Novel Feature Selection Methods for High Dimensional Data

Verónica Bolón-Canedo, Noelia Sánchez-Maroño and Amparo Alonso-Betanzos
Department of Computer Science, University of A Coruña, Campus de Elviña s/n, A Coruña 15071, Spain

1 INTRODUCTION AND STATE OF THE ART

In the past 20 years, the dimensionality of the datasets involved in data mining has increased dramatically, as can be seen in (Zhao and Liu, 2011). This fact is reflected if one analyzes the *dimensionality* (samples $\times$ features) of the datasets posted in the UC Irvine Machine Learning Repository (Frank and Asuncion, 2010). In the 1980s, the maximal dimensionality of the data was about 100; then in the 1990s, this number increased to more than 1500; and finally in the 2000s, it further increased to about 3 million. The proliferation of this type of datasets with very high ($>$ 10000) dimensionality has brought unprecedented challenges to machine learning researchers. Learning algorithms can degenerate their performance due to overfitting, learned models decrease their interpretability as they are more complex, and finally speed and efficiency of the algorithms decline in accordance with size.

Machine learning can take advantage of feature selection methods to be able to reduce the dimensionality of a given problem. *Feature selection* (FS) is the process of detecting the relevant features and discarding the irrelevant and redundant ones, with the goal of obtaining a small subset of features that describes properly the given problem with a minimum degradation or even improvement in performance (Guyon et al., 2006). Feature selection, as it is an important activity in data preprocessing, has been an active research area in the last decade, finding success in many different real world applications, especially those related with classification problems.

There are several situations that can hinder the process of feature selection, such as the presence of irrelevant and redundant features, noise in the data or interaction between attributes. In the presence of hundreds or thousands of features, such as DNA microarray analysis, researchers notice (Yu and Liu, 2004) that is common that a large number of features is not informative because they are either irrelevant or redundant with respect to the class concept. Moreover, when the number of features is high but the number of samples is small, machine learning gets particularly difficult, since the search space will be sparsely populated and the model will not be able to distinguish correctly the relevant data and the noise (Provost, 2000).

Feature selection methods usually come in three flavors: *filter*, *wrapper*, and *embedded* methods (Guyon et al., 2006). The *filter* model relies on the general characteristics of training data and carries out the feature selection process as a pre-processing step with independence of the induction algorithm. On the contrary, *wrappers* involve optimizing a predictor as a part of the selection process. Halfway these two models one can find *embedded* methods, which perform feature selection in the process of training and are usually specific to given learning machines. By having some interaction with the predictor, wrapper and embedded methods tend to obtain higher prediction accuracy than filters, at the cost of a higher computational cost.

There exist numerous papers and books proving the benefits of the feature selection process (Guyon et al., 2006; Dash and Liu, 1997; Kohavi and John, 1997; Zhao and Liu, 2011). However, most researchers agree that there is not a so-called "best method" and their efforts are focused on finding a good method for a specific problem setting. Therefore, new feature selection methods are constantly emerging using different strategies: a) combining several feature selection methods, which could be done by using algorithms from the same approach, such as two filters (Zhang et al., 2008), or coordinating algorithms from two different approaches, usually filters and wrappers (Peng et al., 2010); b) combining feature selection approaches with other techniques, such as feature extraction (Vainer et al., 2011) or tree ensembles (Tuv et al., 2009); c) reinterpreting existing algorithms (Sun and Li, 2006), sometimes to adapt them to specific problems (Sun et al., 2008); d) creating new methods to deal with still unresolved situations (Chidlovskii and Lecerf, 2008; Loscalzo et al., 2009) and e) using an ensemble of feature selection techniques to ensure a better behavior (Saeys et al., 2008).

Bolón-Canedo V., Sánchez-Maroño N. and Alonso-Betanzos A..
Novel Feature Selection Methods for High Dimensional Data.

2 RESEARCH PROBLEM

As mentioned in the introduction, in the last years the dimensionality of datasets involved in data mining applications has increased steadily. This large-scale data carries new opportunities and challenges to computer scientists, giving the opportunity for discovering subtle population patterns and heterogeneities that were not possible with small-scale data. However, the massive sample size and high dimensionality of data introduce new computational challenges. Theoretically, having more data should give more discriminating power. However, the nature of high dimensionality of data can cause the so-called problem of *curse of dimensionality* or *Hughes effect* (Hughes, 1968). This phenomenon occurs when the model has to be learned from a finite number of data samples in a high-dimensional feature space with each feature having a number of possible values, and so an enormous amount of training data are required to ensure that there are several samples with each combination of values. The Hughes effect is therefore known as the situation where with a fixed number of training samples, the predictive power of the learner reduces as the feature dimensionality increases. In this situation, feature selection plays a crucial role.

There exists a vast body of feature selection methods in the literature, including filters based on distinct metrics (e.g. entropy, probability distributions or information theory) and embedded and wrappers methods using different induction algorithms. The proliferation of feature selection algorithms, however, has not brought about a general methodology that allows for intelligent selection from existing algorithms. In order to make a correct choice, a user not only needs to know the domain well, but also is expected to understand technical details of available algorithms (Liu and Yu, 2005). On top of this, most algorithms were developed when dataset sizes were much smaller, but nowadays distinct compromises are required for the case of small-scale and large-scale (big data) learning problems. Small-scale learning problems are subject to the usual approximation-estimation trade-off. In the case of large-scale learning problems, the trade-off is more complex because it involves not only the accuracy of the selection but also other aspects, such as stability (i.e. the sensitivity of the results to training set variations) or scalability.

The objective of this research is two-fold. First, an analysis of classical feature selection is performed, evaluating the adequacy of different methods in different situations. Moreover, the benefits of feature selection have been studied in different data settings: a) datasets with a number of samples much higher than the number of features; b) datasets with a number of features much higher than the number of samples; and c) datasets with a high number of features and samples. After studying the feature selection domain, the second part of the research is devoted to develop novel feature selection to be applied to high dimensional datasets.

