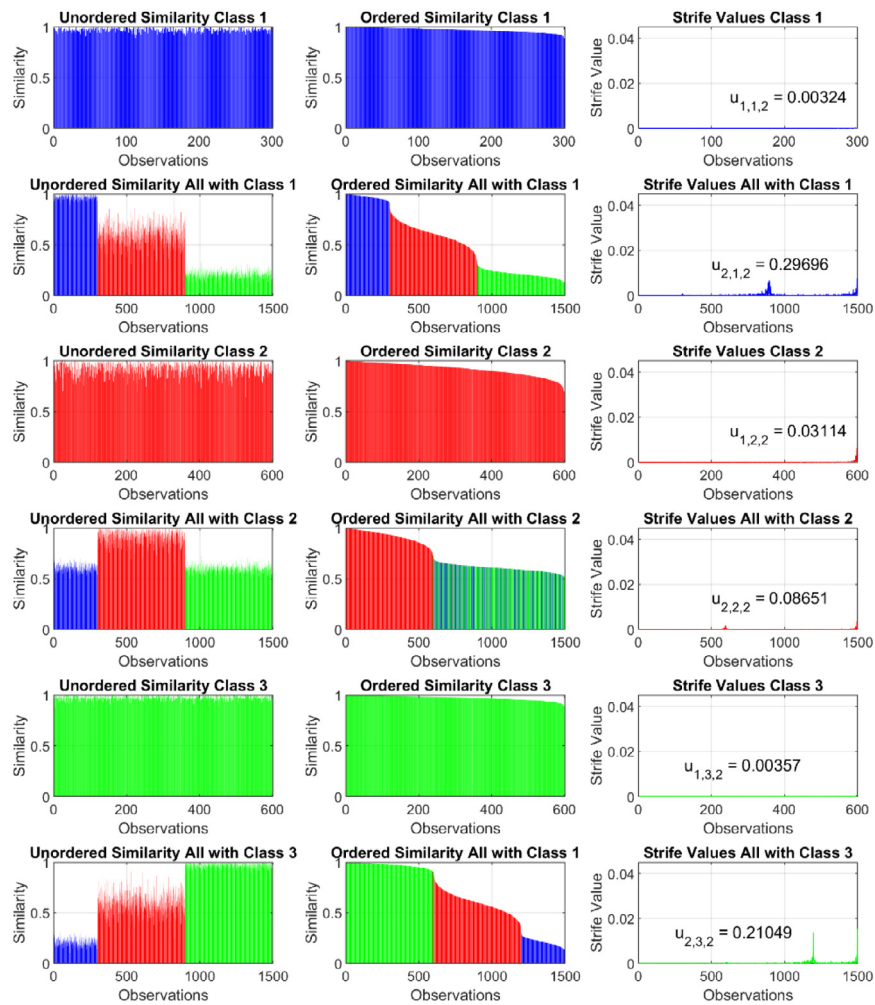


Fig. A.7. Third artificial example strife components for feature 2.

Table A.1

Specifications of the normal distributions for the artificial examples (before standardization).

