Feature Selection with Complexity Measure in a Quadratic Programming Setting

Ricardo Sousa, Hélder P. Oliveira, and Jaime S. Cardoso

INESC Porto, Faculdade de Engenharia, Universidade do Porto, Portugal
{rsousa,jaime.cardoso}@inescporto.pt helder.oliveira@fe.up.pt,
WWW home page: http://www.inescporto.pt/~{rsousa,hfpo,jsc}

Abstract. Feature selection is a topic of growing interest mainly due to the increasing amount of information, being an essential task in many machine learning problems with high dimensional data. The selection of a subset of relevant features help to reduce the complexity of the problem and the building of robust learning models. This work presents an adaptation of a recent quadratic programming feature selection technique that identifies in one-fold the redundancy and relevance on data. Our approach introduces a non-probabilistic measure to capture the relevance based on Minimum Spanning Trees. Three different real datasets were used to assess the performance of the adaptation. The results are encouraging and reflect the utility of feature selection algorithms.

Keywords: Feature Selection, Pattern Recognition, Quadratic Optimisation

1 Introduction

With access to new means of information becoming increasingly easier, there has been a rapid growth of data. Providers gather useful information on customers' tastes and preferences in order to offer products and services similar to penchants of target consumers; bank analysts gather client's profiles based on their bank history and experience; and, with the advent of gene technology, it is now possible to sequence genes with innumerable possibilities of applications. Nevertheless, with the information growth and, consequently, the increased number of characteristics that define the data, new technologies are required to deal with this amount of information efficiently. One way is through the use of feature selection (FS) techniques that try to capture only the relevant information from data.

FS is an important issue in many applications, and in some cases, it is a necessary step for the construction of the predictive model (classification or regression). This became more important especially for data with an high dimensional nature. Moreover, the use of unimportant features can add undesirable complexity to the model and can negatively influence its behaviour by introducing errors in the prediction. Therefore, FS techniques were designed with different purposes, such as preventing overfitting, improving classification performance,

design of faster models and to improve understanding of the processes that generate the data. Furthermore, the usage of FS techniques can induce a reduction of the complexity of the dataset.

Other approaches used for dimensionality reduction, such as principal component analysis, differ from the FS problem. While the first one changes the original representation of the data, the second only uses a subset (containing hopefully the most important information) of the whole feature set. This means that the original semantics are preserved.

FS algorithms can be divided into three different categories depending on how the FS is incorporated into the predictive model construction task. The categories are mainly defined as filter, wrapper and embedded methods. The first one filters the data by removing features with low relevance. This is performed as a preprocessing step and it is independent of the construction phase of the predictive model. Wrapper methods iteratively select a subset of features until a sufficiently 'good' model is constructed. This evaluation is usually performed during the training and validation phase of the model. Finally, the embedded techniques are intrinsically designed with the learning model. A best subset of features is obtained during the training phase.

In [8] a new FS algorithm that captures in just one step the correlated features and those related to the classes is presented. This is achieved with a quadratic programming formulation where a term α controls how much importance the algorithm should give to the within features correlation or to feature-class correlation. To measure this correlation the authors in [8] use either the Pearson or Mutual Information (MI) to measure linear or non-linear relations, respectively. In this article an adaptation is presented motivated by [10] where the major advantage is a non-probabilistic approach for data estimation.

In Section 2 a recent method which tackles the problem of FS through a quadratic programming formulation will be briefly discussed. Some issues regarding the use of MI in the FS context will also be outlined. This strategy tries to identify the non-redundant features and those that are correlated with the class labels. In Section 3 an adaptation of this strategy of capturing the instances correlated with the class labels through a non-probabilistic scheme will be outlined. Finally, in Section 4 some results will be presented and some conclusions will be drawn in Section 5.