3 OUTLINE OF OBJECTIVES

As mentioned above, the main goals of the thesis are to analyze in detail the feature selection domain and then, to develop novel feature selection methods for high-dimensional data. An outline of this can be seen in Figure 1, where each objective is divided in several subobjectives, which will be following described in detail. Notice that the diagram also includes information about the level of completion of each task.

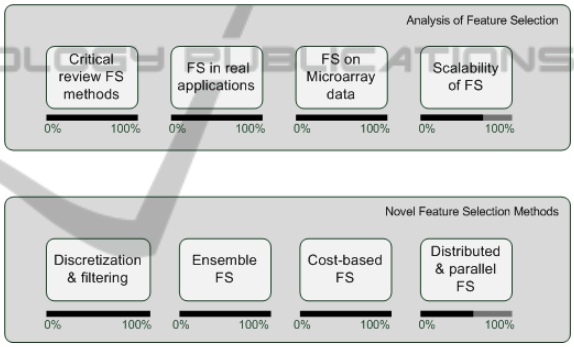


Figure 1: Outline of objectives.

1. Analysis of classical feature selection and application to real problems:
   (a) Critical review of the most popular feature selection methods in the literature by checking their performance in an artificial controlled experimental scenario. In this manner, the ability of the algorithms to select the relevant features and to discard the irrelevant ones without permitting noise or redundancy to obstruct this process is evaluated.
   (b) Application of classic feature selection to real problems in order to check their adequacy. Specifically, testing the effectiveness of feature selection in two problems found in the medical domain: tear film lipid layer classification and K-complex identification in sleep apnea.
   (c) Analysis of the behavior of feature selection in a very challenging field: DNA microarray classification. DNA microarray data is a hard challenge for machine learning researchers due to the high number of features (around 10 000)

but small sample size (typically one hundred or less). For this purpose, it is necessary to review the most up-to-date algorithms developed ad-hoc for this type of data, as well as studying their particularities.

(d) Analysis of the scalability of existing feature selection methods. With the advent of high-dimensionality, machine learning researchers are not focused only in the accuracy of the selection, but also in the scalability of the solution. Therefore, this issue must be addressed, covering the three situations described in Section 2: high number of samples, high number of features, high number of samples and features.

2. Development of novel feature selection methods for high-dimensional data

(a) Development of a new framework which consists of combining discretization and filter methods. This framework is successfully applied to intrusion detection and microarray data classification.

(b) Development of a novel method for dealing with high-dimensional data: an ensemble of filters and classifiers. The idea of this ensemble is to apply several filters based on different metrics and then joining the results obtained after training a classifier with the selected subset of features. In this manner, the user is released from the task of choosing an adequate filter for each dataset.

(c) Proposal for a new framework for cost-based feature selection. In this manner, the scope of feature selection is broaden by taking into consideration not only the relevance of the features but also their associated costs. The proposed framework consists of adding a new term to the evaluation function of a filter method so that the cost is taken into account.

(d) Distributing feature selection. There are two common types of data distribution: (a) horizontal distribution wherein data are distributed in subsets of instances; and (b) vertical distribution wherein data are distributed in subsets of attributes. Both approaches are tested, employing for this sake filter and wrapper methods.

4 STAGE OF THE RESEARCH

This research has been started in 2009 and at present if faces its final stage. As can be seen in Figure 1, the great majority of objectives have been completed. For the analysis of existing feature selection methods, the first three tasks have been addressed. An investigation about the particularities of classical methods has been carried out, allowing to select the most appropriate algorithms to face problems of real systems. Moreover, a study in detail about feature selection in microarray DNA data has been accomplished, since this is a challenging domain which is a trending topic for machine learning researchers. With the appearance of high-dimensionality, it is also necessary to study the scalability of existing feature selection methods. The scalability of filter-based methods has been tackled with successful results. However, the analysis of embedded and wrapper methods is still in progress and some conclusions on this issue are expected to be obtained soon.

The second objective of the thesis is also almost completed. As mentioned in Section 1, the current tendency in feature selection is not toward developing new algorithmic measures, but toward favoring the combination or modification of existing algorithms. For this reason, this goal is focused in exploring different strategies to deal with the new problematics which have emerged derived from the big data explosion. Particularly, a novel method which combines discretization and filter algorithms has been proposed and its effectiveness was demonstrated in different applications, such as intrusion detection or microarray data. Then, based on the assumption that a set of experts is better than a single expert, an ensemble of filters and classifiers was proposed and tested again on microarray data as well as in some classical datasets. Another task that has been concluded was to design methods for cost-based feature selection, which has been motivated by the fact that, in some cases, features has its own risk or cost, and this factor must be taken into account as well as the accuracy. Finally, a recent topic of interest has arisen which consists of distributing the feature selection process. For this sake, several approaches have been proposed, splitting the data both vertically and horizontally. However, in some cases the partitioning of the datasets can introduce some redundancy among features. For solving this problem, new partitioning schemes are being investigated, for example by dividing the features according to some goodness measure. Moreover, a final research line in progress is devoted to perform parallel feature selection using a cluster computing framework called Spark (Spark, nd). This distributed programming model has been proposed to handle large-scale data problems. However, most existing feature selection techniques are designed to run in a centralized computing environment and their implementations have to be adapted to this new technology. By using Spark, the final user will be released of the de-

cision of how to distribute the data.

5 RESULTS

As stated above, this thesis is facing its final stage and therefore the objectives outlined in Section 3 have been addressed. In this section, some experimental results as well as the key publications will be presented.

