Using Information Gain Attribute Evaluation to Classify Sonar Targets

Jasmina Novakovic

***Abstract* – This paper presents an application of Information Gain (IG) attribute evaluation to the classification of the sonar targets with C4.5 decision tree. C4.5 decision tree has inherited ability to focus on relevant features and ignore irrelevant ones, but such method may also benefit from independent feature selection. In our experiments, IG attribute evaluation significantly improves C4.5 decision tree. This research also shows that feature selection helps increase computational efficiency while improving classification accuracy.**

***Keywords* - classification accuracy, C4.5 decision tree, feature selection, IG attribute evaluation.**

I. INTRODUCTION

Feature selection is a fundamental problem in many different areas. For some problems, all features may be important, but for some target concept, only a small subset of features is usually relevant.

Feature selection reduces the dimensionality of feature space, removes redundant, irrelevant, or noisy data. It brings the immediate effects for application: speeding up a data mining algorithm, improving the data quality and thereof the performance of data mining, and increasing the comprehensibility of the mining results.

Finding the best feature subset is usually intractable [1] and many problem related to feature selection have been shown to be NP-hard [2]. Feature selection has been a fertile field of research and development since 1970's in statistical pattern recognition [3]-[5], machine learning and data mining [6]-[11].

Feature selection algorithms may be divided into filters [12], [13], wrappers [1] and embedded approaches [6].

Some classification algorithms have inherited ability to focus on relevant features and ignore irrelevant ones. Decision trees are primary example of a class of such algorithms [14], [15], but also multi-layer perceptron (MLP) neural networks with strong regularization of the input layer may exclude the irrelevant features in an automatic way [16].

Jasmina Novakovic, Faculty of Computer Science, Megatrend University, Bulevar Umetnosti 29, 11070 Beograd, Serbia, (e-mail: jnovakovic@megatrend.edu.rs).

Such methods may also benefit from independent feature selection. On the other hand, some algorithms have no provisions for feature selection. The k-nearest neighbor algorithm is one family of such methods that classify novel examples by retrieving the nearest training example, strongly relaying on feature selection methods to remove noisy features.

Section II presents general feature selection structure. Section III describes one of the ranking methods, IG attribute evaluation. Section IV presents C4.5 decision tree as supervised learning algorithm. Section V describes the experiments and results. Section VI concludes and gives future investigations.

II. GENERAL FEATURE SELECTION STRUCTURE

It is possible to derive a general architecture from most of the feature selection algorithms. General architecture consists of four basic steps: subset generation, subset evaluation, stopping criterion, and result validation [7]. The feature selection algorithms create a subset, evaluate it, and loop until an ending criterion is satisfied [17]. Finally, the subset found is validated with the classifier algorithm on real data.

Subset generation is a search procedure, it generates subsets of features for evaluation. The total number of candidate subsets is 2^N, where N is the number of features in the original data set, which makes exhaustive search through the feature space infeasible with even moderate N. Non deterministic search like evolutionary search is often used to build the subsets [18]. It is also possible to use heuristic search methods. There are two main families of these methods: forward addition [19] (starting with an empty subset, we add features after features by local search) or backward elimination (the opposite).

Each subset generated by the generation procedure needs to be evaluated by a certain evaluation criterion and compared with the previous best one with respect to this criterion. If it is found to be better, then it replaces the previous best subset. A simple method for evaluating a subset is to consider the performance of the classifier algorithm when it runs with that subset. The method is

classified as a wrapper, because in this case, the classifier algorithm is wrapped in the loop. On the contrary, filter methods do not rely on the classifier algorithm, but use other criteria based on correlation notions.

The feature selection process may run exhaustively before it stops without a suitable stopping criterion. A feature selection process may stop under one of the following reasonable criteria: (1) a predefined number of features are selected, (2) a predefined number of iterations are reached, (3) whether addition (or deletion) of any feature does not produce a better subset, (4) an optimal subset according to the evaluation criterion is obtained.

The selected best feature subset needs to be validated by carrying out different tests on both the selected subset and the original set and comparing the results using artificial data sets and real-world data sets.

