Estimating Attributes: Analysis and Extensions of RELIEF

Igor Kononenko

University of Ljubljana, Faculty of Electrical Engineering & Computer Science,
Tržaška 25, SLO-61001 Ljubljana, Slovenia
e-mail: igor.kononenko@ninurta.fer.uni-lj.si

Abstract. In the context of machine learning from examples this paper deals with the problem of estimating the quality of attributes with and without dependencies among them. Kira and Rendell (1992a,b) developed an algorithm called RELIEF, which was shown to be very efficient in estimating attributes. Original RELIEF can deal with discrete and continuous attributes and is limited to only two-class problems. In this paper RELIEF is analysed and extended to deal with noisy, incomplete, and multi-class data sets. The extensions are verified on various artificial and one well known real-world problem.

1 Introduction

This paper deals with the problem of estimating the quality of attributes with strong dependencies to other attributes which seems to be the key issue of machine learning in general. Namely, for particular problems (e.g. parity problems of higher degrees) the discovering of dependencies between attributes may be unfeasible due to combinatorial explosion. In such cases efficient heuristic algorithms are needed to discover the dependencies.

Information gain was proposed as a measure for estimating the attribute's quality by Hunt et al. (1966) and later used by many authors (Quinlan, 1986). The idea is to estimate the difference between the prior entropy of classes C and posterior entropy, given values V of an attribute:

$$Gain = -\sum_C P(C)\log_2 P(C) - \sum_V \left(-P(V) \times \sum_C P(C|V)\log_2 P(C|V) \right) \quad (1)$$

Information gain and similar estimates like gini index (Breiman et al., 1984), distance measure (Mantaras, 1989), and j-measure (Smyth & Goodman, 1990) assume that attributes are independent and therefore are not applicable in domains with strong dependencies between attributes.

Kira and Rendell (1992a,b) developed an algorithm called RELIEF, which was shown to be very efficient in estimating attributes. The key idea of RELIEF is to estimate attributes according to how well their values distinguish among instances that are near each other. For that purpose RELIEF for a given instance searches for its two nearest neighbours: one from the same class (called *nearest*

hit) and the other from different class (called *nearest miss*). In fact, RELIEF's estimate $W[A]$ of attribute A is an approximation of the following difference of probabilities:

$$W[A] = P(\text{different value of A}|\text{nearest instance from different class}) - P(\text{different value of A}|\text{nearest instance from same class}) \quad (2)$$

The rationale is that good attribute should differentiate between instances from different classes and should have the same value for instances from the same class.

Original RELIEF can deal with discrete and continuous attributes and is limited to only two-class problems. It is not clear how RELIEF could be extended to deal with incomplete data and to deal with problems with more than two classes. Straightforward extensions do not give satisfactory results. In this paper RELIEF is analysed and extended to deal with noisy, incomplete, and multi-class data sets. The extensions are verified on various artificial and one well known real-world problem.

In the next section RELIEF is described and extended to k-nearest neighbours search. In section 3, we extend relief to deal with missing data and in section 4 to deal with multi-class problems. Results of experiments with extended versions of RELIEF on artificial data sets and one real world problem are discussed in section 5.

2 Estimating Probabilities with RELIEF

The original algorithm of RELIEF (Kira & Rendell, 1992a,b) is the following

```
set all weights W[A] := 0.0;
for i := 1 to m do
    begin
      randomly select an instance R;
      find nearest hit H and nearest miss M;
      for A := 1 to all_attributes do
        W[A] := W[A] - diff(A,R,H)/m + diff(A,R,M)/m;
    end;
```

where *diff(Attribute,Instance1,Instance2)* calculates the difference between the values of Attribute for two instances. For discrete attributes the difference is either 1 (the values are different) or 0 (the values are equal), while for continuous attributes the difference is the actual difference normalized to the interval $[0, 1]$. Normalization with m guarantees that all weights are in the interval $[-1, 1]$.

Function *diff* is used also for calculating the distance between instances to find the nearest neighbours. The total distance is simply the sum of differences over all attributes.

