Monografías del Seminario Matemático García de Galdeano **33**, 331–340 (2006)

FEATURE SELECTION USING MUTUAL INFORMATION AND NEURAL NETWORKS

O. Valenzuela, I. Rojas, L. J. Herrera, A. Guillén, F. Rojas, L. Marquez and M. Pasadas

Abstract. Reducing the dimensionality of the raw input variable space is an important step in pattern recognition and functional approximation tasks often determined by practical feasibility. The purpose of this study was to investigate an information theoretic approach to feature selection. We will use mutual information (MI) as a pre-processing step for artificial neural networks. The reasons why mutual information is not in wider use currently (except between two scalar variables) lie in computational difficulties. The probability density functions of the variables are required, and MI involves numerical integration of functions of those, which leads to a high computational complexity. Because of the difficulty in directly implementing the maximal dependency condition, we first derive an equivalent form, called minimal redundancy maximal relevance criterion, for first order incremental feature selection.

The feature selection methodology is hybridized with three different classification and universal function approximation paradigms: Multilayer Perceptron, Radial Basis Function and Support Vector Machine. We perform extensive experimental comparison of the proposed hybrid algorithm for different problem: breast cancer classification, diabetes in Pima Indians and arrhythmia.

§1. Introduction

Feature selection is the process of choosing a subset of features relevant to a particular application. During the selection process, a decision criterion is used to remove irrelevant or redundant features. Extensive research has led researchers to appreciate the importance of feature selection when developing Computer-Assisted Diagnostic (CAD) tools. Optimized feature selection reduces data dimensionality and potentially removes noise, thus resulting in CAD tools that are not only more accurate but also more robust. Several CAD applications have demonstrated the positive impact of optimized feature selection. The most popular feature selection algorithm utilized in CAD is the stepwise linear discriminant analysis, borrowed from linear statistics. It is designed to reduce the dimensionality of the feature vector by selecting in stepwise fashion the features that maximize the linear separability of the output classes.

The approach is based on the Mutual Information (MI) concept. MI measures the general dependence of random variables without making any assumptions about the nature of their underlying relationships. Consequently, MI can potentially offer some advantages over feature selection techniques that focus only on the linear relationships of variables. MI accounts for higher-order statistics, not just for second order. In addition, it can also be used as the

basis for non-linear transforms. MI also bounds the optimal Bayes error rate. The reasons why mutual information is not in wider use currently (except between two scalar variables) lie in computational difficulties. The probability density functions of the variables are required, and MI involves numerical integration of functions of those, which leads to a high computational complexity. Given two random variables x and y, their mutual information is defined in terms of their probabilistic density functions $p(x)$, $p(y)$ and $p(x,y)$.

The feature selection methodology is hybridized with three different classification and universal function approximation paradigms: Multilayer Perceptron, Radial Basis Function and Support Vector Machine. Multilayer perceptrons (MLPs) are feedforward neural networks trained with the standard backpropagation algorithm. They are supervised networks so they require a desired response to be trained. They have been shown to approximate the performance of optimal statistical classifiers in difficult problems. Radial basis function (RBF) networks have a static Gaussian function as the nonlinearity for the hidden layer processing elements. The Gaussian function responds only to a small region of the input space where the Gaussian is centred. The Support Vector Machine (SVM) is one of the most successful learning algorithms proposed in recent years. One of the main advantages of the SVM over other networks is that its training is performed through the solution of a linearly constrained convex quadratic programming problem: therefore, only a global (not necessarily unique) minimum exists and, given a fixed tolerance, efficient algorithms can find an approximate solution in a finite number of steps.

We perform extensive experimental comparison of the proposed hybrid algorithm for different problem: function approximation or regression problem, classification (breast cancer, diabetes and thyroid classification) and time series forecasting. This different technique, in order to obtain a self-containing paper, will be explained in the following section. The remainder of this paper is organized as follows: In Section 2, we will present a feature selection algorithm, which selects the more important variables to the artificial intelligent paradigms (different kinds of neural networks). Section 3 presents a brief resume of the main intelligent computation techniques used in this paper. The benchmark problems selected to check the behaviour of the proposed algorithm are detailed in Section 4, whereas the simulation results are given in Section 5. Finally, Section 6 presents the main conclusions.

