ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Cluster homogeneity as a semi-supervised principle for feature selection using mutual information

Frederico Coelho[1] and Antonio Padua Braga[1] and Michel Verleysen[2] *

1- Universidade Federal de Minas Gerais - Brazil

2- Université Catholique de Louvain - Belgium

Abstract. In this work the principle of homogeneity between labels and data clusters is exploited in order to develop a semi-supervised Feature Selection method. This principle permits the use of cluster information to improve the estimation of feature relevance in order to increase selection performance. Mutual Information is used in a Forward-Backward search process in order to evaluate the relevance of each feature to the data distribution and the existent labels, in a context of few labeled and many unlabeled instances.

1 Introduction

The solution of machine learning problems is often hampered by redundant information embeded into a large number of variables, which are usually chosen to represent the problem according to their availability and to some sort of *a priori* knowledge. Reducing the number of variables by Feature Selection (FS) may improve learning performance by smoothing the effects of the well known "curse of dimensionality" and "concentration of the euclidean norm" [1] problems. FS may also contribute to a better understanding of the variable behavior, bringing more clearly physical interpretation of real problems [2];

A common approach to FS is to estimate the relevance and to rank each feature according to their relation (or correlation) with the output targets [3, 4]. This approach is intuitive and easy to implement but it usually fails to consider the relevance of a given feature in the presence of others[2], since most filter methods are univariate [4, 5]. In this context, Mutual Information (MI) [6] arises as a good "relation" criterion, since it is a multivariate measure which is widely used to evaluate relations among sets of features and output labels.

Basically, labels and data are the available sources of information to perform FS. Many methods [3, 7] are able to deal only with labeled data while others only deal with unlabeled data [8, 9]. However, in many real situations, the amount of labeled data is not sufficient to characterize well the relations between input data and output classes. Since labeling by human experts can be costly, it is common in many kinds of problems to have large unlabeled data sets available and very few labeled data. Due to the availability of the large unlabeled data set, the question that arises in such a context is "why not to use information extracted from the unlabeled data in order to estimate feature relevance and

*Work developed with founding support from CAPES - Process BEX 1456105

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

to induce models?" The joint use of labeled and unlabeled data to perform FS characterizes the semi-supervised feature selection paradigm.

Some machine learning approaches include clustering methods in order to label instances. They are based on the assumption that the underlying distributions of the data, and their modes, can be estimated from the sampled data by clustering methods. One of the basic principles of structural data analysis is that labels are consistent with data distributions. Accordingly, the relevance of features to labels should also be reflected by the relevance of features to clusters.

In this work a semi-supervised FS strategy based on MI will be introduced. The basic principle of the method is replace, for unsupervised data, the label information by cluster information in order to estimate the relevance of each feature or feature subset.

This paper is organized as follows: first the FS framework will be summarized. Then the use of unlabeled data into this framework will be detailed. Next some experiments will be presented as well as their results leading to the conclusions.

2 Feature Selection

Feature selection is usually accomplished according to a relevance criterion and to a search strategy. The former aims to assess how relevant a single feature subset is, while the latter aims to guide the search towards the most relevant feature subset, since, in practice, testing all possible subsets (exhaustive search) can be unfeasible even for problems with few variables. In this work a filter method is implemented using MI as a relevance criterion. Roughly speaking, MI measures the amount of information shared among two or more sets of variables [6] capturing even nonlinear relations among them. The multivariate properties of MI makes it an important approach to assess the relevance of subsets of features, since it may be affected by joint behavior of a feature in the presence of others. Equation 1 shows the relevance evaluation between the input data X and the output vector Y:

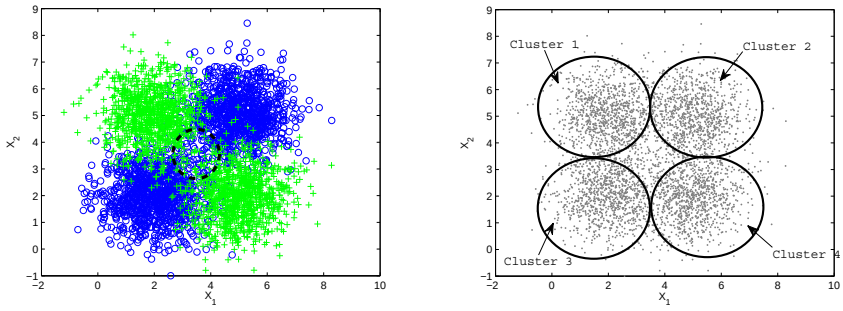
$$r = MI\left(X, Y\right) \ . \tag{1}$$

The implemented search technique is the forward-backward (FB) procedure [10, 2]. The forward strategy has smaller capability to finding more complementary features, compared to backward selection. On the other hand even the smallest nested subset is predictive. The backward strategy, in turn, is capable of finding complementary features, however, its performance is degraded for smallest nested subsets. So, the forward-backward process tries to get the best of both approaches.

3 Using Unlabeled Data

Evaluating feature relevance using MI requires that the data set contains some labeled data; however, small data sets may fail to represent well the general

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.



(a) (b)

Fig. 1: For the two class XOR problem, in 1(a) none of the features alone can explain the distribution of the classes, defined by circles and crosses, and in 1(b), even without the labels, features 1 and 2 are still able to explain data distribution.

relation between input and output variables as shown in the illustrative example[1] of Figure 1(a). Is this example, the distribution of labels is not well represented if a small data set is sampled within the central circle. Since labeling can be costly, it is expected that unlabeled data could provide some information about the posterior probability of labels that could improve FS. The feature selection task could be performed by searching for those features that are important not only for labels, but also for clusters, which are expected to be consistent with labels. The use of both labeled and unlabeled data characterizes the semi-supervised paradigm.

Data distribution information can be useful even when there is a reasonable amount of labeled data. As an example, consider a forward FS procedure applied to a three dimensional problem, for which features X_1 and X_2 together fully explain the labels in Y and X_3 is completely random (Fig. 1(a) shows the relevant features). Individually none of the three features is able to explain the labels, so in the first iteration of the algorithm (that will be univariate), their MI value will be small and, by chance, feature X_3 could be ranked first, resulting in a poor initial subset selection. In such a situation the distribution of the dataset may provide additional information about the relevance of X_1 and X_2.

Features X_1 and X_2, together, are still able to discriminate the instances into four different clusters according to the distribution of the dataset, regardless of labels, as shown at Figure 1(b). So, if we are able to estimate the cluster structure that best fits data generator functions, we can estimate the relevance of each feature subset according to the dataset distribution. Each pattern, especially the unlabeled ones, can be associated to a given cluster and receive a tag according

[1]This is an hypothetical example to illustrate the problem. In real problems labeled and unlabeled data are not expected to be concentrate in different space regions.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

to the cluster number (Figure 1(b)). These "cluster labels" assigned to each unlabeled data, results on the *cluster label vector* Y_{cl}. In addition, the number of clusters N_c should be sufficiently large in order to guarantee label homogeneity within clusters.

In general, the MI between a feature set X and its vector of labels Y can be defined in terms of their joint and marginals probabilities as

$$MI\left(X,Y\right)=\sum_{x\in X}\sum_{y\in Y}p\left(x,y\right)log\frac{p\left(x,y\right)}{p\left(x\right)p\left(y\right)}\ . \tag{2}$$

Equation 2 can be rewritten by splitting the data according to their classes as shown in Equation 3 for a binary case, where superscripts (1) and (-1) indicate respectively the data belonging to classes +1 and -1:

$$MI\left(X,Y\right)=\sum_{x\in X^{(1)}}\sum_{y\in Y^{(1)}}p\left(x,y\right)log\frac{p\left(x,y\right)}{p\left(x\right)p\left(y\right)}+\sum_{x\in X^{(-1)}}\sum_{y\in Y^{(-1)}}p\left(x,y\right)log\frac{p\left(x,y\right)}{p\left(x\right)p\left(y\right)}\ . \tag{3}$$