Obviously, the algorithm tries to approximate the difference (2). Parameter m represents the number of instances for approximating probabilities. The larger

m implies more reliable approximation. However, m cannot exceed the number of available training instances. The obvious choice for relatively small number of training instances is to set m to the upper bound and run the outer loop of the learning algorithms over all available training instances. In all our experiments we used this simplification of the algorithm.

The selection of the nearest neighbours is of crucial importance in RELIEF. The purpose is to find the nearest neighbours with respect to important attributes. Redundant and noisy attributes may strongly affect the selection of nearest neighbours and therefore the estimation of probabilities on noisy data becomes unreliable. To increase the reliability of probability approximation RELIEF can be extended to search for k-nearest hits/misses instead of only one near hit/miss. The extended version of the algorithm, called RELIEF-A, averages the contribution of k nearest hits/misses.

To estimate the contribution of more nearest neighbours we generated 3 data sets. All attributes in these data sets are binary with equal prior probabilities for both values ($P(V1) = P(V2) = 0.5$) except for random attributes which have various prior probability of values, which are however independent of the class. There are two equally probable classes ($P(C1) = P(C2) = 0.5$) and each data set has 200 instances. We compared the *intended information gain* of attributes with the estimates generated by RELIEF-A by calculating the standard linear *correlation coefficient*. The correlation coefficient can show how close are the intended quality and the estimated quality of attributes.

The intended information gain is the one calculated from probabilities that were used to generate artificial data sets. Note that due to random generator (and in later sections due to added noise and incomplete data) the factual information gain may differ from the intended information gain. However, in all our experiments the correlation between the intended information gain and the factual information gain was greater than 0.95. Besides, as we are interested in the "true" probability distribution that was "responsible" for generation of the data, we should consider the intended information gain as the target for an ideal estimator.

First data set contained 5 random binary attributes with different prior probability of values ($P(V1) = 0.5, 0.4, ..., 0.1$) and 5 independent informative attributes. The degrees of information gains were determined with the probability of value $V1$ given class $C1$ ($P(V1|C1) = 0.95, 0.85, ..., 0.55$) and given class $C2$ ($P(V1|C2) = 1 - P(V1|C1)$).

The second data set was obtained from the first data set by replacing each informative attribute with 2 binary attributes (altogether $5 + 5 \times 2 = 15$ attributes). The values of each pair of new attributes were determined by the value of the original attribute using parity relation of second order (EXOR relation). For example, if original attribute has value $V1$ then two new attributes have equal values, otherwise different values. Therefore, the *intended* information gain of two new attributes together is equal to the information gain of the original attribute. Note that information gain calculated with (1) is zero for all new attributes while the intended information gain for the new attribute is half

of the information gain of the original attribute.

The third data set was obtained from the first data set by replacing each informative attribute with 3 binary attributes (altogether 20 attributes), the values of which were determined by the value of the original attribute using parity relation of third order. Therefore, each new attribute in this data set has one third of the information gain of the original attribute. Each data set was also corrupted with 0%, 10% and 20 % class noise. X% of noise means that class was changed for X% of randomly selected instances.

