The final publication is available at http://www.sciencedirect.com/science/article/pii/S0957417415000925
Expert Systems with Applications, Volume 42, Issue 10, 15 June 2015, Pages 4632–4642
doi:10.1016/j.eswa.2015.01.069

Similarity of feature selection methods: an empirical study across data intensive classification tasks

Nicoletta Dessì and Barbara Pes*

Dipartimento di Matematica e Informatica, Università degli Studi di Cagliari,
Via Ospedale 72, 09124 Cagliari, Italy
dessi@unica.it, pes@unica.it
corresponding author (tel: +39 070 6758758, fax: +39 070 6758501)

Abstract

In the past two decades, the dimensionality of datasets involved in machine learning and data mining applications has increased explosively. Therefore, feature selection has become a necessary step to make the analysis more manageable and to extract useful knowledge about a given domain. A large variety of feature selection techniques are available in literature, and their comparative analysis is a very difficult task. So far, few studies have investigated, from a theoretical and/or experimental point of view, the degree of similarity/dissimilarity among the available techniques, namely the extent to which they tend to produce similar results within specific application contexts. This kind of similarity analysis is of crucial importance when two or more methods are combined in an ensemble fashion: indeed the ensemble paradigm is beneficial only if the involved methods are capable of giving different and complementary representations of the considered domain. This paper gives a contribution in this direction by proposing an empirical approach to evaluate the degree of consistency among the outputs of different selection algorithms in the context of high dimensional classification tasks. Leveraging on a proper similarity index, we systematically compared the feature subsets selected by eight popular selection methods, representatives of different selection approaches, and derived a similarity trend for feature subsets of increasing size. Through an extensive experimentation involving sixteen datasets from three challenging domains (Internet advertisements, text categorization and micro-array data classification), we obtained useful insight into the pattern of agreement of the considered methods. In particular, our results revealed how multivariate selection approaches systematically produce feature subsets that overlap to a small extent with those selected by the other methods.

Keywords: Data mining, Knowledge discovery, Feature selection, Similarity measures

1. Introduction

While data intensive applications are fast increasing in scope and sophistication, the extraction of useful knowledge from the large amounts of available data can be a very difficult task (Kumar & Minz, 2014; Liao, Chu, & Hsiao, 2012). One of the most critical issues for data manipulation and analysis is high dimensionality, i.e. the presence of a huge number of attributes (features) that are associated with each problem instance in the dataset. This can cause a number of drawbacks such as reduced performance, large computational time, and the use of features that may be either redundant or irrelevant to the problem at hand.

A lot of research has focused on methods for effectively handling high dimensional datasets (Chandrashekar & Sahin, 2014; Khalid, Khalil, & Nasreen, 2014), with two main approaches existing in literature: mapping the original feature space to a new space with lower dimensions (Wang & Paliwal, 2003) or selecting a meaningful subset of the original features, hence discarding those irrelevant and redundant ones (Guyon & Elisseeff, 2003). This last approach, referred to as feature selection, has proved to be very effective in the context of high dimensional classification problems, enabling to improve predictive performance as well as to obtain faster and more cost-effective predictors, and to achieve a better understanding of the underlying domain.

Though many works have investigated the potential and limits of existing feature selection methods (Bolón-Canedo, Sánchez-Maroño, & Alonso-Betanzos, 2013; Tang, Alelyani, & Liu, 2014), the choice of the most appropriate method for a given task remains difficult. Indeed, while more and more feature selection algorithms are available, there is little theoretical support to find the “right” one for the problem at hand (Liu & Yu, 2005). Increasingly, real-world datasets are being handled by applying a number of selection techniques, instead of a single one, and then integrating their outputs in some way.

As suggested by recent literature (Dittman, Khoshgoftaar, Wald, & Napolitano, 2012), when choosing a set of techniques for a classification task it is beneficial to evaluate their degree of consistency. Different techniques, indeed, may select different features depending on the search strategy and evaluation criteria adopted in the selection process. However, despite their specificities, two techniques can be similar in their behavior, i.e. they can systematically produce results that overlap to a great extent (Cannas, Dessì, & Pes, 2013).

A similarity-based analysis of feature selection techniques can provide useful insight for domain modeling and understanding: if a set of techniques are dissimilar, i.e. exhibit in general a different behavior, then there is more reason to have confidence in a feature selected by all these techniques. On the other hand, it is not surprising if similar techniques select the same features and it doesn't help to confirm the relevance of these features for the considered domain (Dessì, Pascariello, & Pes, 2013).

Furthermore, when multiple feature selection methods are systematically combined in an ensemble fashion (Altidor, Khoshgoftaar, Van Hulse, & Napolitano, 2011a), a similarity evaluation of the methods in the ensemble should not be neglected: it would not be indeed beneficial to combine two or more methods that give almost identical results. Though it is recognized that diversity has a crucial role for the success of an ensemble learning strategy (Dietterich, 2000), most research work on ensemble feature selection has so far not given due consideration to this important issue. Existing ensemble approaches are mainly built on an "ad hoc" basis (Dutkowski & Gambin, 2007; Leung & Hung, 2010; Olsson & Oard, 2006; Yang, Zhou, Zhang, & Zomaya, 2010), depending on the specific problem at hand, and there is a lack of systematic studies aiming at providing insight on which methods should be combined, and how this combination should be made, based on the degree of diversity/similarity of the involved methods.

In this paper, we aim to give a valuable contribution in this direction by investigating the similarity of eight popular feature selection techniques, representatives of different types of selection approaches. Specifically, we consider both univariate methods that evaluate each feature independently from the others as well as multivariate methods that take into account interdependencies among features. The similarity analysis is carried out in two stages: (i) the feature subsets produced by the chosen methods are compared, on a pair-wise basis, using a proper similarity index; (ii) the overall degree of consistency among the eight methods (or a specific group of them) is obtained by averaging similarity values over all the involved pair-wise comparisons. A similarity trend is also derived for feature subsets of increasing size.

The datasets used in the analysis come from three challenging domains: Internet advertisements, text categorization and micro-array data classification. To the best of our knowledge, there is no study in literature that performs such a similarity analysis encompassing different real world application scenarios, as we do in this work.

The paper is organized as follows. Section 2 provides a survey of current literature and discusses related works. Section 3 describes all materials and methods involved in our empirical study, i.e. the adopted methodology, as well as the feature selection techniques and the datasets used for the experiments. The results of the analysis are presented and discussed in section 4. Finally, section 5 contains concluding remarks and future research directions.

2. Literature survey and related work

Feature selection is crucial to the analysis of high dimensional datasets coming from a number of application areas such as bioinformatics and text processing. It involves the exploration of the original feature space and the selection of the optimal feature subset based on a suitable relevance evaluation criterion (Kumar & Minz, 2014). According to whether the dataset is labeled or not, feature selection algorithms can be categorized into supervised (Song, Smola, Gretton, Borgwardt, & Bedo, 2007), unsupervised (Dy & Brodley, 2004) and semi-supervised (Xu, King, Lyu, & Jin, 2010).

Supervised selection methods can be further categorized into *filter*, *wrapper* and *embedded* methods, depending on how they interact with the learning algorithm (classifier) that will be ultimately used to infer a model (Tang et al., 2014). Basically, *filter* approaches (Lazar et al., 2012) assess the relevance of features by looking only at the intrinsic properties of the data, without involving the use of a learning algorithm in the selection stage. In contrast, *wrapper* approaches (Guyon & Elisseeff, 2003) perform a search in the space of feature subsets and evaluate each subset by training and testing a specific classification model; hence wrappers are tailored to a specific learning algorithm, and may achieve better performance than filters methods, but at the price of a greater computational cost. Finally, *embedded* approaches (Ma & Huang, 2008) leverage the internal parameters of a classification algorithm to select relevant features, often providing a good trade-off between computational cost and performance.

A wide literature is currently available on the strengths and weaknesses of different feature selection methods (Bolón-Canedo et al., 2013; Hall & Holmes, 2003; Lazar et al., 2012; Saeys, Inza, & Larranaga, 2007), the choice of the "best" method being dependent on the specific problem at hand. Moreover, with the aim of devising suitable solutions for specific problem settings, new proposals are constantly appearing that exploit different strategies, e.g. (i) using different selection approaches (e.g. a filter and a wrapper) in different search stages (Cannas, Dessì, & Pes, 2011; El Akadi, Amine, El Ouardighi, & Aboutajdine, 2011), (ii) combining the outcomes of different feature selectors in an ensemble fashion (Altidor et al., 2011a; Latkowski & Osowski, 2015) or (iii) combining feature selection with other approaches such as feature extraction (Bharti & Singh, 2015).

