Feature Selection Using Approximate Multivariate Markov Blankets

Rafael Arias-Michel[1], Miguel García-Torres[2(✉)], Christian Schaerer[1], and Federico Divina[2]

[1] Facultad Politécnica, Universidad Nacional de Asunción, P.O. BOX 2111 SL, San Lorenzo, Paraguay
cshaer@pol.una.py
[2] Computer Science, Universidad Pablo de Olavide, 41013 Seville, Spain
{mgarciat,fdivina}@upo.es

Abstract. In classification tasks, feature selection has become an important research area. In general, the performance of a classifier is intrinsically affected by existence of irrelevant and redundant features. In order to find an optimal subset of features, Markov blanket discovery can be used to identify such subset. The Approximate Markov blanket (AMb) is a standard approach to induce Markov blankets from data. However, this approach considers only pairwise comparisons of features. In this paper, we introduce a multivariate approach to the AMb definition, called Approximate Multivariate Markov blanket (AMMb), which takes into account interactions among different features of a given subset. In order to test the AMMb, we consider a backward strategy similar to the Fast Correlation Based Filter (FCBF), which incorporates our proposal. The resulting algorithm, named as FCBF_{ntc}, is compared against the FCBF, Best First (BF) and Sequential Forward Selection (SFS) and tested on both synthetic and real-world datasets. Results show that the inclusion of interactions among features in a subset may yield smaller subsets of features without degrading the classification task.

1 Introduction

In classification problems, the *course of dimensionality* refers to the negative effect on the performance of a classifier when applying on a dataset with too many features. This is due to various reasons, e.g., the search space to explore has a too high dimensionality and that many features may be irrelevant or even redundant, introducing in this way noise in the dataset. In this context, we say that a feature is irrelevant if it does not provide any information on the class, and it is, therefore, not needed for the classification. The problem is how to identify a feature as irrelevant. Different approaches, e.g., [1,2,5,8] have addressed this problem, by considering the relevance of a feature on its contribution to the meaning of the class concept. In this context, feature relevance has arisen as a measure of the amount of relevant information that a feature may contain about the class in classification tasks.

F. Martínez-Álvarez et al. (Eds.): HAIS 2016, LNAI 9648, pp. 114–125, 2016.
DOI: 10.1007/978-3-319-32034-2_10

It follows that, in many cases, a preprocessing phase aimed at the selection of non-redundant and relevant features becomes necessary. Moreover, such a dimensionality reduction would be beneficial also for datasets of lower dimensionality.

Different strategies have been proposed for addressing the above mentioned problems. There are two main approaches used in feature selection: the wrapper and the filter approach [7]. In the wrapper approach, a classifier is used in order to estimate the quality of the selected features. The main advantage of such approach is that the feature selection phase benefits from the direct feedback provided by the classifier. However, the main drawback is that such methods are computationally expensive. Moreover, there is the risk of overfitting. On the other hand, strategies based on the filter approach estimate the quality of the selected features by using some estimations based on the properties of the data, without using any classifiers to this aim. Thus, filter approaches are less computationally expensive than wrapper approaches, and can deal with high-dimension datasets. This fact represents a clear advantage. However, the ausence of a classifier can imply in lower classification results.

The Fast Correlation Based Filter (FCBF) [16] approach has shown excellent performance when applied to high dimensional datasets, being able to select a set of non-redundant and informative features. In particular, this method relies on the concept of Approximate Markov blanket (AMb), proposed in [8] in order to detect redundancy and on the Symmetrical Uncertainty (SU) in order for detecting dependencies among features.

Even if the FCBF has achieved good results, it presents an important limitation: it only considers interaction that can take place between pairs of features, but does not consider interactions among different features. In this work we propose a feature selection strategy that aims at solving this limitation.

Our approach is based on the AMb, and it aims at modifying it in order to consider possible interactions that can exist among different features, and not only interactions among pairs of features. To this aim, we propose to redefine the AMb and use the SU and a normalized version of the total correlation [12,15] (NTC) as cost functions. Then, we use this definition in a variant of the FCBF algorithm denoted as FCBF$_{ntc}$. Results obtained on synthetic and real datasets confirm the effectiveness and potential of the proposed strategy.

