ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

Risk Estimation and Feature Selection

Gauthier Doquire, Benoît Frénay and Michel Verleysen *

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

Abstract.
For classification problems, the risk is often the criterion to be eventually minimised. It can thus naturally be used to assess the quality of feature subsets in feature selection. However, in practice, the probability of error is often unknown and must be estimated. Also, mutual information is often used as a criterion to assess the quality of feature subsets, since it can be seen as an imperfect proxy for the risk and can be reliably estimated. In this paper, two different ways to estimate the risk using the Kozachenko-Leonenko probability density estimator are proposed. The resulting estimators are compared on feature selection problems with a mutual information estimator based on the same density estimator. Along the line of our previous works, experiments show that using an estimator of either the risk or the mutual information give similar results.

1 Introduction

In classification, model performances are usually assessed by the risk or, equivalently, the probability of error. In the context of feature selection, this criterion can e.g. be used to select the best subset of features, for a given number of features. However, the risk is usually not available and has to be estimated from training data. Risk estimation has been tackled in different works [1, 2, 3], which mostly rely on discretising features [4] or on counting errors made by a k nearest-neighbours classifier [5]. Alternatively, mutual information can also be used instead, since it is strongly related to the risk [6, 7]. Based upon the Kozachenko-Leonenko density estimator [8], the variant of the Kraskov estimator [9] proposed by Gomez et al. [10] can be used in classification. Using mutual information gives good results in feature selection [11], even if maximising the mutual information is not always equivalent to minimising the risk [12, 13]. In line with [12, 13], this paper tackles direct risk estimation in a way which allows a fair comparison between risk and mutual information in feature selection.

In this paper, it is proposed to use the Kozachenko-Leonenko estimator [8] to estimate the risk in two different ways. These two estimators and the mutual information estimator of Gomez et al. [10] are compared on feature selection problems. The goal of this paper is to assess whether it is interesting to directly estimate the risk instead of using mutual information and how the risk should be estimated. Since the Kozachenko-Leonenko probability density estimator is at the heart of these three estimators, it allows a fair comparison.

This paper is organised as follows. Section 2 reviews the literature on risk estimation and discusses the use of mutual information as a proxy to the risk.

[*] Gauthier Doquire is funded by a Belgian F.R.I.A grant.

Section 3 proposes two new estimators based on the Kozachenko-Leonenko estimator and discusses a mutual information estimator for classification. These three estimators are compared in Section 4 and Section 5 concludes the paper.

2 Risk and Mutual Information in Feature Selection

Given the random variables $X \in \Re^d$ and $Y \in \mathcal{Y}$ corresponding to the associated class, the classification risk [14] for a given classifier $f : \Re^d \to \mathcal{Y}$ is defined as

$$R(f) = \mathop{\mathbb{E}}_{X,Y} \left[\mathbb{I} \left[y \neq f(x) \right] \right]. \tag{1}$$

where x and y are the values taken by X and Y and $\mathbb{I}[.]$ is the indicator function. The Bayes risk is the optimal risk which can be achieved, i.e.

$$R^* = \min_f R(f) = \mathop{\mathbb{E}}_{X} \left[1 - \max_{y \in \mathcal{Y}} p_{Y|X}(y|x) \right] \tag{2}$$

where $P_{Y|X}$ is the conditional distribution of Y given X. In the above equation, the label $y_{\max}$ which maximises $p_{Y|X}(y|x)$ for a given x is called the Bayes decision. In the rest of this paper, the risk always refers to the Bayes risk, since we are interested in selecting features which lead to the best possible classification performances, i.e. when they are used by an optimal classifier.

The idea of feature selection through risk estimation is not new and dates back to [1, 2]. These papers are based on rectangular Parzen density estimation and require the features to first be discretised, which leads to a loss of information. Moreover, each possible combination of discretised feature values has to be considered, which is not tractable for high-dimensional datasets. Feature discretisation is also needed in [4] which focuses on cancer classification problems. In [3], binary classification problems are tackled through Parzen or k-NN density estimation procedures. Related works also include [5] which counts the number of mistakes made by a weighted 1-NN classifier and [15] which establishes relationships between risk minimisation and the well-known Relief algorithm. Contrarily to the risk estimators reviewed above, those proposed in this paper are able to deal with continuous features and multi-label classification problems.