With such a body of algorithms available, their comparative analysis is a very difficult task. Most of the existing comparative studies focus on a specific application domain, such as text classification (Forman, 2003; Méndez, Fdez-Riverola, Díaz, Iglesias, & Corchado, 2006), genomic analysis (Abusamra, 2013; Bolón-Canedo, Sánchez-Maroño, Alonso-Betanzos, Benítez, & Herrera, 2014), software defect prediction (Khoshgoftaar, Gao, Napolitano, & Wald, 2014), image classification (Staroszczyk, Osowski, & Markiewicz, 2012). A number of studies have been also conducted on artificially generated data (Bolón-Canedo et al., 2013) in order to evaluate the performance of selection methods under specific conditions (e.g. class imbalance, noise, redundancy and interaction between features).

To date, a quite neglected issue in feature selection literature is the theoretical and/or experimental assessment of the degree of consistency among the outputs of different selection methods. Indeed, it is known that different selection techniques may result in different feature subsets, especially when the high dimensionality is coupled with a small sample size (Saeys et al., 2007), but few direct comparisons exist that quantify these differences in a systematic way. Existing studies (as those cited above) mostly focus on comparing the outcomes of different techniques in terms of predictive performance or, less frequently, in terms of stability with respect to sample variation (Haury, Gestraud, & Vert, 2011; Kalousis, Prados, & Hilario, 2007; Wang, Khoshgoftaar, & Liang, 2013). However, as we showed in a previous work (Dessì et al., 2013), selection methods with a similar behavior in terms of accuracy and/or stability do not necessarily select similar feature subsets and, on the other hand, feature subsets with a good degree of overlapping do not necessarily result in similar classification and/or stability performance. Therefore, a systematic analysis of the similarity among the feature subsets selected by different techniques could help in best understanding the behavior of these techniques as well as the extent of their applicability to specific contexts. This is where our paper aims to give its contribution by extending, through more extensive experiments, previous works in this field.

In the context of gene selection from high dimensional micro-array datasets, a first experimental assessment of the overlap between gene lists produced by different feature selection methods is provided by Jeffery, Higgins, and Culhane (2006). Within the same domain, a similarity analysis involving filter-based selection techniques is presented in the work of Dittman et al. (2012): results are here averaged across nine datasets and four feature subset sizes (i.e. 50, 75, 100 and 200 genes), providing a global view on the pattern of agreement of the considered methods but without a discussion about the influence of the subset size on the observed similarity. Still limited to the biological domain, a similarity trend for feature subsets of increasing size is provided by Cannas et al. (2013), who also discuss the characteristics of different similarity measures, giving evidence of the effectiveness of the *Kuncheva index* (Kuncheva, 2007) in evaluating the degree of consistency among feature subsets. In this work, we extend the above contributions by experimentally investigating the pattern of agreement of different methods across classification tasks of different domains (Internet advertisements, text categorization and micro-array data classification), as to best highlight the "intrinsic" similarity among methods, as well as the specific behavior of each method in each domain.

3. Materials and methods

In this work we focus on feature selection methods falling into filter or embedded approaches that, in practice, are often preferred to wrappers, especially in high dimensional classification problems, due to the lower computational complexity (Tang et al., 2014). Specifically, we consider selection techniques that produce as output a *ranked list* where features are ordered based on their relevance for the classification task at hand. Referred in the following as *rankers*, these methods assign a weight to each feature according to some scoring criterion that measures the degree of correlation between that feature and the target class. This process can be carried out in two ways (Saeys et al., 2007): assessing each feature independently from the others (*univariate ranking methods*) or taking into account interdependencies among features (*multivariate ranking methods*). Features with the highest weights appear in the first positions of the ranked list, while features with the lowest weights reside at the last positions. Finally, the list is usually cut at a proper threshold point in order to obtain a subset of highly predictive features.

It has to be noted that, when comparing the outputs of different techniques, the comparison can be carried out in different ways: (i) based on the weights assigned to features; (ii) based on the ranking positions of features; (iii) considering only the features in the highest ranking positions, regardless of their weighting or order (*feature subset selection*). Obviously, this last approach involves the choice of a suitable threshold that may be dependent on the specific problem at hand.

Appropriate similarity measures for feature weighting, ranking and subset selection can be derived from different correlation coefficients (Kalousis et al., 2007; Saeys, Abeel, & Van de Peer, 2008). For feature weighting, the Pearson correlation coefficient can be used while, for feature ranking, the Spearman rank correlation coefficient is a common choice. For feature subsets, a number of metrics have been proposed (Duda, Hart, & Stork., 2001; Kuncheva, 2007) that basically rely on the degree of overlapping among subsets, with proper normalization factors.

To assess the similarity of feature selection techniques, we focus here on comparing feature subsets, as these are most often used by domain experts. Indeed, a comparison based on feature weighting or feature ranking takes into account the complete feature preferences produced by a method, while the subset selection approach gives a more precise picture of greater utility for users and domain experts: the selected features convey the most important information and are ultimately used to infer predictive models.

In the remaining of this section we provide a description of all materials and methods involved in our empirical study. Specifically, we present the methodology used for the similarity analysis (section 3.1), the feature selection techniques that have been the object of this analysis (section 3.2), and the different classification tasks where the analysis has been performed (section 3.3).

3.1. Methodology for the similarity analysis

Our methodological approach is meant to provide a systematic comparison of the feature subsets produced by different selection techniques. Specifically, given a number k of rankers r_i $(i = 1, 2, \dots, k)$ and a dataset d of n features, each r_i is applied to d giving as output a ranked list l_i where the n features appear in descending order of relevance: this results in k distinct ranked lists, each providing a different ordering of the n features. The comparison among the lists is carried out, in a pair-wise fashion, for different values of the cut-off threshold t (setting a given threshold means to set the size of the feature subset that will be ultimately used for predictive purposes).

We denote as set_{it} the feature subset resulting from cutting the list l_i at threshold t. The comparison between two subsets set_{it} and set_{jt} $(i,j = 1, 2, \dots, k)$, of the same size t, is carried out using the Kuncheva index (Kuncheva, 2007) that has proved to be a good choice in the context of such a similarity analysis (Cannas et al., 2013). Specifically, it is defined as follows:

$$Kuncheva(set_{it}, set_{jt}) = \frac{|set_{it} \cap set_{jt}| - t^2/n}{t - t^2/n}$$

where $|set_{it} \cap set_{jt}|$ is the number of features that are present in both the subsets *set*$_{it}$ and *set*$_{jt}$. The Kuncheva measure tends to $|set_{it} \cap set_{jt}|/t$, i.e. to the fraction of overlap between the subsets, when the subset size is small compared to the overall number of features. For increasing values of *t*, the similarity value is corrected by taking into account the probability that a feature is included in both subsets simply by chance (this probability grows as the subset size approaches the dimensionality of the original dataset).

To obtain an overall evaluation of the degree of similarity among the *k* feature subsets (or among a number of them, if the focus is restricted to a specific group of selection methods, as discussed later in section 4), we average over all the involved pair-wise comparisons:

$$S_t = \frac{2 \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} Kuncheva(set_{it}, set_{jt})}{k(k-1)}$$

The resulting similarity S_t refers to a specific threshold *t*, i.e. to a specific subset size: if a suitable range of threshold values is explored, a similarity trend can be derived for feature subsets of increasing size.

3.2. Feature selection techniques

As previously mentioned, we considered both univariate selection techniques, where each single feature is weighted independently from the others, and multivariate selection techniques, where the interdependencies among features are taken into account. For all of them we leveraged the implementation provided by the WEKA package (Bouckaert et al., 2010).

As representatives of the univariate approaches, we chose *Chi Squared* (χ^2), *Information Gain* (IG), *Symmetrical Uncertainty* (SU), *Gain Ratio* (GR) and *OneR* (OR). As representatives of the multivariate approaches, we considered *ReliefF* (RF), *SVM-ONE* and *SVM-RFE*. A brief description of each method is provided in what follows.

