Comprehensible Interpretation of Relief's Estimates

Marko Robnik-Šikonja MARKO.ROBNIK@FRI.UNI-LJ.SI
Igor Kononenko IGOR.KONONENKO@FRI.UNI-LJ.SI
University of Ljubljana, Faculty of Computer and Information Science, Tržaška 25, 1001 Ljubljana, Slovenia

Abstract

Attribute estimation is an important machine learning problem. It is contained in many tasks, e.g., feature subset selection, constructive induction, decision and regression tree building. Relief algorithms are one of the most successful heuristic measures for solving this problem. The quality estimates of the Relief algorithms have commonly been interpreted as the difference of two probabilities, which make them rather difficult for human comprehension. We present a new insight on how these algorithms work by analyzing their behavior with abundance of training data. We show that Relief's weight of an attribute converges to the ratio between the number of explained changes in the concept and the number of examined instances. We show how this new interpretation of quality estimates can be used to explain some behaviors of Relief algorithms and give examples of better human comprehension in two real world problems.

1. Introduction

The problem of estimating the quality of attributes (features) is an important issue in machine learning. There are several important tasks in the process of machine learning e.g., feature subset selection, constructive induction, decision and regression tree building, which contain an attribute estimation procedure as their (crucial) ingredient.

In the literature there are several measures for estimating the quality of the attributes. If a target concept is a discrete variable (classification problem) these are e.g., information gain (Hunt et al., 1966), Gini index (Breiman et al., 1984), Gain ratio (Quinlan, 1993), ReliefF (Kononenko, 1994), and also χ^2 and G statistics are used. If the target concept is presented as a real valued function (continuous class and regression problem) then the estimation heuristics are e.g., the mean squared and the mean absolute error (Breiman et al., 1984), and RReliefF (Robnik Šikonja & Kononenko, 1997).

Most of the heuristic measures for estimating the quality of the attributes assume the independence of the attributes and are therefore less appropriate in problems which possibly involve much feature interaction. Relief algorithms (Relief, ReliefF and RReliefF) do not make this assumption. They are efficient, aware of contextual information, and can correctly estimate the quality of the attributes in problems with strong dependencies between the attributes.

While Relief algorithms have commonly been viewed as a feature subset selection methods that are applied in a prepossessing step before the model is learned (Kira & Rendell, 1992) and are one of the most successful preprocessing algorithms to date (Dietterich, 1997), they are actually general feature estimators and have been used successfully in a variety of settings: to select splits in the building phase of decision tree learning (Kononenko et al., 1997), to select splits and guide the constructive induction in learning of the regression trees (Robnik Šikonja & Kononenko, 1997), as attribute weighting method (Wettschereck et al., 1997) and also in inductive logic programming (Pompe & Kononenko, 1995).

Many machine learning tasks do not care about comprehensiveness of the learned. However, when the learned model is used to analyze and understand some underlying processes, a comprehensible interpretation of the results is crucial. For example, if a human expert is involved she may want to know why the feature subset selection algorithm has selected certain attributes and discarded the others, what is the criteria for ranking the attributes, what is the meaning of the quality estimations, etc. With decision and regression trees there are many more such questions which demand a human-friendly interpretation of the quality estimates in order to be answered satisfactorily.

The quality estimates of the Relief algorithms have commonly been interpreted as the difference of two probabilities: the probability that two instances have different value of the attribute if they have different prediction value and the probability that two instances have different value of the attribute if they have similar prediction values. These two probabilities contain an additional condition that the

instances are close to each other in the problem space and form the estimate of how well the values of the attribute distinguish between the instances that are near to each other.

In this paper we present another interpretation of the Relief's estimates which is based on a theoretical analysis. We analyze the behavior of Relief when the number of the examples approaches infinity i.e., when the problem space is densely covered with the examples. Under this conditions we prove that the quality estimate of the attribute can be interpreted as the ability of the attribute to explain the changes in the predicted value. We present advantages of this interpretation and show some behaviors of the Relief algorithms that can be explained with a help of this property.

We assume that examples $I_1, I_2, ..., I_n$ in the instance space are described by a vector of attributes $A_1, A_2, ..., A_a \in \mathcal{A}$, where a is the number of explanatory attributes, and are labelled with the target value τ_i. The examples are therefore points in the a dimensional space. If the target value is categorical we call the modelling task classification and if it is numerical we call the modelling task regression. In the classification we assume that there are c classes with the prior probabilities $p(c_i)$.

