Two New Metrics for Feature Selection in Pattern Recognition

Pedro Piñero[1], Leticia Arco[2], María M. García[2], Yaile Caballero[3],
Raykenler Yzquierdo[1], and Alfredo Morales[1]

[1]Bioinformatic Laboratory, University of the Informatics Sciences. La Habana, Cuba
{ppp, ryzquierdo, amoraleso}@uci.cu
[2]Artificial Intelligence Laboratory, Central University of Las Villas. Santa Clara, Cuba.
{leticiaa, mmgarcia}@uclv.edu.cu
[3]Informatics Department, University of Camagüey. Camagüey, Cuba
yaile@reduc.cmw.edu.cu

Abstract. The purpose of this paper is to discuss about feature selection methods. We present two common feature selection approaches: statistical methods and artificial intelligence approach. Statistical methods are exposed as antecedents of classification methods with specific techniques for choice of variables because we pretend to try the feature selection techniques in classification problems. We show the artificial intelligence approaches from different points of view. We also present the use of the information theory to build decision trees. Instead of using Quinlan's Gain we discuss others alternatives to build decision trees. We introduce two new feature selection measures: MLRelevance formula and the PRelevance. These criteria maximize the heterogeneity among elements that belong to different classes and the homogeneity among elements that belong to the same class. Finally, we compare different feature selection methods by means of the classification of two medical data sets.

1 Introduction

The Pattern Recognition is an interdisciplinary science, having strong connections with Mathematics, Engineering and Computer Sciences. The following problems can be solved by means of the pattern recognition techniques:

- search of effective object descriptions and
- classification problems.

In classification problems, the studied objects are described in terms of a set of features. Each feature (xi) has a set (Mi) of acceptable values and a comparison criterion (δi) associated to it. Suppose that a given training sample, in a framework of a supervised classification problem, has a (training) matrix representation $I_0(C_1, C_2,\ldots, C_r)$, that is, object descriptions (O1, O2, …, Op) are stored in a matrix with as many columns as features, as many rows as objects in the sample, and they are split in groups corresponding with their respective classes (C_1, C_2, …, C_r). Likewise the succession I(O´$_1$, O´2, …, O´q) of standard descriptions of the objects (O´$_1$, O´$_2$, ...,

A. Sanfeliu and J. Ruiz-Shulcloper (Eds.): CIARP 2003, LNCS 2905, pp. 488–497, 2003.

O´q) such that $O_j^{'} \notin I_0(C_1, C_2, \ldots, C_r)$ with 1<j<q is called control matrix. Usually the feature selection problem appear in the classification problems and in problems of search of effective object descriptions, as a necessary step to reduce the dimensions of the initial space of representation of objects and simplify the classification process complexity. The problem of the selection of features for the classification consists on finding an algorithm ς such that:

First: $\varsigma\,(K(x_1(O), x_2(O),\ldots, x_n(O))) = K\,(x_{i_1}(O), x_{i_2}(O),\ldots, x_{i_p}(O))$

∀O∈I0(C1, C2,…, Cr) where K is a classification criterion It means that the algorithm ς reduces the dimensions of the space without affecting the belonging of each object to its respective class. In other words, using the ς algorithm the belonging r-plus of the initial training matrix remains constant although the dimensions of the space are smaller than the initial dimensions.

Second: Given a classification algorithm A and a function Φ that measures the quality of it.

$$\Phi[A(I_0(C_1,C_2,\ldots,C_r),I(O_1^{'},O_2^{'},\ldots,O_q^{'}))] \le \Phi[A(I_0(C_1,C_2,\ldots,C_r),\zeta\,(I(O_1^{'},O_2^{'},\ldots,O_q^{'})))]$$

For $\zeta\,(I(O_1^{'},O_2^{'},\ldots,O_q^{'}))$ we denote the projection of the control matrix in a new space. This new space is obtained from applying ς to the initial space. Theoretically having more features should give us more discriminating power. However the real world provides us with many reasons why this is not generally the case [1] :

- First: the induction algorithm complexity grows dramatically with the number of features.
- Second: the irrelevant and redundant features also cause problems in the classification context as they may confuse the learning algorithm by helping to obscure the distributions of the small set of truly relevant features for the task at hand.