Example	Class 1	Class 2	Class 3
First Artificial Example	$\mu_1 = \begin{pmatrix} 2 \\ 5 \end{pmatrix}$ $\Sigma_1 = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.5 \end{pmatrix}$ <i>600 Observations</i>	$\mu_2 = \begin{pmatrix} 7 \\ 2 \end{pmatrix}$ $\Sigma_2 = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.5 \end{pmatrix}$ <i>600 Observations</i>	
Second Artificial Example	$\mu_{1,1} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$ $\mu_{1,2} = \begin{pmatrix} 9 \\ 5 \end{pmatrix}$ $\Sigma_{1,1} = \begin{pmatrix} 0.05 & 0 \\ 0 & 0.05 \end{pmatrix}$ $\Sigma_{2,2} = \begin{pmatrix} 0.3 & 0 \\ 0 & 0.3 \end{pmatrix}$ <i>300 Observations</i> <i>600 Observations</i>	$\mu_{2,1} = \begin{pmatrix} 5 \\ 3.5 \end{pmatrix}$ $\mu_{2,2} = \begin{pmatrix} 12 \\ 6.5 \end{pmatrix}$ $\Sigma_{2,1} = \begin{pmatrix} 0.3 & 0 \\ 0 & 0.3 \end{pmatrix}$ $\Sigma_{2,2} = \begin{pmatrix} 0.05 & 0 \\ 0 & 0.05 \end{pmatrix}$ <i>600 Observations</i> <i>300 Observations</i>	
Third Artificial Example	$\mu_1 = \begin{pmatrix} 2 \\ 10 \end{pmatrix}$ $\Sigma_1 = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.1 \end{pmatrix}$ 300 Observations	$\mu_2 = \begin{pmatrix} 5 \\ 6.5 \end{pmatrix}$ $\Sigma_2 = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.8 \end{pmatrix}$ 600 Observations	$\mu_3 = \begin{pmatrix} 10 \\ 3 \end{pmatrix}$ $\Sigma_2 = \begin{pmatrix} 0.8 & 0 \\ 0 & 0.1 \end{pmatrix}$ <i>600 Observations</i>

Table A.2

Ranking and scores for the artificial examples.

Approach	Example 1		Example 2		Example 3	
	Ranking	Scores	Ranking	Scores	Ranking	Scores
Nonspecificity	1	11.9655	1	2.8740	1	19.1781
	2	7.6001	2	2.3602	2	16.5332
Strife	1	0.2871	1	0.0191	1	0.5943
	2	0.2504	2	0.0059	2	0.556
Total Uncertainty	1	12.2526	1	2.8931	1	19.7724
	2	7.8505	2	2.3661	2	17.0892

Table A.3

Adjusted Stability Measure (ASM) for fixed feature subsets for all real-world data sets.

Filter	Cryotherapy	Retinopathy	CTG-1	CTG-2	Kidney	Dermatology	SRBCT	Leukemia1	DLBCL	Prostate
ReliefF (10)	0.36	0.45	0.54	0.54	0.50	0.58	0.63	0.64	0.61	0.68
ReliefF (70,20)	0.31	0.55	0.59	0.59	0.46	0.58	0.63	0.67	0.56	0.69
FES	0.57	0.63	0.49	0.62	0.45	0.55	0.63	0.63	0.61	0.45
FSAE	0.67	0.63	0.59	0.62	0.54	0.60	0.64	0.67	0.61	0.45
Fisher Score	0.41	0.56	0.62	0.57	0.58	0.57	0.67	0.68	0.56	0.67
Nonspecificity	0.57	0.55	0.62	0.46	0.49	0.57	0.68	0.69	0.47	0.36
Strife	0.57	0.49	0.52	0.57	0.56	0.48	0.56	0.55	0.66	0.50
TU	0.57	0.55	0.62	0.46	0.49	0.57	0.68	0.69	0.47	0.36
SU	0.39	0.56	0.49	0.51	0.57	0.50	0.46	0.57	0.30	0.28
No FS	0	0	0	0	0	0	0	0	0	0

Table A.4

Average Spearman rank correlation over all real-world data sets.

Rank Corr.	ReliefF (10)	ReliefF (70,20)	FES	FSAE	Fisher Score	Nonspecificity	Strife	TU	SU
ReliefF (10)	1	0.3	0.0	0.1	0.0	0.0	0.0	0.0	0.0
ReliefF (70,20)	0.3	1	0.0	0.0	0.0	0.0	0.0	0.0	0.1
FES	0.0	0.0	1	0.3	0.0	0.3	0.0	0.3	0.0
FSAE	0.1	0.0	0.3	1	0.1	0.3	0.0	0.3	0.1
Fisher Score	0.0	0.0	0.0	0.1	1	0.0	0.1	0.0	0.0
Nonspecificity	0.0	0.0	0.3	0.3	0.0	1	0.0	0.6	0.1
Strife	0.0	0.0	0.0	0.0	0.1	0.0	1	0.0	-0.1
TU	0.0	0.0	0.3	0.3	0.0	0.6	0.0	1	0.1
SU	0.0	0.1	0.0	0.1	0.0	0.1	-0.1	0.1	1

Table A.5

Average feature subset size and Adjusted Stability Measure (ASM) for the KNN classifier using a hybrid approach.

