The following article has been published in

Cannas L.M, Dessì N., Pes B., *Assessing similarity of feature selection techniques in high-dimensional domains*, Pattern Recognition Letters, Volume 34, Issue 12, 1 September 2013, pp. 1446-1453,

DOI: 10.1016/j.patrec.2013.05.011

Assessing Similarity of Feature Selection Techniques in High-Dimensional Domains

Laura Maria Cannas, Nicoletta Dessì, Barbara Pes

Università degli Studi di Cagliari, Dipartimento di Matematica e Informatica,

Via Ospedale 72, 09124 Cagliari, Italy

{lauramcannas, dessi, pes } @unica.it

Abstract: Recent research efforts attempt to combine multiple feature selection techniques instead of using a single one. However, this combination is often made on an "ad hoc" basis, depending on the specific problem at hand, without considering the degree of diversity/similarity of the involved methods. Moreover, though it is recognized that different techniques may return quite dissimilar outputs, especially in high dimensional/small sample size domains, few direct comparisons exist that quantify these differences and their implications on classification performance. This paper aims to provide a contribution in this direction by proposing a general methodology for assessing the similarity between the outputs of different feature selection methods in high dimensional classification problems. Using as benchmark the genomics domain, an empirical study has been conducted to compare some of the most popular feature selection methods, and useful insight has been obtained about their pattern of agreement.

Keywords: *Feature Selection, Similarity Measures, High-Dimensional Data.*

1 Introduction

Feature selection techniques are critical to the analysis of high dimensional datasets coming from a number of application areas such as text processing, combinatorial chemistry and bioinformatics. In the context of classification problems, the objective of feature selection is three-fold (Guyon and Elisseeff, 2003): improving prediction performance, providing faster and more cost-effective predictors, and achieving a better knowledge of the underlying domain.

Feature selection techniques can be broadly divided into three classes, depending on how they interact with the classifier (Saeys et al., 2007). Filter methods look at the intrinsic properties of the data, independently of the classifier design, and provide a feature weighting, ranking or feature subset as output. In contrast, wrapper methods perform a search in the space of feature subsets, guided by the outcome of a classification model. They often give better results than filter methods, but at the price of a greater computational cost. Finally, embedded methods use the internal parameters of a classification model to perform feature selection, often achieving a good trade-off between performance and computational cost.

Instead of choosing one particular feature selection method, and accepting its outcome as the final subset, different methods can be combined using ensemble approaches (Saeys et al., 2008). Indeed, it is recognized that (Yang et al., 2005) there is not a single universally optimal feature selection technique and that (Yeung et al., 2005) more than one subset of features may discriminate the data equally well, especially in high dimensional/small sample size domains. Moreover, different algorithms may select feature subsets that can be considered local optima in the space of feature subsets, and ensemble approaches might give a better approximation to the optimal subset or ranking of features (Saeys et al., 2008).

A crucial issue in designing an ensemble strategy for feature selection is the degree of diversity among the selectors to be combined (Dietterich, 2000). However, research efforts that exploit multiple feature selection methods instead of using a single one are mainly built on an “ad hoc” basis (Dessì and Pes, 2009; Dutkowski and Gambin, 2007; Leung and Hung, 2010; Tan et al., 2006; Yang et al., 2010), depending on the specific classification problem at hand. There is a lack of systematic studies aiming at proving insight on which methods should be combined, and how this combination should be made, based on the degree of diversity/similarity of the different methods. Moreover, though it is recognized that different techniques may return quite dissimilar outputs when applied to high dimensional/small sample size domains, few direct comparisons exist that quantify these differences and their implications on classification performance.

This paper aims to provide a contribution in this direction by proposing a general methodology for assessing the similarity between the outputs of different feature selection methods in high dimensional classification problems. Specifically, we focus on selection methods that produce a ranking of features based on some scoring criterion that measures the importance of each feature for the predictive task at hand. The resulting ranked list, where features appear in descending order of relevance, can be cut at a proper threshold point in order to produce a subset of highly predictive features. Leveraging on a number of similarity measures proposed in literature, our methodology enables a systematic comparison of the subsets produced by different selection methods, for different values of the cut-off threshold, as to derive a similarity trend for feature subsets of increasing size.

The proposed approach has been evaluated on high dimensional/small sample size datasets from the genomics domain. Specifically, we worked with benchmark datasets deriving from DNA micro-array experiments. An empirical study has been conducted to compare some of the most popular feature selection methods, and useful insight has been obtained about their pattern of agreement.

The paper is organized as follows. Section 2 describes the proposed methodology. Section 3 gives details on datasets and methods used in the empirical study. Experimental results are presented and discussed in section 4. Finally, section 5 contains some final remarks as well as future research directions.

2 The Proposed Approach

As previously mentioned, our methodological approach focuses on feature selection methods that output a ranked list of features. The degree of similarity between lists produced by different methods can be assessed in terms of how consistently the features are ordered in these lists. In particular, when two lists are cut at a given threshold point, the similarity of the resulting feature subsets can be evaluated in terms of their degree of overlapping: a number of similarity measures have been proposed (Kuncheva, 2007; Saeys et al., 2008) that basically compare two feature subsets by evaluating the number of elements that they have in common and by introducing some normalization factor.

The methodology adopted for the analysis is illustrated in Fig. 1 and further detailed by the pseudocode in Fig.2. It has as input a *dataset D* of N features, a *set Met = {Met$_1$, Met$_2$, … , Met$_M$}* of feature selection methods, a *set Thr = {Thr$_1$, Thr$_2$, …, Thr$_T$}* of threshold values, and a *similarity measure I.*

As a first step, each method Met$_m$ (m = 1, …, M) is applied to D in order to produce a ranked list where the N features appear in descending order of relevance. This process is carried out independently for each method resulting in M distinct ranked lists {Ranked$_1$, Ranked$_2$, … , Ranked$_M$}, each expressing a different ordering of the N features (lines 2-4).

