Detection of Interdependences in Attribute Selection

Javier Lorenzo, Mario Hernández and Juan Méndez

Dpto. de Informática y Sistemas
Univ. de Las Palmas de Gran Canaria
35017 Las Palmas, Spain
jlorenzo@dis.ulpgc.es

Abstract. A new measure for attribute selection, called GD, is proposed. The GD measure is based on Information Theory and allows to detect the interdependence between attributes. This measure is based on a quadratic form of the Mántaras distance and a matrix called Transinformation Matrix. In order to test the quality of the proposed measure, it is compared with other two feature selection methods, namely Mántaras distance and Relief algorithms. The comparison is done over 19 datasets along with three different induction algorithms.

1 Introduction

Knowledge Discovery in Databases and the part of Machine Learning dealing with learning from examples overlap in the algorithms and the problems addressed. In both areas induction algorithms have to deal with attributes or fields that are not relevant to the definition of a concept, so an important task is to filter those attributes to avoid the noise that they can introduce. The irrelevant or redundant attributes do not affect the ideal Bayesian classifier because the addition of new attributes never decreases the performance of the classifier. However, many practical classifiers decrease its performance when irrelevant or redundant attributes arise. To overcome this problem, different approaches have been proposed to select the more relevant attributes that define a concept (or class) [4]. Some works on attribute selection were the WINNOW algorithm proposed by Littlestone [16], the FOCUS algorithm proposed by Almuallim and Dietterich [3] and the Relief algorithm proposed by Kira and Rendell [11]. All these algorithms have a common feature, they do not include the performance of the classifier as a measure to guide the selection of the attributes. John et al. [10] propose the *wrapper* method that utilizes the performance of the classifier to carry out the selection of the attributes. There is much evidence that wrapper method give good results [1, 10]. However, due to the computational cost, the wrapper methods can only be applied in combination with classifiers of low complexity.

The method proposed in this work utilizes a measure based on Information Theory to guide the selection of the attributes. The use of Information Theory

concepts in feature selection is not recent. Quinlan [21] proposed a measure called Gain Ratio that corresponds to the ratio between the mutual information and the entropy [24]. Another approach based on Information Theory is proposed by López de Mántaras [17], where a distance measure is defined and analyzed in its relationship with the Gain Ratio measure proposed by Quinlan. Wettschereck and Dieterich [23] demonstrated that the performance of the k-NN and Nearest-Hyperrectangle classifiers increases when the attributes are weighted by mutual information. Similar conclusions are presented by Daelemans [7], when the features used in the Exemplar-based Generalization algorithm are weighted with the mutual information in a problem of assignment of syllable boundaries in Dutch. Koller and Sahami [14] utilize the Kullback-Liebler distance as a "correlation" metric in an approximation to the Markov blankets for feature selection.

In the previous works the attributes are considered independent, they do not take into account the relation among them. In this work a measure called *GD*, between an attribute subset and the class is proposed. This measure, unlike the Gain Ratio, tries to get the possible interdependence among attributes and is based on a quadratic form of the distance proposed by López de Mántaras and a matrix called *Transinformation Matrix*.

The organization of the rest of this paper is as follows. The GD measure is defined Sect. 2, and a comparison with the Mántaras distance and Relief algorithm on several datasets is presented in Sect. 3. The selection of the attributes is carried out from a set of labeled samples. Each sample of the data set is composed of a n dimensional vector of attributes $X = \{X_1, X_2, \ldots, X_n\}$ and a label Y which indicates the class the sample belongs to.

2 GD Measure

López de Mántaras [17] proposes a distance measure between two partitions to select the attributes associated with the nodes of a decision tree. In each node, it is chosen the attribute that produces the partition closest to the correct partition of the samples subset in the examples.

The Mántaras distance has the following expression:

$$d_{LM}(P_A, P_B) = H(P_A/P_B) + H(P_B/P_A) \tag{1}$$

where $H(P_A/P_B)$ and $H(P_B/P_A)$ correspond to the entropy of a partition when the other is known. This measure fulfils the properties of a metric distance function. If we change the references to partitions by attributes and class in the definition of the Mántaras distance, we get the following expression

$$d_{LM}(X_i, Y) = H(X_i/Y) + H(Y/X_i) = H(X_i, Y) - I(X_i; Y) \tag{2}$$

where $H(X_i, Y)$ is the joint entropy of X_i and Y, and $I(X_i; Y)$ is the mutual information of X_i over Y.