- *Chi Squared* (χ^2) evaluates features individually by measuring their chi-squared statistic with respect to the class: the larger the chi-squared, the more important a feature is for the classification task at hand (Liu & Setiono, 1995). Specifically, once a feature has been discretized into a number I of intervals, its χ^2 value is computed as:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{K} \frac{(A_{ij} - \frac{R_i \cdot B_j}{N})^2}{\frac{R_i \cdot B_j}{N}}$$

where N is the number of instances, K the number of classes, R_i the number of instances in the ith interval, B_j the number of instances in the jth class, and A_{ij} the number of instances in the ith interval and jth class.

- *Information Gain* (IG) is grounded on the information-theoretical concept of entropy (Hall & Holmes, 2003). Basically, IG measures how the entropy of the class C (i.e. the degree of uncertainty about its prediction) decreases when the value of a given feature X is known:

$$IG = H(C) - H(C|X)$$

where $H(C)$ and $H(C|X)$ denote the entropy of the class before and after observing the feature X.

- *Symmetrical Uncertainty* (SU), as IG, relies on the entropy concept to evaluate the correlation between a feature X and the class C (Witten, Frank, & Hall, 2011). In more detail, it is obtained by:

$$SU = 2 \cdot \frac{H(C) - H(C|X)}{H(X) + H(C)}$$

i.e. the information gain associated to a feature is normalized by taking into account both the intrinsic entropy of the feature as well as the entropy of the class. This compensates for the information gain's bias toward features with more values (Senthamarai Kannan & Ramaraj, 2010).

- *Gain Ratio* (GR) is in turn a refinement of IG. Specifically, GR corrects the IG score of a feature X using the information split (Quinlan, 1993):

$$InfoSplit = -\sum_{i=1}^{r} \frac{|X_i|}{N} \log \frac{|X_i|}{N}$$

where $|X_i|$ is the number of samples where X takes the value of X_i, r is the number of distinct values of X, and N is the total number of samples in the dataset. The GR score of a feature, obtained as *IG/InfoSplit*, is sensitive to how broadly and uniformly the feature splits the data (Witten et al., 2011).

- *OneR* (OR) derives the level of significance of each feature using a simple rule-based classifier (Holte, 1993). For each attribute in the training data, the algorithm creates one rule by determining the most frequent class for each attribute value (a rule is simply a set of attribute values bound to their majority class). Then the classification accuracy of each rule is calculated, and the attributes are ranked (i.e. ordered by relevance) based on the accuracy of the corresponding rules.

- *ReliefF* (RF) estimates the relevance of features based on their ability to distinguish between instances that are near to each other (Hall & Holmes, 2003). Basically, given a probe instance, the weight *w* of a feature *X* is obtained by evaluating how the instance's nearest hit (one from the same class) differs from its nearest miss (one from a different class). The rationale is that a predictive feature should differentiate between instances from different classes and have the same value for instances from the same class. Based on a suitable number of probe instances, RF attempts to approximate the following difference of probabilities:

$$w(X) = P(\textit{different value of } X \mid \textit{nearest instance from different class}) \\ - P(\textit{different value of } X \mid \textit{nearest instance from same class})$$

  Originally defined for two-class problems, the method was later extended to handle noise and multi-class datasets (Kononenko, 1994). The RF approach can be also adjusted to weight nearest neighbors by their distance (Robnik-Sikonja & Kononenko, 2003).

- *SVM-ONE* uses a linear SVM classifier to derive the relevance of each feature (Rakotomamonjy, 2003). Basically, a linear SVM looks for an optimal hyperplane as a decision function for separating instances in the input space. In that function, each feature (dimension) is assigned a weight that expresses how the feature contributes to the multivariate decision of the classifier:

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$$

where **x** is an instance vector in the n-dimensional space of input features, **w** is the weight vector and *b* denotes a constant (bias). Hence, based on their weights, the features can be ranked from the most important to the least important.

- *SVM-RFE* relies on a linear SVM classifier, as the SVM_ONE method, but adopts a backward elimination strategy to obtain the final ranking of features (Guyon, Weston, Barnhill, & Vapnik, 2002). First, a linear SVM is trained on the input space and features are sorted according to their weights in the hyperplane function. Then, the least important features are removed and the overall weighting process is repeated on the remaining features. The computational complexity of the method is greatly influenced by the fraction of features removed at each iteration: we set this parameter as 10%, as suggested by recent literature (Kalousis et al., 2007; Saeys et al, 2008).

3.3. Classification tasks

Table 1 lists the datasets used in this study and provides their characteristics in terms of the total number of features, number of instances, class distribution and percentage of minority instances. Specifically, we considered classification problems from three challenging domains, as detailed in what follows.

(1) Internet advertisements. This dataset represents a set of possible advertisements on Internet pages (Kushmerick, 1999). The features encode the geometry of the image as well as phrases occurring in the document's URL or the image's URL, alt text, anchor text, and words near the anchor text. The task is to predict whether an image is an advertisement ("ad") or not ("nonad").

(2) Reuters-21578 ModApte. This is one of the most celebrated text datasets (Lewis, 1997). It consists of 12902 documents manually classified across 135 categories. According to the widely used *ModApte split*, 9603 documents are used as training set and the remaining 3299 as test set. Once discarded categories with no document in the test set, the remaining training set consists of 9598 documents. Specifically, the 10 categories with the highest number of positive training examples have been considered in this study. For each category we performed feature selection

experiments on a binary dataset where the training documents related to that category are labeled as "yes", and the others as "no". Each dataset was preliminarily preprocessed by removing stop-words and extracting n-grams, defined as sequences of maximum three words consecutively occurring within a document (Pietramala, Policicchio, Rullo, & Sidhu, 2008).

(3) Biomedical datasets. Five biomedical datasets coming from DNA micro-array experiments are included in this study: *Colon Tumor* (Alon et al., 1999), *DLBCL* (Shipp et al., 2002), *Leukemia* (Golub et al., 1999), *Lung Cancer* (Beer et al., 2002) and *Central Nervous System* (Pomeroy et al., 2002). Instances correspond here to biological samples that are distinguished between normal and tumor tissues (*Colon Tumor* and *Lung Cancer* datasets) or between different tumor types (*DLBCL* and *Leukemia* datasets) or between different treatment outcomes (*Central Nervous System* dataset), while features correspond to levels of expression of different genes. The high dimensionality (thousands of gene expression values) coupled with the small sample size (a few dozen of biological samples) makes these datasets a very interesting benchmark for feature selection and classification experiments.

Table 1. *Datasets used in the empirical study.*

4. Experimental results and discussion

The results of our empirical study are summarized in what follows. According to the methodological approach presented in section 3.1, we measured the similarity among the ranked lists produced by eight selection techniques (see section 3.2) in terms of the Kuncheva similarity among the feature subsets resulting from cutting these lists at a given threshold. The threshold values explored in this study are 10, 20, 30, 50, 100 and 150 (larger values are rarely used in practice since high-ranked features are more likely to be useful for predictive purposes as well as for domain modeling).

For each threshold, i.e. for each subset size, we evaluated the similarity among the resulting feature subsets on a pair-wise basis, i.e. by considering each pair of subsets. The results of this pair-wise analysis are reported, in terms of similarity matrices, in section 4.1. Next, we explored the similarity trend by averaging results over all pair-wise comparisons, as well as over pair-wise comparisons involving only ranking methods with a specific selection approach (i.e. pairs of univariate methods, pairs of multivariate methods, and "hybrid" pairs where one method is univariate and the other multivariate). This way, besides the overall degree of consistency among the considered methods, we can also investigate the specific behavior of univariate and multivariate methods, as discussed later in section 4.2. Finally, section 4.3 focuses on how the behavior of certain methods (in particular, multivariate methods) may be influenced by the setting of their main parameters.