Assuming that, after clustering procedures, clusters C_i, $i=1,2,...,k$ are homogeneous and correspond to instances from the same class, i.e., they were generated in such a way that $\{C_1,C_2,...,C_i\}\subset Y^1$ and $\{C_{i+1},...,C_k\}\subset Y^{-1}$, the MI can be rewritten as

$$\begin{aligned}MI\left(X,Y\right)=&\sum_{x\in X^{(1)}}\sum_{c\in C_1}p\left(x,c\right)log\frac{p\left(x,c\right)}{p\left(x\right)p\left(c\right)}+\ldots+\sum_{x\in X^{(1)}}\sum_{c\in C_i}p\left(x,c\right)log\frac{p\left(x,c\right)}{p\left(x\right)p\left(c\right)}+\\&\sum_{x\in X^{(-1)}}\sum_{c\in C_{i+1}}p\left(x,c\right)log\frac{p\left(x,c\right)}{p\left(x\right)p\left(c\right)}+\ldots+\sum_{x\in X^{(-1)}}\sum_{c\in C_k}p\left(x,c\right)log\frac{p\left(x,c\right)}{p\left(x\right)p\left(c\right)}\end{aligned}\ . \tag{4}$$

In such situation, $MI(X,Y_{cl})=MI(X,Y)$. In practice, as we are dealing with unlabeled data, if the number of clusters is defined sufficiently large to allow that clusters encompass mostly instances from same class, we have that $MI(X,Y_{cl})\approx MI(X,Y)$. Equation 1 can now be rewritten as

$$r_s=MI\left(X^{(\ell)}\cup X^{(u)},Y\cup Y_{cl}\right)\ , \tag{5}$$

where $X^{(\ell)}$ and $X^{(u)}$ are respectively the labeled and unlabeled data sets, Y is the label vector and Y_{cl} is the vector of cluster labels. Equation 5 can be directly used in our forward-backward FS filter method. In this way more information about the relevance of each feature subset is provided taking into account the cluster information. Therefore cluster information replaces the "label" information for unlabeled data in order to consider them in the evaluation of the MI.

4 Experiments and Results

The experiments consist in comparing the performances of feature subsets selected according to a pure supervised approach and the semi-supervised method

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

presented in this paper. A sequential FB FS strategy [10, 2] was implemented and applied to some real and synthetic datasets, using a MI estimator tailored to classification problems. This estimator was developed by Goméz et al. [10] which has high performance even in a context of scarce data. Data was clustered with K-means algorithm [11]; the number of clusters N_c is shown in Table 1. N_c was empirically chosen in such way to be sufficiently large in order to guarantee label homogeneity within clusters.

The final results aim at comparing the final feature subset obtained when using only labeled data S_ℓ, with the one obtained using both labeled and unlabeled data $S_{\ell u}$. After sampling the two sets, the Linear Discriminant Analysis Method (LDA) was used in order to classify the test set in three different conditions: considering only S_ℓ, $S_{\ell u}$ or the set F of all features. The mean classification accuracy and standard deviation for 10 different trials are presented in Table 2. LDA was chosen to perform the classification tests due its simplicity and robustness.

Three data sets were used in the experiments. The first one (FBench) is a synthetic data set, originally developed for benchmark regression problems [12], whose output is a function of some of their random input variables. Its output was discretized into two classes (1 for $Y > 0$ and -1 for $Y < 0$) in order to transform it into a classification problem. Two other problems come from the UCI Machine Learning Repository (www.ics.uci.edu/∼mlearn/): the sonar data set, composed by instances of a sonar response from rocks and mines, and the Pen-Based Handwritten Digits data set, composed by digit samples from 44 different writers. For this last problem we considered only instances of digits 1 and 2 in the experiments.