Similar to the Gain Information proposed by Quinlan, the Mántaras distance can be considered an attribute selection measure because for each node of the tree the most relevant attribute is selected.

The use of Gain Ratio and the Mántaras distance has the drawback of operating over isolated attributes because they assume that attributes are independent and therefore these methods do not detect the interdependencies that could exist between attributes. A way to get into account the dependencies between attributes it is to compute the mutual information for each pair of attributes $I(X_i; X_j)$. These interdependencies of attributes can be represented with the aid of *Transinformation Matrix* T, a square matrix of dimension n (number of attributes) where each element *(i,j)* of the matrix is the mutual information between attributes *i-th* and *j-th.*

$$T = [I(X_i; X_j)]_{i,j=1,n}$$

Once the Transinformation matrix has been defined, it is necessary to find an expression for the GD measure that includes the Transinformation matrix and the distance (2). This expression must be defined in such a way that subsets of attributes with strong dependencies between attributes yields lower values than other ones without these strong dependencies. A solution come from the analogy to significance level between the Transinformation matrix and the covariance matrix (Σ) of two random variables. This analogy can be established because both matrices measure interrelation between variables. In the Mahalanobis distance [8], the covariance matrix is utilized to correct the effects of cross covariances between two components of a random variable. The expression of the Mahalanobis distance between two samples *(X,Y)* of a random variable is:

$$d_{Mahalanobis}(X,Y) = (X-Y)^t \Sigma^{-1} (X-Y) \tag{3}$$

Therefore the GD measure can be defined in a similar way to the Mahalanobis distance, using the Transinformation matrix instead of the covariance matrix and the distance (2) instead of the Euclidean distance. The GD measure $d_{GD}(X,Y)$ between the set of attributes X and the class Y is expressed as:

$$d_{GD}(X,Y) = D(X,Y)^t T^{-1} D(X,Y) \tag{4}$$

where $D(X,Y) = (d_{LM}(X_1,Y), d_{LM}(X_2,Y), \ldots, d_{LM}(X_n,Y))^T$, is a vector whose *i-th* element is the distance (2) between the attribute X_i and the class, and T is the Transinformation matrix of the set of attributes X. From the equation (2) we can observe that the elements of the *D(X,Y)* vector are smaller as the information that the attribute gives about the class is greater. This statement can be proved from the expression of the mutual information [6], that is $I(X_i,Y) = I(Y,X_i) = H(Y) - H(Y/X_i)$ where $H(Y)$ is the entropy of Y and $H(Y/X_i)$ is the conditional entropy of Y when X_i is known.

Given a set of attributes and the associated Transinformation matrix, the GD measure fulfils the following properties:

1. $d_{GD}(X,Y) \geq 0$, $\forall X,Y$ and $d_{GD}(X,X) = 0$
2. $d_{GD}(X,Y) = d_{GD}(Y,X)$, $\forall X,Y$

The demonstration of the two previous properties is trivial if we take into account the properties of the distance (2) and the properties of the Transinformation matrix [18]. The triangle inequality property has not been demonstrated for the GD measure yet, and so it can be considered as an semimetric distance function. The GD measure satisfies the monotonicity property that states that the measure increases with dimensionality. This property is easily probed using the expression 4 and the properties of the Transinformation matrix. Therefore, only subsets with the same cardinality can be compared between one another.

The use of GD measure for feature selection is based on the fact that the distances $d_{LM}(X_i, Y)$ decreases as the information of an attribute subset about the class increases. On the other hand, if an element of the Transinformation matrix is large (it indicates that the mutual dependence between two attributes is high) then the GD measure increases. Therefore it can be concluded that a low value of the GD measure between an attribute subset and the class indicates that the attributes give a lot of information about the class and that high dependencies do not exist between the attributes.

