UDK 519.226.3

Matjaž Gams, Matija Drobnič
Jozef Stefan Institute

ABSTRACT. One of the recently developed systems for machine learning (GINESYS) significantly outperformed all compared systems including theoretically optimal Bayesian classifier, which was the second in both tests. We tested several options in Bayesian classifier to investigate the real cause for nonoptimal results and to estimate the upper limit in classification accuracy. The conclusion is that while it is possible to achieve even higher classification accuracy with suitable parameter adjustment in Bayesian classifier, it seems that GINESYS practically achieved the optimal classification accuracy.

Sistem za empirično učenje GINESYS je v praktičnih meritvah presegel primerjane sisteme z Bayesovim klasifikatorjem vred. Podrobna analiza kaže, da je dosežene rezultate mogoče preseči, vendar so že zelo blizu optimalne meje.

1. INTRODUCTION

Machine learning is a quickly developing area of Artificial Intelligence [Winston]. According to the major inference type used it can be divided into rote learning, learning from instruction, learning by deduction, by analogy, from examples and from observation and discovery [Carbonell et al; Michalski]. The scope of this article is learning from examples or Empirical Learning (EL). The aim of EL is to induce general descriptions of concepts from examples (instances) of these concepts. Examples are usually objects of a known class described in terms of attributes and values. The final product of learning are symbolic descriptions in human understandable forms. Induced descriptions of concepts, representing different classes of objects, can be used for classifying new objects. EL systems basically perform the same task (classification) as statistical methods and can be directly compared to them from the point of classification accuracy. On the other hand, EL systems offer further advantages, namely a) explanation during classification of new examples and b) the insight into the laws of the domain by observing classification rules. Explanation during classification (a) is important since it enables the user to check the line of reasoning and verify the system's decision. The knowledge base (b) can be viewed as a new representation of the domain knowledge, which can be of great value to domain experts, especially in domains that are not yet well formalized and understood.

2. A SIMPLE EXAMPLE

For a simple example let us consider a case where we have a device with 8 binary switches representing 256 legal combinations. Device reports errors in some combinations and we want to find out what subsequence causes them.

	SWITCHES 1 2 3 4 5 6 7 8	STATUS
1	0 1 1 0 1 0 0 1	ERROR
2	1 0 1 1 1 1 1 0	OK
3	0 0 1 0 1 1 0 1	ERROR
4	1 0 1 1 0 0 1 1	OK
5	1 0 1 1 1 0 1 1	ERROR

Table 1. Device reports error in some combinations of switches. Which subsequence of switches causes them?

Probably the most common answer in EL systems would be, that error is reported when switches 5 and 8 are on (=1).

In practical tasks EL systems deal with domains with 10 to 10.000 examples (typically some hundred) with 2 to 500 attributes (typically ten or some ten) [Breiman et al; Quinlan; Lavrac et al]. Attributes can be real, integer or categorical with many possible values.

3. EMPIRICAL LEARNING

The whole process of empirical learning consists of four steps:
- preprocessing of learning examples,
- construction of a classification rule,
- classification of new instances and
- analysing the laws of the domain.

Detailed description can be found elsewhere, e.g. [Kononenko] or [Gams, Lavrac] with detailed overview of some well known algorithms - C4 [Quinlan], CART [Breiman et al], ASSISTANT 86 [Cestnik et al], CN2 [Clark, Niblett], AQ15 [Carbonell et al]. We shall formally represent here only a domain area and a classification rule.

A set of learning examples L = {(x,c)} consists of pairs (x,c), where x is a vector (denoting properties of the object) in a measurement space X and c represents the index of the class of example x.

Components of vectors x are called attributes or variables. The values of attributes can be numerical or categorical.

A classification or a decision rule d(x) is a mapping which maps every x from X into some c from C or into the probability distribution (p1,p2...pJ) where pi is a real number between 0 and 1.

A classification rule d(x) splits the whole space X into spaces X1,X2,...XJ, such that for every Xi only a certain subset of d(x) is relevant.

