Prog Artif Intell (2016) 5:65–75
DOI 10.1007/s13748-015-0080-y

REGULAR PAPER

Feature selection for high-dimensional data

Verónica Bolón-Canedo[1] · Noelia Sánchez-Maroño[1] · Amparo Alonso-Betanzos[1]

Received: 1 December 2015 / Accepted: 29 December 2015 / Published online: 15 February 2016

Abstract This paper offers a comprehensive approach to feature selection in the scope of classification problems, explaining the foundations, real application problems and the challenges of feature selection in the context of high-dimensional data. First, we focus on the basis of feature selection, providing a review of its history and basic concepts. Then, we address different topics in which feature selection plays a crucial role, such as microarray data, intrusion detection, or medical applications. Finally, we delve into the open challenges that researchers in the field have to deal with if they are interested to confront the advent of "Big Data" and, more specifically, the "Big Dimensionality".

Keywords Feature selection · High-dimensional data · Big Data

1 Introduction

We are in a data-driven era, in which the size of digital data available in the world is continuously growing since data are acquired for countless purposes. Hence, machine learning algorithms have to cope with data sets whose volume and complexity are also increasing. Among the many machine learning algorithms, feature selection is characterized by electing the attributes that allow to clearly define a problem, apart from those that are irrelevant or redundant. The many existing feature selection methods need to evolve to deal with this new context. In this paper, we describe the current context of feature selection, starting with a brief definition and presenting the main categories under which existing methods are divided. Subsequently, we introduce the trending topics and the open challenges that must be considered to make feature selection indispensable in the era of Big Data.

The purpose of this introductory section was to briefly present this new term that has captured the attention of the scientific community: Big Data. The volume of the new data sets makes dimensionality reduction, and thus feature selection, a necessity. Further, we present some inherent difficulties that current data sets may have and, therefore, constitute a challenge for any machine learning technique, including feature selection.

1.1 A new scenario: Big Data

Since the late past century, enterprises have stored data to extract information in a near future, but without a clear idea of the potential usefulness of such amount of data. In addition to this, the growing popularity of the Internet has generated data in many different formats (text, multimedia, etc.) and from many different sources (systems, sensors, mobile devices, etc.). Different studies have tried to determine the size of this digital universe, i.e., these digital bits created, replicated, and consumed annually. According to one of the most recent studies [71], the digital universe is doubling in size every two years; then, by 2020, it is expected to reach 44 zettabytes (10^{21} bytes). In a more visual way, if the digital universe is represented by the memory in a stack of tablets (iPad Air 0.29$''$ thick and 128 Gb), by 2020 there would be 6,6 stacks from the Earth to the Moon.

In this context, a new concept is born: Big Data. The first documented use of the term "Big Data" appeared in 1997

✉ Verónica Bolón-Canedo
vbolon@udc.es

Noelia Sánchez-Maroño
nsanchez@udc.es

Amparo Alonso-Betanzos
ciamparo@udc.es

[1] Departamento de Computación, Universidade da Coruña, Campus de Elviña s/n, A Coruña 15071, Spain

[19] and it was referred to the area of scientific visualization where the data sets are usually very large, specifically the authors say “When data sets do not fit in main memory (in core), or when they do not fit even on local disk, the most common solution is to acquire more resources.” This first definition of Big Data refers to one of its main characteristic: Volume. However, this concept rapidly expands by including two more properties: velocity and variety [41]. The former refers to the speed of data creation, streaming, and aggregation and the latter is a measure of the richness of the data representation (text, images video, audio, etc). These three properties were called the 3 V’s that define Big Data. Another “V” property was included afterwards: Value. Nowadays, value is considered the most important V because it refers to the process of discovering huge hidden values from large data sets with various types and rapid generation [32].

Big Data can also be classified into different categories to better understand their characteristics. This classification is based on five aspects: (i) data sources, (ii) content format, (iii) data stores, (iv) data staging and (v) data processing [32]. Machine learning algorithms usually cope with data staging (cleaning, transforming and normalizing data) and data processing (batch or real time). However, machine learning is still in its early stages of development [16]. Many algorithms do not scale beyond data sets of a few million elements found in real-world data, i.e., they cannot deal with Big Data. Then, further research is required to develop algorithms that apply in real-world situations and on data sets of trillions of elements. The automated or semi-automated analysis of enormous volumes of data lies at the heart of big-data computing for all application domains.

1.2 Why is feature selection important?

The size of a data set can be measured in two dimensions, number of samples/instances (n) and number of features/attributes (m). Both m and n can be enormously large [47]. However, researchers in the data analytics community have largely taken a one-sided study of volume, which refers to the “instance size” of the data. On the other hand, the corresponding factor of “Big Dimensionality”, i.e., the feature size, has received much lesser attention [75].

Nevertheless, in the past few years, the number of data sets with a ultra-high number of features is growing. For instance, there are 18 data sets with more than 5000 features at the UCI machine learning repository [45]. Table 1 shows the maximum number of features of the data sets posted in the UCI repository since 2008.

Other popular repositories, such as the LIBSVM Database [18], include data sets with more than 29 million features (KDD Cup 2010 dataset). In particular, seven of the data sets posted there have more than 1 million features.

Table 1 Maximum number of features of the data sets posted in the UCI repository [45] since 2008

Year	Name	No. of features
2008	Bag of Words and Dorothea	100, 000
2009	URL Reputation	3, 231, 961
2010	p53 mutants	5409
2011	PEMS-SF	138, 672
2012	CNAE-9	857
2013	Gas sensor open	1, 950, 000
2014	Gas sensor flow	150, 000
2015	Electricity load diagrams	140, 256

A specific problem arises when these data sets, instead of being large in both dimensions, have a number of features much larger than the number of samples, hindering the posterior learning process. The best known example is microarray data sets [14]; this type of data usually has very small samples (often less than 100), whereas the number of features ranges from 6000 to 60,000, since it measures a gene expression. Due to the complex problem addressed, such data sets have been widely studied in the literature [8]. Moreover, there are data sets with similar characteristics in other areas such as image classification, face recognition and text classification [7]. In fact, data sets with large ratio features/samples are constantly appearing, see for example, the recent data sets about electricity load (140,256 features/370 samples) or gas sensor array (120,432 features/58 samples) at UCI machine learning repository [45]. Under this situation, feature selection methods whose objective is to select a minimal subset of features according to some reasonable criteria become indispensable to achieve a simpler data set. Then, this reduced data set is a better representative of the whole population, and hence it may lead to more concise results and better comprehensibility [47]. For that reason, new feature selection methods continue to emerge and the importance of these techniques has not stopped growing.

