AN EXPERIMENT IN AUTOMATIC LEARNING OF DIAGNOSTIC RULES

I. BRATKO, P. MULEC

FACULTY OF ELECTRICAL ENG. AND J. STEFAN
INSTITUTE E. KARDELJ UNIVERSITY, LJUBLJANA'
YUGOSLAVIA

UDK:681.3:616-071

The paper reports on an experiment in automatic learning of classification rules for medical diagnosis. The input to the learning process is a set of examples, i.e. already diagnosed patients. The output is a diagnostic rule, in the form of a decision tree, for diagnosing unknown examples. As a learning method we employed a slightly modified Quinlan's algorithm ID3. The lymphographic investigation served as a problem-domain for the experiment. We used the data about 150 patients, each of them described by a set of 18 discrete attributes and classified into one of 9 alternative diagnoses. The average precision of automatically derived rules obtained in a series of experiments was about 80% when diagnosing unknown patients, which compares favourably to the estimated precision of human diagnosticians. This is between 60 and 85% depending on experience.

POSKUS Z AVTOMATSKIM UČENJEM DIAGNOSTIČNIH PRAVIL. Članek opisuje poskus z avtomatskim učenjem diagnostičnih pravil za diagnosticiranje v medicini. Vhod v proces učenja je množica primerov, to je pacientov z znanimi diagnozami. Izhod je diagnostično pravilo v obliki odločitvenega drevesa za diagnosticiranje neznanih primerov. Kot metodo učenja smo uporabili nekoliko modificiran Quinlanov algoritem ID3, kot problemsko področje za naš poskus pa je služila limfografska preiskava. Uporabili smo podatke o 150 pacientih, opisanih z 18 diskretnimi atributi in klasificiranih v 9 možnih alternativnih diagnoz. Povprečna natančnost diagnostičnih pravil, avtomatsko generiranih v zaporednih poskusih, je bila okrog 80% pri diagnosticiranju neznanih primerov. Ocenjena natančnost diagnostika - zdravnika leži med 60 in 85%.

Introduction

One problem arising in the development of computer applications such as expert information systems is: How to get the problem-domain knowledge into the system? The usual way is that the human domain-expert himself describes his or her own knowledge in some suitable formal language. It often turns that this is a difficult task since the knowledge used by the expert is often intuitive, not systematic, and/or poorly formalised. Examples of problem-domains in which human experts typically use nonformalised knowledge are: medical diagnosis, economic forecasts, playing chess etc.

Another, attractive way of getting the knowledge into the system is based on the use of automatic learning from examples and counter-examples. The domain-expert's task here becomes simpler as he is no more requested to systematically formalise his entire knowledge, but only to provide the system with an adequate set of examples. This set should, hopefully, be sufficient for the system to autonomously recognise the regularities underlying the examples.

In this paper we report on an experiment in automatic learning of medical diagnosis. The diagnostic domain chosen for the experiment was lymphographic investigation. As examples for learning we used old medical data with known correct diagnoses. The result of the learning process was a diagnostic rule in the form of a decision tree. This decision tree defines a mapping between lymphographic data and the corresponding diagnosis, and can thus be used for automatic diagnosis.

Our learning algorithm was based on the Quinlan's automatic learning program ID3 (e.g. Quinlan 1979, Quinlan 1980), which had to be generalised to classification into any number of classes (ID3 could originally deal with two classes only). The results of the experiment indicated that the precision of the automatically learned diagnostic rule superseded that of an average physician - practitioner in this field, and that it is only slightly worse than the precision of best specialists for lymphographic investigation

The learning algorithm

The algorithm used in our experiment is a version of Quinlan's ID3 system, which is based on Hunt's CLS (Concept Learning System, Hunt et. al. 1966).

The input to the algorithm are examples together with their class membership. Each example is described by a set of discrete attributes. Each attribute has typically a few values. All examples are specified by the values of all the attributes (i.e. each example is completely specified), and by the class to which the example belongs. Quinlan's original algorithm works with two classes only. As our problem of lymphographic diagnosis required 9 classes, ID3 had to be modified accordingly. The appropriate generalisation from 2 to N classes of ID3's information-theoretic evaluation function was straightforward.

