A Powerful Feature Selection approach based on Mutual Information

Ali El Akadi[†], *Abdeljalil El Ouardighi*[††], *and Driss Aboutajdine*[†††]

[†]*GSCM-LRIT, Faculty of Sciences, Mohammed V University, Rabat, Morocco*
[††]*M2CE, GSCM-LRIT, Faculty of Economic Sciences, Hassan I University, Settat, Morocco*
[†††]*GSCM-LRIT, Faculty of Sciences, Mohammed V University, Rabat, Morocco*

Summary

Feature selection aims to reduce the dimensionality of patterns for classificatory analysis by selecting the most informative instead of irrelevant and/or redundant features. In this paper we propose a novel feature selection measure based on mutual information and takes into consideration the interaction between features. The proposed measure is used to determine relevant features from the original feature set for a pattern recognition problem. We use a Support Vector Machine (SVM) classifier to compare the performance of our measure with recently proposed information theoretic criteria. Very good performances are obtained when applying this method on handwritten digital recognition data.

Key words:

Feature selection, Feature interaction, Mutual Information, Interaction Gain, Handwritten Digit recognition

1. Introduction

Feature selection is a very important step in classification since the inclusion of irrelevant and redundant features often degrade the performance of a classification algorithm both in speed and prediction accuracy.

In the case of a pattern recognition problem, the objective of feature selection is to find the smallest subset of features that maximizes the pattern recognition ability. Ideally, this can be achieved by examining all possible subsets and finding the one that satisfies the above criterion. This approach is known as exhaustive feature selection. Even with a moderate number of features, the exhaustive selection is impractical because of its computational requirements. Other feature selection methods were developed to reduce computational complexity by compromising performance.

All feature selection methods need to use an evaluation function together with a search procedure to obtain the optimal feature set. The evaluation function measures how good a specific subset can be in discriminating between classes and can be divided into two main groups: filters and wrappers. Filters measure the relevance of feature subsets independently of any classifier, whereas wrappers use the classifier's performance as the evaluation measure. Search procedures on the other hand, are methods that only consider small portion of all possible subsets. In this paper, our objective is to develop an evaluation function that can be used with any search procedure. We will consider filter evaluation measures because they are faster than wrapper and can handle large datasets [1]. A variety of filter-based measures have already been proposed in the literature. The most popular fall under the following three categories: distance measures, consistency measures and information measures.

This paper will focus on the information measure that is based on the concept of mutual information. The drawback of the recently proposed information measures is that they don't take into consideration the interaction between features. Indeed, a single feature can be considered irrelevant based on its correlation with the class; but when combined with other features, it becomes very relevant. Unintentional removal of these features can result in a loss of useful information and thus may cause poor classification performance. This is studied in [3] as attribute interaction.

We will propose a new information based evaluation function called IGFS (Interaction Gain for Feature Selection) that overcomes the drawbacks of the other functions. Our proposed method will be used to determine the relevant features from the original feature set for a pattern recognition problem and will be compared with three recently proposed information theoretic criteria.

The rest of the paper is organized as the follows: Some information theoretic notions for feature selection and the state of the art about the recently proposed information theoretic criteria for feature selection are addressed in the section two. Section three presents our proposed evaluation function based on Interaction Gain (IGFS). Experimental results on handwritten digital recognition data and comparison in term of classification accuracy between our proposed method and three recently proposed information theoretic criteria is presented in section four. The last section summarizes the finding and gives some perspectives that can follow up on this work.

Manuscript received April 18, 2008.
Manuscript revised April 20, 2008.

2. Information theoretic for Feature Selection

2.1 Definitions and measurements

(i) Mutual information and conditional mutual information:

The first goal of a prediction model is to minimize the uncertainty on the dependent variable. A good formalization of the uncertainty of a random variable is given by Shannon and Weaver's [4] information theory. While first developed for binary variables, it has been extended to continuous variables. Let X and Y be two random variables (they can have real or vector values). We denote $\mu_{X,Y}$ the joint probability density function of X and Y. We recall that the marginal density functions are given by:

$$\mu_X(x) = \int \mu_{X,Y}(x,y)dy \qquad (1)$$

$$\mu_Y(y) = \int \mu_{X,Y}(x,y)dx \qquad (2)$$

Let us now recall some elements of information theory. The uncertainty on Y is given by its entropy defined as:

$$H(Y) = -\int \mu_Y(y)\log\mu_Y(y)dy \qquad (3)$$

If we get knowledge on Y indirectly by knowing X, the resulting uncertainty on Y knowing X is given by its conditional entropy, that is:

$$H(Y/X) = -\int \mu_X(x)\int \mu_Y(y/X=x)\log\mu_Y(y/=x)dydx \qquad (4)$$

The joint uncertainty of the (X,Y) pair is given by the joint entropy, defined as:

$$H(X,Y) = -\int \mu_{X,Y}(x,y)\log\mu_{X,Y}(x,y)dxdy \qquad (5)$$

The mutual information between X and Y can be considered as a measure of the amount of knowledge on Y provided by X (or conversely on the amount of knowledge on X provided by Y). Therefore, it can be defined as [5]:

$$I(X;Y) = H(Y) - H(Y/X) \qquad (6)$$

Which is exactly the reduction of the uncertainty of Y when X is known. If Y is the dependant variable in a prediction context, the mutual information is thus particularly suited to measure the pertinence of X in a model for Y [6]. Using the properties of the entropy, the mutual information can be rewritten into:

$$I(X;Y) = H(Y) + H(Y) - H(X,Y) \qquad (7)$$

That is, according to the previously recalled definitions, into [7]:

$$I(X;Y) = -\int \mu_{X,Y}(x,y)\log\frac{\mu_{X,Y}(x,y)}{\mu_X(x)\mu_Y(y)}dxdy \qquad (8)$$

The conditional mutual information is defined as:

$$\begin{aligned} I(X_1;Y/X_2) &= H(X_1/Y) - H(X_1/Y,X_2) \\ &= I(X_1/Y,X_2) - I(X_1/Y) \end{aligned} \qquad (9)$$

This value quantifies how much information is shared between X_1 and Y, given the value of X_2. Another way to see it, as it is decomposed above, is as the difference between the information required to describe X_1 given X_2, and the information to describe X_1 given both X_2 and Y. If Y and X_2 carry the same information about X_1, the two terms on the right are equal, and the conditional mutual information is zero. On the opposite, if both Y and X_2 bring information, and if those informations are complementary, the difference is large.

(ii) Feature interaction and Interaction Gain:

Feature selection is one effective mean to remove irrelevant features [8]. Optimal feature selection requires an exponentially large search space $(O(2*N)$, where N is the number of features) [9]. Researchers often resort to various approximations to determine relevant features (e.g., relevance is determined by correlation between individual features and the class) [10], [11]. However, a single feature can be considered irrelevant based on its correlation with the class; but when combined with other features, it becomes very relevant. An illustration of feature interaction is given by the well-known XOR problem [12], [13]:

X_1	1	1	0	0
X_2	1	0	1	0
$X_1 \oplus X_2$	0	1	1	0

We see that X_1 and X_2 have null mutual information with the output, once they are taken individually (i.e $I(X_1;Y)=0$, $I(X_2;Y)=0$). However, when they are taken together, the mutual information

$I(X_1, X_2;Y) = H(Y) > 0$ of the subset is positive. Interaction explains why an apparently irrelevant combination of variables can eventually perform efficiently in a learning task. To decide, whether there is interaction between two attributes, [14] propose an heuristic test, called interaction gain. It is based on the well-known idea of information gain. Information gain can be regarded as a measure of the strength of a 2-way interaction between an attribute X and the class Y. In this spirit, we can generalize it to 3-way interactions by introducing the interaction gain [14]:

$$I(X_1; X_2; Y) = I(X_1, X_2; Y) - I(X_1; Y) - I(X_2; Y) \quad (10)$$

Interaction gain can be understood as the difference between the actual decrease in entropy achieved by the joint attribute $X_1 X_2$ and the expected decrease in entropy with the assumption of independence between attributes X_1 and X_2. The higher the interaction gain, the more information was gained by joining the attributes in the Cartesian product, in comparison with the information gained from single attributes. It is quite easy to see that when interaction gain is negative, context decreased the amount of dependence. When the interaction gain is positive, context increased the amount of dependence. When the interaction gain is zero, context did not affect the dependence between the two attributes. Interaction gain is identical to the notion of interaction information [13] and mutual information among three random variables [14], [15].