5.1 Review of Feature Selection Methods

The first step when dealing with feature selection should be to review the existing algorithms and to check their performance under different situations. In (Bolón-Canedo et al., 2013d) a review of 11 classical feature selection methods were applied over 11 synthetic and 2 real datasets was presented. The main objective of this work is to provide the user some recommendations about which feature selection method is the most appropriate under a given type of data.

The suite of synthetic datasets chosen covers phenomena such as the presence of irrelevant and redundant features, noise in the data or interaction between attributes. A scenario with a small ratio between number of samples and features where most of the features are irrelevant was also tested. It reflects the problematic of datasets such as microarray data, a well-known and hard challenge in the machine learning field where feature selection becomes indispensable.

Within the feature selection field, three major approaches were evaluated: filters (correlation-based feature selection – CFS, consistency-based, INTERACT, Information Gain, ReliefF, minimum Redundancy Maximum Relevance – mRMR and M_d), wrappers (with a support vector machine – SVM and a C4.5 tree) and embedded methods (SVM recursive feature elimination – SVM-RFE and feature selection perceptron – FS-P). To test the effectiveness of the studied methods, an evaluation measure was introduced trying to reward the selection of the relevant features and to penalize the inclusion of the irrelevant ones. Besides, four classifiers were selected (C4.5, SVM, IB1 and naive Bayes) to measure the effectiveness of the selected features and to check if the true model was also unique.

Table 1 shows the behavior of the different feature selection methods over the different problems studied, where the larger the number of dots, the better the behavior. To decide which methods were the most suitable under a given situation, it was computed a trade-off between the proposed index of success and the classification accuracy. In light of these results, ReliefF turned out to be the best option independently of the particulars of the data, with the added benefit that it is a filter, which is the model with the lowest computational cost. However, SVM-RFE with a non-linear kernel showed outstanding results, although its computational time is in some cases prohibitive (in fact, it could not be applied over some datasets). Wrappers have proven to be an interesting choice in some domains, nevertheless they must be applied together with their own classifiers and it has to be reminded that this is the model with the highest computational cost. In addition to this, Table 1 provides some guidelines for specific problems.

Table 1: Summary.

Method	Correlation & redundancy	Non Linearity	Noise Inputs	Noise Target	No. feat >> No. samples
CFS	•	•	•	•••	••••
Consistency	•	•	•	•••	••
INTERACT	•	•	•	•••	•••
InfoGain	•	•	•	•••	•••
ReliefF	••••	•••••	•••••	•••••	••
mRMR	••••	•••	•••••	••	•
M_d	••••	••	•••	•••	•••
SVM-RFE	••••	•	•	••••	•••••
SVM-RFEnl	••••	•••••	•••	••••	–
FS-P	•••••	••	••••	••••	•
Wrapper SVM	•	•	•••	••••	••
Wrapper C4.5	••	•••	•••	•••	•••

The feature selection methods were also tested over two real datasets, demonstrating the conclusions extracted from this theoretical study over real scenarios, and proving the effectiveness of feature selection. A preliminary study on this topic was published in (Bolón-Canedo et al., 2011c).

5.2 Application of Feature Selection to Real Problems

After reviewing the behavior of the most famous feature selection methods over synthetic datasets, it is necessary to prove their benefits on real problems. This section will present real applications of this discipline, reporting success in different domains such as classification of the tear film lipid layer and the K-complex classification.

In (Bolón-Canedo et al., 2012; Remeseiro et al., 2013) a fast and automatic tool is presented to classify the tear film lipid layer. The time required by previous approaches prevented their clinical use because it was too long to allow the software tool to work in real time. To solve this problem, feature selection plays a crucial role since it reduces the number of input features and, consequently, the processing time. Three of the most popular feature selection methods were chosen for this research: CFS, Consistency-based and INTERACT. Those methods were tested over the fea-

tures extracted from the images using co-occurence features, a popular texture analysis method, in the Lab colour space. Results showed that the CFS filter surpass previous results in terms of processing time whilst maintaining classification accuracy. In clinical terms, the manual process done by experts can be now automatized with the benefits of being faster, with maximum accuracy over 96% and with a processing time under 1 second. The clinical significance of these results should be highlighted, as the agreement between subjective observers is between 91%-100%. Thus, it is completely recommended the use of this application for clinical purposes as a supporting tool to diagnose evaporative dry eye.

The second real scenario was the K-complex classification (Hernández-Pereira et al., 2014), a key aspect in sleep studies. The same three filter methods were applied combined with five different machine learning algorithms, trying to achieve a low false positive rate whilst maintaining the accuracy. When feature selection was applied, the results improved significantly for all the classifiers. It is remarkable the 91.40% of classification accuracy obtained by CFS, reducing in 64% the number of features.

Notice that both problems are within the medical field, and in both cases the experts can take advantage of the feature selection. Not only are they benefited from improvements in classification accuracy, but also from the model simplification, leading in some cases to a better understanding of it.

5.3 Feature Selection on DNA Microarray Classification

Among the different problems which have been brought with the explosion of high-dimensional data, one of the most important and studied is the analysis of DNA microarray data. The key point to understand all the attention devoted to this field is the challenge that their problematic poses. Besides the obvious disadvantage of having so much features for such a small number of samples, researchers have to deal also with classes which are very unbalanced, training and test datasets extracted under different conditions, dataset shift or the presence of outliers. This is the reason because new methods emerge every year, not only trying to improve previous results in terms of classification accuracy, but also aiming to help biologists to identify the underlying mechanism that relates gene expression to diseases.

The research presented in (Bolón-Canedo et al., 2013) reviews the up-to-date contributions of feature selection research applied to DNA microarray analysis, as well as the datasets used. Since the infancy of microarray data classification, feature selection became an imperative step, in order to reduce the number of features (genes).