Instead of the risk, mutual information (MI) has often been used as a feature selection criterion. MI is a symetrical quantity measuring the amount of information that two variables carry about each other. It is formally defined as

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

where

$$H(X) = -\int_X p_X(x) \log p_X(x) dx \tag{4}$$

is the entropy of the continuous random variable X and

$$H(X|Y) = \sum_{y \in \mathcal{Y}} p_Y(y) H(X|Y = y) \tag{5}$$

is the conditional entropy of X given Y is known (for Y assumed to be discrete) [16]. One of the main reasons for the use of MI in feature selection is the existence of an upper and a lower bound on the Bayes risk R^* as a function of the conditional entropy (and thus equivalently of the MI) [6, 7]. However, as demonstrated in [12, 13], MI is not an ideal proxy for mutual information in feature selection. Indeed, in some specific situations, a feature subset having a higher MI with the class labels than another one could actually lead to a higher risk. MI is thus not always optimal from the risk point of view.

3 Using the Kozachenko-Leonenko Estimator

The Kozachenko-Leonenko estimator [8] is a nearest neighbours density estimator which can e.g. be used to estimate mutual information [9, 10]; it assumes that p_X remains constant in a small hypersphere with diameter $\epsilon_k(i)$ containing exactly the k nearest neighbours of the ith sample. Using this hypothesis, Kozachenko and Leonenko obtain the following estimate

$$\log \hat{p}_X(x_i) = \psi(k) - \psi(n) - \log c_d - d \log \epsilon_k(i) \tag{6}$$

where ψ is the digamma function and c_d is the volume of the d-dimensional unit hypersphere. The Kozachenko-Leonenko estimator can been used to estimate mutual information [9, 10], since one can write

$$\hat{I}(X;Y) = \hat{H}(X) - \sum_{y \in \mathcal{Y}} \hat{p}_Y(y) \hat{H}(X|Y=y); \tag{7}$$

using the density estimator defined in Equation (6), one eventually obtains

$$\hat{I}(X;Y) = \psi(n) - \frac{1}{n} \sum_{y \in \mathcal{Y}} n_y \psi(n_y) + \frac{d}{n} \left[\sum_{i=1}^{n} \log \epsilon_k(i) - \sum_{y \in \mathcal{Y}} \sum_{i|y_i = y} \log \epsilon_k(i|y) \right] \tag{8}$$

where n_y is the number of samples which belong to class y and $\epsilon_k(i|y)$ is the diameter of the hypersphere containing the k nearest neighbours in that class.

This paper proposes to estimate the risk using the Kozachenko-Leonenko estimator. Indeed, Bayes' rule allows one to obtain the estimate

$$\hat{p}_{Y|X}(y|x) = \frac{\hat{p}_{X|Y}(x|y)\hat{p}_Y(y)}{\sum_{y \in \mathcal{Y}} \hat{p}_{X|Y}(x|y)\hat{p}_Y(y)}, \tag{9}$$

which can in turn be used to estimate the risk in two possible ways.

Firstly, one can simply count misclassifications using Equation (9), i.e. use

$$\widehat{R^*} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I} \left[y_i \neq \arg\max_{y \in \mathcal{Y}} \hat{p}_{Y|X}(y|x) \right] \tag{10}$$

which is an empirical estimator [17] of the true risk and is very similar to what is commonly used to estimate the risk of classifiers on test instances.

Secondly, one can also rely on the alternative empirical estimator of the risk

$$\widehat{R^*} = \frac{1}{n}\sum_{i=1}^{n}\left[1 - \max_{y\in\mathcal{Y}} \hat{p}_{Y|X}(y|x_i)\right]. \tag{11}$$

The main difference between the estimators (10) and (11) is that the former uses the training labels, whereas the latter uses the estimated class memberships.