2 Problems that feature selection has to deal with

In the previous section, we have pointed out the importance of feature selection. In this section, we will present the inherent difficulties that data may have due to the different methods of acquisition. For the sake of brevity, we have not focused on all issues; some not covered issues such as outliers, data complexity or reduced sample size can be found in our book [13].

2.1 Class imbalance

Data are said to suffer the Class imbalance problem when the class distributions are highly imbalanced. This occurs when a data set is dominated by a major class or classes which have significantly more instances than the other rare/minority classes in the data. Often the minority class is very infrequent, such as 1 % of the data set. If one applies most traditional classifiers on the data set, they are likely to predict everything as the majority class [46]. However, typically, people are more interested in learning rare classes. For example, applications such as medical diagnosis prediction of rare but important diseases, such as cancer. Similar situations are observed in other areas, such as detecting fraud in banking operations, detecting network intrusions, anomaly detection and so on [24]. Throughout the past years, in the machine learning community, many solutions have been proposed to deal with this problem and they can be categorized into three major groups : (i) data sampling; (ii) algorithm modification and (iii) cost-sensitive learning [49].

Similarly to classic algorithms, learning algorithms adapted to this kind of problem can benefit from an adequate selection of features. Therefore, it is important to select features that can capture the high skew in the class distribution.

2.2 Data set shift

In real environments, samples may be collected under different conditions. For example, in an image classification task, there may be differences in lighting conditions. This sometimes makes the data set used for training the machine learning method differ significantly from the test set. Many machine learning competitions try to reflect this real situation. For example, in the KDDCup 99 competition task [39], the aim was to build a network intrusion detector, a predictive model capable of distinguishing between intrusions or attacks and "good" normal connections. The test set provided at this competition includes specific attack types not present in the training data to illustrate a realistic situation where new attacks continuously appear.

These data sets can suffer what is known as data set shift, which is defined as "a challenging situation where the joint distribution of inputs and outputs differs between the training and test stages" [58]. Moreno et al. [51] state that there is no standard term to refer to this situation; therefore, there are numerous terms in the bibliography, such as "concept shift" or "concept drift", "changes of classification", "changing environments", etc. These authors suggest to maintain the term "data set shift" for "any situation in which training and test data follow distributions that are in some way different".

Numerous alternatives to address the "data set shift" problem can be found in [51,58]. However, it has to be noticed that this problem may hinder the process of feature selection and classification.

2.3 Incremental learning

Traditionally, machine learning consists of attempting to learn concepts from a static data set. This data set is, therefore, assumed to contain all information necessary to learn the relevant concepts. This model, however, has proven unrealistic for many real-world scenarios where the data flow continuously or come in separated batches over time [34], such as financial analysis, climate data analysis, bank fraud protection, traffic monitoring, predictive customer behavior, etc.

Learning under such conditions is known as incremental learning. According to the definition by Muhlbaier et al. [52], a learning algorithm is incremental if, for a sequence of training instances (potentially batches of instances), it satisfies the following criteria:

- It produces a sequence of hypotheses such that the current hypothesis describes all data seen thus far.
- It only depends on the current training data and a limited number of previous hypotheses.

Given this definition, the stability–plasticity dilemma [28] arises, i.e., it is necessary to design a learning system that remains stable and unchanged to irrelevant events (e.g., outliers), while plastic (i.e., adaptive) to new, important data (e.g., changes in concepts). Online learning algorithms are those where the system is adapted immediately upon seeing the new instance and the instance is then immediately discarded. The study of online learning algorithms is an important domain in machine learning, one that has interesting theoretical properties and practical applications [66]. This kind of learning has become a trending area in the past few years since it allows to solve important problems such as concept drift (see Sect. 2.2). This phenomenon happens when the underlying data distribution changes, and these changes make the model built on old data inconsistent with the new data, and a regular updating of the model is necessary [70]. Applied to feature selection, a concept drift may cause that the subset of relevant features changes over the time, so different sets of features become important for classification and some totally new features with high predictive power may appear.

2.4 Noisy data

It is almost inevitable that there is some noise in most of the collected data, except in the most structured and synthetic environments. This "imperfect data" can be due to many sources, for instance, faulty measuring devices, transcription errors, and transmission irregularities. However, the performance of a learning algorithm may greatly depend on the quality of the data used during the training phase, so a model built from a noisy training set might be less accurate and less compact than one built form the noise-free version of the same data set using an identical algorithm [17].

Imperfections in a data set can be dealt with in four broad ways: (i) leave the noise in, (ii) data cleaning, i.e., filter the noise out, (iii) data transformation, i.e., correct the noise and (iv) data reduction, that is, to reduce the amount of data by aggregating values or removing and clustering redundant attributes [69]. Any of these techniques has its advantages and disadvantages [13]. However, in the current context, where the data sets are so large that they become unmanageable, data reduction seems an appropriate strategy. And, among the existing techniques, dimensionality reduction—and consequently feature selection—is one of the most popular to remove noisy (i.e., irrelevant) and redundant features. However, while feature selection has been the target of many works, very little study has been done to systematically address the issue of feature relevancy in the presence of noisy data.

2.5 Budget constraints

Learning and decision-making under budget constraints in uncertain and dynamic environments have gained attention in several communities including machine learning, signal processing and information theory (see for example the recent workshops on this topic at the International Conference on Machine Learning [36]). Learning problems under budget constraints arise in a number of large-scale real-world industrial applications ranging from medical diagnosis to search engines and surveillance. In these applications budget constraints arise as a result of limits on computational cost, delay, throughput, power and monetary value. For instance, in search engines CPU cost during test-time must be budgeted and accounted for.