In the GD measure an important aspect is the singularity of the Transinformation matrix. It has been analytically demonstrated that the Transinformation matrix is non singular for dimension two and three. An analytical demonstration have not been found for greater dimensions yet, but all the matrices generated in the examples of Sect. 3 were found non-singular.

3 Results

In this section we compare the quality of the selected attributes by the GD measure with the selected attributes by other two methods. The other two methods chosen in the comparative are the Mántaras distance and the ReliefF algorithm. On the one hand, the Mántaras distance has been chosen because it has a conceptual resemblance with the GD measure. On the other hand, the ReliefF method has been chosen because it has been widely referenced in the bibliography [5, 22, 12]. The ReliefF method is a version of the Relief method due to Kononenko [15] that allows attributes with missing values and multiclass problems. The quality of each selected attribute subset was tested by means of the accuracy that three classifiers yields. As we are interested in comparing the selection methods, we do not make any optimization in the classifiers to avoid bias in the accuracy due to the optimizations. The classifiers are: the Naive Bayes classifier [8], a decision tree induced with the ID3 method [21] and the IB1 algorithm [2]. The implementation of the induction algorithms was carried out with the $\mathcal{MLC}$++ library [13] and the comparison was performed with 19 databases of the UCI Machine Learning Databases Repository [19]. The databases are the following ones: Breast Cancer Ljubljana (BC), Breast Cancer Wisconsin (BW), Credit Card (CR), Glass (GL), Glass2 (G2), Heart Disease (HD), Iris (IR), Led (LE), Liver Disorder (LD), Monks (M1,M2,M3), Parity5+5 (P5), Pima Indian Diabetes (PI), Post-operative (PO), Tic-Tac-Toe (TT), Voting (VO), Wine (WI), Zoo (ZO).

Table 1. Results on Naive Bayes classifier and branch-and-bound

	d_{GD}		*ReliefF*			d_{LM}		
	# attr.	Acc.	# attr.	Acc.	Concl.	# attr.	Acc.	Concl.
BC	4	74.09±0.80	8	74.13±0.73	=	7	74.05±0.79	=
BW	10	96.30±0.22	7	96.74±0.23	<	10	96.30±0.22	=
CR	1	86.37±0.37	1	86.37±0.37	=	1	86.37±0.37	=
GL	8	48.08±1.04	8	48.08±1.04	=	3	49.18±1.10	=
G2	4	69.92±1.10	4	65.11±1.10	>	2	61.67±1.10	>
HD	8	85.04±0.62	11	85.45±0.65	=	12	85.15±0.62	=
IR	2	96.00±0.49	2	96.00±0.49	=	2	96.00±0.49	=
LE	7	74.99±0.44	7	74.99±0.44	=	7	74.99±0.44	=
LD	6	55.58±0.82	3	59.45±0.91	<	6	55.58±0.82	=
M1	1	75.00±0.60	3	75.00±0.60	=	3	75.00±0.60	=
M2	1	67.14±0.74	1	67.14±0.74	=	1	67.14±0.74	=
M3	2	97.22±0.23	2	97.22±0.23	=	4	97.22±0.23	=
P5	1	45.13±0.24	1	44.84±0.24	=	1	45.13±0.24	=
PI	6	75.67±0.46	2	76.60±0.50	<	4	75.70±0.47	=
PO	1	70.49±1.68	1	71.18±1.64	<	1	69.47±1.67	>
TT	6	72.12±0.43	6	73.00±0.43	<	6	73.02±0.44	<
VO	1	95.63±0.28	1	95.63±0.28	=	1	95.63±0.28	=
WI	6	97.53±0.34	13	97.43±0.34	=	12	97.59±0.35	=
ZO	7	93.65±0.74	8	93.75±0.73	=	8	93.75±0.73	=

All the previous databases have less than 10% of samples with missing values that were removed. With respect to the continuous attributes, they were discretized with the simple equal width discretization method with 10 intervals. The process we follow to test the quality of the attributes selected with the GD measure was the following. For all datasets, we selected the best attribute subset according to each of the three methods compared. Then we estimate the accuracy that each classifier yields with the selected attributes. The accuracy was estimated taking the mean of ten runs of a 10 k-fold cross validation.

