

J. Laurikkala¹, M. Juhola¹,
S. Lammi², K. Viikki¹

Comparison of Genetic Algorithms and Other Classification Methods in the Diagnosis of Female Urinary Incontinence

¹Department of Computer Science,
University of Tampere,
²Department of Computer Science
and Applied Mathematics,
University of Kuopio,
Finland

Abstract: Galactica, a newly developed machine-learning system that utilizes a genetic algorithm for learning, was compared with discriminant analysis, logistic regression, k-means cluster analysis, a C4.5 decision-tree generator and a random bit climber hill-climbing algorithm. The methods were evaluated in the diagnosis of female urinary incontinence in terms of prediction accuracy of classifiers, on the basis of patient data. The best methods were discriminant analysis, logistic regression, C4.5 and Galactica. Practically no statistically significant differences existed between the prediction accuracy of these classification methods. We consider that machine-learning systems C4.5 and Galactica are preferable for automatic construction of medical decision aids, because they can cope with missing data values directly and can present a classifier in a comprehensible form. Galactica performed nearly as well as C4.5. The results are in agreement with the results of earlier research, indicating that genetic algorithms are a competitive method for constructing classifiers from medical data.

Keywords: Urinary Incontinence, Computer-Assisted Diagnosis, Genetic Algorithms, Machine Learning, Comparison

1. Introduction

The differential diagnosis of female urinary incontinence is a difficult classification problem, which involves the evaluation of medical history, manual measurements and urodynamical testing, to assign the data of an incontinent woman into one of the diagnostic classes. We intended to develop a medical expert system, not solely to assist physicians who are specialized in this area, but also to help general practitioners who, from time to time, encounter women suffering from urinary incontinence. A newly built genetic algorithm-based machine-learning system, Galactica [1, 2], will be included as an integral part in the expert system and will be used to discover automatically expert knowledge from databases.

The purpose of this study was to compare the Galactica system with a

variety of methods used in medicine to build classifiers [3] that predict the class of a case on the basis of case information. In classification terminology, a case is often referred to as an example and it is described by attributes (parameters used in the diagnosis) that have one or more values, selected from a specified set of domain values. Methods chosen for comparison fall basically into two categories, being either statistical techniques or machine learning. Although occasionally it is impossible to allocate a particular method into a single category, some general guidelines exist to help in the differentiation [4]. Many of the statistical methods are parametric, assuming some form of the model and then estimating appropriate values for the model's parameters from the data. Statistical methods also typically focus on tasks in which all attributes have ordinal or continuous values.

Statistical methods include logistic regression and discriminant analysis [5], two basic tools of medical research, and k-means cluster analysis [6], which also possesses some characteristics of unsupervised machine-learning methods. The C4.5 decision-tree generator [4] was selected as a competing machine-learning technique, because it is one of the most powerful learning algorithms available and is widely used as a benchmark for new approaches. Whitley et al. recommended that a simple hill-climbing algorithm should be included in any comparison involving evolutionary algorithms, such as genetic algorithms, to provide baseline results for the test problem [7]. For this reason, Davis' random bit climber (RBC) [8] was also included as one of the comparison methods.

A neural network [9] is probably the best known machine-learning ap-

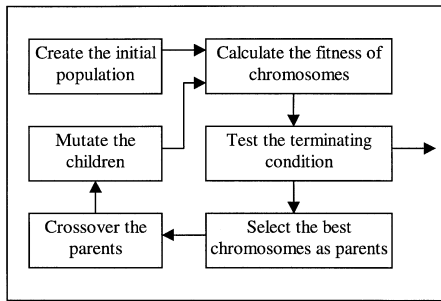


Fig. 1 The typical genetic algorithm.

proach, but this technique had to be excluded because there were not enough cases. A well-known heuristic method for defining the minimum number of examples needed to train a neural network is as follows: Sum the connections between subsequent layers and multiply by 10 [10]. For example, a neural network having 13 input nodes, a hidden layer of 10 nodes and three output nodes would perhaps have been appropriate. However, for this network the training alone would have required 1,600 $([13 \times 10 + 10 \times 3] \times 10)$ cases, which is approximately 3-times the number of the available female urinary incontinence cases.