Learning under test-time budgets departs from the traditional machine learning setting and introduces new exciting challenges. For instance, features are accompanied by costs [9] (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) and their amortized sum is constrained at test-time. In other settings, a system must maintain a throughput constraint to keep pace with arriving traffic. All settings have in common that they introduce a new tradeoff between accuracy and cost [9]. Studying this tradeoff between cost and performance is an inherent challenge that should be investigated in a principled fashion.

3 Foundations of feature selection

As discussed in the previous section, feature selection is an area of growing interest in the field of machine learning; therefore there are many trending topics and open challenges. But before exposing them, we briefly present the definition of feature selection and the available methods for its successful application.

3.1 What is feature selection?

The ultrahigh dimensionality of actual data sets not only incurs unbearable memory requirements and high computational cost in training, but also deteriorates the generalization ability of learning algorithms because of the “curse of dimensionality” issue. This term, coined by Richard Bellman in [4], indicates the difficulty of optimization by exhaustive enumeration on product spaces. Considering that a data set can be represented by a matrix where the rows are the recorded samples and the columns are the features, to tackle the “curse of dimensionality” issue, we can find “narrower” matrices that in some sense are close to the original. Since these narrower matrices have a smaller number of features, they can be used much more efficiently than the original matrix. The process of finding these narrow matrices is called dimensionality reduction. There are two main techniques to achieve this dimensionality reduction: feature extraction and feature selection. Feature extraction consists of reducing the feature space by deriving new features transforming the existing ones; these new features are intended to be informative and non-redundant. On the other hand, feature selection (FS) is defined as the process of detecting relevant features and discarding irrelevant and redundant features with the goal of obtaining a subset of features that accurately describe a given problem with a minimum degradation of performance [29]. Both techniques are aimed at improving the performance of machine learning methods by using simpler models, probably gaining training speed. However, the main advantages of selection against extraction are [29] as follows:

- Data understanding, gaining knowledge about the process and perhaps helping to visualize it.
- Data reduction, limiting storage requirements and perhaps helping in reducing costs.

Therefore, feature selection is the elected technique in those contexts where it is important to maintain the representativeness of the problem or where the cost of acquisition and/or maintenance of the features is high such as clinical problems.

3.2 Classification of feature selection methods

From a functional point of view, FS methods can work in two different ways [44]. Some methods assign weights to each feature, in such a way that the order corresponding to their theoretical relevance is preserved. Methods that follow this approach are known as continuous, individual evaluation or ranking methods. The second set of methods are known as binary or subset evaluation methods. First, they produce candidate feature subsets using search strategies. Then, the subsets are assessed by an evaluation function which determines the final selected subset of features. Moreover, methods can be uni or multivariate, depending on whether they consider each feature independently of the rest or not.

From a structural point of view, FS methods can be classified in three major groups [29] (see Fig. 1). Filter methods perform the feature selection step as pre-processing, before the learning step. The filter is independent of the learning algorithm and relies on underlying attributes of data. Wrapper methods use the learning algorithm as a subroutine, measuring the usefulness of the features with the prediction performance of the learning algorithm over a validation set. In embedded methods, the FS process is specifically built into the machine learning method, in such a way that the search is guided by the learning process itself. Each of these approaches has its advantages and disadvantages. The main factors are the speed of computation and the probability of overfitting. Filters are faster than embedded methods, and the latter are faster than wrappers. Regarding overfitting, wrappers are more likely to overfit than embedded methods, which are more likely to overfit than filter methods. In general, filters are relatively inexpensive in terms of computational efficiency.

Given the importance of FS, numerous FS methods exist. Without being an exhaustive list, Table 2 illustrates some of the best known methods (where n is the number of samples and m is the number of features). Note that there are no wrapper methods in this table because they are formed by combining a search strategy with an induction algorithm, so there are as many as combinations of both techniques. In general, filters are relatively inexpensive in terms of computational efficiency; they are simple and fast and, therefore, most of the designed methods pertain to this category.

4 Trending topics

As mentioned before, the advent of high-dimensional data has brought unprecedented challenges to machine learning researchers, making the learning task more complex and computationally demanding. In this scenario, feature selection might play a crucial role and that is why its use as preprocessing technique has been growing in importance along the past years. In this section, we will discuss some of the topics in which the use of feature selection is in the spotlight.

4.1 Does the "best feature selection method" exist?

Feature selection has been an active and fruitful field of research in machine learning. Its importance is indisputable and it has proven effective in increasing predictive accuracy and reducing complexity of machine learning models. For this reason, new feature selection methods have been steadily appearing during the past few decades. The pro-

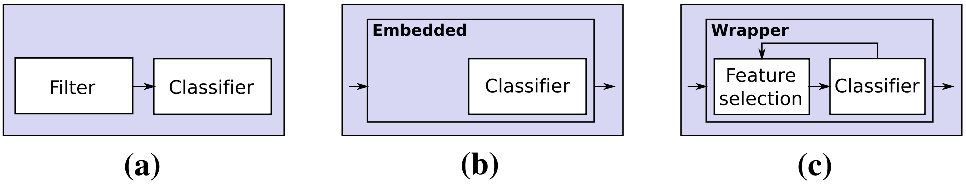

Fig. 1 Feature selection techniques. **a** Filter. **b** Embedded. **c** Wrapper

Table 2 Frequently used feature selection methods

	Uni/multivariate	Functional view	Structural view	Complexity
Chi-squared [48]	Univariate	Ranker	Filter	nm
F score (Fisher score) [21]	Univariate	Ranker	Filter	nm
Information gain [59]	Univariate	Ranker	Filter	nm
ReliefF [39]	Multivariate	Ranker	Filter	n^2m
mRMR [55]	Multivariate	Ranker	Filter	nm^2
SVM-RFE [30]	Multivariate	Ranker	Embedded	$\max(n, m)m^2$
CFS [31]	Multivariate	Subset	Filter	nm^2
FCBF [43]	Multivariate	Subset	Filter	$nmlogm$
INTERACT [77]	Multivariate	Subset	Filter	nm^2
Consistency [20]	Multivariate	Subset	Filter	nm^2

liferation of feature selection algorithms, however, has not brought about a general methodology that allows for intelligent selection from existing algorithms. There are some rules of thumb, such as choosing filters when dealing with extremely large data sets and using embedded and wrappers when the computational burden is not an issue, but even so some level of expertise is necessary to choose a specific method. In the machine learning field, it is common to deal with some factors that might affect the performance of a feature selection method, such as the proportion of irrelevant features present in the data or the interaction between attributes. Another important factor to take into account is the noise in the data. In the current scenario of Big Data, huge amounts of data are continuously generated, so it becomes more and more difficult to rely on the correctness of the provided label assignments (a problem known as label-noise [23]).