Hybrid Approach	Cryotherapy		Retinopathy		CTG-1		CTG-2		Kidney		Dermatology		SRBCT		Leukemia1		DLBCL		Prostate	
	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM
KNN + ReliefF (10)	3.6	0.35	5.0	0.61	8.0	0.55	9.8	0.48	3.0	0.84	23.7	0.58	13.5	0.83	58.0	0.73	558.6	0.67	24.9	0.85
KNN + ReliefF (70,20)	3.5	0.39	5.9	0.59	8.6	0.53	8.0	0.51	2.8	0.84	22.7	0.57	17.6	0.83	107.4	0.76	552.6	0.66	6.8	0.87
KNN + FES	3.8	0.39	16.2	0.58	19.2	0.91	19.5	0.68	8.8	0.47	29.5	0.35	641.9	0.52	607.8	0.71	643.7	0.64	3971.3	0.45
KNN + FSAE	3.9	0.37	15.8	0.54	18.0	0.83	18.7	0.68	3.6	0.70	26.3	0.71	513.5	0.56	373.2	0.78	771.9	0.63	4288.2	0.46
KNN + Fisher Score	3.9	0.40	6.2	0.60	9.6	0.51	6.1	0.67	1.4	0.93	24.2	0.68	20.2	0.90	35.6	0.92	112.5	0.67	121.6	0.96
KNN + Nonspecificity	3.4	0.50	10.8	0.42	17.5	0.79	17.8	0.75	12.5	0.47	26.0	0.56	151.7	0.78	347.8	0.76	624.4	0.61	263.3	0.70
KNN + Strife	3.5	0.28	4.3	0.66	9.5	0.49	6.7	0.59	6.9	0.61	28.4	0.71	70.6	0.77	76.6	0.83	418.6	0.79	9.1	0.93
KNN + Total Uncertainty	3.4	0.50	10.8	0.42	16.8	0.72	17.7	0.74	12.3	0.46	26.0	0.56	151.7	0.78	283.4	0.77	663.0	0.60	199.7	0.69
KNN + SU	3.6	0.30	8.5	0.54	16.4	0.54	9.7	0.46	4.0	0.78	28.3	0.75	120.6	0.72	189.2	0.83	1435.3	0.31	1246.2	0.39
KNN + No FS	6	0	19	0	21	0	21	0	24	0	34	0	2308	0	5327	0	5469	0	10509	0

Table A.6

Average feature subset size and Adjusted Stability Measure (ASM) for the decision tree classifier using a hybrid approach.

Hybrid Approach	Cryotherapy		Retinopathy		CTG-1		CTG-2		Kidney		Dermatology		SRBCT		Leukemia1		DLBCL		Prostate	
	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM
DT + ReliefF (10)	2.9	0.49	7.4	0.47	15.9	0.68	14.7	0.59	2.0	0.91	23.1	0.63	4.8	0.87	3.8	0.58	34.6	0.65	63.4	0.93
DT + ReliefF (70,20)	3.0	0.49	5.5	0.61	16.0	0.69	15.6	0.64	1.9	0.91	22.0	0.59	4.6	0.86	4.3	0.58	28.4	0.69	68.2	0.95
DT + FES	2.9	0.43	15.0	0.68	19.4	0.91	19.4	0.72	7.4	0.57	23.9	0.65	760.0	0.45	1579.1	0.51	1774.8	0.42	4395.8	0.48
DT + FSAE	2.7	0.41	14.7	0.66	18.6	0.84	18.6	0.69	2.7	0.84	23.0	0.65	148.3	0.74	356.5	0.79	892.6	0.58	3648.3	0.52
DT + Fisher Score	2.0	0.63	4.8	0.66	13.8	0.57	11.8	0.49	1.0	0.96	23.7	0.66	10.1	0.85	11.1	0.87	13.9	0.50	35.2	0.99
DT + Nonspecificity	2.9	0.50	13.4	0.54	18.2	0.81	19.1	0.68	7.8	0.50	23.1	0.62	130.0	0.80	146.1	0.77	230.5	0.72	387.1	0.67
DT + Strife	1.9	0.61	2.7	0.84	17.0	0.72	9.2	0.48	3.9	0.68	19.3	0.61	93.1	0.75	48.7	0.82	19.3	0.81	21.7	0.96
DT + Total Uncertainty	2.9	0.50	13.4	0.54	18.2	0.81	19.1	0.68	7.2	0.45	23.0	0.62	109.9	0.81	131.6	0.76	286.4	0.75	81.2	0.67
DT + SU	2.6	0.51	9.1	0.48	17.2	0.79	16.8	0.39	3.1	0.81	29.9	0.56	48.4	0.91	10.0	0.75	79.0	0.48	55.0	0.38
DT + No FS	6	0	19	0	21	0	21	0	24	0	34	0	2308	0	5327	0	5469	0	10509	0

