ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

Mutual information based feature selection for mixed data

Gauthier Doquire and Michel Verleysen *

Université Catholique de Louvain - ICTEAM/Machine Learning Group
Place du Levant 3, 1348 Louvain-la-Neuve - Belgium

Abstract. The problem of feature selection is crucial for many applications and has thus been studied extensively. However, most of the existing methods are designed to handle data consisting only in categorical or in real-valued features while a mix of both kinds of features is often encountered in practice. This paper proposes an approach based on mutual information and the maximal Relevance minimal Redundancy principle to handle the case of mixed data. It combines aspects of both wrapper and filter methods and is well suited for regression problems. Experiments on artificial and real-world datasets show the interest of the methodology.

1 INTRODUCTION

Feature selection is a task of great importance when mining datasets of high dimension. Indeed, getting rid of redundant or irrelevant features generally increases the performances of a predictive model and make it more interpretable and less prone to overfitting [1]. Moreover, dimensionality reduction also decreases the computational load of models. Eventually, on a more practical point of view, feature selection can prevent from collecting and storing data whose measurement can either be expensive or hard to perform.

These reasons lead to the development of a huge number of feature selection algorithms in the past few years. The large majority of them assumes the datasets are either continuous, i.e. all the features are real-valued variables, or categorical, i.e. all the features take values amongt a finite set of categories or symbols. However, in many practical applications, data come in a mixed way. As an example, medical surveys can include continuous features as the height or the heart rate of a patient, together with categorical ones as the type of diabete encountered. Socio-economic data also often contain discrete variables about individuals like the level of education, the relationship status or the sex, as well as continuous features like the age or the income.

The first obvious way to handle this kind of data is to code the categorical attributes into discrete numerical values before applying an algorithm designed for continuous data. However, this approach is not appropriate since the distance between samples would then have no meaning. Indeed, using the Euclidean distance, two different codings would lead to different distances which is an undesirable property.

*G. Doquire is funded by a Belgian F.R.I.A grant.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

On the other hand, using an algorithm handling categorical data after the discretization of the continuous features [2] could easily lead to a loss of information. Algorithms able to handle mixed attributes are thus needed.

Some algorithms have been proposed to this end, essentially for classification problems, e.g. [3, 4]. The first one [3] uses the error probability of a classification model, while the other one [4] is based on rough sets theory.

To the best of our knowledge, the only attempt to achieve feature selection for regression problems with mixed data is the work of Hall [5] which essentially consists in a maximal Relevance minimal Redundancy (mRmR) approach with the correlation coefficient as criterion.

In this paper, a mRmR approach is also followed but there are two major differences with [5]. First the approach is based on mutual information (MI). MI has the advantage over correlation to detect non-linear relationships between attributes. Moreover, at each step of the algorithm either the best categorical or the best continuous feature is selected in a wrapper-like procedure.

The remaining of the paper is organized as follows. Section 2 briefly recalls fundamental notions about MI. Section 3 introduces the algorithm. Section 4 presents experimental evidences of the interest of the method, while Section 5 concludes the work.

2 MUTUAL INFORMATION

Shannon's MI [6] has been used successfully in many feature selection algorithms [7]. It is a symmetric measure of the dependance between two random variables X and Y whose formal definition is:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{1}$$

where $H(X)$ is the entropy of X. The entropy can be understood as the degree of uncertainty we have on the values taken by X:

$$H(X) = -\int f_X(\zeta_X) \log f_X(\zeta_X)\, d\zeta_X, \tag{2}$$

f_X being the probability density function (pdf) of X.

MI can also be expressed in term of conditionnal entropies, making its interpretation for feature selection quite obvious:

$$I(X;Y) = H(Y) - H(Y|X). \tag{3}$$

Indeed, a feature having a high MI with the output reduces the uncertainty about this ouptut we want to predict.

MI can eventually be expressed as:

$$I(X;Y) = \int\int f_{X,Y}(\zeta_X, \zeta_Y) \log \frac{f_{X,Y}(\zeta_X, \zeta_Y)}{f_X(\zeta_X) f_Y(\zeta_Y)}\, d\zeta_X\, d\zeta_Y. \tag{4}$$

As f_X, f_Y and $f_{X,Y}$ are not known in most cases, MI cannot be analytically computed but has to be estimated from the data.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

3 METHODOLOGY

A natural objective of feature selection is to build a set of sufficiently informative features which is as compact as possible. Indeed, a feature very informative about the output to predict is useless and should thus not be selected if it carries the same information as another previously selected feature.