Then, we set a threshold value Thr$_t$ (t = 1, …, T) (line 5) and consider only the first Thr$_t$ features from each list Ranked$_m$ (m = 1, …, M), thus obtaining M sub-lists, i.e. feature subsets, of size Thr$_t$ (lines 6-9). The subset resulting from the m-th list in correspondence of the t-th threshold value is denoted as FS$_{tm}$, while the set {FS$_{t1}$, FS$_{t2}$, …, FS$_{tM}$} containing the M subsets of size Thr$_t$ is denoted as FSset$_t$.

Afterwards, the M subsets in the FSset$_t$ are compared, in pairs, using the similarity measure I that provides, as previously mentioned, a criterion to evaluate the degree of overlapping between two subsets. Specifically, it assumes values in the range [0, 1], where 0 means the absence of similarity (no features in common) and 1 the maximum similarity (identical subsets). Leveraging on I, we build a similarity matrix SM$_t$, of size M×M, that stores the similarity value for each pair of the M subsets {FS$_{t1}$, FS$_{t2}$, …, FS$_{tM}$} of size Thr$_t$. Specifically, the element SM$_t$[j][k] of the matrix represents the similarity between FS$_{tj}$ and FS$_{tk}$, as measured by I. Since the matrix is symmetric (SM$_t$[j][k] = SM$_t$[k][j]), only the positions in the upper-right triangular block are considered (lines 10-14).

Finally, to have an overall evaluation of the degree of similarity between the M subsets in the FSset$_t$, an average similarity value S$_t$ is calculated on the matrix SM$_t$ (lines 15-16) as follows:

$$S_t = \frac{2\sum_{j=1}^{M}\sum_{k=j+1}^{M} SM_t[j][k]}{M(M-1)} \qquad \textbf{(1)}$$

i.e. as the average over all pairwise similarity comparisons between the M subsets in the FSset_t.

This analysis is performed for different threshold values Thr_t (t = 1, …, T) and, for each of these, we obtain the corresponding FSset_t, SM_t and S_t. As a consequence, the outputs of our procedure are a list of T FSsets, a list of T similarity matrices and a list of T average similarity values. This way we can easily derive a similarity trend for feature subsets of increasing size.

3 Empirical Study: Methods and Datasets

Consistently with the methodology described in the previous section, an empirical study has been conducted to compare some popular feature selection techniques in the context of high dimensional classification problems. The settings for the analysis have been established as follows.

Feature Selection Methods. With regard to the set $Met = \{Met_1, Met_2, \ldots, Met_M\}$, we considered M = 7 techniques representative of different classes of selection methods. In particular, we chose to experiment with both univariate methods, that rank each feature separately, and multivariate methods, that take into account feature dependencies. For all of them we used the implementation provided by the WEKA machine learning environment (Hall et al., 2009).

Among the univariate techniques, we considered: *Chi Squared (χ^2)* (Liu and Setiono, 1995) as representative of statistic methods; *Information Gain (IG)* (Quinlan, 1986), *Symmetrical Uncertainty (SU)* (Press et al., 1998), and *Gain Ratio (GR)* (Quinlan, 1993) as representative of entropic methods; and finally *OneR (OR)* (Holte, 1993) as representative of methods that incorporate a classification technique (in this case, a simple rule-based classifier).

Among the multivariate techniques, we considered *ReliefF (RF)* (Kononenko, 1994) and *SVM_RFE (SVM)* (Guyon et al., 2002). The basic idea of ReliefF is to estimate the relevance of features based on their ability to distinguish between instances that are near to each other. As representative of embedded techniques, SVM_RFE uses a linear SVM classifier to derive the weights of features. Then, based on these weights, the least important features are removed and the procedure is iteratively repeated on the remaining features.

Similarity Measures. As concerns the *similarity measure I*, we chose to validate our methodology with three well known metrics: the *Overlap index* (Stiglic and Kokol, 2010), the *Jaccard index* (Saeys et al., 2008) and the *Kuncheva index* (Kuncheva, 2007), that are calculated as follows:

$$\text{Overlap} = \frac{| FS_{tj} \cap FS_{tk} |}{Thr_t} \quad (\mathbf{2})$$

$$\text{Jaccard} = \frac{| FS_{tj} \cap FS_{tk} |}{| FS_{tj} \cup FS_{tk} |} \quad (\mathbf{3})$$

$$\text{Kuncheva} = \frac{| FS_{tj} \cap FS_{tk} | - \frac{Thr_t^{\,2}}{N}}{Thr_t - \frac{Thr_t^{\,2}}{N}} \quad (\mathbf{4})$$

where FS_{tj} and FS_{tk} are feature subsets of size Thr_t, obtained from a dataset of dimensionality N. Basically, the above measures evaluate the amount of overlapping between the two subsets.

In particular, the Overlap index is calculated by simple counting of the features that are present in both the subsets and dividing them by the subset size. The Jaccard index divides the number of features that are present in both the subsets by the number of features obtained by the union of the two subsets. Both Overlap and Jaccard measures have a tendency to increase when the size of subsets approaches the total number of features N. To avoid this behavior, the Kuncheva index introduces a correction term that takes into account the probability that a feature is selected by chance: this ensures that the similarity has high value only if it exceeds the similarity by chance (or by design).