The rest of the paper is organized as follows. In Sect. 2 we provide the basements the theoretic foundations of our proposal and a description of FCBF$_{ntc}$. The data used in this work are introduced in Sect. 3. Then, Sect. 4 provides an experimental validation of our proposal. In Sect. 5 we present the main conclusions dentifying possible future developments.

2 Theoretical Foundations

In this paper, we will use the following notation. $\mathcal{X}$ denotes the n dimensional set of features, while $x_i \in \mathcal{X}$ is used for representing its elements. $\mathcal{E}$ stands for the set of samples, and a single sample is represented by the pair (x_i, y_i)

where $y_i \in \mathcal{Y}$ is a known class label of x_i. The classification mapping F is denoted as:

$$F : \mathcal{X} \rightarrow \mathcal{Y}. \tag{1}$$

Feature selection can be formulated as the problem of finding a subset of features $\mathcal{S} \subset \mathcal{X}$ that minimizes a given cost function $J(\mathcal{X}, \mathcal{Y})$, with respect to the expression (1), i.e.:

$$\min_{x_i \in \mathcal{X},\, y_i \in \mathcal{Y}} J(x_i, y_i) \tag{2}$$

$$\text{subject to } F(x_i) = y_i. \tag{3}$$

Typically, feature selection algorithms aim at finding a set of features that minimize the cost function $J(\mathcal{X}, \mathcal{Y})$.

Several works [1,2,5,8] have made an effort for classifying the features according to the contribution to the meaning of the class concept. In this context, feature relevance has arisen as a measure of the amount of relevant information that a feature may contain about the class in classification tasks.

In this context, a feature is considered irrelevant if it contains no information about the class and therefore it is not necessary at all for the predictive task. Removing this type of features may improve the predictive model as well as the speed of the learning algorithm. In contrast relevant features are those that embodies information about the class concept. However, for minimizing the error rate it may not be necessary to select all relevant features; as a subset with the most predictive power may be sufficient. Furthermore, such subset of features may not be unique due to redundancy.

In order to identify redundant features, Holler and Sahami [8] proposed to use the concept of feature Markov blanket.

Definition 1 (Markov blanket). *Given a feature x_i, $\mathcal{M}_i \subset \mathcal{X}$ ($x_i \notin \mathcal{M}_i$) is said to be a Markov blanket for x_i iff*

$$P(\mathcal{X} - \mathcal{M}_i - \{x_i\}, \mathcal{Y} \,|\, x_i, \mathcal{M}_i) = P(\mathcal{X} - \mathcal{M}_i - \{x_i\}, \mathcal{Y} | \mathcal{M}_i). \tag{4}$$

According to this definition, a set of features $\mathcal{M}_i$ is a Markov blanket for a feature x_i if x_i is conditionally independent of $\mathcal{X} - \mathcal{M}_i - \{x_i\}$. If so, then x_i is also conditionally independent of $\mathcal{Y}$. Such condition is stronger than the conditional independence between x_i and $\mathcal{Y}$ given $\mathcal{M}_i$. It requires that $\mathcal{M}_i$ subsume not only the information that x_i has about $\mathcal{Y}$ but also about all of the other features. Therefore, given a subset $S \subseteq \mathcal{X}$, a feature $x_i \in \mathcal{S}$ can be removed from $\mathcal{S}$ if we find a Markov blanket $\mathcal{M}_i$ for x_i within $\mathcal{S}$. In this case we can say that x_i is a redundant feature of $\mathcal{S}$ and so removing it from the subset will not affect the predictive power of the classification model.