The analysis of the techniques and traditional criteria for feature selection will be exposed with details in other sections of this paper. The second section shows the most popular statistic and artificial intelligence techniques used to solve the feature selection problems. In the third section, we'll introduce two new criteria related with the relevance or the irrelevance of features, which are valid for any feature selection technique. Finally, we'll show comparison results of algorithms that use different criteria of feature selection.

2 Statistical, Artificial Intelligence and Logical Combinatorial Pattern Recognition Techniques for Feature Variable Selection

2.1 Statistic Techniques for Feature Variable Selection

The feature selection appears in the classical statistics, in relation with all the techniques of the multivariate analysis, from the most elementary techniques: the analysis of variance (ANOVA) and the regression. In fact, in the multiple lineal regression theory some new feature selection procedures appear, influencing the ulterior development of the multivariate statistic. We talk about the step-to-step methods as a way to get the better equation of regression among all the possible, keeping in mind the correlation among variables. The procedures "step-to-step" are easily extended to the classification statistical techniques: the discrimination analysis, the logistic regression, the decision trees, etc. [2]

There isn't a strong criterion to divide the classification procedures in separate groups. In a way sense, they are always extensions of some statistical techniques such as: the discriminate analysis, the methods based on decision trees (CHAID technique: Chi-square Automatic Interaction Detector), the methods of estimating of densities (KNN: *k*-nearest neighbors), or the techniques of hierarchical group formations.

These four procedures (linear discrimination, decision-tree, *k*-nearest-neighbors and clustering) are prototypes for four kinds of classification procedures. Not surprisingly they have been refined and extended, but they still represent the major strands incurrent classification practice and research. Then, it may be a good criterion of classification. However, in [3], the authors preferred create groups of methods around the more traditional heading of classical statistics, modern statistical techniques, Machine Learning and neural networks.

2.2 Artificial Intelligence Techniques for Feature Variable Selection

There are a lot of applications of the heuristic search methods to solve the feature variable selection problems [4]. To characterize the feature selection algorithms four issues should be defined in [5].

Other approaches to solve the feature selection problems have as principal idea to apply a weighting function to features. The weighting schemes generally are easier to implement than the others machine learning methods. They are frequently more difficult to understand because usually work as a black box.

Perceptron is a well-known feature weighting method, which adds or eliminates weights on the linear threshold unit in response to errors that occurs during the classification process. Many learning algorithms such as: back-propagation algorithm and least-mean square algorithm have been well studied. The results of Perceptron-weighting techniques can be affected when the number of irrelevant features grows. To decrease the sensibility of the Perceptron algorithm the Winnow algorithm is proposed in [5].

Other approach to solve the problems related to the relevant features in classification problems is the filter methods [6]. This viewpoint divides the feature selection process and the induction process. These methods make a preprocessing of the training data and filter out the irrelevant features before the induction process occurs. The filter methods work independent of the induction methods. They can be used in combination with different induction methods. Besides, the filter methods evaluate each feature based on its correlation with set of classes choosing the suitable number of relevant features. Two of the most well known filter methods for feature selection are RELIEF [7] and FOCUS [8].

Other feature selection methodology, which has recently received much attention, is the wrapper model. This model searches through the space of feature subsets using the estimated accuracy from an induction algorithm as the measure of goodness for a particular feature subset [6]. Actually the wrapper methods are well known in statistic and pattern recognition. The principal notion in the wrapper methods is to determine the feature subset that allows us better estimations than separate measures. The major disadvantage of wrapper methods over filter methods is the high computational cost of them. The wrapper methods similar to the filter methods can be used in combination with different induction methods. In fact the OBLIVION [9] wrapper algorithm combines the wrapper notion with the nearest-neighbor method.

The embedded approaches to determine relevant features are popular methods too. A clearest example of feature selection methods embedded within a basic induction algorithm, are the "methods for inducing logical descriptions". For these algorithms the space of hypotheses is described by the partial ordering and the algorithms use this ordering to organize their search for concept descriptions. The core of these algorithms is to add or remove features from the concept description in response to prediction errors on new instances. For example, recursive partitions methods for induction, such as Quinlan's ID3 Quinlan [10] , C4.5 [11] and CART [12] carry out a greedy search through the space of decision trees, at each node using an evaluation criterion to choose the feature having the best ability to discriminate among classes.

