Inductive and Bayesian Learning in Medical Diagnosis

Igor Kononenko

University of Ljubljana

Faculty of Electrical and Computer Engineering

Trzaska 25, 61001 Ljubljana, Slovenia

Abstract. Although successful in medical diagnostic problems, inductive learning systems were not widely accepted in medical practice. In this paper two different approaches to machine learning in medical applications are compared: the system for inductive learning of decision trees Assistant, and the naive Bayesian classifier. Both methodologies were tested in four medical diagnostic problems: localization of primary tumor, prognostics of recurrence of breast cancer, diagnosis of thyroid diseases, and rheumatology. The accuracy of automatically acquired diagnostic knowledge from stored data records is compared and the interpretation of the knowledge and the explanation ability of the classification process of each system is discussed. Surprisingly, the naive Bayesian classifier is superior to Assistant in classification accuracy and explanation ability, while the interpretation of the acquired knowledge seems to be equally valuable. In addition, two extensions to naive Bayesian classifier are briefly described: dealing with continuous attributes, and discovering the dependencies among attributes.

KEYWORDS: machine learning, induction of decision trees, naive Bayesian classifier, explanation, medical diagnosis

1 Introduction

In recent years a lot of different inductive learning algorithms were developed [33,34] mostly influenced by Quinlan' s ID3 inductive algorithm for generating decision (classification, diagnostic) trees from examples of solved problems [45,46]. Some of these systems are ACLS [43], Assistant [30,2,10], CART [5], C4 [46], AQ [35,36,39], CN2 [12,13,14], LogArt [9] and ITRULE [19,52,53]. The technology of inductive learning is already appropriate for routine use. Several commercial systems are available with advanced user interfaces that enable the development of applications in various fields.

Inductive learning is well known approach to automatic knowledge acquisition for expert systems which overcomes the so called Feigenbaum bottleneck of knowledge acquisition from human experts [37,54]. Instead of time consuming process of extracting the knowledge from experts and professional literature, an inductive learning system can be used to generate the knowledge, usually in the form of decision rules, from the concrete problems already solved by some experts.

It seems that this technology is well suited for medical diagnosis in small specialized diagnostic problems. Data about correct diagnoses are often available in the form of medical archives in specialized hospitals or their departments. All that has to be done is to type the data into the computer in the appropriate form and run the system for inductive learning. The derived set of decision rules can be used to reveal the basic relations and laws in the problem domain in an explicit and transparent form and, of course, can be used for diagnosing new patients.

In several medical domains the inductive learning systems were actually applied, e.g. in oncology [2,4,15,30,55], liver pathology [32], prognosis of the survival in hepatitis [30], urology [2,30,49], diagnosis of thyroid diseases [20,21,48], rheumatology [22,23,31], diagnosing craniostenosis syndrome [1], dermatoglyptic diagnosis [11], cardiology [3,6,12], neuropsychology [40], gynaecology [41], and perinatology [23]. Typically, automatically generated diagnostic rules slightly outperformed the diagnostic accuracy of physicians specialists.

Although the results seem excellent, this technology was not widely accepted in medical practice for several reasons:

- Inflexibility of the knowledge representation. The set of attributes that describe the patients must be fixed. The information that is used by the rules to derive the final diagnosis is limited to strictly defined parameters while subjective, informal and fuzzy opinions (like intuition, impression, etc.) can not be represented in a formal and symbolic way.

- Learning and classification is sensitive to missing data [47] which is often the case in medical data.

- The generated decision rules typically include too few attributes [44]. The explanation of decisions is therefore rather limited and does not sufficiently support typically exact decisions of generated diagnostic rules.

- Non-negligible is also subjective resistance of physicians to new diagnostic technology. However, they should be aware that expert systems for medical diagnosis are in fact new tools for supporting and improving diagnostic process and in no way can replace the physicians.

In this paper two approaches to learning in medical diagnosis are compared: ID3-like inductive learning system Assistant [2,10] and naive Bayesian classifier [8,17,18,24,30] with some extensions that include advantages of both methodologies. Basically, the two approaches are much different, according to knowledge representation as well as the explanation ability. From training examples Assistant generates a decision tree while knowledge generated (and/or used) by the naive Bayesian classifier is simply a table of conditional and prior probabilities. Using a decision tree, a new problem can be solved by following the appropriate path from the root to the leaf of the tree. The path corresponds to the decision rule and itself represents the explanation of the solution of the problem. On the other hand, the naive Bayesian classifier solves the new problem by appropriately combining the probabilities from the table obtained in the learning phase. The explanation of the solution is the sum of information gains from different attributes (features, parameters, descriptors) to the conclusion.

In the next section both methodologies are briefly described. In section 3 the experiments in four medical diagnostic problems are described and in section 4 the different knowledge representations are discussed as well as the estimation of both approaches by physicians experts. In section 5 the extensions of the naive Bayesian classifier are described. Finally, in section 6 some conclusions are given and further research is proposed.

2 Assistant and Naive Bayesian Classifier

The classification problem discussed in this paper is the following: given a set of training instances, each described with a set of n attributes and each belonging to exactly one of a certain number of possible classes, learn to classify new, unseen objects. In addition, each attribute has a fixed number of possible values. For medical diagnosis, training instances are descriptions of patients with known final diagnoses, attributes are symptoms, anamnestic data, and results of laboratory tests, and classes are possible diagnoses.

Let C represent one of the possible classes. Let V_{i,J_i} be a Boolean variable having value 1 if the current instance has value J_i of i-th attribute and 0 otherwise. The conditional probability of class C given the values of all attributes is given with the following formula derived from the Bayesian rule [17] (for brevity conditions $V_{ij} = 1$ will be written simply as V_{ij}):

$$P(C|V_{1,J_1},\ldots,V_{n,J_n}) = P(C)\prod_{i=1}^{n} Q_i(C,J_i) \tag{1}$$

where

$$Q_i(C,J_i) = \frac{P(V_{i,J_i}|C,V_{1,J_1},\ldots,V_{i-1,J_{i-1}})}{P(V_{i,J_i}|V_{1,J_1},\ldots,V_{i-1,J_{i-1}})} = \frac{P(C|V_{1,J_1},\ldots,V_{i,J_i})}{P(C|V_{1,J_1},\ldots,V_{i-1,J_{i-1}})} \tag{2}$$

and $P(C)$ is the prior probability of class C. It was shown in [24] that the classification with ID3 like inductive learning system [45] can be described with (1). The basic learning algorithm is as follows:

```
(* initialization P'(C) := P(C); *)
if (P'(C) = 1) or (P'(C) = 0) or no more attributes
   then generate a leaf
    else
     begin
       select attribute i that minimizes the expected entropy;
       split the learning instances according to attribute's values;
       for each attribute's value $J_i$ do
          recursively construct a subtree with $P'(C) := P'(C) * Q_i(C,J_i)$
            from the corresponding subset of training instances
     end
```

Assistant [2,10] was derived from ID3 and incorporates mechanisms for dealing with incomplete and noisy domains. The main advantages with respect to ID3 are on-line binarization of attributes and pre and post-pruning techniques for discovering and pruning the unreliable parts of decision trees.