In the next Section we briefly present the Relief algorithms and explain why they work. Section 3 gives an overview of the existing probabilistic interpretation and in Section 4 we analyze a behavior of Relief algorithms when the number of the examples is large. We derive a theoretical property of the Relief's quality estimates and offer an interpretation for it which is comprehensible for human experts. Section 5 contains examples of Relief's behavior which can be explained with the help of this property. Final section summarizes the contributions of the paper.

2. Relief algorithms

In this Section we will describe the Relief algorithms and discuss their similarities and differences. Because of the easier understanding of the main idea we will first present the original Relief algorithm (Kira & Rendell, 1992) and give an account on how and why it works. Afterwards we will discuss its extension ReliefF (Kononenko, 1994), which is more robust and can deal with multi class problems and an adaptation to regression problems called RReliefF (Robnik Šikonja & Kononenko, 1997).

A main idea of the original Relief algorithm (Kira & Rendell, 1992), given in Figure 1, is to estimate the quality of the attributes according to how well their values distinguish between the instances that are near to each other. For that purpose, given a randomly selected instance R_i (line 3), Relief searches for its two nearest neighbors: one from the same class, called *nearest hit* H, and the other from the different class, called *nearest miss* M (line 4). It updates the quality estimation $W[A]$ for all the attributes A depending on their values for R_i, M, and H (lines 5 and 6). If the instances R_i and H have different values of an attribute A then the attribute A separates two instances of the same class which is not desirable so we decrease the quality estimation $W[A]$. On the other hand if the instances R_i and M have different values of an attribute A then the attribute A separates two instances with the different class values which is desirable so we increase the quality estimation $W[A]$. The process is repeated m times, where m is a user-defined parameter.

Function $\mathrm{diff}(A, I_1, I_2)$ calculates the difference between the values of the attribute A for two instances I_1 and I_2. For nominal attributes it was originally defined as:

$$\mathrm{diff}(A, I_1, I_2) = \begin{cases} 0 & ; value(A, I_1) = value(A, I_2) \\ 1 & ; otherwise \end{cases} \tag{1}$$

and for numerical attributes as:

$$\mathrm{diff}(A, I_1, I_2) = \frac{|value(A, I_1) - value(A, I_2)|}{max(A) - min(A)} \tag{2}$$

The function diff is used also for calculating a distance between the instances to find the nearest neighbors. The total distance is simply the sum of distances over all attributes (Manhattan distance).

The original Relief can deal with nominal and numerical attributes. However, it cannot deal with incomplete data and is limited to two-class problems. Its extension, ReliefF (Kononenko, 1994) deals with multi-class problems, is more robust and can deal with incomplete and noisy data. The selection of k hits and misses is the basic difference to Relief which ensures greater robustness of the algorithm concerning a noise.

In regression problems the predicted value τ is continuous, therefore the (nearest) hits and misses cannot be used. To solve this difficulty, instead of requiring the exact knowledge of whether two instances belong to the same class or not, a kind of probability that the predicted values of two instances are different is introduced. This probability can be modelled with the relative distance between the predicted values of the two instances. The actual algorithm is called RReliefF (Regressional ReliefF) (Robnik Šikonja & Kononenko, 1997).

We will not go into detail of ReliefF and RReliefF (various parameters, distance handling, metrics) but rather concentrate on possible interpretations of their quality estimates.

Algorithm Relief
Input: for each training instance a vector of attribute values and the class value
Output: the vector W of estimations of the qualities of attributes

1. set all weights $W[A] := 0.0$;
2. **for** i := 1 **to** m **do begin**
3. randomly select an instance R_i;
4. find nearest hit H and nearest miss M;
5. **for** A := 1 **to** a **do**
6. $W[A] := W[A] - \mathrm{diff}(A, R_i, H)/m + \mathrm{diff}(A, R_i, M)/m$;
7. **end**;