The output of the algorithm is a decision tree. The nodes of this tree correspond to tests of attributes. The arcs stemming from nodes in the tree correspond to the values of the attribute corresponding to the node. Each leaf of the tree is assigned a class in such a way that this class conta ins all the examples which, according to their attribute values, fall into this leaf.

The algorithm for constructing a decision tree from examples is very simple and efficient. First, a subset, called a "window", of the example set is chosen. A decision tree which "explains" this window is constructed. Then this tree is tested against the whole example set. If the tree explains the whole set (i.e. correctly classifies all the examples in the set) then this tree is the final result of the learning process. If not, then the window is modified by the inclusion of some examples which contradict the current decision tree, whereby possibly deleting some of the members of the old window. A new decision tree is constructed for the new window, then tested against the complete example set, etc.

A decision tree for a given window is constructed in a top-down fashion. First, one of the attributes is selected to become the root of the tree. This attribute partitions the window into "subwindows", so that each subwindow contains examples with the same value of this attribute. Then, subtrees are constructed for all the subwindows. The subtrees are connected to corresponding arcs stemming from the root.

Attributes to become roots of the (sub)trees are chosen by a heuristic criterion: that attribute is chosen which most reduces the information content of the (sub)window.

An implementation of this algorithm is in more detail documented in Mulec 1980.

The problem of Lymphographic diagnosis

In the lymphographic investigation, 18 symptoms are considered. Symptoms correspond to attributes, as referred to in the previous section. There are 9 possible alternative diagnoses; that is: each example is classified into one of 9 classes. Table 1 shows a form which is to be filled in by a physician when diagnosing a lymphograph. The data in this form defines one example for our learning algorithm.

Experiment and results

In the experiment, we used the archive data about 150 patients who were lymphographically investigated at the Institute of Oncology, Ljubljana, over a 3 year period. Fig. 1 shows the diagnostic rule produced by the learning algorithm if all 150 samples were used as training examples.

By the definition of the Quinlan's algorithm, the diagnostic rule has to correctly diagnose all the examples used for training. It is interesting, however, how successfully this diagnostic rule classifies unknown samples. To investigate this question empirically, we randomly permuted all 150 examples, then used the first 100 examples as a training set for the derivation of a diagnostic rule, and then tested the rule on the remaining 50 samples as unknown cases. To eliminate the risk of pathological permutations, this experiment was repeated 10 times, each time with another

