Supervised Nonparametric Information Theoretic Classification

Cédric Archambeau[†], Torsten Butz[‡], Vlad Popovici[‡], Michel Verleysen[*,†], Jean-Philippe Thiran[‡]

[†] Université catholique de Louvain
Machine Learning Group
Place du Levant, 3
1348 Louvain-la-Neuve, Belgium
{archambeau,verleysen}@dice.ucl.ac.be

[‡] Swiss Federal Institute of Technology (EPFL)
Signal Processing Institute
1015 Lausanne, Switzerland
{torsten.butz,vlad.popovici,jp.thiran}@epfl.ch

Abstract

In this paper, supervised nonparametric information theoretic classification (ITC) is introduced. Its principle relies on the likelihood of a data sample of transmitting its class label to data points in its vicinity. ITC's learning rule is linked to the concept of information potential and the approach is validated on Ripley's data set. We show that ITC may outperform classical classification algorithms, such as probabilistic neural networks and support vector machines.

1. Introduction

Clustering and classification algorithms are fundamental ingredients in pattern recognition, data mining, knowledge discovery and other related fields [4, 1]. Whereas clustering is an unsupervised way of grouping data using a measure of similarity, classification can be seen as its supervised version. Indeed, when the labels of the class prototypes are available, one should use this additional knowledge in order to construct or improve a classification rule. Subsequently, this classification rule may be used to automatically classify new data samples. This rule is therefore often called the generalization rule and the set of labelled prototypes can be seen as the learning data set.

A common technique for clustering and classification involves the estimation of the underlying probability density functions (PDFs) of the classes. Two types of approaches can be considered. In parametric PDF estimation, it is assumed that the data is drawn from a specific density model. The model parameters are then fitted to the data. Unfortunately, an a priori choice of the PDF model is not always suited in practice, as it might provide a false representation of the true PDF. By contrast, we can build nonparametric PDF estimators, as for example the Parzen window estimator [6]. Such techniques do not assume any functional form of the PDF and allow its shape to be entirely determined from the data.

Recently, a clustering evaluation function based on the *Information Potential* was proposed [5] in order to perform nonparametric clustering tasks. In [2] the formalism of information theoretic clustering is generalized by considering nonparametric classification error estimation and by linking it to information theoretical concepts, such as error transmission and distortion.

Within this framework, we propose a classification criterion strongly related to information theory, as it allows to minimize the error transmission. Actually, assigning the optimal class label to a data sample can be seen as minimizing the erroneous transmission of the real class label of that sample. In addition, the proposed classification algorithm is nonparametric. As a consequence, the shape of the true, but unknown underlying PDFs is not enforced a priori.

This paper is organized as follows. In Section 2, we introduce nonparametric information theoretic classification (ITC), which is based on the local class label transmission. Subsequently, in Section 3, we recall the K-nearest neighbor exchange algorithm [4], which is used for optimizing the learning rule. Finally, in Section 4, we show simulation results using Ripley's synthetic data set [8]. We compare the performance of ITC to probabilistic neural networks [9], which are nothing else than Bayesian clas-

[*] M.V. is a Senior Research Associate of the Belgian National Fund for Scientific Research.

This work was partially supported by the European Commission (IST-2000-25145), the Belgian FNRS and the Swiss NCCR IM2, funded by the Swiss National Science Foundation.

0-7695-2128-2/04 $20.00 (C) 2004 IEEE

sification combined to nonparametric class PDF estimation, and support vector machines [10, 3], using Gaussian kernels. All the considered methods are supervised.

2. Information theoretic classification

In this section, the concepts of local and global class label transmission are first defined. Subsequently, an information theoretic learning rule related to error transmission and the information potential is constructed. Finally, a classification criterion based on local class label transmission is introduced.

2.1. Local class label transmission

Consider the d-dimensional feature vector $\mathbf{x} \in \Omega_{\mathbf{X}}$. Let us define the following likelihood function:

$$\mathrm{L}(\mathbf{x}|C, C') = p(\mathbf{x}|C) \cdot p(\mathbf{x}|C'), \tag{1}$$

where $C \in \Omega_C$ and $C' \in \Omega_C$ are the true class label and the estimated class label of $\mathbf{x}$ respectively. The quantity $\mathrm{L}(\mathbf{x}|C, C')$ corresponds to the likelihood of having the data sample $\mathbf{x}$ given C and C'. In other words, it is the likelihood of having $\mathbf{x}$ knowing that its true class label is C and supposing we assign it the label C'. Therefore (1) characterizes the *local transmission* of the label C to the estimated label C' in the vicinity of $\mathbf{x}$.