Information theory is one approach to solve the information uncertainty problems; however, it's not a tool for manipulating uncertain knowledge. Instead, it's a tool for measuring uncertainty. In information theory, uncertainty is measured by a quantity called "entropy". It's similar to, but not the same as, the concept of entropy in physics [13]. An example of the entropy computation is presented in the selection variable building decision trees. In fact, Quinlan propose the ID3 algorithm to induce classifications rules in form of decision tree [11, 14]. In recent years, Quinlan introduces the algorithms C4.5 [11, 14] and C5.0 [15]. These Quinlan's algorithms improve the ID3 algorithm because they work with numeric and symbolic data and manipulate cases with missing values.

In the information theory approach many other measures have been proposed, for example, instead of using Quinlan's Gain, Mántaras [16] propose two-feature selection measures based on a distance between partitions.

2.3 Logical Combinatorial Pattern Recognition and Testor Theory in the Feature Variable Selection

Some problems related to the feature selection can be solved in the context of the testor theory. This is a branch of Mathematical Logic that began in the Soviet Union at the end of the 50's. I. A. Cheguis and S. V. Yablonskii [17] were the first researchers that developed this theory. Their works were motivated by the problem of fault detection in logical schemes, particularly applied to computer logical circuits.

In the middle of the 60's, Y. I. Zhuravlev adapted the testor concept to pattern recognition [18].

Testor definition (Zhuravlev): If the complete set of features R allows us to distinguish between objects (rows of MI) from different classes, then R is a testor. Furthermore, any non-empty feature subset of R, that satisfies this property, is a testor. Others Testor‘s concepts, that improve the original Zhuravlev’s concept are proposed in [19] [20]

3 Two New Alternative Criteria for Feature Variable Selection

In this section we propose two alternative criteria to choose the relevant features in classification problems. Some theoretical results obtained from the analysis of this measure of relevance are presented.

3.1 The MLRelevance Criterion

Suppose a feature (A) with $i = 1,2,...,k$ acceptable values, S set of samples and S_i subset of S that contains the samples having the value i in the feature A. Then the expression $|S_i|/|S|$ is the relative frequency of the value i in S.

Then Equation 1 shows a measure that determines the relevance of the feature A.

MLRelevance measure $$R(A)=\sum_{i=1}^{K}\frac{|S_i|}{|S|}e^{(1-C_i)} \quad (1)$$

where $R(A)$ is the relevance measure of the feature A on set S, k is the number of different values for the feature A and C_i is the number of different classes presented in objects having the value i for the feature A.

Let us begin by saying some general aspects of our measure:

- its principal idea is to maximize the heterogeneity among elements that belong to different classes and the homogeneity among elements that belong to the same class and,
- $0 \leq R(A) \leq 1$ and $\sum_{i=1}^{K}\frac{|S_i|}{|S|}=1$. Consequently, the feature that maximizes $R(A)$ is better.
- The Equation 1 will always be defined for any set S that is a good property of this equation.

3.2 The PRelevance Criterion

Another criterion is a lineal combination of the MLRelevance criterion and a heuristic. The core of this second metric deals with to determine the relevance of an attribute a as a lineal combination of the relevance of the isolated attribute “a” and the relevance of the groups of attributes B such that $a \in$ B.

3.2.1 Preliminary Concepts

The heuristic that we use in the *PRelevance* computation is based on the rough sets theory [21].

Lets the decision system $W = (U, A \cup D)$, and the sets $B \subseteq A$ y $S \subseteq U$. We can approximate S using only the information contained in B by constructing the *B-lower* and the *B-upper* approximations of S, denoted (B_*) and (B^*) respectively. A rough set is any set S, $S \subseteq U$ defined from its *B-lower* and *B-upper* approximations [22].

We'll define indiscernibility, this is the fundamental notion in the rough sets theory. The objects that are characterized by the same information are indiscernible (similar) in the view of the information that is available.