The syntax of a classification rule d(x) is:

```
<d>       ::= <Rule> | <Rule> and  <d>
classification rule

<Rule>    ::= <Class> if  <Cpx>
rule

<Cpx>     ::= <Sel> | <Sel> and  <Cpx>
complex

<Sel>     ::=    | Atr <op>  <Values>
selector

<Values> ::= Val | Val or  <Values>
values

<Class>  ::=  1|2|3...|J
class

<op>      ::= < | = |  >
operators
```

Atr corresponds to the name of the attribute and Val is a categorical or numerical value.

This syntax is transformable into DNF and is similar to the syntax of most rule-based systems or expert systems [Waterman, Hayes-Roth]. Note that the actual syntax is slightly more complicated [Gams].

4. DOMAIN DESCRIPTION

We performed practical measurements on two real-world domains. Data were obtained by I. Kononenko and represent descriptions and diagnoses of patients from the Oncological Institute Ljubljana. The only correction was replacement of missing values by the most probable ones for a given class. More detailed description is in [Gams], here we present only cumulative data about these domains:

Domain 1

number of attributes	18
no. of possible values per attribute	2 - 8 (average 3.3)
number of classes	9
total number of examples	150

distribution of examples amongst classes
number of examples in C1 to C9

1	2	3	4	5	6	7	8	9
2	1	12	8	69	53	1	4	0

importance of attributes - A1 to A18 : none of them is redundant

Domain 2

number of attributes	17
no. of possible values per attribute	2 - 3 (average 2.2)
number of classes	22
total number of examples	339

distribution of examples amongst classes
number of examples in C1 to C22

1	2	3	4	5	6	7	8	9	10	11
84	20	9	14	39	1	14	6	0	2	28

12	13	14	15	16	17	18	19	20	21	22
16	7	24	2	1	10	29	6	2	1	24

importance of attributes - A1 to A17 (counting how many examples overlap when omitting the i-th attribute)

1	2	3	4	5	6	7	8	9
80	80	58	85	60	53	65	55	68

10	11	12	13	14	15	16	17
63	55	60	53	54	57	65	65

5. GINESYS

5.1. ALGORITHM DESCRIPTION

The top level description of GINESYS (Generic INductive Expert SYstem Shell) is as follows:

```
repeat
  initialize Rule;
  generate Rule;
  add Rule to d(x);
  L = L - {examples covered by Rule}
until satisfiable(d(x))
```

In this general view GINESYS represents a prototype of a unifying algorithm for empirical learning covering many other systems. In a slightly more specified description we obtain the following algorithm:

```
repeat
  generalize Rule;
  repeat
    specialize Rule
  until stop(Rule);
  postprocess(Rule);
  add Rule to d(x);
  L = L - {examples covered by Rule}
until satisfiable(d(x))
```

The main difference between other EL systems and GINESYS is in "confirmation rules". Basic idea of confirmation rules is using several sources of information for classification. That seems to be common practice in every day life. For example when we try to predict the weather, we look at the official weather report, but also look at the sky and ask our neighbour. The implementation of this idea in GINESYS is that instead of using only one rule for classification several rules confirm or confute the first one. In case of a confrontation between these rules the Bayesian classifier is consulted as a method of a conflict resolution [Waterman, Hayes-Roth]. One confirmation rule in our simple example in Table 1 could be : Error is reported, when switches 3, 5 and 8 are on. This rule could be redundant or even wrong, but on the other hand it could be the only correct one! From examples in Table 1 it is not clear which of these possibilities is the right one, so both (and other) rules are stored and consulted. In more detailed tests [Gams] it was shown that this method of consulting several rules (= using different kinds of information) significantly improved classification accuracy.

5.2. COMPARATIVE RESULTS

A detailed comparison was made with other well known EL systems in two noisy medical domains (oncology). Table 2 shows results in classification accuracy.

	domain 1	domain 2
GINESYS	69.9	51.9
BAYES	68.4	50.1
other systems	67.3	48.7

Table 2. Classification accuracy measured as the percentage of correct guesses.