III. Information Gain attribute evaluation

Diverse feature ranking and feature selection techniques have been proposed in the machine learning literature. The purpose of these techniques is to discard irrelevant or redundant features from a given feature vector. The following attribute evaluations are used: IG, gain ratio, symmetrical uncertainty, relief-F, one-R and chi-squared. In this paper, we consider evaluation of the practical usefulness of IG attribute evaluation.

Entropy is a commonly used in the information theory measure, which characterizes the purity of an arbitrary collection of examples. It is in the foundation of the IG attribute ranking methods. The entropy measure is considered as a measure of system's unpredictability. The entropy of Y is

$$H(Y) = -\sum_{y \in Y} p(y) \log_2 (p(y)) \qquad (1)$$

where $p(y)$ is the marginal probability density function for the random variable Y. If the observed values of Y in the training data set S are partitioned according to the values of a second feature X, and the entropy of Y with respect to the partitions induced by X is less than the entropy of Y prior to partitioning, then there is a relationship between features Y and X. Then the entropy of Y after observing X is:

$$H(Y/X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y/x) \log_2 (p(y/x)) \qquad (2)$$

where $p(y|x)$ is the conditional probability of y given x.

Given the entropy as a criterion of impurity in a training set S, we can define a measure reflecting additional information about Y provided by X that represents the amount by which the entropy of Y decreases. This measure is known as IG. It is given by

$$IG = H(Y) - H(Y/X) = H(X) - H(X/Y) \qquad (3)$$

IG is a symmetrical measure (refer to equation (3)). The information gained about Y after observing X is equal to the information gained about X after observing Y. A weakness of the IG criterion is that it is biased in favor of features with more values even when they are not more informative.

IV. C4.5 Decision Tree

Different methods exist to build decision trees, but all of them summarize given training data in a tree structure, with each branch representing an association between feature values and a class label. One of the most famous and representative amongst these is the C4.5 decision tree [20]. The C4.5 decision tree works by recursively partitioning the training data set according to tests on the potential of feature values in separating the classes. The decision tree is learned from a set of training examples through an iterative process, of choosing a feature and splitting the given example set according to the values of that feature. The most important question is which of the features is the most influential in determining the classification and hence should be chosen first. Entropy measures or equivalently, information gains are used to select the most influential, which is intuitively deemed to be the feature of the lowest entropy (or of the highest information gain). This learning algorithm works by: a) computing the entropy measure for each feature, b) partitioning the set of examples according to the possible values of the feature that has the lowest entropy, and c) for each are used to estimate probabilities, in a way exactly the same as with the Naive Bayes approach. Although feature tests are chosen one at a time in a greedy manner, they are dependent on results of previous tests.

V. EXPERIMENTS AND RESULTS

Connectionist Bench (Sonar, Mines vs. Rocks) data set was used for IG attribute evaluation with C4.5 decision tree, taken from the UCI repository of machine learning databases [20]. This is the data set used by Gorman and Sejnowski in their study of the classification of sonar signals using a neural network [21]. The task is to train a network to discriminate between sonar signals bounced off a metal cylinder and those bounced off a roughly cylindrical rock.

This data set contains 111 patterns obtained by bouncing sonar signals off a metal cylinder at various angles and under various conditions, and 97 patterns obtained from rocks under similar conditions. The transmitted sonar signal is a frequency-modulated chirp, rising in frequency. The data set contains signals obtained from a variety of different aspect angles, spanning 90 degrees for the cylinder and 180 degrees for the rock.

Each pattern is a set of 60 numbers in the range 0.0 to 1.0, where each number represents the energy within a particular frequency band, integrated over a certain period of time.

If the object is a rock, the label associated with each record contains the letter "R" and if it is a mine (metal cylinder) "M". The numbers in the labels are in increasing order of aspect angle, but they do not encode the angle directly.

Fig. 1 shows a sample return from the rock and the cylinder. The preprocessing of the raw signal was based on experiments with human listeners. The temporal signal was first filtered and spectral information was extracted and used to represent the signal on the input layer.