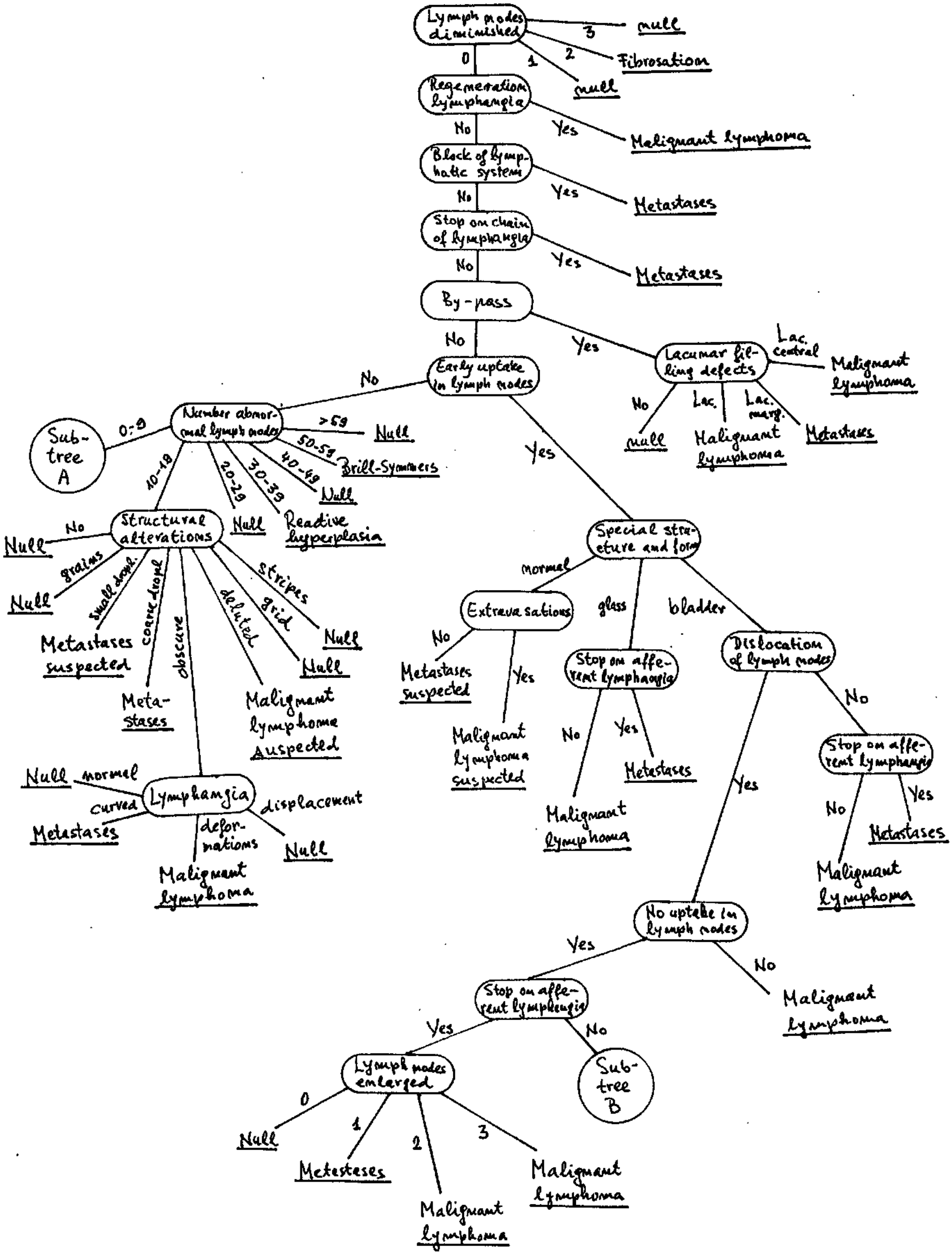

Figure 1: A diagnostic rule for lymphographic investigation.