Genetic algorithms [11-15] are robust search algorithms, utilizing loosely the principles of natural selection and natural genetics. The four distinct properties [13] of the genetic algorithms are the sub-symbolic coding of a problem, search from the population of chromosomes which are also known as solutions or individuals, “blind” search, based only on the fitness of the chromosomes and use of stochastic operators. Figure 1 illustrates the functioning of a typical genetic algorithm. First, the initial population, which usually consists of binary strings, is created. Then, the fitness of each chromosome in the population is calculated. The fittest chromosomes are selected as the parents for the next generation. Pairs of children are produced by exchanging the parts of parents with the crossover operator. In addition, small changes are made to the children with the mutation operator. The genetic algorithm loops until the pre-determined terminating condition is fulfilled.

Genetic algorithms offer many advantages [13, 16, 17], such as conceptual

Attribute	Values	Missing values (%)
Urine in vagina	Yes, No	0.2
Urgency score	Low, High	27.4
Post voiding residual	Normal, High	20.4
Probability of motor urge incontinence	High, Low	63.4
Cystometry	Normal, Abnormal	20.9
Pressure transmission ratio	Normal, Abnormal	34.5
Minimum urethral closure pressure	Negative, Positive	24.0
Stress sign	Yes, No	21.9
Mobility of urethrovesical junction	Normal, Abnormal	36.2
Uroflowmetry	Normal, Abnormal	97.0
Cystoscopy	Normal, Abnormal	90.4
Stress symptom	Yes, No	0.4
Continuous loss of urine	Yes, No	0
Difficulties with voiding	Yes, No	0
Urge symptom	Yes, No	0.8

Table 1 Names, values and the amount of missing data of the differential diagnostic attributes of female urinary incontinence.

simplicity, robustness, broad applicability and the potentiality to hybridize with other methods, but they also suffer from some drawbacks [7, 18, 19] including limitations in current test sequences, optimal parameter settings and, to some extent, weak theoretical foundations. There are few comparative studies for the use of genetics-based machine learning in medicine. Bonelli et al. [20] compared a classifier system [13] with a CN2 logic-reduction system and a neural network for three medical domains. All systems performed very similarly in induction tasks with a slight advantage in favor of neural networks. Janikow [21] found that GIL system was able to perform competitively with human experts, the AQ15 system and the ASSISTANT decision-tree generator, when applied to learning recurrence prognosis for breast cancer from patient data. Congdon [22] showed that

genetic algorithms outperformed other methods in the diagnosis of coronary artery disease in terms of descriptive ability, although decision trees (ID3, GID3* and OBtree) were faster and produced simpler individual rules. Unsupervised learners Autoclass and Cobweb were recognized by her as being inappropriate for the needs of common-disease researchers. All systems predicted unseen cases poorly, probably because of the complexity of the problem.

2. Materials and Methods

2.1 Incontinence Data

The classification methods were tested on the female urinary incontinence data set that was collected retrospectively in the Department of Obstet-

Diagnosis	Learning set			Testing set			Both sets		
	P	N	Total	P	N	Total	P	N	Total
Stress	228	142	370	95	65	160	323	207	530
Mixed	96	269	365	44	121	165	140	390	530
Sensory urge	21	349	370	12	148	160	33	497	530

Table 2 Number of positive P and negative N examples in the learning and testing data sets.

rics and Gynecology of Kuopio University Hospital, Finland. The examples in the data set are characterized by 15 binary-valued attributes which are described in Table 1. In addition, a class attribute identifies the class (diagnosis) of an example. More detailed representations of the class attribute and diagnostic attributes can be found in [23, 24].