Figure 1. Pseudo code of the basic Relief algorithm

3. Probabilistic interpretation

Let us remember how the Relief algorithm updates its quality estimates (see line 6 in Figure 1). The positive updates of the weights are actually forming the estimate of the probability that the attribute discriminates between the instances with the different class, while the negative updates are forming the estimate of the probability that the attribute separates the instances with the same class. The Relief's estimate $W[A]$ of the quality of the attribute A is therefore the approximation of the following difference of probabilities (Kononenko, 1994):

$$W[A] = P(\text{diff. value of A}|\text{nearest inst. from diff. class}) - P(\text{diff. value of A}|\text{nearest inst. from same class}) \quad (3)$$

To be more general we rewrite Equation (3) into the form suitable also for regression:

$$W[A] = P(\text{diff. value of A}|\text{near inst. with diff. prediction}) - P(\text{diff. value of A}|\text{near inst. with same prediction}) \quad (4)$$

This Equation forms a basis for the common interpretation of the quality estimations of Relief algorithms which are explained as the difference of two probabilities: the probability that two instances have different values of the attribute if they have different prediction values and the probability that two instances have different values of the attribute if they have similar prediction values. These two probabilities contain an additional condition that the instances are close to each other and form the estimate of how well the values of the attribute distinguish between the instances that are near to each other.

While this interpretation is well suited to explain certain properties of the Relief algorithms it sometimes fails. Concerning human comprehension of Equation (4) it turns out that it is difficult to understand. The negated similarity (different values) and subtraction of probabilities are difficult to comprehend for experts in real world problems.

We present another interpretation of the Relief's estimates based on the theoretical analysis.

4. Interpretation as the portion of the explained concept changes

We analyze the behavior of Relief when the number of the examples approaches infinity i.e., when the problem space is densely covered with the examples. We present the necessary definitions and prove that the Relief's quality estimates can be interpreted as the ratio between the number of the explained changes in the concept and the number of the examined instances. Then we present advantages of this interpretation and show some behaviors of the Relief algorithms that can be explained with the help of this property. The exact form of this property differs between Relief, ReliefF and RReliefF.

We start with Relief and claim that in a classification problem as the number of the examples goes to infinity the Relief's weights for each attribute converge to the ratio between the number of class label changes the attribute is responsible for and the number of the examined instances. If a certain change can be explained in several different ways, all the ways share the credit for it in the quality estimate. If several attributes are involved in one way of the explanation all of them get the credit in their quality estimate. We formally present the definitions and the property.

Definition 4.1 *Let $B(I)$ be the set of instances from $\mathcal{I}$ nearest to the instance $I \in \mathcal{I}$ which have different prediction value τ than I:*

$$B(I) = \{Y \in \mathcal{I}; \mathrm{diff}(\tau, I, Y) > 0 \wedge Y = \arg\min_{Y \in \mathcal{I}} \delta(I, Y)\} \quad (5)$$

Let $b(I)$ be a single instance from the set $B(I)$ and $p(b(I))$

the probability that it is randomly chosen from $B(I)$. Let $A(I, b(I))$ be a set of attributes with different value between instances I and $b(I)$.

$$A(I, b(I)) = \{A \in \mathcal{A}; \quad b(I) \in B(I) \wedge \operatorname{diff}(A, I, b(I)) > 0\} \tag{6}$$

We say that attributes $A \in A(I, b(I))$ are responsible for the change of the predicted value of the instance I to the predicted value of $b(I)$ as the change of their values is one of the minimal number of changes required for changing the predicted value of I to $b(I)$. If the sets $A(I, b(I))$ are different we say that there are different ways to explain the changes of the predicted value of I to the predicted value $b(I)$. The probability of certain way is equal to the probability that $b(I)$ is selected from $B(I)$.

Let $A(I)$ be a union of sets $A(I, b(I))$:

$$A(I) = \bigcup_{b(I) \in B(I)} A(I, b(I)) \tag{7}$$

We say that the attributes $A \in A(I)$ are responsible for the change of the predicted value of the instance I as the change of their values is the minimal necessary change of the attributes' values of I required to change its predicted value. Let the quantity of this responsibility take into account the change of the predicted value and the change of the attribute:

$$r_A(I, b(I)) = p(b(I)) \cdot \operatorname{diff}(\tau, I, b(I)) \cdot \operatorname{diff}(A, I, b(I)) \tag{8}$$