The two risk estimators discussed in this section are compared for feature selection in the rest of this paper. The estimators (10) and (11) are similar to the approaches used e.g. in [1, 2, 3, 4, 5], but they rely on the Kozachenko-Leonenko estimator which (i) is an actual density estimator contrarily to some k-neighbours estimators, (ii) gives good results in feature selection [11] and (iii) can deal with high-dimensional data. Moreover, using the Kozachenko-Leonenko estimator for the estimators (8), (10) and (11) allows a fair comparison in Section 4. Notice that using Equation (11) is more costly than Equations (8) and (10), since $n|\mathcal{Y}|$ conditional probabilities have to be estimated in the former case, whereas only n conditional probabilities are needed in the latter case.

4 Experiments

This section compares the three quantities[1] introduced in Section 3, i.e. the mutual information (8) and the two risk estimators (10) and (11), as feature selection criteria. Feature selection has consequently been carried out using these three criteria with a greedy backward search procedure. Backward search starts with all features and recursively eliminates the one whose removal leads to the highest value of mutual information or to the lowest value of risk, according to the considered criterion. While other procedures such as the forward search could be used instead, it has been suggested in [12, 13] that with the mutual information, backward procedures are expected to produce better results.

The criterion of comparison is the balanced classification rate (the class-mean of the percentage of the samples of a particular class correctly classified) of a 1-nearest neighbor classifier, as a function of the number of selected features, obtained on a test set independent of the training set. The 1-NN classifier has been chosen for both its simplicity and its sensitivity to irrelevant features. Indeed, it gives the same weight to each feature and is not able to perform any kind of embedded feature selection. The results have been obtained through a 10-fold cross-test procedure. To avoid any problem in the determination of the nearest neighbours in the MI or risk estimators, a small random zero-mean Gaussian noise with variance 10^{-3} has been added to the features of each training set before the feature selection process. The noisy datasets are only used for feature selection, while the noise-free datasets are considered for classification.

Figure 1 shows the performances of the three approaches on 6 datasets from the UCI repository [18]. As it can be seen, the results obtained with the three methods are quite similar. Mutual information (8) and error counting (10)

[1] MATLAB and Python implementations are available at `http://www.ucl.ac.be/mlg`.

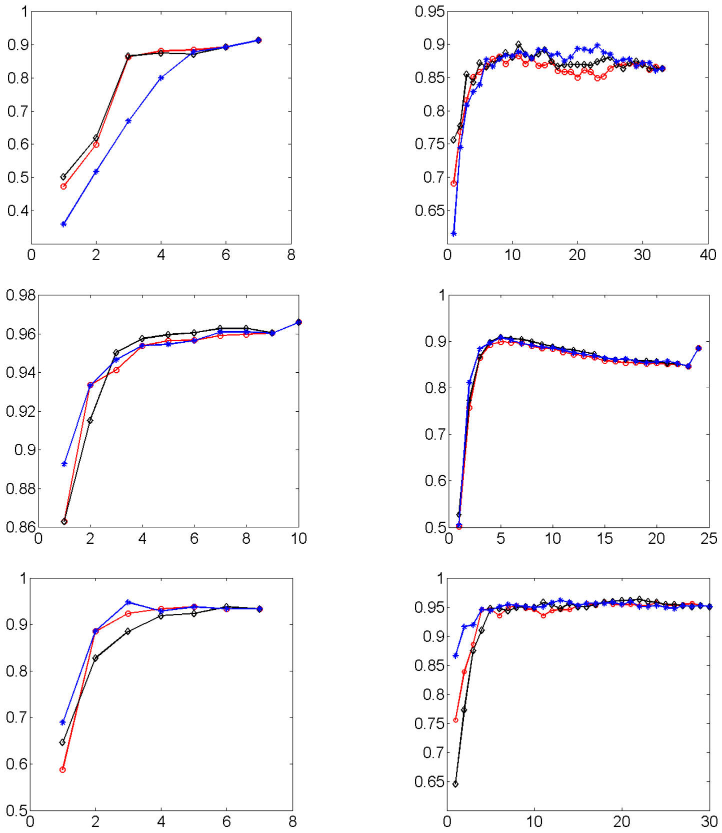

Figure 1: Balanced classification rate of a 1-nearest neighbour classifier as function of the number of selected features obtained with the mutual information (8) (o), the misclassification count (10) ($\diamond$) and the direct risk estimation (11) (*).