It is inevitable that most real-world data sets contain missing values. Also the incontinence data set has a substantial amount of data missing for a variety of reasons (Table 1). Sometimes, the diagnosis has been reached without all possible tests or measurements. It is also possible that there has not been time for some measurements, or equipment has been unavailable or out of order. Since there were not enough complete examples for a reliable statistical analysis, some method had to be applied to replace the missing values. One method is simply to remove an attribute from the data. This policy was applied to two attributes (Uroflowmetry and Cystoscopy) since these had the most missing values.

The missing values of the remaining 13 attributes were replaced with the means calculated from all the available values in the data set. The means were rounded off to the nearest integer value. There exist more advanced methods, such as regression imputation, Expectation-Maximization (EM) imputation, and multiple imputation, to treat missing values in data [25, 26]. However, means imputation was used, because it is a quick and simple method to replace missing values. In addition, we found in recent tests that means imputation produced results quite similar to the results obtained from the more advanced methods [24].

2.2 Comparison Metric

Accuracy (the percentage of correctly classified examples from all examples) [21, 22] was selected as the metric for the comparison of classifiers produced by different methods. In our research, we were mainly interested in prediction accuracy [22], because usually classifiers are evaluated on the basis of their ability to identify correctly unseen cases. Descriptive accuracy [22], which measures the classification accuracy in

known data, was of lesser importance. Examples were assigned randomly in the learning and testing data sets in the 70%/30% ratio often employed [20, 21], and during the splitting process the class attribute was converted to binary form by marking examples as positive or negative with respect to the diagnosis to be learned. Data sets were formed in this way for the three most frequent diagnoses in the data set: stress, mixed and sensory urge incontinence (Table 2). Since other diagnostic classes, motor urge incontinence (N=16) and normal (N=18) were rare, classifiers were not built for these diagnoses in the comparison.

2.3 Machine Learning Systems

C4.5 [4] is an inductive decision-tree generator, which takes as its learning input examples described with attributes having a discrete or continuous domain. All examples belong to one of mutually exclusive classes. The learning output is a decision tree where leaves represent classes and nodes are tests based on attributes. The decision tree is constructed using a top-down approach which starts from the root of the tree and is applied recursively until the tree is complete. At each step of the building process, the attribute which divides the learning set in the best possible way, in terms of gain ratio, is selected to be the test of the node. Overly complex decision trees can be reduced with pruning and when further simplification is needed, unpruned trees may be converted to rules. The conversion to rules is not straightforward knowledge reformulation, because rules are also generalized by deleting conditions, which seem to be irrelevant to the classification.

Galactica [1, 2] utilizes a genetic algorithm to learn inductively rules from pre-classified examples, which are characterized by discrete attributes. Knowledge is represented in sub-symbolic chromosomes for the genetic algorithm and in symbolic rules for humans. For example, a rule for the stress incontinence diagnosis might be: *[Urine in vagina = No] AND [Stress symptom = Yes] AND [Difficulties with voiding = No] AND [Urge symptom = No]*. The genetic algorithm maintains a population of variable-length chromosomes

from which the fittest individuals are selected for reproduction. The parents are modified with the crossover and mutation operators to generate a new generation. After the genetic algorithm has terminated, chromosomes can be decoded as symbolic rules for examination.

The chromosomes are coded as binary strings so that for each condition there are reserved as many bits as the corresponding attribute has values, and the condition bit is set to one if the attribute has a corresponding value, otherwise the bit is set to zero [21, 27]. If an attribute does not exist in a rule, all of its value bits are turned to one, that is, the attribute can have any value and is therefore meaningless. Since the 13 attributes were binary valued, the chromosomes consisted of 26 ones or zeroes. The sub-symbolic coding for the previous rule would be: 0111111111111111111010101. The order of the attributes is the same as shown in Table 1. The first two bits are reserved for the urine-in-vagina attribute: the first bit corresponds to value "Yes" and the second bit corresponds to value "No". The last two bits in the chromosomes are reserved for the urge symptom attribute.