While GINESYS achieved best results and Bayesian classifier the second ones, none of the compared systems [Gams] outperformed results in the last row in Table 2. These results are actually an average over ten runs on randomly chosen 70% of data for learning and remaining 30% of data for testing. In further tests (t-tests, [Gams]) it was shown that the number of tests, distribution and the difference between classification accuracies was sufficient to ensure that differences are a result of some deeper cause (e.g. better algorithm) and not a chance choice.

Other measurements proved superiority not only from the point of classification accuracy, but also in generality, complexity of classification rule and explanation [Gams]. GINESYS and other algorithms discussed in this paper were implemented in Pascal on VAX 11/750.

5.3. IRREPROACHABILITY OF MEASUREMENTS

We argue that our measurements are irreproachable (unbiased) since:

- all systems were measured on exactly the same data
- no "cleaning" of data was performed
- no special form of data was allowed
- no unusual method of measuring classification accuracy was used
- no domain dependent parameters were allowed
- the number of data and tests was sufficient (t-tests) to avoid chance choice
- results were strictly checked and verified by many supervisors from the program source level to the level of classification trace.

5.4. DISCUSSION ABOUT RESULTS

Some of the systems for empirical learning achieved good results in practical testing in several real life domains, practically approaching or even outperforming domain experts and statistical methods [Kononenko; Michalski, Chilausky; Breiman et al; Gams]). More acceptable is the opinion [Breiman et al], that although all methods are more or less domain dependent, EL systems in general achieve about the same classification accuracy as other statistical methods. In our measurements some of the EL systems, especially those without special mechanisms for noisy domains, gave unexpectedly poor performance compared to the results published by the originators of algorithms. Since our implementations of those systems were the same as published, several possible explanations remain. It might be that actual implementations use some unpublished extra features, maybe the domains used for testing were especially suitable for specific algorithms etc.

The authors of this article also find questionable comparing between the system and the expert, since we regard EL systems mainly as a helping tool and not as a stand-alone program. The other reason is that fair comparison between machine and human is extremely difficult. The correct comparison should be (system + user) : user.

In most complex realistic domains mechanisms for dealing with noise are of greatest importance as independently discovered in [Breiman et al; Kononenko] and it is not realistic to achieve even tolerable results without them [Kononenko; Gams].

6. BAYESIAN CLASSIFIER

6.1. THEORETICAL FOUNDATIONS

The concept of the Bayes rule is one of the most important concepts in the field of classification and also learning. For the data drawn from a probability distribution P(A,j), the most accurate rule can be given in the terms of P(A,J) and this rule is called the Bayes rule. It is normally denoted by dB(x).

Precisely, suppose that (x,y), $x \in X$, $y \in Y$ is a random sample from the probability distribution $P(A,j)$ on $X \times C$, i. e., $P(x \in A, y=j)=P(A,j)$. Then we define dB(x) as the Bayes rule if for any other classifier d(x),

$$P(d_B(x) \neq c(x)) < P(d(x) \neq c(x))$$

Let us assume that X is N-dimensional euclidean space and for every j, $j=1, \ldots ,J$, $P(A/j)$ has the probability density fj(x) and for sets $A \subset X$

$$P(A/j) = \int_A f_j(x)dx$$

Then we can prove the following theorem:

$$d_B(x) = (j \,;\, f_j(x)P_e(j) = \max_i f_i(x)P\,(i))$$

where J is the number of classes and P(j) is the prior probability of the class j. The proof can be found in [Breiman et al].

In practice, neither the P(j) nor the fj(x) are known. The three most common classification procedures, used to approximate the Bayes rule by using the learning sample data, are discriminant analysis, kernel density estimation and k-th nearest neighbour. Accuracy of two of them have been compared with the results of Ginesys on both domains.