The ratio between the responsibility of the attribute A for the predicted values of the set of cases $\mathcal{S}$ and the cardinality m of that set is therefore:

$$R_A = \frac{1}{m} \sum_{I \in \mathcal{S}} r_A(I, b(I)) \tag{9}$$

Property 4.1 *Let the concept be described with the attributes $A \in \mathcal{A}$ and n noiseless instances $I \in \mathcal{I}$; let $\mathcal{S} \subseteq \mathcal{I}$ be a set of randomly selected instances used by Relief (line 3 on Figure 1) and let m be the cardinality of that set. If Relief randomly selects the nearest instances from all possible nearest instances then for its quality estimate $W[A]$ the following property holds:*

$$\lim_{n \to \infty} W[A] = R_A \tag{10}$$

The quality estimate of the attribute can therefore be explained as the ratio of the predicted value changes the attribute is responsible for to the number of the examined instances.

Proof. Equation (10) can be explained if we look into the spatial representation. There are a number of different characteristic regions of the problem space which we usually call peaks. The Relief algorithms selects an instance R from $\mathcal{S}$ and compares the value of the attribute and the predicted value of its nearest instances selected from the set $\mathcal{I}$ (line 6 on Figure 1), and than updates the quality estimates according to these values. For Relief this mean: $W[A] := W[A] + \operatorname{diff}(A, R, M)/m - \operatorname{diff}(A, R, H)/m$, where M is the nearest instance from the different class and H is the nearest instance from the same class.

When the number of the examples is sufficient ($n \to \infty$), H must be from the same characteristic region as R and its values of the attributes converge to the values of the instance R. The contribution of term $-\operatorname{diff}(A, R, H)$ to the $W[A]$ in the limit is therefore 0.

Only the terms $\operatorname{diff}(A, R, M)$ contribute to $W[A]$. The instance M is randomly selected nearest instance with different prediction than R, therefore in the noiseless problems there must be at least some difference in the values of the attributes and M is therefore an instance of $b(R)$ selected with probability $p(M)$. As M has different prediction value than R the value $\operatorname{diff}(\tau, R, M) = 1$. The attributes with different values at R and M constitute the set $A(R, M)$. The contribution of M to $W[A]$ for the attributes from $A(R, M)$ equals $\operatorname{diff}(A, R, M)/m = \operatorname{diff}(\tau, R, M)) \cdot \operatorname{diff}(A, R, M))/m$ with probability $p(M))$.

Relief selects m instances $I \in \mathcal{S}$ and for each I randomly selects its nearest miss $b(I)$ with probability $p(b(I))$. The sum of updates of $W[A]$ for each attribute is therefore: $\sum_{I \in \mathcal{S}} p(b(I)) \operatorname{diff}(\tau, R, b(I)) \operatorname{diff}(A, R, b(I))/m) = R_A$, which concludes the proof. ■

Let us show an example, which illustrates the idea. We have a Boolean domain where the class value is defined as $\tau = (A_1 \wedge A_2) \vee (A_1 \wedge A_3)$. Table 1 gives the tabular description of the problem. The right most column shows which of the attributes is responsible for the change of the predicted value.

Table 1. Tabular description and responsibility of the attributes for the change of the predicted value.

line	A_1	A_2	A_3	τ	responsible attributes
1	1	1	1	1	A_1
2	1	1	0	1	A_1 or A_2
3	1	0	1	1	A_1 or A_3
4	1	0	0	0	A_2 or A_3
5	0	1	1	0	A_1
6	0	1	0	0	A_1
7	0	0	1	0	A_1
8	0	0	0	0	(A_1, A_2) or (A_1, A_3)

In line 1 we say that A_1 is responsible for the class assignment because changing its value to 0 would change τ to 0,

while changing only one of A_2 or A_3 would leave τ unchanged. In line 2 changing any of A_1 or A_2 would change τ too, so A_1 and A_2 represent two manners how to change τ and also share the responsibility. Similarly we explain lines 3 to 7, while in line 8 changing only one attribute is not enough for changing τ. However, changing A_1 and A_2 or A_1 and A_3 changes τ. Therefore the minimal number of the required changes is 2 and the credit (and the updates in the algorithm) goes to both A_1 and A_2 or A_1 and A_3. There are 8 peaks in this problem which are equiprobable so A_1 gets the estimate $\frac{4+2\cdot\frac{1}{2}+2\cdot\frac{1}{2}}{8} = \frac{3}{4} = 0.75$ (it is alone responsible for lines 1, 5, 6, and 7, shares the credit for lines 2 and 3 and cooperates in both credits for line 8). A_2 (and similarly A_3) gets estimate $\frac{2\cdot\frac{1}{2}+\frac{1}{2}}{8} = \frac{3}{16} = 0.1875$ (it shares the responsibility for lines 2 and 4 and cooperates in one half of line 8).