The fitness of a chromosome is mostly based on the number of positive and negative examples covered in the learning set. The fitness increases as the positive cover grows and the negative cover diminishes. In addition, the complexity of chromosome affects slightly the fitness of a chromosome; the simpler chromosomes are considered better than the more complex ones. A binary coded example is covered by the chromosome when the logical AND operation between the example and the chromosome returns the example unchanged. In other words, the example is covered if the attribute values both in the example and chromosome match. Missing values do not prevent classification, because match is considered to happen when an attribute is absent in the example or chromosome, or when the attribute is absent both in the example and chromosome.

Random bit climber (RBC) [8] modifies rules which are presented in binary strings as in the Galactica system. Search starts from a random position

Method	Diagnosis						Mean	
	Stress		Mixed		Sensory urge		accuracy	
	D	P	D	P	D	P	D	P
k-means cluster analysis	73	74	64	71	65	59	67	68
Discriminant analysis	87	96	86	81	89	89	87	89
Logistic regression	89	93	87	86	95	92	90	90
C4.5 (tree)	90	93	89	82	97	93	92	89
C4.5 (rule)	90	93	89	86	97	93	92	91
Galactica	86	95	86	82	93	94	88	90
Random bit climber	81	88	85	82	92	92	86	87
Mean accuracy	85	90	84	81	90	87		

Table 3 Descriptive D and prediction P accuracy (%) of classifiers obtained with different methods from the repaired female urinary incontinence data.

3. Results

The prediction and descriptive accuracy of the classifiers obtained from the repaired data are shown in Table 3. Discriminant analysis and Galactica had the best prediction accuracy for female urinary stress incontinence. Logistic regression and C4.5 with rules produced the most accurate classifiers for mixed incontinence. C4.5 and Galactica classified the sensory urge diagnosis with the highest prediction accuracy. The best overall performer was C4.5 with rules, the second best being Galactica and logistic regression. RBC achieved unexpectedly good results. Instead, k-means cluster analysis achieved the lowest prediction accuracy in all data sets and was also the worst overall performer. The machine-learning systems and RBC were also tested on learning classifiers from the incomplete original data (Table 4). Both trees and rules of C4.5 slightly outperformed Galactica, while RBC produced less accurate results than the other methods. Decision trees produced by C4.5 had the highest prediction accuracy in all data sets formed from the original data.

Due to their stochastic nature, Galactica and RBC occasionally produced very accurate results. For example, the best rule that RBC learned from original stress incontinence data had 94% prediction accuracy. Unfortunately, randomness leads also to clearly sub-optimal results. Therefore, the accuracy of Galactica and RBC shown in Tables 3 and 4 are the means of the best results from 30 independent runs (Table 5). Galactica produced more coherent results than RBC, especially with the stress and mixed incontinence data. The replication of runs was carried out so that Galactica and RBC were run in batch mode where the methods repeated a given number of runs and saved the results into a log file without human interaction.

In contrast to the results of Congdon [22], the prediction accuracy was high in the repaired and original data. The difference may be explained by the complexity of the coronary artery disease diagnosis problem. The prediction accuracy of diagnostic rules found in our earlier experiments [1, 2] was slightly lower with original stress inconti-

where a bit is flipped, and also the following bits are selected randomly. If a change produces a better solution, it is accepted and search continues using this solution until a new improvement is found or every bit has been changed. When all bits have been flipped, a new random sequence is generated. If no improvement was found during bit-flipping, the search has located a local optimum and, therefore, a new random solution is generated [7].

2.4 Experimental Setup

The statistical analysis was conducted with the repaired stress, mixed and sensory urge data using the statistical software package SPSS for Windows 6.1.3. A logistic regression model was constructed with the Backward elimination method and discriminant analysis was done with the Stepwise method to leave only the statistically significant attributes in the model. The accuracy of the models generated by both methods

was calculated in the testing sets using the function coefficients and the cut-off values obtained from the learning sets. In cluster analysis, all attributes were left in the model and the final clusters formed from the learning set were used to classify examples in the testing sets to either of the two clusters.