2.1 Bivariate Approach for Feature Redundancy

Redundancy is generally defined in terms of feature correlation and it is widely accepted that two features are redundant if their values are correlated. However, linear correlations may not be sufficient to detect non-linear dependencies

between features. In this work we use a non-linear correlation measure based on entropy. By considering each feature as a random variable, the uncertainty about the values of a random variable (r.v.) X is measured by its entropy $H(X)$, where $H(X)$ is defined as:

$$H(X) := -\sum_i P(x_i)\log_2(P(x_i)), \qquad \forall i = 1, \ldots, n. \tag{5}$$

Given another random variable Y, the conditional entropy $H(X|Y)$ measures the uncertainty about the value of X given the value of Y and is defined as:

$$H(X|Y) := -\sum_j P(y_j)\sum_i P(x_i|y_j)\log_2(P(x_i|y_j)), \tag{6}$$

where $P(y_j)$ is the prior probability of the value y_j of Y, and $P(x_i|y_j)$ is the posterior probability of a given value x_i of variable X given the value of Y. Information Gain [13] of a given variable X with respect to variable Y (IG(Y;X)) measures the reduction in uncertainty about the value of X given the value of Y and is defined as:

$$IG(X|Y) := H(X) - H(X|Y). \tag{7}$$

Therefore IG can be used as correlation measure. For instance, given the random variables X, Y and Z, X is considered to be more correlated to Y than Z, if $IG(Y|X) > IG(Z|X)$. IG is a symmetrical measure; which is a desired property for a correlation measure. However it is biased in favor of r.v. with more values. Such values have to be normalized to ensure the values are comparable with each other. In order to do so, IG can be normalized using the corresponding entropies. To this aim, the Symmetrical Uncertainty (SU) measure can be used:

$$SU(X,Y) := 2\left[\frac{IG(X|Y)}{H(X)+H(Y)}\right]. \tag{8}$$

SU is preferred to IG since it compensates for IG's bias and restricts its values to the range $[0, 1]$. A value of 1 indicates that X and Y are completely correlated, while a value of 0 indicates that X and Y are independent. Therefore, SU can be used as a correlation measure between features.

Based on the SU correlation measure, authors of [16] introduced the Approximate Markov blanket concept to analyze the feature redundancy:

Definition 2 (Approximate Markov blanket (AMb)). *Given two features X_i and X_j $(i \neq j)$ so that $SU(X_j, \mathcal{Y}) \geq SU(X_i, \mathcal{Y})$, and given a class label set $\mathcal{Y}$ then X_j forms an approximate Markov blanket for X_i iff $SU(X_i, X_j) \geq SU(X_i, \mathcal{Y})$.*

Based on the definitions above, the authors proposed the fast correlation based filter - FCBF - algorithm, whose pseudocode is the following:

Algorithm 1. FCBF

```
   Input: S(X_1, X_2, ..., X_N, Y)                          // a training data set
          δ                                                 // a predefined threshold
   Output: S_best                                           // a selected subset
 1 begin
 2   for i ← 1 to N do
 3     calculate SU(X_i, Y) for X_i;
 4     if SU(X_i, C) > δ then
 5       append X_i to S'_list;
 6     end if
 7   end for
 8   order S'_list in descending SU(X_i, Y) value;
 9   X_j ← getFirstElement(S'_list);
10   repeat
11     X_i ← getNextElement(S'_list, X_j);
12     if X_i ≠ NULL then
13       repeat
14         if SU(X_i, X_j) ≥ SU(X_i, Y) then
15           remove X_i from S'_list;
16         end if
17         X_i ← getNextElement(S'_list, X_i);
18       until X_i = NULL;
19     end if
20     X_j ← getNextElement(S'_list, X_j);
21   until X_j = NULL;
22   S_best = S'_list;
23 end
```

As it can be seen, the algorithm starts by ordering the features according to their symmetrical uncertainty with respect to the class. In this step, a feature is considered relevant if it exceeds a predefined threshold. Let X_1 be the first feature from S'_{list}. Then the algorithm verifies if X_1 is a AMb of $X_i (i > 1)$. If it is, then X_i is removed from the set of relevant features. The above process is repeated until there are no features in S'_{list}.

