Benchmarking Indicates Relevance of Multiple Knowledge

Matjaž Gams
Jožef Stefan Institute, Jamova 39, 61000 Ljubljana, Slovenia
Phone: +386 61 125 91 99, Fax: +386 61 219 677
E-mail: matjaz.gams@ijs.si

Keywords: artificial intelligence, multiple knowledge, multistrategy learning

Edited by: Anton P. Železnikar
Received: March 17, 1994 **Revised:** November 11, 1994 **Accepted:** November 15, 1994

Over the last 7 years, detailed measurements of available learning systems were performed on two real-life medical domains with the purpose to verify the importance of multiple knowledge. The performance of the combined system GINESYS, consisting of an artificial intelligence and a statistical method, was analysed with and without multiple knowledge and by varying the number of learning examples, the amount of artificially added noise, the impurity and the error estimate functions. These measurements and those of other researchers indicate that multiple knowledge can provide essential improvements. Measurements also indicate that improvements over "one-level" or monostrategy knowledge representation representations are quite common in real-life noisy and incomplete domains.

1 Introduction

In easing the bottleneck of knowledge acquisition in expert systems (Harmon et al. 1988), automatic knowledge construction from examples has proven useful in many practical tasks. Quite often, examples are described in terms of attributes and their values and each example belongs to a certain class. The task of the system is to induce concept descriptions from examples. First systems were designed for exact domains like chess end-games and constructed trees (ID3 - Quinlan 1983) or lists of rules (AQ11 - Michalski & Larson 1983). But in many real-life domains (Gams & Karalič 1989), because of noise or incomplete description (Manago & Kodratoff 1987) specialised mechanisms have to be applied. In noisy domains, longer rules (or longer branches in trees) perform better on learning examples while truncated rules (pruned tees with shorter branches) perform better on unseen examples. On the basis of this principle, the second group of inductive systems emerged (CART - Breiman et al. 1984; AQ15 - Michalski et al. 1986; ASSISTANT - Kononenko 1985; CN2 - Clark & Niblett 1989; C4 - Quinlan 1987). Around five years ago the third group of systems began emerging (GINESYS - Gams 1988; 1989; LOGART - Cestnik, Bratko 1988; new CN2 - Clark & Boswell 1991), based on the explicit use of multiple knowledge.[1] Each of these groups of systems usually achieves better performance than previous. Better performance of multiple knowledge systems was especially noticeable in classification accuracy, also in better comprehensibility (although more difficult to measure) when compared to the other two groups. At the same time, their efficiency remained similar to those in the second group.

With measurements presented in this paper we give additional arguments for successfulness of multiple knowledge by explicitly measuring the influence of the number of learning examples and the influence of noise, as well as the influence of the error estimate and impurity functions. Benchmarking was performed on two often used domains - lymphography and primary tumor (Clark & Niblett 1989; Michalski et al. 1986; Cestnik & Bratko 1988; Gams 1988).

[1] By 'multiple knowledge' we refer to multiple models, multiple systems or multiple methods.

Here we present results of benchmarking over a period of 7 years. Testing was performed always on the same two oncological domains. Altogether, around 20 systems were benchmarked. Our system GINESYS was constructed on the bases of first benchmarking of around 10 systems in 1987 from a frustration since statistical systems have regularly achieved better accuracy than single AI systems. GINESYS is described in Section 3, benchmarking in Sections 4 and 5.

2 Multiple Knowledge and Multistrategy Learning

Even first expert systems like MYCIN (Shortliffe 1976) and most rule-based systems already enabled a certain amount of multiplicity, i.e. redundancy, since rules can be more or less multiple. Newer systems like CN2 (Clark & Niblett 1989) or C4 (Quinlan 1987) contain similar amount of redundancy which is probably one of the reasons for their successful behaviour in noisy domains. In (Catlett & Jack 1987) it was reported that constructing a separate decision tree for each class with the same method as when constructing one decision tree for all classes significantly increased accuracy. Similar conclusion was derived by Clark & Boswell (1991) when constructing several lists of rules and by Buntine (1989) when combining 5 decision trees with different roots.

In communications, the positive effect of using redundant bits is known for decades and even simple ID numbers in banking have additional digits in order to improve the robustness of the whole system. Theoretical aspects of redundancy in such cases are described e.g. in (Shannon & Weaver 1964).

In most every-day activities, people use multiple knowledge whenever there is any possibility of biasing (Utgoff 1986). For example, when hiring a new employee, one checks several reports which are basically multiple (e.g. biography, recommendations etc.). When bringing an important decision, humans often discuss possibilities in groups of relevant people. A council of physicians is consulted when dealing with difficult or important cases. One physician suffices for most of normal activities since one is substantially cheaper than a group of them.

It is commonly accepted that cross-checking of several knowledge sources is generally better than using one source of knowledge alone. Humans are intrinsically multiple. They apply multiple strategies in every-day activities without paying much attention to that phenomenom. Therefore, machine and human multistrategy learning have natural interrelationship and potential benefits in both directions.

In recent years there were several distinguished events related to mutistrategy learning. Among them: a book edited by R. Michalski & G. Tecuci: *Machine Learning, A Multistrategy Approach*, Vol. IV, Morgan Kaufmann (1994), specialised international workshops on multistrategy learning organised by George Mason University, special issue of Informatica (Tecuci 1993), and IJCAI-93 workshop on integration of machine learning and knowledge acquisition (Tecuci, Kedar & Kodratoff 1993).