Threshold values. For the set $Thr = \{Thr_1, Thr_2, \ldots, Thr_T\}$, we considered numerous threshold values starting from 1 and reaching values that are close to N (i.e. the dataset dimensionality). To best detail the similarity trend for low thresholds, i.e. feature subsets of small size, we used an increasing grain to choose the elements in *Thr*.

Datasets. We chose to perform the analysis on three datasets deriving from DNA micro-array experiments. These are significant examples of high dimensional/small sample size datasets, as they usually store thousands of gene expression profiles measured on a few dozen of samples. The limited number of samples is typical of this domain, and it is due to the high cost associated with the data extraction procedure.

As detailed in Table 1, we worked with: *Leukemia* (Golub et al., 1999), containing 7129 features and 72 samples belonging to patients suffering from acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL); *Colon* (Alon et al., 1999) containing 2000 features and 62 samples distinguished between tumor and normal colon tissues; *Prostate* (Singh et al., 2002), containing 12600 features and 102 samples differed between healthy and tumor prostate tissues. Features correspond to levels of expression of different genes and are continuous.

4 Results and Discussion

In this section we present a summary of the most significant experimental results. The focus here is on the systematic comparison of different ranked lists for different threshold values: setting a given threshold means to set the size of the feature subset that results from cutting a list in correspondence of that threshold. Hence, by spanning a wide range of threshold values, we can derive the pattern of agreement between different ranking methods in terms of the similarity of the resulting feature subsets.

Firstly, we show one of the similarity matrices obtained for Leukemia dataset (Table 2.a), Colon dataset (Table 2.b) and Prostate dataset (Table 2.c). For the sake of space, only the matrices based on the Overlap index are shown, in correspondence of the threshold value of 20: we are comparing here, in pairs, seven subsets of the same size (20 features) as produced by the ranking methods presented in section 3, i.e. *Chi Squared* (χ^2), *Information Gain* (IG), *Symmetrical Uncertainty* (SU), *Gain Ratio* (GR), *OneR* (OR), *ReliefF* (RF) and *SVM_RFE* (SVM). Different shades of gray are used to highlight different similarity ranges: the darker the gray, the higher the similarity values.

Results in Table 2 give useful insight about the pattern of agreement of the considered methods. As regards the univariate approaches (χ^2 , IG, SU, GR, OR), where each feature is evaluated independently from the others, a first evidence is that the χ^2 statistic produces results quite similar to entropic methods IG and SU (the degree of overlapping among the resulting feature subsets is superior to 0.75

for all the datasets). The other entropic method, i.e. GR, turns out similar to IG and SU in Leukemia (Table 2.a) and Prostate (Table 2.c) datasets, but exhibits a different behavior in the Colon dataset (Table 2.b) which is recognized as a more noisy benchmark. Globally, the univariate methods are more similar to each other than to the multivariate approaches, i.e. RF and SVM_RFE, that work differently since take into account feature dependencies. Indeed, both RF and SVM_RFE produce feature subsets that overlap to a lesser extent with the subsets selected by other methods. These findings are confirmed by the matrices (here omitted) derived using different threshold values and different metrics (Jaccard and Kuncheva indexes).

Interestingly, the method that gives the most dissimilar results, i.e. SVM_RFE, turns out very effective in identifying highly predictive features, as witnessed by literature (Guyon et al., 2002; Zhou and Tuck, 2007) as well as by some experiments we performed to evaluate the classification accuracy trend for feature subsets of increasing size. Specifically, for each of the considered feature selection methods, we cut the resulting ranked list in correspondence of different threshold values and, for each threshold, we evaluated the accuracy achieved by a K-NN classifier on the resulting feature subset (using a 10-fold cross-validation setting). Fig. 3 shows, for each of the considered datasets, a comparison among the accuracy on the subsets selected by SVM_RFE and those selected by OR and IG (as representative of different classes of selection methods). As we can see, when feature selection is performed by SVM_RFE, a small number of features is sufficient to reach a very high accuracy: 100% for Leukemia with a threshold (i.e. subset size) of 20; 94% for Colon with a threshold of 10; 98% for Prostate with a threshold of 30. In the same threshold conditions, the other selection methods result in a lower classification performance, as witnessed by IG and OR curves in Fig. 3 (as well as by curves relative to χ^2, SU, GR and RF, omitted in the figure for more readability).

It is not a novelty in literature that (Saeys et al., 2007) there not exists a threshold value (i.e. subset size) “optimal” for all the ranking methods and that, in correspondence of the same threshold, different methods can result in a different predictive performance. The added value of our approach is to help explain these differences in terms of the degree of similarity/dissimilarity between the subsets selected, for a given threshold, by different methods.

To globally evaluate the pattern of agreement between the seven methods here considered (i.e. χ^2, IG, SU, GR, OR, RF, SVM_RFE), we systematically investigated the overall similarity trend for feature subsets of increasing size. In more detail, according to the methodology stated in section 2, we explored a wide range of threshold values and, for each of these, we derived a similarity matrix (where the subsets selected by the considered methods are compared in pairs) and the corresponding average similarity (obtained as the average over all the pairwise comparisons between the involved subsets). Results on the average similarity are summarized in Fig. 4 and Fig. 5.

Specifically, curves in Fig 4.a, Fig. 4.b and Fig. 4.c show (for Leukemia, Colon and Prostate respectively) the overall similarity trend obtained for different similarity measures, i.e. Overlap, Jaccard and Kuncheva. In all the datasets, the Overlap and the Jaccard curves exhibit an analogous behavior (though the Jaccard index results in lower similarity values due to a different normalization factor). In particular, both these measures have a drawback: they tend to increase as the size of subsets approaches the total number of features. The Kuncheva curve, instead, coincides with the Overlap curve for low threshold values (meaning feature subsets of small size) but exhibits a very different behavior for higher threshold values, due to the correction introduced for the probability that a feature is selected by chance (see section 3): this probability obviously grows as the subset size approaches the dimensionality of the original dataset. Results in Fig. 4 clearly show the effectiveness of the correction term used in the Kuncheva approach.