2.2 Multivariate Approach

In order to assess the dependency among features from a subset, the concept of total correlation [12,15] provides an effective way to compute it. Given $X_1, \ldots, X_n$, denoted as $X_{1:n}$, the total correlation is defined as follows:

$$C(X_{1:n}) := \sum_{i=1}^{n} H(X_i) - H(X_{1:n}) \tag{9}$$

with $H(X_{1:n})$ the mulivariate joint entropy. It can be noted that for $n = 2$, the total correlation is equivalent to the bivariate mutual information.

In order to restrict the values of $C(X_{1:n})$ to the range $[0, 1]$, as in the case of SU, we have to normalize. The Normalized Total Correlation (NTC) can be defined as:

$$NTC(X_{1:n}) := \frac{C(X_{1:n}) - C_{min}}{C_{max} - C_{min}} \tag{10}$$

The maximum total correlation occurs when one of the variables determines all of the other variables and is given by Eq. 11, while the minimum is given by Eq. 12.

$$C_{max} := \sum_{i=1}^{n} H(X_i) - \max_{X_i} H(X_i) \tag{11}$$

$$C_{min} := \{0; \sum_{i=1}^{n} H(X_i) - \log_2 m\} \tag{12}$$

With m the sample size. If the sample is large enough, $C_{min} = 0$. However, the number of samples necessary increase exponentially with the number of features. In general samples are not so large and, therefore, $C_{min} \neq 0$. The lower values total correlation can obtain in these cases is given by $C(X_{1:n}) = \sum_{i=1}^{n} H(X_i) - \log_2 m$. However, This expression can be negative if m is large enough. Threfore, we have to select between 0 and the minimum possible value, Now we can use this measure to define the Approximate Multivariate Markov blanket (AMMb) as follows:

Definition 3 (Approximate Multivariate Markov blanket (AMMb)). *Given a feature X_i and a subset S_j ($X_i \notin S_j$), and given a class label $\mathcal{Y}$ then S_j forms an approximate Markov blanket for X_i iff $NTC(X_i, \mathcal{Y}) \leq NTC(X_i, S_j)$.*

We can use the AMMb in the FCBF algorithm. The pseudocode of the resulting algorithm, denoted as FCBF$_{ntc}$, is presented in Algorithm 2 and differs, from the original strategy in the us of NTC instead of SU and in the following. Let X_1 be the first feature from S'_{list}. Then the algorithm verifies if X_1 is a AMMb

Algorithm 2. FCBF$_{ntc}$

```
   Input: S(X_1, X_2, ..., X_N, Y)                          // a training data set
          δ                                                 // a predefined threshold
   Output: S_best                                           // a selected subset
 1 begin
 2     for i ← 1 to N do
 3         calculate NTC(X_i, Y) for X_i;
 4         if NTC(X_i, C) > δ then
 5             append X_i to S'_list;
 6         end if
 7     end for
 8     order S'_list in descending SU(X_i, Y) value;
 9     X_j ← getFirstElement(S'_list);
10     S ← X_j;
11     repeat
12         X_i ← getNextElement(S'_list, X_j);
13         if X_i ≠ NULL then
14             repeat
15                 if NTC(X_i, S) ≥ NTC(X_i, Y) then
16                     remove X_i from S'_list;
17                 end if
18                 X_i ← getNextElement(S'_list, X_i);
19             until X_i = NULL;
20         end if
21         X_j ← getNextElement(S'_list, X_j);
22         S ← X_j;
23     until X_j = NULL;
24     S_best = S'_list;
25 end
```

of $X_i (i > 1)$. If it is, then X_i is removed from the set of relevant features. In the second iteration, let X_3 be the next feature from the S'_{list}. Then, the algorithm verifies if $\{X_1, X_3\}$ is a AMMb of $X_i (i > 3)$ and this process is repeated until there are no features in S'_{list}.

