A Geometric Approach to Feature Selection

Tapio Elomaa and Esko Ukkonen

Department of Computer Science, P. O. Box 26 (Teollisuuskatu 23)
FIN-00014 University of Helsinki, Finland
{elomaa, ukkonen}@cs.helsinki.fi

Abstract. We propose a new method for selecting features, or deciding on splitting points in inductive learning. Its main innovation is to take the positions of examples into account instead of just considering the numbers of examples from different classes that fall at different sides of a splitting rule. The method gives rise to a family of feature selection techniques. We demonstrate the promise of the developed method with initial empirical experiments in connection of top-down induction of decision trees.

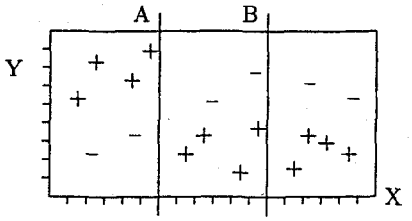
1 Introduction

Traditional feature selection methods are based on statistical evaluation heuristics, which base their ranking solely on the observed frequencies of instances. I.e., some information content measure (e.g., entropy) is computed for the splitting of the training set a feature introduces if applied [1, 4]. Typically, no other information, except an example's value for the feature is utilized in this decision-making. Alternative methods are given in [2, 3]. In this paper we introduce a method that also considers the position of examples in the instance space as the basis of its decision-making. We concentrate on decision tree learning in this paper, even though our method can be applied to other inductive learning schemes, like "separate-and-conquer" rule learners, as well.

The main idea of our technique is to take pairs of examples that lie (roughly) perpendicularly to a splitting boundary into account when ranking the splitting rule. We try to avoid dispersing example clusters in which the members are all of the same class. On the other hand, we try to place the boundary in between examples from different classes. Empirical comparison with the information gain heuristic demonstrates that the new method can improve the prediction accuracy and reduce the size of the produced decision trees in some domains.

2 Rationale for inspecting the frontiers

Consider the two-dimensional space of the following figure, where we have two choices of a linear separator to be applied as splitting rule: A and B. The two features X and Y are of ordered type. Using only frequencies there is no way to discriminate between A and B. However, we can easily see that A would be a much better choice than B, since it readily clusters the examples in the 'opposite' direction, while B breaks such clusters needlessly. If split B is chosen, then there does not exist any linear division along the dimension Y leading to a perfect clustering.



How could we tell A apart from B without breaking promising clusters in the dimension Y? Our approach is to compute the number of pairs of examples that fall at different sides of a splitting rule. Every pair of examples from the same class that the rule breaks up diminishes its goodness count, while breaking up a pair of examples of different classes is rewarded. We do not count all example pairs but, rather, limit our attention to certain regions of the space. These regions of our interest are the *frontiers* of a splitting rule.

3 A geometric approach to feature selection

Let us now define the evaluation function for features based on the geometric properties of the data. An *instance* is an n-dimensional vector of feature values. An *example* is an instance with an associated classification. Let a be an example and X a feature. Then, by $a(X)$ we denote the projection of vector a into the dimension X, i.e., the value of feature X in the example a, and by a_X we denote the vector a excluding its dimension X.

We are considering only binary splits on ordered features at the moment. Hence, a splitting rule S concerning the value of feature X is of the form $X \leq \textbf{constant}$. Geometrically the inequality defines a halfspace in the instance space. Let $S(a)$ be the truth value of applying the splitting rule S to the example a. By $S(X)$ we denote the constant appearing in the splitting rule S.

We define that a pair of examples (a, b) is an *internal pair* (of class C) if both a and b are of the same class (C), otherwise (a, b) is an *external pair*. The pair (a, b) belongs to the (k, l)-*frontier* of a splitting rule S concerning the value of feature X if

$$d(a_X, b_X) \leq l \text{ and } |S(X) - a(X)| \leq k \text{ and } |S(X) - b(X)| \leq k \text{ and } S(a) \neq S(b).$$

The function d approximates the examples' deviation from the perpendicular direction. In other words, a pair of examples belongs to the (k, l)-frontier of a splitting rule if the line connecting the two examples (as determined by the function d) deviates from the direction of the normal of the splitting boundary by at most an angle l, both are within the distance k from the splitting boundary in the dimension under consideration, and they fall at different sides of the hyperplane. Varying the distance bound k and the angle bound l gives rise to a family of techniques for frontier inspection.

There are several possible ways to choose the distance measure d. We elect to use threshold (m-of-n) functions; that is, we require that m features out of the total n features under consideration must be equal. Then the first condition of (k, l)-frontier takes the form $d(a_X, b_X) \leq n - m$, where $n = \|a\| - 1$. Threshold functions let us

circumvent the problem of nonuniformity of feature ranges. They remove the need of normalizing the feature ranges.

A splitting rule S interfears with a cluster of examples in the direction of its normal least when it breaks up as few internal pairs in its (k,l)-frontier as possible. Furthermore, segmentation along the dimension under consideration is at greatest when as many external pairs as possible are broken up. To evaluate a splitting rule S we subtract the number, $I_{k,l}(S)$, of internal pairs within the (k,l)-frontier of S from the number, $E_{k,l}(S)$, of external pairs within its (k,l)-frontier. We try to maximize this difference. That is, the splitting rule S registering the largest value for the function

$$Q(S) = E_{k,l}(S) - I_{k,l}(S)$$

is chosen as the new branching rule.