We also observe that, in all the datasets, the Kuncheva curve reaches a peak in correspondence of a threshold value (1025 for Leukemia, 135 for Colon, 2228 for Prostate) after which χ^2 and entropic methods assign a null weight to features. Moreover, all the similarity curves reach a maximum in the range of very low threshold values: in the neighborhood of 10 for Leukemia and Colon, while for Prostate we can see a peak at threshold 2 and another local maximum in the range 14-18.

The behavior at low threshold values, most interesting when looking at small subsets of highly informative features, is best highlighted in Fig. 5 where only the Overlap curve is shown. These results seem to suggest that, for small subset sizes, it is possible to identify a threshold point (or a range of thresholds) where the different feature selection methods reach, on average, a higher agreement, despite the specific behavior of every single method.

5 Concluding Remarks and Future Work

A general methodology has been presented for assessing the similarity between the outputs of different feature selection methods in high dimensional classification problems. Leveraging on a number of metrics from literature, the proposed approach allows to derive a similarity trend for feature subsets of increasing size, obtained cutting the original ranked lists at different threshold points.

Through an empirical evaluation on the genomics domain, useful insight has been obtained about the pattern of agreement of some popular feature selection techniques. For example, the considered multivariate methods, i.e. SVM_RFE and ReliefF, turn out quite dissimilar to each other as well as to the considered univariate methods. Furthermore, among the entropic univariate methods, the Gain Ratio is the one that exhibits the greatest differences. However, despite the specific behavior of every single technique, a range of threshold values can be identified where the different techniques reach a higher consensus.

As regards the choice of the similarity measure, our experimental results confirm the effectiveness of the Kuncheva index in evaluating the degree of consistency between a pair of feature subsets with a proper correction for the probability that a feature is included in those subsets purely by chance. Hence, it can be regarded as a suitable choice when evaluating the similarity between the outputs of two feature selection processes.

The similarity analysis presented in this study can be preliminary to properly devising an ensemble strategy for feature selection. When aggregating the outputs of different methods (e.g. for deriving a consensus feature subset containing the most frequently selected features) we cannot prescind from considering the degree of diversity/similarity of the involved methods. Indeed, combining two or more feature selection techniques that give almost identical results would not be beneficial. In an ensemble perspective, the aim should be to reach a consensus result among methods that are capable of giving different and complementary representations of the considered domain. In our opinion, the approach presented in this paper might represent an useful contribution to this issue.

For future work, we plan to extend our research in two main directions. First, ensemble feature selection will be empirically explored by examining different ways of combining feature ranking techniques on the basis of their degree of similarity/dissimilarity, as assessed according to the methodology here proposed.

As a second research direction, further experiments will be performed on different application domains. In particular, text categorization is a very challenging classification task (Joachims, 1998; McCallum and Nigam, 1998): in this context documents are represented by a *bag-of-words*, that is a vector whose elements are frequency-based weights of the words in the text. The vector dimension, corresponding to the size of the considered vocabulary, can be in the order of hundreds of thousands of words, making it imperative to apply suitable feature selection strategies in order to achieve good classification performance.

Inspired by the success in the text categorization field, the *bag-of-words* representation has become one of the most popular methods for representing image content (Zhang et al., 2010) and has been successfully applied to visual categorization and object recognition problems (Csurka et al., 2004; Sivic and Zisserman, 2003). In this context, local image features are clustered into “visual words” to be used for classification/recognition purposes (the feature values are the normalized histogram bin counts of the visual words). The accuracy of the *bag-of-words* classifiers, however, is often limited by the presence of uninformative features extracted from the background or irrelevant image segments. Hence, as witnessed by recent studies (Creusen et al., 2009; Liu et al., 2008; Naikal et al., 2011; Turcot and Lowe, 2009), feature selection can significantly improve the accuracy achieved in categorization and recognition tasks.

Despite their specificities, the aforementioned application domains can all benefit from the availability of different feature selection techniques as well as of a methodology, as the one here presented, to compare the outputs of these techniques. Our future experiments will be then focused on comparing and combining several selection methods in application areas different from the genomics domain explored in this paper. In such experimental studies, it could be interesting to consider a larger number of selection methods, including novel learning algorithms such as (Prinzie and Van den Poel, 2008; Ting et al., 2010) that are specifically designed to handle high-dimensional feature spaces and include the automatic relevance detection of features in those spaces.

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”.

References

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

Creusen, I.M., Wijnhoven, R.G.J., de With, P.H.N., 2009. Applying Feature Selection Techniques for Visual Dictionary Creation in Object Classification, Proceedings of the 2009 International Conference on Image Processing, Computer Vision, & Pattern Recognition, IPCV, 722-727, CSREA Press.

Csurka, G., Dance, C. R., Fan L., Willamowski J., Bray C., 2004. Visual Categorization with Bags of Keypoints, In Workshop on Statistical Learning in Computer Vision, ECCV, 1-22.

Dessì, N., Pes, B., 2009. An Evolutionary Method for Combining Different Feature Selection Criteria in Microarray Data Classification. Journal of Artificial Evolution and Applications, Volume 2009, Article ID 803973, Hindawi.

Dietterich, T., 2000. Ensemble methods in machine learning. Proceedings of the 1st International Workshop on Multiple Classifier Systems, 1-15.