Table A.7

Average feature subset size and Adjusted Stability Measure (ASM) for the similarity classifier using a hybrid approach.

Hybrid Approach	Cryotherapy		Retinopathy		CTG-1		CTG-2		Kidney		Dermatology		SRBCT		Leukemia1		DLBCL		Prostate	
	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM
Sim + ReliefF (10)	1.5	0.68	1.5	0.87	12.5	0.38	7.2	0.58	4.4	0.74	22.7	0.58	23.9	0.85	8.2	0.80	182.7	0.64	16.6	0.90
Sim + ReliefF (70,20)	1.7	0.63	1.6	0.87	10.0	0.43	8.4	0.52	5.3	0.69	22.0	0.57	23.8	0.84	12.2	0.82	214.2	0.66	9.5	0.92
Sim + FES	3.4	0.43	11.0	0.50	15.4	0.71	17.6	0.76	3.0	0.83	30.3	0.35	1283.6	0.40	1674.6	0.52	1147.9	0.54	1007.7	0.79
Sim + FSAE	2.8	0.49	10.5	0.51	12.7	0.56	15.5	0.61	2.3	0.86	27.8	0.58	632.5	0.56	468.5	0.74	868.7	0.59	905.8	0.80
Sim + Fisher Score	1.4	0.71	1.0	0.95	8.4	0.54	5.0	0.73	2.6	0.84	25.7	0.71	15.2	0.84	24.6	0.89	9.7	0.51	3.3	0.98
Sim + Nonspecificity	3.6	0.40	7.0	0.56	12.3	0.52	8.3	0.56	13.5	0.44	28.0	0.24	343.9	0.67	565.2	0.73	114.2	0.58	30.5	0.66
Sim + Strife	1.7	0.61	1.1	0.94	6.4	0.62	6.4	0.65	6.7	0.51	24.2	0.62	66.6	0.83	68.9	0.82	247.1	0.86	5.0	0.93
Sim + Total Uncertainty	3.6	0.40	7.0	0.56	12.3	0.52	8.3	0.56	11.5	0.43	28.0	0.23	322.9	0.68	558.9	0.75	109.2	0.62	26.4	0.64
Sim + SU	3.0	0.50	3.8	0.75	15.4	0.67	7.8	0.54	1.8	0.90	29.4	0.61	77.1	0.78	210.0	0.78	25.0	0.44	57.5	0.43
Sim + No FS	6	0	19	0	21	0	21	0	24	0	34	0	2308	0	5327	0	5469	0	10509	0

Table A.8

Average feature subset size and Adjusted Stability Measure (ASM) for the support vector machine using a hybrid approach.