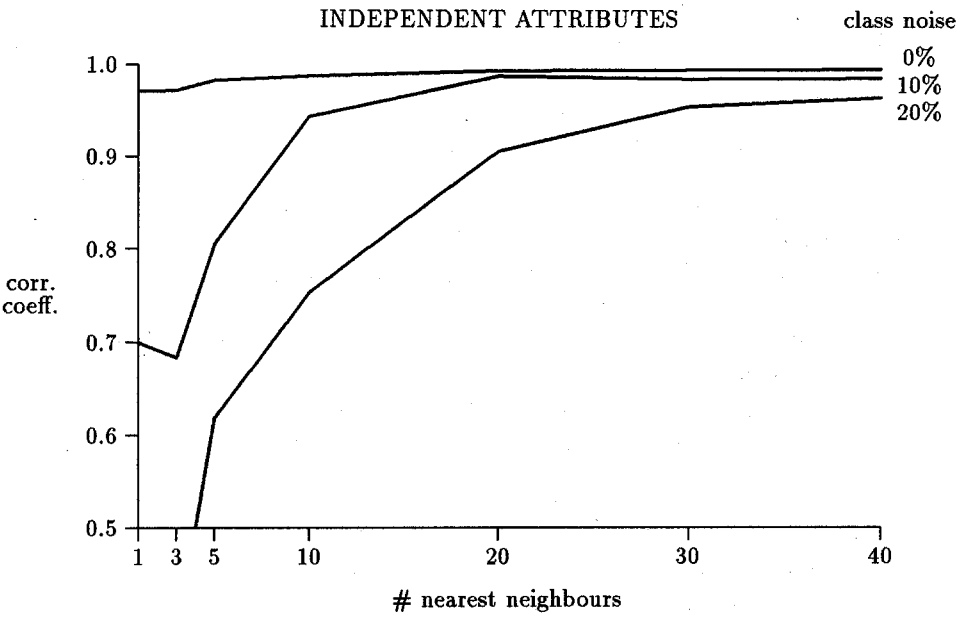

Figure 1: Results of experiments with RELIEF-A using the data set with independent and informative attributes.

The results for each data set are presented on figures 1-3. The results clearly show that the higher number of nearest neighbours in RELIEF-A improves the estimates of attributes even for noise free data sets. However, the improvement is more drastic for data sets with noise.

For independent attributes (figure 1) the quality of estimate monothonously increases with the number of nearest neighbours. This can be formally explained with the following derivation. When the number of nearest neighbours increases, equation (2) becomes:

$$W[A] = P(\text{different value of A}|\text{different class}) - P(\text{different value of A}|\text{same class}) \quad (3)$$

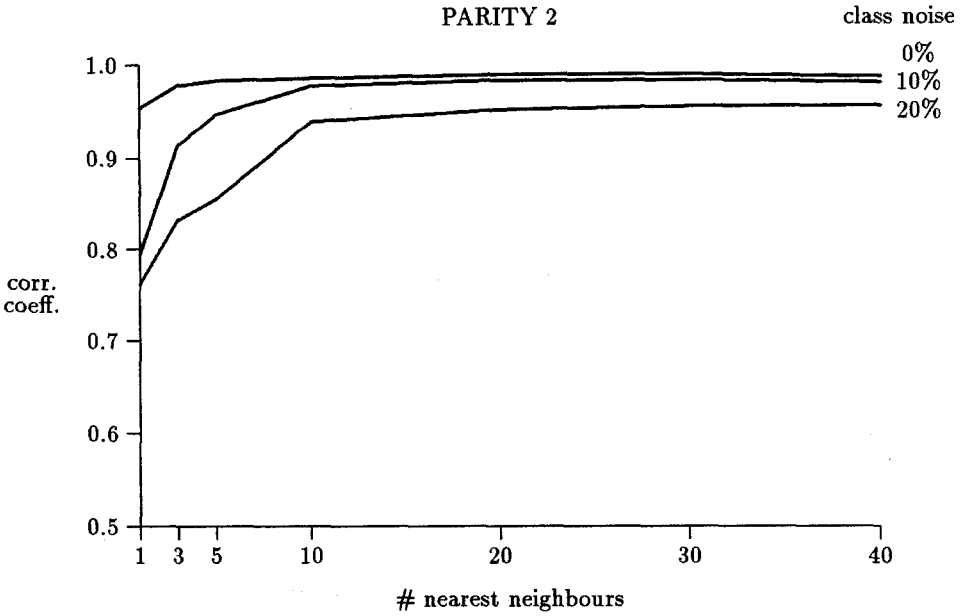

Figure 2: Results of experiments with RELIEF-A using the data set with pairwise informative attributes.