§2. Feature selection: Mutual information and Minimal Redundancy Maximal Relevance Criterion

Mutual information is a good indicator of relevance between variables, and has been used as a measure in several feature selection algorithms. However, calculating the mutual information is difficult, and the performance of a feature selection algorithm depends on the accuracy of the mutual information [2]. In this section we use a method of calculating mutual information between input and class variables based on the Parzen window [3]. In the breast cancer classification problem presented in this paper, the mutual information between the input features X and the class C can be represented as follows:

$$I(X;C) = H(C) - H(C|X). \tag{1}$$

In this equation, because the class is a discrete variable, the entropy of the class variable $H(C)$ can be calculated as:

$$H(X) = -\sum_{x \in \overline{X}} p(x) \log p(x). \tag{2}$$

Where the discrete variable X has $\overline{X}$ alphabets and the probability density function (pdf) is $p(x) = \Pr\{X = x\}$, $x \in \overline{X}$. The hard problem is to compute the conditional entropy:

$$H(C|X) = -\int_X p(x) \sum_{c=1}^{N} p(c|x) \log p(c|x)\, dx. \tag{3}$$

Because it is not easy to estimate $p(c|x)$, being N the number of classes. By the Bayesian Rule, the conditional probability $p(c|x)$ can be written as

$$p(c|x) = \frac{p(x|c)p(c)}{p(x)}. \tag{4}$$

Using the Parzen Window method [3], is it possible to estimate the conditional pdf $\widehat{p}(x|c)$, an using the estimate the conditional pdf $\widehat{p}(x|c)$, the conditional probability is

$$\widehat{p}(c|x) = \frac{\sum_{i \in I_c} \exp\left(-\frac{(x-x_i)^T \sum^{-1} (x-x_i)}{2h^2}\right)}{\sum_{k=1}^{N} \sum_{i \in I_k} \exp\left(-\frac{(x-x_i)^T \sum^{-1} (x-x_i)}{2h^2}\right)}. \tag{5}$$

Therefore, using n training samples, the conditional entropy, assuming that each sample has the same probability is

$$\widehat{H}(C|X) = -\sum_{j=1}^{n} \frac{1}{n} \sum_{c=1}^{N} \widehat{p}(c|x_j) \log \widehat{p}(c|x_j). \tag{6}$$

Where x_j is the jth sample of the training data. Therefore, given the input data D tabled as N samples and M features $X = \{x_i,\ i = 1, \ldots, M\}$, and the target classification variable c, feature selection problem is to find from the M-dimensional observation space, RM, a subspace of m features, Rm, that "optimally" characterizes c. Given a condition defining the "optimal characterization", a search algorithm is needed to find the best subspace. Because the total number of subspaces is $2M$, and the number of subspaces with dimensions no larger than m is $\sum_{l=1}^{m} \binom{M}{l}$, it is hard to search the feature subspace exhaustively. Alternatively, many sequential-search based approximation schemes have been proposed, including best individual features, sequential forward search, sequential forward floating search, etc (see [6, 10] for detailed comparison.). Max-Relevance is to search features satisfying the following equation:

$$\max D(S,c),\ D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i;c), \tag{7}$$

and satisfying the *minimal redundancy* (Min-Redundancy) condition expressed as

$$\max R(S),\ R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i, x_j). \tag{8}$$

The criterion combining the above two constraints is called *minimal-redundancy-maximal relevance* (mRMR) [1]. This is the criterion that has been used in the proposed paper.

§3. Paradigms from Artificial Intelligence

3.1. Multilayer Perceptron

The essential idea of a feed-forward ANN is that each neuron outputs a smoothly rising function of the sum of its weighted inputs, e.g. $F(a*w1+b*w2+c*w3)$. The weighted sum in the brackets also equals the scalar product of the data and weight vectors, $d.w$, which in turn equals $D*W*\cos(\text{angle between } d \text{ and } w)$, where d is (a,b,c) and w is $(w1,w2,w3)$. This is worth knowing because it demonstrates that the neuron is effectively detecting the feature w.