2 Quadratic Programming for Feature Selection

The authors in [8] proposed a new way of performing FS using a quadratic programming formulation. The FS problem can be described as a two-fold problem. First, one tries to eliminate the similar variables (redundancy) and secondly, to capture how correlated each feature is with the class (relevance). In [8] the authors propose to tackle the FS problem in one-step process through quadratic programming. The quadratic term (Q in Equation (1)) would capture the redun-

dancy whereas the linear term (F in Equation (1)) would capture the relevance.

$$\min_{\mathbf{x}} \left\{ \frac{1}{2}(1-\alpha)\mathbf{x}'Q\mathbf{x} - \alpha F'\mathbf{x} \right\} \tag{1}$$

The α constant can be considered as a trade-off between redundancy and relevance. The $\mathbf{x}$ magnitude values show how important each feature is to the problem. In order to capture these two different pieces of information, the authors first used the Pearson correlation measure. In doing so, the quadratic term (Q) and linear term (F) would be defined as follows:

$$Q_{ij} = \rho_{ij} = \frac{\sum_{m=1}^{M}(v_{mi}-\overline{v_i})(v_{mj}-\overline{v_j})}{\sqrt{\sum_{m=1}^{M}(v_{mi}-\overline{v_i})^2 \sum_{m=1}^{M}(v_{mj}-\overline{v_j})^2}} \qquad F_i = \sum_{k=1}^{C} P(K=k)|\rho_{iC_k}| \tag{2}$$

where M is the number of samples, v_{ki} is the k^{th} sample of random variable v_i and $\overline{v_i}$ is the sample mean of the random variable v_i. P is the empirical prior probability and C_k is the class represented by a 1-of-K coding scheme.

However, this measure only captures the linear relations among instances and between instances and classes. To retrieve the non-linear relations the authors opted to use Mutual Information (MI). This method is borrowed from the information theory fields and measures how much information two random variables share. In this setting, this type of metric would be used to assess redundancy and relevance. It is therefore not surprising that entropy measures such as MI can also be viewed as feature dependence estimators where its broadly use can be associated to the good performances achieved.

In spite of the concept of independence being mathematically well defined, the use of the opposite for dependence assessing does not provide how much in value one variable depends on the other. Hence, the dependence concept in the feature selection setting raises some questions. In [9] a set of postulates proposed by Rényi to define dependence is revised. In this work it is also argued that a measure that holds on those postulates is valid for capturing dependence in the feature selection context. In other words, given a specific set of statistical properties to evaluate dependence, a good estimator would have to satisfy its corresponding properties. It is also argued that this is not valid for MI, since despite most commonly used MI techniques are non-negative and symmetric, they are not invariant to one-to-one transformations and do not reach a maximum when features are highly correlated. Moreover, in [9] it is also argued that highly complex learning machines could easily cope with the data complexity and infer a linear relation with the features and output, or more precisely, perform overfitting on the data.

Due to the issues mentioned above, a more pragmatic approach could be more convenient. For this purpose we have selected MST (Minimum Spanning Trees). Even though MST was introduced as a generalisation of the Wald-Wolfowitz run test, it provides the means for our measure to assess the increase of complexity when a subset of features is removed. Furthermore, such measure can easily be fitted into the Equation (1) to assess data relevance.

Despite the elegant approach [8], the quadratic programming problems can be highly computational heavy. To overcome this difficulty, the authors [8] also

proposed an iterative optimisation strategy. Due to time restrictions, it was not possible to explore this approach. Nonetheless, for the proposal delved in the following section and with regards to the experiments performed, a generic quadratic programming solver is sufficient.

3 Feature Selection based on Redundancy and Relevance

In [8], the Pearson correlation was first used to measure the linear relation on each pair of features (*redundancy*). In other words, features that are highly correlated can be discarded since they do not provide a meaningful informative gain. However, this measure may not be appropriate if features are not linearly correlated. To overcome this problem, the MI was used. With regards to the correlation of features with the class (*relevance*), the Pearson and MI were used in the same manner. When measuring the linear relation between features and classes, a 1-of-K coding scheme for the labels was used [8].