3 GINESYS

GINESYS (Generic INductive Expert SYstem Shell) is one of the oldest systems actively utilising multiple knowledge representations (Gams 1988). It consists of two systems (i.e. methods), one from AI and one from statistics. There were sensible reasons for combining methods from different fields. First of all, artificial intelligence methods enable construction of knowledge bases which are typically very transparent and understandable; therefore, it was hoped that a combination would still be more understandable than statistical knowledge bases. A statistical method was chosen on the basis of the hypothesis that knowledge representations should be as different as possible.

GINESYS utilises two different strategies on the basis of these two systems: the AI system constructs and consults lists of rules, and the statistical system multiplies probabilities according to the distribution of classes corresponding to each attribute of the tested example. Both single systems already implicitly utilise multiple knowledge – the AI part through a couple (typically 5) of rules attached to the main one which are triggered when classifying, and the statistical part through combining probabilities relating to the value of each attribute of the tested example.

The AI part of GINESYS is named INESYS

```
d := (()); (*d is initialised*)
repeat
  Star := (NP); BestRules := (NP);
  repeat
    for all Rule_j from Star generate all specialisations
    NewRule_i that do not fulfil the stopping criteria;
    Star := ();
    put into Star at best MAXSTAR the best NewRule_i
    evaluated by user defined impurity function;
    from rules in BestRules and significant NewRule_i choose
    the best MAXBEST rules, evaluated by user defined
    error estimate function, and put them into BestRules
  until Star is empty;
  add BestRules into d;
  L := L - examples, covered by the best evaluated rule from
  BestRules
until L is empty
```

The INESYS algorithm

(see top of the page). It reimplements many of mechanisms of the ID3 and AQ (CN2) family of algorithms. It was primarily designed as an attempt to fully simulate the family of ID3 and AQ inductive empirical learning systems (Gams & Lavrač 1987). Theoretically, it simulates N^M different algorithms where M is the number of modules of the algorithm and N is the number of variations of each module (Gams 1989). The actual number of different variations of GINESYS can be estimated to several hundreds.

INESYS constructs rules with a beam search over all possible combinations of attributes. In addition, it utilises several search-guiding and error-estimate functions such as informativity, the Gini index, Laplacean error estimate and significance. Due to elaborate mechanisms for noise handling, INESYS typically constructs a small number of short rules, i.e. with a small number of attributes. For example, on average, 5.1 main rules with 1.4 attributes in a rule were constructed in lymphography. In primary tumor, there were 11.0 main rules with 2.3 attributes in a rule. Therefore, a typical rule had the form:

if $(A_i = v_{ij})\&(A_k = v_{kl})$ *then* $Distribution_n$
where

- $(A_i = v_{ij})$ is a Boolean test whether attribute i has value j, and
- $Distribution_n$ is a class probability distribution corresponding to the condition part of the rule, i.e. a complex.

A general description of INESYS is:

```
repeat
  construct Rule(s);
  add Rule(s) to d;
  L := L - ExamplesCoveredByRule
until satisfiable d
```

where L is the set of learning examples, d is constructed knowledge in the form of trees or lists of rules and Rule(s) is one or many branches in a tree or one or many rules in the list of rules. A procedural description of INESYS is presented at the top of the page where d is the constructed knowledge in the form of an ordered list of ordered lists of rules, Star and BestRules are ordered lists of rules and L is the set of learning examples. NP is a rule with an uninstantiated complex and class probability distribution of L.

In INESYS, the main improvement regarding existing rule-based systems are rules attached to the main rule. The aim of these multiple rules is twofold. First, to give the user more rules and thus more opportunities to analyse the laws of the domain. Second, to improve classification accuracy by cross-checking the matched rule with con-

firmation rules. This mechanism already enables the use of multiple knowledge to a certain degree:

```
if Complex1 then Class1
   (Complex11 then Class11
        ........
    Complex1R then Class1R)
   else if Complex2 then Class2
          (Complex21 then Class21
               ........
           Complex2R then Class2R)
```

Classification in INESYS starts by sequentially checking main rules. When the first main rule matches a new example, corresponding multiple rules that match the new example add their class probability distribution according to the formula for the union of independent events

$$p_{12} = p_1 + (1 - p_1) \times p_2.$$

Probabilities are multiplied by error estimates in order to calibrate the effect of rules with different credibility, and finally normalised. There are two threshold parameters that present a heuristic estimate of the goodness of classification by a rule: the smallest necessary percentage of the majority class (MINACC) and the smallest difference between the percentage of the majority class and the second to majority class (MINDIFF). Each constructed rule in GINESYS has to satisfy both conditions. Parameter MINDIFF additionally affects the classification process in the sense that the class probability distribution of a combined main and confirmation rules must satisfy it.

The second method in GINESYS is the approximation of the Bayesian classifier which assumes independence of attributes. It is often referred to as "naive Bayes"(Good 1950), in this paper also "Bayes". Naive Bayes constructs all possible rules with only one attribute in the complex. Therefore, the form of these rules is:

if $(A_i = v_{ij})$ *then* $Distribution_n$.

The classification schema is as follows: all rules, that match a new example, are taken in consideration. The probability of each class c is computed by the following formula:

$$P(c|A) = P_a(c) \times (P(A_1|c)/P_a(A_1)) \times \ldots$$
$$\times (P(A_v|c)/P_a(A_v)) \quad (Eq.1)$$

where $P(c|A)$ denotes probability of class c given attributes and values A of the tested example, $P_a(c)$ denotes the a priori probability of class c, $P(A_i|c)$ the probability that attribute A_i has the same value as the classified example regarding the class c, $P_a(A_i)$ the same as before, but regardless of class, and v is the number of attributes. By calculating probabilities of all classes by (Eq.1.), a class probability is obtained. Therefore, although naive Bayes constructs rules similar to INESYS, in the process of classification all attributes are considered in Bayes and on average only around two in INESYS.