It was natural that numerous reviews of the existing feature selection methods appeared during the past few years. A common problem, however, when testing the effectiveness of feature selection methods is that it is not always possible to know the relevant features a priori, so their performance clearly rely on the performance of the learning method used afterwards and it can vary notably from one method to another. Apart from this, the performance of the different methods can be measured using many different metrics such as computer resources (memory and time), accuracy, ratio of features selected, etc. So, for example, Molina et al. [50] assessed the performance of fundamental feature selection algorithms in a controlled scenario, taking into account data set relevance, irrelevance and redundancy. Saeys et al. [63] created a basic taxonomy of classical feature selection techniques, discussing their use in bioinformatics applications. Hua et al. [35] compared some basic feature selection methods in settings involving thousands of features, using both model-based synthetic data and real data. Brown et al. [15] presented a unifying framework for information theoretic feature selection, bringing almost two decades of research into heuristic filter criteria under a single theoretical umbrella. Along the same line, Vergara and Estévez [72] presented a review of the state of the art of information-theoretic feature selection methods. Finally, García et al. [25] dedicated a chapter in their data preprocessing book to a discussion of feature selection and an analysis of its main aspects and methods.

In addition to these works, we have performed our own review [6], in which we evaluated the performance of several state-of-the-art feature selection algorithms in an artificial controlled scenario, checking their efficiency in tackling problems such as redundancy between features, nonlinearity, noise in inputs and in the class label and a higher number of features than samples (as happens with DNA microarray classification).

The conclusion of these works is that existing feature selection methods have their merits and demerits and, although some suggestions are made to help the user, there is no "best" feature selection method in general.

4.2 Can feature selection be helpful in real applications?

Feature selection may be very useful in real domains, since it allows the storage costs to decrease, the performance of a classifier to improve and a good understanding of the model to be obtained. In this section, we will comment two case studies in which we obtained promising results thanks to the application of feature selection.

4.2.1 Tear film lipid layer classification

Evaporative dry eye (EDE) is a symptomatic disease which affects a wide range of population and has a negative impact on their daily activities, such as driving or working with computers. Its diagnosis can be achieved by several clinical tests, one of which is the analysis of the interference pattern. A methodology for automatic tear film lipid layer (TFLL) classification was developed in [62], based on color texture analysis. However, the best accuracy results were obtained at the expense of a too long processing time (38 s) because many features had to be computed. This fact makes this methodology unfeasible for practical applications and prevents its clinical use. Reducing processing time is a critical issue in this application which should work in real-time to be used in the clinical routine. Therefore, we decided to apply feature selection methods in an attempt to decrease the number of features and, consequently, the computational time without compromising the classification performance.

The results of our study [61] after applying the CFS filter were able to surpass previous results in terms or processing time whilst maintaining classification accuracy. In clinical terms, the manual process done by experts can be automated with the benefits of being faster and unaffected by subjective factors, with maximum accuracy over 97 % and processing time under 1 s.

4.2.2 K-complex classification

K-complex is one of the key features that contributes to sleep stages assessment. Unfortunately, their visual identification is very time-consuming and rather dependent on the knowledge and experience of the clinician since it cannot be performed on regular basis. This is the reason why automatic identification of K-complexes is of great interest. For this reason, in [33] we presented a methodology for the automatic classification of K-complexes, making use of three feature selection filters, and five different classification algorithms. Our objective was to achieve a low false-positive rate (very important in this scenario) 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 was obtained by the CFS filter, reducing in 64 % the number of features.

4.3 Is feature selection paramount when dealing with microarray data?

Over the past two decades, the advent of DNA microarray data sets has stimulated a new line of research in bioinformatics and in machine learning. This type of data is used to collect information from tissue and cell samples regarding gene expression differences that could be useful for diagnosing disease or for distinguishing a specific tumor type. Although there are usually very few samples (often fewer than 100 patients) for training and testing, the number of features in the raw data ranges from 6000 to 60,000, since it measures the gene expression en masse, and converts this problem to the classical example of reduced sample size. A typical classification task is to separate healthy patients from cancer patients based on their gene expression "profile" (binary approach). There are also data sets where the goal is to distinguish between different types of tumors (multiclass approach), making the task even more complicated.

Since the introduction of this type of data, feature selection has been considered a de facto standard in the field, and a huge number of feature selection methods were utilized trying to reduce the input dimensionality while improving the classification performance. In [8], we have reviewed the up-to-date contributions of feature selection research applied to the field of DNA microarray data analysis, as well as analysing the intrinsic difficulties of this type of data (such as class imbalance, data set shift or the presence of outliers). In addition to this, we provided a practical evaluation of several well-known feature selection methods on a suite of nine widely used binary data sets. The conclusions of this comprehensive study were

- Support vector machines showed their superiority over other classifiers in this domain, as previously noticed by Gonzalez [27]. On the other hand, decision trees such as C4.5 may be affected by their embedded feature selection, in some cases leading to an extremely reduced set of features which can degrade the classification accuracy.
- Regarding the different feature selection methods tested, the filters that returned a subset of features showed an outstanding behavior, especially CFS and INTERACT. Notice that those methods which return a ranker require a threshold to decide the number of features to keep and, since this number has to be set a priori, it may prove too small or too large, the main disadvantage of using these types of methods.