The capture of the non-linear relations through information theory techniques such as mutual information can be very appealing (and widely used, e.g. [1, 4] to name a few). However, and as stated in Section 2, the use of MI for feature selection can, in theory, encompass some issues as argued in [9].

Several works presented in the literature tackle this situation by introducing new techniques to capture the non-linear relations. In the next section, the incorporation of a non-probabilistic based technique in the formulation presented in Equation (1) will be outlined.

3.1 Capturing Relevance through Classifier Complexity

This study was initially motivated by the work presented in [10]. Here, FS is defined as a method which aims to select a subset of features that minimises the overall classification complexity. To achieve this goal, the authors [10] propose a multiresolution approach on the feature space. However, such approach requires high computational effort. Therefore, they suggest techniques such as MST as an alternative [10].

Distribution free procedures are already well covered in the literature. In [6] a technique is proposed to measure the multivariate randomness by using a MST. This method constructs a minimum weight graph that connects all of the N nodes (instances) with E edges. The weights in our study were defined as the euclidean distance between a pair of instances. By using MST over all of our instances, the edges that connect nodes with different classes labels can afterwards be counted [6]. The rationale is, the more edges connecting nodes with different labels one has, the more complex the dataset is. The complexity of an MST is, in the worst case, of the order $\mathcal{O}(E + X \log(N))$, where X is the number of edges no longer than the longest edge in the MST. Our first trial encompassed the direct implementation of this setting to estimate features relevance. This analysis is performed feature by feature in the linear term in Equation (1) where

F_i is inversely proportional to the number of nodes with different labels which share an edge.

Based on this complexity analysis, one can also think how less (or more) complex our dataset can become if we discard a feature, say f_i. For this approach, one starts by measuring the complexity of the dataset with the whole features, $\mathbf{f}$. The gain would be expressed by the complexity increase with the removal of one feature. Formally,

$$F_{-i} = \frac{f_i - \mathbf{f}}{\mathbf{f}} = \Delta f_i$$

Although these techniques can be very complex in computational terms, they do not assume any probabilistic distribution of the data. For both adaptations, this problem was tackled from a wrapper perspective were the parameter α needed to be tuned.

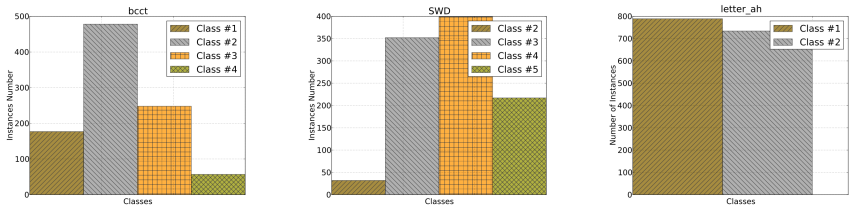
4 Experiments

The aim of this experimental study was to evaluate the usage of other measures to capture the relevance of each feature. Aiming towards the identification of the best features that describe the data, the question of which classifier was used is not considered as the authors did in [8]. Nevertheless, the same learning scheme was used in all experiments.

For the real data we tested the new measures on the Letter, SWD and BCCT datasets [2,3,5]. The first dataset, which is publicly available on the UCI machine learning repository, is composed of $20,000$ instances with 16 features describing the 26 capital letters of the English alphabet. Each instance is mainly defined by statistical moments and edge counts. In our experiments we used a subset of the whole dataset comprehending only the discrimination of the letter A versus the letter H. The SWD dataset contains real-world assessments of qualified social workers regarding the risk facing children if they stayed with their families at home and is composed by 10 features and 4 classes. The last dataset encompasses on 113 patients and expresses the aesthetic evaluation of Breast Cancer Conservative Treatment (BCCT) [3]. For each patient there were 8 observers that manually identified several fiducial points. Based on these points, an automated system automatically recorded 30 features, capturing visible breast alterations such as: breast asymmetry, skin colour changes due to the radiotherapy treatment and surgical scar appearance. The aesthetic outcome of the treatment for each patient was classified in one of the four categories: Excellent, Good, Fair and Poor. For this specific set, the same experiment procedure was conducted as described in [7].