In the following section, we will proceed to a critical survey of information theoretic approaches existing in literature, by stressing when and where the notion of interaction is taken into account.

2.2 State of the Art

As mutual information can measure relevance, this quantity is currently used in literature for performing feature selection. One of the main reasons for adopting it is its low complexity computational complexity cost $(O(d * N))$ where d is the number of variables and N is the number of samples) in the case of discrete variables. The following sections will sketch three state-of-the-art filter approaches that use this quantity.

(i) Variable Ranking (RANK):

The ranking method returns a ranking of variables on the basis of their individual mutual information with the output. This means that, given n input variables, the method first computes n times the quantity $I(X_i;Y)$, $i = 1 \ldots n$, then ranks the variables according to this quantity and eventually discards the least relevant ones [16].

The main advantage of the method is its rapidity of execution. Indeed, only n computations of mutual information are required for a resulting complexity $(O(n * 2 * N))$. The main drawback derives from the fact that possible redundancies between variables is not taken into account. Indeed, two redundant variables, yet highly relevant taken individually, will be both well ranked. As a result, a model that uses these two variables is dangerously prone to an increased variance without any gain in terms of bias reduction. On the contrary, two variables can be complementary to the output (i.e. highly relevant together) while each of them appears to be poorly relevant once taken individually. As a consequence, these variables could be badly ranked, or worse eliminated, by the ranking filter. Although the variable ranking algorithm is reputed to be fast, it may be poorly efficient as it only relies on individual relevance. Recently, new algorithms that combine relevance and redundancy analysis offer a good compromise between accuracy and computational load as the Fast Correlation Based Filter [6]. Also, some heuristic search methods such as the best first search (also known as the forward selection) can be combined efficiently with information theoretic criteria in order to select the best variable given a previously selected subset.

In the next sections, two theoretic criteria existing in the literature and that can be easily combined with the forward selection, are presented.

(ii) Minimum Redundancy - Maximum Relevance criterion (MRMR):

The minimum redundancy - maximum relevance criterion [17] consists in selecting the variable X_i among the not yet selected features X_{-S} that maximizes $u_i - z_i$ where u_i is a relevance term and z_i is a redundancy term. More precisely, u_i is the relevance of X_i to the output Y alone, and z_i is the mean redundancy of X_i to each variable $X_i \in X_S$ already selected.

$$u_i = I(X_i;Y) \quad (11)$$

$$z_i = \frac{1}{d} \sum_{X_j \in X_S} I(X_i;X_j) \quad (12)$$

$$X_{MRMR} = \arg \max_{X \in X_{-S}} (u_i - z_i) \quad (13)$$

At each step, this method selects the variable which has the best compromise relevance-redundancy. This selection criterion is fast and efficient. At step d of the forward search, the search algorithm computes $n - d$ evaluations where each evaluation requires the estimation of $d+1$ bi-variate densities (one for each already selected variables plus one with the output). It has been shown in [17] that the MRMR criterion is an optimal first order approximation of the conditional relevance criterion. Furthermore, MRMR avoids the estimation of multivariate densities by using multiple bi-variate densities. Note that, although the method aims to address the issue of redundancy between variables through the term z_i, it is not able to take into account the interactions between variables.

(iii) Conditional Mutual Information Maximization Criterion (CMIM):

This approach [18] proposes to select the feature $X_i \in X_{-S}$ whose minimal conditional relevance $I(X_i;Y/X_j)$ among the selected features $X_j \in X_S$, is maximal. This requires the computation of the mutual information of X_i and the output Y, conditional on each feature $X_j \in X_S$ previously selected. Then, the minimal value is retained and the feature that has a maximal minimal conditional relevance is selected. The variable returned according to the CMIM criterion is:

$$X_{CMIM} = \arg \max_{X \in X_{-S}} (\min_{X_j \in X_S} (I(X_i;Y/X_j))) \tag{14}$$

This selection criterion is powerful. It selects relevant variables, it avoids redundancy, it avoids estimating high dimensional multivariate densities and unlike the previous method, it does not ignore variable interaction. However, it will not necessary select an interacting variable with the already selected variables. Indeed, a variable that has a high interaction with an already selected variable will be characterized by a high conditional mutual information with that variable but not necessarily by a high minimal conditional information. In terms of complexity, note that at the d^{th} step of the forward search, the algorithm computes $n-d$ evaluations where each evaluation following CMIM requires the estimation of d tri-variate densities (one for each previously selected variable).