When networks are built using three layers, the middle layer is called "the hidden layer". Each unit within the hidden layer may act as a feature detector, responding to features appearing within the input data. This neural network structure is usually called a multi-layer perceptron (MLP). The MLP architecture is the most popular in real world applications. Each layer is fully connected to the next. The results of many authors working with MLPs over many years, including the authors of this paper, tends towards the optimistic view that simple monotonic functions like the sigmoid or the $\tan(h)$ are widely applicable.

The connection weights of a neural network need to be discovered for a correct solution to any problem and this is called training. Where the interpretation of a set of (training) data is known, it is appropriate to use supervised learning; whereas if there are no available interpretations for the data, supervised learning cannot be used and unsupervised learning can be useful. One of the most popular algorithms to adapt the weights of a multilayer Perceptron is based on the Generalized Delta Rule (GDR) [7]. With the GDR, small updates are made to each weight such that the updates are proportional to the backpropagated error term at the node. The update rule for the GDR is

$$\Delta w_{ij}(t) = \eta \delta_j x_{ji}. \tag{9}$$

The problem with the gradient descent approach is in choosing η: we'd like it to be small, to ensure we make progress moving downhill, but we'd also like it to be big so that we converge to the solution quickly. Solutions to this dilemma include varying η in response to how well previous steps worked, or iteratively finding the minimum in the direction of the gradient (i.e. "minimization").

The Levenberg-Marquardt method takes a different approach, by recognizing that the curvature of the function gives us some information about how far to move along the slope of the function. This has been the method use in the present paper.

3.2. Radial Basis Function

Broomhead and Lowe[4] were the first to exploit the use of radial basis functions in the design of neural networks and to show how RBF networks model nonlinear relationships and implement generalization or interpolation between data points. Radial basis function (RBF) neural networks consist of neurons which are locally tuned. An RBF network can be regarded as a feedforward neural network like a multilayer perceptron (MLP) with a single layer of hidden units, whose responses are the output of radial basis functions. Both MLP and RBF architectures have the capability of approximating mathematically well-behaved

functions to any desired degree of accuracy, provided there are enough nodes in the network. Theoretical frameworks [5] show that RBF networks with Gaussian units having different kernel widths are universal approximators with respect to the uniform norm for continuous functions defined on a compact convex set. Experience shows that RBF response networks often give approximations that are as good as or better than MLP, with one or two orders of magnitude less training effort. The output of the networks is defined as the linear combination of the radial basis function layer, as follows:

$$\tilde{F}_{RBF}(X) = \sum_{i=1}^{N} w_i \phi_i(X, C_i, \sigma_i) + \lambda_0, \tag{10}$$

where the radial $\mathbb{R}^n$ basis functions ϕ_i are the nonlinear functions, which depend on the parameters $C_i \in \mathbb{R}^n$ that represent the centre of the basis function and $\sigma_i \in \mathbb{R}^n$, the dilation or scaling factor. The basis function is expressed as

$$\phi_i(X, C_i, \sigma_i) = \phi_i(\|X - C_i\| / \sigma_i), \tag{11}$$

with $\|.\|$ being the norm used. This is the expression of the weighted sum of the radial basis function ($\tilde{F}_{RBF}$).

3.3. Support Vector Machine

We are given a set of N data points $\{x_i, y_i\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^n$ is the ith input data, and $y_i \in \{-1, +1\}$ is the label of the data. The Support Vector Machine (SVM) approach aims at finding a classifier of form [9, 10]:

$$y(x) = \text{sign}\left[\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b\right], \tag{12}$$

where α_i are positive real constants and b is also a real constant, in general, and $K(x_i, x) = \langle \phi(x_i), \phi(x) \rangle$, $\langle ., . \rangle$ is inner product and $\phi(x)$ is the nonlinear map from original space to the high dimensional space. In the high dimensional space, we assume the data can be separated by a linear hyperplane, this will cause:

$$\begin{cases} w^T \phi(x_i) + b \geq 1, & \text{if } y_i = +1, \\ w^T \phi(x_i) + b \leq -1, & \text{if } y_i = -1, \end{cases} \tag{13}$$

which is equivalent to

$$y_i[w^T \phi(x_i) + b] \geq 1, \quad i = 1, \ldots, N. \tag{14}$$

In case of such separating hyperplane does not exist, we introduce a so called slack variable ξ_i such that

$$\begin{cases} y_i[w^T \phi(x_i) + b] \geq 1 - \xi_i, & i = 1, \ldots, N \\ \xi_i \geq 0, & i = 1, \ldots, N. \end{cases} \tag{15}$$