Since the end of the nineties, when microarray datasets began to be dealt with, a large number of feature selection methods were applied. In the literature one can find both classical methods and methods developed especially for this kind of data. Due to the high computational resources that these datasets demand, wrapper and embedded methods have been mostly avoided, in favor of less expensive approaches such as filters.

The recent literature has been analyzed in order to give the reader a brushstroke about the tendency in developing feature selection methods for microarray data. Furthermore, a summary of the datasets used in the last years is provided. In order to have a complete picture on the topic, we have also mentioned the most common validation techniques. Since there is no consensus in the literature about this issue, we have provided some guidelines.

Finally, a framework for feature selection evaluation in microarray datasets has been proposed and a practical evaluation where the results obtained are analyzed. This experimental study tries to show in practice the problematics that have been explained in theory. For this sake, a suite of 9 widely-used binary datasets was chosen to apply over them 7 classical feature selection methods. For obtaining the final classification accuracy, 3 well-known classifiers were used. This large set of experiments aims also at facilitating future comparative studies when a researcher proposes a new method.

Regarding the opportunities for future feature selection research, the tendency is toward focusing on new combinations such as hybrid or ensemble methods. This type of methods are able to enhance the stability of the final subset of selected features, which is also a trending topic in this domain. Another interesting line of future research might be to distribute the microarray data vertically (i.e. by features) in order to reduce the heavy computational burden when applying wrapper methods.

5.4 Scalability of Feature Selection Methods

When dealing with the performance of machine learning algorithms, most papers are focused on the accuracy obtained by the algorithm. However, with the advent of high dimensionality problems, researchers must study not only accuracy but also scalability. Aiming at dealing with a problem as large as possible, feature selection can be helpful as it reduces the input

dimensionality and therefore the run-time required by an algorithm.

In (Bolón-Canedo et al., 2011a; Peteiro-Barral et al., 2013) the effectiveness of feature selection on the scalability of training algorithms for artificial neural networks (ANNs) was evaluated, both for classification and regression tasks. Since there are no standard measures of scalability, those defined in the PASCAL Large Scale Learning Challenge (Sonnenburg et al., 2009) were used to assess the scalability of the algorithms in terms of error, computational effort, allocated memory and training time. Results showed that feature selection as a preprocessing step is beneficial for the scalability of ANNs, even allowing certain algorithms to be able to train on some datasets in cases where it was impossible due to the spatial complexity. Moreover, some conclusions about the adequacy of the different feature selection methods over this problem were extracted.

The next step was to evaluate the scalability of the feature selection methods without the influence of machine learning methods. An algorithm is said to be scalable if it is suitable, efficient and practical when applied to large datasets. However, the current state is that the issue of scalability is far from being solved although is present in a diverse set of problems. Research on this topic has been collected in (Peteiro-Barral et al., 2012; Bolón-Canedo et al., 2013; Rego-Fernández et al., 2014). An analysis of the scalability of feature selection methods, which has not received much consideration in the literature, has been presented. Eight well-known filter-based feature selection algorithms were evaluated, covering both ranking and subset methods. A suite of ten artificial datasets was chosen, so as to be able to assess the degree of closeness to the optimal solution in a confident way. For determining the scalability of the methods, several new measures are proposed, based not only in accuracy but also in execution time and stability, and their adequacy was demonstrated. In light of the experimental results, the fast correlation-based filter (FCBF) seems to be the most scalable subset filter. As for the ranker methods, ReliefF is a good choice when having a small number of features (up to 128) at the expense of a long training time. For this reason, when dealing with extremely-high datasets, Information Gain demonstrated better scalability properties.

5.5 Combination of Discretization and Filter

The KDD (Knowledge Discovery and Data Mining Tools Conference) Cup 99 dataset is a well-known benchmark dataset with 5 million samples and 41 features, which can be regarded as a multiclass or binary problem. Since some of its features are very unbalanced, discretization is necessary prior to feature selection. A method based on the combination of discretization, filtering and classification methods that maintains the performance results using a reduced set of features is published in (Bolón-Canedo et al., 2011b). The results obtained in the binary approach (Bolón-Canedo et al., 2009; Bolón-Canedo et al., 2010b) outperformed the KDD Cup 99 competition winner result in performance, while using only 17% of the total number of features. Also, the KDD Cup 99 dataset has been studied as a multiple class problem, distinguishing among normal connections and four types of attacks. Multiple class problems can be dealt with by means of two different approaches: using a multiple class algorithm and using multiple binary classifiers. Both approaches were tested, but with the first one, none of the results achieved improved those obtained by the KDD winner.

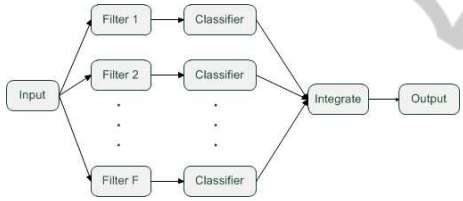
For the multiple binary classifiers approach, two class binarization techniques were utilized, namely *One vs Rest* and *One vs One*. One of the results obtained by *One vs Rest* and eight of the results obtained by *One vs One* got a better score than the KDD winner, so if the results of this research were in the original contest, the KDD winner will be the tenth entry of the competition. It is specially important the result obtained with the *One vs One* approach, combining the Proportional k-Interval Discretization (PKID) discretizer, the Consistency-based filter, the C4.5 classifier and the Accumulative Sum decoding technique. This result achieved a score of 0.2132 with only a third of the input features, that improves the KDD winner score in 0.0199. It is necessary to bear in mind that the difference between the winner and the second was only 0.0025.

This combination of discretizator and filter was also applied successfully to several multiclass problems (Sánchez-Maroño et al., 2010) and to gene selection of microarray data (Bolón-Canedo et al., 2010a; Porto-Díaz et al., 2011). Remind that DNA microarray data is a hard challenge for machine learning researchers due to the high number of features (around 10 000) but small sample size (typically one hundred or less).