3 Data

This section describes the datasets used in this work for assessing the goodness of our proposal. First, we analyze the AMb approach using synthetic data. This kind of data provides a controlled environment scenario for analyzing potencial strengths and limitations of the proposal. In a second step, we tested the proposal on real datasets taken from the UCI repository [11]. The characteristics of the datasets are explained below.

3.1 Synthetic Datasets

We use the LED synthetic data, which consist of 7 boolean features and 10 concepts, the set of decimal digits. LED displays contain 7 light-emitting diodes and so all the 7 features are relevant. We add irrelevant features for testing if the proposal is able to find the relevant ones. We also add noise in order to further test our approach. Given a percentage of noise P_n a feature will have a a probability equal to P_n of being inverted. We generate several versions of this dataset as specified in Table 1.

Table 1. Summary of synthetic datasets.

dataset	#samples	#irr. features	noise (%)
LED	1000	$\{0, 50, 100\}$	$\{0, 10, 20\}$

3.2 UCI Datasets

Table 2 summarises the characteristics of the chosen datasets. The first two columns correspond to the name of the datasets as it appears in the UCI repository and the identifier (id) used in forthcoming tables. The following two columns show the total number of instances and the number of features. Finally, the last column presents the number of labels.

4 Experiments and Results

This section describes the experiments performed for testing our proposal. In particular, the objectives of the experimentation are:

Table 2. Characteristics of the datasets.

Dataset	id	#inst.	#feat.	#labels
Balance scale	bsc	625	4	3
Nursery	nur	12960	8	5
Pima diabetes	pdi	768	8	2
Page blocks	pbl	5473	10	5
Wine	win	178	13	3
Ionosphere	ion	351	34	2
Spambase	spa	4601	57	2
Sonar	son	208	60	2

– to evaluate the proposed multivariate AMb approach and compare it with the original proposal.
– to test the performance of $FCBF_{ntc}$ compared to state of the art algorithms from the feature selection problem.

As already stated, in the first experiment, we use synthetic datasets and analyze the number of relevant features identified, as well as the total number of features selected. Since we want to compare the proposed measure with the orginal one, we compare the FCBF_{ntc} with the FCBF. In the second set of experiments, we considered real world datasets. Classification error, the number of features selected and the robustness of the solutions found are used for assessing the performances of the different strategies tested. 10 fold cross-validation is used in order to assess model quality. In this case we compare the proposal with FCBF, Best First (BF) and Sequential Forward Selection (SFS).

Robustness or stability [6,9,10] of feature subset selection strategies measures the sensitivity to variations of a feature selection algorithm. We quantify the robustness with the Jaccard index [14], which is defined as the size of the intersection divided by the size of the union of the sets. Let A and B be subsets of features such that $A, B \subseteq \mathcal{X}$. The Jaccard index for such subsets $\mathcal{I}_J(A, B)$ is defined as:

$$\mathcal{I}_J(A, B) = \frac{|A \cap B|}{|A \cup B|}. \tag{13}$$

Given a set of solutions $\mathcal{S} = \{S_1, \ldots, S_m\}$, the approach for estimating the stability, $\Sigma(S)$, among this set of solutions consists of averaging the pairwise $\mathcal{I}_J(\cdot, \cdot)$ (Σ) similarities

$$\Sigma(\mathcal{S}) = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \mathcal{I}_J(S_i, S_j).$$

Higher values correspond to more stable subsets.

For computing the classification error, we use the Naive Bayes classifier due to its popularity and good results achieved. We compare our proposal against

the Sequential Forward Selection (SFS) and the Best First (BF) due to their well performance in general. All experiments have been developed using `Weka` [4], and the source code is available upon request.

Finally, we applied statistical tests to support the conclusions. Following the guidelines proposed by Demšar [3], we apply the Wilcoxon signed-ranks test, which is a non-parametric alternative to the paired t-test.

4.1 Synthetic Datasets

Table 3 shows the results obtained by FCBF$_{ntc}$ and FCBF on synthetic datasets. The first column refers to the number of features. Then, the percentage of noise followed by the number of features selected by each algorithm. As we can see, both algorithms find the 7 relevant features except for FCBF that fails on the first dataset. However, FCBF adds irrelevant features when increasing the complexity of the dataset, both the dimensionality and the noise.