According to the structural risk minimization principle, the risk bound is minimized by the following minimization problem:

$$\min_{w,\xi} J_1(w,\xi) = \frac{1}{2} w^T w + c \sum_{i=1}^{N} \xi_i \tag{16}$$

subject to (15).

§4. Dataset used

In this section we present the medical diagnosis problem and the results of the different approaches presented in the bibliography.

4.1. Wisconsin breast cancer dataset

The presence of a breast mass is an alert sign, but it does not always indicate a malignant cancer. Fine needle aspiration (FNA) of breast masses is a cost-effective, non-traumatic, and mostly non-invasive diagnostic test that obtains information needed to evaluate malignancy.

The Wisconsin breast cancer dataset was obtained from repository of machine learning database University of California, Irvine. This data set has 32 attributes (30 real-valued input features) and 569 instances of which 357 benign and 212 malignant class. However, diagnostic decisions are essentially black boxes, with no explanation as to how they were attained.

Nine visually assessed characteristics of an FNA sample considered relevant for diagnosis were identified, and assigned an integer value between 1 and 10. The measured variables are as follows: 1. Clump Thickness (V_1); 2. Uniformity of Cell Size (V_2); 3. Uniformity of Cell Shape (V_3); 4. Marginal Adhesion (V_4); 5. Single Epithelial Cell Size (V_5); 6. Bare Nuclei (V_6); 7. Bland Chromatin (V_7); 8. Normal Nucleoli (V_8); 9. Mitosis (V_9). The diagnostics in the Wisconsin breast cancer dataset were furnished by specialists in the field. The database itself contains 683 cases, with each entry representing the classification for a certain ensemble of measured values.

4.2. Diabetes in Pima Indians

This database is taken from the UCI repository (see [8]) and is a fairly well know bench mark problem in machine learning. A population of women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona, was tested for diabetes according to World Health Organization criteria. The data were collected by the US National Institute of Diabetes and Digestive and Kidney Diseases. The diagnostic, binary-valued variable investigated is whether the patient shows signs of diabetes according to World Health Organisation criteria. The database contains details of 768 females all of which are older than 21. This was split into a training and test set each containing 384 instances. There are 8 attributes: 1. Number of times pregnant; 2. Plasma glucose concentration; 3. Diastolic blood pressure, 4. Triceps skin fold thickness; 5. 2-Hour serum insulin; 6. Body mass index; 7. Diabetes pedigree function; 8. Age.

4.3. Arrythmia database

In order to assess the ability of techniques considered in this work to deal with incomplete or ambiguous biosignal data from multiple patients in a real-world setting, we use the UCI Arrythmia dataset developed by Guvenir *et. al.* for our simulation experiment. This database contains 279 attributes, 206 of which are linear valued and the rest are nominal. Concerning the study of H. Altay Guvenir: "The aim is to distinguish between the presence and absence of cardiac arrhythmia and to classify it in one of the 16 groups. Class 01 refers to 'normal' ECG classes 02 to 15 refers to different classes of arrhythmia and class 16 refers to the rest of unclassified ones. For the time being, there exists a computer program that makes such a classification. However there are differences between the cardiolog's and the programs classification. Taking the cardiolog's as a gold standard we aim to minimise this difference by means of machine learning tools".

§5. Simulation results

The proposed methodology consist of two different phases: in the first one, the feature selection algorithm is applied in order to obtain the most relevant attributes, in the second phase, a soft-computing paradigm is used for classification purpose. The feature selection algorithm used tries to select a feature that minimizes the redundancy and maximizes the relevance, instead of selecting just independent features. In fact, in real problems, features selected using the feature selection algorithm presented in section 2 will have more or less correlation with each other. However, experiments show that the joint effect of these features can lead to very good classification accuracy. A set of features that are completely independent of each other usually would be less optimal. In the second phase, powerful classifier paradigms, as Multilayer perceptron, Radial Basis Function Neural Networks or Support Vector Machine are used with the most important attributes selected in previous phase. The simulation result, comparing the different paradigms for the three different problems are presented in the figures 1 to 3.