2.2. Global class label transmission

We may extend the concept of local label transmission by defining the *global transmission* of the class labels as the expected likelihood of transmitting the true class label C to the estimated label C' over the entire feature domain $\Omega_{\mathbf{x}}$:

$$\begin{aligned} T(C, C') &= \int_{\Omega_{\mathbf{x}}} \sum_{C \in \Omega_C} \sum_{C' \in \Omega_C} D(C, C') \\ &\quad \cdot \mathrm{L}(\mathbf{x}|C, C') P(C) P(C') d\mathbf{x}. \end{aligned} \tag{2}$$

In this equation, $P(C)$ and $P(C')$ are the true and estimated class priors, respectively. The function $D(C, C')$ is the similarity function.

In this paper, we focus on the minimization of the error transmission, i.e. we want to maximize the global transmission of true class labels. Besides, we do not assume any class-dependent similarity for simplicity. Therefore, we assume $D(C, C')$ equals 1 when $C = C'$ and 0 otherwise. For other similarity choices, we refer to [2]. The main advantage of the similarity function is that it increases the flexibility of the classification algorithm by making it application-dependent.

2.3. Classification learning rule

The Parzen window estimator [6] is one of the most popular techniques for estimating a PDF nonparametrically. It consists of placing a well-defined kernel function on each data sample and then determining a common kernel width σ. In practice, Gaussian kernels are often used. The estimated PDF is defined as the sum of all Gaussian kernels, multiplied by a normalization factor.

Both conditional PDFs appearing in the expression of the likelihood (1), can be estimated nonparametrically from the data by the Parzen window estimator:

$$\hat{p}(\mathbf{x}|C) = \frac{1}{|S_C|} \sum_{\mathbf{x}_i \in S_C} N\left(\mathbf{x} - \mathbf{x}_i, {\sigma_i}^2\right), \tag{3}$$

$$\hat{p}(\mathbf{x}|C') = \frac{1}{|S_{C'}|} \sum_{\mathbf{x}_j \in S_{C'}} N\left(\mathbf{x} - \mathbf{x}_j, {\sigma_j}^2\right), \tag{4}$$

where the Gaussian kernels are defined as follows:

$$N\left(\mathbf{x} - \mathbf{m}, \sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left[-\frac{\|\mathbf{x} - \mathbf{m}\|^2}{2\sigma^2}\right]. \tag{5}$$

The sets S_C and $S_{C'}$ contain the data samples with class labels C and C' respectively, and $|S_C|$ and $|S_{C'}|$ are their cardinality.

Posing M the number of prototypes, substituting the PDF estimates (3) and (4) into expression (2) of the global class label transmission $T(C, C')$, and replacing the priors $P(C)$ and $P(C')$ by $\frac{|S_C|}{M}$ and $\frac{|S_{C'}|}{M}$ respectively, leads to:

$$\begin{aligned} T(C, C') &= \int_{\Omega_{\mathbf{x}}} \sum_{C \in \Omega_C} \sum_{C' \in \Omega_C} D(C, C') \\ &\quad \cdot \frac{1}{M} \sum_{\mathbf{x}_i \in S_C} N\left(\mathbf{x} - \mathbf{x}_i, {\sigma_i}^2\right) \\ &\quad \cdot \frac{1}{M} \sum_{\mathbf{x}_j \in S_{C'}} N\left(\mathbf{x} - \mathbf{x}_j, {\sigma_j}^2\right) d\mathbf{x}. \end{aligned}$$

After integration, we can write the following *information theoretic learning rule*:

$$\hat{C}' = \arg\max_{C'} \sum_{C \in \Omega_C} \sum_{C' \in \Omega_C} D(C, C') V(S_C, S_{C'}), \tag{6}$$

where

$$V(S_C, S_{C'}) = \frac{1}{M^2} \sum_{\mathbf{x}_i \in S_C} \sum_{\mathbf{x}_j \in S_{C'}} N\left(\mathbf{x}_i - \mathbf{x}_j, {\sigma_i}^2 + {\sigma_j}^2\right).$$

The quantity $V(S_C, S_{C'})$ is the *information potential* [5] and is closely related to Renyi's entropy [7].