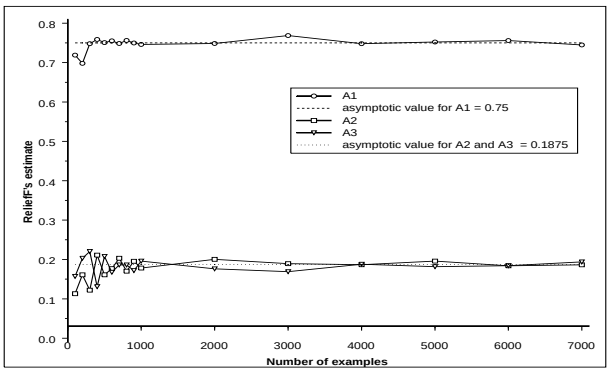

Figure 2. The estimates of the attributes by Relief and ReliefF are converging to the ratio between class labels they explain and the number of examined instances.

Figure 2 shows the estimates of the quality of the attributes for this problem for Relief (and also ReliefF). As we wanted to scatter the concept we added besides three important attributes also five random binary attributes to the problem description. We can observe that as we increase the number of the examples the estimate for A_1 is converging to 0.75, while the estimates for A_2 and A_3 are converging to 0.1875 as we expected. The reason for rather slow convergence is in the random sampling of the examples, so we need much more examples than the complete description of the problem (256 examples).

For ReliefF this property is somehow different. Due to the lack of space we omit the exact definitions and proof and give only a summary and results here. Their exact formulations can be found in (Robnik-Šikonja & Kononenko, 2001). ReliefF searches for k nearest instances for each of the misses and weights their contribution to $W[A]$ with the prior probabilities of the class values. If $R_A(i,j)$ represents the sum of the responsibilities of the attribute A for the change of the class value of m instances from the class i to the class j then ReliefF has the following property:

Property 4.2 *Let $p(c_i)$ represent the prior probability of the class c_i. Under the conditions of Property 4.1, algorithm ReliefF behaves as:*

$$\lim_{n\to\infty} W[A] = \sum_{i=1}^{c}\sum_{\substack{j=1\\ j\neq i}}^{c} \frac{p(c_i)p(c_j)}{1-p(c_i)} R_A(i,j) \qquad (11)$$

We can therefore explain the quality estimates as the ratio of class values changes the attribute is responsible for to the number of the examined instances weighted with the prior probabilities of class values.

Corollary 4.3 *In two class problems where* diff *function is symmetric:* $\mathrm{diff}(A, I_1, I_2) = \mathrm{diff}(A, I_2, I_1)$ *Property 4.2 is equal to Property 4.1.*

In a Boolean noiseless case as in the example above the fastest convergence would be with only 1 nearest neighbor. With more nearest neighbors (default with the algorithms ReliefF and RReliefF) we need more examples to see this effect as all of them has to be from the same/nearest peak.

The interpretation of the quality estimates with the ratio of the explained changes in the concept is true for RReliefF as well, as it also computes Equation (4), however, the updates are proportional to the size of the difference in the prediction value. The exact formulation and proof remain for the further work.

Note that the sum of the expressions (10) and (11) for all attributes is usually greater than 1. In certain peaks there are more than one attribute responsible for the class assignment i.e., the minimal number of attribute changes required for changing the value of the class is greater than 1 (e.g., line 8 in Table 1). The total number of explanations is therefore greater (or equal) than the number of inspected instances. As Relief algorithms normalize the weights with the number of inspected instances m and not the total number of possible explanations, the quality estimations are not proportional to the attributes's responsibility but present rather a portion of the explained changes. For the estimates to represent proportions we would have to change the algorithm and thereby lose probabilistic interpretation of the attributes' weights.