Moreover, feature selection algorithms require a way to search the feature space and to build the set of selected features since it is not possible in practice to test the $2^n - 1$ possible feature sets (n being the number of features). A simple and fast solution is to employ greedy search procedures such as a forward search strategy. It consists in adding at each step the best feature according to a particular criterion. This choice is never questionned again later.

To combine these two ideas, a quite simple solution is to look at each step for the feature whose difference between its relevance and its redundancy with already selected features is maximal. In this context, MI can be used to evaluate both the relevance and the redundancy [8]. More precisely, if Y is the output to predict, F the set of indices of all features and S the set of indices of already selected features, a possible approach [8] is to give each feature f_j, $j \in F \backslash S$, the score :

$$Score(f_j) = I(f_j; Y) - \frac{1}{|S|} \sum_{s \in S} I(f_j; f_s). \tag{5}$$

The feature with the highest score is then selected and a new score is computed for all of the unselected feature.

Since the problem we are considering here involves categorical and continuous features, it is necessary to make a distinction between them since the range of the score (5) each kind of feature can take could be completely different. Such a difference would of course bias the selection procedure.

To circumvent this, once the score of each feature has been computed at one step of the algorithm, we consider the best categorical and the best continuous features and choose the one whose addition to the current set leads to the best prediction performances of a particular model (measured here by the root mean squared error). This methodology thus combines aspects of filter and wrapper, but keeps the number of models to build relatively small (at most $2n - 2$) compared to a pure wrapper approach and a forward search strategy ($(n(n+1)/2) - 1$).

4 EXPERIMENTAL RESULTS

In this section experimental results are given to demonstrate the interest of the proposed approach. First, the algorithm is tested on three artificial datasets with both relevant and irrelevant features. Then it is shown that the prediction performances of two models can benefit from the described methodology.

In this paper, a nearest neighbors based estimator introduced by Kraskov et al. [9] is used to estimate the MI between continuous features. The estimator is

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

defined as:

$$\hat{MI}(X,Y) = \psi(N) + \psi(K) - \frac{1}{K} - \frac{1}{N}\sum_{i=1}^{N}(\psi(\tau_x(i)) + \psi(\tau_y(i))) \quad (6)$$

where ψ is the digamma function, N the number of points, K the number of neighbors considered and $\tau_x(n)$ the number of points whose distance from x_n, the n^{th} point of X, is not greater than $0.5\max(\epsilon_x(n), \epsilon_y(n))$. $\epsilon_x(n)$ is the distance between x_n and its K^{th} nearest neighbor. When estimating MI between a categorical and a continuous feature, a modified version of [9] is employed, which has originally been introduced for classification problems [10]. The value of the parameter K in both MI estimators was set to 6 as prescribed in [9].

4.1 Artificial datasets

Each of the artificial datasets consists in 4 random binary variables $X_1 \ldots X_4$ taking value 1 or 0 with the same probability, and 4 random continuous variables $X_5 \ldots X_8$, uniformly distributed on [0 , 1]. The sample size is 100. Three outputs are built from this dataset:

$$Y_1 = (X_1 \times X_5) + (X_2 \times X_6), \quad (7)$$

$$Y_2 = 2 \times \sin(X_1 X_5) + 2 \times \cos(X_2 X_6), \quad (8)$$

$$Y_3 = 4 \times X_1 \times \sin(X_2 X_5) + X_6. \quad (9)$$

As the method requires the use of a specific prediction model, it has been tested with the m5' regression tree [11] using the implementation by Jekabsons [12], as well as with a 5-nearest neighbors (5nn) prediction model where the metric used is the Heterogeneous Euclidean-Overlap Metric (HEOM) [13]. Table 1 indicates, for 50 randomly generated datsets, how often the only 4 relevant features have been selected first. The algorithm is denoted by Mixed MI-knn or ixed MI-tree according to the model employed.

For comparison, results obtained with the CFS algorithm [5] using the correlation as relevance criterion are also presented. The approach followed is exactly the one described in the original paper except that in the second phase, all features are added in an order depending on the difference between their correlation with the output and the maximal correlation with a previously selected feature. This allows a fair comparison between the two methods without using the stopping criterion in [5].

The MI-based algorithm clearly outperforms the CFS algorithm for the three cases with both models. As already stated, this is most probably due to the ability of MI to detect non-linear relationships between features.