From (1) the naive Bayesian classifier, as used by Kononenko et al. [30], is obtained if the independence of attributes is assumed. Eq. (1) remains unchanged except that factors Q_i defined with (2) are replaced with Q'_i (we will refer to changed equation (1) with (1')):

$$Q'_i(C, J_i) = \frac{P(V_{i,J_i}|C)}{P(V_{i,J_i})} = \frac{P(C|V_{i,J_i})}{P(C)} \tag{3}$$

The probabilities necessary to calculate (3) are approximated with relative frequencies from the training set. A new object is classified by calculating the probability for each class using equation (1'). The object is classified into the class that maximizes the calculated probability.

Cestnik [8] has shown that the type of approximation of probabilities in (3) considerably influences the classification accuracy of the naive Bayesian classifier. Let $N(C, V_{i,J_i})$ be the number of training instances with J_i-th value of i-th attribute and belonging to class C and $N(V_{i,J_i})$ the number of training instances with J_i-th value of i-th attribute. Usually the probability is approximated with relative frequency, i.e.

$$P(C|V_{i,J_i}) = \frac{N(C, V_{i,J_i})}{N(V_{i,J_i})} \tag{4}$$

However, if the training set is relatively small, the corrections are needed with respect to the assumption of initial distribution [18]. Cestnik [8] used the so called *m-estimate* stemming from the assumption that initial distribution of classes is equal to $P(C)$:

$$P(C|V_{i,J_i}) = \frac{N(C, V_{i,J_i}) + m \times P(C)}{N(V_{i,J_i}) + m} \tag{5}$$

where the prior probability of class C is calculated using the Laplace law of succession [17]:

$$P(C) = \frac{N(C) + 1}{N + 2} \tag{6}$$

Cestnik has shown some nice properties of using approximation (5) in formula (3) and has shown experimentally, that naive Bayesian classifier using approximation (5) performs significantly better than if (4) is used. Parameter m is used to balance the influence of prior probability versus relative frequency of succesful experiments. In our experiments m was set to 2. Formula (5) was used also by Smyth & Goodman [52].

3 Experiments in Medical Diagnostics

We experimented with two learning systems described in the previous section in four medical diagnostic problems: diagnosing the location of primary tumor, prognosing the recurrence of breast cancer, diagnosing thyroid diseases and rheumatology. The data was collected at the University Medical Center in Ljubljana. The following are brief descriptions of diagnostic problems:

Localization of primary tumor: The medical treatment of patients with metastases is much more successful if the location of the primary tumor in the body of the patient is known.

The diagnostic task is to determine one of 22 possible locations of the primary tumor on the basis of age, sex, histological type of carcinoma, the degree of differentiation and 13 possible locations of discovered metastases. From The Institute of Oncology in Ljubljana the data for 339 patients with known location of the primary tumor was provided which was used in our experiments.

Prognostics of breast cancer recurrence: Among patients with removed breast cancer the disease recurs in five years after the operation in about 20% of cases. For better treatment it is necessary to prognose the possibility of the recurrence on the basis of age, size and location of the tumor and the data about lymphatic nodes. The problem is rather difficult for physicians specialists since due to long time observations (five years) little practical experience can be obtained. Furthermore, if a diagnosis is unreliable usually further examinations (e.g. laboratory tests) can be performed to verify the diagnosis. In prognostics there is no such possibility. For that reason the prognostic problems are even more attractive for machine learning than diagnostic problems [55]. From The Institute of Oncology in Ljubljana the data for 288 patients with known recurrence of breast cancer five years after the operation was used in our experiments.

Thyroid diseases: The diagnostic problem is to determine one of four possible diagnoses from age, sex, histological data, and results of laboratory tests. However, in everyday practice physicians use much more additional information for diagnosis, which was not available for computer processing. From The Clinic for Nuclear Medicine in University Clinical Center in Ljubljana the data for 884 patients with known final diagnoses was obtained and used in our experiments.

Rheumatology: The diagnostic problem is to select one of six groups of possible diagnoses from anamnestic data and status data. There is over two hundred diagnoses used by physicians specialists in rheumatology. However, general practitioners have to decide among rheumatological and orthopedical diseases for patients to be further investigated and treated by specialists. Such decisions are unreliable and, by the opinion of the physician specialist in rheumatology, in more than 30% of cases wrong. From Clinic for Rheumatology in University Medical Center in Ljubljana the data for 355 patients with known final diagnoses was provided for our experiments. All diagnoses were verified with additional observations, laboratory tests and Roentgen.

Table 1: Characteristics of four medical data sets.

domain name	# attributes	# diagnoses	# cases	entropy	maj. class
primary tumor	17	22	339	3.64bits	25%
breast cancer	10	2	288	0.72bits	80%
thyroid diseases	15	4	884	1.59bits	56%
rheumatology	32	6	355	1.70bits	66%

The characteristics of data sets used in our experiments are summarized in table 1. The entropy of distribution of classes (defined with $\sum_C P(C) \times log_2 P(C)$ $[bit]$) is interpreted as the expected amount of information necessary to correctly classify one instance. Together with the number of diagnoses, it shows the difficulty of the diagnostic problem. The number of attributes approximately tells how well the patients are described. The majority class is the prior probability of the most probable diagnosis and is in fact the classification accuracy of a simple classifier which for each patient selects the same most probable diagnosis.

One run was performed by randomly selecting 70% of instances for learning and 30% for testing. Results are averages of 10 runs and are given in table 2. The average percent of correct guesses is given together with the average *information score per answer* [29]. The average information score is the measure that eliminates the influence of prior probabilities of classes and can be applied to various kinds of incomplete and probabilistic answers. For completeness, its definition is provided in the appendix. This measure is necessary since in each domain a classifier that classified each instance into the majority class would achieve high classification accuracy.