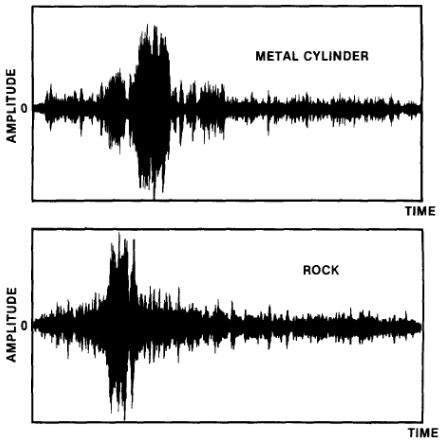

Fig. 1. Amplitude displays of a typical return from the cylinder and the rock as a function of time [21].

The preprocessing used to obtain the spectral envelope is indicated schematically in Fig. 2 where a set of sampling apertures (Fig. 2a) are superimposed over the 2D display of a short-term Fourier Transform spectrogram of the sonar return. The spectral envelope, $P_{t0,v0}(\eta)$, was obtained by integrating over each aperture (Fig. 2b and c).

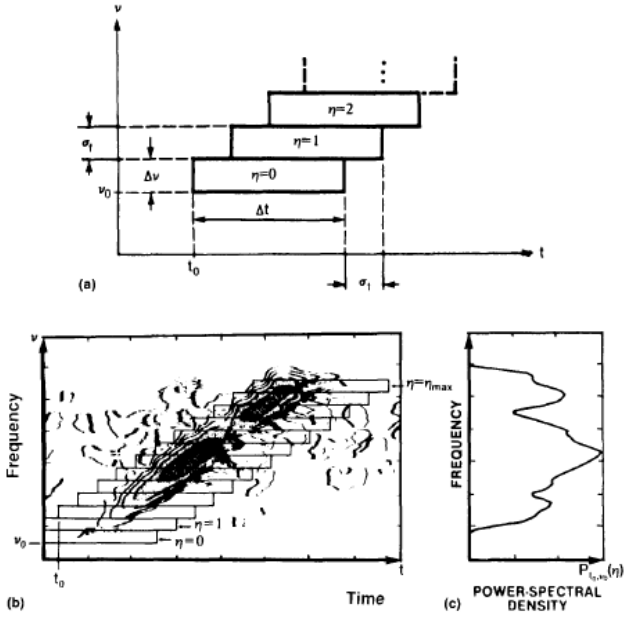

Fig. 2. The preprocessing of the sonar signal produces a sampled spectral envelope. (a) The set of sampling apertures offset temporally to correspond to the slope of the FM chirp, (b) sampling apertures superimposed over the 2D display of the short-term Fourier transform, (c) the spectral envelope obtained by integrating over each sampling aperture [21].

A supervised learning algorithm, C4.5 decision tree is adopted here to build model. The purpose of the experiments described in this section is to empirically test the claim that IG attribute evaluation can improve the accuracy of classification algorithm C4.5 decision tree. The performance of learning algorithms with and without feature selection is taken as an indication of IG attribute evaluation success in selecting useful features, because the relevant features are often not known in advance for natural domains. Classification accuracy was estimated using ten-fold cross validation.

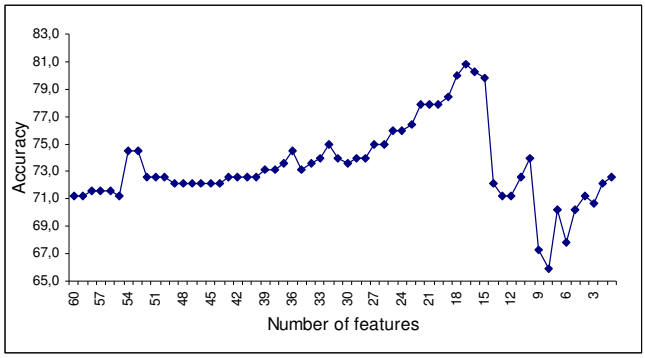

Fig. 3. Classification accuracy of C4.5 decision tree with IG attribute evaluation.

Fig. 3 shows for C4.5 decision tree, how much data set accuracy was improved and degraded by IG attribute evaluation. IG attribute evaluation maintains or improves the accuracy of C4.5 decision tree if we used more than 9 relevant features and degrades, maintains or improves its accuracy if we used less than 9 relevant features. The accuracy of C4.5 decision tree significantly improves more than 10% on this data set with IG attribute evaluation. Evaluation of selecting features is fast.