Cooperation between the AI and the statistical system is relevant only when they propose different classes. In that case, the goodness of triggered rules in INESYS is estimated by the simple heuristics mentioned above. If the goodness of combined rules exceeds the value of a given threshold (parameter MINDIFF), classification by INESYS is adopted. Otherwise, the classification by naive Bayes prevails. In other words: If class probability distribution of combined rules is estimated as unreliable, the statistical method is called as a supervisor to decide which class is estimated as the most probable.

The combining schema is based on the following reasoning: When multiple rules confirm the main ones, classification is very likely to be correct. If a significant disagreement occurs then the list of rules is not credible and the other method using different knowledge representation should be consulted. It was expected that short rules constructed by INESYS will be more successful when they have high confidence in their prediction, and the approximation of the Bayesian classifier to be more successful when dealing with difficult cases where truncated rules capturing the main and most important laws of the domain are not predicting with great certainty.

4 Benchmarking

Since 1987, systematic measurements are being performed on two oncological domains, lymphography and primary tumor. Data were obtained from real patients from the Oncological institute Ljubljana (Kononenko 1985; Cestnik & Bratko 1988). Unknown values of attributes were re-

SYSTEM	LYMPHOGRAPHY			PRIMARY TUMOR		
	class.acc.	no.rules	no.att.	class.acc.	no.rules	no.att.
GINESYS*	70.5	5.1	7	52.2	11.0	25
GINESYS	70.5	5.1	7	52.0	11.0	25
BAYES	68.6	56.0	56	50.1	37.0	37
CN2-new1	68.7			50.3		
GB*	67.4	5.1	7	47.6	11.0	25
CN2-new1'	65.6			46.9		
NEAREST NEIG.	72.9			40.4		
C4.5-rules	64.7			38.2		
C4.5-trees-u	63.1			48.9		
C4.5-trees-p	66.7			48.8		
CN2-like1	67.3	4.8	8	48.7	11.4	27
CN2-like1'	66.1	5.0	6	45.6	10.8	22
ID3-like	61.8	25.0	110	48.7	28.6	129
CN2-like2	66.8	10.8	21	45.7	19.3	70
CN2-like2'	65.0	9.4	16	46.2	19.4	68
AQ-like1	60.6	7.0	80	48.8	16.0	423
AQ-like2	55.2	7.0	80	32.0	16.0	423

Table 1: Benchmarking systems on two oncological domains.

placed by the most common values regarding the class.

4.1 Domain Description

Basic statistics of the whole set of data are:

LYMPHOGRAPHY
18 attributes
2 - 8 (average 3.3) values per attribute
9 classes
150 examples
distribution: 2 1 12 8 69 53 1 4 0
all examples differ even if one attribute is deleted

PRIMARY TUMOR
17 attributes
2 - 3 (average 2.2) values per attribute
22 classes
339 examples
distribution: 84 20 9 14 39 1 14 6 0 2 28 16 7 24 2 1 10 29 6 2 1 24
75 examples in the data set have another example with the same values of attributes and different class; if we delete one attribute, this number is: 114 111 81 122 84 75 93 79 97 91 77 83 76 77 79 94 94

4.2 Benchmarked Systems

On the benchmark domains, around 20 AI and statistical systems were compared over more than half of a decade. All the systems were given the same set of 10 random distributions of data, each time taking 70% of data for learning and 30% of data for testing. Results of relevant systems are presented in Table 1. The row in the middle of the Table divides multiple and single systems, i.e. those that use only one rule or combine many rules during one classification.

GINESYS* is a version of GINESYS using "negation" multiple rules, which try to confront the main rule if possible. BG* is GINESYS* without the statistical method, i.e. INESYS with functions B and G. First nearest neighbour algorithm classifies with the class of the nearest nei-

LYMPHOGRAPHY			PRIMARY TUMOR		
FUNCTIONS	INESYS**	GINESYS**	FUNCTIONS	INESYS**	GINESYS**
AB	68.4	69.7	BA	48.3	52.3
GB	67.4	69.9	BG	48.1	51.8
BB	66.4	70.8	GB	47.6	52.0
BG	62.6	68.4	AB	46.6	51.3
BA	62.4	68.4	BB	46.4	52.5

Table 2: Accuracy under different impurity and error estimate functions.

ghbour where distance is measured by the number of attributes with different values. BAYES is an approximation of the Bayesian classifier using an assumption that attributes are independent. ID3-like is a version of the ASSISTANT system using cross-validation pruning. CN2-like systems are different modifications of the CN2 algorithm, and CN2-new systems are latest versions. C4.5-rules constructs rules, C4.5-trees-u unpruned trees, and C4.5-trees-p pruned trees. AQ-like systems are modifications of the AQ15 systems.

Classification accuracy (column 1 in each domain in Table 1) was measured as an average percentage of correct classifications in ten test runs. The second column in each domain represents the average number of rules in a rule list or branches in the tree. The third column is a product of the number of rules (branches) times the average length of a rule (branch) times the number of internal disjunctions.

The relations between systems are similar to those observed in other measurements (Clark & Niblett 1989; Rendell et al. 1987; Rendell et al. 1988). Systems of the AQ family usually achieve lower classification accuracy than CN2 or ASSISTANT, while ASSISTANT and CN2 achieve similar classification accuracy. AQ-like1 represents an estimate of the upper possible classification accuracy of the rules, constructed by the AQ-like system. BAYES achieved better results than other systems except GINESYS. Nearest neighbour algorithm seems to be very domain dependent. GINESYS achieved the best average classification accuracy over both domains.