Due to the monotonicity property we implement the branch-and-bound method [20] to search the subset with minimum value of GD measure. For the ReliefF algorithm we sort the attributes in decreasing order of relevance and take in each case the number of attributes we were considering. With the Mántaras distance we do the same but sorting the attributes in increasing value of the distance. The best results obtained for each classifier are shown in Tables 1, 2 and 3.

To assess the results, two paired t statistical tests with a confidence level of 90% were accomplished. Under the null hypothesis of the first statistical test, the subsets selected by the two methods yield the same accuracy. If the null hypothesis of this first statistical test is rejected, another statistical test is accomplished in which the null hypothesis is that the subset selected by the GD measure yields lower or equal to the another method.

Table 2. Results on ID3 classifier and branch-and-bound

	d_{GD}		*ReliefF*			d_{LM}		
	# attr.	Acc.	# attr.	Acc.	Concl.	# attr.	Acc.	Concl.
BC	4	75.77±0.70	1	72.50±0.80	>	4	75.05±0.77	=
BW	2	95.32±0.23	3	95.46±0.23	=	10	94.72±0.29	>
CR	1	86.37±0.37	1	86.37±0.37	=	1	86.37±0.37	=
GL	6	69.72±0.96	6	69.72±0.96	=	8	69.27±0.98	=
G2	5	82.56±0.96	3	83.42±0.95	=	6	88.20±0.81	<
HD	4	79.07±0.71	5	81.41±0.72	<	8	78.96±0.75	=
IR	2	95.93±0.49	2	95.93±0.49	=	2	95.93±0.49	=
LE	7	73.44±0.42	7	73.44±0.42	=	7	73.44±0.42	=
LD	4	64.37±0.78	4	64.37±0.78	=	5	63.51±0.75	=
M1	5	99.93±0.07	3	100.0±0.00	=	5	99.93±0.07	=
M2	5	77.80±0.69	5	77.80±0.69	=	5	77.80±0.69	=
M3	5	100.0±0.00	6	100.0±0.00	=	5	100.0±0.00	=
P5	8	99.96±0.04	5	100.0±0.00	=	8	99.96±0.04	=
PI	8	70.65±0.49	8	70.65±0.49	=	8	70.65±0.49	=
PO	4	70.56±1.62	1	71.18±1.64	=	1	69.47±1.67	=
TT	9	85.46±0.33	9	85.46±0.33	=	9	85.46±0.33	=
VO	1	95.63±0.28	1	95.63±0.28	=	1	95.63±0.28	=
WI	3	94.61±0.49	5	95.28±0.53	=	3	94.61±0.49	=
ZO	9	96.42±0.59	8	96.03±0.60	=	8	96.03±0.60	=

The conclusions obtained from these statistical tests appear in Tables 1, 2, 3 under the column labeled "Concl.".

If we compare the accuracy of the three classifiers (Naive Bayes, ID3 and IB1) with the three selection methods (GD measure, ReliefF and Mántaras distance) on the 19 databases we get 114 results.

In 9 (7.9%) of the 114 cases, the set of attributes selected by the GD measure yields better accuracy than the two other methods. In 90 (78.9%) of the 114 cases, the set of attributes selected by the GD measure yields a accuracy that is equal to the two other methods. Considering the comparative with the two methods separatively we get that with respect to the ReliefF method in 4 (7%) of the cases the set of attributes selected by the GD measure is better than the selected by ReliefF and in 43 (75.4%) is equal. On the other hand, with respect to the Mántaras distance in 5 (8.8%) of the cases the results obtained by the GD measure improve the results of the Mántaras distance, and in 47 (82.4%) cases the results are equals.