To sum up, the conclusions of this work were that since the infancy of microarray data classification, feature selection has become an imperative preprocessing step, not only to improve the classification performance but also to help biologists identify the underlying mechanism that relates gene expression to diseases. Due to the high computational resources that these data sets demand, wrapper and embedded methods have been mostly avoided, in favor of less expensive approaches such as filters.

Regarding the opportunities for future feature selection research in this topic, there tends to be a focus on new combinations such as hybrid or ensemble methods. These types of methods are able to enhance the robustness of the final subset of selected features. Another interesting line of future research might be to distribute the microarray data vertically (i.e., by features) to reduce the heavy computational burden when applying wrapper methods.

4.4 The other side of the coin: feature selection for problems with large number of samples

It seems natural to believe that feature selection is more effective when the number of features is extremely high, but the truth is that it can be also helpful even when the number of features is relatively small but the number of samples is high, since it contributes to reduce the general dimension of the data.

An example of this type of problem can be the classification of intrusion detection systems; in particular, the KDD (Knowledge Discovery and Data Mining Tools Conference) Cup 99 data set [38] is a well-known benchmark for machine learning researchers. This data set contains five million samples represented by 41 features, with the aim of categorizing each connection in one of the following classes: normal connection (around 20 % of the connections), Denial of Service (DoS) attacks, Probe attacks, Remote-to-Local (R2L) attacks and User-to-Root (U2R) attacks. Usually, people deal with a smaller subset provided in the competition as training set (494 021 instances) and a test set containing 331 029 patterns.

Although the number of features in this data set cannot be considered extremely high (41), this data set is a good candidate for feature selection because of the characteristics of its input attributes. There are two features that are constant and some that are almost constant. Apart from constant features, the KDD Cup 99 data set has continuous features that are very skewed and for which a possible solution could be discretizing numeric data. So, in a previous work [10], we proposed a three-step methodology which consisted of (i) applying a discretizer method; (ii) after discretization, feature selection was applied using filters; and (iii) finally, a classifier was applied.

Table 3 shows the best results of our approach (two first columns) compared with other results in the literature as well

Table 3 Test results on KDD Cup 99 data set, comparison with other authors

Method	Error	TP	FP
Disc+Cons+C4.5	**5.14**	**94.08**	1.92
Disc+INT+C4.5	6.74	91.73	**0.44**
KDD Winner	6.70	91.80	0.55
5FNs_poly	6.48	92.45	0.86
5FNs_fourier	6.69	92.72	0.75
5FNs_exp	6.70	92.75	0.75
SVM Linear	6.89	91.83	1.62
SVM 2poly	6.95	91.79	1.74
SVM 3poly	7.10	91.67	1.94
SVM RBF	6.86	91.83	1.43
ANOVA ens.	6.88	91.67	0.90
Pocket 2cl.	6.90	91.80	1.52
Pocket mcl.	6.93	91.86	1.96

Best performance values marked in bold font

as with the winner of the KDD Cup 99 competition. Details of the experimental settings can be consulted in [10,13].

As can be seen in Table 3, the combination Disc+Cons+C4.5 obtains the best error and true positive rate employing only 6 features (14 % of total). Nevertheless, this improvement has a negative impact on the false-positive rate, although it is not the worst value in the table (SVM 3poly and Pocket mcl). The lowest FP rate is achieved using Disc+INTERACT+C4.5. It can be verified that these results outperform the results achieved by the KDD Winner. Error and TP rate are very similar and, in addition to this, a decrement in the FP rate is obtained using this combination. It must be emphasized that the FP rate is a measure of immense importance in determining the quality of an intrusion detection system. Moreover, this combination uses only 7 features, while the KDD winner employs the whole feature set (41). Therefore, a better result is obtained with a simpler model that only needs 17 % of the total features, enhancing the appropriateness of feature selection in this kind of data sets.

So far, this paper was devoted to studying feature selection methods and their adequacy for being applied to data of high dimensionality. However, there are still an important number of emerging challenges that researchers need to deal with and that will be outlined in the next section.

5 Open challenges

As mentioned already in the Introduction section, large-scale data are getting common nowadays in most contexts, due to the new possibilities available in sensoring and computing technologies. The so-called big-dimensional data might thus be created by handheld devices, social networks, internet of things, multimedia, and many other new applications with the well-known characteristics of volume, velocity and variety. The Terabyte (10^{12} bytes) is being gradually seen as "medium size", and now the usual measures are in PetaByte (10^{15} bytes) and progressively turning to Zettabytes (10^{21} bytes). Furthermore, problems like incomplete and inconsistent data, present already in small and medium data sets, will appear even more frequently, because data are obtained from different sensors and systems. The impact of noise, outliers, incomplete and inconsistent data, as well as redundant data (see Sect. 1.2), will be increased. Therefore, how to mitigate their impact will be an open issue. Consequently, feature selection methods probably will become one of the must-do preprocessing steps in handling these data sets to be able to obtain accurate, efficient and interpretable learning models. Ironically, though, most feature selection algorithms are not applicable in these high-dimensional data sets, due to their excessive temporal requirements. Thus, there are some open issues on the field:

– Using efficient methods to reduce the computational time will play an important role. Regarding this challenge, one of the paths that have been addressed already by a few papers is to consider parallel computing environments to re-implement feature selection algorithms [56,60,68]. Other authors, however, had tried to comply with the extreme dimensions of the new data sets using on-line or incremental feature selection methods [73,74]. The latter could also be used for those cases in which non-stationary distributions are confronted, also an important new line of improvement for FS methods. Some other works [3,5,67,76] address the distribution of the data sets both vertically (features) and/or horizontally (samples), as well as the methods that can be used to join the partial results obtained. In this way, an important reduction in the computational time is achieved, and more important, data sizes that were previously unapproachable become feasible, while accuracy is maintained and, in some cases, even improved.
– Scalability and stability of FS methods are other aspects that deserve deeper research. In large-scale problems, not only accuracy is important, but also a tradeoff between the latter and the computational complexity of the methods employed should be taken into account. In this sense, feature selection algorithms should be studied to reflect their sensitivity to variations in the training sets (stability). Only a few studies are available in this respect [15]. Scalability, that is, studying the behavior of the different algorithms when data set sizes increase, is also of great importance. In this case, univariate methods are generally more scalable, but as they do not take into account feature dependencies, the performance results of the subsequent machine learning algorithms are lower. Thus, if accuracy is to be preferred over temporal constraints, multivariate techniques are to be used [57]. Ensemble