AQ-like systems construct more complex rules than other systems. However, the third column might be misleading for tree constructing algorithms like ID3-like because it represents tree as a list of separated branches. GINESYS* and GINESYS are measured only by the main rules and not by the multiple ones. On the other side, from the results in Table 1 it follows that systems like GINESYS and CN2 construct smaller number of shorter main rules while AQ-like systems construct more complex rules.

The efficiency of the benchmarked algorithms was also analysed. AQ systems are about an order of magnitude slower than ASSISTANT, CN2 and GINESYS, and these are about an order of magnitude slower than BAYES. Results are similar to other measurements when having in mind that our versions of CN2 and GINESYS use a data compression mechanism which speeds up the algorithm roughly five times. GINESYS PC, another version of GINESYS, runs on IBM PC computers and is available as a free scientific software.

4.3 Varying Impurity and Error Estimate Functions

In order to verify whether improvements in GINESYS were caused by multiple knowledge or by domain-dependent parameters, several parameters were varied, and functions were the first among them. GINESYS uses two different groups of functions: informativity functions and error estimate functions. Informativity functions strategically guide search by trying to determine the amount of impurity. Error estimate functions try to estimate classification error. Four functions were used in all 16 possible combinations in each domain. Classification accuracy of GI-

LYMPHOGRAPHY	% OF LEARNING EXAMPLES						
SYSTEM	20%	30%	40%	50%	60%	70%	80%
GINESYS	52.8	58.2	63.1	63.7	60.1	70.5	75.3
BAYES	52.8	59.3	60.8	61.2	58.2	68.6	72.1
INESYS	39.2	51.7	54.1	62.6	59.0	67.4	74.3
ASSISTANT	53.9	60.5	57.9	57.5	55.2	62.1	65.2
ASSIST 0	53.2	60.7	57.4	57.8	55.9	62.4	66.8

Table 3: Accuracy in lymphography at different percentages of learning data.

NESYS with (GINESYS**) and without (INESYS**) top-level multiple knowledge was compared. In Table 2 we present only the best three combinations of INESYS** in both domains. The four functions used were: I - informativity (Quinlan 1986); A - % of majority class; G - Gini index (Breiman et al. 1984); B - Laplacean error estimate (Niblett & Bratko 1986). The first letter denotes the impurity function and the second letter the error estimate function.

Measurements presented in Table 2 indicate that Laplacean error estimate is one of the most successful functions used for impurity or error estimates. Informativity is unexpectedly not present in the best three combinations. Default functions for GINESYS systems (GB) were taken in advance from the literature (Breiman et al. 1984; Niblett & Bratko 1986).

4.4 Varying Percentage of Learning Examples

Benchmarks in sections 4.2 and 4.3 were performed on 10 distributions of data each time taking 70% of data for learning and 30% of data for testing. In Table 3 and 4 we varied the percentage of learning data from 20% to 80% and used the remaining data for testing. Graphical representation of data in Table 4 is shown in Figure 1. Systems in Figure 1 are denoted as in column 1 of Table 4. ASSIST 0 is ASSISTANT without pruning and INESYS is GINESYS without the statistical method.

Probably the main reason for unproportionally low classification accuracy of INESYS with small number of learning examples are functions which work well only with several ten examples. But even then there are some cases when INESYS classifies better than BAYES. The combining mechanism usually decides well when to choose the right method. The performance of INESYS increases with the number of learning examples, and the gain of GINESYS over BAYES also proportionally increases. In lymphography, ASSISTANT prunes the tree by approximately 50% and achieves very similar classification accuracy as ASSIST 0. In primary tumor, the pruned tree constructed by ASSISTANT is roughly 4 times smaller than the tree of ASSIST 0 which besides constructing more complex trees also achieves lower classification accuracy.

The improvement of GINESYS over the best of it's two subparts was typically around 1-2% leading to a conclusion that the combining mechanism performed well when changing the number of learning examples.

4.5 Varying Additional Noise

Noise was introduced into the lymphography and primary tumor domain to attributes and classes in the learning and test examples. For example, 1% of noise means that, on average, each hundred's value of attribute and each hundred's class was randomly changed in learning and test data. Average results of 10 tests (see section 4.2) are presented in Tables 5 and 6, and in Figure 2.

When the amount of noise increases, the performance of INESYS relatively improves and achieves even better classification accuracy that GINESYS. As expected, in a very noisy situation, a small number of short rules performs the best. Si-

P. TUMOR	% OF LEARNING EXAMPLES						
SYSTEM	20%	30%	40%	50%	60%	70%	80%
GINESYS (G)	41.9	44.6	48.1	49.0	48.1	52.0	52.3
BAYES (B)	41.8	45.2	47.5	48.0	47.2	50.1	50.3
INESYS (I)	26.9	35.6	33.8	43.5	41.2	45.9	46.7
ASSISTANT (A)	39.8	43.5	43.5	45.9	44.3	47.9	49.2
ASSIST 0 (A0)	39.6	41.6	39.9	41.1	39.6	41.3	41.7

Table 4: As in Table 3, but for the primary tumor domain.