TABLE 1: GENERATING DECISION RULES

Number of most relevant features	*Number of leaves*	*Size of tree*
60 - 52	18	35
51 - 49	17	33
48 - 36	18	35
35 - 34	17	33
33 - 33	19	37
32 - 29	16	31
28 - 22	17	33
21 - 21	18	35
20 -18	19	37
17 - 17	20	39
16 - 14	18	35
13 - 13	19	37
12 - 11	20	39
10 - 10	23	45
9 - 9	21	41
8 - 8	19	37
7 - 5	14	27
4 - 4	8	15
3- 1	2	3

C4.5 decision tree without feature selections is generated 18 rules, and size of the tree is 35. Table 1 shows that IG attribute evaluation changes the size of the trees induced by C4.5 decision tree depends on number of most relevant features. Rules for this data set obtained by C4.5 decision tree without feature selections are:

If f_11 = <0.197 and f_1 <= 0.0392 and f_4 <= 0.0539 and f_28 <= 0.9578 and f_27 <= 0.2771 and f_2 <= 0.0378
Then Classification = Mine (2.0);

If f_11 = <0.197 and f_1 <= 0.0392 and f_4 <= 0.0539 and f_28 <= 0.9578 and f_27 <= 0.2771 and f_2 > 0.0378
Then Classification = Rock (2.0);

If f_11 = <0.197 and f_1 <= 0.0392 and f_4 <= 0.0539 and f_28 <= 0.9578 and f_27 > 0.2771
Then Classification = Rock (56.0);

If f_11 = <0.197 and f_1 <= 0.0392 and f_4 <= 0.0539 and f_28 > 0.9578 and f_2 <= 0.0253
Then Classification = Rock (2.0);

If f_11 = <0.197 and f_1 <= 0.0392 and f_4 <= 0.0539 and f_28 > 0.9578 and f_2 > 0.0253
Then Classification = Mine (3.0);

If f_11 = <0.197 and f_1 <= 0.0392 and f_4 > 0.0539 and f_21 <= 0.7894 and f_18 <= 0.2613
Then Classification = Mine (2.0);

If f_11 = <0.197 and f_1 <= 0.0392 and f_4 > 0.0539 and f_21 <= 0.7894 and f_18 > 0.2613
Then Classification = Rock (6.0);

If f_11 = <0.197 and f_1 <= 0.0392 and f_4 > 0.0539 and f_21 > 0.7894
Then Classification = Rock (6.0);

If f_11 = <0.197 and f_1 > 0.0392
Then Classification = Mine (8.0/1.0);

If f_11 > 0.197 and f_27 <= 0.8145 and f_54 <= 0.0205 and f_53 <= 0.0166 and f_21 <= 0.5959
Then Classification = Rock (14.0);

If f_11 > 0.197 and f_27 <= 0.8145 and f_54 <= 0.0205 and f_53 <= 0.0166 and f_21 > 0.5959 and f_51 <= 0.0153 and f_23 <= 0.7867
Then Classification = Rock (13.0/1.0);

If f_11 > 0.197 and f_27 <= 0.8145 and f_54 <= 0.0205 and f_53 <= 0.0166 and f_21 > 0.5959 and f_51 <= 0.0153 and f_23 > 0.7867
Then Classification = Mine (6.0/1.0);

If f_11 > 0.197 and f_27 <= 0.8145 and f_54 <= 0.0205 and f_53 <= 0.0166 and f_21 > 0.5959 and f_51 > 0.0153
Then Classification = Mine (7.0);

If f_11 > 0.197 and f_27 <= 0.8145 and f_54 <= 0.0205 and f_53 > 0.0166
Then Classification = Mine (12.0/1.0);

If f_11 > 0.197 and f_27 <= 0.8145 and f_54 > 0.0205
Then Classification = Mine (13.0);

If f_11 > 0.197 and f_27 > 0.8145 and f_8 <= 0.0697 and f_2 <= 0.0222
Then Classification = Mine (3.0);

If f_11 > 0.197 and f_27 > 0.8145 and f_8 <= 0.0697 and f_2 > 0.0222
Then Classification = Rock (2.0);

If f_11 > 0.197 and f_27 > 0.8145 and f_8 > 0.0697
Then Classification = Mine (51.0);

The experiments presented in this article show that IG attribute evaluation's ability to select useful features improves C4.5 decision tree.