Hybrid Approach	Cryotherapy		Retinopathy		CTG-1		CTG-2		Kidney		Dermatology		SRBCT		Leukemia1		DLBCL		Prostate	
	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM	Avg.	ASM
SVM + ReliefF (10)	2.7	0.44	13.0	0.38	19.6	0.72	17.1	0.59	2.5	0.83	21.9	0.58	9.8	0.78	371.7	0.78	35.5	0.59	37.7	0.86
SVM + ReliefF (70,20)	2.8	0.44	11.4	0.43	19.2	0.69	15.7	0.68	2.8	0.85	22.7	0.58	9.5	0.80	288.0	0.77	28.9	0.60	26.3	0.82
SVM + FES	3.6	0.40	17.6	0.37	20.1	0.20	19.6	0.54	6.7	0.56	27.3	0.32	181.6	0.75	222.3	0.80	980.0	0.52	3025.7	0.54
SVM + FSAE	3.3	0.39	17.4	0.36	19.2	0.53	18.5	0.51	3.0	0.75	26.1	0.55	159.0	0.76	127.9	0.82	873.0	0.56	1933.7	0.63
SVM + Fisher Score	2.3	0.53	15.5	0.65	18.9	0.16	13.4	0.52	3.2	0.79	26.6	0.69	13.8	0.89	281.5	0.84	58.2	0.70	46.5	0.90
SVM + Nonspecificity	3.5	0.40	17.9	0.40	20.5	0.21	19.2	0.36	13.5	0.47	25.7	0.54	85.4	0.78	63.9	0.74	835.1	0.61	1425.9	0.56
SVM + Strife	2.2	0.54	15.8	0.30	19.0	0.39	18.9	0.39	4.6	0.74	23.8	0.60	33.1	0.72	91.3	0.70	110.8	0.82	19.5	0.87
SVM + Total Uncertainty	3.5	0.40	17.9	0.40	20.5	0.21	19.2	0.36	13.9	0.48	26.4	0.54	84.4	0.78	60.0	0.74	884.5	0.60	1378.0	0.56
SVM + SU	3.5	0.28	14.8	0.38	19.9	0.61	14.9	0.56	2.8	0.81	29.8	0.35	38.7	0.77	531.6	0.66	349.1	0.57	166.5	0.46
SVM + No FS	6	0	19	0	21	0	21	0	24	0	34	0	2308	0	5327	0	5469	0	10509	0

References

- Almuallim, H., Dietterich, T.G., 1991. Learning with many irrelevant features. In: Proc Ninth Natl Conf Artif. Intell., Vol. 91, pp. 547–552 <https://doi.org/10.5555/1865756.1865761>.
- Ambrose, C., McLachlan, G.J., 2002. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci U S A, <http://dx.doi.org/10.1073/pnas.102102699>.
- Antal, B., Hajdu, A., 2014. Diabetic retinopathy debrecen data set. UCI mach. Learn. Repos. <https://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy+Debrecen+Data+Set>.
- Bandemer, H., Näther, W., 1992. Fuzzy Data Analysis. Springer ScienceBusiness Media, Dordrecht.
- Bins, J., Draper, B.A., 2001. Feature selection from huge feature sets. Proc. IEEE Int. Conf. Comput. Vis. <http://dx.doi.org/10.1109/ICCV.2001.937619>.
- Blum, A.L., Langley, P., 1997. Selection of relevant features and examples in machine learning. Artif. Intell. 97, 245–271. [http://dx.doi.org/10.1016/S0004-3702\(97\)00063-5](http://dx.doi.org/10.1016/S0004-3702(97)00063-5).
- Bolón-Canedo, V., Alonso-Betanzos, A., 2018. Recent advances in ensembles for feature selection.
- Bolón-Canedo, V., Alonso-Betanzos, A., 2019. Ensembles for feature selection: A review and future trends. Inf. Fusion 52, 1–12. <http://dx.doi.org/10.1016/j.inffus.2018.11.008>.
- Bolón-Canedo, V., Sánchez-Marño, N., Alonso-Betanzos, A., 2013. A review of feature selection methods on synthetic data. Knowl. Inf. Syst. 34, 483–519. <http://dx.doi.org/10.1007/s10115-012-0487-8>.
- Bolón-Canedo, V., Sánchez-Marño, N., Alonso-Betanzos, A., 2014a. Data classification using an ensemble of filters. Neurocomputing 135, 13–20. <http://dx.doi.org/10.1016/j.neucom.2013.03.067>.
- Bolón-Canedo, V., Sánchez-Marño, N., Alonso-Betanzos, A., et al., 2014b. A review of microarray datasets and applied feature selection methods. Inf. Sci. (Ny) 282, 111–135. <http://dx.doi.org/10.1016/j.ins.2014.05.042>.
- Bonferroni, C., 1950. Sulle medie multiple di potenze. Bolletino. Mat. Ital. 5, 267–270.
- Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and regression trees.