When we omit the assumption of the sufficient number of the examples then the estimates of the attributes can be greater than their asymptotic values because the instances more distant than the minimal number of required changes might be selected into the set of nearest instances and the attributes might be positively updated also when they are responsible for the changes which are more distant than the minimal number of required changes.

The behavior of the Relief algorithm in the limit (Equation (4.1)) is the same as the the asymptotic behavior of the algorithm Contextual Merit (CM) (Hong, 1997) which uses only the contribution of nearest misses. In multi class problems CM searches for nearest instances from different classes disregarding the actual class they belong, while ReliefF selects equal number of instances from all different classes and normalizes their contribution with their prior probabilities. The idea is that the algorithm should estimate the ability of attributes to separate each pair of the classes regardless of which two classes are closest to each other. It was shown (Kononenko, 1994) that this approach is superior and the same normalization factors occur also in asymptotic behavior of ReliefF given by Equation (4.2).

Based on the asymptotic properties one could come to, in our opinion, the wrong conclusion that it is sufficient for estimation algorithm to consider only nearest instances with different class. While the nearest instances with the same prediction have no effect when the number of the instances is unlimited they nevertheless play an important role in problems of practical sizes. E.g., in parity problems with three important attributes CM separates the important from unimportant attributes for a factor of 2 to 5 and only 1.05-1.19 with numerical attributes while with the Relief algorithms under the same conditions this factor is over 100 due to the additional information it extracts from the nearest instances with the same prediction.

Clearly the interpretation of Relief's weights as the ratio of explained concept changes is more comprehensible than the interpretation with the difference of two probabilities. The responsibility for the explained changes of the predicted value is intuitively clear. Equation 10 is non probabilistic, unconditional, and contains a simple ratio, which can be understood taking the unlimited number of the examples into account. The actual quality estimates for the attributes in given problem are therefore approximations of these ideal estimates which occur only with abundance of data.

5. Understandability of new interpretation

While we are aware that it is not possible to measure understandability in a clear and unambiguous way we still think that it is possible to compare it. First, the concept of the responsibility for the explained changes of the predicted value is intuitively clear while it is much harder to form mental picture what a difference of two conditional probabilities might represent. Second, Equation (10) (compared to Equation (4)) is non probabilistic, unconditional, and contains a simple ratio, which can be understood taking the unlimited number of examples into account. The third criterion we argue for is the human perception, which certainly is subjective and vague but perhaps the best we can hope for with the lack of exact definition. We present two experiences with human experts on real world problems where an error minimization was not as much of a concern as the comprehensiveness of the learned models and their explanation power.

5.1 Wound healing problem

The first problem we are going to describe is medical (Cukjati et al., 2001), namely we evaluated various prognostic factors which affect a wound healing rate and tried to predict it.

We started with a database of 266 patients with 390 wounds which were recorded in more than a decade lasting clinical study of the use of electrical stimulation to accelerate chronic wound healing at Institute of the Republic of Slovenia for Rehabilitation in Ljubljana. The controlled study involved the conventional conservative treatment, sham treatment, biphasic pulsed current, and direct current electrical stimulation. Each patient and wound were registered and a wound healing process was weekly followed. Many patient and wound data were missing and not all wounds were followed regularly or until the complete wound closure which is relatively common problem of clinical trials. We selected 214 patients with 300 wound cases who obeyed the inclusion criteria: initial wound area larger than $1cm^2$ and at least four weeks of the observation or until the complete wound closure. Among these 300 wound cases, the observation periods in 174 cases lasted until the complete wound closure and was shorter in the rest 126 cases. In these cases the time to complete wound closure was estimated from the wound area measurements obtained during the observation period. The measure of the wound healing rate was defined as an average advance of the wound margin towards the wound center and it was calculated as the average wound radius (the initial wound area divided by the initial perimeter and multiplied by 2) divided by the time to complete wound closure.

The problem was described with three types of attributes: wound attributes, patient attributes and treatment attributes. We selected the following wound attributes: wound area, wound grade, wound type, location, elapsed time from spinal cord injury to wound appearance, elapsed time from wound appearance to the beginning of treatment, and wound aetiology. The recorded patient attributes were age, sex, total number of wounds, diagnosis, and, in case of spinal cord injured patient, degree of spasticity. The wounds were randomly assigned into four treatment groups: conservative treatment, sham treatment, direct and biphasic current stimulation of wound healing. Besides the type of treatment we also took into account daily duration of the electrical stimulation.