Four physicians specialists in each domain were tested to estimate their diagnostic accuracy. From a set of training data a subset of patients was randomly selected and their description printed on paper without the final diagnosis. The physicians were asked to select the most probable diagnosis for each patient. The performances of physicians in table 2 are the averages of four physicians specialists in each domain that were tested at The University Medical Center in Ljubljana.

Both, Assistant and the naive Bayesian classifier, outperformed physicians specialists but naive Bayes was also better than Assistant. However, as already mentioned in section 1, such automatically generated knowledge base should be considered only as a tool for supporting and improving the diagnostic process and not as a replacement for physicians. Although in breast cancer and

Table 2: The comparison of performance of different classifiers in four medical domains.

classifier	primary tumor	breast cancer	thyroid	rheumatology
naive Bayes	49% 1.59bits	78% 0.15bits	70% 0.79bits	67% 0.51bits
Assistant	44% 1.38bits	77% 0.07bits	73% 0.87bits	61% 0.46bits
physicians	42% 1.22bits	64% 0.05bits	64% 0.59bits	56% 0.26bits

rheumatology the diagnosing of a patient on the paper is somewhat unnatural, for other two domains it often occurs in practice. However, in each diagnostic problem, physicians are able to take into account other informations that can not be used by a computer. The accuracy of physicians should therefore be considered as an estimate of how good the algorithms perform and not how bad the physicians diagnose.

4 Knowledge Representation and Explanation Ability

4.1 Knowledge derived by Assistant

The knowledge generated by Assistant is in the form of a decision tree. The root of a tree is usually the most important attribute for the given classification problem. In fact it is the attribute which minimizes the expected entropy, i.e. the expected amount of information, necessary to classify a new object. The physicians typically agree that the attribute at the root of the tree contains the most important information. Table 3 represents the ordering of attributes in the primary tumor domain by Assistant and by a physician specialist with respect to their significance. Similar table in the problem of diagnosing the craniostenosis syndrome is given in [1].

The physicians felt that the structure of the top of decision trees makes sense and is an interesting representation of the decision process. However, although a decision tree outperformed the physicians experts with respect to diagnostic accuracy the physicians were not prepared to use them in practice. The rules (paths from the root to the leaves of the tree) were too short, containing only few, although most informative, attributes and were too poorly describing the patient to make the reliable decision [44]. The physicians typically use all available information to make a decision and they are also able to estimate the reliability of the diagnosis. If the reliability is not high enough then additional examinations are needed.

Table 3: The comparison of the ordering of attributes with respect to their significance in the primary tumor domain by Assistant and by human expert.

physician's ordering	Assistant's ordering
histologic type	histologic type
sex	sex
brain	axillar
lung	bones
neck	differentiation
bones	diaphragm
age	peritoneum
skin	neck
peritoneum	lung
...	...

4.2 Explanation ability of naive Bayes

The knowledge generated and used by the naive Bayesian classifier is simply a table of prior and conditional probabilities approximated with relative frequencies from the training set. Physicians evaluated the table as providing useful information. E.g., the prior probability of location of the primary tumor in lung is $P(lung) = 0.25$ while if the "histologic type" is known to be "adeno" the priobability changes into $P(lung|histologictype = adeno) = 0.14$.

Furthermore, the classification process of the naive Bayesian classifier can be naturally interpreted using the definition of information [51]. The logarithm of (3) is interpreted as the information gain by i-th attribute to the conclusion that an object belongs to class C:

$$\log_2 Q'_i(C, J_i) = -\log_2 P(C) - (-\log_2 P(C|V_{i,J_i})) \tag{7}$$

i.e. the prior information necessary to classify into class C minus the posterior information necessary to make that decision. The minus logarithm of (1') gives:

$$-\log_2 P(C|V_{1,J_1}, \ldots, V_{n,J_n}) = -\log_2 P(C) - \sum_{i=1}^{n} \log_2 Q'_i(C, J_i) \tag{8}$$

which is interpreted as the amount of information necessary to find out that an object belongs to class C given the values of all attributes. This amount of information before the classification is equal to $-\log_2 P(C)$. Therefore the right hand side of (8) is interpreted as the *sum of information gains* from all attributes to the conclusion that an object belongs to class C.

Eq. (8) is appropriate for explanation of a decision if $P(C|V_{1,J_1},\ldots,V_{n,J_n}) > P(C)$. If $P(C|V_{1,J_1},\ldots,V_{n,J_n}) < P(C)$ the information is obtained for class $\overline{C}$. Therefore eq. (8) is changed into:

$$-\log_2 P(\overline{C}|V_{1,J_1},\ldots,V_{n,J_n}) = -\log_2(1-P(C)) - \sum_{i=1}^{n}(-\log_2(1-P(C)) + \log_2(1-P(C|V_{i,J_i}))) \quad (9)$$

Equations (8) and (9) are appropriate for explanation of decisions of the naive Bayesian classifier as the sum of information gains from all attributes. A similar scheme is proposed by Smyth et al. [53], and for decision trees using the *weight of evidence* by Michie & Al Attar [38]. The analogy between decision tree and naive Bayesian approaches is analyzed in more detail in [24].

To estimate the understandability, correctness and usefulness of such an explanation, the following experiments in four medical domains were performed. A set of all instances was randomly divided into 70% of instances for training and 30% for testing. After the training each testing instance (with removed diagnosis) was classified with the naive Bayesian classifier. The explanation form of one classification is given in Table 4. We randomly selected 5 testing patients correctly diagnosed with the naive Bayesian classifier and 5 incorrectly diagnosed. It is assumed that a diagnosis is correct if the most probable diagnosis returned by the classifier (the first appearing in the form in figure 1) is the correct diagnosis. In each medical domain there was therefore 10 explanations of diagnostic decisions.

These explanations were then estimated by physicians specialists. Physicians did not know the correct diagnoses of the patients. They had to estimate the correctness of decisions, the correctness of the explanation and the understandability. In addition, they were asked to estimate the usefulness of such on-line support of the diagnostic process for physicians specialists, nonspecialists and for education of students. The results of their estimations are given in table 5. For thyroid diseases and rheumatology there is no estimation of usefulness because the data used in our experiments do not follow the usual way the physicians work and do not include all the attributes needed for reliable diagnosis.