Table 3. Summary of the results achieved by FCBF$_{ntc}$ and FCBF on synthetic datasets.

#feats.	noise (%)	FCBF$_{ntc}$	FCBF
7	0	7	6
	10	7	7
	20	7	7
50	0	7	22
	10	7	23
	20	7	24
100	0	7	30
	10	7	31
	20	7	32

4.2 UCI Datasets

Table 4 presents the accuracy achieved by each feature selection algorithm. The first column refers to the id of the dataset. Then, for each algorithm, the mean accuracy achieved on each dataset is reported, together with its respective standard deviation.

In general, all algorithms show similar results except on the nur, spa and son datasets. On nur, FCBF$_{ntc}$ and FCBF outperform BF and SFS. On spa, BF and SFS achieve better result while on son, FCBF is the best algorithm. However, on average, all algorithms achieve very close values. Moreover, differences are not statistical significant.

The size of the solutions found is presented in Table 5. As it can be noticed, the FCBF$_{ntc}$ is the algorithm the achieves, on average, the higher reduction.

Table 4. Mean accuracy values with their respective standard deviation obtained by NB classifier after applying algorithms FCBF$_{ntc}$, FCBF, BF and SFS.

id	Accuracy			
	FCBF$_{ntc}$	FCBF	BF	SFS
bsc	90.56 ± 1.75	90.56 ± 1.75	90.56 ± 1.75	90.56 ± 1.75
nur	89.90 ± 0.73	90.31 ± 0.50	70.97 ± 1.02	70.97 ± 1.02
pdi	76.17 ± 4.54	77.21 ± 4.18	76.04 ± 4.64	76.04 ± 4.64
hpbl	94.13 ± 0.87	93.20 ± 1.26	94.52 ± 0.52	94.52 ± 0.52
win	97.19 ± 3.96	97.75 ± 3.92	96.67 ± 4.68	96.67 ± 4.68
io	88.61 ± 6.85	89.17 ± 5.00	89.17 ± 6.29	89.17 ± 6.29
spa	79.09 ± 1.58	77.00 ± 1.06	86.70 ± 1.22	86.70 ± 1.22
son	65.76 ± 14.78	70.62 ± 7.54	67.24 ± 8.66	67.24 ± 8.66
mean	85.18	85.73	83.98	83.98
P-val		0.553	0.673	0.673

Only the differences between FCBF$_{ntc}$ and FCBF were found significant with a confidence level of $\alpha = 0.95$.

Finally, the robustness is shown in Table 6. As with the accuracy, all algorithms show similar results. Although on average BF and SFS are more stable, these differences are not statistically significant.

Table 5. Dimensionality reduction achieved, with their respective standard deviation, achieved by algorithms FCBF$_{ntc}$, FCBF, BF and SFS.

id	#features			
	FCBF$_{ntc}$	FCBF	BF	SFS
bsc	4.0 ± 0.00	4.0 ± 0.00	4.0 ± 0.00	4.0 ± 0.00
nur	7.0 ± 0.00	8.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.00
pdi	3.0 ± 0.00	3.8 ± 0.42	3.2 ± 0.42	3.2 ± 0.42
pbl	4.0 ± 0.00	3.7 ± 0.48	6.0 ± 0.00	6.0 ± 0.00
win	4.3 ± 0.67	9.6 ± 0.52	8.2 ± 0.63	7.8 ± 0.92
ion	4.1 ± 0.57	5.3 ± 0.67	12.6 ± 2.27	12.6 ± 2.27
spa	5.0 ± 0.00	15.2 ± 0.92	10.0 ± 0.00	10.0 ± 0.00
son	5.0 ± 0.00	9.6 ± 0.7	17.7 ± 1.42	17.7 ± 1.42
mean	4.55	7.4	7.84	7.79
P-val		0.035	0.151	0.151