We employed machine learning algorithms to estimate the quality of the attributes and to build tree based models (regression and classification trees) to predict the wound healing rate based on the initial wound, patient and treatment data. We also considered the estimated wound healing rate based on a mathematical model and built trees for prediction of the wound healing rate after one, two, three, four, five and six weeks of follow-up. We tested several algorithms for attribute estimation which were used in feature subset selection and for selection of splits in interior nodes of tree based models. In the earlier stages of our data exploration we found that RReliefF for the regression and ReliefF for classification problems, were the most effective, concerning the error and comprehensiveness of the learned models. In the regression we considered also the mean squared error and mean absolute error as the attribute estimators (Breiman et al., 1984) and in the classification Gain ratio (Quinlan, 1993) was tested.

The attributes' quality estimates calculated using RReliefF revealed that the initial wound area, followed by the patient's age and the time from wound appearance to treatment beginning are the most prognostic attributes. Important prognostic attributes are also wound shape, location and type of treatment. The human experts agreed with this selection and the ranking of the attributes findings but also wanted to understand the quantities of the Relief's weights. We first tried to explain the quality estimates with the difference of two probabilities (Equation (4)) but this did not help them to build a mental picture. Shortly afterwards we came out with the ratio of the explained concept changes interpretation and the experts liked this interpretation. It helped them to compare also the sizes of the quality weights. As they understood the meaning of the weights they were also much more confident about the structure and performance of the regression and decision trees we build in the second stage of our study.

5.2 Modelling of the apparent photosynthesis

Our second experience is with an ecological problem. We wanted to model the apparent photosynthesis of *Stipa bromoides* grass (Dalaka et al., 2000). The aim of our study was to compare machine learning tools with ordinary statistical models which were constructed based on background knowledge concerning the response of photosynthesis to irradiance. The data set consisted from 777 measurements of the apparent photosynthesis of the leaves of Stipa bromoides, air and leaf temperature, relative air humidity, measuring site, leaf category, and photosynthetically active radiation. We used 10-fold cross validation to split the data into the training and testing set and built the regression tree on each training set. As feature estimators in the interior nodes of the tree we used RReliefF or the mean squared error. In the leaves of the trees we modelled the response variable with linear equations, mean, or median value of the cases in the leaf. We obtained a minimal error with the linear equations in the leaves and using RReliefF as the partitioning criterion. The field expert liked the structure of the trees produced by RReliefF. They even discovered new knowledge with the help of splitting points determined by the RReliefF as the estimator, namely that there is a threshold value close to 35% for the relative humidity which determines the influence of the irradiance and temperature on the apparent photosynthesis.

We explained the quality estimates of RReliefF as the difference of two probabilities, however this interpretation was not acquired into the mental representation of the problem the expert had. They understood what mean square error does as an attribute estimator but not the meaning of the RReliefF's weights. As we could not offer any better explanation at the time they regarded the weights as useful but without the proper meaning.

When we later presented the ratio of the explained concept changes interpretation of the RReliefF's quality estimates to one of the involved field experts he was very interested in it and considered the RReliefF's ranking of the attributes as the true assessment of their practical importance for determining the apparent photosynthesis.

6. Explained behavior

We describe a behavior of Relief algorithms which can be explained with the help of Equation (10). Let us have a family of regression problems described with numerical attributes with values on interval $[0, 1]$ and 1000 examples. Besides I important attributes there are also 10 random attributes in each problem. They are defined as

$$\mathrm{LinEq-I}: \qquad \tau = \sum_{j=1}^{I} A_j \tag{12}$$

In this problem all attributes are equally important. We want to check how many important attributes we can have in the problem for RReliefF still to separate them from the unimportant attributes. Figure 3 shows two RReliefF's quality estimate curves: one for the worst estimated important attribute and the other for the best estimated random attribute.