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

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., 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., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software: an update. SIGKDD Explorations, vol. 11, no. 1, 10-18.

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

Joachims, T., 1998. Text categorization with support vector machines: Learning with many relevant features. In Proceedings of the 10th European Conference on Machine Learning, 137-142, Chemnitz, Germany.

Kononenko, I., 1994. Estimating Attributes: Analysis and Extensions of RELIEF. Proceedings of the European Conference on Machine Learning, ECML-94, 171-182.

Kuncheva, L.I., 2007. A Stability Index for Feature Selection, International Multi-Conference: Artificial Intelligence and Applications, 390-395, ACTA Press Anaheim, CA, USA.

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, Vol. 7, No. 1, 108-117.

Liu, D., Hua, G., Viola, P., Chen, T., 2008. Integrated Feature Selection and Higher-order Spatial Feature Extraction for Object Categorization, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1-8.

Liu, H. Setiono, R., 1995. Chi2: Feature selection and discretization of numeric attributes, Proc. IEEE 7th International Conference on Tools with Artificial Intelligence, 338-391.

McCallum, A., Nigam, K., 1998. A comparison of event models for naive bayes text classification, AAAI Workshop on Learning for Text Categorization, 41-48, Madison, WI.

Naikal, N., Yang, A.Y., Sastry, S.S., 2011. Informative Feature Selection for Object Recognition via Sparse PCA, Proceedings of the 2011 International Conference on Computer Vision, 818-825.

Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T., 1998. Numerical Recipes in C, Cambridge University Press.

Prinzie, A., Van den Poel, D., 2008. Random forests for multiclass classification: Random MultiNomial Logit, Expert Systems with Applications, Vol. 34, Issue 3, pp. 1721-1732.

Quinlan, J.R., 1986. Induction of decision trees, Machine Learning, vol. 1, no. 1, pp. 81-106.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.

Saeys, Y., Inza, I., Larranaga P., 2007. A review of feature selection techniques in bioinformatics. Bioinformatics, vol. 23 no.19, 2507-2517.

Saeys, Y., Abeel, T., Van de Peer, Y., 2008. Robust Feature Selection Using Ensemble Feature Selection Techniques. Proceedings of ECML PKDD '08, LNAI 5212, 313-325, Springer.

Singh D., Febbo P.G., Ross K., Jackson D.G., Manola J., Ladd C., Tamayo P., Renshaw A.A., D'Amico A.V., Richie J.P., Lander E.S., Loda M., Kantoff P.W., Golub T.R., Sellers W.R., 2002. Gene Expression Correlates of Clinical Prostate Cancer Behavior, Cancer Cell, 1(2):203-209.

Sivic, J., Zisserman, A., 2003. Video Google: A Text Retrieval Approach to Object Matching in Videos, Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV), 1470-1477.

Stiglic, G., Kokol, P., 2010. Stability of Ranked Gene Lists in Large Microarray Analysis Studies, Journal of Biomedicine and Biotechnology, Volume 2010, Article ID 616358, Hindawi.

Tan, F., Fu, X.Z., Zhang, Y., Bourgeois, A.G., 2006. Improving Feature Subset Selection Using a Genetic Algorithm for Microarray Gene Expression Data. IEEE Congress on Evolutionary Computation, Vancouver, BC, Canada, 2529- 2534.

Ting, J., D'Souza, A., Vijayakumar, S., Schaal, S., 2010. Efficient Learning and Feature Selection in High-Dimensional Regression, Neural Computation, Vol. 22, Issue 4, pp. 831-886.

Turcot, P., Lowe, D.G., 2009. Better matching with fewer features: The selection of useful features in large database recognition problems, IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops), 2109- 2116.

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 2010, 11 (Suppl 1):S5.

Yang, Y.H, Xiao, Y., Segal, M.R., 2005. Identifying differentially expressed genes from microarray experiments via statistic synthesis. Bioinformatics, 21(7), 1084-1093.

Yeung, K.Y., Bumgarner, R.E., Raftery, A.E., 2005. Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. Bioinformatics, 21, 2394-2402.

Zhang, Y., Jin, R., Zhou, Z., 2010. Understanding Bag-of-Words Model: A Statistical Framework, International Journal of Machine Learning and Cybernetics, Vol. 1, Issue 1-4, 43-52.

Zhou, X., Tuck, D.P., 2007. MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data, Bioinformatics 23(9): 1106-1114.

Table 1. Micro-array datasets used in the empirical study.

Dataset	No. of samples	Distribution among classes	No. of features	Reference
Leukemia	72	47 ALL + 25 AML	7129	Golub et al., 1999
Colon	62	40 tumor + 22 normal	2000	Alon et al., 1999
Prostate	102	52 tumor + 50 normal	12600	Singh et al., 2002

Table 2. Similarity matrices for Leukemia dataset (a), Colon dataset (b), and Prostate dataset (c); results are obtained with the Overlap index, in correspondence of the threshold value (i.e. subset size) of 20

(a)

	X2	**IG**	**SU**	**GR**	**OR**	**RF**	**SVM**
X2	1	0.85	0.80	0.75	0.70	0.45	0.40
IG	0.85	1	0.90	0.70	0.70	0.40	0.40
SU	0.80	0.90	1	0.80	0.70	0.45	0.35
GR	0.75	0.70	0.80	1	0.60	0.45	0.35
OR	0.70	0.70	0.70	0.60	1	0.50	0.40
RF	0.45	0.40	0.45	0.45	0.50	1	0.45
SVM	0.40	0.40	0.35	0.35	0.40	0.45	1

(b)