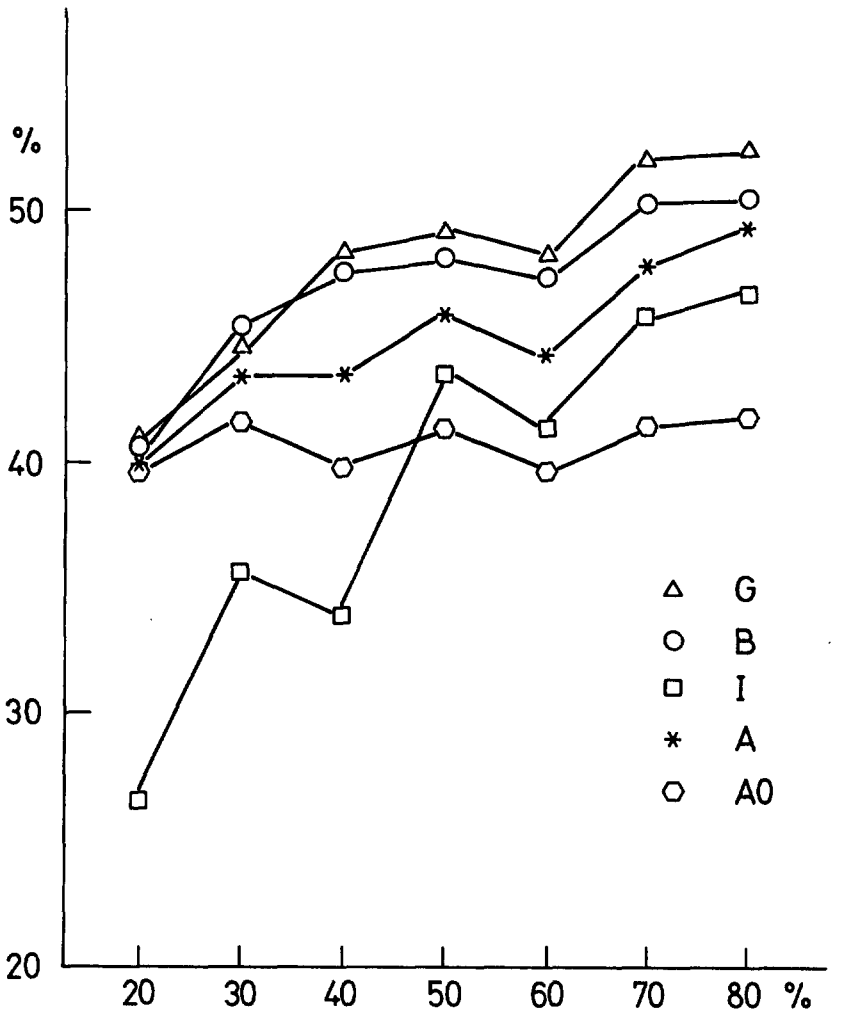

Figure 1: Graphical representation of data in Table 4. On the x-axis is the percentage of learning data and on the y-axis is classification accuracy.

LYMPHOGRAPHY	% OF ADDITIONAL NOISE						
SYSTEM	0%	1%	5%	10%	20%	35%	50%
GINESYS	70.5	65.3	63.7	53.1	43.8	28.9	21.1
BAYES	68.6	65.8	61.7	51.1	41.8	28.0	20.7
INESYS	67.4	63.4	59.1	53.0	41.4	30.3	25.4
ASSISTANT	62.1	60.2	52.8	34.1	33.3	23.4	18.4
ASSIST 0	62.4	60.5	51.8	41.6	29.9	23.5	17.6

Table 5: The influence of additional noise – lymphography.

P. TUMOR	% OF ADDITIONAL NOISE					
SYSTEM	0%	1%	5%	10%	20%	35%
GINESYS (G)	52.0	50.6	42.6	35.2	23.5	13.8
BAYES (B)	50.1	47.8	40.3	33.5	23.6	13.9
INESYS (I)	45.9	43.5	36.2	30.7	20.0	16.1
ASSISTANT (A)	47.9	44.9	39.4	30.5	16.7	8.4
ASSIST 0 (A0)	41.3	39.1	32.4	25.3	14.5	8.7

Table 6: The influence of additional noise – primary tumor.

milar effect is noticeable in the lymphography domain especially compared to ASSISTANT and is probably connected to the fact that ASSISTANT constructs a tree of several tens of leaves while INESYS constructs from 2 to 5 rules. With a growing amount of noise, the gain of GINESYS slowly decreases but remains around 2% as long as any rule of INESYS can be trusted as the meaningful one.

5 New Measurements

In further attempts to verify the obtained results presented in Section 4, GINESYS and benchmark data were around five years ago sent to over 50 laboratories and declared to be freely available for scientific purposes. The obtained answers can be clustered into two groups: several laboratories benchmarked systems on the proposed two domains, or at least approved the approach. On the other hand, there were some researchers who considered proposed benchmarking of classification accuracy as a numerical measurement belonging to statistics. In their opinion, artificial intelligence methods should be evaluated mainly at the level of ideas. Indeed, measuring only classification accuracy does not consider several important advantages of artificial intelligence, e.g., the transparency of the constructed knowledge base or the comprehensibility of classifications. However, in the last two years we have observed a constant shift in a direction which accepts such verifications as crucial in evaluating quality.