3. Interaction Gain Based Feature Selection (IGFS)

The new proposed evaluation measure for a given feature X will be based on the individual Mutual Information and a compromise between features redundancy and features interaction. The compromise is made by the mean of Interaction Gain. In formal notation, the variable returned according to the IGFS criterion is:

$$X_{IGFS} = \arg \max_{X \in X_{-S}} (I(X_i;Y) + \frac{1}{d} \sum_{X_j \in X_S} I(X_i;X_j;Y)) \tag{15}$$

The main advantage in using this criterion for selecting variables is that an interacting variable of an already selected one has a much higher probability to be selected than with other criteria. The relevance of each feature can be indicated by its Mutual Information with class labels $I(X_i;Y)$. The second term makes a compromise between redundancy and interaction. A negative Interaction Gain indicates that the features are redundant and a positive one indicates that the features work well together.

4. Experimental Results

In this section, we perform comprehensive experiments on handwritten digital recognition dataset to compare the IGFS selection algorithm with the three state of the art approaches discussed above: The Ranking algorithm, the Minimum Redundancy Maximum Relevance criterion and the Conditional Mutual Information Maximization criterion.

4.1 Dataset Description

We have used the dataset of handwritten numeral recognition from UCI Machine Learning Repository [19]. It consists of 649 features on handwritten numerals (‘0’–‘9’). These 649 features distribute over the following feature sets: 76 Fourier coefficients of the character shapes, 216 profile correlations, 64 Karhunen-Love coefficients, 240 pixel averages in 2x3 windows, 47 Zernike moments, 6 morphological features. There are 200 patterns per class (for a total of 2,000 patterns).

4.2 Classifier Description

SVM (Support Vector Machine) is a relatively new and promising classification method [20]. It is a margin classifier that draws an optimal hyper-plane in the feature vector space; this defines a boundary that maximizes the margin between data samples in two classes, therefore leading to good generalization properties. A key factor in SVM is to use kernels to construct nonlinear decision boundary.

In this experimentation, we used the Weka [21] version of LIBSVM [22] which allow us to directly construct a multiclass SVM with exponential kernel.

4.3 Assessment measure

We assessed classification performance using K-fold cross validation. In this assessment method the original sample is partitioned into k sub-samples. Of the k sub-samples, a single sub-sample is retained as the validation data for testing the model, and the remaining $k-1$ sub-samples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k sub-samples used exactly once as the validation data. The k results from the folds then can be averaged (or otherwise combined) to produce a single estimation. Cross validation accuracy provides more realistic assessment of classifiers which generalize well to unseen data. We used 10-fold cross validation [23], [24].

4.4 Results

Each selection method stops after those thirty variables have been selected. Then, the evaluation of the selection is done by using a 10-fold cross validation with a SVM learning algorithm. The accuracy of classification (recognition rate) relatively to the step by step introduction of the variables is computed and the evolution of the recognition rate using different feature selection algorithm is reported in the fig.1.

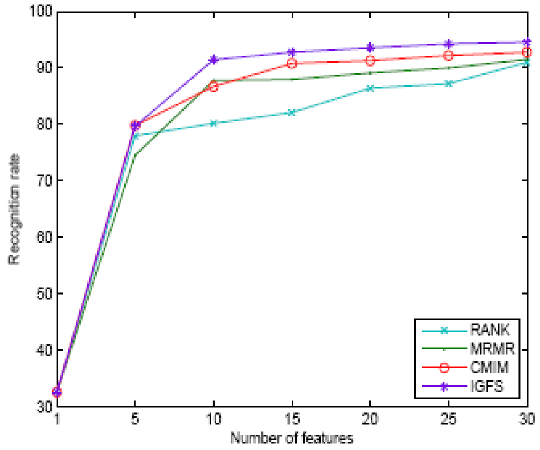

Fig. 1 Evolution of the 10-fold cross validation accuracy of the SVM learning algorithm

The above graph show the strength of our proposed measure compared with the three well known feature selection algorithms. In addition IGFS is better than the other algorithm by at least 2% of the recognition rate.