In Fig. 1 the class frequency distribution for each dataset is depicted.

The training was performed on 20% of the data. The splitting of the data into training and test sets was repeated 5 times in order to obtain more stable results for accuracy by averaging and also to assess the variability of this measure. The best parametrisation of each model was found by a 'grid-search' based on a 5-fold cross validation scheme conducted on the training set.



(a) Frequency for BCCT dataset.

(b) Frequency for SWD dataset.

(c) Frequency for Letter AH dataset.

Fig. 1: Real datasets frequency values.

For these experiments the wrapper approach of the original method was selected . Therefore the best α value had to be tuned. We performed a search over the range 0.2 — 0.6 with a 0.2 step. Finally, the error of the model was estimated on the test set using the Misclassification Error Rate (MER) measure.

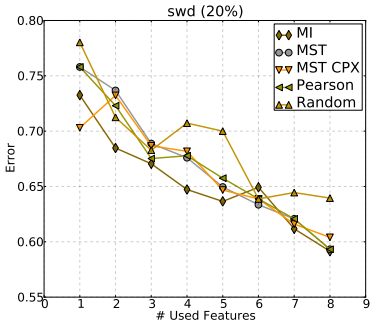
4.1 Results

We started our analysis on the SWD dataset. One can easily see that MI achieves the best results on the major subset of features—see Fig. 2a. However, when using more than 6 features from the whole set of features, MST, MST CPX and MI obtained similar results. The same behaviour can be stated on the letter dataset—see Fig. 2b. Here MST CPX attains a slightly better performance than all of the other techniques when using more than 8 features.

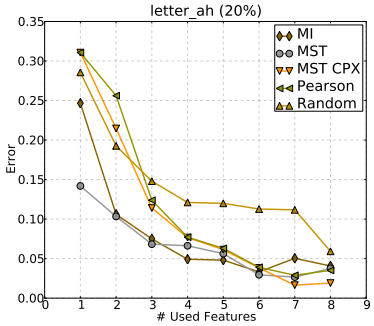
Afterwards, the performance of all feature selection techniques on the BCCT dataset was analysed. One can see that the minimum error attained was with the MI using two features—see Fig. 2c. However, the performance gain was subtle being the difference from the second best method of approximately 3%.

Being already defined some prior knowledge for this problem [7], an experiment exploring the benefit of such technical information was conducted. To achieve this goal, the prior knowledge was introduced as a postprocessing phase of the FS algorithm. This postprocessing can be described as follows. There are three important consequences of the BCCT treatment and they must all be considered. However, there are some features that translate these differences better than others. In order of importance, from the the most to the least important, we have the asymmetry alteration, colour change and scar appearance. For example, if four features had to be selected, we would use 2 asymmetry features, 1 colour feature and 1 of scar feature, in order of importance given by the FS algorithm.

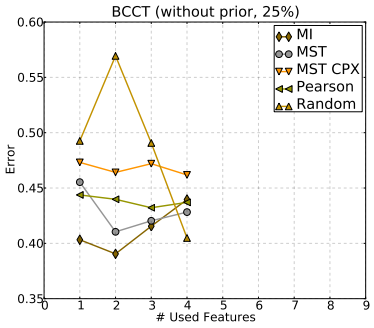
From the prior knowledge perspective, it clearly attained the best results—see Fig. 2d. Even with the prior knowledge, MI and MST based approaches were not able to attain such results. The best selected features were ρBRA and BCD for MST, and BCD and LBC for MI. For the Pearson measure, the best selected features were ρLBC, sX2L, cX2b and ρBCE.



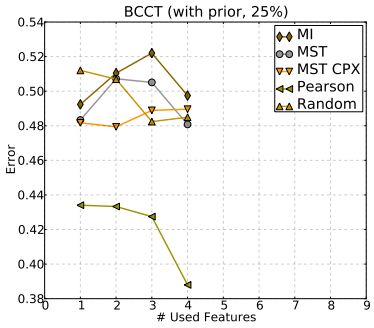
(a) Results for SWD dataset.