Note that in classical nonparametric PDF estimation, the kernel width enforces a smooth density model. However, within the information theoretic framework, it can be interpreted in a slightly different way. Here, the kernel width controls the extent of the local domain around each labelled prototype, rather than the degree of overlapping of the kernels. In other words, it regulates the region of label transmission of each data point.

2.4. Classification criterion

By solving the maximization problem posed by the information theoretic learning rule, we can automatically classify the class prototypes. Yet, whenever one wants to classify new data samples the whole optimization procedure of the classification algorithm should be repeated, which is not suited in practice. In addition, when the kernel widths are not chosen properly, we might easily overfit the data, resulting in a classification scheme not representative of the true classes. Therefore, class prototypes are used for learning the classification task, while test samples are kept for evaluating its performance. The optimal kernel widths can then be selected according to the best generalization performance. It remains that, in order to test the algorithm, we need a valid classification rule.

Consider again the information theoretic learning rule. By applying this rule, we maximize the global transmission of the correct class labels over the entire feature space domain. As a result, we assign the most probable class labels $\hat{C}'_l$ to the class prototypes $\{\mathbf{x}_l\}_{l=1}^{M_l}$, according to their neighborhood. By analogy to (2) where $C = \hat{C}'_l$ and $C' = \hat{C}'_l$, we can classify the test samples $\{\mathbf{x}_t\}_1^{M_{t=1}}$ according to the likelihood of their local estimated class label transmission, weighted by the class priors:

$$\hat{C}'_t = \arg\max_c \hat{L}(\mathbf{x}_t | c = \hat{C}'_l, c = \hat{C}'_l) P(\hat{C}'_l) P(\hat{C}'_l). \quad (7)$$

Although the classification learning rule and the classification criterion are not identical, it was found experimentally that both are in agreement when the kernel widths are not overestimated.

3. K-nearest neighbor exchange algorithm

The learning rule (6) can be optimized by the K-nearest neighbor exchange algorithm [4]. This optimization procedure allows escaping from local minima and saves computations by labelling data groups instead of the individual data points. As the contribution to the global label transmission of data samples lying closely to each other and belonging to the same classes is large, the classification algorithm assigns a common class label to samples in the same neighborhood. Meanwhile, class boundaries are positioned along regions of low sample density.

The optimization algorithm iterates over the following steps:

1. Initialize randomly a class label to each sample of the learning data set $\{\mathbf{x}_l\}_{l=1}^{M_l}$ and choose an initial group size $K = K_0$;

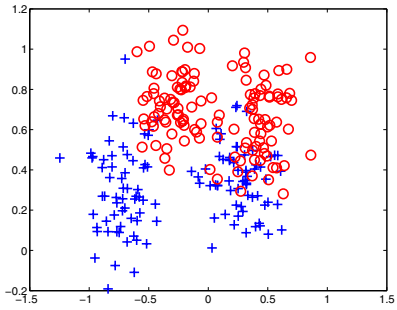

Figure 1. Ripley's synthetic learning data set. The true class of the prototypes are denoted by '$\circ$' and '$+$' respectively.

2. Create M_l groups by searching for each sample $\mathbf{x}_l$ the K-nearest neighbors with the same class label;
3. Remove identical groups, resulting in P_l groups;
4. Repeat until no improvement:
   - For each group p, change its class label to any class label and record improvement if any;
   - If in the previous step, any improvement was recorded, permute randomly the group indices;
5. If $K > 1$, divide the group size K by two, and go to step 2;
6. End of algorithm.

In practice, the initial group size K_0 can be initialized for example as $\frac{M_l}{|\Omega_C|}$.

4. Results and discussion

The proposed supervised information theoretic classification algorithm is validated on Ripley's data set [8], shown in Figure 1. It is a synthetic data set, composed of a learning set of size $M_l = 250$ and a test set of size $M_t = 1000$. Each class is a mixture of two Gaussian distributions. The two equally sized classes are strongly overlapping.