Machine learning systems and RBC were applied to both the incomplete original and the complete repaired data. C4.5 was run on a Sun SparcStation with basic configuration to produce pruned trees and rules. Galactica had the following control parameter values: population size = 100, generations = 100, probability of mutation = 0.001 and probability of crossing-over chromosomes = 0.5 [12]. Initial populations were generated in a totally random manner and the best solution in each generation was moved without modifications into the next generation. RBC was allowed to pass the random mutation sequence 100 times.

Method	Diagnosis						Mean	
	Stress		Mixed		Sensory urge		accuracy	
	D	P	D	P	D	P	D	P
C4.5 (tree)	88	97	89	86	96	93	91	92
C4.5 (rule)	87	96	87	80	96	93	90	90
Galactica	85	93	87	82	89	87	87	87
Random bit climber	77	84	86	82	89	87	84	84
Mean accuracy	84	93	87	83	93	90		

Table 4 Descriptive D and prediction P accuracy (%) of classifiers obtained with different methods from the original female urinary incontinence data

Method		Diagnosis								
		Stress			Mixed			Sensory urge		
		min	max	std dev	min	max	std dev	min	max	std dev
Galactica	R	68	96	5	82	82	0	87	94	2
	O	68	94	6	82	82	0	75	91	5
RBC	R	41	96	19	69	82	3	87	94	3
	O	41	94	19	73	82	2	75	93	4

Table 5 The minimum and maximum prediction accuracy (%) and standard deviation (%) of 30 of the best solutions obtained with Galactica and RBC from the repaired R and original O stress, mixed and sensory urge incontinence data.

rules which C4.5 and Galactica learned from the repaired data had the highest overall prediction accuracy, being 91% and 90%, respectively. Logistic regression was as accurate as Galactica, but it is probable that, if complete real-world data had been available, logistic regression and discriminant analysis would have achieved somewhat lower accuracy, because replacement of missing values with means reduces the variance of attribute values and makes model building easier for statistical methods.

Presentation of the information in a comprehensible manner is crucial in areas such as decision making in medicine, where humans must fully understand the classifiers [28]. Logistic regression and discriminant analysis create a mathematical model which gives valuable information, for example, about the dependencies between diagnos-

nence data (89%) and somewhat higher with original mixed incontinence data (86%) than the prediction accuracy of Galactica's rules reported in this paper. These differences are due to the random manner in which the learning and testing sets are created. When new learning and testing sets differ from the previous sets, it is also likely that the accuracy of the rules will differ slightly.

Further statistical analysis showed that k-means cluster analysis produced a classifier from repaired data, having a significantly ($p < 0.05$) lower prediction accuracy than the other classifiers (Fig. 2). There were practically no statistically significant differences between the remaining methods. In original data, the accuracies of C4.5 generated trees and RBC's rules were significantly different in favor of C4.5.

4. Discussion

The best methods were logistic regression, discriminant analysis, C4.5 and Galactica. It is difficult to select a single method as victor, because there were no clear statistically significant differences between these methods. Probably the ideal situation would be that both statistical techniques and a machine-learning system could be used together to solve the problem. However, if one must choose between these approaches, we consider that the machine-learning approach is preferable over statistical techniques, especially in the automatic building of medical decision aids, because of missing data and the need to present the classifiers in a form which is comprehensible to human beings.