feature selection is another relatively new approach that has appeared from the rising interest from various application fields, most notably from bioinformatics, due, as mentioned before, to the very high dimensionality of the data. In [1,11,12,40,63], the different possibilities of this approach for improving accuracy and stability are emphasized. Ensemble feature selection applies feature selection methods multiple times and combines the results into one decision; thus the final feature list should be more stable. Ensemble FS is similar to ensemble classification, and normally the procedures are either emphasizing data diversity (using the same methods over different samples of the data set), functional diversity (using different methods over the same data set), or in a hybrid manner. In all cases, the stability of the results is higher than in each of the methods individually [64], but other interesting advantages are that the performance is comparable, if not better, than that of the methods alone, while the non experienced users are exonerated from the task of selecting the adequate feature selection method for each problem at hand.

- Most of the successful stories of feature selection are related to classification problems, and in a lesser extent to regression. However, the application of FS to other areas is scarcely studied. In [53], the organizers of a NIPS workshop, Caruana and Joachims, stated that it is of great interest to explore the supervised learning problems that go beyond the standard value prediction model, such as those in which either the learning goal or the input to the learner is more complex than that in classification and regression. Some of these problems are ordinal regression, graph learning, learning partial or complete ranking preferences, one-class or anomaly detection problems, etc. Despite the time that has passed since that workshop, and although some new learning algorithms of the types mentioned were developed, specific feature selection aimed at dealing with their characteristics that could enhance their generalization capabilities are very few [2,26,37,42]. One interesting line of research could be not only a deeper study for designing new powerful FS algorithms for these problems, but also how to manage them in distributed scenarios.
- Better techniques for visualization of the features involved in a problem and their relations is undoubtedly another of the most interesting latest challenges. The availability of tools of this type will allow for better understanding of the problems, accessing insight of the data available [7]. However, although visualizations are allowed for several feature extraction techniques [22,65], that is not the case for feature selection methods that are preferably used when model interpretability is necessary, as the former transform the original features into a new set of features, while the latter maintain the original features, eliminating the redundant and irrelevant ones. Defiance is allowing useful, user-friendly visualization of results, as it is formulated for example, in several of the H2020 calls related to Big Data. Consequently, two areas that have not interacted frequently, such as feature selection methods and visualization techniques, with a widespread use in areas of Business Intelligence, should find an intersection to follow a common route that perhaps will take a leading role in the present real-world high-dimensional scenarios.
- Finally, the emergence of Big Data has also had a great influence on the fields of computer vision and multimedia analysis. A new term has been coined, named "Visual Big Data", which is devoted to visual information such as image and videos. In this scenario, dimensionality reduction techniques also play a important role, as demonstrated by the recent special issue in Neurocomputing on "Dimensionality reduction for visual Big Data" [54]. Papers in this special issue include subspace learning for visual Big Data, non-negative matrix factorization for visual Big Data, sparse representation for visual Big Data, feature extraction and selection for visual Big Data, metric learning for visual Big Data and applications of dimensionality reduction for visual Big Data. As can be seen, researchers in both communities (visual Big Data and dimensionality reduction) can benefit from huge opportunities as well as new challenges.

In conclusion, although feature selection is a field of machine learning that has been applied for decades, it is still in the spotlight due to the advent of Big Data and the appearance of new scenarios—not only related with massive volumes or stream data, but also with other aspects such as unbalanced classes, uncertain and partial labels, non-stationary distributions, etc.—which open new lines of research in which the use of feature selection is, perhaps, more necessary than ever. This new scenario offers both opportunities and challenges to machine learning researchers, who should embrace this opportunity to launch new lines of research.

Acknowledgments This research has been economically supported in part by the Ministerio de Economía y Competitividad of the Spanish Government through the research project TIN 2012-37954, partially funded by FEDER funds of the European Union, and by the Consellería de Industria of the Xunta de Galicia through the research project GRC2014/035. V. Bolón-Canedo acknowledges support of the Xunta de Galicia under postdoctoral Grant code ED481B 2014/164-0.

References

1. Awada, W., Khoshgoftaar, T.M., Dittman, D., Wald, R., Napolitano, A.: A Review of the Stability of Feature Selection Techniques for Bioinformatics Data. In: Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on, pp. 356–363 (2012)