4 Discussion

Multiway splitting does not pose a major problem in case of ordered features, since the subsets introduced by the splitting rule will also be ordered along the dimension under consideration. Hence, if the data is split into n subsamples we apply the binary splitting procedure to the $n-1$ pairs of adjoining subsets and sum together all values $Q(S_i)$, where S_i is one of the $n-1$ splitting boundaries of the rule.

With nominal features we cannot set any bounds to the distance of two examples. Moreover, we cannot tell which of the subsets of the training data introduced by the splitting lie next to each other when there are more than two possible values. However, computing the deviance of an example from the normal of the split is not problematic when threshold functions are applied.

Since the distance of two examples cannot be bound on nominal features we elect to treat them as a special case: we do not try to bind the distance but, rather, treat the (k,l)-frontier as if it was (∞,l)-frontier for nominal features. Deciding on the adjacency of subsets introduced by splitting on a nominal feature is a bit more difficult problem. We could choose a random order of the subsamples and treat it as a set of ordered subsamples, or we could go through all the possible orders, count the value of function Q for each of them and choose a suitable measure out of the function values (e.g., the best function value, the average over all function values). The random ordering scheme is applied in the subsequent experiments.

5 Experiments

The algorithms in the following experiments are an implementation of ID3 [4] with information gain as the feature selection heuristic, ID3G is the basic ID3 with the addition of the geometric heuristic to resolve ties of the gain heuristic, and finally, in IDG the geometric heuristic has replaced the information gain; otherwise the program is unchanged, in particular the χ^2 test is still in action.

The domains are two Boolean functions: six-bit multiplexor and four-bit exclusive-or with four irrelevant bits; the difficult (the second) domain of the MONK's problems [5]; a breast cancer domain from the Helsinki University Central Hospital and four of the standard UCI repository databases.

Table 1. The average (over 20 runs) accuracies and sizes of the trees built by the three algorithms on test data and the time taken to build the tree.

	Accuracy (%)			Size (# of nodes)			Time (1/100 s.)		
Domain	ID3	ID3G	IDG	ID3	ID3G	IDG	ID3	ID3G	IDG
6-Multiplexor	78.6	80.0	98.1	35.0	34.4	21.2	1.5	2.8	4.1
XOR 4+4	70.3	81.2	100.0	156.7	116.9	31.0	10.3	21.5	57.3
MONK 2	69.2	69.7	78.2	173.0	171.0	179.0	6.0	8.5	38.4
LED	68.9	67.0	67.6	103.3	103.0	111.1	6.2	7.5	35.3
Voting	94.4	94.8	92.8	25.5	25.9	90.2	8.0	12.4	291.3
Hepatitis	78.7	73.7	84.1	45.4	47.1	66.5	3.5	8.3	176.9
Tumor	25.7	25.0	29.5	199.3	205.7	437.9	21.7	69.6	665.7
Breast cancer	54.2	53.0	54.5	399.1	381.9	992.4	29.8	69.6	374.3
Average	67.5	68.1	75.6	142.2	135.7	241.2	10.9	25.0	205.4

From Table 1 we observe that in the first three domains, which contain all possible feature value combinations, the geometric approach achieves significant advantage over the information gain in prediction accuracy. The induced decision trees are either optimal or close to optimal in size for these concepts. IDG takes only some 3–6 times the construction time of ID3 in tree building. No angle relaxation is required in these complete domains; i.e., the angle bound $l = 0$.

IDG's behavior in the remaining five test domains is determined by the characteristics of the domain. In LED and Tumor domains, which have many possible classes (10 and 22, respectively), it is neither able to increase the classification accuracy significantly nor able to reduce the classifier size by relaxing the angle bound. In the Hepatitis domain good results in both respects are achieved, while in the Voting domain only the tree size reduces by angle relaxation. What is the significant difference of these two domains is not yet clear to us. In the Breast cancer domain, on the other hand, some advantage in tree accuracy can be achieved by angle relaxation at the expense of increased classifier complexity.

Our initial experiments with the new method clearly demonstrate the potential of it. However, further analysis and development is required before consistent behavior can be expected.

References

1. Breiman, L., Friedman, J., Olshen, R., Stone, C.: *Classification and Regression Trees.* Wadsworth, Pacific Grove, CA, 1984
2. Fayyad, U., Irani, K.: The attribute selection problem in decision tree generation. *Proc. Tenth National Conference on Artificial Intelligence* (pp. 104–110). Morgan Kaufmann, San Mateo, CA, 1992
3. Kira, K., Rendell, L.: A practical approach to feature selection. *Proc. Ninth Intl. Workshop on Machine Learning* (pp. 249–256). Morgan Kaufmann, San Mateo, CA, 1992
4. Quinlan, R.: Induction of decision trees. Mach. Learn. 1 (1986) 81–106
5. Thrun, S. *et al.*: The MONK's problems – a performance comparison of different learning algorithms. Report CMU-CS-91-197. Carnegie Mellon University