Physicians found such explanation of classification as natural and similar to their classification. They also sum up the evidence for/against a diagnosis. They were immediately prepared to use the naive Bayesian classifier with the incorporated facility of explaining the classification by means of equations (8) and (9) as a tool for enhancing the reliability of a diagnostic process. To improve understandability the physicians suggested that the description of the patients should include, besides general attributes (age, sex, ...), only attributes with pathological values. Although such

Table 4: A form containing the explanation of a diagnosis of one patient. YY is the prior probability of a given diagnosis and XX is the probability returned by the classifier. Ni is the strength of confirmation of a decision by i-th attribute and Mj is the strength of rejection by j-th attribute.

description of a patient:

attribute	value	attribute	value	attribute	value
Att 1	Val i	Att 2	Val j	Att 3	Val k
...					

====================================

diagnosis of a patient			———-EXPLANATION————–			
poss.diag.	prob.	(prior)	confirm. attribute		reject.attribute	
Diag D1	XX %	(YY %)	Att i	Ni	Att j	Mj
			Att k	Nk	Att l	Ml
			...		...	
...						

Table 5: The results of physicians' estimation of explanation form for 4 medical diagnostic problems. Each estimation is in the form ALL/WR where ALL is the average over all 10 explanation forms and WR is the average over 5 incorrectly classified patients. One estimate is the number between 0 and 10.

	primary tumor	breast cancer	thyroid	rheumat.
corr. of diagnosis	7.7/6.0	8.0/8.0	7.9/7.6	5.2/3.6
corr. of explanation	5.9/5.2	7.4/7.8	8.2/8.2	3.8/4.3
understandability	7.6/6.4	8.4/8.6	7.9/8.2	5.0/5.0
useful for specialist	4.5/5.0	5.8/5.4	-	-
useful for nonspecialist	6.0/5.6	7.3/7.0	-	-
useful for students	7.6/6.2	8.2/8.2	-	-

reduction of information is not drastic, it may have critical influence on classification accuracy in nontrivial cases. This may also explain poor performance of physicians as compared with naive Bayesian classifier in table 2. The other suggestion was that the trivial and immediately obvious decisions should be omitted (e.g. that a female patient can not have a primary tumor in a prostate).

From table 5 it can be concluded that explanation of the naive Bayesian classifier decision is understandable to physicians and that the explanation shows the similarity between the system's and the physicians' way of making decisions. Only in rheumatology the results are not good enough. A detailed analysis showed that the explanation is correct if the information gain of attributes is observed *independently.* Because of some strong correlations between attributes in this domain attributes must not be observed independently. The attempt to overcome this problem is described in section 5.2. In table 5 one can see that there is no significant difference between the estimation of systems performance for correct and incorrect decisions. Therefore when the limited amount of information is present the system and the human experts make similar mistakes.

The overall physician's impression was that the form with explanation typically replicates their way of diagnosing, i.e. the summation of evidence for/against the diagnosis. Only in the 'breast cancer' domain physicians found certain decisions as new information interesting for further investigation. In all other domains the physicians felt that nothing new is contained in the explanations that they did not know before. They had the opinion that in cases where they did not agree with the explanation, the system made a mistake.

Table 6: Characteristics of two classifiers

	naive Bayes	Assistant
		(non-naive Bayes)
generates	probabilities	decision tree
knowledge	implicit	explicit
explanation	inf. gains	if-then rule
domains	inexact	exact
# atts used	all	few
missing data	insensitive	sensitive
prob.approx.	reliable	unreliable
speed	fast	slow
incremental	yes	no
mulival. atts	sensitive	insensitive
independence	assumed	not assumed

5 Extensions of naive Bayesian classifier

In table 6 the characteristics of naive Bayes and Assistant (which represents non-naive Bayes, see section 2) are sketched. The generated knowledge by Assistant is in the form of a decision tree while naive Bayes generates probabilities. The top part of a decision tree typically shows the structure of the problem. The decision tree can be used without a computer to classify new objects. Therefore, it is a kind of *explicit knowledge.* On the other hand, the probabilities generated by naive Bayes cannot be directly used to classify new objects. This kind of knowledge is *implicit.* The physicians found both types of knowledge interesting and useful.

The explanation of classification of a new object in Assistant is simply the if-then rule used for the classification while in naive Bayes the explanation is the sum of information gains from all attributes for/against the conclusion. Physicians preferred the sum of information gains as more natural explanation, similar to the way physicians diagnose. While if-then rules typically include too few attributes for reliable classification [44], naive Bayes uses all available attributes. Besides, learning of decision rules and classification with decision rules is very sensitive to missing data [47] and the learning algorithm is slower and not incremental. A missing value of an attribute in naive Bayes is simply ignored and the learning process is relatively fast and essentially incremental.

The major advantage of naive Bayes is the reliability of approximation of probabilities. Due to small number of training instances covered by single decision rule the final decision of the rule is unreliable. Pruning of decision trees partially overcomes this problem. However, due to pruning, rules are shortened and more attributes are discarded from diagnostic process.

The only advantages of Assistant over the naive Bayesian classifier are appropriateness for dealing with *continuous attributes* and *"non-naivety"* (there is no independence assumption). In next subsections the extensions to the naive Bayesian classifier are described that overcome these problems.

5.1 Naive Bayes and continuous attributes

In rheumatology and thyroid diagnostic problem a lot of attributes are continuous (7 in thyroid and 22 in rheumatology data). Assistant with on-line binarization of attributes [2,30] successfully solves the problem of continuous and multivalued attributes (as shown in the case of thyroid diseases, see table 2). The naive Bayesian classifier assumes that all attributes are discrete. Therefore, continuous attributes must be converted to discrete by introducing a number of fixed bounds. This can be done using one of the algorithms for discretization of continuous attributes [6,7], or can be done manually by a human expert. The latter approach was used in our experiments described in section 3. However, the approach with fixed (exact) bounds destroys the integrity of the training set and ignores the order of values.

Better approach is to use *fuzzy bounds* for continuous attributes that overcomes both the loss of the information about the order of values as well as the loss of integrity of the training set. In the following one of possible algorithms for dealing with fuzzy bounds is briefly described together with results in two medical domains with continuous attributes. The method is described in more details in [28]. The task is to calculate the probabilities of all classes of an object with a given value of a continuous attribute. These probabilities should be approximated with relative frequencies calculated from the distribution of training instances with the similar value of the attribute. It is expected that small variations of the value of the attribute should have small effects on the probabilities. As opposed to exact bounds, where slightly different value can have drastic effects on the calculated probabilities, the bounds of intervals are assumed to be fuzzy.