4.1 Similarity matrices

Results of our similarity analysis are here reported separately for the three application domains considered in this study, i.e. *Internet advertisements*, *Reuters datasets* and *biological datasets* (see section 3.3). As concerns the Internet advertisements dataset, Table 2 shows the pair-wise similarity between the subsets of 20 features selected by univariate methods (χ^2, IG, SU, GR, OR) and multivariate methods (RF, SVM_RFE, SVM_ONE). Specifically, each entry in the table represents the Kuncheva similarity value obtained on the pair of subsets selected by the rankers in the corresponding row and column. Being the resulting matrix symmetric, only positions in the upper-right triangular block are filled. A first evidence from Table 2 is that univariate approaches produce results quite similar to each other. In particular, χ^2, IG and OR generate identical feature subsets, while GR seems to exhibit a more specific behavior. In contrast, each multivariate method produces a feature subset that overlaps to a little extent with the subsets selected by the other methods. Indeed, all pair-wise comparisons involving a multivariate approach (both univariate-multivariate and multivariate-multivariate comparisons) result in a small degree of similarity.

Also for the Reuters datasets, we concentrate here on feature subsets of 20 features. Specifically, we built a similarity matrix for each of the ten datasets of this domain and then averaged, for each combination of ranking methods, the similarity values obtained across the ten datasets so as to obtain a synthetic and comprehensive view of the pattern of agreement of the considered methods in the context of text categorization. Results are shown in Table 3. As regards the univariate approaches, the degree of consistency among selected subsets is globally lower than that observed in Table 2. Indeed, similarity among χ^2, IG and SU is still high, while OR in this case produces results that are only partially consistent with those of the other univariate methods. GR, in turn, exhibits a quite different behavior though its ranking algorithm is grounded on the same concept of entropy as IG and SU. Indeed, due to the ‘split information’ included as normalization factor in GR (see section 3.2), the weight assigned to a feature may be strongly dependent on how broadly and uniformly it splits the data. Among multivariate approaches, RF selects features that don’t overlap at all with those selected by the other methods. As well, SVM_RFE and SVM_ONE produce results that overlap only to a small extent with those of the other methods, revealing a strong dependence of the selection outcome on the specific multivariate logic used to evaluate the interdependences among features.

Finally, Table 4 shows the results obtained in the biological domain. Each entry in the matrix expresses the similarity between the outputs (subsets of 20 features) of two methods, averaged across all the five biological datasets considered in this study. Once again, χ^2, IG and SU turn out similar to each other (though their pair-wise Kuncheva values in this case are lower). GR confirms to be somewhat different from the other entropic methods, i.e. IG and SU, even if there is not a so pronounced dissimilarity as in the Reuters datasets. Moreover, as seen in Table 2 and in Table 3, the pair-wise comparisons that involve a multivariate approach (univariate-multivariate and multivariate-multivariate pairs) result in lower similarity values, though the behavior of RF in this domain is a little more consistent with the other methods. There is also more consistency between SVM_RFE and SVM_ONE.

For the sake of space, the similarity matrices built for different subset sizes are here omitted, but next section discusses the influence of the subset size when averaging similarity values across all pair-wise comparisons (or a group of them).

Table 2. Internet advertisements dataset: similarity matrix for subsets of 20 features. *Each cell in the matrix reports the Kuncheva similarity for the pair of subsets selected by the methods in the corresponding row and column.*

Table 3. Reuters datasets: similarity matrix for subsets of 20 features. *Each cell in the matrix reports the Kuncheva similarity, averaged across all ten Reuters datasets, for the pair of subsets selected by the methods in the corresponding row and column.*

Table 4. Biological datasets: similarity matrix for subsets of 20 features. *Each cell in the matrix reports the Kuncheva similarity, averaged across all five biological datasets, for the pair of subsets selected by the methods in the corresponding row and column.*

4.2 Average similarity trend

For each similarity matrix, we calculated the average similarity over all 28 cells, i.e. over all pair-wise comparisons, in order to evaluate the global degree of consistency among the eight selection methods used in this study. Moreover, we considered the average similarity (i) across the 10 cells that represent pairs of univariate methods (*univariate-univariate* comparisons), (ii) across the 3 cells that represent pairs of multivariate methods (*multivariate-multivariate* comparisons), and (iii) across the 15 cells that represent “hybrid” pairs (*univariate-multivariate* comparisons). We performed this analysis for feature subsets of increasing size: the results are shown in Fig.1, Fig.2 and Fig.3, for Internet advertisements, Reuters datasets and biological datasets respectively.

In the Internet advertisements classification task (Fig.1), the average similarity among univariate methods is very high, ranging from 0.73 for subsets of 10 features to 0.91 for subsets of 150 features. On the other hand, multivariate-multivariate comparisons result in an average similarity that ranges from 0.13 for subsets of 10 features to 0.31 for subsets of 150 features, and hybrid pairs exhibit an even lower similarity, making the overall similarity not high at all (between 0.36 and 0.42).

As regards the Reuters datasets (Fig.2), the average similarity among univariate methods is quite lower (between 0.53 and 0.46) than that observed previously in Fig.1. However, univariate-univariate pairs confirm to be, in average, more similar than univariate-multivariate and multivariate-multivariate pairs. In particular, for feature subsets of small size, multivariate-multivariate comparisons result in average similarity values lower than 0.10, meaning that there is almost no overlap between the selected subsets.

Interestingly, in the biological datasets, multivariate methods are not so inconsistent with one another (Fig.3), probably due to the existence in this domain of relatively small-sized groups of features (i.e. genes) that strongly interact to determine a given pathological state (Saeys et al., 2007). Hence, the stronger the interactions among features, the higher seems to be the similarity among multivariate methods. However, univariate-univariate comparisons again result in the highest similarity values (irrespective of the considered subset size).

A comparison of the overall similarity on Internet advertisement, Reuters and biological datasets is shown in Fig.4. It is clear that the text categorization domain is the one where the outputs of the different selection approaches are less consistent with one another, while the overall similarity trend is almost the same for the other two domains, despite the specificities previously discussed. This reveals that the behavior of the considered selection methods is somewhat “domain dependent”. Indeed, besides the high number of features (words), as well as the high number of instances, text classification problems are also characterized as frequently having a high degree of class imbalance

(Forman, 2007). For example, in the Reuters datasets here considered, the percentage of minority instances (i.e. positive instances), is lower than 5% in seven cases out ten (see Table 1). This may lead to significantly more uncertainty in the frequency estimates of words in the positive class, making the learning task intrinsically more difficult and thus sharpening discrepancies among different algorithms.

Despite this, there are also a number of strong analogies among what observed across the different classification tasks, such as the higher similarity among the univariate approaches compared to the highly specific behavior of each multivariate method (that leads to a lower similarity both in multivariate-univariate comparisons as well as in multivariate-multivariate comparisons). Another interesting observation is that, in all the considered application scenarios, the average similarity depends only to a small extent on the number of selected features (Fig.1-Fig.4), suggesting that selection methods have in some way an “intrinsic” degree of similarity/diversity (at least in the range of feature subset sizes that are usually used in practical applications).

Fig. 1. Internet advertisements dataset: average similarity vs subset size. *For a given subset size, the corresponding similarity matrix is built (as shown in Table 2), and the average similarity is computed over all pair-wise comparisons (ALL), over the pair-wise comparisons where both methods are univariate (Uni-Uni), over the pair-wise comparisons where both methods are multivariate (Multi-Multi), and over the pair-wise comparisons where one method is univariate and the other multivariate (Uni-Multi).*

Fig. 2. Reuters datasets: average similarity vs subset size. *For a given subset size, the corresponding similarity matrix is built across all ten Reuters datasets (as shown in Table 3), and the average similarity is computed over all pair-wise comparisons (ALL), over the pair-wise comparisons where both methods are univariate (Uni-Uni), over the pair-wise comparisons where both methods are multivariate (Multi-Multi), and over the pair-wise comparisons where one method is univariate and the other multivariate (Uni-Multi).*

Fig. 3. Biological datasets: average similarity vs subset size. *For a given subset size, the corresponding similarity matrix is built across all five biological datasets (as shown in Table 4), and the average similarity is computed over all pair-wise comparisons (ALL), over the pair-wise comparisons where both methods are univariate (Uni-Uni), over the pair-wise comparisons where*

both methods are multivariate (Multi-Multi), and over the pair-wise comparisons where one method is univariate and the other multivariate (Uni-Multi).