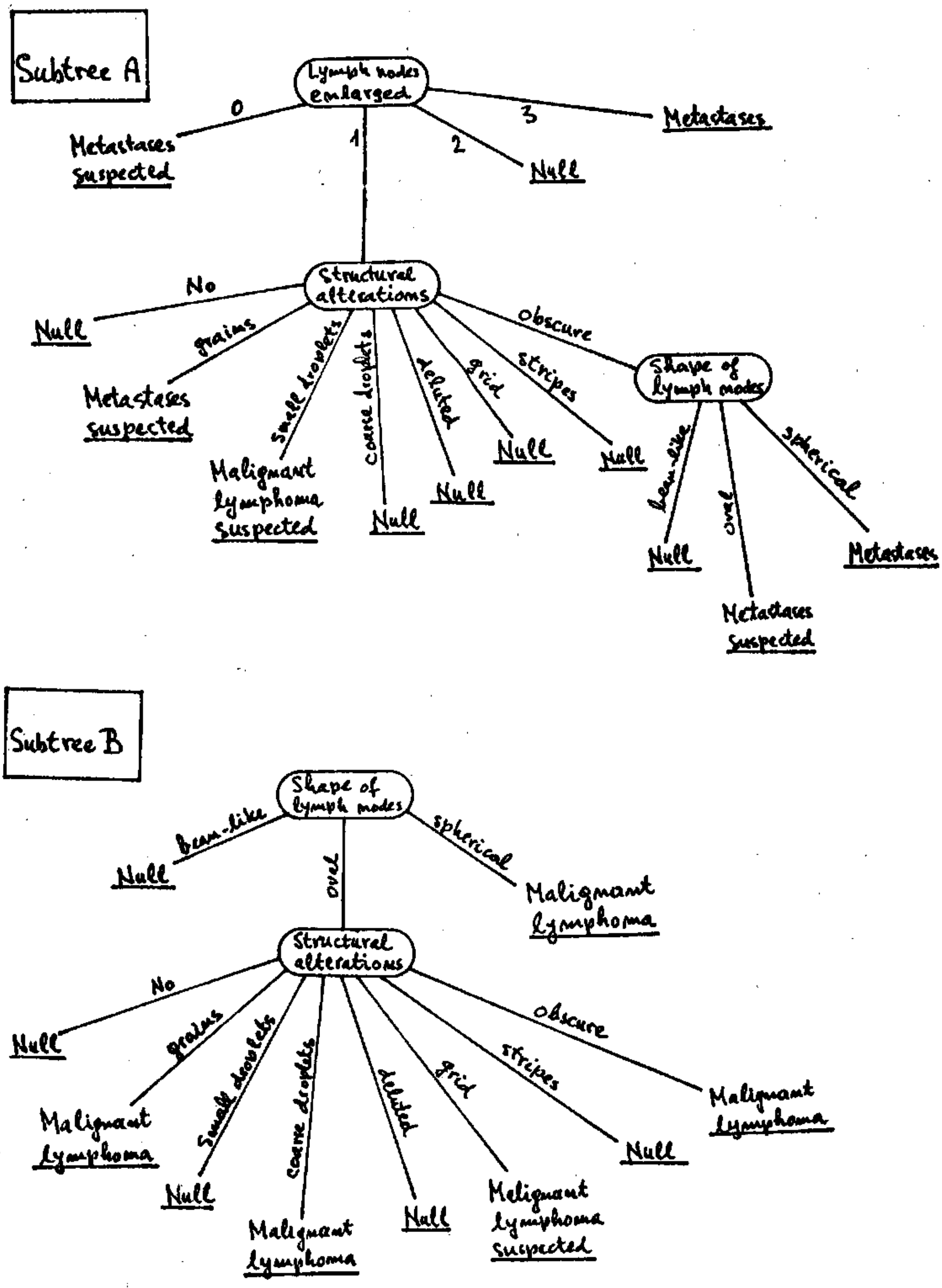

Figure 1: Continued.

Lymphographic attributes

1. lymphangia:

0	normal
1	curved
2	deformations
3	displacement

2. Stop on afferent lymphangia:

1	no
2	yes

3. Stop on chain of lymphangia:

1	no
2	yes

4. Block of lymphatic system:

1	no
2	yes

5. By-pass:

1	no
2	yes

6. Extravasations:

1	no
2	yes

7. Regeneration lymphangia:

1	no
2	yes

8. Early uptake in lymph-nodes:

1	no
2	yes

9. Lymph nodes diminished:

0
1
2
3

10. Lymph nodes enlarged:

0
1
2
3

11. Shape of lymph nodes:

1	bean-like
2	oval
3	spherical

12. Various filling defects:

1	no
2	folicular
3	big central
4	small defects

13. Lacunar filling defects:

1	no
2	lacunar
3	lacunar marginal
4	central

14. Structural alterations:

1	no
2	grains
3	small droplets
4	course droplets
5	deluted
6	grid
7	stripes
8	obscure

15. Special structure and form:

1	glass
2	bladder

16. Dislocation of lymph nodes:

1	no
2	yes

17. No uptake in lymph nodes:

1	no
2	yes

18. Number of abnormal lymph nodes:

0	0-9
1	10-19
2	20-29
3	30-39
4	40-49
5	50-59
6	more than 69

Diagnoses

1	normal
2	reactive hyperplasia
3	metastases suspected
4	malignant lymphoma suspected
5	metastases
6	malignant lymphoma
7	Brill-Symmers
8	fibrosation
9	other diseases

Table 1: Symptoms and diagnoses in lymphographic investigation.

random permutation of the data.