2. Bahamonde, A., Bayn, G. F., Dez, J., Quevedo, J.R., Luaces, O., Del Coz, J.J., Goyache, F.: Feature subset selection for learning preferences: A case study. In: Proceedings of the International conference on Machine learning, p. 7. ACM (2004)
3. Banerjee, M., Chakravarty, S.: Privacy preserving feature selection for distributed data using virtual dimension. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, pp. 2281–2284. ACM (2011)
4. Bellman, R.E.: Adaptive control processes: a guided tour, vol. 4, p. 5. Princeton University Press (1961)
5. Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A.: Distributed feature selection: an application to microarray data classification. Appl. Soft Comput. 30
6. Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A.: A review of feature selection methods on synthetic data. Knowl. Inf. Syst. **34**(3), 483–519 (2013)
7. Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A.: Recent advances and emerging challenges of feature selection in the context of big data. Knowl. Based Syst. (2015)
8. Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., Benítez, J.M., Herrera, F.: A review of microarray datasets and applied feature selection methods. Inf. Sci. **282**, 111–135 (2014)
9. Bolón-Canedo, Verónica, Porto-Díaz, Iago, Sánchez-Maroño, Noelia, Alonso-Betanzos, Amparo: A framework for cost-based feature selection. Pattern Recognit. **47**(7), 2481–2489 (2014)
10. Bolon-Canedo, Veronica, Sanchez-Marono, Noelia, Alonso-Betanzos, Amparo: Feature selection and classification in multiple class datasets: An application to kdd cup 99 dataset. Expert Syst. Appl. **38**(5), 5947–5957 (2011)
11. Bolón-Canedo, Verónica, Sánchez-Maroño, Noelia, Alonso-Betanzos, Amparo: An ensemble of filters and classifiers for microarray data classification. Pattern Recognit. **45**(1), 531–539 (2012)
12. Bolón-Canedo, Verónica, Sánchez-Maroño, Noelia, Alonso-Betanzos, Amparo: Data classification using an ensemble of filters. Neurocomputing **135**, 13–20 (2014)
13. Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A.: Feature selection for high-dimensional data. Springer (2015). doi:10.1007/978-3-319-21858-8
14. Broad institute.: Cancer Program Data Sets. http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi. Accessed Jan 2016
15. Brown, G., Pocock, A., Zhao, M., Luján, M.: Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. J. Mach. Learn. Res. **13**(1), 27–66 (2012)
16. Bryant, R., Katz, R.H., Lazowska, E.D.: Creating revolutionary breakthroughs in commerce, science and society. Big-data Comput (2008)
17. Choh M.T.: Combining noise correction with feature selection. In: Data Warehousing and Knowledge Discovery, pp. 340–349. Springer (2003)
18. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) **2**(3), 27 (2011)
19. Cox, M., Ellsworth, D.: Application-controlled demand paging for out-of-core visualization. In: Proceedings of the 8th conference on Visualization'97, p. 235-ff. IEEE Computer Society Press (1997)
20. Dash, Manoranjan, Liu, Huan: Consistency-based search in feature selection. Artif. Intell. **151**(1), 155–176 (2003)
21. Duda, Richard O, Hart, Peter E, Stork, David G: Pattern classification, 2nd edn. Wiley, NY (2010)
22. Flach, P.: Machine Learning: The art and science of algorithms that make sense of data. Cambridge University Press, Cambridge (2012)
23. Frénay, Benoît, Verleysen, Michel: Classification in the presence of label noise: a survey. Neural Netw. Learn. Syst. IEEE Trans. **25**(5), 845–869 (2014)
24. Galar, Mikel, Fernández, Alberto, Barrenechea, Edurne, Bustince, Humberto, Herrera, Francisco: A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **42**(4), 463–484 (2012)
25. Garcia, S., Luengo, J., Herrera, F.: Data preprocessing in data mining. Springer, Switzerland (2015)
26. Geng, X., Liu, T. Y., Qin, T., Li, H.: Feature selection for ranking. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in Information Retrieval, p. 407–414. ACM (2007)
27. González Navarro, F.F.: Feature selection in cancer research: microarray gene expression and in vivo 1H-MRS domains. PhD thesis, Universitat Politècnica de Catalunya (2011)
28. Grossberg, Stephen: Nonlinear neural networks: Principles, mechanisms, and architectures. Neural Netw. **1**(1), 17–61 (1988)
29. Guyon, Isabelle, Gunn, Steve, Nikravesh, Masoud, Zadeh, Lofti A: Feature extraction: foundations and applications, vol. 207. Springer, Berlin, Heidelberg (2008)
30. Guyon, Isabelle, Weston, Jason, Barnhill, Stephen, Vapnik, Vladimir: Gene selection for cancer classification using support vector machines. Mach. Learn. **46**(1–3), 389–422 (2002)
31. Hall, M.A.: Correlation-based feature selection for machine learning. PhD thesis, The University of Waikato (1999)
32. Hashem, Ibrahim Abaker Targio, Yaqoob, Ibrar, Anuar, Nor Badrul, Mokhtar, Salimah, Gani, Abdullah, Khan, Samee Ullah: The rise of ''big data'' on cloud computing: review and open research issues. Inf. Syst. **47**, 98–115 (2015)
33. Hernández-Pereira, Elena, Bolón-Canedo, Veronica, Sánchez-Maroño, Noelia, Álvarez-Estévez, Diego, Moret-Bonillo, Vicente, Alonso-Betanzos, Amparo: A comparison of performance of k-complex classification methods using feature selection. Inf. Sci. **328**, 1–14 (2016)
34. Hoens, T.Ryan, Polikar, Robi, Chawla, Nitesh V.: Learning from streaming data with concept drift and imbalance: an overview. Progress in. Artifi. Intell. **1**(1), 89–101 (2012)
35. Hua, J., Tembe, W.D., Dougherty, E.R.: Performance of feature-selection methods in the classification of high-dimension data. Pattern Recognit. **42**(3), 409–424 (2009)
36. ICML workshop on Learning with Test-Time Budgets. https://sites.google.com/site/budgetedlearning2013/. Accessed Jan 2016
37. Jeong, Y.S., Kang, I.H., Jeong, M.K., Kong, D.: A new feature selection method for one-class classification problems. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **42**(6), 1500–1509
38. KDD Cup 99 Dataset. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html. Accessed Jan 2016
39. Kononenko, I: Estimating attributes: analysis and extensions of relief. In: Machine Learning: ECML-94, pp. 171–182. Springer (1994)
40. Kuncheva, L.: Combining pattern classifiers. Methods and algorithms. Wiley, Hoboken, NJ (2014)
41. Laney, Doug: 3d data management: Controlling data volume, velocity and variety. META Group Res. Note **6**, 70 (2001)
42. Laporte, L., Flamary, R., Canu, S., Djean, S., Mothe, J.: Nonconvex regularizations for feature selection in ranking with sparse SVM. Neural Netw. Learn. Syst. IEEE Trans. **25**(6), 1118–1130 (2014)
43. Lei, Yu., Liu, Huan: Feature selection for high-dimensional data: A fast correlation-based filter solution. ICML **3**, 856–863 (2003)
44. Lei, Yu., Liu, Huan: Efficient feature selection via analysis of relevance and redundancy. J. Mach. Learn. Res. **5**, 1205–1224 (2004)
45. Lichman, M.: UCI machine learning repository, 2013. http://archive.ics.uci.edu/ml. Accessed Jan 2016
46. Ling, C.X., Sheng, V.S.: Class imbalance problem. In Encyclopedia of Machine Learning, pp. 171–171. Springer (2010)