Fig. 4. Comparison of average similarity trends on Internet advertisements, Reuters and biological datasets. *For a given subset size, the average similarity is computed over all pair-wise comparisons in the corresponding similarity matrix.*

4.3 Sensitivity analysis with respect to internal parameters

The experimental results discussed in sections 4.1 and 4.2 show that different feature selection techniques may result in quite different feature subsets, with a more pronounced dissimilarity in the context of multivariate feature selection. As an extension of the similarity analysis previously presented, it would be interesting to investigate the extent to which the same feature selection technique may produce different results depending on the chosen setting of the parameters of the method. Indeed, when using popular data mining suites, practitioners often rely on default settings without awareness of what the implications might be on the selection outcome.

Since multivariate methods have shown to exhibit a more specific behavior, we conducted a number of experiments to study their sensitivity with respect to internal parameters. In the context of SVM embedded feature selection, the most critical parameter is the fraction of features iteratively removed when adopting a backward elimination strategy (SVM_RFE approach). In previous experiments (sections 4.1 and 4.2), we set this parameter as 10% (*10%-RFE setting*) being this a common choice in literature (Kalousis et al., 2007; Saeys et al, 2008). Further experiments were performed removing 50% of features at each iteration (*50%-RFE setting*), as used for example by Abeel, Helleputte, Van de Peer, Dupont, and Saeys (2010). The resulting feature subsets were then compared based on the Kuncheva index (section 3.1), i.e. for each dataset and for each threshold value (subset size) we evaluated the degree of similarity between the subset produced with the 10%-RFE setting and the subset produced with the 50%-RFE setting. Finally, similarity values were averaged across the datasets of the same domain. Results are summarized, for each domain (Internet advertisements, Reuters and biological datasets), in Table 5.

As we can see, the fraction of features discarded at each iteration of RFE may sensibly influence the selection outcome, especially in the text categorization domain. Even in the case of biological datasets, when considering feature subsets of small size, there is not a high degree of similarity among the results produced by SVM_RFE in the two settings here explored. This suggests that the choice of the internal parameters of a selection method may be as important as the choice of the method itself, in terms of the implications on the selection outcome.

The other multivariate approach used in this study, i.e. ReliefF (RF), seems to have a more stable behavior with respect to changes in the internal parameters. In particular, we performed preliminary experiments by introducing a weighting mechanism that takes account for the distance between probe instances and their neighbors (see section 3.2). The resulting feature subsets are mostly consistent with those selected by the simple version of RF previously considered. With further experiments, we plan to extend this part of our experimental study, so as to have a more complete picture of the sensitivity of the different selection approaches with respect to their internal settings.

Table 5. *SVM_RFE approach: similarity analysis between 10%-RFE setting and 50%-RFE setting.*

5. Conclusions and future research directions

The explosive growth in databases has created a need to develop methods and technologies that use information and knowledge intelligently (Liao et al., 2012). Therefore, data mining has become an increasingly important research area. Indeed, data mining techniques, including feature selection techniques, play a crucial role to modern knowledge-based systems (Akerkar & Sajja, 2009). As recognized by recent literature (Bolón-Canedo, Sánchez-Maroño, & Alonso-Betanzos, 2011; Danenasa & Garsvaa, 2015; Eesa, Orman, & Brifcani, 2015; Nassirtoussi, Aghabozorgi, Wah, & Ngo, 2014), there are an increasing number of application scenarios where feature selection is exploited in automatic knowledge discovery to complement and enrich knowledge acquired from domain experts.

While new feature selection approaches are constantly appearing in the literature, there is no well-established theoretical and technical framework for evaluating how the choice of a specific method may influence the outcome of the analysis (Bolón-Canedo et al., 2013). As well, there are only partial investigations that focus on assessing the degree of consistency/similarity among the outputs of different selection methods within specific application contexts. Such a kind of similarity analysis would be of great importance especially in high dimensional problems, where the application of different selection techniques and the comparison of their outputs may help to identify the most relevant features (Dittman et al., 2012).

To give a contribution in this direction, our paper has proposed an empirical approach to evaluate the extent to which different selection methods may exhibit a similar/dissimilar behavior i.e. produce similar/dissimilar results. Specifically, we compared, on a pair-wise basis, the subsets of features provided by eight popular selection techniques (χ^2, IG, SU, GR, OR, RF, SVM_RFE, SVM_ONE) in the context of three challenging domains, i.e. Internet advertisements, text categorization and micro-array data classification.

Despite the specificity of each domain, a number of general indications have emerged about the pattern of agreement of the considered methods. For example, the univariate approaches have proved to be quite similar to each other; in particular, χ^2, IG, and SU form a “cluster” of methods that provide, in each of the datasets here considered, a highly consistent output. Though exploiting the same concept of entropy as IG and SU, GR exhibits a somewhat different behavior, with a more pronounced dissimilarity in the context of text categorization. Actually, from a different point of view, other studies have highlighted the peculiarity of the GR function, which theoretically could lead to a more fair feature weighting (Debole & Sebastiani, 2003) but practically often results in an higher sensitivity to noise (Altidor, Khoshgoftaar, & Van Hulse, 2011b).

On the other hand, multivariate methods select feature subsets that are, in most cases, quite dissimilar from those produced by the other methods, revealing a strong dependence of the selection outcome on the specific multivariate logic used to evaluate the interdependences among features. We also observed that multivariate approaches, especially SVM-based methods, are quite sensitive to the setting of their internal parameters, i.e. different parameters values may lead to feature subsets that overlap only partially.

Furthermore, our empirical analysis has shown that the consistency among the outputs of the considered methods is influenced only to a small extent by the number of selected features: this seems to suggest that these methods have in some way an “intrinsic” degree of similarity/diversity (at least in the range of feature subset sizes that are usually used in practical applications).

There are a number of directions in which our work can be extended. First, it would be interesting to include in the experimental study more datasets from different domains as well as other selection techniques, so as to strengthen the findings so far obtained. Second, even for the selection methods here considered, further experiments should be conducted to more comprehensively evaluate their sensitivity with respect to internal parameters, which we only partially addressed in this work. But, most importantly, the approach presented in this paper can be a starting point for defining ensemble strategies that effectively exploit multiple selection methods to improve classification performance and robustness of the chosen features (Latkowski & Osowski, 2015). Indeed, as previously observed, the degree of diversity of the involved methods is of crucial importance in an ensemble perspective: if two feature selection criteria produce similar results, fusion of such two criteria does not help. Hence, starting from the empirical study here proposed, we plan to explore different ensemble combinations and to evaluate how the overall ensemble performance is affected by the degree of similarity/diversity of the involved methods. In particular, hybrid combinations involving both univariate and multivariate techniques seem to be a promising avenue to maximize the ensemble diversity.

Acknowledgments

This research was supported by RAS, Regione Autonoma della Sardegna (Legge regionale 7 agosto 2007, n. 7 “Promozione della ricerca scientifica e dell’innovazione tecnologica in Sardegna”) in the project “DENIS: Dataspaces Enhancing the Next Internet in Sardinia”.

Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this work.

Authors’ contributions

Nicoletta Dessì coordinated the research and wrote parts of the article.

Barbara Pes performed the similarity analysis and wrote parts of the article.

Both authors approved the final article.

References

Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., & Saeys, Y. (2010). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, *26*(3), 392–398.

Abusamra, H. (2013). A Comparative Study of Feature Selection and Classification Methods for Gene Expression Data of Glioma. *Procedia Computer Science, 23*, 5-14.

Akerkar, R. A., & Sajja, P. S. (2009). *Knowledge-based systems*. Sudbury, MA, USA: Jones & Bartlett Publishers.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A.J. (1999). Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays. *PNAS*, *96*, 6745-6750.

Altidor, W., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2011a). Ensemble Feature Ranking Methods for Data Intensive Computing Applications. In B. Furth & A. Escalante (Eds), *Handbook of Data Intensive Computing* (pp. 349-376), Springer.

Altidor, W., Khoshgoftaar, T. M., & Van Hulse, J. (2011b). Robustness of Filter-Based Feature Ranking: A Case Study. In *Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference* (pp. 453-458).

Beer, D. G., Kardia, S. L. R., Huang, C., Giordano, T. J., Levin, A. M., Misek, D. E., … Hanash, S. (2002). Gene-expression Profiles Predict Survival of Patients with Lung Adenocarcinoma. *Nature Medicine*, *8*, 816-824.