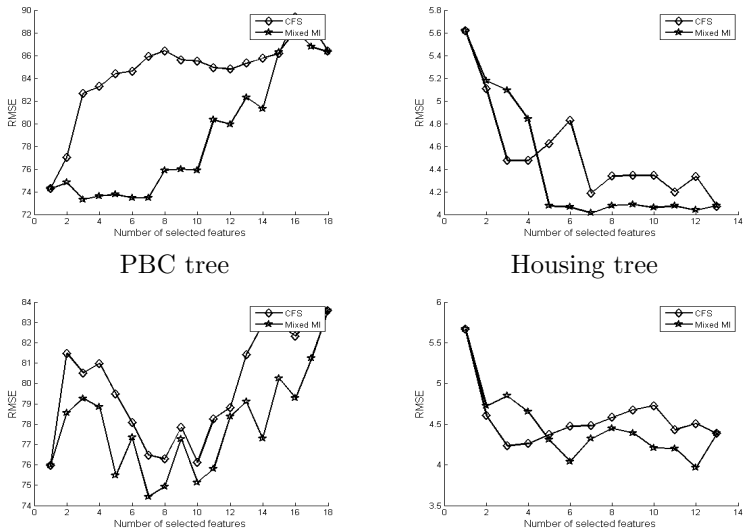
4.2 Real-world datasets

Two real-world datasets are used for further assesment of the algorithm. The first one is the PBC dataset, containing 10 continuous and 8 categorical features.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

	CFS	Mixed MI-tree	Mixed MI-knn
Y_1	41	47	47
Y_2	21	47	49
Y_3	36	47	47

Table 1: Number of times each algorithm correctly identifies the four informative features out of 50 experiments.



PBC tree Housing tree

PBC 5nn Housing 5nn

Fig. 1: RMSE of a 5 nearest neighbors predictor (5nn) and a m5' regression tree (tree) as a function of the number of selected features.

The sample size is 276. The dataset is available on the StatLib datasets archive website[1] . The second one is the well known Boston Housing dataset. The task here is to predict the prices of houses in the suburbs of Boston according to 1 binary and 12 continuous socio-economic and demographic features. In this work, only the 1000 first samples have been considered. The dataset is available from the UCI Machine Learning Repository [2] .

Figures 1 shows the 5-fold cross-validation error as a function of the number of selected features. Here again, results advocate in favour of the Mixed MI approach as it always leads to the global smallest RMSE. For the PBC dataset, the algorithm leads to better performances for any number of features, except when working with one and all the features (in this last case the RMSE is of course equal for both methods).

[1] http://lib.stat.cmu.edu/datasets/
[2] http://archive.ics.uci.edu/ml/index.html

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

5 CONCLUSIONS AND FUTURE WORK

In this work, an algorithm for feature selection in regression problems with both categorical and continuous data is introduced. It is based on the mRmR principle where both the relevance and the redundancy are evaluated using MI. It also includes a wrapper step in order to handle possible different score ranges for continuous and categorical features. However, this approach keeps affordable the number of models to build.

The methodology is tested on 3 artificial and 2 real-world datasets. The results show that the approach is more able to identify relevant features than the CFS algorithm when non-linear relationships exist between the features and the output. Moreover, for the considered real-world datasets, it leads to improved precision for two prediction models in terms of RMSE.

All the experiments led in this paper are devoted to regression problems, since they have been much less studied than classification ones. However, with some adaptations of the MI estimators, the same methodology could be tested on classifications problems too [10].

References

[1] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[2] H. Liu, F. Hussain, C. L. Tan, and M. Dash. Discretization: An enabling technique. *Data Mining and Knowledge Discovery*, 6(4):393–423, 2002.

[3] W. Tang and K. Z. Mao. Feature selection algorithm for mixed data with both nominal and continuous features. *Pattern Recogn. Lett.*, 28(5):563–571, 2007.

[4] Q. Hu, J. Liu, and D. Yu. Mixed feature selection based on granulation and approximation. *Know.-Based Syst.*, 21(4):294–304, 2008.

[5] M. A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning*, pages 359–366, 2000.

[6] C. E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423, 623–656, 1948.

[7] R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5:537–550, 1994.

[8] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27:1226–1238, 2005.

[9] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Phys. Rev. E*, 69(6):066138, 2004.

[10] V. Gomez-Verdejo, M. Verleysen, and J. Fleury. Information-theoretic feature selection for functional data classification. *Neurocomputing*, 72(16-18, Sp. Iss. SI):3580–3589, 2009.

[11] R. J. Quinlan. Learning with continuous classes. In *5th Australian Joint Conference on Artificial Intelligence*, pages 343–348, Singapore, 1992. World Scientific.

[12] G. Jekabsons. M5primelab: M5' regression tree and model tree toolbox for Matlab/Octave, 2010.

[13] R. Wilson and Martinez T. R. Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*, 6:1–34, 1997.