In this paper, the performances of information theoretic classification (ITC), probabilistic neural networks (PNNs) and support vector machines (SVMs) are compared. We consider PNNs using Gaussian activation functions. Basically, such PNNs perform Bayesian classification after estimating the PDF of each class by means of the class prototypes [9]. The class densities are estimated nonparametrically by Parzen. SVMs are supervised classification technique based on Vapnik's statistical learning

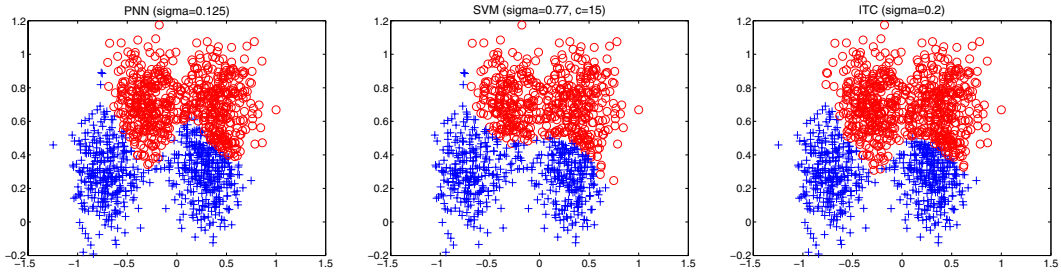

Figure 2. Classification of Ripley's synthetic test set. The left figure shows the classification by PNN, the one in the middle by SVM and the figure on the right by ITC.

theory [10, 3]. The goal is to find a large margin separation hyper-plane between the two classes. As the classes are usually not linearly separable in the input space, data is projected into a high-dimensional space through the so-called kernel-trick, where the linear classification can be done. Again, we consider Gaussian kernels.

In Figure 2, the classification result of Ripley's test set is shown, using PNN, SVM and ITC. In the latter, we have taken $\sigma_i = \sigma_j = \sigma_{ITC}$ for simplicity. By performing an exhaustive search, it was found that the optimal kernel width for PNN is $\sigma_{PNN} = 0.125$ and for ITC $\sigma_{ITC} = 0.2$. As noted in Section 2.3, this difference can be explained by the information theoretical interpretation given to the kernel width in the context of ITC, i.e. it controls the region of label transmission around each prototype. The optimal kernel width for SVM is $\sigma_{SVM} = 0.75$ and the regularization factor $c = 15$. For each classification algorithm, the average classification error E was computed. For PNN we have obtained $E_{PNN} = 9.3\%$, for SVM $E_{SVM} = 9.7\%$, and for ITC $E_{ITC} = 9.1\%$.

PNN and ITC perform better than SVM. When comparing PNN and ITC, one can see that ITC performs slightly better than PNN. This can be explained as follows. The effect of learning the classification task by maximizing the global class label transmission is the removal of class outliers when estimating the class PDFs. We denote by the term 'class outlier' a data sample located within a foreign class. Hence, the class transitions are steeper, reducing the uncertainty on the location of the class boundaries. The classification result is therefore less sensitive to atypical data samples, as they do not contribute to the PDF estimation of their respective classes.

5. Conclusion

In this paper, supervised nonparametric information theoretic classification was introduced. The resulting classification algorithm exploits the likelihood of each class prototype of transmitting its class label to a data point located in its vicinity. Based on this principle, an information theoretic learning rule and a related classification rule were proposed. The former is linked to the concept of information potential. The relevance of the approach was demonstrated by simulation. Indeed, information theoretic classification outperforms probabilistic neural networks and support vector machines on Ripley's well-known synthetic data set.

References

[1] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford university press, 1995.

[2] T. Butz. *From Error Probability to Information Theoretic Signal and Image Processing*. Ph.D. dissertation, Swiss Federal Institute of Technology (EPFL), Switzerland, 2003.

[3] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge university press, 2000.

[4] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley, 1973.

[5] E. Gokcay and J. C. Principe. Information theoretic clustering. *IEEE Patt. Anal. Mach. Int.*, 24(2):158–171, 2002.

[6] E. Parzen. On estimation of a probability density function and mode. *Ann. Math. Stat.*, 33:1065–1076, 1962.

[7] A. Renyi. On measures of entropy and information. In *Proc. of 4th Berkeley Symp. on Prob. Math. Stat.*, volume 1, pages 547–561, 1960.

[8] B. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

[9] D. F. Specht. Probabilistic neural networks. *Neural Networks*, 3:109–118, 1990.

[10] B. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.