The problem of missing data in medical data sets was evident in this comparison: none of the statistical methods could perform reliably without replacing the missing values. In contrast, C4.5 and Galactica were able to produce accurate trees and rules from the incomplete original data. Furthermore,

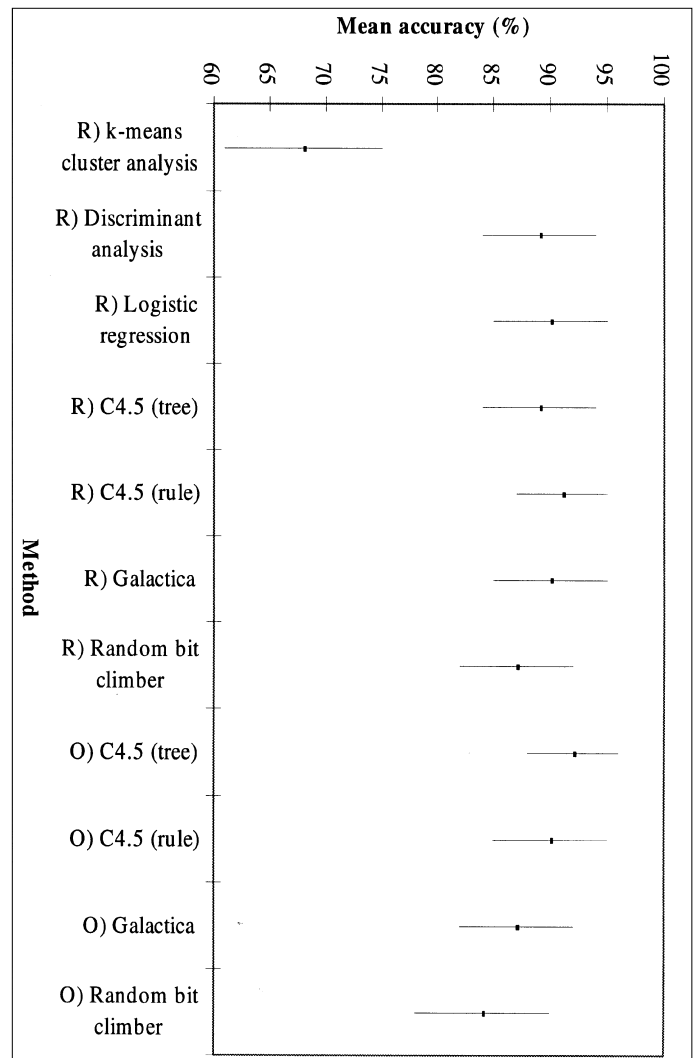


Fig. 2 95% confidence intervals for the prediction accuracy of the compared methods. R) Repaired data. O) Original data.

tic parameters and diagnosis. Understanding and thus correct interpretation of these non-symbolic models is usually quite difficult for individuals who do not have considerable statistical knowledge. Consequently, medical decision-support systems constructed from the results of statistical methods are black boxes having limited capabilities to explain their decisions. Moreover, rigorous evaluation and testing may be far more difficult than with transparent systems [29]. Conversely, decision trees and rules are easily understandable and no additional expertise is needed. An expert in the problem area, for example a physician, can directly evaluate and verify this type of classifier.

The results showed that C4.5 achieved the highest overall prediction accuracy both from the complete repaired and incomplete original data. The success of C4.5 was expected, because this system is mature and previous studies have proven its power. Incomplete original data was easier for C4.5 than for Galactica: the prediction accuracy of C4.5 trees, C4.5 rules and Galactica rules were 92%, 90% and 87%, respectively. This difference may be explained by the different policies used to treat missing attribute values. C4.5 uses a sophisticated probabilistic approach to cope with missing values both in the tree-building process and classification [4]. Since Galactica does not employ any special policy for incomplete data, missing values may lead to overgeneralization of a rule, thus reducing its classification ability. This is due to the fitness criteria which favors general rules covering both general and specific examples in data [1, 2].

The most unexpected result was the good performance of RBC, which was originally included to provide a baseline result. This result suggests that the current set of diagnostic parameters may be too small to allow the more complex methods to express their full power. It is likely that in a larger search space RBC would not perform as well as in this study in comparison with the other methods. Clearly, k-means cluster analysis was the worst method in this comparison. The unsupervised manner in which the clusters are formed without considering the class information might predictably lead to a poor performance.