	X2	**IG**	**SU**	**GR**	**OR**	**RF**	**SVM**
X2	1	0.80	0.80	0.40	0.65	0.50	0.20
IG	0.80	1	0.95	0.40	0.55	0.35	0.20
SU	0.80	0.95	1	0.45	0.55	0.35	0.20
GR	0.40	0.40	0.45	1	0.40	0.35	0.15
OR	0.65	0.55	0.55	0.40	1	0.45	0.20
RF	0.50	0.35	0.35	0.35	0.45	1	0.25
SVM	0.20	0.20	0.20	0.15	0.20	0.25	1

(c)

	X2	**IG**	**SU**	**GR**	**OR**	**RF**	**SVM**
X2	1	0.85	0.75	0.65	0.70	0.45	0.10
IG	0.85	1	0.90	0.80	0.60	0.50	0.15
SU	0.75	0.90	1	0.90	0.55	0.45	0.15
GR	0.65	0.80	0.90	1	0.45	0.40	0.20
OR	0.70	0.60	0.55	0.45	1	0.45	0.15
RF	0.45	0.50	0.45	0.40	0.45	1	0.30
SVM	0.10	0.15	0.15	0.20	0.15	0.30	1

Fig.1. The methodology.

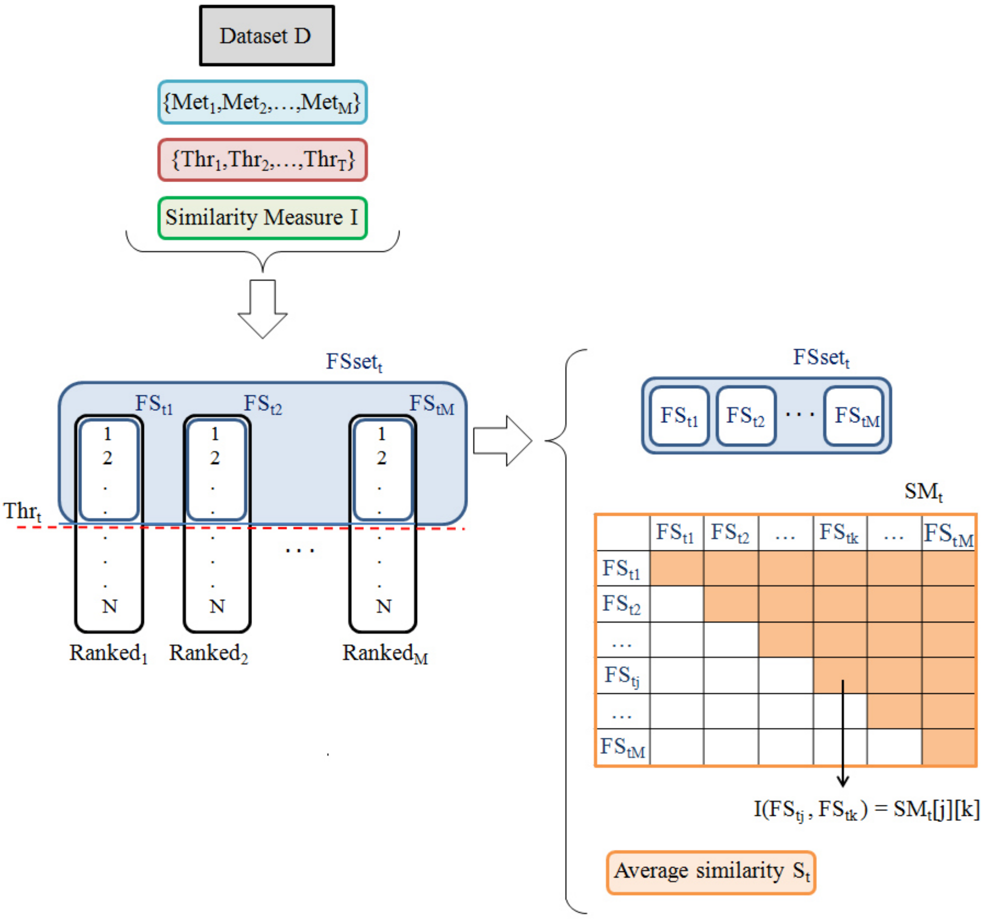


Fig.2. Pseudocode describing the methodology.

```
Input:
D – Dataset of N features
Met – Set of M filter methods: Met = {$Met_1$,$Met_2$,…,$Met_M$}
Thr – Set of T threshold values: Thr = {$Thr_1$,$Thr_2$,…,$Thr_T$}
I – Similarity measure

begin
for m: 1 to M
  $Ranked_m$ = rank the N features according to $Met_m$
end for
for t: 1 to T
  // create FSset
  for m: 1 to M
    $FS_{tm}$ = select the first $Thr_t$ features from $Ranked_m$
  end for
  // create matrix
  for j: 1 to M
    for k: j+1 to M
      $SM_t$[j][k] = evaluate similarity between $FS_{tj}$ and $FS_{tk}$
                     using I
    end for
  end for
  // calculate average similarity
  $S_t$ = calculate average similarity on $SM_t$
end for
end

Output:
List of T FSset
List of T similarity matrices SM
List of T average similarity values S
```

Fig. 3. Classification accuracy vs threshold (i.e. subset size) for Leukemia dataset (a), Colon dataset (b) and Prostate dataset (c).