We can see that the difference between the quality estimates of the important and unimportant attributes is decreasing and eventually one of the unimportant attributes become estimated as better than the worst one of the important attributes. In our case this happens with 10 important attributes. This is not surprising considering the properties of Relief: each attribute gets a weight according to the number of the explained function values so increasing the number of the important attributes make the weights for each of

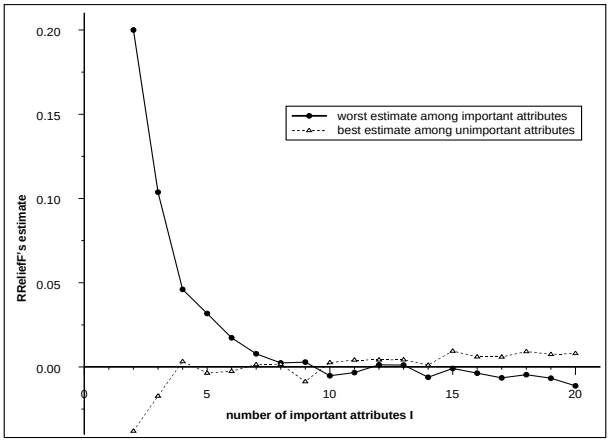

Figure 3. RReliefF's quality estimates in $\mathrm{LinEq}-\mathrm{I}$ problems with 1000 examples.

them decreasing and approaching zero. If we increase the number of examples to 4000, we can reliably estimate the attributes up to 16 important attributes.

If the important attributes are not equally important then the Relief algorithms ideally would share the credit among them according to the number of concept changes explanations. In practice less important attributes get less than this. The reason for this is that the differences in the predicted value caused by less important attribute (see Eq. (4)) are overshadowed by larger changes caused by more important attributes. The details can be found in (Robnik-Šikonja & Kononenko, 2001).

Therefore, in practice the Relief algorithms tend to underestimate less important attributes, while they successfully recognize more important attributes even in difficult conditions.

7. Conclusions

We have presented the analysis of the Relief algorithms which shows that when the number of the examples is unlimited, the quality estimates of the attributes represent the ratio between the number of changes in the predicted value which these attributes explain and the number of the inspected instances. In other words: the attribute gets the quality estimate according to its ability to explanation the particular concept.

This property helps us to understand the bias of Relief algorithms towards more important attributes and against less important attributes.

We believe that the presented interpretation of the Relief's quality estimates is comprehensible and our experience with domain experts in two real world problems confirms it.

References

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth

Cukjati, D., Robnik-Šikonja, M., Reberšek, S., Kononenko, I., & Miklavčič, D. (2001). Prognostic factors, prediction of chronic wound healing and electrical stimulation. *Medical & Biological Engineering & Computing*. (submitted).

Dalaka, A., Kompare, B., Robnik-Šikonja, M., & Sgardelis, S. (2000). Modeling the effects of environmental conditions on apparent photosynthesis of Stipa bromoides by machine learning tools. *Ecological Modelling*, *129*, 245–257.

Dietterich, T. G. (1997). Machine learning research: Four current directions. *AI Magazine*, *18*, 97–136.

Hong, S. J. (1997). Use of contextual information for feature ranking and discretization. *IEEE transactions on knowledge and data engineering*, *9*, 718–730.

Hunt, E. B., Martin, J., & Stone, P. J. (1966). *Experiments in induction*. New York: Academic Press.

Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. *Machine Learning: Proceedings of International Conference (ICML'92)* (pp. 249–256).

Kononenko, I. (1994). Estimating attributes: analysis and extensions of Relief. *Machine Learning: ECML-94* (pp. 171–182). Springer Verlag.

Kononenko, I., Šimec, E., & Robnik-Šikonja, M. (1997). Overcoming the myopia of inductive learning algorithms with RELIEFF. *Applied Intelligence*, *7*, 39–55.

Pompe, U., & Kononenko, I. (1995). Linear space induction in first order logic with ReliefF. In G. Della Riccia, R. Kruse and R. Viertl (Eds.), *CISM courses and lectures no. 363*. Springer Verlag.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann.

Robnik Šikonja, M., & Kononenko, I. (1997). An adaptation of Relief for attribute estimation in regression. *Machine Learning: Proceedings of the Fourteenth International Conference (ICML'97)* (pp. 296–304).

Robnik-Šikonja, M., & Kononenko, I. (2001). Theoretical and empirical analysis of ReliefF and RReliefF. (http://ai.fri.uni-lj.si/rmarko/papers/.

Wettschereck, D., Aha, D. W., & Mohri, T. (1997). A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artificial Intelligence Review*, *11*, 273–314.