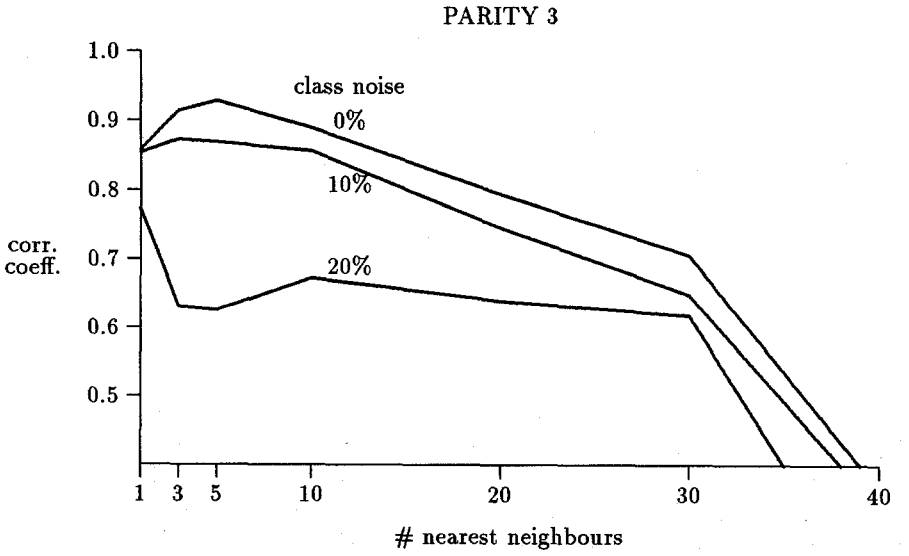

Figure 3: Results of experiments with RELIEF-A using the data set with triplets of informative attributes.

$$= P(\text{equal value of A}|\text{same class})$$

$$-P(\text{equal value of A}|\text{different class})$$

If we rewrite

$$P_{equal} = P(\text{equal value of A})$$

$$P_{samecl} = P(\text{same class})$$

and

$$P_{samecl|equal} = P(\text{same class}|\text{equal value of A})$$

we obtain using Bayes rule:

$$W[A] = \frac{P_{samecl|equal}P_{equal}}{P_{samecl}} - \frac{(1 - P_{samecl|equal})P_{equal}}{1 - P_{samecl}}$$

Using equalities

$$P_{samecl} = \sum_C P(C)^2$$

$$P_{samecl|equal} = \sum_V \left(\frac{P(V)^2}{\sum_V P(V)^2} \times \sum_C P(C|V)^2 \right)$$

we obtain:

$$W[A] = \frac{P_{equal} \times Gini'(A)}{P_{samecl}(1 - P_{samecl})} \tag{4}$$

where

$$Gini'(A) = \sum_V \left(\frac{P(V)^2}{\sum_V P(V)^2} \times \sum_C P(C|V)^2 \right) - \sum_C P(C)^2 \tag{5}$$

is highly correlated with gini-index (Breiman et al., 1984) for classes C and values V of attribute A. The only difference is that instead of factor

$$\frac{P(V)^2}{\sum_V P(V)^2}$$

gini-index uses

$$\frac{P(V)}{\sum_V P(V)} = P(V)$$

Gini index is one of impurity functions that is in turn highly correlated with information gain as defined with (1). The denominator of equation (4) is constant for all attributes and therefore does not influence the correlation factor. Factor P_{equal} is a kind of normalization factor for multi valued attributes. In our experiments all attributes were binary with equal prior probabilities for both values, which gives constant $P_{equal} = 0.5$. Therefore, with increasing number of nearest neighbours, the estimates of RELIEF-A are highly correlated with gini-index and information gain.

For dependent attributes the quality increases up to a maximum but later decreases as the number of nearest neighbours exceeds the number of instances that belong to the same peak in the distribution space for a given class. This effect can be seen on figure 3 while for PARITY-2 problem we observed this effect for larger number of nearest neighbours.

It is interesting that, for smaller number of nearest hits/misses used by the algorithm, noise more drastically affects results on independent data sets than results on data sets with dependent attributes. This may be explained by the fact that an incorrect class label implies incorrect attribute values for only one half/third of attributes for parity-2/parity-3 problems. Again, using higher number of nearest neighbours helps: it drastically reduces this effect.