Definition 1 Indiscernibility: To each set of the attribute B such that $B \subseteq A$, is associated an indiscernible binary relation denoted by I_B. This relation allows us to determine which objects are indiscernible from the others by the relation. $I_B = \{ (x,y) \in UxU: f(x,a_i)=f(y,a_i)$ para todo $a_i \in B\}$. If $(x,y) \in I_B$ we said that the objects x and y were B- indiscernible.

The lower approximation of a set S respect to a set of attributes B is defined as the collection of objects which equivalences classes are contained completely in the set; whereas the upper approximation is defined as the collection of objects which equivalences classes are partially contained in the set. Formally,

$$B_*(S)= \{x \in U \mid B(x) \subseteq S \} \quad (2)$$

$$B^*(S)= \{x \in U \mid B(x) \cap S \neq \phi\} \quad (3)$$

Now, we can define the boundary region on S as:

$$BN_B(S)= B^*(S) - B_*(S) \quad (4)$$

If the set BN_B is empty then the set S is exact respect to the equivalence relation B. In any other case $BN_B(X) \neq \phi$, the set S is inexact, vague, rough; respect to B. Using the lower and upper approximations of a concept, three regions are defined:

I Positive region: $POS(X) = B_*(X)$.

II Boundary region: $BN_B(X)$.

III Negative region: $NEG(X)=U-B^*(X)$

3.2.2 Dependences between Attributes

Intuitively, a set of decision attributes D, depends totally on a set of B attributes, denoted by $B \Rightarrow D$, if all the values of the D attributes are univocally determined by the values of the attribute in B.

In other words, D depends totally on b, if there is a functional dependency between the values of D and B [22].

Definition 2: Dependency in *k* grade.

It's said that D depends on B in a k grade $(0 \leq k \leq 1)$, denoted by $B \Rightarrow_k D$, by the k value, and defined by the expression 5.

$$k = \frac{|POS_B(D)|}{|U|} \quad (5)$$

$$\text{Donde } POS_B(D) = \bigcup_{X \in U/D} B_*(X) \tag{6}$$

If *k=1* then it's said that D depends totally on B, while if *k<1* it's said that D depends partially on B.

3.2.3 PRelevance Computation

From what it's been defined till now, so far the calculus of *PRelevance* with respect to an attribute "*a*" it's defined as *RP(a)* expression:

$$\text{PRelevance} \quad RP(a) = R(a) + H(a) \tag{7}$$

Where *R(a)* is the function of the equation 1 and *H(a)* is calculated as it shown in the algorithm 1. The attribute that maximizes *RP(a)* is the most relevant attribute.

Algorithm 1

Step1: it is calculated the vector $R(T) = (R(a_1), R(a_2), R(a_3), \ldots, R(a_{r(a)}))$ with $T \subseteq A$

Step2: It's determined the n best attributes, begin the best those which maximize $R(a_i)$. As a result of this step the vector, $RA = (R(a_i), R(a_j), \ldots, R(a_t))$ with $n = | RA |$, is obtained.

Step3: The n combinations are determined in p from the attributes selected in the step2. A vector of combinations is obtained: $Comb = (\{a_i, a_j, a_k\}, \ldots \{a_i, a_t\, a_p\})$

An example of it being, n = 4 and p = 3 and being the selected attributes in the step 2 (a_1, a_3, a_5, a_8) the combination vector has $C_p^n = \frac{n!}{p!(n-p)!} = 4$ components which would be $Comb = (\{a_1, a_3, a_5\}, \{a_1, a_3, a_8\}, \{a_3, a_5, a_8\}, \{a_1, a_5, a_8\})$.

Step 4: We calculated the independency grade of the classes with respect to each of the obtained combinations in the previous step. As a result of this step we obtain the vector of dependencies $DEP = (k(Comb_1, d), k(Comb_2, d), \ldots k(Comb_r, d))$.