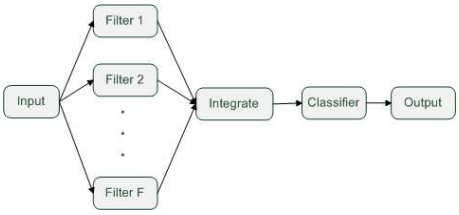
5.6 An Ensemble of Filters and Classifiers

There is a vast body of feature selection methods in the literature, based on different metrics, and to choose the adequate method for each scenario is not an easy-to-solve question. The proliferation of feature

selection algorithms has not brought about a general methodology that allows for intelligent selection from existing algorithms. For a specific dataset, employing one or another feature selection method varies the selected subset of features and, consequently, the performance result obtained by a machine learning algorithm. In order to reduce the variability associated to feature selection, an ensemble of filters has been proposed and published in (Bolón-Canedo et al., 2012) in order to obtain good performance independently on the dataset. The idea was to combine several filters, employing different metrics and performing a feature reduction. Each filter selects a subset of features and this subset is used for training the classifier. There will be as many outputs as filters employed and the result of the filters and classifier will be combined using simple voting (see Figure 2(a)). The experimental results on DNA microarray data showed that, although in some specific cases there is a filter that performs better than the ensemble, there is not a better filter in general, and the ensemble seems to be the most reliable alternative when a feature selection process has to be carried out.



(a) Ensemble 1



(b) Ensemble 2

Figure 2: Implementations of the ensemble.

Then, the previous study was extended with another approach (Bolón-Canedo et al., 2011d; Bolón-Canedo et al., 2013a), changing the role of the classifier. Ensemble1 (see Figure 2(a)) classifies as many times as there are filters, whereas Ensemble2 (see Figure 2(b)) classifies only once with the result of joining the different subsets selected by the filters. For Ensemble1, two methods for combining the outputs of the classifiers were studied, as well as the possibility of using an adequate specific classifier for each filter. A total of five different implementations of the two approaches of ensemble were proposed: *E1-sv*, which is Ensemble1 using simple voting as combination method; *E1-cp*, which is Ensemble1 using cumulative probabilities as combination method; *E1-ni*, which is Ensemble1 with specific classifiers naive Bayes and IB1; *E1-ns*, which is Ensemble1 with specific classifiers naive Bayes and SVM and *E2*, which is Ensemble2.

Results over synthetic data showed the adequacy of the proposed methods on this controlled scenario since they selected the correct features. The next step was to apply these approaches to 5 UCI classical datasets. Experimental results demonstrated that one of the ensembles (E1-cp) combined with C4.5 classifier was the best option when dealing with this type of dataset. Finally, the ensemble configurations were tested over 7 DNA microarray data. As expected, using an ensemble was again the best option. Specifically, the best performance was achieved again with E1-cp but this time combined with SVM classifier. It should be noted that some of these datasets presented a high imbalance of the data. To overcome this problem, an oversampling method was applied after the feature selection process. The result was that once again one of the ensembles achieved the best performance, and that this was even better than the one obtained with no preprocessing, showing the adequacy of the ensemble combined with over-sampling methods. Thus, the appropriateness of using an ensemble instead of a single filter remained demonstrated, considering that for all scenarios tested, the ensemble was always the more successful solution.

Regarding the different implementations of the ensemble tested, several conclusions can be drawn. There is a slight difference between the two combiner methods employed with Ensemble1 (simple voting and cumulative probability), although the second one obtained the best performance. Among the different classifiers chosen for this study, it appeared that the type of data to be classified determines significantly the error achieved, so it is responsibility of the user to know which classifier is more suitable for a given type of data. The authors recommend using E1-cp with C4.5 when classifying classical datasets (with more samples than features) and E1-cp with SVM when dealing with microarray dataset (with more features than samples). In complete ignorance of the particulars of the data, we suggest using E1-ns, which releases the user from the task of choosing a specific classifier.

5.7 Cost-based Feature Selection

There is a broad suite of filter methods, based on different metrics, but the most common approaches are to find either a subset of features that maximizes a

given metric or either an ordered ranking of the features based on this metric. However, there are some situations where a user is not only interested in maximizing the merit of a subset of features, but also in reducing costs that may be associated to features. For example, for medical diagnosis, symptoms observed with the naked eye are costless, but each diagnostic value extracted by a clinical test is associated with its own cost and risk. In other fields, such as image analysis, the computational expense of features refers to the time and space complexities of the feature acquisition process. This is a critical issue, specifically in real-time applications, where the computational time required to deal with one or another feature is crucial (see Section 5.2), and also in the medical domain, where it is important to save economic costs and to also improve the comfort of a patient by preventing risky or unpleasant clinical tests (variables that can be also treated as costs).

In (Bolón-Canedo et al., 2014) a new framework for cost-based feature selection is proposed. The objective is to solve problems where not only it is interesting to minimize the classification error, but also reducing costs that may be associated to input features. This framework consists of adding a new term to the evaluation function of any filter feature selection method so that it is possible to reach a trade-off between a filter metric (e.g. correlation or mutual information) and the cost associated to the selected features. A new parameter, called λ, is introduced in order to adjust the influence of the cost into the evaluation function, allowing the user fine control of the process according to his needs.

In order to test the adequacy of the proposed framework, two well-known and representative filters are chosen: CFS (belonging to the subset feature selection methods) and mRMR (belonging to the ranker feature selection methods). Experimentation is executed over a broad suite of different datasets. Results after performing classification with a SVM display that the approach is sound and allow the user to reduce the cost without compromising the classification error significantly, which can be very useful in fields such as medical diagnosis or real-time applications.