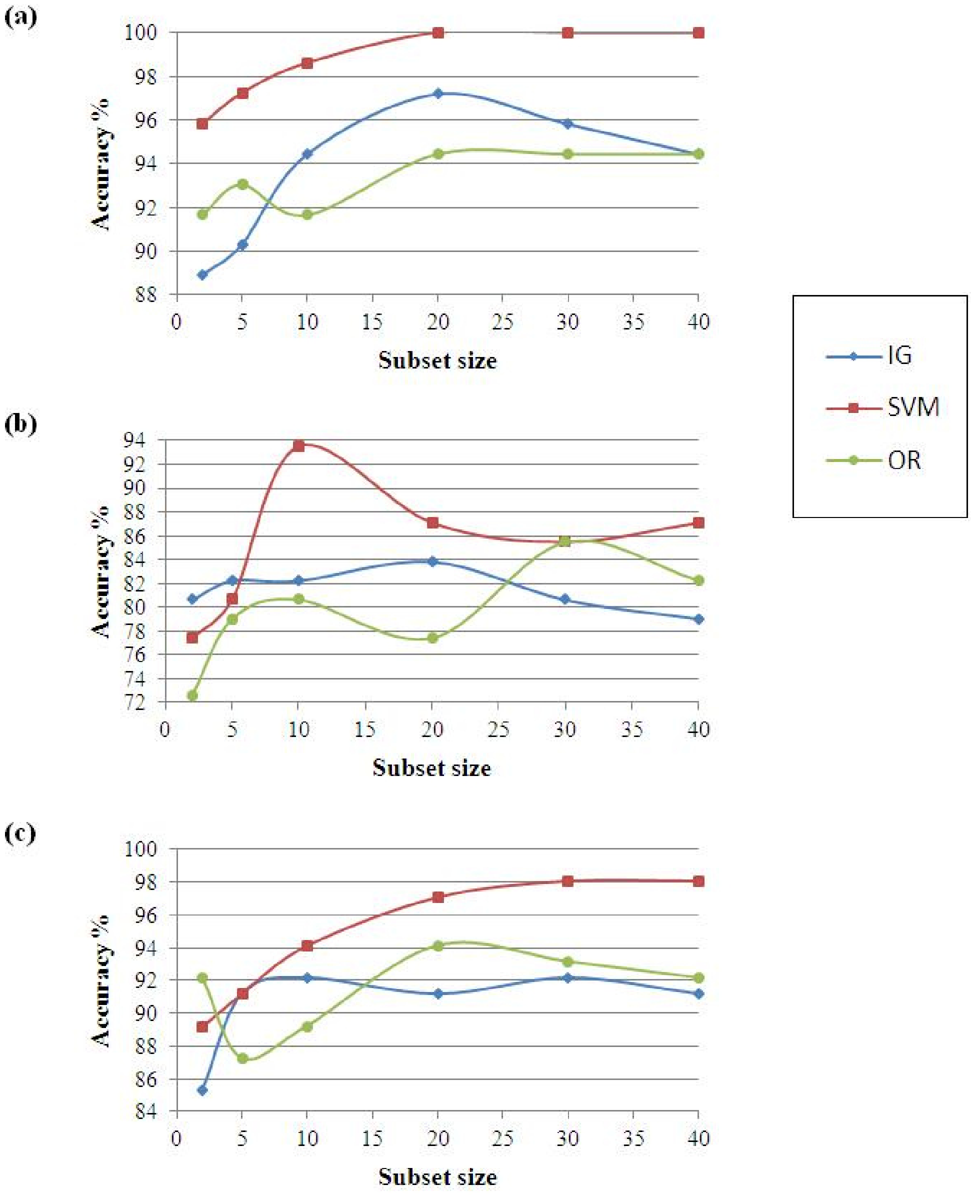


Fig. 4. Average similarity vs threshold (i.e. subset size) for Leukemia dataset (a), Colon dataset (b) and Prostate dataset (c).