On each trial a very small portion of data N_ℓ was chosen as labeled data, another N_t quantity was selected as a test set and the rest N_u instances was considered as unlabeled data, so their labels were not considered in the FS task.

Table 1: Data and algorithm parameters: N is the total number of instances an N_f is the total number of features

Problem	N_f	N	N_ℓ	N_u	N_t	N_c	S_ℓ	$S_{\ell u}$
FBench	10	10000	49	7952	1999	100	4 1 5	1-5-4-10-2-3
Sonar	60	208	11	147	41	40	46	46-36-20-27-30-16-43-24
Pen	16	2287	114	1717	456	30	4	4-15

Table 2: Shows the results for each test, where $\#_f$ is the final number of features of each subset.

Problem	$(\#_f)$ $Accuracy \pm \sigma$		
	F	S_ℓ	$S_{\ell u}$
FBench	(10) 0.8502±0.0123	(3) 0.8199±0.0127	(6) 0.8504±0.0122
Sonar	(60) 0.7117±0.6667	(1) 0.6052±0.1343	(8) 0.6924±0.0803
Pen	(16) 0.9808±0.0075	(1) 0.8478±0.0251	(2) 0.8780±0.0264

In all experiments the obtained accuracy for the subset $S_{\ell u}$ is higher than those obtained using the features selected when using only the labeled data. It is possible to observe in Table 2 that, for Fbench and Sonar problems, there is no representative accuracy loss when using only the features in $S_{\ell u}$ instead of using

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

all features. Only for the Pen data set there is a loss with respect to F. However, there is an improvement in accuracy with respect to using only the supervised set S_ℓ, as expected, since the objective here is to show that cluster information from unlabeled data, and consequently the proposed method, conveys information to improve FS.

5 Conclusion

This work proposes a semi-supervised FS method based on the principle of homogeneity between labels and data clusters. According to this principle the label distribution is consistent and coherent with the distribution of data. In that sense, estimation of data clusters can provide some hints about the posterior label distribution. Therefore, features that are relevant to labels are also relevant to data distribution and, consequently to clusters. The results show that information retrieved from clusters can improve the estimation of feature relevance and of feature selection tasks, specially when the number of labeled data is too small and the unlabeled data is numerous.

References

[1] Michel Verleysen. *Learning high-dimensional data*, pages 141–162. IOS Press, Amsterdam, 2003.

[2] Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lotfi A. Zadeh. *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[3] Kenji Kira and Larry A. Rendell. A practical approach to feature selection. In *ML92: Proceedings of the ninth international workshop on Machine learning*, pages 249–256, San Francisco, CA, USA, 1992. Morgan Kaufmann Publishers Inc.

[4] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical recipes in C (2nd ed.): the art of scientific computing*. Cambridge University Press, New York, NY, USA, 1992.

[5] Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals Eugen.*, 7:179–188, 1936.

[6] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 1991.

[7] C. Krier, D. François, F. Rossi, and M. Verleysen. Feature clustering and mutual information for the selection of variables in spectral data. *Neural Networks*, pages 25–27, 2007.

[8] Pabitra Mitra, C. A. Murthy, and Sankar K. Pal. Unsupervised feature selection using feature similarity. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(3):301–312, 2002.

[9] E. Llobet, O. Gualdron, J. Brezmes, X. Vilanova, and X. Correig. An unsupervised dimensionality-reduction technique. In *Sensors, 2005 IEEE*, 30 2005.

[10] Vanessa Gómez-Verdejo, Michel Verleysen, and Jérôme Fleury. Information-theoretic feature selection for functional data classification. *Neurocomput.*, 72:3580–3589, October 2009.

[11] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, March 1982.

[12] Jerome H. Friedman. Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1):1–67, 1991.