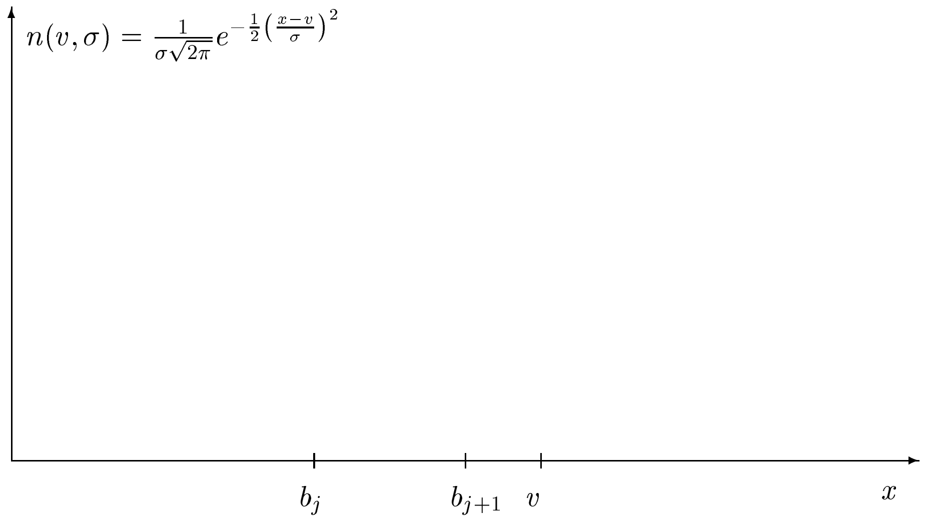
$n(v,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-v}{\sigma}\right)^2}$
b_j
b_{j+1}
v
x

Figure 1: The normal distribution of the influence of:
- an instance over intervals of a continuous attribute *for fuzzy learning*,
- an interval on classification for *fuzzy classification*

Learning and classification phase of the algorithm are modified to assume fuzzy bounds of continuous attributes. The pessimistic set of possible bounds is given in advance either by a human expert or with a simple algorithm, that returns bounds with the uniform distribution of instances over all intervals. In our experiments the former approach was used. The set of bounds is pessimistic in the sense that more bounds are given than probably needed (e.g. all attributes have in advance 20 possible intervals, which is typically a too detailed split). However, exact values of these initial bounds are not important and may vary without significant changes in performance since later there is effectively interval fusion.

First, the *fuzzy learning* is performed by calculating the probability distribution for a given interval from all training instances rather than from instances that have value of a given continuous attribute in this interval. The influence of an instance is assumed to be normally distributed with mean value equal to the value of the regarded attribute and with given σ. σ is the parameter to the learning algorithm and is used to control the 'fuzziness' of the bounds. As shown in figure 1, the influence of a given instance with value v of the given continuous attribute on the distribution of interval $(b_j..b_{j+1})$ is proportional to the following expression:

$$P(v,\sigma,j) = \int_{b_j}^{b_{j+1}} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-v}{\sigma}\right)^2} dx \qquad (10)$$

If $\sigma = 0$ then the usual exact bounds result and the distribution over classes in the given interval is calculated from relative frequency of training instances belonging exactly to that interval. The greater σ implies fuzzier bounds for continuous attributes. Therefore, for N learning instances, each with influence $P(v_k, \sigma, J_i)$, $k = 1..N$ on J_i-th interval of i-th attribute the probability that an instance belongs to that interval is calculated with:

$$\hat{P}(V_{i,J_i}) = \frac{\sum_{k=1}^{N} P(v_k, \sigma, J_i)}{N} \tag{11}$$

and for N_C learning instances that belong to class C, each with influence $P(v_{k_C}, \sigma, J_i)$, $k_C = 1..N_C$ on J_i-th interval of i-th attribute the probability that an instance belongs to class C and that interval is calculated with:

$$\hat{P}(C\&V_{i,J_i}) = \frac{\sum_{k_C=1}^{N_C} P(v_{k_C}, \sigma, J_i)}{N} \tag{12}$$

Each interval corresponds to one attribute's value. With exact bounds a given value of a continuous attribute influences exactly one interval. By making bounds fuzzy it influences all the intervals (values) for a given attribute and the influence is normally distributed.

The *fuzzy classification* is performed by calculating the probability of all classes for a given object, given the value of the continuous attribute, from all intervals of that attribute rather than from the interval to which the object belongs. The influence of intervals is assumed to be normally distributed with mean value equal to the value of the regarded attribute of an object and with given σ. σ is like in fuzzy learning the parameter to the classification algorithm and is used to control the 'fuzziness' of the bounds. As shown in figure 1, the influence of a given interval on the probability of all classes is proportional to the expression (10). The expression (3) in equation (1') is here replaced by:

$$Q_i"(C) = \sum_{j=1}^{NV_i} P(v, \sigma, j) \times \frac{\hat{P}(C|V_{i,j})}{\hat{P}(C)} \tag{13}$$

where v is the value of i-th attribute of a given object. Like in fuzzy learning, for $\sigma = 0$ the usual exact bounds are assumed and the probabilities of all classes for the given object are calculated from one interval only.

In experiments in two medical domains parameter σ was determined with the following formula for each continuous attribute A_i:

$$\sigma_i = SIG \times \frac{upperbound_i - lowerbound_i}{\#intervals_i}$$

where SIG is parameter that was varied in our experiments. If $SIG = 1$ then σ_i is equal to the average interval length for i-th attribute. Therefore fuzziness is the function of average length of intervals. In table 7, the results are given for various values of parameter SIG. Results are again

Table 7: Results of naive Bayesian classifier with fuzzy bounds for continuous attributes.

	thyroid		rheumatology	
SIG	acc(%)	inf.(bit)	acc(%)	inf.(bit)
0.0	70	0.79	67	0.51
0.3	72	0.85	69	0.58
0.4	72	0.85	69	0.57
0.5	72	0.86	69	0.58
0.6	72	0.85	69	0.59

averages over 10 runs with randomly selected training and testing sets. Obviously, results are not very sensitive to parameter SIG as long as $SIG > 0$.