perform better on the Ecoli dataset (Fig. 1(a)), while risk estimation (11) is slightly better on the Ionosphere (Fig. 1(b)) and Seeds (Fig. 1(e)) datasets. Performances are equivalent on the other datasets. To asses the significance of the results, a two sample test of means has been carried out, following prescriptions in [19]. MI performances are significantly better for the first three feature subsets for the Ecoli Dataset. No other significative differences can be observed, except for very small features subsets of one or two features in some datasets. It is worth noting that the error counting (10) seems sufficient to get good results.

5 Conclusion

This paper proposes two estimation procedures for the Bayes classification risk using the Kozachenko-Leonenko density estimator. The risk estimators do not

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

require any feature discretisation and can deal with multi-class problems. The interest of the proposed estimators is illustrated in a feature selection context, where their performances are shown to be comparable to the ones of the mutual information criterion estimated based on the same entropy estimator. This observation is in good agreement with our previous work and shows again the strong relationships existing between these two criteria. Besides feature selection, the proposed risk estimators could as well be used in another area; for instance they could easily be applied to instance selection in active learning.

References

[1] P.J. Min. A non-parametric method for feature selection. In *Adaptive Processes, 1968. Seventh Symposium on*, volume 7, page 34, 1968.

[2] K. S. Fu, P. J. Min, and T. J. Li. Feature selection in pattern recognition. *IEEE T. Syst. Man Cyb.*, 6:33–38, 1970.

[3] Keinosuke Fukunaga and Donald M. Hummels. Bayes error estimation using parzen and k-nn procedures. *IEEE T. Pattern Anal.*, 9(5):634 –643, 1987.

[4] Jian Li, Jin-Mao Wei, Tian Yu, and Hai-Wei Zhang. Feature selection based on bayes minimum error probability. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on*, pages 706 –710, 2012.

[5] Peng-Fei Zhu, Tian-Hang Meng, Yun-Long Zhao, Rui-Xian Ma, and Qing-Hua Hu. Feature selection via minimizing nearest neighbor classification error. In *Machine Learning and Cybernetics (ICMLC), 2010 International Conference on*, pages 506 –511, 2010.

[6] R. Fano. *Transmission of Information: A Statistical Theory of Communications.* The MIT Press, Cambridge, MA, 1961.

[7] M. E. Hellman and J. Raviv. Probability of error, equivocation and the chernoff bound. *IEEE T. Inform. Theory*, 16:368–372, 1970.

[8] L. F. Kozachenko and N. Leonenko. Sample estimate of the entropy of a random vector. *Problems Inform. Transmission*, 23:95–101, 1987.

[9] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Phys. Rev. E*, 69(6):066138, 2004.

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

[11] F. Rossi, A. Lendasse, D. Francois, V. Wertz, and M. Verleysen. Mutual Information for the Selection of Relevant Variables in Spectrometric Nonlinear Modelling. *Chemometr. Intell. Lab.*, 80:215–226, 2006.

[12] B. Frénay, G. Doquire, and M. Verleysen. Theoretical and empirical study on the potential inadequacy of mutual information for feature selection in classification. *Accepted for publication in the Special Issue for ESANN 2012 of Neurocomputing.*

[13] B. Frénay, G. Doquire, and M. Verleysen. On the potential inadequacy of mutual information for feature selection. In *Proceeding of ESANN*, 2012.

[14] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd Edition).* Wiley-Interscience, 2001.

[15] Shuang Hong Yang and Bao-Gang Hu. Discriminative feature selection by nonparametric bayes error minimization. *IEEE T. Knowl. Data En.*, 24(8):1422 –1434, 2012.

[16] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, 1991.

[17] B. Schölkopf and A.J. Smola. *Learning With Kernels: Support Vector Machines, Regularization, Optimization and Beyond.* Mit Press, 2002.

[18] D.J. Newman A. Asuncion. UCI machine learning repository, 2007.

[19] R.H. Riffenburgh. *Statistics in Medicine.* Academic Press, 2012.