In 1990 we received the first, and so far only report of a system, NAIVE BAYES* (Cestnik 1992), which achieved better accuracy than GINESYS in both domains (54.1% in primary tumor and 70.9% in lymphography). The improvement is based on a correction of the weakness of NAIVE BAYES which happens whenever there is a gap in the data, meaning there is no example with the particular value of the attribute. Then, one factor in the product becomes 0 and the resulting product (Eq.1) becomes 0. This was already

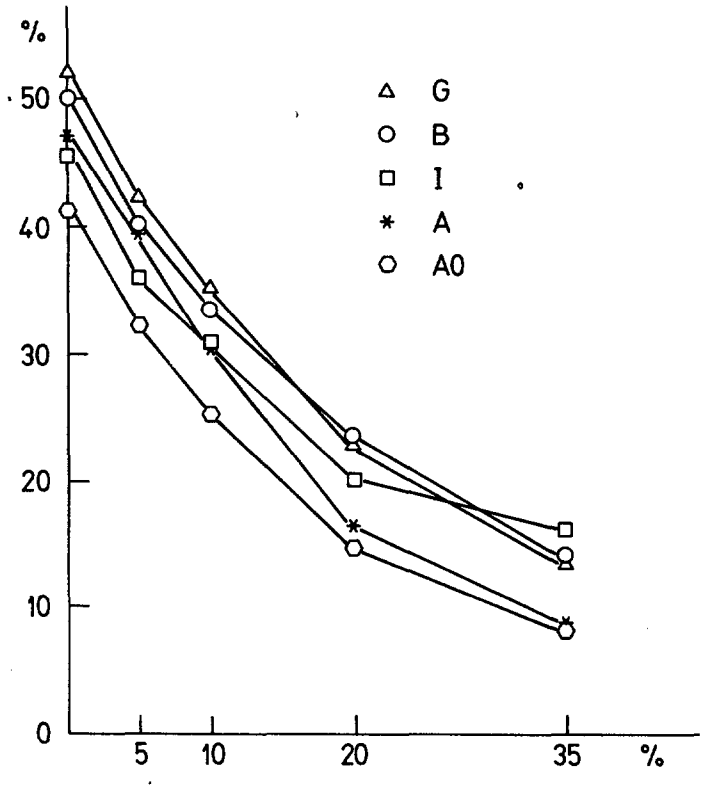

Figure 2: Graphical representation of data in Table 6. On the x-axis is the percentage of additional noise and on the y-axis is classification accuracy.

observed in (Gams & Drobnič 1988; Gams et al. 1991) where ϵ was used instead of 0. In NAIVE BAYES*, the Laplacean estimate is introduced for a correction instead of ϵ.

The reported improvements enabled additional experiments in trying to construct a multiple system, achieving even better classification accuracy. In the first attempt, NAIVE BAYES* was directly embedded into GINESYS, but the observed classification accuracy was lower than that of NAIVE BAYES*. Obviously, a smaller number of stronger rules had to be constructed since NAIVE BAYES* achieved significantly better classification accuracy than GB. Several parameters in GINESYS deal with rules, such as significance (Kalbfleish 1979), modified Laplacean error estimate (Niblett & Bratko 1986) or MINDIFF and MINACC. In the second attempt, MINDIFF was set to 0.5 instead of the previous 0.3, and MINACC to 0.7. Consequently, GINESYS90 achieved an additional 0.8% increase in primary tumor and 1.3% in lymphography over NAIVE BAYES*. Later it was found that the values of MINACC and MINDIFF belong to the set of optimal combinations, as can be observed in Tables 8 and 9.

The updated versions of NAIVE BAYES and GINESYS achieve the best two classification accuracies (compare Table 1 and Table 7). The percentage of corrections by NAIVE BAYES was 8% in lymphography and 27% in primary tumor in GINESYS and, correspondingly, 25% and 45% in GINESYS90.

New values of parameters MINDIFF and MINACC force GINESYS90 to construct a smaller number of longer rules. Also, rules are usually roughly twice more often corrected by NAIVE BAYES* than in GINESYS. To a great extent, this is due to the increased average number of classifications performed by the null or uninstantiated rule, i.e. the last rule in a rule list. This number increased from 9.2 to 15.9 in lymphography (45 classifications), and from 18.0 to 55.1 in primary tumor (102 classifications). Understandably, the last uninstantiated rule is always considered as unreliable in GINESYS and GINESYS90. But in the INESYS and INESYS90 algorithm, the classification is still performed by corresponding null-rule class distribution which is typically only slightly better than the default rule. Therefore, it is understandable that on average accuracy of

SYSTEM	LYMPHOGRAPHY			PRIMARY TUMOR		
	class.acc.	no.rules	no.att.	class.acc.	no.rules	no.att.
INESYS90	63.7	3.8	7	36.3	6.9	19
NAIVE BAYES*	70.9	56.0	56	54.1	37. 0	37
GINESYS90	72.2	59.6	63	54.9	44.3	56

Table 7: Accuracy, number of rules, of all attributes.

	LYMPHOGRAPHY									
	ACC.	.1	.2	.3	.4	.5	.6	.7	.8	.9
.9	71.0	+								
.8	72.0	♡							♡	+
.7	71.6	+								
.6	72.0	♡						♡	+	+
.5	72.2	♡				♡	♡	♡	♡	+
.4	70.2	−			−	−		−	+	
.3	70.7	−		−			−	−	+	+
.2	70.2	−								
.1	68.2	−				−			+	

Table 8: Influence of the goodness criterion, GINESYS90, lymphography.

INESYS90 decreased from 67.4% to 63.7% in lymphography and more, from 45.9% to 36.3% in primary tumor. This should not blur the fact that the effective part of INESYS90 which takes part in classifications of GINESYS90 actually achieves better classification accuracy than INESYS.

The influence of the MINDIFF and the MINACC parameters on the classification accuracy of GINESYS90 was further measured, and it was found that there is a wide range of possible combinations which enable similar improvements (see Tables 8 and 9).

The x-axis in Tables 8 and 9 corresponds to MINACC and the y-axis corresponds to MINDIFF ranging from 0.1 to 0.9. The second column of classification accuracies in each Table represents accuracy with current MINDIFF and MINACC <= MINDIFF. Each mark in Tables 8 and 9 represents one ten-runs measurement as follows (in percents):

− bellow 70.9 in lymphography, bellow 54.1 in primary tumor

\+ between 70.9 and 71.9, between 54.1 and 54.6 correspondingly and,

♡ over 71.9 (+1) in lymphography and over 54.6 (+0.5) in primary tumor.