Table 6. Robustness achieved by the algorithms $FCBF_{ntc}$, FCBF, BF and SFS.

id	Accuracy			
	$FCBF_{ntc}$	FCBF	BF	SFS
bsc	1.00	1.00	1.00	1.00
nur	1.00	1.00	1.00	1.00
pdi	1.00	0.91	0.91	0.91
pbl	0.92	0.60	1.00	1.00
win	0.79	0.93	0.89	0.82
ion	0.66	0.60	0.58	0.58
spa	1.00	0.91	1.00	1.00
son	0.42	0.57	0.81	0.81
mean	0.85	0.82	0.90	0.89
P-val		0.834	0.590	1.00

5 Conclusions and Future Works

In this work we address the feature selection problem by extending the concept of approximate Markov blanket in order to consider the interaction among subsets of features. We have extended the symmetrical uncertainty measure to the multivariate case by using the total correlation measure.

Results show that the multivariate measure overcomes the limitation of the bivariate one detecting interactions not observable by the bivariate case. As a consecuence, $FCBF_{ntc}$ can find smaller subsets of features with similar predictive power to the bivariate case. Moreover, the robustness of the solution is also similar.

When compared to other popular strategies, our proposal is also competitive as it achieves a similar accuracy and robustness while reducing, on average, the size of the solutions.

As future work, we will study the performance on high dimensional datasets from several domains and study the theoretical properties of the multivariate SU.

Acknowledgment. This work has been partially supported by the project TIN2015-64776-C3-2-R. Miguel García-Torres acknowledges the financial support of CONACyT-Paraguay (14-VIN-009). Christian E. Schaerer acknowledges PRONII-CONACyT-Paraguay. Part of the computer time was provided by the Centro Informático Científico de Andalucía (CIC).

References

1. Bell, D., Wang, H.: A formalism for relevance and its application in feature subset selection. Mach. Learn. **41**(2), 175–195 (2000)
2. Caruana, R., Freitag, D.: How useful is relevance?. In: Working Notes of the AAAI Fall Symposium on Relevance, pp. 25–29 (1994)

3. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2006)
4. Hall, M.A., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: An update. SIGKDD Explor. **11**(1), 10–18 (2009)
5. John, G.H., Kohavi, R., Pfleger, K.: Irrelevant features and the subset selection problem. In: Proceedings of the Eleventh International Conference on Machine Learning, pp. 121–1129. Morgan Kaufmann (1994)
6. Kalousis, A., Prados, J., Hilario, M.: Stability of feature selection algorithms: A study on high-dimensional spaces. Knowl. Inf. Syst. **12**(1), 95–116 (2007)
7. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artif. Intell. **97**(1–2), 273–324 (1997)
8. Koller, D., Sahami, M.: Toward optimal feature selection. In: Proceedings of the Thirteenth International Conference on Machine Learning, pp. 284–292 (1996)
9. Křížek, P., Kittler, J., Hlaváč, V.: Improving stability of feature selection methods. In: Kropatsch, W.G., Kampel, M., Hanbury, A. (eds.) CAIP 2007. LNCS, vol. 4673, pp. 929–936. Springer, Heidelberg (2007)
10. Kuncheva, L.I.: A stability index for feature selection. In: Proceedings of the 25th IASTED International Multi-Conference, pp. 390–395 (2007)
11. Lichman, M.: UCI Machine Learning Repository. Kluwer Academic, Dordrecht (2013)
12. McGill, W.J.: Multivariate information transmission. Trans. IRE Prof. Group Inf. Theor. **4**, 93–111 (1954)
13. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco (1993)
14. Saeys, Y., Abeel, T., Van de Peer, Y.: Robust feature selection using ensemble feature selection techniques. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part II. LNCS (LNAI), vol. 5212, pp. 313–325. Springer, Heidelberg (2008)
15. Watanabe, S.: Information theoretical analysis of multivariate correlation. IBM J. Res. Develop. **4**(1), 66–82 (1960)
16. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. J. Mach. Learn. Res. **5**, 1205–1224 (2004)