5.2 Semi-naive Bayesian classifier

If attributes are human defined (as was the case in medical data used in our experiments) attributes are usually relatively independent, as humans tend to think linearly. However, independence assumption is often not justified. To avoid this problem, the algorithm should detect the dependencies among attributes and join dependent attributes together. Besides, instead of joining whole attributes, only single values of different attributes can be joint, which is more flexible.

When calculating the probability of class C_j in (1') the influence of attributes A_i and A_l is defined with:

$$\frac{P(C_j|V_{i,J_i})}{P(C_j)} \times \frac{P(C_j|V_{l,J_l})}{P(C_j)} \tag{14}$$

If, instead of assuming the independence of values V_{i,J_i} and V_{l,J_l}, the values are joint, the corrected influence is given with:

$$\frac{P(C_j|V_{i,J_i}V_{l,J_l})}{P(C_j)} \tag{15}$$

For joining the two values two conditions should be satisfied: the values of (14) and (15) should be sufficiently different while the approximation of $P(C_j|V_{i,J_i}V_{l,J_l})$ should be sufficiently reliable. The *semi-naive Bayesian classifier*, described in more details in [26], tries to solve this *trade-off between non-naivety and reliability of approximations of probabilities.*

With semi-naive Bayesian classifier the same experiments were repeated in four medical diagnostic problems. In primary tumor and breast cancer the results were the same as those with the naive Bayesian classifier. The results were expected as the physicians claimed that attributes in these two domains are in fact independent. In other two domains (thyroid and rheumatology) the classification accuracy was in average better for one percent. This indicates that there are some dependencies among attributes which are not very strong. However, the explanation ability of the semi-naive Bayesian classifier may be better than that of naive Bayesian classifier in these two domains (see table 5).

6 Discussion

Table 6 shows the characteristics of both approaches described in this paper. The only disadvantages of the naive Bayesian classifier are inability to deal with continuous attributes and naivety. Fuzzy bounds for continuous attributes as described in section 5.1 overcome the former disadvantage and the semi-naive Bayesian classifier as described in section 5.2 the latter. Therefore, Bayesian approach seems more appropriate for automatic diagnosis in medicine than inductive learning approach.

When classifying new objects with a decision tree the values of only few attributes are examined, those corresponding to the path from the root of a tree to the leaf. In some cases this is an advantage as the values of attributes can be expensive to obtain. However, if attributes are available it is better to take them into account. Note that medical data are not exact and few attributes do not suffice to discriminate between different diagnoses. Besides, if the value for some attribute in the tree is missing (e.g. for the attribute at the root of the tree) the classification becomes unreliable [2,47]. Physicians felt that the patient cannot be reliably diagnosed with only a few attributes [44].

On the other hand, naive Bayesian classifier takes into account all the available attributes. If the value of an attribute is missing, such attribute is simply omitted from equation (1'). The problem with naive Bayes is the independence assumption. In some cases this may be too unrealistic assumption. But it seems that in the data used by human experts there are no strong dependencies between attributes because attributes are properly defined. With the independence assumption the reliability of approximating factors (3) with relative frequencies is much greater than the reliability of approximating factors (2). This is supported with experimental results. Naive Bayesian classifiers despite its naivety achieved better classification accuracy. There is a trade-off between the reliability of approximating probabilities and the errors due to the independence assumption. Semi-naive Bayesian classifier [26] tries to optimize this trade-off. By decreasing the reliability of probability approximations the 'non-naivety' increases, which can be useful for *exact* domains. For *inexact* (fuzzy) domains, like medical diagnosis, the reliability of probability approximations should be higher. Naive Bayes is due to the independence assumption more appropriate for inexact domains while Assistant is appropriate for exact domains with, ideally, complete set of attributes *and* complete set of training instances.

Assistant derives from training instances *explicit* if-then rules tied together in a decision tree. Rules are easy to understand and can be used without a computer, simply listed on the paper. Assistant tends to generate a *few general rules.* A decision tree represents the global structure of the classification problem.

The naive Bayesian classifier's knowledge can be viewed as *implicit*, as it can not be directly applied without calculation. A table of (conditional) probabilities represents a detailed and distributed knowledge without any global picture about the structure of the problem. Naive Bayes performs like if it knew the rules. However, the rules that it uses for classifying objects are typically different for each object from a domain. Therefore, naive Bayes uses a *lot of specialized rules.*

Such kind of knowledge is typical for neural networks [42,50]. It was shown in [24,25,28] that the naive Bayesian classifier can be naturally implemented with the *Bayesian neural network.* Such a network is able to explain its decisions (which is normally not the case with neural networks) while preserving the advantages of neural networks: multidirectional classification, robustness with respect to noisy and missing data, and fast and incremental learning. Induction of decision trees is relatively slow as compared to naive Bayesian classifier (however fast enough to be applicable to thousands of examples), non-incremental, and sensitive to missing and noisy data.

Better classification accuracy is not enough for an expert system to be used in practice. It must be able to explain its decisions and it must be flexible enough to provide all possible alternatives. The overall physician's impression was that the form of naive Bayesian classifier's explanation typically replicates their way of diagnosing, i.e. the summation of evidence for/against the diagnosis. In the 'breast cancer' domain physicians found certain decisions as new information interesting for further investigation.

Further development of machine learning techniques is concerned with the so called *redundant* or *multiple* knowledge [9,16,27,53]. The idea is to generate redundant sets of decision rules which can be dynamically combined to classify new objects in a problem domain. This approach avoids the problem of missing data as well as the problem of bad explanation by decision rules in a similar way as the naive Bayesian classifier does.

Acknowledgements

Collecting and assembling the experimental data would not be possible without the invaluable help of physicians specialists Dr. Matjaž Zwitter, Dr. Sergej Hojker and Dr. Vlado Pirnat from the University Medical Center in Ljubljana. I thank them for providing and interpreting the data, for testing the diagnostic performance of physicians specialists from the University Medical Center, for the interpretation of results and for the estimation of the explanation abilities of the naive Bayesian classifier. This research was supported by Slovenian Research Community. The reported work was done in the Artificial Intelligence Laboratory at the Faculty of Electrical and Computer Engineering in Ljubljana. I would like to thank Prof. Ivan Bratko for providing the environment for efficient scientific work and Matevž Kovačič for his comments on the manuscript. I thank also the anonymous reviewer for valuable comments and suggestions.