Bharti, K. K., & Singh, P. K. (2015). Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering. *Expert Systems with Applications*, *42*(6), 3105–3114.

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2011). Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset. *Expert Systems with Applications*, *38*, 5947–5957.

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. *Knowledge and Information Systems*, 34(3), 483-519.

Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., Benítez, J. M., & Herrera, F. (2014). A review of microarray datasets and applied feature selection methods. *Information Sciences*, *282*, 111–135.

Bouckaert, R. R., Frank, E., Hall, M. A., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I.H. (2010). WEKA - Experiences with a Java Open-Source Project. *Journal of Machine Learning Research*, *11*, 2533-2541.

Cannas, L. M, Dessì, N., & Pes, B. (2011). A Hybrid Model to Favor the Selection of High Quality Features in High Dimensional Domains. In *Proceedings of the 12th International Conference on Intelligent Data Engineering and Automated Learning, IDEAL 2011* (pp. 228-235).

Cannas, L. M, Dessì, N., & Pes, B. (2013). Assessing similarity of feature selection techniques in high-dimensional domains. *Pattern Recognition Letters*, *34*(12), 1446-1453.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers and Electrical Engineering*, *40*, 16-28.

Danenasa, P., & Garsvaa, G. (2015). Selection of Support Vector Machines based classifiers for credit risk domain. *Expert Systems with Applications*, *42*, 3194–3204.

Debole, F., & Sebastiani, F. (2003). Supervised term weighting for automated text categorization. In *Proceedings of SAC-03, 18th ACM Symposium on Applied Computing* (pp. 784–788).

Dessì, N., Pascariello, E., & Pes, B. (2013). A Comparative Analysis of Biomarker Selection Techniques. *BioMed Research International*, *2013*, Article ID 387673.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Proceedings of the 1st International Workshop on Multiple Classifier Systems, MCS 2000* (pp. 1-15).

Dittman, D., Khoshgoftaar, T. M., Wald, R., & Napolitano, A. (2012). Similarity analysis of feature ranking techniques on imbalanced DNA microarray datasets. In *Proceedings of the 2012 IEEE International Conference on Bioinformatics and Biomedicine* (pp. 1-5).

Duda, R., Hart, P., & Stork., D. (2001). *Pattern Classification and Scene Analysis.* New York: John Willey and Sons.

Dutkowski, J., & Gambin, A. (2007). On consensus biomarker selection. *BMC Bioinformatics*, *8* (Suppl 5):S5.

Dy, J. G., & Brodley, C. E. (2004). Feature selection for unsupervised learning. *Journal of Machine Learning Research*, *5*, 845-889.

Eesa, A. S., Orman, Z., & Brifcani, A. M. A. (2015). A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems. *Expert Systems with Applications*, *42*, 2670–2679.

El Akadi, A., Amine, A., El Ouardighi, A., & Aboutajdine, D. (2011). A two-stage gene selection scheme utilizing MRMR filter and GA wrapper. *Knowl Inf Syst*, *26*(3), 487–500.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, *3*, 1289-1305.

Forman, G. (2007). Feature Selection for Text Classification. In H. Liu & H. Motoda, (Eds.), *Computational Methods of Feature Selection* (pp. 257–276). Boca Raton, FL: Chapman and Hall/CRC Press.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., … Lander, E. S. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, *286*, 531-537.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*, 389-422.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157-1182.

Hall, M. A., & Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering*, *15*(6), 1437–1447.

Haury, A. C., Gestraud, P., & Vert, J. P. (2011). The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures. *PLOS ONE*, *6*(12), e28210.

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, *11*, 63-91.

Jeffery, I. B., Higgins, D. G., & Culhane, A. C. (2006). Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics*, *7*:359.

Kalousis, A., Prados, J., & Hilario, M. (2007). Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and Information Systems, 12*(1), 95-116.

Khalid, S., Khalil, T., & Nasreen, S. (2014) A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning. In *Proceedings of the Science and Information Conference, SAI 2014* (pp. 372 – 378).

Khoshgoftaar, T. M., Gao, K., Napolitano, A., & Wald, R. (2014). A comparative study of iterative and non-iterative feature selection techniques for software defect prediction. *Information Systems Frontiers*, *16*(5), 801-822.

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of relief. In *Proceedings of the Seventh European Conference on Machine Learning* (pp. 171-182).

Kumar, V., & Minz S. (2014). Feature Selection: A literature Review. *Smart Computing Review*, *4* (3), 211-229.

Kuncheva, L. I. (2007). A Stability Index for Feature Selection. In *Proceedings of the 25th IASTED International Multi-Conference: artificial intelligence and applications* (pp. 390-395).

Kushmerick, N. (1999). Learning to remove Internet advertisements. In *Proceedings of the 3rd International Conference on Autonomous Agents* (pp. 175-181).

Latkowski, T., & Osowski, S. (2015). Data mining for feature selection in gene expression autism data. *Expert Systems with Applications*, *42*, 864–872.

Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., … Nowé, A. (2012). A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis, *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9*(4), 1106-1119.

Leung,Y., & Hung, Y. (2010). A Multiple-Filter-Multiple-Wrapper Approach to Gene Selection and Microarray Data Classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 7*(1), 108-117.

Lewis, D. D. (1997). *Reuters-21578 text categorization test collection*. Distribution 1.0. http://www.daviddlewis.com/resources/testcollections/reuters21578

Liao, S. H., Chu, P. H., & Hsiao, P. Y. (2012). Data mining techniques and applications – A decade review from 2000 to 2011. *Expert Systems with Applications*, *39*, 11303–11311.

Liu, H., & Setiono, R. (1995). Chi2: Feature selection and discretization of numeric attributes. In *Proceedings of the 7th International Conference on Tools with Artificial Intelligence, ICTAI'95* (pp. 338-391).

Liu, H., & Yu., L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans. on Knowledge and Data Engineering*, *17*(3), 1–12.

Ma, S., & Huang, J. (2008). Penalized feature selection and classification in bioinformatics. *Briefings in Bioinformatics*, *9*(5), 392-403.

Méndez, J. R., Fdez-Riverola, F., Díaz, F., Iglesias, E. L., & Corchado, J. M. (2006). A Comparative Performance Study of Feature Selection Methods for the Anti-spam Filtering Domain, In *Proceedings of the 6th Industrial Conference on Data Mining, ICDM 2006* (pp. 106-120).

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, *41*, 7653–7670.

Olsson, J. O. S., & Oard, D. W. (2006). Combining feature selectors for text classification. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management, CIKM '06* (pp. 798–799).

Pietramala, A., Policicchio, V. L., Rullo, P., & Sidhu, I. (2008). A Genetic Algorithm for Text Classification Rule Induction. In *Proceedings of ECML PKDD '08* (pp. 188-203).

Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., … Golub, T. R. (2002). Prediction of Central Nervous System Embryonal Tumour Outcome Based on Gene Expression. *Nature*, *415*, 436-442.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Francisco: Morgan Kaufmann.

Rakotomamonjy, A. (2003). Variable selection using SVM based criteria. *Journal of Machine Learning Research*, *3*, 1357–1370.

Robnik-Sikonja, M., & Kononenko, I. (2003). Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning*, *53*, 23–69.

Saeys, Y., Inza, I., & Larranaga P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507-2517.

Saeys, Y., Abeel, T., & Van de Peer, Y. (2008). Robust Feature Selection Using Ensemble Feature Selection Techniques. In *Proceedings of ECML PKDD '08* (pp. 313-325).

Senthamarai Kannan, S., & Ramaraj, N. (2010). A novel hybrid feature selection via Symmetrical Uncertainty ranking based local memetic search algorithm. *Knowledge-Based Systems*, *23*, 580-585.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C, … Golub, T. R. (2002). Diffuse Large B-cell Lymphoma Outcome Prediction by Gene-expression Profiling and Supervised Machine Learning. *Nature Medicine, 8*(1), 68-74.

Song, L., Smola, A., Gretton, A., Borgwardt, K., & Bedo J. (2007). Supervised feature selection via dependence estimation. In *Proceedings of the 24th International Conference on Machine Learning* (pp. 823-830).