Step 5: For each attribute "a" we calculate the value of H(a) following the equation 8:

$$H(a) = \sum_{\forall i \,/\, a \in Comb_i} k(Comb_i, d) \tag{8}$$

As it can be appreciated in the computation of PRelevance for an attribute, is very expensive and depends on *(|A|), (n),(p)* and *(|d|)*. These parameters depend on the real problem that we can to solve. Also, if we want to use the *PRelevance* metric to build decision trees then the expensive procedure is repeated and the cost of our learning increases a lot. In order to reduce the learning duration we propose a PRelevance's implementation using a parallel platform (MPI, PVM)

4 Comparisons between Different Feature Selection Methods

In this section we shall compare different feature selection methods using data of two medical domains. We use in our comparisons: the thyroid database provided by the

Garvan Institute of Medical Research, Sydney and the heart database assays from the European Statlog project, Dept. Statistics and Modeling Science, Strathclyde University in 1993. Both medical databases appear in UCI Repository of Machine Learning databases, University of California [23].

We compare the correctness percents of classification among systems C5.0 [15], KNN IB4 implementation [24], MLClassif (VCramer), MLClassif (Mantaras) and MLClassif (MLRelevance). We use the VCramer formula; this is a measure of interrelation between variables [25] [2]. The C5.0 system developed by Ross Quinlan creates a decision tree based on Quinlan's gain. The MlClassif system developed by our team creates partitions by recursive sorting of the training set. To rearrange each partition an appropriate feature is selected. To choose the most relevant feature in each moment we use: Mántaras's distance, the VCramer formula or the MLRelevance measure.

From each database we create randomly ten partition pairs (Table 1), having each partition pair 75% of elements for train and the rest for test. We execute the algorithms in each partition. The values that we show in the table 1 represent the percent of correct classification obtained from each algorithm in the partition.

To compare the algorithms results we applied the Kruskal-Wallis Test for each variable; we used the Monte Carlo method for computing the significance level and considered 99% as confidence interval for the significance.

The superscript letters used in tables 1 and 2 represent different sets. These sets were obtained from to apply the Kruskal-Wallis test. Values having the same superscript belong to the same set. It means that these values have not a significant difference.

To compare two algorithms we used the Mann-Whitney U test for each variable; we used the Monte-Carlo method for computing the significance level and considered 99% as confidence interval.

In thyroid significant differences are found regard train and test variables, however in heart only is found significant differences regards train variable. The Table 2 shows the algorithms grouping.

Table 1. Partition 1 experimental results

Partition 1, accuracy results	Thyroids database		Heart database	
	Train	Test	Train	Test
MLClassif (Vcramer)	30.12^{a}	33.38^{a}	89.16^{b}	88.06^{a}
MLClassif (Mantaras)	85.46^{b}	87.45^{b}	92.61^{c}	89.55^{a}
MLClassif (MLRelevance)	96.39^{c}	95.26^{c}	93.1^{C}	88.6^{a}
KNN IB4	87.4^{b}	83.9^{b}	71.8^{a}	88.1^{a}
C5.0	98.3^{c}	95.4^{c}	93.1^{c}	87.7^{a}

Table 2. Resulting groups from applying the statistic tests to the classification results

Group	Thyroids (Train and test)	Heart (Train)
1	C5.0^c, MLClassif (MLRelevance)^c	C5.0^c, MLClassif (MLRelevance)^c, MLClassif (Mantaras)^c
2	KNN IB4^b, MLClassif (Mantaras)^b	MLClassif (VCramer)^b
3	MLClassif (VCramer)^a	KNN IB4^a

As a conclusion of the above tables: the methods of group1, are better than the methods of group2, likewise the group2 methods are better than the group3 methods and the methods that belong to the same group don't have significant differences.

5 Conclusions

The purpose of this paper was to discuss about feature selection methods. We presented there common feature selection approaches: statistical methods, logical combinatorial pattern recognition approach and artificial intelligence approach. For each approach we discussed some methods and algorithms.

Statistical methods are presented as antecedents of the other methods with their specific techniques for choice and transformation variables. In the logical combinatorial pattern recognition we discuss the testor theory and its application to the classification and feature selection problems. Different artificial intelligence techniques are presented and its properties briefly discussed.

We introduce two new relevance criteria the *MLRelevance* $R(A)$ and the *PRelevance RP(A)*.

These feature selection criteria maximizes the heterogeneity among elements that belong to different classes and the homogeneity among elements that belong to the same class. $R(A)$ always will be defined for any set S, $0 \leq R(A) \leq 1$ and is not sensitive to the number of features values. The *RP(A)* computation is very expensive and we propose a PRelevance's implementation using a parallel platform.