6.2. PRACTICAL IMPLEMENTATIONS

The k-th nearest neighbour method [Fix, Hodges] was implemented as simple as possible. The algorithm searches through the set of learning examples and determinates distance between learning and test example as the number of mismatches in their attribute values. Test example is then classified by its first nearest neighbour, and if there are more equally distant neighbour, the last one found is picked for classification. It is so called Nearest-neighbour classifier [Batcheles 1974]. Although it is so primitive, this method classifies test examples with 72.9% average accuracy in the domain 1, what is even better than GINESYS. However, in the domain 2, which is far more complex, the classification accuracy is only 40.4% what is considerably lower than that of the other methods.

The following approximation of the Bayes rule [Clark, Niblett; Kononenko] is one of the most commonly used. In general, the rule is formed as

$$P(c/A) = P_a(c)\frac{P(A/c)}{P(A)} = P_a(c)\frac{P(\bigwedge A_i/c)}{P(\bigwedge A_i)}$$

At this point the assumption is made, that all attributes are independent :

$$P(c/A) = P_a(c)\frac{\prod_i P(A_i/c)}{\prod_i P(A_i)} = P_a(c)\frac{\prod_i P(A_i/c)}{\prod_i P_a(A_i)} \qquad (1)$$

When classifying a new example we need to evaluate formula (1). One practical solution when dealing with categorical values is to store all factors into a 3-dimensional table TB[i,j,k] with the following indexes :

i - attribute index
j - attribute's value's index
k - class index

TB[i,j,k] is the number of examples in the learning set with the properties, denoted by index values.

When evaluating formula (1) during classification of a new example, one of the factors can be 0. The result can be either 0 or undefined. The solution in the second case is obvious - delete this attribute from the formula. The same solution is sometimes used when the result is 0.

In Table 2 GINESYS (without domain dependent parameters or other adjustments [Gams]) achieved higher classification accuracy than the practical implementation of theoretically optimal Bayesian classifier. The reason for this must be in practical implementation, especially in

a) approximation of probabilities from the learning set,
b) assumption, that attributes are independent,
c) practical solutions to numerical problems.

Problem (a) can be discarded, since all systems [Gams] processed exactly the same data. But it could be the case, that different classifiers (also different implementations of Bayesian classifier) are more and other less sensible to the number and distribution of input data.

6.3. A PRACTICAL EXAMPLE

Problems appearing during the evaluation of formula (1) can be shown in a simple example. Let us try to classify examples e1 and e2 from data obtained from Table 1.

e1 = 0 0 0 0 0 0 0 0
e2 = 1 0 1 1 1 0 1 1

$$P(OK/e1) = \frac{2}{5}\cdot\frac{\frac{0}{2}\frac{2}{2}\frac{0}{2}\frac{0}{2}\frac{1}{2}\frac{1}{2}\frac{0}{2}\frac{1}{2}}{\frac{2}{5}\frac{1}{5}\frac{0}{5}\frac{2}{5}\frac{1}{5}\frac{3}{5}\frac{2}{5}\frac{1}{5}} = ?(2.6)$$

$$P(ERROR/e1) = \frac{3}{5}\cdot\frac{\frac{2}{3}\frac{2}{3}\frac{0}{3}\frac{2}{3}\frac{0}{3}\frac{2}{3}\frac{2}{3}\frac{0}{3}}{\frac{1}{5}\frac{1}{5}\frac{0}{5}\frac{2}{5}\frac{1}{5}\frac{3}{5}\frac{2}{5}\frac{1}{5}} = ?$$

$$P(ERROR/e2) = \frac{3}{5}\cdot\frac{\frac{1}{3}\frac{2}{3}\frac{3}{3}\frac{1}{3}\frac{3}{3}\frac{2}{3}\frac{1}{3}\frac{3}{3}}{\frac{3}{5}\frac{4}{5}\frac{5}{5}\frac{3}{5}\frac{4}{5}\frac{3}{5}\frac{3}{5}\frac{4}{5}} = 0.149$$

$$P(OK/e2) = \frac{2}{5}\cdot\frac{\frac{2}{2}\frac{2}{2}\frac{2}{2}\frac{2}{2}\frac{1}{2}\frac{1}{2}\frac{2}{2}\frac{1}{2}}{\frac{3}{5}\frac{4}{5}\frac{5}{5}\frac{3}{5}\frac{4}{5}\frac{3}{5}\frac{3}{5}\frac{4}{5}} = 0.754$$

None of the two examples gives the sum of all classes equal to 1. Even if we delete all columns having 0 we obtain results like P(OK/e1) = 2.6. Furthermore, in case of example e2 the probability for class OK is greater than for the class ERROR, although example e2 is the same as example e5 from Table 1, belonging to class ERROR. However note that the nearest neighbour method would classify correctly in this case.