(b) Results for Letter (A vs. H subproblem) dataset.



(c) Results for BCCT dataset (without prior).



(d) Results for BCCT dataset (with prior).

Fig. 2: Results for real datasets performed with the several FS techniques.

Since a wrapper methodology was used, a best classifier would be required to assess each subset of features used. However, due to time restrictions such analysis could not be performed. Therefore, further studies could improve the assessment of this behaviour. Nevertheless, either MST or MST CPX attain very similar results when compared to MI.

5 Conclusion

This paper presents an adaptation of a recent FS method [8] by exchanging the relevance term with a MST as a complexity measurement. The goal of this FS scheme is to quantify the within variables similarity (redundancy) and the correlation between features and the class labels (relevance) through a quadratic optimisation problem. A constant α is used as a trade-off term to define the importance of the redundancy or relevance for the optimisation problem. After selecting a subset of features, they were used during the learning model construction and their use on a test dataset was assessed using the Misclassification

Error Rate (MER). For these experiments, three real (Letter AH, SWD and BCCT) datasets were used.

A preliminarily study shows that MST provides just as good results as MI. However, MST has the advantage of not assuming any data density distribution. It was also interesting to see that on the BCCT database the results show that the baseline method with prior knowledge performs better in terms of classification error when compared to the other approaches.

Regarding to future work, this work can be extended in several directions. One of them passes through the stability assessment of this FS method [11]. Another one is the analysis of the non-linearity within the features. Finally, experiments on larger datasets are required.

Acknowledgements This work was partially funded by Fundação para a Ciência e Tecnologia (FCT) with reference PTDC/EIA/64914/2006 and SFRH/BD/43772/2008. The authors would also like to thank Sohan Seth for the mathematical clarifications.

References

1. Balagani, K.S., Phoha, V.V.: On the feature selection criterion based on an approximation of multidimensional mutual information. IEEE Trans. Pattern Anal. Mach. Intell. 32, 1342–1343 (July 2010)
2. Ben-David, A., Sterling, L.: Generating rules from examples of human multiattribute decision making should be simple. Expert Systems with Applications 31(2), 390 – 396 (2006)
3. Cardoso, J.S., Cardoso, M.J.: Towards an intelligent medical system for the aesthetic evaluation of breast cancer conservative treatment. Artificial Intelligence in Medicine 40, 115–126 (2007)
4. Estévez, P.A., Tesmer, M., Perez, C.A., Zurada, J.M.: Normalized mutual information feature selection. Trans. Neur. Netw. 20, 189–201 (February 2009)
5. Frey, P.W., Slate, D.J.: Letter recognition using holland-style adaptive classifiers. Mach. Learn. 6, 161–182 (March 1991)
6. Friedman, J.H., Rafsky, L.C.: Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. Annals of Statistics 7(4), 697—717 (1979)
7. Oliveira, H.P., Magalhaes, A., Cardoso, M.J., Cardoso, J.S.: An accurate and interpretable model for bcct.core. In: Proceedings of the 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 6158–6161 (2010)
8. Rodriguez-Lujan, I., Huerta, R., Elkan, C., Cruz, C.S.: Quadratic programming feature selection. Journal of Machine Learning Research 11, 1491–1516 (April 2010)
9. Seth, S., Príncipe, J.C.: Variable Selection: A Statistical Dependence Perspective. In: Proceeding of the Ninth International Conference on Machine Learning and Applications. pp. 931—936 (2010)
10. Singh, S.: Multiresolution estimates of classification complexity. IEEE Trans. Pattern Anal. Mach. Intell. 25, 1534–1539 (December 2003)
11. Somol, P., Novovicova, J.: Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. Pattern Analysis and Machine Intelligence, IEEE Transactions on 32(11), 1921 –1939 (2010)