- Caruana, R., Freitag, D., Greedy Attribute Selection. In: *Int. Conf. Mach. Learn.*, Vol. 48, pp. 28–36. doi: 10.1.1.41.3576.
- Caruana, R., Freitag, D., 1994b. How Useful is Relevance? *AAAI Fall Symp Tech Rep FS-94-02*.
- Chandrasekar, G., Sahin, F., 2014. A survey on feature selection methods. *Comput. Electr. Eng.* 40, 16–28. <http://dx.doi.org/10.1016/j.compeleceng.2013.11.024>.
- Cortes, C., Vapnik, V., 1995. Support-vector networks. *Mach. Learn.* 20, 273–297. <http://dx.doi.org/10.1023/A:1022627411411>.
- Cover, T.M., Hart, P.E., 1967. Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory* 13, 21–27. <http://dx.doi.org/10.1109/TIT.1967.1053964>.
- Das, S., 2001. Filters, wrappers and a boosting-based hybrid for feature selection. In: *Proc 18th Int. Conf. Mach. Learn.*, pp. 74–81. <https://doi.org/10.5555/645530.658297>.
- Dash, M., Liu, H., 1997. Feature selection for classification. *Intell. Data Anal.* 1, 131–156. <http://dx.doi.org/10.3233/IDA-1997-1302>.
- Dash, M., Liu, H., 2003. Consistency-based search in feature selection. *Artif. Intell.* 151, 155–176. [http://dx.doi.org/10.1016/S0004-3702\(03\)00079-1](http://dx.doi.org/10.1016/S0004-3702(03)00079-1).
- Dessi, N., Pes, B., 2015. Similarity of feature selection methods: An empirical study across data intensive classification tasks. *Expert. Syst. Appl.* 42, 4632–4642. <http://dx.doi.org/10.1016/j.eswa.2015.01.069>.
- Dua, D., Graff, C., 2019. UCI machine learning repository. UCI mach. Learn. Repos.
- Duch, W., 2006. Filter methods. In: Guyon, I., Nikravesh, M., Gunn, S., Zadeh, L.A. (Eds.), *Feature Extraction: Foundations and Applications*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 89–117.
- Duda, R.O., Hart, P.E., Stork, D.G., 2012. *Pattern Classification*. John Wiley Sons, Inc, New York.
- Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3, 1157–1182. <http://dx.doi.org/10.5555/944919.944968>.
- Hartley, R.V.L., 1928. Transmission of information. *Bell Syst. Tech. J.* <http://dx.doi.org/10.1002/j.1538-7305.1928.tb01236.x>.
- Hilario, M., Kalousis, A., 2008. Approaches to dimensionality reduction in proteomic biomarker studies. *Brief. Bioinform* 9, 102–118. <http://dx.doi.org/10.1093/bib/bbn005>.
- Ilter, N., Guvenir, H.A., 1998. Dermatology data set. UCI mach. Learn. Repos. <https://archive.ics.uci.edu/ml/datasets/Dermatology>. (Accessed 5 March 2017).
- John, G.H., Kohavi, R., Pfleger, K., 1994. Irrelevant features and the subset selection problem. In: *Machine Learning Proceedings*, pp 121–129.
- Khozeimeh, F., Layegh, P., Alizadehsani, R., Roshanzamir, M., 2018. Cryotherapy data set. UCI Mach. Learn.Repos <https://archive.ics.uci.edu/ml/datasets/Cryotherapy+Dataset+>. (Accessed 20 November 2019).
- Kira, K., Rendell, L., 1992a. A practical approach to feature selection.
- Kira, K., Rendell, L., 1992b. The feature selection problem: traditional methods and a new algorithm. In: *AAAI-92 Proc.* pp. 129–134.
- Klir, G.J., Yuan, B., 1995. *Fuzzy sets and fuzzy theory*. In: *Theory and Applications*. Prentice Hall, New Jersey.
- Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. *Artif. Intell.* 97, 273–324. [http://dx.doi.org/10.1016/S0004-3702\(97\)00043-X](http://dx.doi.org/10.1016/S0004-3702(97)00043-X).