Then, in (Bolón-Canedo et al., 2014), a modification of the ReliefF filter for cost-based feature selection, called mC-ReliefF, is proposed. Twelve different datasets, covering very diverse situations, were selected to test the approach. Results after performing classification with a SVM and Kruskal-Wallis statistical tests, again demonstrated the adequacy of the cost-based feature selection. Finally, the method was applied to the real problem presented in Section 5.2: the tear film lipid layer classification. In this scenario the time required to extract the features prevented clinical use because it was too long to allow the software tool to work in real time. mC-ReliefF permits to automatically decrease the required time (from 38 seconds to less than 1 second, that is in 92%) while maintaining the classification performance. Notice that this reduction in time is very important since interviews with optometrists revealed that a scale of computation time over 10 seconds per image makes the system not usable.

5.8 Distributed Feature Selection

Traditionally, feature selection methods are applied in a centralized manner, i.e. a single learning model to solve a given problem. However, when dealing with large amounts of data, distributed feature selection seems to be a promising line of research since allocating the learning process among several workstations is a natural way of scaling up learning algorithms. Moreover, it allows to deal with datasets that are naturally distributed, a frequent situation in many real applications (e.g. weather databases, financial data or medical records). There are two common types of data distribution: (a) horizontal distribution wherein data are distributed in subsets of instances; and (b) vertical distribution wherein data are distributed in subsets of attributes.

The great majority of approaches distribute the data horizontally, since it constitutes the most suitable and natural approach for most applications. In (Bolón-Canedo et al., 2013), a methodology is proposed which consists of applying filters over several partitions of the data, combined in the final step into a single subset of features. The idea of distributing the data horizontally builds on the assumption that combining the output of multiple experts is better than the output of any single expert. There are three main stages: (i) partition of the datasets; (ii) application of the filter to the subsets; and (iii) combination of the results. An experimental study was carried out on six datasets considered representative of problems from medium to large size. In terms of classification accuracy, our distributed filtering approach obtains similar results to the centralized methods, even with slight improvements for some datasets. Furthermore, the most important advantage of the proposed method is the dramatically reduction in computational time (from the order of hours to the order of minutes).

While not common, there are some other developments that distribute the data by features. In (Bolón-Canedo et al., 2013b) the data are distributed vertically in order to have the feature selection process distributed. This approach is especially suitable for mi-

croarray data since in this manner we will deal with subsets with a more balanced features/samples ratio and avoid overfitting problems.

The partition of the dataset consists of dividing the original dataset into several disjoint subsets of approximately the same size that cover the full dataset. Two different methods were used for partitioning the data: (a) performing a randomly partition and (b) ranking the original features before generating the subsets. The second option was introduced trying to improve the performance results obtained by the first one. By having an ordered ranking, features with similar relevance to the class will be in the same subset, which will facilitate the task of the subset filter which will be applied later. These two techniques for partitioning the data will generate two different approaches for the distributed method: Distributed Filter (DF) with the randomly partition and Distributed Ranking Filter (DRF) associated to the ranking partition.

After this step, the data is split by assigning groups of k features to each subset, where the number of features k in each subset is half the number of samples, to avoid overfitting. When opting for the randomly partition (DF), the groups of k features are constructed randomly, having into account that the subsets have to be disjoint. In the case of the ranking partition (DRF), the groups of k features are generated sequentially over the ranking, so features with a similar ranking position will be in the same group. Notice that the random partition is equivalent to obtain a random ranking of the features and then follow the same steps as with the ordered ranking. Figure 3 shows a flow chart which reflects the two algorithms proposed, DF and DRF. After having several small disjoint datasets D_i, the filter method will be applied to each of them, returning a selection S_i for each subset of data. Finally, to combine the results, a merging procedure using a classifier will be executed.

The experiments on eight microarray datasets showed that this proposal was able to reduce the running time significantly with respect to the standard (centralized) filtering algorithms. In terms of execution time, the behavior is excellent, being this fact the most important advantage of our method. Furthermore, with regard of classification accuracy, our distributed approach was able to match and in some cases even improve the standard algorithms applied to the non-partitioned datasets. This situation is reflected in Figure 4, where the best result among all the classifiers is displayed for any dataset and the consistency-based filter. It is easy to see at a glance that the accuracies fall into similar values (being most of the time a distributed approach the best option)

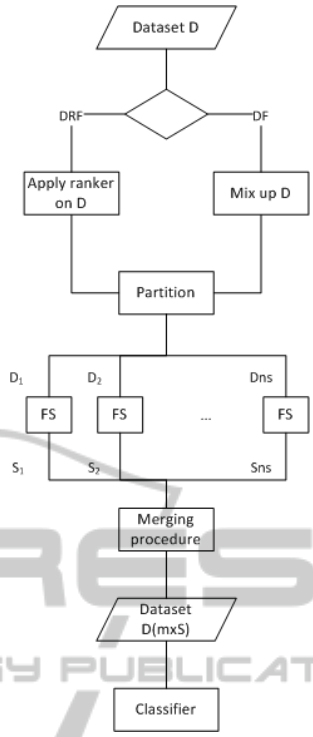


Figure 3: Flow chart of distributed filter approach.

whilst the differences in time are outstanding.

Finally, in (Bolón-Canedo et al., 2013c), the adequacy of a distributed approach for wrapper feature selection was tested over four datasets considered representative of problems from medium to large size. The goal was to design a distributed wrapper which would led to a reduction in the running time as well as in the storage requirements while the accuracy would not drop to inadmissible values. Again, the experiments showed that our method was able to shorten the execution time impressively compared to the standard wrapper algorithms. Furthermore, our distributed wrapper achieved a similar performance to the original wrapper. In terms of test accuracy, the proposed distributed wrapper is able to match and in some cases even to improve the standard results applied to the non-partitioned datasets.