VI. CONCLUSIONS

IG attribute evaluation may filter features leading to reduce dimensionality of the feature space. In our experiments, IG attribute evaluation significantly improves C4.5 decision tree, in spite of the fact that C4.5 decision tree has inherited ability to focus on relevant features and ignore irrelevant ones. In this research feature selection helps increase computational efficiency while improving classification accuracy. These conclusions will be tested on larger data sets using various classification algorithms in the near future.

REFERENCES

[1] Kohavi, R., and John, G.H. "Wrappers for feature subset selection", *Artificial Intelligence*, vol. 97, 1997, 273-324.

[2] Blum, A.L., and Rivest, R.L. "Training a 3-node neural networks is NP-complete", *Neural Networks*, 5:117–127, 1992.

[3] Wyse, N., Dubes, R., and Jain, A.K. "A critical evaluation of intrinsic dimensionality algorithms", In E.S. Gelsema and L.N. Kanal, editors, *Pattern Recognition in Practice*, Morgan Kaufmann Publishers, Inc., 1980, 415–425.

[4] Ben-Bassat, M. "Pattern recognition and reduction of dimensionality", In P. R. Krishnaiah and L. N. Kanal, editors, *Handbook of statistics-II*, North Holland, 1982, 773–791.

[5] Siedlecki, W., and Sklansky, J. "On automatic feature selection", *International Journal of Pattern Recognition and Artificial Intelligence*, 2:197–220, 1988.

[6] Blum, A.I., and Langley, P. "Selection of relevant features and examples in machine learning", *Artificial Intelligence*, vol 97, 1997, 245-271.

[7] Dash, M., and Liu, H. "Feature selection methods for classifications", *Intelligent Data Analysis: An International Journal*, 1(3), 1997. http://www-east.elsevier.com/ida/free.htm.

[8] Dy, J. G., and Brodley, C. E. "Feature subset selection and order identification for unsupervised learning". In *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000, 247–254.

[9] Kim, Y., Street, W., and Menczer, F. "Feature selection for unsupervised learning via evolutionary search", In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000, 365–369.

[10] Das, S. "Filters, wrappers and a boosting-based hybrid for feature selection", In *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

[11] Mitra, P., Murthy, C. A., and Pal, S. K. "Unsupervised feature selection using feature similarity", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):301–312, 2002.

[12] Almuallim, H., and Dietterich, T.G. "Learning with many irrelevant features", In: *Proc. AAAI-91*, Anaheim, CA, 1991, 547-552.

[13] Kira, K., and Rendell, L.A. "The feature selection problem: tradional methods and a new algorithm", In: *Proc. AAAI-92*, San Jose, CA, 1992, 122-126.

[14] Breiman, L., Friedman, J.H., Olshen, R.H., and Stone, C.J., *Classification and regression trees*, Wadsworth and Brooks, Monterey, CA, 1984.

[15] Quinlan, J.R. *C4.5: Programs for machine learning*, San Mateo, Morgan Kaufman, 1993.

[16] Duch, W., Adamczak, R., and Grabczewski, K. "A new methodology of extraction, optimization and application of crisp and fuzzy logical rules", *IEEE Transactions on Neural Networks*, vol. 12, 2001, 277-306.

[17] Liu, H., and Motoda, H. *Feature selection for knowledge discovery and data mining*, Kluwer Academic Publishers, 1998.

[18] Yang, J., and Honavar, V. "Feature subset selection using a genetic algorithm", *IEEE Intelligent Systems* 13:44-49, 1998.

[19] Koller, D., and Sahami, M. "Toward optimal feature selection", In *International Conference on Machine Learning*, 1996, 284-292.

[20] Merz, C. J. and Murphy P. M. *UCI Repository of machine learning databases*, http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998.

[21] Gorman, R. P., and Sejnowski, T. J. "Analysis of hidden units in a layered network trained to classify sonar targets" in Neural Networks, Vol. 1, 1988, pp. 75-89.