Top-level or global multiplicity in any version of GINESYS can be estimated by the percentage of different classifications of both single systems. In Table 10, it is presented for GINESYS90 in both domains with MINDIFF = 0.3 and 0.5 (MINDIFF = MINACC) on training and testing examples.

Let us measure the internal multiplicity of each monostrategy system. INESYS90 constructs a list of sublists of rules. However, the order of rules is important and the confirmation rules are attached to the main rules. Therefore, each sublist of rules corresponds to a particular subset of train-

	PRIMARY TUMOR									
	ACC.	.1	.2	.3	.4	.5	.6	.7	.8	.9
.9	54.3	+								
.8	54.3	+							+	
.7	54.3	+							+	
.6	54.5	+							+	
.5	54.9	♡				♡	♡	♡	+	
.4	53.5	−			−	−		−	+	
.3	53.6	−		−						
.2	53.1	−								
.1	51.9	−				−		−	+	+

Table 9: As in Table 8, but primary tumor.

MINDIFF	LYMPHOGRAPHY		PRIMARY TUMOR	
	train	test	train	test
0.3	28	26	28	34
0.5	26	29	44	49

Table 10: Percentage of different classifications, i.e. top-level or global multiplicity in GINESYS90.

ing data and there seems to be no natural way to extract many knowledge bases such that each covers the whole measurement space. On the other hand, rules in both NAIVE BAYES and NAIVE BAYES* have the form
$if(A_i = v_{ij})$ *then* $Distribution_n$
and are constructed on the whole training data. Therefore, a list of rules with the same attribute and all possible values of that attribute represents one knowledge base covering the whole measurement space. The average percentage of different classifications of each such knowledge base and the combined knowledge base is presented in Table 11. It should be observed that the same single knowledge bases are used in NAIVE BAYES and NAIVE BAYES*, but they are differently combined. Whatever the case, both NAIVE BAYES and NAIVE BAYES* can be regarded as internally consisting of multiple knowledge bases. Furthermore, these knowledge subbases are quite independent of each other, although they are constructed on the same training data.

Overall, finding a reasonable combination of the two knowledge bases, i.e. GINESYS90, took only one day of work and resulted in achieving an average 1% increase in classification accuracy. The amount of efforts needed was evidently small because only already existing systems had to be modified.

6 Discussion

Multiple knowledge has proven useful in many measurements, first in (Brazdil & Torgo 1990; Buntine 1989; Catlett & Jack 1987; Cestnik & Bratko 1988; Clark & Boswell 1991; Gams 1988; 1989; Gams, Drobnič & Petkovšek 1991), and followed by tens of reports in the last couple of years. In our measurements, classification accuracy of the combined knowledge base was typically better than the accuracy of each single knowledge

SYSTEM	LYMPHOGRAPHY		PRIMARY TUMOR	
	train	test	train	test
NAIVE BAYES	48	50	72	70
NAIVE BAYES*	44	46	65	66

Table 11: Percentage of different classifications in BAYES, i.e. internal multiplicity.

base. However, due to a relatively high standard deviation the statistical significance of this improvement cannot be proved in 10 tests (Gams 1989). On the other hand, additional measurements were performed by varying parameters of GB (form and number of multiple rules, goodness of rules, factor of significance, impurity functions, error estimate functions) and domain parameters (percentage of training and testing data, percentage of additional noise). In this paper we present over 200 measurements each time averaging 10 tests. If we delete measurements with more than 20% of additional noise and those with less than 70 learning examples, we obtain 167 measurements with only 3 cases where (a version of) GINESYS has not achieved the best classification accuracy. The improvement was typically around 1%.

Therefore, the improvement in 167 measurements (each time averaging over 10 tests) is statistically highly significant. Although more intensive measurements were performed in recent years, e.g., (Brazdil et al. 1994), measurements in this paper present one of the longer-lasting efforts.

Besides better classification accuracy, improved explainability and understandability were also reported. Indeed, the informativity of the knowledge base with multiple rules seems to be much better than without them. Multiple rules can be trimmed off and a "usual" knowledge base is obtained as a downgraded version. Since a user can define the number of multiple rules, the preference function and other parameters, it enables a thorough extradition of most valuable rules. The efficiency of the learning algorithms remains practically the same when using multiple knowledge.

In conclusion, more and more indications emerge that "single-knowledge" systems in general do not achieve the performance of "multiple-knowledge" systems. Therefore, multiple knowledge is becoming regularly implemented in recent systems. The reported gains are usually substantial at small additional cost.

While research on monostrategy methods and one-level knowledge representations continues to be of great importance to the machine learning community, the interest and amount of research work in multistrategy learning and multiple knowledge representations rapidly increases over the last couple of years. Expansion is accompanied by great diversification and new approaches.

In general, multiple systems enable greater competence than monostrategy systems relying on one knowledge representation and one computing mechanism. On the other hand, multiple systems demand more understanding of capacities, limitations and cooperation between single systems. Due to the constant growth of computer power, speed and memory requirements have to a great extend diminished, thus bringing the focus to essential research and engineering questions.

ACKNOWLEDGMENTS

This work was supported by the Ministry of Science, Research and Technology, Republic of Slovenia and was carried out as a part of European Project ESPRIT II Basic Research Action number 3095, Project ECOLES. Research facilities were provided by the "Jozef Stefan" Institute. Data were provided by the Oncological institute of the University medical centre in Ljubljana. We are grateful for suggestions from prof. Ivan Bratko.