A small number of examples is insufficient for most statistical methods and also for Bayesian classifier. In real measurements in domain 1 and 2 the number of examples was always greater than one hundred and was considered sufficient. Nevertheless these counterexamples show that better classification accuracy is possible.

7. ADJUSTING PARAMETERS IN BAYESIAN CLASSIFIER

Probabilities used in evaluating formula (1) are approximated by prior probabilities in the learning set, what yields some error in classification. The formula is then evaluated as follows

$$P(c/A) = P_a(c) * \frac{\prod_i \frac{a_i}{b}}{\prod_i \frac{c_i}{d}} \qquad (2)$$

where

- ai is the number of examples of the class c with the same value of the i-th attribute as the test example,
- b is the number of examples of the class c,
- ci is the number of all examples with the same value of the i-th attribute as the test example, and
- d is the number of all learning examples.

When dealing with noisy data, errors may occur during evaluation of formula (2). Two methods have been used to avoid this errors.

7.1. OMITTING THE UNRELIABLE FACTORS

The main idea of this method is that the accuracy of estimations in formula (2) grows with the number of examples. Therefore, if b or d (only b in practice) is smaller than parameter MINN, which is set before evaluation, then the factor with this b is omitted during evaluation of formula (2) and the class probability is estimated by its prior probability. Table 3 shows the results in classification accuracy with different values of MINN.

MINN :	domain 1	domain 2
0	68.4 (0.0)	50.1 (0.0)
1	68.4 (11.1)	50.1 (4.5)
2	68.4 (47.8)	49.7 (31.8)
3	69.3 (52.2)	49.7 (32.3)
4	68.7 (56.7)	49.8 (33.6)
6	68.0 (60.0)	50.9 (47.7)
10	66.7 (75.6)	48.7 (60.5)
15	67.1 (77.8)	46.2 (72.7)

Table 3. Classification accuracy measured as the percentage of correct guesses. Values in brackets are percentages of classifications with prior probability.

It is interesting that accuracy is almost independent of the number of classifications with prior probabilities and it decreases only if we classify approximately 75% of examples this way. Yet we can see that accuracy on both domains reaches its maximum when approximately 50% of classifications are done by the prior probability of classes and this maximal accuracy is near the accuracy of GINESYS.

7.2. ADJUSTING ZERO FACTORS

A problem occurs, what to do when ai in the formula (2) is 0. One solution is to omit this factor from the formula. Another idea could be to set ai to a very small number EPS in such case. The idea is that this zero is the result of domain noise and, with more learning examples, we would sooner or later find such example and therefore we made almost no mistake and we also don't lose the information contained in the distribution of other attributes. The results of this method are shown in Table 4.

EPS :	domain 1	domain 2
0.00	68.4	50.1
0.01	68.7	52.8
0.05	71.1	52.9
0.10	72.2	51.9
0.50	24.2	27.7
1.00	2.2	7.5

Table 4. Classification accuracy achieved by adjusting zero factors in formula (2) by some small value EPS.

In both cases this method achieves accuracy higher than GINESYS. But, rapid drop of accuracy also shows that this method is very sensible to the value of EPS. Whenever we set zero factor to some value different from 0 we introduce an error into the evaluation of formula (2) and if the value of EPS is too big the results are no more reliable at all.

7.3. COMBINATION OF BOTH METHODS

In this case, both methods described in 7.1. and 7.2. are used together. First we look whether b and d are bigger than MINN and than we set zero factors in formula (2) to EPS where needed. The results of measurements on both domains are shown in Table 5 and Table 6.