References

[1] Baim P.W., A Method for Attribute Selection in Inductive Learning Systems, *IEEE Trans. on PAMI*, Vol.10, No. 6, 1988, pp.888-896.

[2] Bratko I. & Kononenko I., Learning Rules from Incomplete and Noisy Data, in B. Phelps (ed.) *Interactions in Artificial Intelligence and Statistical Methods*, Hampshire: Technical Press, 1987.

[3] Bratko I., Mozetič I., Lavrač N., *KARDIO: A study in deep and qualitative knowledge for expert systems*, Cambridge,MA: MIT Press, 1989.

[4] Bratko I., Mulec P., An Experiment in Automatic Learning of Diagnostic Rules, *Informatica*, Ljubljana, Vol.4, No.4, 1980, pp. 18-25.

[5] Breiman L., Friedman J.H., Olshen R.A., Stone C.J., *Classification and Regression Trees*, Wadsforth International Group, 1984.

[6] Catlett J., On changing continuous attributes into ordered discrete attributes, *Proc. European Working Session on Learning-91*, Porto, March 4-6, 1991, pp. 164-178.

[7] Cestnik B., Informativity-based splitting of numerical attributes into intervals, *Proc. IASTED internat. conf. Expert Systems & Applications*, Zurich, June 26-28, 1989, pp. 59-62.

[8] Cestnik B., Estimating Probabilities: A Crucial Task in Machine Learning, *Proc. European Conf. on Artificial Intelligence*, Stockholm, August, 1990, pp. 147-149.

[9] Cestnik B., Bratko I., Learning redundant rules in noisy domains, *Proc. European Conf. on Artificial Intelligence*, Munich, 1988, pp.348-351.

[10] Cestnik B., Kononenko I.& Bratko I., ASSISTANT 86 : A knowledge elicitation tool for sophisticated users, in: I.Bratko, N.Lavrac (eds.): *Progress in Machine learning*, Wilmslow: Sigma Press, 1987.

[11] Chan K.C.C. & Wong A.K.C., Automatic Construction of Expert Systems from Data: A Statistical Approach, *Proc. IJCAI Workshop on Knowledge Discovery in Databases*, Detroit, Michigan, August, 1989, pp.37-48.

[12] Clark P. & Boswell R., Rule Induction with CN2: Some Recent Improvements, *Proc. European Working Session on Learning-91*, Porto, Portugal, March, 1991, pp.151-163.

[13] Clark, P., Niblett, T., Learning if then rules in noisy domains. In: B. Phelps (ed.), *Interactions in Artificial Intelligence and Statistical Methods.* Hampshire, England: Technical Press, 1987.

[14] P.Clark, T.Niblett, The CN2 Induction Algorithm, Machine Learning, Vol 3, 1989, pp. 261-283.

[15] Elomaa T., Holsti N., An Experimental Comparison of Inducing Decision Trees and Decision Lists in Noisy Domains, *Proc. 4th European Working Session on Learning*, Montpeiller, Dec. 4-6, 1989, pp.59-69.

[16] Gams M., New Measurements Highlight the Importance of Redundant Knowledge, *Proc. 4th European Working Session on Learning*, Montpellier, Dec. 4-6, 1989, pp. 71-80.

[17] Good, I.J., *Probability and the Weighing of Evidence*, London: Charles Griffin, 1950.

[18] Good I.J., *The Estimation of Probabilities*, Cambridge: M.I.T. Press, 1965.

[19] Goodman R.M.F & Smyth P., ITRULE: An Information Theoretic Rule-Induction Algorithm, *Proc. 1st European Workshop on Knowledge Acquisition for Knowledge-based Systems*, Reading University, Sept. 2-3, 1987.

[20] Hojker S., Kononenko I., Jauk A., Fidler V. & Porenta M., Expert System's Development in the Management of Thyroid Diseases, *Proc. European Congress for Nuclear Medicine*, Milano, Sept., 1988.

[21] Horn K.A., Compton P., Lazarus L., Quinlan J.R., An Expert System for the Interpretation of Thyroid Assays in a Clinical Laboratory, *The Australian Computer Journal*, Vol. 17, No. 1, 1985, pp.7-11.

[22] Karalič A., Pirnat V., Significance Level Based Classification with Multiple Trees, *Informatica*, Ljubljana, Vol.15, No. 1, 1991, pp.54-58.

[23] Kern J., Deželič G., Težak-Benčič M., Durrigl T., Medical Decision Making Using Inductive Learning Program (in Croatian), *Proc 1st Congress on Yougoslav Medical Informatics*, Beograd, Dec. 6-8, 1990, pp.221-228.

[24] Kononenko I., ID3, Sequential Bayes, Naive Bayes and Bayesian Neural Networks. *Proc. 4th European Working Session on Learning*, Montpeiller, Dec. 4-6, 1989, pp.91-98.

[25] Kononenko I., Bayesian neural networks, *Biological Cybernetics*, Vol. 61, 1989, pp.361-370.

[26] Kononenko I., Semi-naive Bayesian classifier, *Proc. European Working Session on Learning-91* (Y.Kodratoff (ed.), Springer-Verlag), Porto, March 4-6 1991, pp.206-219.

[27] Kononenko I., An experiment in machine learning of redundant knowledge, *Proc. Intern. Conf. MELECON 1991*, Ljubljana, May 1991, pp.1146- 1149.

[28] Kononenko I., Feedforward Bayesian Neural Networks and Continuous Attributes, *Proc. Int.Joint Conference on Neural Networks IJCNN-91*, Singapore, 18-21 Nov. 1991, pp146-151.

[29] Kononenko I. & Bratko I., Information based evaluation criterion for classifier's performance, *Machine Learning*, Vol.6, No.1, 1991, pp.67-80.

[30] Kononenko I., Bratko I., Roškar E.: Experiments in automatic learning of medical diagnostic rules, International School for the Synthesis of Expert's Knowledge Workshop, Bled, August, 1984.

[31] Kononenko I., Jauk A. & Janc T., Induction of Reliable Decision Rules, International School for the Synthesis of Expert's Knowledge Workshop, Udine, 10-13 Sept., 1988.

[32] Lesmo L., Saitta L., Torasso P., Learning of Fuzzy Production Rules for Medical Diagnoses, In: Gupta M.M. & Sanchez E.(eds.) *Approximate reasoning in Decision Analysis*, North-Holland, 1982.