This result seems to confirm the conclusions by Congdon, indicating that unsupervised learning methods are likely to perform poorly in terms of descriptive accuracy, especially when compared to genetic algorithms and decision trees [22]. Additional information, such as setting the number of clusters equal to the number of diagnoses in data and initializing the centers, allow a clustering algorithm to achieve better results [23].

The majority of classifiers for female urinary stress incontinence had DESCRIPTIVE accuracies higher than the PREDICTION accuracies. Usually, the known data can be classified with higher accuracy than the unseen cases. This result is probably due to the heterogeneity of the stress-incontinence cases. Sometimes the stress incontinence-related symptoms are so obvious that the patient can be referred to surgery without further tests and measurements. However, most of the stress-incontinent women undergo urodynamic measurements. As a result, the classifiers must identify cases having most attributes absent and cases having nearly all attributes present. When the original data were used, the machine learning methods generated general classifiers, which could classify the heterogeneous data. When the data were replaced using the rounded means, many new complete stress incontinence cases that resembled each other and the original complete cases were introduced into the data. Since the variance of data was reduced, especially the classification of the repaired stress-incontinence data became easier. Although the cases were randomly placed in the learning and testing sets, it may also be possible that the cases were divided so that the classification of the unseen stress-incontinence cases was easier than the classification of the other cases.

5. Conclusions

In conclusion, the results are in agreement with the results of earlier research [20-22], indicating that genetic algorithms are a competitive method to generate classifiers from medical data. Although there were no clear statistically significant differences in prediction accuracy between the best methods

(logistic regression, discriminant analysis, C4.5 and Galactica), we propose that machine-learning systems C4.5 or Galactica are a preferable choice for building classifiers to aid decision making in medicine. Firstly, these systems are able to process medical data directly without compulsory application of replacement methods for missing values. Secondly, machine learning systems produce classifiers that humans find readily comprehensible and verifiable.

C4.5 produced slightly more accurate classifications than Galactica, but the performance difference was small. Moreover, Galactica is a novel system (the first version of our method) and we are confident that further research is likely to make it more efficient. For example, it would be interesting to hybridize RBC with Galactica to produce a better initial population. RBC might also be used to perform a local search after the genetic algorithm has located a near optimal solution.

Acknowledgments

This study was funded by grants from the Academy of Finland, the Technology Development Centre (TEKES) and Oskar Öflund Foundation. The authors are grateful to physicians Jorma Penttinen, MD, and Pauliina Aukee, MD, for their aid in medical issues. We would like to thank Erkki Pesonen, PhD, for comments concerning the treatment of missing data values.

REFERENCES

1. Laurikkala J, Juhola M. Learning diagnostic rules from a urological database using a genetic algorithm. In: Alander JT, eds. *Proceedings of The Third Nordic Workshop on Genetic Algorithms and their Applications*. Helsinki: Finnish Artificial Intelligence Society, 1997: 233-44.
2. Laurikkala J, Juhola M. A genetic-based machine learning system to discover the diagnostic rules for female urinary incontinence. *Comput Programs Biomed* 1998; 55: 217-28.
3. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Belmont: Wadsworth International Group, 1984.
4. Quinlan RJ. *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann, 1993.
5. Armitage P. *Statistical Methods in Medical Research*. Oxford: Blackwell Scientific Publications, 1994.
6. Norušis MJ. *SPSS for Windows Professional Statistics Release 6.0*. Chicago: SPSS Inc., 1993.
7. Whitley D, Rana S, Dzuberka J, Mathias KE. Evaluating evolutionary algorithms. *Artif Intel* 1996; 85: 245-76.