3 Incomplete Data Sets

To enable RELIEF-A to deal with incomplete data sets, the function

$$diff(Attribute, Instance1, Instance2)$$

should be extended to missing values of attributes. We compared 3 versions of RELIEF:

RELIEF-B: If at least one of two instances has unknown value for a given attribute, the diff is set to $1 - \frac{1}{\#\_values\_of\_attribute}$.

RELIEF-C: Same as RELIEF-B except that during updating the estimates $W[A]$ the contributions of such differences (calculated from instances with at least one unknown value for the given attribute) are ignored, with appropriate normalization. The idea is that unknown values should be ignored from the estimates and if enough training instances is provided, the resulting estimates should converge to correct estimates.

RELIEF-D: Calculate the probability that two given instances have different values for the given attribute:

- if one instance (e.g. I1) has unknown value:

$$diff(A, I1, I2) = 1 - P(value(A, I2)|class(I1))$$

- if both instances have unknown value:

$$diff(A, I1, I2) = 1 - \sum_{V}^{\#values(A)} \left(P(V|class(I1)) \times P(V|class(I2))\right)$$

The conditional probabilities are approximated with relative frequencies from the training set.

To estimate the performance of three algorithms 0%, 10%, 20 % and 30% of values of informative attributes were replaced with unknown values. It turned out that for noise free data, there was no significant difference between the performance of above three algorithms. However, for incomplete and noisy data,

RELIEF-D performed significantly better. A typical picture is shown on figure 4 for the data set with independent informative attributes. The number of nearest hits/misses was set to 10 and there were 30% of unknown values. For other data sets and other values of parameters the pictures of results are similar.

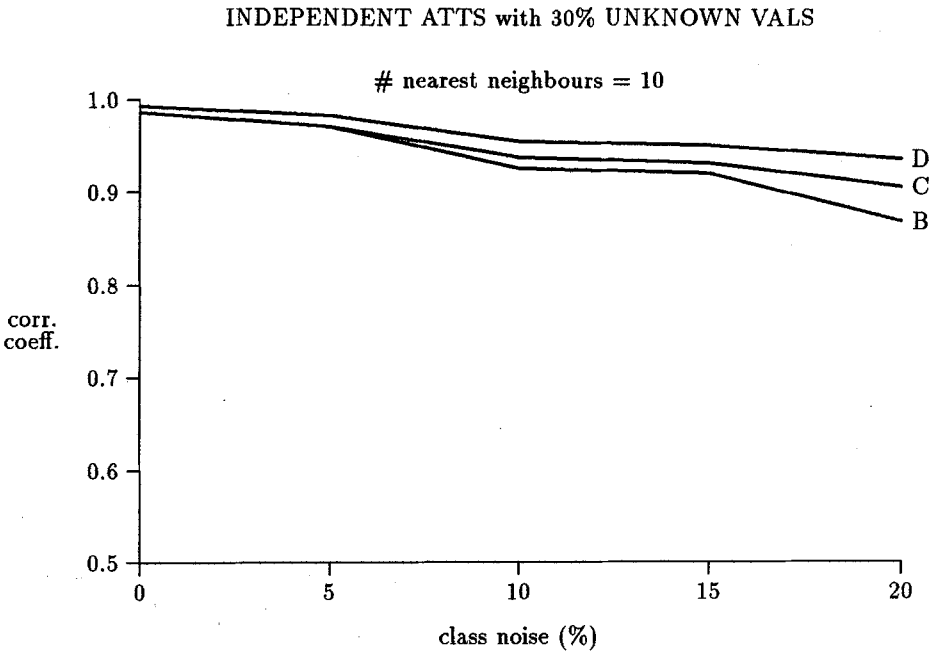

Figure 4: Results of different versions of RELIEF on incomplete data set (30% of unknown attribute values) with independent attributes. The number of nearest hits/misses is 10.

4 Multi-Class Problems

Kira and Rendell (1992a,b) claim that RELIEF can be used to estimate attributes of data sets with more than two classes by splitting the problem into a series of 2-class problems. This solution seems unsatisfactory. In order to use RELIEF in practice it should be able to deal with multi class problems without any prior changes in knowledge representation that could affect the final outcomes.