From Figure 1, it is important to note that due the small number of attributes presented in the data base, reducing the number of features contribute to continuously decrease the error rate for all the classifier used. Figure 2 shows the evolution of the error rate and the comparison of feature classification accuracies of different paradigms for the Arrhythmia problem. It is important to note that this benchmark has a big number of attributes for just a small number of instances. Therefore, even if the classification paradigm has more input nodes (big number of features), the error rate is not decreased because overfitting is produced. In fact, the critical issue in developing a neural network or kernel method (as SVM) is generalization: how well will the network make predictions for cases that are not in the training set? Neural network, like other flexible nonlinear estimation methods such as kernel regression and smoothing splines, can suffer from either underfitting or overfitting. A network that is not sufficiently complex can fail to detect fully the signal in a complicated data set, leading to underfitting (analysing Figure 2, this is the behaviour presented with only 1 or 2 features). A network that is too complex may fit the noise, not just the signal, leading to overfitting. Overfitting is especially dangerous because it can easily lead to predictions that are far beyond the range of the training data with many of the common types of NNs. Overfitting can

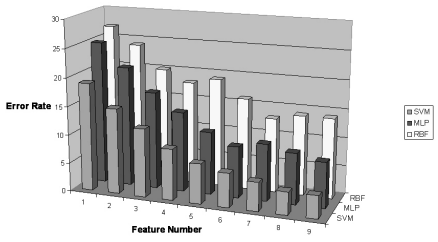

Figure 1: Comparison of feature classification accuracies of different paradigms for the Breast-Cancer problem

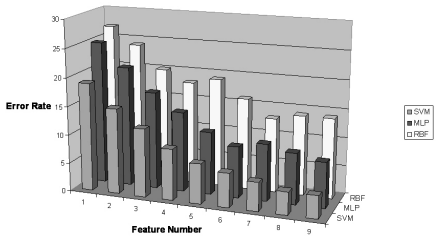

Figure 2: Comparison of feature classification accuracies of different paradigms for the Diabetes problem

also produce wild predictions in multilayer perceptrons even with noise-free data (analysing Figure 2, this is the behaviour presented with classifiers with a big number of features, around 15). For decision-making, the probability output from the classifier is rounded to $1, 2, \ldots, 16$ depending on the classification probability threshold. Examining results in detail, it was observed that misclassifications were mainly concentrated on the area where probabilities are estimated to be between integer number (for example, between 1.4 and 1.7). It is therefore concluded that the classification results around and real number, with the non-integer part around 0.5 have higher risk of misclassification.

This also suggests that if the misclassifications between output result with non-integer part in the interval $[0.4, 0.7]$ could be ignored, the misclassification would be reduced substantially. The randomness of the occurrence of arrhythmic beats suggests that a static analysis, based only on the features of the current beat, might be appropriate. The high intra- and interpatient variability of the beat shape suggests an approach that takes the patient as a reference of himself or herself. However, in clinical domain, it is hard for a physician to build a new model for every patient, especially when we want to monitor a patient' condition in the real-time but have no his/her data to train our classifier and build our model. Therefore, the use of inter-patient data included high noisy component and incomplete information becomes very important in practical domain.

Finally, Figure 3 shows the behaviour of the error rate for the diabetes problem. The evolution of the error is similar to the Breast-Cancer problem.