Diagnostic rules were evaluated in two ways: by "absolute precision" and by "relative precision". The relative precision was based on the physicians judgement on the seriousness of particular errors in diagnosis. Thus each possible case of misclassification was assigned a penalty value according to the physician's feeling of how serious was the difference between the wrong and the correct diagnosis.

Absolute precision is the percentage of unsuccessfully diagnosed samples. The following cases were counted as unsuccessful diagnosis:

- the patient falls into a leaf of the decision-tree labelled by another diagnosis;
- the patient falls into a leaf of the decision tree labelled by "null" (that is a leaf which did not match any example in the training set, and therefore the class of this leaf was not known);
- the patient falls into a leaf labelled "search" (that means that in this case the attributes are insufficient for unambigous diagnosis; this situation arises if patients with the same symptoms in the training set were diagnosed differently).

The last case above indicates a sort of insufficiency or incosistency of the training set. It never occured in our set of 150 patients.

The relative precision is computed so that each incorrect diagnosis (the first one of the above three cases) is penalised by a penalty value between 0 and 1. For example, to diagnose a "normal" patient "metastases" is considered to be a most serious error and is therefore penalised by 1. On the other hand, the interchange of the diagnoses "metastases" and "metastases suspected" is a small mistake (penalty 0.1). Table 2 is a penalty matrix for our experiment as proposed by a physician specialised in lymphographic diagnosis.

Table 3 contains some characteristics of the learnt diagnostic rules for all 10 experiments. Columns in the table correspond to the experiments. Each experiment is described by the following parameters:

- the size of the diagnostic rule, i.e. the number of nodes in the decision tree;
- the necessary size of the data-base, i.e. the number of examples in the window which was sufficient for the construction of a decision tree to explain all 100 examples in the training set;
- the number of unknown testing samples which matched a leaf labelled "null";
- the number of unknown testing samples which match a leaf labelled "search" (this was always 0 as our example set was "consistent");
- the number of incorrectly diagnosed samples (case 1 above);

- absolute precision (percentage);
- relative precision (percentage).

Comparatively poor precision in the first experiment can be explained by the fact that the examples in this experiment were not randomly permuted. They were chronologically ordered, covering a few years period. During this period, the human diagnostician's criteria for recognising some of the symptoms were probably changing, which made symptom-patterns of patients, distant in time, incompatible to some extent. The average absolute precision was about 80%, the average relative precision was 88%.

Discussion

To evaluate the above results let us compare the precision of our automatically learned diagnostic rules to that attained by the physicians in practice, and to that of another learning method.

The absolute precision of the lymphographic diagnosis attained by physicians - practitioners in the field, is between 60% and 85%, depending on how experienced is the diagnostician. The 80% average precision of our system compares quite favourably with this 60 - 85% interval.

M.Soklič carried out, at the Institute of Oncology, another experiment in automatic learning using the same medical data and employing his own learning method based on quasi-spherical partitioning of the pattern-space (Raziskovalna skupnost Slovenije, 1978). The precision obtained by that method was: absolute 62%, relative 70%.

These comparisons indicate that our automatically derived diagnostic rule could be successfully applied in the practice of lymphographic diagnosis. Unfortunately a straightforward use of our decision tree by the physician would still require considerable physician's knowledge about lymphographic investigation. This knowledge is necessary for the recognition of symptoms (i.e. attribute values) in lymphographs. It seems that for a really helpful application in this diagnostic problem, a mucn more sophisticated system would be needed. Such a system should guide the user also in recognising particular symptoms, or should itself be capable of recognising visual patterns.

Acknowledgement

The authors would like to thank dr. G. Klanjšček and dr. M.Soklič for advice, advice, and medical data used in our experiment, and dr. M.Zwitter for his suggestions in the preparation of this paper.

References

Hunt, E.B., Martin, J., Stone, P. (1966) Experiments in Induction, Academic Press.

Mulec, P. (1980) Algorithms for automatic learning (Undergraduate thesis). Ljubljana: Faculty of Electrical Eng. (in Slovenian).