\ EPS : MINN :	0.00	0.01	0.05	0.10
0	0.0	0.3	2.7	3.8
2	0.0	1.4	1.1	1.1
4	0.3	1.2	0.9	0.3
6	−0.4	0.0	−0.4	−0.4

Table 5. Increase of classification accuracy by combination of both methods in domain 1 (basic accuracy 68.4%).

\ EPS : MINN :	0.00	0.01	0.05	0.10
0	0.0	2.7	2.8	1.8
2	−0.4	1.9	2.2	2.0
4	−0.3	1.8	2.1	2.0
6	0.8	2.7	3.2	3.2

Table 6. Increase of classification accuracy by combination of both methods in domain 2 (basic accuracy 50.1%).

The accuracy of GINESYS is in both cases exceeded by more than 1% what is also the difference between basic Bayesian classifier and GINESYS. Yet there is a problem which combination of EPS and MINN values to choose and how far this decision is domain independent. Therefore, GINESYS can still be considered to reach the practical upper bound of classification accuracy.

8. EXTERNAL RULE DIRECTED BAYESIAN CLASSIFIER

Bayesian classifier itself does not derive any explicit rules and therefore rules generated by some other system (in our case GINESYS) can be used to control the evaluation of formula (1). Two such methods have been tested. The idea of the first one is that any (more or less successful) rule denotes a complex of attributes which are logically connected and therefore a deviation from the optimal Bayesian classifier is somewhat corrected. The second one is an attempt to cross GINESYS and Bayes together to yield better results.

8.1. CLASSIFICATION WITH IMPORTANT ATTRIBUTES

This method uses rules, generated by GINESYS. During evaluation of formula (1) the rule which matches the current example is searched for. If it is found, we calculate adequate probabilities by searching through the table for entire complex and not by decomposing the attribute complex to basic attributes. On the other hand, if the matching rule is not found, undisturbed evaluation of formula (1) follows. The results are shown in Table 7.

Measurements show that introducing of externally generated rules into Bayesian classifier only slightly disturbs its classification accuracy.

8.2. CLASSIFICATION WITH IMPORTANT ATTRIBUTES ONLY

The main idea of this method is to use important attribute complexes in classification if possible. For each example classified we first search for the matching GINESYS rule. If such one is found, it is used for classification. If not (when only Null rule of GINESYS is found), the classification is carried out by formula (1). The results are shown in Table 7.

	domain 1	domain 2
Bayes	68.4	50.1
Rule Directed Bayes	68.4	49.2
Important Attr. & Bayes	69.3	47.9

Table 7. Classification accuracy measured as the percentage of correct guesses.

9. COMPARISON BETWEEN EL SYSTEMS AND STATISTICAL CLASSIFIERS

Let us summarize conclusions from previous paragraphs:

- It is possible to further improve classification accuracy of implementations of Bayesian classifier, even to overpass best results of GINESYS.
- Among compared methods without domain-independent parameters GINESYS performs best and is very close to the practical limit in classification accuracies in measured domains.

Statistical classifiers are basically unable to perform explanation during classification and to build a human understandable knowledge base. Besides these, other disadvantages can be pointed out [Breiman et al; Gams]:

- they can not deal with domains with small number of learning examples;
- it is difficult to deal with unusual situations (deleting by 0, unknown values, ...);
- their results vary according to the suitability of problem domain.

It is only fair to notice that more advanced statistical methods eliminate some of these disadvantages. However these properties remain basically unchanged.

Another reason for so good results of EL systems like GINESYS compared to statistical methods is shown in Figure 1. Real-life complex domains probably contain logical laws which cover greater areas regardless of noise in given examples. On the contrary statistical methods depend on variations of probability distribution. In Figure 1 the correct probability distribution for classes 1 and 2 is presented by bold lines. Dotted lines represent probability distribution, obtained from given examples. Because of the fact that probability distribution is more sensible to chance choice it is possible that dotted lines 1 and 2 overlap causing incorrect classification.