§6. Conclusions

We have presented a method for feature extraction using as criterion an approximation of the mutual information between features and class labels. This approximation is inspired by the minimal-redundancy-maximal-relevance condition. The purpose of this study was to investigate an information theoretic approach to feature selection for classification and function approximation problems and its hybridization with artificial neural networks. The approach is based on the mutual information (MI) concept. MI measures the general dependence of random variables without making any assumptions about the nature of their underlying relationships. Consequently, MI can potentially offer some advantages over feature selection

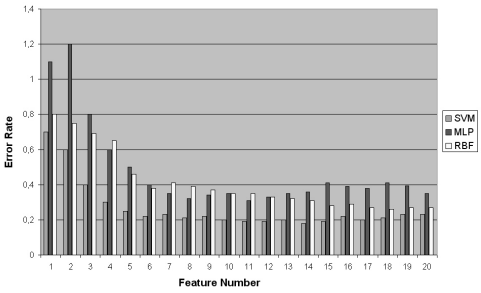

Figure 3: Comparison of feature classification accuracies of different paradigms for the Arrhythmia problem

techniques that focus only on the linear relationships of variables. MI accounts for higher-order statistics, not just for second order. In addition, it can also be used as the basis for non-linear transforms. MI also bounds the optimal Bayes error rate. The reasons why mutual information is not in wider use currently (except between two scalar variables) lie in computational difficulties. The probability density functions of the variables are required, and MI involves numerical integration of functions of those, which leads to a high computational complexity. Given two random variables x and y, their mutual information is defined in terms of their probabilistic density functions $p(x)$, $p(y)$ and $p(x,y)$. We present a theoretical analysis of the minimal-redundancy-maximal-relevance (MRMR) condition and its hybridising with different paradigms from the artificial intelligence, as Multilayer Perceptron, Radial Basis Function and Support Vector Machine. We perform extensive experimental comparison of the proposed hybrid algorithm for different problem: breast cancer classification, Diabetes in Pima Indians and Arrhythmia. The experimental results show that for benchmark problem with a small number of attributes, increasing the number of selected features, the error index decrease. However, in real problem with big number of attributes, although in general more selected features will lead to a smaller classification error, the decrement of error might not be significant for each additional feature, or occasionally there could be fluctuation of classification errors. For example, in Figure 2, the 10th selected feature seemingly has not led to a major reduction of the classification error produced with the first 11th features.

Acknowledgements

This work has been partially supported by the Spanish CICYT Project TIN2004-01419.

References

[1] DING, C., AND PENG, H. C. Minimum redundancy feature selection from microarray gene expression data. *Proceedings of the Computational Systems Bioinformatics (CSB'03)*, 2003.

[2] BABICH, G. A., AND CAMPS, O. I. Weighted Parzen window for pattern classification. *IEEE Trans. Pattern Analysis and Machine Intelligence 18*, 5 (1996), 567–570.

[3] KWAK, N., AND CHOI, C-H. Input feature selection by mutual information based on Parzen window, *IEEE Trans. Pattern Analysis and Machine Intelligence 24*, 12 (2002), 1667–1671.

[4] BROOMHEAD, D. S., AND LOWE, D. Multivariable functional interpolation and adaptive networks, *Comple Syst. 2* (1988), 321–355.

[5] HARTMAN, E., KEELER, J. D., AND KOWALSKI, J. Layered neural networks with gaussian hidden units are universal approximators. *Neural Comput. 2* (1990).

[6] RIVERA RIVAS, A., ORTEGA LOPERA, J., ROJAS, I., AND DEL JESUS, M.Co-evolutionary algorithm for RBF by self-organizing population of neurons. *Lecture Notes in Computer Science 2686*, Springer-Verlag Heidelberg, (2003).

[7] KARAYIANNIS, N. B., AND WEIQUN MI, G. Growing radial basis neural networks: Merging supervised and unsupervised learning with network growth techniques. *IEEE Transaction on Neural Networks 8*, 6 (1997), 1492–1506.

[8] BLAKE, C. L., AND MERZ, C. J. UCI repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science, 1998. `http://www.ics.uci.edu/mlearn/MLRepository.html`.

[9] CRISTIANINI, N., AND SHAWE-TAYLOR, J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[10] SCHLKOPF, B., AND SMOLA, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. MIT Press, 2001.

O. Valenzuela, L. Marquez and M. Pasadas
Department of Applied Mathematics
University of Granada, Spain

I. Rojas, L. J. Herrera, A. Guillén and F. Rojas
Department of Architecture and Computer Technology
University of Granada, Spain