After the previous global evaluation of the results, we are going to focus on two databases where the ability of the GD measure for filtering irrelevant and redundant attributes is stated. The BW database has a completely irrelevant attribute that is the identifier of each sample. This attribute has been the last selected attribute by ReliefF and the GD measure whereas the Mántaras distance selects it firstly. In the CRX database, the attributes A4 and A5 are

Table 3. Results on IB1 classifier and branch-and-bound

	d_{GD}		*ReliefF*			d_{LM}		
	# attr.	Acc.	# attr.	Acc.	Concl.	# attr.	Acc.	Concl.
BC	6	73.53±0.73	8	74.03±0.72	=	7	74.08±0.75	=
BW	9	95.81±0.23	5	96.38±0.21	<	7	95.90±0.23	=
CR	11	83.49±0.43	9	82.92±0.40	=	1	82.31±1.23	=
GL	6	76.53±0.89	6	76.53±0.89	=	9	68.59±0.99	>
G2	5	87.85±0.73	5	87.85±0.73	=	4	86.68±0.83	=
HD	11	80.78±0.69	5	79.30±0.73	=	8	76.52±0.71	>
IR	4	95.73±0.50	4	95.73±0.50	=	4	95.73±0.50	=
LE	7	64.81±0.66	7	64.81±0.66	=	7	64.81±0.66	=
LD	5	66.01±0.81	5	66.01±0.81	=	5	66.01±0.81	=
M1	5	99.79±0.12	3	100.0±0.00	<	5	99.79±0.12	=
M2	5	67.01±0.70	5	67.01±0.70	=	5	67.01±0.70	=
M3	5	99.66±0.15	2	96.94±0.28	>	5	99.66±0.15	=
P5	8	99.92±0.05	5	100.0±0.00	=	8	99.92±0.05	=
PI	8	70.63±0.44	8	70.63±0.44	=	8	70.63±0.44	=
PO	1	50.08±2.37	2	56.52±1.69	<	1	54.76±2.74	<
TT	9	80.77±0.38	8	81.98±0.37	<	8	81.94±0.37	<
VO	14	93.98±0.33	10	93.77±0.33	=	10	93.84±0.39	=
WI	11	97.54±0.37	11	96.69±0.41	>	9	98.26±0.32	<
ZO	11	97.71±0.44	10	97.03±0.48	=	13	97.51±0.48	=

completely correlated and one of them has been removed in some distributions of this database. This correlation between attributes A4 and A5 is detected by the GD measure and selects the A5 attribute in last position, however ReliefF and the Mántaras distance do not take into account the correlation between the attributes and they select both of them before other attributes.

In Table 2, it can be noticed that the attribute selection do not improve significantly the performance of the induced decision tree, in 89.5% of cases using ID3 we get the same accuracy with the three methods. However, the attribute selection can reduce the size of the induced tree. To assess this statement, we accomplish two statistical tests similar to the used with the accuracy but now making reference to the number of nodes of the generated decision trees. Focusing on the 34 cases of databases of Table 2 whose accuracies are not statistically different, we observe that in 5 (14.7%) cases the trees induced with the attribute set obtained with the GD measure has less nodes than the obtained with the attributes selected by the other two methods, and in 22 (64.7%) the number of nodes is equal.

With the Naive Bayes classifier, when the GD measure yields the same accuracy that the other methods the cardinality of the selected attribute sets is smaller. So the computational cost of the induced classifier will be lower, giving the same accuracy. On the contrary, the GD measure and Mántaras distance use more attribute than ReliefF to yield the same accuracy. This fact is due to the

nature of the ReliefF algorithm that is very similar to the schema of the IB1 classifier.

4 Conclusions

From the above results in the datasets, we conclude that the GD measure does not improve the performance of the ReliefF algorithm significatively. Although the number of selected attributes is smaller and so the induced classifier will have a lower computational cost. Also, we have found that the proposed measure detects irrelevant and redudant attributes where the other two methods fail.

On the other hand, the results obtained with the GD measure improve the results obtained with the Mántaras distance and with less attributes, so we can conclude that the use of the Transinformation matrix improve the performance of the Mántaras distance.

The smaller cardinality of the selected attribute sets is due to the nature of the GD measure that allows to detect the dependencies between attributes. So, redundant attributes are rejected and a reduction of the cardinality of the selected sets is achieved.