The analysis of this graph allowed us to take out the following results:

1) The measures based on the mutual information can be used for performing feature selection for the problem of pattern recognition;

2) The analysis of the interaction between features must be taken into consideration when selecting features for pattern recognition.

4. Conclusion and future work

In this paper, we proposed a new evaluation function, called IGFS, based on the concept of mutual information and interaction gain. The function takes into consideration the interaction between features. When the function was used with the stepwise selection procedure in the problem of pattern recognition, it improves classification accuracy with a lesser number of features compared to the other methods. The main advantage of the proposal measure is that it takes into account different features interaction without increasing the computational complexity.

Further experiments will focus on other pattern recognition problems. Moreover, other search strategies than the forward selection in order to validate the criterion in a wider range of domains.

References

[1] Dash, M., Liu, H. ”Feature selection for classification”. Intelligent Data Analysis 1 pp 131-156 (1997).

[2] Isabelle Guyon Andr Elisseeff, ”An Introduction to Variable and Feature Selection” Journal of Machine Learning Research 3 pp 1157-1182 (2003).

[3] Aleks Jakulin and Ivan Bratko. Analyzing attribute dependencies. In PKDD, 2003.

[4] C.E. Shannon, W. Weaver, The Mathematical Theory of Communication, University of Illinois Press, Urbana, IL, 1949.

[5] T.M. Cover, J.A. Thomas, Elements of Information Theory,Wiley, New York, 1991.

[6] Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research 5 (2004) 12051224

[7] C.H. Chen, Statistical Pattern Recognition, Spartan Books, Washington, DC, 1973.

[8] A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning. Artificial Intelligence, 97:245271, 1997.

[9] H. Almuallim and T. G. Dietterich. Learning boolean concepts in the presence of many irrelevant features. Artificial Intelligence, 69(1- 2):279305, 1994.

[10] M. A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In ICML, 2000.

[11] L. Yu and H. Liu. Feature selection for highdimensional data: a fast correlation-based filter solution. In ICML, 2003.

[13] Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence 97(1-2) (1997) 273-324.

[13] McGill, W.J.: Multivariate information transmission. Psychometrika 19 (1954) 97-116.

[14] Jakulin, A.: Attribute interactions in machine learning. Master’s thesis, University of Ljubljana, Faculty of Computer and Information Science (2003).

[15] Yeung, R.W.: A new outlook on Shannon’s information measures. IEEE Transactions on Information Theory 37 (1991) 466-474.

[16] Duch, W., Winiarski, T., Biesiada, J., Kachel, A.: Feature selection and ranking filters. In: International Conference on Artificial Neural Networks (ICANN) and International

Conference on Neural Information Processing (ICONIP). (2003) 251 254.

[17] Peng, H., Long, F.: An efficient max-dependency algorithm for gene selection. In: 36th Symposium on the Interface: Computational Biology and Bioinformatics. (2004).

[18] Fleuret, F.: Fast binary feature selection with conditional mutual information. Journal of Machine Learning Research 5 (2004) 15311555.

[19] C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998.

[20] Vapnik V, The Nature of Statistical Learning Theory, New York: Springer, 1995.

[21] Ian H. Witten and Eibe Frank. Data mining:Practical machine learning tools and techniques with Java implementations. Morgan Kaufman, San Francisco, CA, USA, 2000. http://www.cs.waikato.ac.nz/ml/weka/.

[22] Yasser EL-Manzalawy (2005). WLSVM. URLhttp://www.cs.iastate.edu/ yasser/wlsvm/.

[23] Moore, A.W. and Lee, M.S., ”Efficient algorithms for minimizing cross validation error.” In: Proceedings of Eleventh International Conference on Machine Learning, Morgan Kaufmann, New Brunswick, New Jersey, 190-198, (1994).

[24] Thomas G. Dietterich ”Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms” Neural Computation 10, 1895-1923(1998).

[25] T. Mitchell, ” Machine Learning” McGraw-Hill (1997).