Staroszczyk, T., Osowski, S., & Markiewicz, T. (2012). Comparative Analysis of Feature Selection Methods for Blood Cell Recognition in Leukemia. In *Proceedings of the 8th International Conference on Machine Learning and Data Mining in Pattern Recognition, MLDM 2012* (pp 467-481).

Tang, J., Alelyani, S., & Liu, H. (2014). Feature Selection for Classification: A Review. In C.C. Aggarwal (Eds.), *Data Classification: Algorithms and Applications* (pp. 37-64), CRC Press.

Wang, X., & Paliwal, K. K. (2003). Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition. *Pattern Recognition*, *36*(10), 2429–2439.

Wang, H., Khoshgoftaar, T. M., & Liang, Q. A. (2013). A Study of Software Metric Selection Techniques: Stability Analysis and Defect Prediction Model Performance. *International Journal on Artificial Intelligence Tools*, *22*(5), 1360010.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd edition). San Francisco: Morgan Kaufmann.

Xu, Z., King, I., Lyu, M., & Jin, R. (2010). Discriminative semi-supervised feature selection via manifold regularization. *IEEE Transactions on Neural Networks, 21*(7), 1033 - 1047.

Yang, P., Zhou, B. B., Zhang, Z., & Zomaya, A. Y. (2010). A multi-filter enhanced genetic ensemble system for gene selection and sample classification of microarray data. *BMC Bioinformatics*, *11* (Suppl 1):S5.

Table 1. *Datasets used in the empirical study.*

Dataset name	***Number of features***	***Number of instances and class distribution***	***% of minority instances***
Internet advertisements	1559	3279 (458 ad, 2821 nonad)	14.0%
Reuters_acq	7495	9598 (1650 yes, 7948 no)	17.2%
Reuters_corn	8302	9598 (181 yes, 9417 no)	1.9%
Reuters_crude	14466	9598 (389 yes, 9209 no)	4.1%
Reuters_earn	9500	9598 (2877 yes, 6721 no)	30.0%
Reuters_grain	12473	9598 (433 yes, 9165 no)	4.5%
Reuters_interest	10458	9598 (347 yes, 9251 no)	3.6%
Reuters_money-fx	7757	9598 (538 yes, 9060 no)	5.6%
Reuters_ship	9930	9598 (197 yes, 9401 no)	2.1%
Reuters_trade	7600	9598 (369 yes, 9229 no)	3.8%
Reuters_wheat	8626	9598 (212 yes, 9386 no)	2.2%
Colon Tumor	2001	62 (22 normal, 40 tumor)	35.5%
Diffuse Large B-Cell Lymphoma (DLBCL)	7130	77 (19 FL, 58 DLBCL)	24.7%
Leukemia	7130	72 (25 AML, 47 ALL)	34.7%
Lung Cancer	7130	96 (10 normal, 86 tumor)	10.4%
Central Nervous System	7130	60 (21 survivors, 39 failures)	35.0%

Table 2. Internet advertisements dataset: similarity matrix for subsets of 20 features. *Each cell in the matrix reports the Kuncheva similarity for the pair of subsets selected by the methods in the corresponding row and column.*

subsets of 20 features		**Univariate**					**Multivariate**		
		χ^2	**IG**	**SU**	**GR**	**OR**	**RF**	**SVM_RFE**	**SVM_ONE**
Univariate	χ^2		1.00	0.85	0.70	1.00	0.09	0.14	0.19
	IG			0.85	0.70	1.00	0.09	0.14	0.19
	SU				0.85	0.85	0.04	0.19	0.09
	GR					0.70	0.04	0.14	0.09
	OR						0.09	0.14	0.19
Multivariate	**RF**							0.09	0.14
	SVM_RFE								0.24
	SVM_ONE								

Table 3. Reuters datasets: similarity matrix for subsets of 20 features. *Each cell in the matrix reports the Kuncheva similarity, averaged across all ten Reuters datasets, for the pair of subsets selected by the methods in the corresponding row and column.*

subsets of 20 features		**Univariate**					**Multivariate**		
		χ^2	**IG**	**SU**	**GR**	**OR**	**RF**	**SVM_RFE**	**SVM_ONE**
Univariate	χ^2		0.80	0.92	0.27	0.52	0.00	0.22	0.28
	IG			0.81	0.21	0.36	0.00	0.20	0.27
	SU				0.24	0.49	0.00	0.22	0.29
	GR					0.46	0.00	0.19	0.17
	OR						0.00	0.21	0.23
Multivariate	**RF**							0.00	0.00
	SVM_RFE								0.24
	SVM_ONE								

Table 4. Biological datasets: similarity matrix for subsets of 20 features. *Each cell in the matrix reports the Kuncheva similarity, averaged across all five biological datasets, for the pair of subsets selected by the methods in the corresponding row and column.*

subsets of 20 features		**Univariate**					**Multivariate**		
		χ^2	**IG**	**SU**	**GR**	**OR**	**RF**	**SVM_RFE**	**SVM_ONE**
Univariate	χ^2		0.74	0.75	0.59	0.59	0.36	0.17	0.14
	IG			0.81	0.58	0.52	0.31	0.16	0.13
	SU				0.71	0.55	0.30	0.16	0.13
	GR					0.46	0.29	0.15	0.13
	OR						0.31	0.18	0.17
Multivariate	**RF**							0.29	0.35
	SVM_RFE								0.57
	SVM_ONE								

Table 5. *SVM_RFE approach: similarity analysis between 10%-RFE setting and 50%-RFE setting.*

Subset size **Domain**	***10***	***20***	***30***	***50***	***100***	***150***
Internet Advertisements	0.6	0.85	0.66	0.86	0.85	0.87
Reuters datasets	0.47	0.44	0.45	0.48	0.52	0.57
Biological datasets	0.42	0.53	0.57	0.63	0.69	0.77

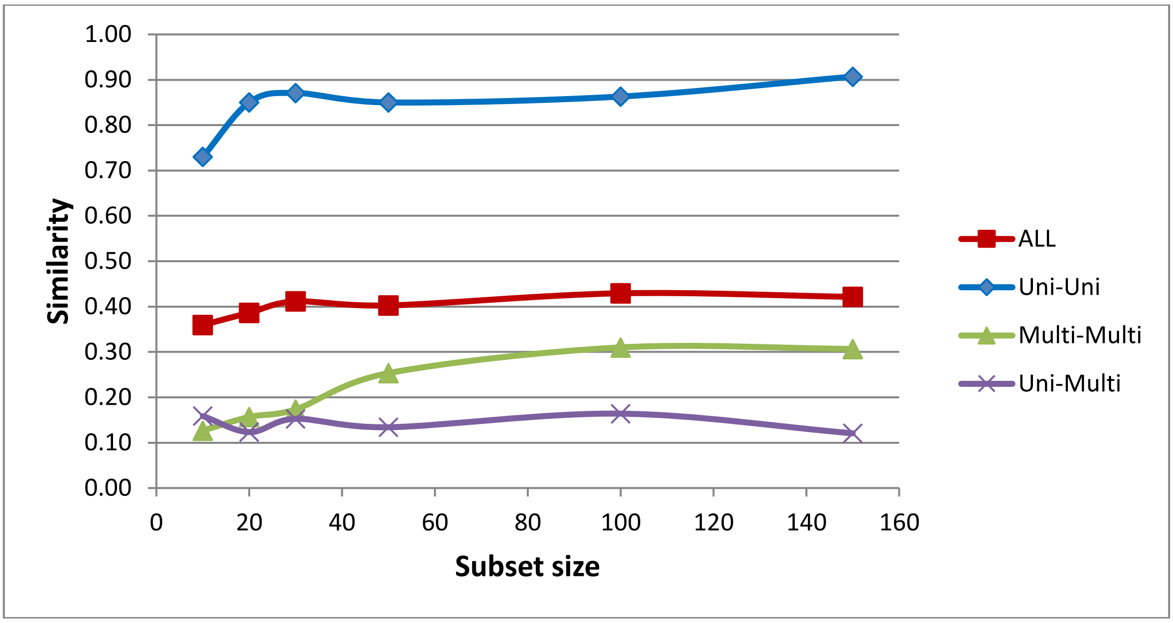

Fig. 1. Internet advertisements dataset: average similarity vs subset size. *For a given subset size, the corresponding similarity matrix is built (as shown in Table 2), and the average similarity is computed over all pair-wise comparisons (ALL), over the pair-wise comparisons where both methods are univariate (Uni-Uni), over the pair-wise comparisons where both methods are multivariate (Multi-Multi), and over the pair-wise comparisons where one method is univariate and the other multivariate (Uni-Multi).*