As future works we are interested in testing the performance of the GD measure with other discretization methods as the proposed by Fayyad and Irani [9]. Another important point to test is the strength of the measure with datasets with missing values. Finally, it is interesting to find a method that permits to establish the dimension of the attribute subset from the GD measure automatically.

Acknowledgements

This work was supported by the Spanish Ministry of Education under project TAP95-0288. We want to thank Prof. López de Mántaras for his helpful suggestions on the experiments of this work.

References

1. D. W. Aha and R. L. Bankert. Feature selection for case-based classification of cloud types: An empirical comparison. In *Proc. of the 1994 AAAI Workshop on Case-Based Reasoning*, pages 106–112. AAAI Press, 1994.
2. D. W. Aha, Dennis Kibler, and M. K. Albert. Instance-based learning algorithms. *Machine Learning*, 6:37–66, 1991.
3. H. Almuallim and T.G. Dietterich. Learning with many irrelevant features. In *Proc. of the Ninth National Conference on Artificial Intelligence*, pages 547–552. AAAI Press, 1991.
4. A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97:245–271, 1997.
5. R. Caruana and D. Freitag. Greedy attribute selection. In *Proc. of the 11th International Machine Learning Conference*, pages 28–36, New Brunswick, NJ, 1994. Morgan Kaufmann.

6. T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley & Sons Inc., 1991.
7. Walter Daelemans and Antal van den Bosch. Generalization performance of backpropagation learning on a syllabification task. In *Proc. of the Third Twente Workshop on Language Technology*, pages 27–38, 1992.
8. R. Duda and P. Hart. *Pattern Classification and Scene Analysis.* John Willey and Sons, 1973.
9. U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proc. of the 13th Int. Joint Conference of Artificial Intelligence*, pages 1022–1027, 1993.
10. G. H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In W. William and Haym Hirsh, editors, *Procs. of the Eleventh International Conference on Machine Learning*, pages 121–129. Morgan Kaufmann, San Francisco, CA, 1994.
11. K. Kira and L. A. Rendell. The feature selection problem: Traditional methods and a new algorithm. In *Proc. of the 10th National Conf. on Artificial Intelligence*, pages 129–134, 1992.
12. R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, December 1997.
13. R. Kohavi, D. Sommerfield, and J. Dougherty. Data mining using MLC++: A machine learning library in C++. In *Tools with Artificial Intelligence*, pages 234–245. IEEE Computer Society Press, 1996. Received the best paper award.
14. D. Koller and M. Sahami. Toward optimal feature selection. In *Proc. of the 13th Int. Conf. on Machine Learning*, pages 284–292. Morgan Kaufmann, 1996.
15. I. Kononenko. Estimating attributes: Analysis and extensions of relief. In F. Bergadano and L. de Raedt, editors, *Machine Learning: ECML-94*, pages 171–182, Berlin, 1994. Springer.
16. N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.
17. R. Lopez de Mántaras. A distance-based attribute selection measure for decision tree induction. *Machine Learning*, 6:81–92, 1991.
18. J. Lorenzo, M. Hernández, and J. Méndez. GD: A Measure based on Information Theory for Attribute Selection. In Helder Coelho, editor, *Proc. of the 6th Ibero-American Conference on Artificial Intelligence*, Lectures Notes in Artificial Intelligence, Springer Verlag, 1998.
19. C. J. Merz and P.M. Murphy. *UCI Repository of machine learning databases.* Irvine, CA: University of California, Department of Information and Computer Science., 1996.
20. P.M. Narendra and K. Fukunaga. A branch and bound algorithm for feature selection. *IEEE Trans. on Computers*, 26:917–922, 1977.
21. J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.
22. D. Wettschereck and D. W. Aha. Weighting features. In *Proc. of the First Int. Conference on Case-Based Reasoning*, pages 347–358, 1995.
23. D. Wettschereck and T. G. Dieterich. An experimental comparison of the nearest-neighbor and nearest-hyperrectangle algorithms. *Machine Learning*, 19:5–27, 1995.
24. A. P. White and W. Z. Liu. Bias in information-based measures in decision tree induction. *Machine Learning*, 15:321–329, 1994.