6 CONCLUSIONS

Continual advances in computer-based technologies have enabled researchers and engineers to collect data at an increasingly fast pace. To address this challenge, feature selection becomes an imperative preprocessing step which needs to be adapted and improved to handle high-dimensional data.

This work is devoted to study feature selection and

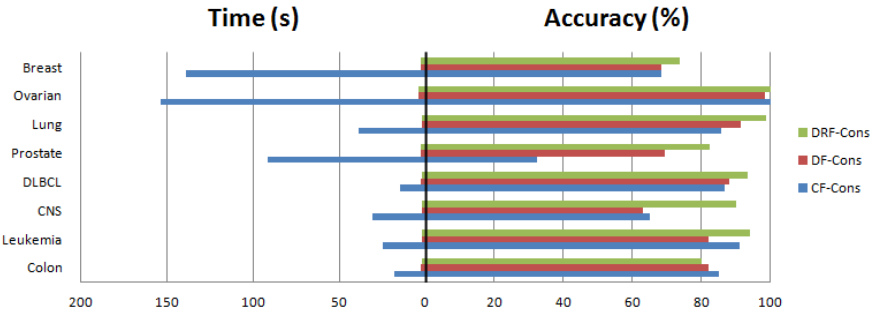


Figure 4: Comparison of accuracy and time for consistency-based filter.

its adequacy to large-scale data. The tendency nowadays is two-fold: on the one hand, to improve and extend the existing methods to address the new challenges associated to high-dimensionality. And, on the other hand, to develop novel techniques to directly solving the arising challenges.

First, a critical analysis of existing feature selection was performed, to check their adequacy toward different challenges and be able to provide some recommendations. Bearing this analysis in mind, the most adequate techniques were applied to several real-life problems, obtaining a notable improvement in performance. Apart from efficiency, another critical issue in large-scale applications which is scalability. The effectiveness of feature selection methods may be significantly downgraded, if not totally inapplicable, when the data size increases steadily. For this reason, a stability analysis in detail of the most famous techniques was done.

Then, new techniques for large-scale feature selection were proposed. In the first place, as most of the existing feature selection techniques need data to be discrete, a new approach was proposed that consists in a combination of a discretizer, a filter method and a very simple classical classifier, obtaining promising results. Another proposal was to employ a ensemble of filters instead of a single one, releasing the user from the decision of which technique is the most appropriate for a given problem. An interesting topic is also to consider the cost related with the different features, therefore a framework for cost-based feature selection was proposed, demonstrating its adequacy in a real-life scenario. Finally, it is well-known that a manner of handling large-scale data is to transform the large-scale problem into several small-scale problems, by distributing the data. With this aim, several approaches for distributed and parallel feature selection have been proposed.

As can be seen, this thesis covers a broad suite of problems arisen from the advent of high-dimensionality. The proposed approaches have demonstrated to be sound, and it is expected that their contribution will be important in the next years, since feature selection for large-scale data is likely to continue to be a trending topic in the near future.

ACKNOWLEDGEMENTS

This research has been partially funded by the Secretaría de Estado de Investigación of the Spanish Government and FEDER funds of the European Union through the research projects TIN 2012-37954 and PI10/00578; and by the Consellería de Industria of the Xunta de Galicia through the research projects CN2011/007 and CN2012/211. Veronica Bolon-Canedo acknowledges the support of Xunta de Galicia under *Plan I2C* Grant Program.

REFERENCES

Bolón-Canedo, V., Peteiro-Barral, D., Alonso-Betanzos, A., Guijarro-Berdiñas, B., and Sánchez-Maroño, N. (2011a). Scalability analysis of ann training algorithms with feature selection. In *Advances in Artificial Intelligence*, pages 84–93. Springer.

Bolón-Canedo, V., Peteiro-Barral, D., Remeseiro, B., Alonso-Betanzos, A., Guijarro-Berdinas, B., Mosquera, A., Penedo, M. G., and Sánchez-Maroño, N. (2012). Interferential tear film lipid layer classification: an automatic dry eye test. In *Tools with Artificial Intelligence (ICTAI), 2012 IEEE 24th International Conference on*, volume 1, pages 359–366. IEEE.

Bolón-Canedo, V., Porto-Díaz, I., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2014). A framework for cost-based feature selection. *Pattern Recognition (In press)*.

Bolón-Canedo, V., Rego-Fernández, D., Peteiro-Barral, D., Alonso-Betanzos, A., Guijarro-Berdinas, B., and Sánchez-Maroño, N. (2013). On the scalability of filter techniques for feature selection on big data. *IEEE Computational Intelligence Magazine Special Issue on Computational Intelligence in Big Data (Under Review)*.

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2009). A combination of discretization and filter methods for improving classification performance in kdd cup 99 dataset. In *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*, pages 359–366. IEEE.

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2010a). On the effectiveness of discretization on gene selection of microarray data. In *International Joint Conference on Neural Networks. IJCNN 2010*, pages 3167–3174. IEEE.

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2011b). Feature selection and classification in multiple class datasets: An application to kdd cup 99 dataset. *Expert Systems with Applications*, 38(5):5947–5957.

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2011c). On the behavior of feature selection methods dealing with noise and relevance over synthetic scenarios. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 1530–1537. IEEE.

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2011d). Toward an ensemble of filters for classification. In *Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on*, pages 331–336. IEEE.

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2012). An ensemble of filters and classifiers for microarray data classification. *Pattern Recognition*, 45(1):531–539.

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2013a). Data classification using an ensemble of filters. *Neurocomputing (In Press)*.

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2013b). A distributed filter approach for microarray data classification. *Applied Soft Computing (Under Review)*.

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2013c). A distributed wrapper approach for feature selection. In *European Symposium on Artificial Neural Networks, ESANN 2013*, pages 173–178.