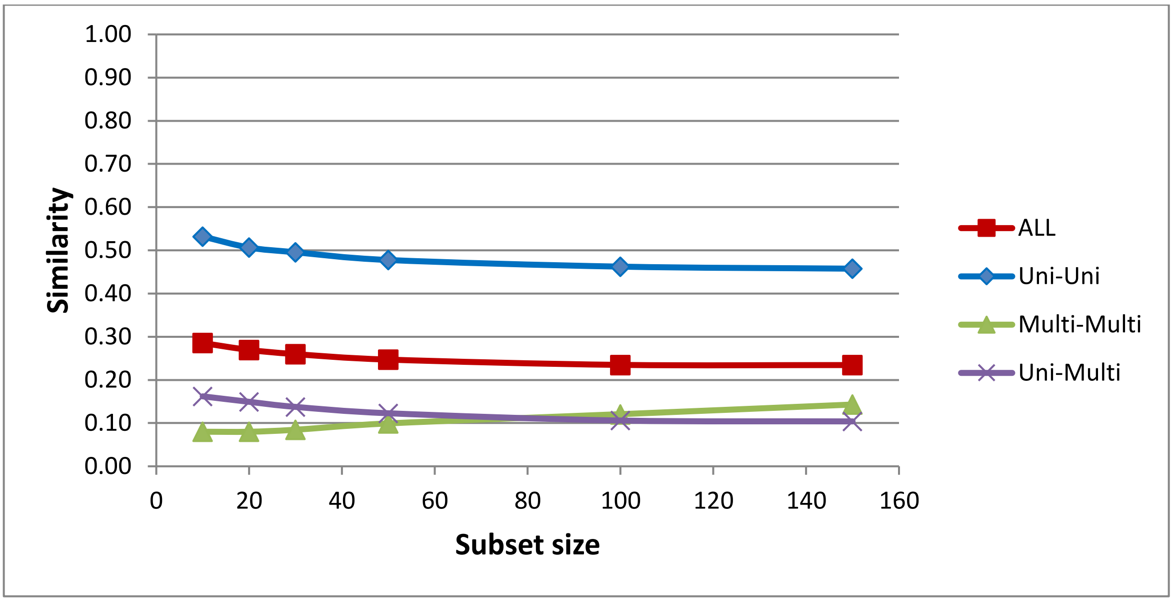

Fig. 2. Reuters datasets: average similarity vs subset size. *For a given subset size, the corresponding similarity matrix is built across all ten Reuters datasets (as shown in Table 3), and the average similarity is computed over all pair-wise comparisons (ALL), over the pair-wise comparisons where both methods are univariate (Uni-Uni), over the pair-wise comparisons where both methods are multivariate (Multi-Multi), and over the pair-wise comparisons where one method is univariate and the other multivariate (Uni-Multi).*

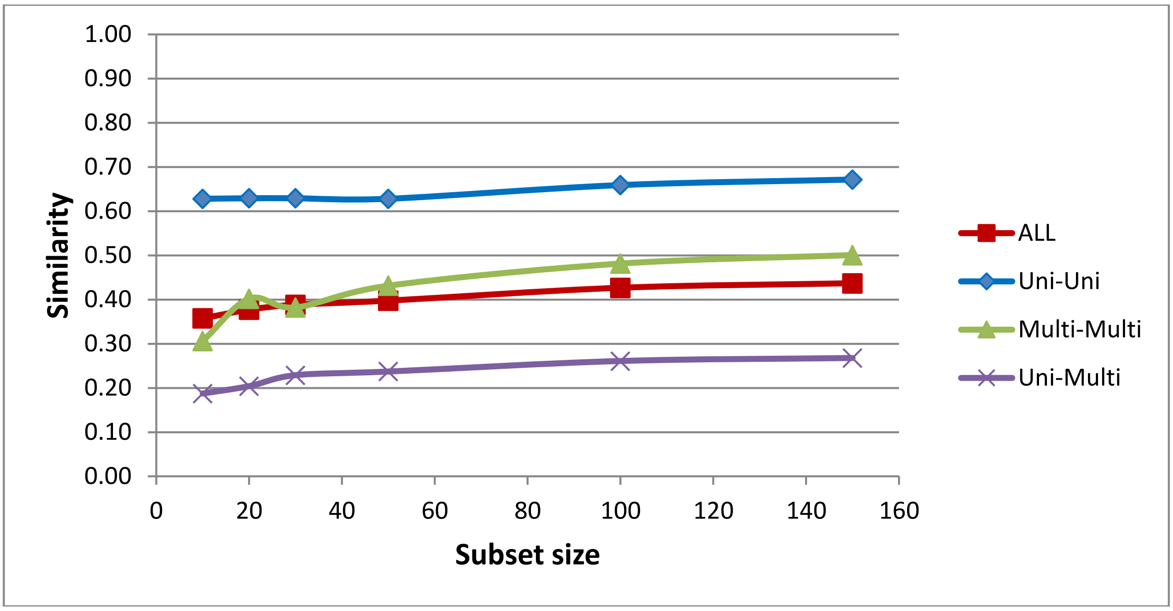

Fig. 3. Biological datasets: average similarity vs subset size. *For a given subset size, the corresponding similarity matrix is built across all five biological datasets (as shown in Table 4), and the average similarity is computed over all pair-wise comparisons (ALL), over the pair-wise comparisons where both methods are univariate (Uni-Uni), over the pair-wise comparisons where both methods are multivariate (Multi-Multi), and over the pair-wise comparisons where one method is univariate and the other multivariate (Uni-Multi).*

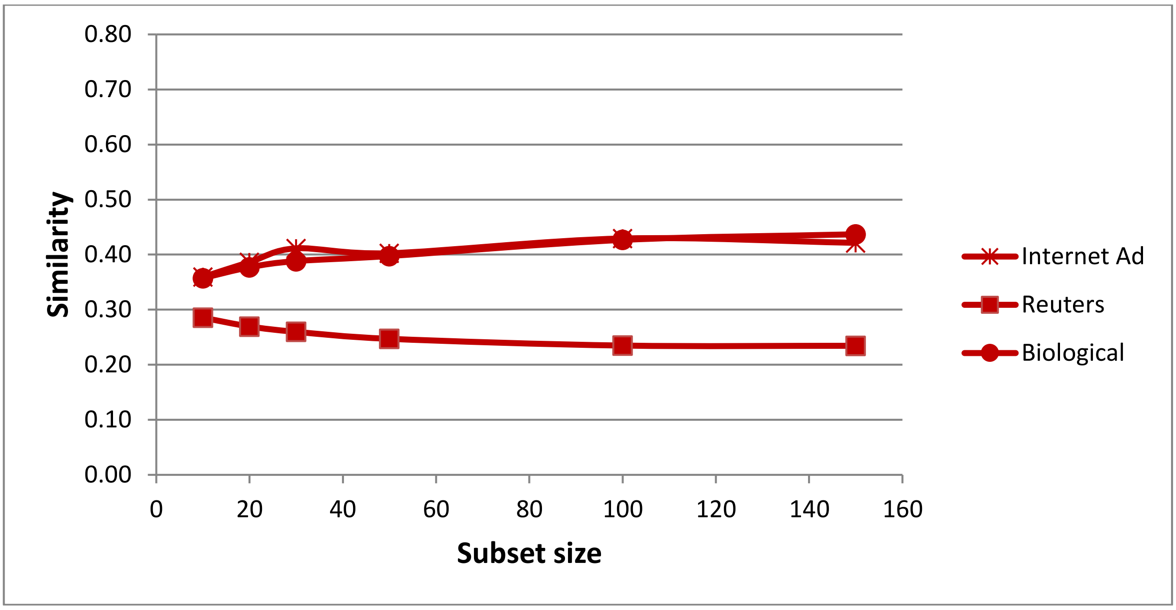

Fig. 4. Comparison of average similarity trends on Internet advertisements, Reuters and biological datasets. *For a given subset size, the average similarity is computed over all pair-wise comparisons in the corresponding similarity matrix.*