We experimented with two extensions of RELIEF-D for multi-class problems:

RELIEF-E: Near miss of the given instance I is defined as the nearest neighbour from different class. This is straightforward generalization of RELIEF.

RELIEF-F: Instead of finding one near miss M from different class, the algorithm finds one near miss $M(C)$ for each different class and averages their contribution for updating estimates $W[A]$. The average is weighted with the prior probability of each class:

$$W[A] := W[A] - diff(A, R, H)/m + \sum_{C \neq class(R)} [P(C) \times diff(A, R, M(C))]/m$$

The idea is that the algorithm should estimate the ability of attributes to separate each pair of classes regardless of which two classes are closest to each other.

In order to compare the performance of above two algorithms we generated four additional data sets. First two data sets have 3 and 4 equally probable classes, respectively, 3 random attributes, and 3 informative attributes for each pair of classes. The data set with 3 classes has $3 + 3 \times \frac{3 \times 2}{2} = 12$ binary attributes, and the one with 4 classes have $3 + 3 \times \frac{4 \times 3}{2} = 21$ attributes. The attributes were made informative by means of the prior probability of one of the attribute's values given the class. E.g. for attribute that is informative for separating class 1 and 3 we have $P(V1|C1) = 0.95, 0.75, 0.55$, $P(V1|C3) = 1 - P(V1|C1)$ and $P(V1|C2) = P(V1|C4) = 0.5$.

The other two data sets were obtained from the first two by replacing each informative attribute with 2 binary attributes in the same way as in the second data set described in section 2. Therefore, the third data set has 21 attributes and the last one has 39 attributes.

Results are given in figures 5 and 6. The results show clear advantage of RELIEF-F both in noise free and noisy data.

5 Discussion

RELIEF is efficient heuristic estimator of attributes that is able to deal with data sets with dependent and independent attributes. Its extensions incorporated in RELIEF-F enable it to deal with noisy and incomplete data sets and, what is probably the most important contribution of RELIEF-F, it can efficiently deal with multi class problems.

To verify this conclusions, drawn from experiments with artificial data sets, we ran different versions of RELIEF on one well known medical data set. However, for real world data sets the intended (true) information gain of attributes is unknown. For "primary tumor" data set physicians claim, that attributes are independent, and this was also confirmed with the experiments with semi-naive Bayesian classifier (Kononenko, 1991). Therefore, for this data set, it is acceptable to use information gain calculated with (1) as an estimate of the target attribute quality.

To estimate the performance of different versions of RELIEF, we calculated the correlation coefficient between factual information gain of attributes and RELIEF's estimates $W[A]$. Results are given in figure 7. Results on the real