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2013d). A review of feature selection methods on synthetic data. *Knowledge and information systems*, 34(3):483–519.

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2014). mc-relieff: An extension of relieff for cost-based feature selection. In *6th International Conference on Agents and Artificial Intelligence (ICAART) (Accepted)*.

Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., Benítez, J., and Herrera, F. (2013). An insight into microarray datasets and feature selection methods: a framework for ongoing studies. *Information Sciences (Under review)*.

Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., and Hernandez-Pereira, E. (2010b). Feature selection and conversion methods in KDD Cup 99 dataset: A comparison of performance. In *Proceedings of the 10th IASTED International Conference*, pages 58–66.

Bolón-Canedo, V., Sánchez-Maroño, N., and Cerviño-Rabuñal, J. (2013). Scaling up feature selection: A distributed filter approach. In *Advances in Artificial Intelligence*, pages 121–130. Springer.

Chidlovskii, B. and Lecerf, L. (2008). Scalable feature selection for multi-class problems. In *Machine Learning and Knowledge Discovery in Databases*, pages 227–240. Springer.

Dash, M. and Liu, H. (1997). Feature selection for classification. *Intelligent data analysis*, 1(3):131–156.

Frank, A. and Asuncion, A. (2010). UCI Machine Learning Repository. http://archive.ics.uci.edu/ml. [Online; accessed December-2013].

Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. (2006). *Feature extraction: foundations and applications*, volume 207. Springer.

Hernández-Pereira, E., Bolón-Canedo, V., Sánchez-Maroño, N., Álvarez-Estévez, D., Moret-Bonillo, V., and Alonso-Betanzos, A. (2014). A comparison of performance of k-complex classification methods using feature selection. *Information Sciences (Under review)*.

Hughes, G. (1968). On the mean accuracy of statistical pattern recognizers. *Information Theory, IEEE Transactions on*, 14(1):55–63.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324.

Liu, H. and Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *Knowledge and Data Engineering, IEEE Transactions on*, 17(4):491–502.

Loscalzo, S., Yu, L., and Ding, C. (2009). Consensus group stable feature selection. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 567–576. ACM.

Peng, Y., Wu, Z., and Jiang, J. (2010). A novel feature selection approach for biomedical data classification. *Journal of Biomedical Informatics*, 43(1):15–23.

Peteiro-Barral, D., Bolon-Canedo, V., Alonso-Betanzos, A., Guijarro-Berdinas, B., and Sánchez-Maroño, N. (2012). Scalability analysis of filter-based methods for feature selection. *Advances in Smart Systems Research*, 2(1):21–26.

Peteiro-Barral, D., Bolón-Canedo, V., Alonso-Betanzos, A., Guijarro-Berdiñas, B., and Sánchez-Maroño, N. (2013). Toward the scalability of neural networks through feature selection. *Expert Systems with Applications*, 40(8):2807–2816.

Porto-Díaz, I., Bolón-Canedo, V., Alonso-Betanzos, A., and Fontenla-Romero, O. (2011). A study of performance on microarray data sets for a classifier based on information theoretic learning. *Neural Networks*, 24(8):888–896.

Provost, F. (2000). Distributed data mining: Scaling up and beyond. *Advances in distributed and parallel knowledge discovery*, pages 3–27.

Rego-Fernández, D., Bolón-Canedo, V., and Alonso-Betanzos, A. (2014). Scalability analysis of mrmr for microarray data. In *6th International Conference*

on Agents and Artificial Intelligence (ICAART) (Accepted).

Remeseiro, B., Bolón-Canedo, V., Peteiro-Barral, D., Alonso-Betanzos, A., Guijarro-Berdinas, B., Mosquera, A., Penedo, M. G., and Sánchez-Maroño, N. (2013). A methodology for improving tear film lipid layer classification. *IEEE Journal of Biomedical and Health Informatics (In Press)*.

Saeys, Y., Abeel, T., and Van de Peer, Y. (2008). Robust feature selection using ensemble feature selection techniques. In *Machine Learning and Knowledge Discovery in Databases*, pages 313–325. Springer.

Sánchez-Maroño, N., Alonso-Betanzos, A., García-González, P., and Bolón-Canedo, V. (2010). Multiclass classifiers vs multiple binary classifiers using filters for feature selection. In *International Joint Conference on Neural Networks. IJCNN 2010*, pages 2836–2843. IEEE.

Sonnenburg, S., Franc, V., Yom-Tov, E., and Sebag, M. (2009). PASCAL Large Scale Learning Challenge. *Journal of Machine Learning Research*.

Spark (n.d.). Apache Spark - Lightning-Fast Cluster Computing. http://spark.incubator.apache.org. [Online; accessed December-2013].

Sun, Y. and Li, J. (2006). Iterative relief for feature weighting. In *Proceedings of the 23rd international conference on Machine learning*, pages 913–920. ACM.

Sun, Y., Todorovic, S., and Goodison, S. (2008). A feature selection algorithm capable of handling extremely large data dimensionality. In *SDM*, pages 530–540.

Tuv, E., Borisov, A., Runger, G., and Torkkola, K. (2009). Feature selection with ensembles, artificial variables, and redundancy elimination. *The Journal of Machine Learning Research*, 10:1341–1366.

Vainer, I., Kraus, S., Kaminka, G. A., and Slovin, H. (2011). Obtaining scalable and accurate classification in large-scale spatio-temporal domains. *Knowledge and information systems*, 29(3):527–564.

Yu, L. and Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, 5:1205–1224.

Zhang, Y., Ding, C., and Li, T. (2008). Gene selection algorithm by combining relieff and mrmr. *BMC genomics*, 9(Supl 2):S27.

Zhao, Z. and Liu, H. (2011). *Spectral Feature Selection for Data Mining*. Chapman & Hall/Crc Data Mining and Knowledge Discovery. Taylor & Francis Group.