Finally we compare different features selection measures by means of two medical databases. We compare the measures: VCramer (statistic measure), C5.0 algorithm (Quinlan's gain), Mántaras and MLRelevance. We conclude: C5.0 and MLRelevance obtain the best results and VCramer obtains the worse results in Thyroid database; while in Heart database, C5.0, MLRelevance and Mántaras obtain the best results and KNN obtains the worse results.

References

1. Koller, D. and S. Mehran, *Toward Optimal Feature Selection*. 1997, Computer Science Department Stanford University: Stanford.
2. Grau, R., *Estadística aplicada con ayuda de paquetes de software*. 1994, Editorial Universitaria: Jalisco México.
3. Michie D. Spiegelhalter J., T.C.C., *Machine Learning, Neural and Statistical Classification*. 1994: Springer.
4. Bello., R., *Métodos de Solución de Problemas para la Inteligencia Artificial*. 1998, Santa clara: Universidad Central de Las Villas.
5. Blum, A. and P. Langley, *Selection of relevant features and examples in mechine learning*. Artificial Intelligence 97, 1997: p. 245–271.
6. John, G., R. Kohavi, and K. Pfleger. *Irrelevant features and the subset selection problems*. in *Proceedings 11th International conferences on Machine Learning*. 1994. New Brunswick, NJ.
7. Kira, K. and L.A. Rendell. *A practical approach to feature selection*. in *Proceedings 9th International Conference on Machine Learning*. 1992. Aberdeen, Scotland.
8. H. Almuallim, T.G.D. *Learning with many irrelevant features*. in *Proceedings of AAAI-92*. 1992: MIT Press.
9. Langley, P. and S. Sage. *Oblivious decision trees and abstract cases*. in *Working Notes of the AAAI-94, Workshop on Case Base Reasoning*. 1994. Seattle.
10. Quinlan, J.R., *Induction of Decision Trees.* Machine Learning, 1986: p. 81–106.
11. Quinlan, J.R., *Improved Use of Continuous Attributes in C4.5.* Research Journal of Artificial Intelligence, 1996. **4**: p. 77–90.
12. Breiman, L.F., J.H. Olshen, R.A. Stone, C.J. *Classification and Regression Trees.* Wadsworth Belmont, CA. 1984.
13. Brender, J. *Measuring quality of medical knowledge*. in *In Proceeding of the Twelfth International Congress of the European Federation for Medical Informatics*. 1994.
14. Quinlan, J., *C4.5: Programs for Machine Learning*. 1993, San Mateo: Morgan Kaufman.
15. Quinlan, J.R., *See5/C5.0*. 2002,.
16. Mántaras, R.L., *A Distance-Based Attribute Selection Measure for Decision Tree Induction.* Machine Learning, 1991.
17. Cheguis, I. and S.Yablonskii, *K-Testor*. 1958, Moscow: Trudy Matematicheskava Instituta imeni V. A. Steklova LI. 270–360.
18. Zhuravlev, Y.I. and S.E. Tuliaganov, *Measures to Determine the Importance of Objects in Complex Systems*. Vol. 12. 1972, Moscu. 170–184.
19. Aizenberg, N.N. and A.I. Tsipkin, *Prime Tests*. Vol. 4. 1971: Doklady Akademii Nauk. 801–802.
20. Ruiz-Shulcloper, J. and M.L. Cortés, *K-testores primos.* Revista Ciencias Técnicas Físicas y Matemáticas, 1991. 9: p. 17–55.
21. Pawlak, Z., *Rough Sets- Theorical Aspects of Reasoning about Data*. 1991, Dondrecht: Kluwer Academic.
22. Komorowski, J., et al., *A Rough Set Perspective on Data and Knowledge*, in *The HandBook of DataMining and Knowledge Discovery*, W. Klosgen, Editor. 1999, Oxford University Press.
23. Blake, C.L. and C.J. Merz, *UCI Repository of Machine Learning databases*. 2003, University of California, Department of Information and Computer Science.
24. Aha, D.W., *Case-Based Learning Algorithm*. 1991.
25. Jabson, D., *Applied Multivariate Data Analysis*. Vol. 2 Categorical and Multivariate methods. 1992: Springer.