Quinlan, J.R. (1979) Discovering rules by induction from large collections of examples, in Expert Systems in the Microelectronic Age (ed. D.Michie). Edinburgh: University Press.

Quinlan, J.R. (1980) Semiautonomous knowledge acquisition, in Expert Systems. London: Infotech.

Diagnosis	1	2	3	4	5	6	7	8	9
1 normal	-	0.33	0.66	0.66	1.00	0.85	0.66	0.50	0.66
2 reactive hyperplasia	0.33	-	0.10	0.33	0.66	0.50	0.10	1.00	0.33
3 metastases suspected	0.66	0.10	-	0.50	0.10	0.50	0.50	0.85	0.33
4 malignant lymphoma suspected	0.66	0.33	0.50	-	0.75	0.10	0.15	0.15	0.50
5 metastases	1.00	0.66	0.10	0.75	-	0.75	0.66	0.50	0.50
6 malignant lymphoma	0.85	0.50	0.50	0.10	0.75	-	0.33	0.15	0.33
7 Brill-Symmers	0.66	0.10	0.50	0.15	0.66	0.33	-	0.85	0.50
8 fibrosation	0.50	1.00	0.85	0.15	0.50	0.15	0.85	-	0.66
9 other diseases	0.66	0.33	0.33	0.50	0.50	0.33	0.50	0.66	-

Table 2: Seriousness of errors in diagnosis.

Index of experiment	1	2	3	4	5	6	7	8	9	10
Rule size	88	80	74	68	53	78	53	58	64	68
Data-base size	82	75	63	62	68	68	65	62	56	62
Null	0	5	3	1	0	6	1	4	4	3
Search	0	0	0	0	0	0	0	0	0	0
Wrong diagnosis	22	6	9	6	10	6	5	8	4	4
Absolute precision (%)	56	78	76	86	80	76	88	76	84	84
Relative precision (%)	80	85	88	91	90	82	93	85	90	91

Table 3: Results in repeated experiments.

APPENDIX: Symptoms and diagnoses in lymphographic investigation
Original form as used at the Institute of Oncology, Ljubljana (in Slovenian)

Limfografski simptomi

1. Mezgovnice:

0 normalno
1 loki
2 deformacije
3 odriv

2. Blok dovodnih mezgovnic:

1 ga ni
2 je

3. Blok mezgovnic verige:

1 ga ni
2 je

4. Blok limfatičnega sistema:

1 ga ni
2 je

5. Obvoz - by pass:

1 ni
2 je

6. Ekstravazati - jezerca:

1 jih ni
2 so

7. Regeneracijske mezgovnice:

1 jih ni
2 so

8. Zgodnje kopičenje v bezgavkah:

1 ga ni
2 je

9. Velikost bezgavk - zmanjšanje:

0
1
2
3

10. Velikost bezgavk - povečanje:

0
1
2
3

11. Sprememba oblike bezgavk:

1 fižol
2 ovalna
3 okrogla

12. Polnitveni defekti razni:

1 jih ni
2 folikularni
3 veliki centralni
4 drobci

13. Polnitveni defekti lakularni:

1 jih ni
2 lakunarni
3 lakunarni marginalni
4 lakunarni centralni

14. Sprememba strukture kopičenja:

1 je ni
2 zrnata
3 drobno kapljasta
4 grobo kapljasta
5 razredčena
6 mrežasta
7 progasta
8 zabrisana

15. Posebna struktura in oblika:

1 kelih
2 mehur

16. Dislokacija - odriv bezgavk:

1 ga ni
2 je

17. Izpad kopičenja bezgavk:

1 ga ni
2 je

18. Število prizadetih bezgavk:

0 0-9
1 10-19
2 20-29
3 30-39
4 40-49
5 50-59
6 več kot 59

Diagnoze

1 normalni izvid
2 reaktivna hiperplazija
3 sumljiv na metastaze
4 sumljiv na maligni limfom
5 metastaze
6 maligni limfom
7 Brill-Symmers
8 fibrozacija
9 ostale bolezni