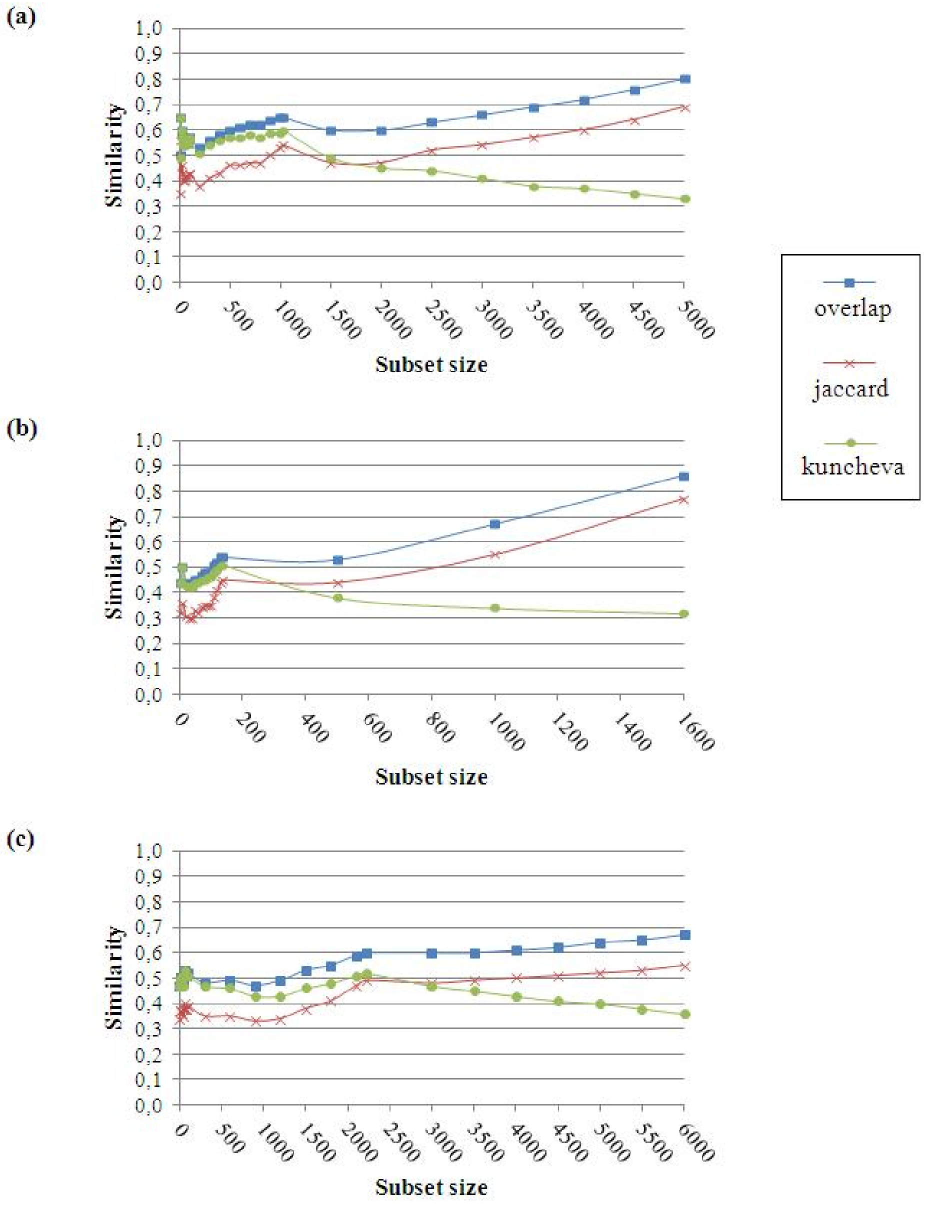
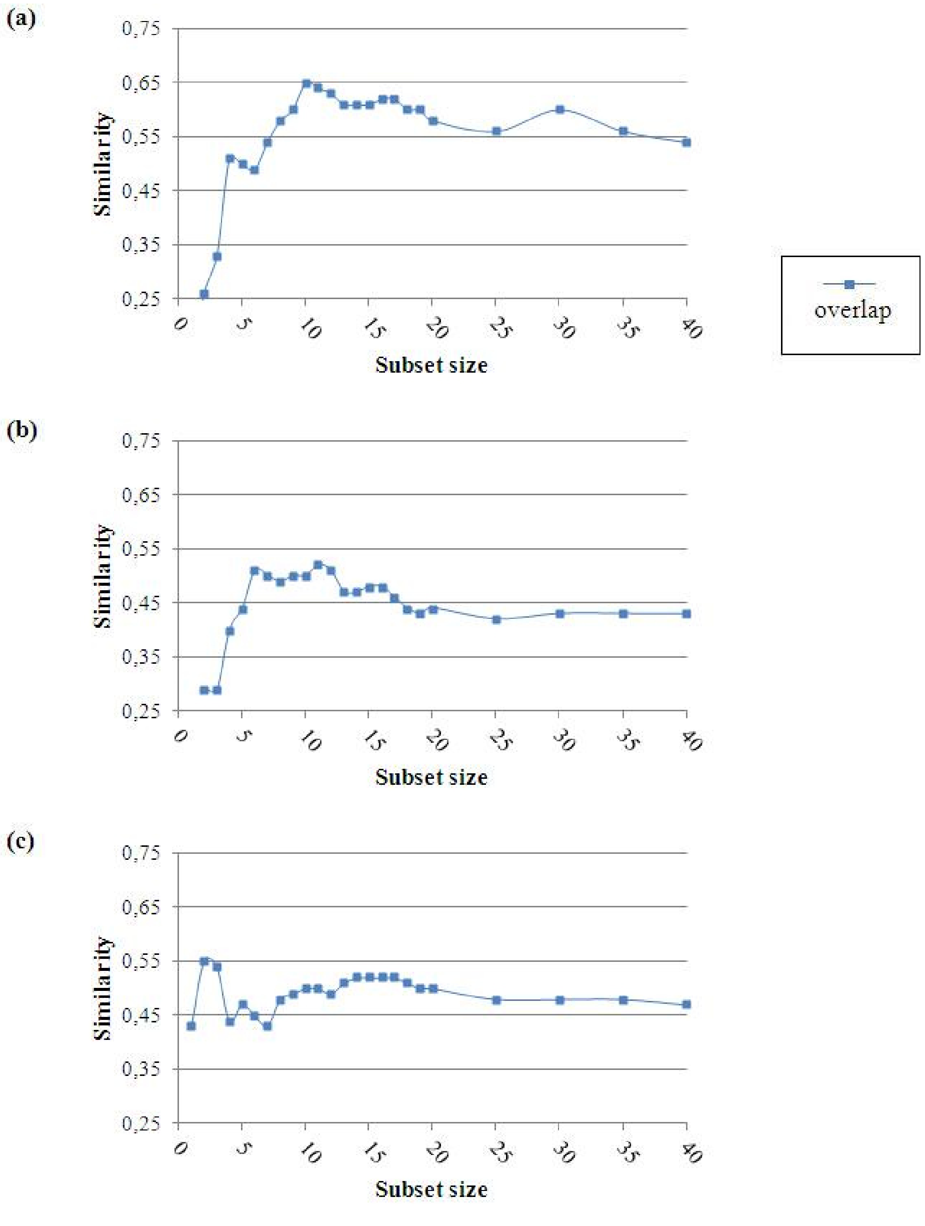


Fig. 5. Average similarity vs threshold (i.e. subset size) in the range of low threshold values, for Leukemia dataset (a), Colon dataset (b) and Prostate dataset (c).


(a)
(b)
(c)
Similarity
Subset size
overlap