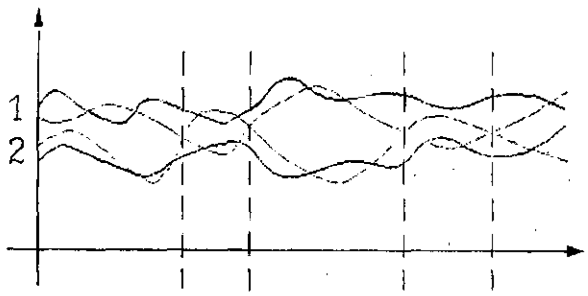

Figure 1: A graphical representation of one of the possible reasons why GINESYS performs so well compared to statistical methods.

10. CONCLUSION AND DISCUSSION

Older systems for empirical learning (EL) outperformed the statistical methods from the point of explanation during classification and possibility of building a human understandable knowledge base. While it was in some cases reported that older EL systems outperformed statistical methods as well as domain experts this opinion is not undoubtedly shared with the authors of this article. More acceptable is the conclusion [Breiman et al], that the best EL systems achieve about the same classification accuracy as statistical methods. Nevertheless it seems that the new breed of EL systems with GINESYS as one of the most promising representatives outperforms statistical methods even in classification accuracy (at least in so far measured domains).

ACKNOWLEDGES

We are grateful for suggestions to prof. Ivan Bratko. Marjan Petkovsek provided mathematical background for our analysis. Students Izidor Jerebic, Borut Znidar, Aram Karalic and Darko Zupanic were of great help in programming tasks. Igor Kononenko and the Oncological Institute Ljubljana enabled us to do the measurements on real-world domains. This work was partly supported by the Slovene Research Council and COST-13. Research facilities were provided by the "Jozef Stefan" Institute.

REFERENCES

Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984): "Classification and Regression Trees", Wardsworth Int. Group.

Carbonell J.G., Michalski R.S., Mitchell T.M. (1983): "An Overview of Machine Learning", in Michalski R.S., Carbonell J.G., Mitchell T.M. (ed.), Machine Learning: an Artificial Intelligence Approach, Tioga Publishing.

Cestnik B., Kononenko I., Bratko I. (1987): "ASSISTANT 86: A Knowledge Elecitation Tool For Sophisticated Users", Progress in Machine Learning, Sigma Press.

Clark P., Niblett P. (1987): "Induction in Noisy Domains", Progress in Machine Learning, SIGMA Press.

Kononenko I. (1985): "Razvoj sistema za induktivno učenje ASISTENT", magistrsko delo, Fakulteta za eletrotehniko, Ljubljana.

Kononenko I. (1985): "Strukturno avtomatsko ucenje", Informatica 3, str. 44 - 56.

Gams M. (1987): "Principi poenostavljanja v sistemih za avtomatsko ucenje", doktorska disertacija, Ljubljana.

Gams M., Lavrač N. (1987): "Review of Five Empirical Leaning Systems Within a Proposed Schemata", Progress in Machine Learning, (ed. Bratko I., Lavrac N.), Sigma Press.

Lavrač N., Varšek A., Gams M., Kononenko I., Bratko I. (1986): "Automatic construction of the knowledge base for a steel classification expert system", The 6th International Workshop on Expert Systems, Avignon.

Michalski R.S. (1987): "Machine Learning", Tutorial 6, IJCAI 1987, Milano.

Michalski R.S., Chilausky L.R. (1980): "Learning by Being Told and Learning from Examples: an Experimental Comparison for Soybean Disease Diagnosis", Policy Analysis and Information Systems, Vol. 4, No 2.

Quinlan J.R. (1986): "Induction of Decision Trees", AI Summer Seminar, Dubrovnik.

Waterman D.A., Hayes-Roth F. (ed.) (1977): "Pattern-Directed Inference Systems", Academic Press.

Winston H.P. (1984): "Artificial Intelligence", Addison-Wesley.