[33] Michalski, R.S., Carbonell, J.G. & Mitchell, T.M. (eds.), *Machine Learning: An Artificial Intelligence Approach*, Tioga Publ. Comp., 1983.

[34] Michalski, R.S., Carbonell, J.G. & Mitchell, T.M. (eds.), *Machine Learning: An Artificial Intelligence Approach, Volume II*, Morgan Kaufmann, 1986

[35] Michalski, R.S., Chilausky, R.L., Learning by Being Told and Learning from Examples: An Experimental Comparison of the Two Methods of Knowledge Acquisition in the Context of Developing An Expert System for Soybean Disease Diagnosis. *International Journal of Policy Analysis and Information Systems.* Vol.4, No.2., pp. 125-161.

[36] Michalski, R.S., Mozetič, I., Hong, J., Lavrač, N., The Multi Purpose Incremental Learning System AQ15 and Its Testing Application to Three Medical Domains. *Proc. of the National Conf. on Artificial Intelligence AAAI 86.* Philadelphia, August, 1986, pp.1041-1047.

[37] Michie D., Machine learning and knowledge acquisition. In: *International Handbook of Information Technology and Automated Office Systems.* North-Holland, Elsevier Science Publishers, 1985.

[38] Michie, D., A. Al Attar, Use of Sequential Bayes with Class Probability Trees. In: J.E. Hayes-Michie, D.Michie & E. Tyugu (eds.) *Machine Intelligence 12*, Oxford: Oxford University Press, 1991.

[39] Mozetič, I., Lavrač, N., Kononenko, I., Automatic construction of diagnostic rules. *Proc. of IV. Mediterranean Conf. on Medical & Biological Engineering.* Sevilla, Spain, Sept. 9-12, 1986.

[40] Muggleton S., *Inductive Acquisition of Expert Knowledge*, Turing Institute Press & Addison-Wesley, 1990.

[41] Nunez M., Decision Tree Induction Using Domain Knowledge, In: Wielinga B. et al. (eds.) *Current Trends in Knowledge Acquisition*, Amsterdam: IOS Press, 1990.

[42] Pao Y.H., *Adaptive Pattern Recognition and Neural Networks.* Addison-Wesley, 1989.

[43] Paterson, A., Niblett, T., The ACLS User Manual, Intelligent Terminals Ltd, Glasgow, 1982.

[44] Pirnat V., Kononenko I., Janc T. & Bratko I., Medical Estimation of Automatically Induced Decision Rules, *Proc. of 2nd Europ. Conf. on Artificial Intelligence in Medicine*, City University, London, August 29-31, 1989, pp.24-36.

[45] Quinlan J.R., Discovering rules by induction from large collections of examples, in D.Michie (ed.): *Expert systems in the Micro Electronic Age*, Edinburgh University Press, 1979.

[46] Quinlan J.R., Induction of Decision Trees. *Machine Learning.* Vol. 1, No. 1, 1986, pp. 81-106.

[47] Quinlan J.R., Unknown attribute values in induction, *Proc. 6th Int.Workshop on Machine Learning*, Cornell University, Ithaca, June 26-27, 1989, pp.164-168.

[48] Quinlan R., Compton P., Horn K.A., Lazarus L., Inductive knowledge acquisition: A case study, in: J.R.Quinlan (ed.) Applications of expert systems, Turing Institute Press & Addison- Wesley, 1987. (Also: *Proc. 2nd Australian Conf. on Applications of Expert Systems*, Sydney, May 14-16, 1986)

[49] E.Roškar, P.Abrams, I.Bratko, I.Kononenko, A.Varšek, MCUDS - An expert system for the diagnostics of lower urinary tract disorders, *Journal of Biomedical Measurements, Informatics and Control*, Vol. 1, No. 4, 1986, pp. 201 - 204.

[50] Rumelhart, D.E. & McClelland, J.L. (eds.), *Parallel Distributed Processing, Vol. 1: Foundations.* Cam-

bridge: MIT Press, 1986.

[51] Shannon, C.E. & Weaver, W., *The mathematical theory of communications.* The University of Illinois Press, Urbana, 1949.

[52] Smyth P. & Goodman R.M., Rule Induction Using Information Theory, In: G.Piarersky-Shapiro, W.Frawley (eds.) *Knowledge Discovery in Databases*, The MIT Press, 1990.

[53] Smyth P., Goodman R.M., Higgins C., A hybrid Rule-based Bayesian Classifier, *Proc.European Conf. on Artificial Intelligence*, Stockholm, August, 1990, pp. 610-615.

[54] Steels L., Second generation expert systems. *Future Generation Computer Systems*, Vol.1, No.4, 1985, pp.213-221.

[55] Zwitter M., Bratko I., Kononenko I., Rational and Irrational Reservations Against the Use of Computer in Medical Diagnosis and Prognosis, *Proc. 3. Mediterranean conf. on medical and biological engineering*, Yugoslavia: Portorož, Sept. 5-9, 1983.

Appendix: Definition of the information score of an answer

A fair evaluation criterion has to exclude the influence of the prior probabilities of classes which may enable a completely uninformed classifier to trivially achieve high classification accuracy. The measure of information score of the classifier's answer defined below excludes the influence of prior probabilities, deals with various types of imperfect and probabilistic answers and can be used also for comparing the performance in different domains. Its interpretation is natural.

Definition [29]:

Let the correct class of an instance be C, $P(C)$ be the prior probability of class C and $P'(C)$ the probability of class C returned by a classifier. The information score I of classifier's answer is defined as follows (note that in our experiments only exact classification is considered, i.e. $P'(C)$ is always either 1 or 0):

a) if $P'(C) > P(C)$ then

$$I = -\log_2 P(C) + \log_2 P'(C) \quad [bits]$$

i.e., the amount of obtained information is the entire amount of information necessary to correctly classify an instance into class C minus the remainder of information necessary to correctly classify that instance.

b) if $P'(C) = P(C)$ then $I = 0[bits]$

i.e., the system did not change the prior probability of the correct class therefore we did not obtain any information.

c) if $P'(C) < P(C)$ then

$$I = -(-\log_2(1 - P(C)) + \log_2(1 - P'(C))) \quad [bits]$$

i.e., the amount of information returned by the system is the entire amount of information necessary to decide that an instance does not belong to class C minus the remainder of information necessary to make that decision. As this information is in fact wrong the information score of the system's answer in this case is defined as negative.