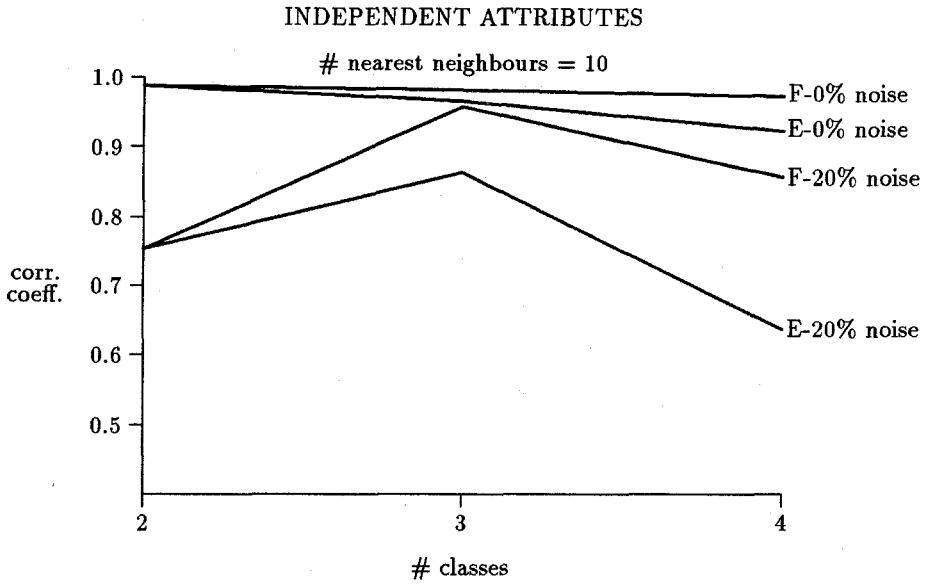

Figure 5: Results of different versions of RELIEF in multi-class problems with independent attributes without noise and with 20% class noise. The number of nearest hits/misses is 10.

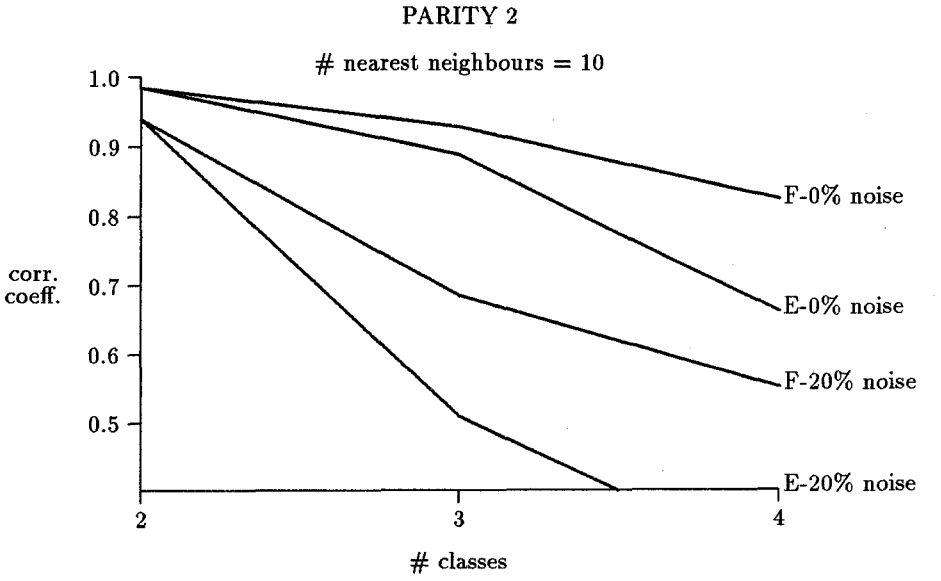

Figure 6: Results of different versions of RELIEF in multi-class problems with dependent attributes without noise and with 20% class noise. The number of nearest hits/misses is 10.

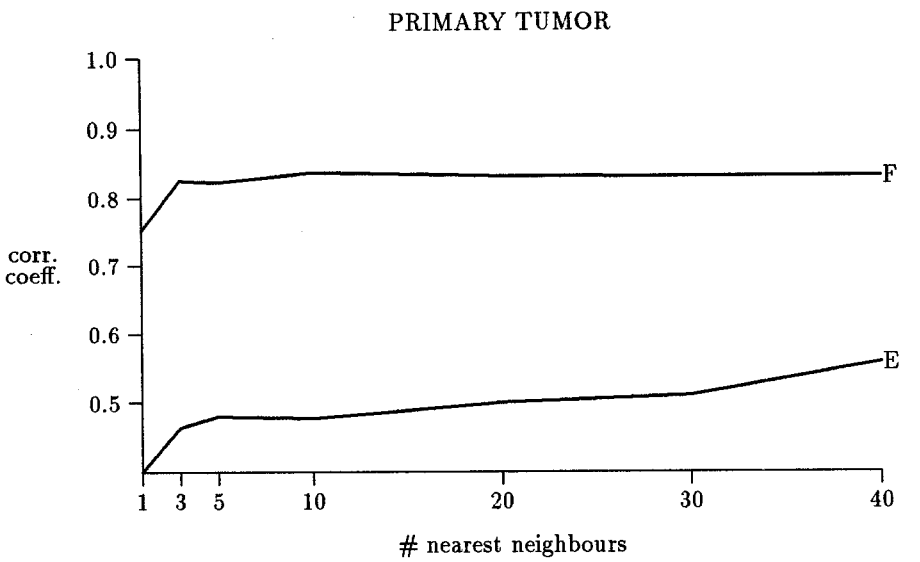

Figure 7: Results of different versions of RELIEF on "primary tumor" medical data set. Versions of RELIEF-E with the same algorithm for unknown values as RELIEF-B and RELIEF-C have all their correlation coefficients less than 0.4.