8. Davis L. Bit-climbing, representational bias, and test suite design. In: Belew RK, Booker LB, eds. *Proceedings of the Fourth International Conference on Genetic Algorithms*. San Mateo: Morgan Kaufmann Publishers, 1991: 18-23.
9. Wasserman PD. *Neural Computing: Theory and Practice*. New York: Van Nostrand Reinhold, 1989.
10. Swingler K. *Applying Neural Networks, a Practical Guide*. London: Academic Press, 1996.
11. Holland JH. *Adaptation in Natural and Artificial Systems*. Ann Arbor: The University of Michigan Press, 1975.
12. De Jong KA. *An Analysis of the Behaviour of a Class of Genetic Adaptive Systems*. PhD Thesis. University of Michigan, 1975.
13. Goldberg DE. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading: Addison-Wesley, 1989.
14. Mitchell M. *An Introduction to Genetic Algorithms*. Cambridge, MA: MIT Press, 1996.
15. Bäck T, Fogel DB, Michalewicz Z. *Handbook of Evolutionary Computation*. Bristol: Institute of Physics Publishing and Oxford University Press, 1997.
16. Goldberg DE. Genetic and evolutionary algorithms come of age. *Comm ACM* 1994; 37(3): 113-9.
17. Fogel DB. The advantages of evolutionary computation. In: Lundh D, Olsson B, Narayanan A, eds. *Biocomputing and Emergent Computation: Proceedings of BCEC97*. World Scientific, 1997.
18. Bäck T, Hammel U, Schwefel H-P. Evolutionary computation: comments on the history and current state. *IEEE Transactions on Evolutionary Computation*, 1997; 1: 3-17.
19. Schwefel H-P. Advantages (and disadvantages) of evolutionary computation over other approaches. In: Bäck T, Fogel BF and Michalewicz Z, eds. *Handbook of Evolutionary Computation*. Bristol: Institute of Physics Publishing and Oxford University Press, 1997: A 1.3: 1-2.
20. Bonelli P, Parodi A. An efficient classifier system and its experimental comparison with two representative learning methods on three medical domains. In: Belew RK, Booker LB, eds. *Proceedings of the Fourth International Conference on Genetic Algorithms*. San Mateo: Morgan Kaufmann Publishers, 1991: 288-95.
21. Janikow CZ. A knowledge-intensive genetic algorithm for supervised learning. *Mach Learn* 1993; 13: 189-228.
22. Congdon CB. *A Comparison of Genetic Algorithms and Other Machine Learning Systems on a Complex Classification Task from Common Disease Research*, PhD Thesis. University of Michigan, Department of Computer Science and Engineering, 1995.
23. Laurikkala J, Juhola M, Penttinen J, Aukee P. Parameter evaluation of the differential diagnosis of female urinary incontinence for the construction of an expert system. In: Pappas C, Maglaveras N, Scherrer JR, eds. *Proceedings of Medical Informatics Europe '97*. Amsterdam: IOS Press, 1997: 671-5.
24. Laurikkala J, Juhola M, Lammi S, Penttinen J, Aukee P. Statistical evaluation of the parameters in the differential diagnosis of female urinary incontinence for the construction of an expert system (in press).
25. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, 1987.
26. Schafer JL. *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall, 1997.
27. De Jong KA, Spears WA, Gordon DF. Using genetic algorithms for concept learning. *Mach Learn* 1993; 13: 161-88.
28. Michalski RS, Kodratoff Y. Research in machine learning. In: Michalski RS, Kodratoff Y, eds. *Machine learning: An Artificial Intelligence Approach, vol. 3*. San Mateo: Morgan Kaufmann Publishers, 1990: 3-30.
29. Hart A, Wyatt J. Evaluating black-boxes as medical decision aids: issues arising from a study of neural networks. *Med Inf* 1990; 15: 229-36.

Address of the authors:
 Jorma Laurikkala,
 University of Tampere,
 Department of Computer Science,
 P.O. Box 607,
 FIN- 33101, Tampere,
 Finland
 E-mail: Jorma.Laurikkala@cs.uta.fi