- Kononenko, I., 1994. In: De Raedt, L., Bergadano, F. (Eds.), *Estimating attributes: analysis and extensions of Relief, ECLML-94*. Machine Learning. pp. 171–182. http://dx.doi.org/10.1007/3-540-57868-4_57.
- Kononenko, I., Simec, E., Robnik-Sikonja, M., 1997. Overcoming the myopia of inductive learning algorithms with RELIEFF. *Appl. Intell.* 7, 39–55. <http://dx.doi.org/10.1023/A:1008280620621>.
- Krawczuk, J., Łukaszuk, T., 2016. The feature selection bias problem in relation to high-dimensional gene data. *Artif. Intell. Med.* <http://dx.doi.org/10.1016/j.artmed.2015.11.001>.
- Kuncheva, L.I., 2007. A stability index for feature selection. In: *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications, AIA 2007*. <https://doi.org/10.5555/1295303.1295370>.
- Lazar, C., Taminiau, J., Meganck, S., et al., 2012. A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 9, 1106–1119. <http://dx.doi.org/10.1109/TCBB.2012.33>.
- Li, J., Cheng, K., Wang, S., et al., 2017. Feature selection: A data perspective. *ACM Comput. Surv.* 50 (94), <http://dx.doi.org/10.1145/3136625>.
- Liu, H., Motoda, H., 2001. *Feature extraction*. In: *Construction and Selection: A Data Mining Perspective*.
- Liu, H., Setiono, R., 1996. A probabilistic approach to feature selection - a filter solution. In: *Proc. 13th Int. Conf. Mach. Learn.* <https://doi.org/10.5555/3091696.3091735>.
- Liu, H., Yu, L., 2005. Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans. Knowl. Data Eng.* 17, 491–502. <http://dx.doi.org/10.1109/TKDE.2005.66>.
- Lohrmann, C., Luukka, P., 2020. MATLAB code files for strife, nonspecificity and total uncertainty. GitHub. <https://github.com/christophLUT/Nonspecificity-strife-and-total-uncertainty-in-supervised-feature-selection> (Accessed 24 August 2020).
- Lohrmann, C., Luukka, P., Jablonska-Sabuka, M., Kauranne, T., 2018. A combination of fuzzy similarity measures and fuzzy entropy measures for supervised feature selection. *Expert. Syst. Appl.* 110, <http://dx.doi.org/10.1016/j.eswa.2018.06.002>.
- Lukasiewicz, J., 1970. *Selected Work*. Cambridge University Press, Cambridge.
- Lustgarten, J.L., Gopalakrishnan, V., Visweswaran, S., 2009. Measuring stability of feature selection in biomedical datasets. In: *AMIA Annu Symp Proc.* pp. 406–410.
- Luukka, P., 2011. Feature selection using fuzzy entropy measures with similarity classifier. *Expert. Syst. Appl.* 38, 4600–4607. <http://dx.doi.org/10.1016/j.eswa.2010.09.133>.
- Luukka, P., Lohrmann, C., 2019. Information transmission and nonspecificity in feature selection. In: R., Kearfott, B. I., R. M., et al. (Eds.), *Fuzzy Techniques: Theory and Applications*. IFSA/NAFIPS 2019. *Advances in Intelligent Systems and Computing*, Vol 1000. Springer, pp. 340–350.
- Luukka, P., Saastamoinen, K., Könönen, V., 2001. A classifier based on the maximal fuzzy similarity in the generalized Lukasiewicz-structure. In: *10th IEEE Int Conf Fuzzy Syst.* <https://doi.org/10.1109/FUZZ.2001.1007281>.
- Miller, A., 2002. *Subset Selection in Regression*. Chapman & Hall/CRC, Boca Raton.
- Opitz, D.W., 1999. Feature selection for ensembles. In: *Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference*. American Association for Artificial Intelligence, USA, pp. 379–384.
- Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 1988. *Numerical Recipes in C. The Art of Scientific Computing*. Cambridge University Press, Cambridge.
- Primekangkang, 2019. Gene data. GitHub <https://github.com/primekangkang/Genedata>.
- Robnik-Šikonja, M., Kononenko, I., 2003. Theoretical and empirical analysis of ReliefF and RReliefF. *Mach. Learn.* 53, 23–69. <http://dx.doi.org/10.1023/A:1025667309714>.
- Roffo, G., Melzi, S., Castellani, U., Vinciarelli, A., 2017. Infinite latent feature selection: A probabilistic latent graph-based ranking approach. *Proc. IEEE Int. Conf. Comput. Vis.* 1, 1407–1415. <http://dx.doi.org/10.1109/ICCV.2017.156>.
- Marques de Sá, J.P., Bernardes, J., Ayres de Campos, D., 2010. Cardiotocography data set. UCI mach. Learn. Repos. <https://archive.ics.uci.edu/ml/datasets/Cardiotocography>. (Accessed 20 November 2019).
- Saeys, Y., Inza, I., Larrañaga, P., 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics* 23, 2507–2517. <http://dx.doi.org/10.1093/bioinformatics/btm344>.
- Shannon, C.E., 1948. A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423. <http://dx.doi.org/10.1145/584091.584093>.
- Singhi, S.K., Liu, H., 2006. Feature subset selection bias for classification learning. In: *ACM International Conference Proceeding Series*. <https://doi.org/10.1145/1143844.1143951>.
- Sosa-Cabrera, G., García-Torres, M., Gómez-Guerrero, S., et al., 2019. A multivariate approach to the symmetrical uncertainty measure: Application to feature selection problem. *Inf. Sci. (Ny)* 494, 1–20. <http://dx.doi.org/10.1016/j.ins.2019.04.046>.
- Soundarapandian, P., Rubini, L., 2015. Chronic kidney disease data set. UCI mach. Learn. Repos. https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease. (Accessed 5 March 2017).
- Souza, J.T., Matwin, S., Japkowicz, N., 2006. Parallelizing feature selection. *Algorithmica* 45, 433–456. <http://dx.doi.org/10.5555/3118749.3118952>.
- Thrun, S., 1992. MONK's problems data set. UCI mach. Learn. Repos. <https://archive.ics.uci.edu/ml/datasets/MONK's+Problems>. (Accessed 19 May 2021).
- Varma, S., Simon, R., 2006. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* <http://dx.doi.org/10.1186/1471-2105-7-91>.
- Vergara, J.R., Estévez, P.A., 2014. A review of feature selection methods based on mutual information. *Neural Comput. Appl.* 24, 175–186. <http://dx.doi.org/10.1007/s00521-013-1368-0>.
- Wood, I.A., Visscher, P.M., Mengersen, K.L., 2007. Classification based upon gene expression data: Bias and precision of error rates. *Bioinformatics* <http://dx.doi.org/10.1093/bioinformatics/btm117>.
- Yager, R.R., 1988. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Trans. Syst. Man Cybern.* 18, 183–190. <http://dx.doi.org/10.1109/21.87068>.
- Yang, P., Yang, Y.H., Zhou, B., Zomaya, A.Y., 2010. A review of ensemble methods in bioinformatics. *Curr. Bioinform* 5, 296–308. <http://dx.doi.org/10.2174/157489310794072508>.
- Yu, L., Liu, H., 2004. Efficient feature selection via analysis of relevance and redundancy. *J. Mach. Learn. Res.* 5, 1205–1224. <http://dx.doi.org/10.5555/1005332.1044700>.