world data set confirm the advantage of RELIEF-F over other versions and also support the conclusions drawn from experiments with artificial data sets.

In this paper we did not address the problem of multi valued attributes. Information gain (1) and gini-index tend to overestimate multi valued attributes and various normalization heuristics are needed to avoid this tendency (e.g. gain ratio (Quinlan, 1986) and binarization of attributes (Kononenko et al., 1984)). RELIEF with k-nearest hits/misses implicitly uses prior probability that two instances have equal values (P_{eqval}) (see equation (4)) for such normalization, which seems to be appropriate. The major difference between information gain and estimates by RELIEF-F in "primary tumor" problem is in estimates of two most significant attributes. Information gain overestimates one attribute with 3 values (by the opinion of physicians specialists). RELIEF-F and normalized versions of information gain correctly estimate this attribute as less important.

Inductive learning algorithms typically use variants of greedy search strategy to overcome the combinatorial explosion during the search for good hypotheses. The major role in the greedy search has a heuristic function that estimates the potential successors of the current state in the search space. RELIEF-F seems to be very promising heuristic function that may overcome the myopy of current inductive learning algorithms. Kira and Rendell used RELIEF as a preprocessor to eliminate irrelevant attributes from the data description before learning. RELIEF-F is general, efficient and reliable enough that can be used inside the learning process to guide the search.

Acknowledgements

Part of this work was done during the author's stay at California Institute of Technology in Pasadena, CA. I would like to thank Padhraic Smyth and Prof. Rodney Goodman for enabling my work in their Micro Systems Research Laboratory. This work was supported by Slovenian Ministry of Science. I thank Matevž Kovačič and Uroš Pompe for comments on earlier versions of the paper. One of the reviewers was slightly imprecise when suggesting me to read strongly related papers in journals IEEE Trans. on PAMI, Pattern Recognition and Pattern Recognition Letters in years 1978-1990. I would appreciate if any reader can make this information more precise.

References

1. Breiman L., Friedman J.H., Olshen R.A., Stone C.J.: Classification and Regression Trees. Wadsforth International Group 1984
2. Hunt E., Martin J & Stone P.: Experiments in Induction. New York: Academic Press 1966
3. Kira K. & Rendell L.: A practical approach to feature selection. In: Proc. Intern. Conf. on Machine Learning. (Aberdeen, July 1992) D.Sleeman & P.Edwards (eds.), Morgan Kaufmann 1992, pp.249–256
4. Kira K. & Rendell L.: The feature selection problem: traditional methods and new algorithm. In: Proc. AAAI'92. San Jose, CA, July 1992
5. Kononenko I., Bratko I., Roškar E.: Experiments in inductive learning of medical diagnostic rules. In: Proc. International School for the Synthesis of Expert Knowledge Workshop. Bled, Slovenia, August 1984
6. Kononenko I.: Semi-naive Bayesian classifier. In: Proc. European Working Session on Learning, (Porto, March 1991), Y.Kodratoff (ed.), Springer Verlag 1991, pp.206–219
7. Mantaras R.L.: ID3 Revisited: A distance based criterion for attribute selection. In: Proc. Int. Symp. Methodologies for Intelligent Systems. Charlotte, North Carolina, U.S.A., Oct. 1989
8. Quinlan R.: Induction of decision trees. Machine learning 1, 81–106 (1986)
9. Smyth P. & Goodman R.M.: Rule induction using information theory. In. G.Piatetsky-Shapiro & W.Frawley (eds.): Knowledge Discovery in Databases. MIT Press 1990