References

[1] Brazdil P., Gama J. & Henery B. (1994),

"Characterizing the Applicability of Classification Algorithms Using Meta-Level Learning", *Proc. of ECML-94*, Italy.

[2] Brazdil P.B. & Torgo L. (1990) "Knowledge Acquisition via Knowledge Integration", *Proc. of EKAW-90.*

[3] Breiman L., Friedman J.H., Olshen R.A. & Stone C.J. (1984) *Classification and Regression Trees*, Wadsworth International Group.

[4] Buntine W. (1989) "Learning Classification Rules Using Bayes", *Proceedings of the 6th International Workshop on Machine Learning*, Ithaca, New York.

[5] Catlett J. & Jack C. (1987) "Is it Better to Learn Each Class Separately?", Technical report.

[6] Cestnik, B. (1992), *Probability Estimations in Automatic Learning*, Ph.D. Dissertation.

[7] Cestnik B. & Bratko I. (1988) "Learning Redundant Rules in Noisy Domains", *Proc. of ECAI*, Munich.

[8] Clark P. & Boswell R. (1991) "Rule Induction with CN2: Some Recent Improvements", *Proceedings of EWSL-91*, Porto.

[9] Clark P. & Niblett P. (1989) "The CN2 Induction Algorithm", *Machine Learning*, Vol. 3, No. 4, Kluwer Academic Press.

[10] Gams M. (1988) *Unifying Principles in Automatic Learning*, Ph.D. thesis, Ljubljana.

[11] Gams M. (1989) "New Measurements Highlight the Importance of Redundant Knowledge", *Proc. of EWSL-89*, Montpellier.

[12] Gams M. & Drobnič M. (1988) Approaching the Limit of Classification Accuracy, *Informatica*, No. 2.

[13] Gams M., Drobnič M. & Petkovšek M. (1991) "Learning from Examples - a Uniform View", *International Journal of Man-Machine Studies*, Vol. 34, No. 1.

[14] Gams M. & Karalič A. (1989) "New Empirical Learning Mechanisms Perform Significantly Better in Real Life Domains", *Proc. of the International Workshop on Machine Learning*, Ithaca, New York.

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

[16] Good I.J. (1950) *Probability and the Weighting of Evidence*, Charles Griffing & Co. Limited, London.

[17] Harmon P., Maus R. & Morrisey W. (1988) *Expert Systems, Tools and Applications*, John Wiley & sons.

[18] Kalbfleish J. (1979) *Probability and Statistical Inference II*, Springer-Verlag.

[19] Kononenko I. (1985) "ASSISTANT: A System for Inductive Learning", *M.Sc. thesis*, Ljubljana.

[20] Manago M.V. & Kodratoff Y. (1987) "Noise and Knowledge Acquisition", *Proc. of IJCAI*, Milano.

[21] Michalski, R. (1994) "Inferential Theory of Learning: Developing Foundations of Multistrategy Learning", in Michalski, Tecuci (ed.), *Machine Learnine, A Multistrategy Approach*, Vol. IV, Morgan Kaufmann.

[22] Michalski R.S. & Larson J. (1983) "Incremental Generation of VL1 Hypotheses: The Underlaying Methodology and Description of the Program AQ11", Technical Report ISG 83-5, Urbana: University of Illinois.

[23] Michalski R.S., Mozetič I., Hong J. & Lavrač N. (1986) "The Multi-purpose Incremental Learning System AQ15 and its Testing Application to three Medical Domains", *Proc. of AAAI 86*, Philadelphia, USA.

[24] Michalski, R. & Tecuci G. (ed.) (1994) *Machine Learning, A Multistrategy Approach*, Vol. IV, Morgan Kaufmann.

[25] Niblett T. & Bratko I. (1986) "Learning Decision Rules in Noisy Domains", *Expert Systems*, Brighton, UK.

[26] Quinlan J.R. (1983) "Learning Efficient Classification Procedures and Their Application to Chess End Games", in Michalski R.S., Carbonell J.G. & Mitchell T.M. (Eds.), *Machine Learning: an Artificial Intelligence Approach*, Tioga Publishing, Palo Alto, USA.

[27] Quinlan J.R. (1986) "Induction of Decision Trees", *Machine Learning 1*, Kluwer Academic Publishers.

[28] Quinlan J.R. (1987) "Generating Production Rules From Decision Trees", *Proc. of IJCAI*, Milano.

[29] Rendell L., Powell B., Cho H. & Seshu R. (1988) "Improving the Design of Rule-Learning Systems", *Proc. of The 8th International Workshop on Expert Systems & Their Applications*, Avignon, France.

[30] Rendell L., Seshu R. & Tcheng D. (1987) "Layered Concept-Learning and Dynamically-Variable Bias Management", *Proc. of IJCAI*, Milano.

[31] Shannon C.E. & Weaver W. (1964) *The Mathematical Theory of Communications*, Urbana, Illinois, University of Illinois Press.

[32] Shortliffe E.H. (1976) *Computer-Based Medical Consultations: MYCIN*, American Elsevier Publishing.

[33] Tecuci G. (ed.) (1993) "Special Issue: Multistrategy Learning", *Informatica*, 4 (1993).

[34] Tecuci G., Kedar S. & Kodratoff, Y. (ed.) (1993) *Proc. of IJCAI-93 Workshop Machine Learning and Knowledge Acquisition*, France.

[35] Utgoff P.E. (1986) *Machine Learning of Inductive Bias.*, Kluver Academic Publishers.