47. Liu, H,, Motoda, H.: Feature selection for knowledge discovery and data mining, volume 454. Springer Science and Business Media (2012)
48. Liu, H, Setiono, R.: Chi2: Feature selection and discretization of numeric attributes. In tai, p. 388. IEEE (1995)
49. López, Victoria, Fernández, Alberto, García, Salvador, Palade, Vasile, Herrera, Francisco: An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Inf. Sci. **250**, 113–141 (2013)
50. Molina, L.C., Belanche, L., Nebot, A.: Feature selection algorithms: a survey and experimental evaluation. In: Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on, pp. 306–313. IEEE (2002)
51. Moreno-Torres, Jose G., Raeder, Troy, Alaiz-RodríGuez, RocíO, Chawla, Nitesh V., Herrera, Francisco: A unifying view on dataset shift in classification. Pattern Recognit. **45**(1), 521–530 (2012)
52. Muhlbaier, Michael D., Topalis, Apostolos, Polikar, Robi: Learn. nc: Combining ensemble of classifiers with dynamically weighted consult-and-vote for efficient incremental learning of new classes. Neural Netw. IEEE Trans. **20**(1), 152–168 (2009)
53. NIPS 2002 Workshop: Beyond Classification and Regression: Learning Rankings, Preferences, Equality Predicates, and Other Structures. http://www.cs.cornell.edu/People/tj/ranklearn/. Accessed Jan 2016
54. Pang, Y., Shao, L.: Special issue on dimensionality reduction for visual big data. Neurocomputing **173**(Part 2), 125–126 (2016)
55. Peng, Hanchuan, Long, Fuhui, Ding, Chris: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. Pattern Anal. Mach. Intell. IEEE Trans. **27**(8), 1226–1238 (2005)
56. Peralta, S., Río, S., Ramírez-Gallego, I., Triguero, J.M., Benítez, Herrera, F.: Evolutionary feature selection for big data classification: a mapreduce approach. Math. Prob. Eng. (2015)
57. Peteiro-Barral, D., Boln-Canedo, V., Alonso-Betanzos, A., Guijarro-Berdiñas, B., Sánchez-Maroño, N.: Scalability analysis of filter-based methods for feature selection. Adv. Smart Syst. Res. **2**(1), 21–26 (2012)
58. Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D.: Dataset shift in machine learning. The MIT Press (2009)
59. Ross Quinlan, J.: Induction of decision trees. Machine Learn. **1**(1), 81–106 (1986)
60. Ramírez-Gallego, S., García, S., Mouriño-Talín, H., Martínez-Rego, D., Bolón-Canedo, V. D., Alonso-Betanzos, A., Benítez, J.M., Herrera, F.: Data discretization: taxonomy and big data challenge. WIREs Data Min. Knowl. Discov. **6**(1), 5–21 (2016)
61. Remeseiro, B., Bolon-Canedo, V., Peteiro-Barral, D., Alonso-Betanzos, A., Guijarro-Berdinas, B., Mosquera, A., Penedo, M.G., Sanchez-Marono, N.: A methodology for improving tear film lipid layer classification. Biomed. Health Inf. IEEE J. **18**(4), 1485–1493 (2014)
62. Remeseiro, B., Ramos, L., Penas, M., Martinez, E., Penedo, M.G., Mosquera, A.: Colour texture analysis for classifying the tear film lipid layer: a comparative study. In: Digital Image Computing Techniques and Applications (DICTA), 2011 International Conference on, p. 268–273. IEEE (2011)
63. Saeys, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics **23**(19), 2507–2517 (2007)
64. Seijo-Pardo, B., Bolón-Canedo, V., Porto-Díaz, I., Alonso-Betanzos, A.: Ensemble feature selection for ranking of features. In 2015 International Work Conference on Artificial Neural Networks (IWANN) 2015, pp. 29–42 (2015)
65. Shalev-Shwartz, S., Ben-David., S.: Understanding Machine Learning: From theory to algorithms. Cambridge University Press, Cambridge (2014)
66. Shalev-Shwartz, Shai: Online learning and online convex optimization. Found. Trends Mach. Learn. **4**(2), 107–194 (2011)
67. Sharma, A., Imoto, S., Miyano, S.: A top-r feature selection algorithm for microarray gene expression data. IEEE/ACM Trans. Comput. Biol. Bioinf. **9**(3), 754–764 (2012)
68. Spark implementations of Feature Selection methods based on information Theory. https://github.com/sramirez/spark-infotheoretic-feature-selection. Accessed Jan 2016
69. Tan, Kay Chen, Teoh, Eu Jin, Yu, Q., Goh, K.C.: A hybrid evolutionary algorithm for attribute selection in data mining. Expert Syst. Appl. **36**(4), 8616–8630 (2009)
70. Tsymbal, A.: The problem of concept drift: definitions and related work. Computer Science Department, Trinity College Dublin **106**, (2004)
71. Vernon T., John F.G., David R., Stephen M.: The digital universe of opportunities: rich data and the increasing value of the internet of things. International Data Corporation, White Paper, IDC_1672 (2014)
72. Vergara, Jorge R., Estévez, Pablo A.: A review of feature selection methods based on mutual information. Neural Comput. Appl. **24**(1), 175–186 (2014)
73. Wang, J., Zhao, P., Hoi, S.C., Jin, R.: Online feature selection and its applications. IEEE Trans. Knowl. Data Eng. p. 114 (2013)
74. Wu, X., Yu, K., Ding, W., Wang, H., Zhu, X.: Online feature selection with streaming features. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 11781192 (2013)
75. Yiteng, Z., Yew-Soon, O., Tsang, I.W.: The emerging "big dimensionality". Computational Intelligence Magazine, IEEE **9**(3), 14–26 (2014)
76. Zhao, Z., Zhang, R., Cox, J., Duling, D., Sarle, W.: Massively parallel feature selection: an approach based on variance preservation. Mach. Learn. **92**(1), 195–220 (2013)
77. Zhao, Zheng, Liu, Huan: Searching for interacting features. IJCAI **7**, 1156–1161 (2007)