Breast tumor malignancy modelling using evolutionary neural logic networks

ATHANASIOS TSAKONAS[1], GEORGIOS DOUNIAS[2], GEORGIA PANAGI[3] and EVANGELIA PANOURGIAS[4]

[1]Artificial Intelligence and Information Analysis Laboratory, Department of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki; [2]Department of Financial and Management Engineering, University of the Aegean, 31 Fostini Str.; [3]Department of Radiology, General Hospital of Chios 'Skilitsion', 82100 Chios; [4]Department of Radiology, Euroclinic Hospital, 9 Athanasiadou Str., 11521 Athens, Greece

Received September 6, 2005; Accepted September 28, 2005

Abstract. The present work proposes a computer assisted methodology for the effective modelling of the diagnostic decision for breast tumor malignancy. The suggested approach is based on innovative hybrid computational intelligence algorithms properly applied in related cytological data contained in past medical records. The experimental data used in this study, were gathered in the early 1990s in the University of Wisconsin, based in post diagnostic cytological observations performed by expert medical staff. Data were properly encoded in a computer database and accordingly, various alternative modelling techniques were applied on them, in an attempt to form diagnostic models. Previous methods included standard optimisation techniques, as well as artificial intelligence approaches, in a way that a variety of related publications exists in modern literature on the subject. In this report, a hybrid computational intelligence approach is suggested, which effectively combines modern mathematical logic principles, neural computation and genetic programming in an effective manner. The approach proves promising either in terms of diagnostic accuracy and generalization capabilities, or in terms of comprehensibility and practical importance for the related medical staff.

Introduction

Breast cancer diagnosis consists one of the major fields of interest in modern oncology, either as part of ongoing conventional medical research, or through a number of approaches involving computational analysis and decision support systems. Computer assisted medical diagnosis gained increased acknowledgement in related literature during the last decade, although real-world applications are still at the research level, rather than forming attractive commercial products, ready to be embodied in wider hi-tech medical systems. Nevertheless, the research results are definitely encouraging, as we are able of construct methodologies which produce effective generalized diagnostic models and in some cases can manage even to discover new expert knowledge, hidden inside past medical records. The proposed data analysis methodologies vary in literature from standard optimization techniques and classical statistics, to modern data mining algorithms and computational intelligence schemes. Basic presuppositions for applying such data analysis techniques are the medical doctors' collaboration and expertise, and also the existence of complete, if possible computerized, related medical records. Recently, hybrid data analysis approaches are appearing in literature, trying to effectively combine more than one known data analysis methodologies.

In this work, such a hybrid computational intelligence approach is suggested, combining modern mathematical logic principles, neural computation and evolutionary computation in an effective manner. Specifically, the evolutionary methodology guides the formation and the tuning of a neural computation model, constructing finally an evolutionary neural logic network (ENLN), described in detail previously (1). An advanced evolutionary computation approach is used, namely grammar-guided genetic programming (GGGP), using cellular encoding. The neural computation model used in this hybrid intelligent approach is called neural logic network (NLN). The NLN approach is considered advantageous for incorporating 3-valued mathematical logic principles. In other words, 'true', 'false' and 'do not know' values can be given to related decision concepts and attributes, a feature which seems to be very close to real-world practice in medical diagnosis tasks. The NLN uses properly encoded medical data as input and is being trained on them over time, in order to produce an effectively generalized decision mechanism, i.e., the best possible classifier for the given problem. The effectiveness of the overall architecture depends on the size, the variety and the reliability of the available past data and approximately reflects the real frequency distribution of the suspected cases arrived at hospital to be diagnosed for breast

Correspondence to: Dr Georgios Dounias, Department of Financial and Management Engineering, University of the Aegean, 31 Fostini Str., 82100 Chios, Greece
E-mail: g.dounias@aegean.gr

Key words: computer assisted breast cancer diagnosis, evolutionary neural logic networks, grammar guided genetic programming, 3-valued logic

Table I. Input features of the breast cancer database.

Variable	Feature	Value range
T1	Clump thickness	1-10
T2	Uniformity of cell size	1-10
T3	Uniformity of cell shape	1-10
T4	Marginal adhesion	1-10
T5	Single epithelial cell size	1-10
T6	Bare nuclei	1-10
T7	Bland chromatin	1-10
T8	Normal nucleoli	1-10
T9	Mitoses	1-10

cancer. In fact, such a mechanism corresponds to a computer-based knowledge generation model which reflects high-level human expertise on specific domains of application (in our case on breast cancer diagnosis from cytological information).

A significant part of the success of the proposed methodology lies on the proper and systematic preparation, modelling and repeated experimentation of the applied hybrid intelligent scheme. The outcome (i.e. the generalized classifier produced) has the additional advantage to be represented as a set of readable first-order logic rules. Thus, it is accessible and interpretable for the medical staff, while it can also be used as a fast and handy 'second-opinion' solution, as it can be modelled as a single line of code in a personal computer.

Materials and methods

The database of breast cancer patients has been previously created in the Medical University of Wisconsin (2,3). The diagnosis is concerned with the classification of a tumor as benign or malignant. In the past, a part of this data has been investigated using theory of linear programming (4,5) to construct a generalized decision model and the diagnostic accuracy obtained in a total of 169 records ranged between 93.5% and 95.9% depending on tuning details of the system. In a larger data set coming from the same source (369 records in total) specific machine learning techniques were applied (6) in order to form a general decision model for the problem. In a test set of 169 records (considered as unknown data for the decision model), the diagnostic accuracy ranged between 92.2% and 93.7% depending again on the system tuning. In the full data that are now available for experimentation (699 records), a subset of 458 (65.5%) concern benign tumors and the rest 241 (34.5%) correspond to malignant ones. The database features are integer numbers in [1,10] corresponding to a quantitative characterization of nine laboratory measurements (T1-T9) of the cells, presented in detail in Table I. To avoid overfitting during the training phase of our system, we made use of a validation set. Half (i.e. 50%) of the total data set were used as training data, two subsets of 25% of the total set each (i.e. 174 cases), were used to form the validation set, as well as the testing set. All these sets were created randomly. The split of the entire dataset into training, testing and validating subsets (usually according to the abovementioned 50-25-25 analogy), is a common practice in genetic programming experimentation, instead of using typical cross validation schemes. This is necessary due to the fact that each complete GP-training cycle is a very time-consuming process, growing exponentially with the complexity and the size of the data set. Finally, note that the missing data percentage in our case was rather low (≤1‰).

As mentioned above, this study makes use of neural logic networks and genetic programming. Below, we briefly explain fundamental concepts of the methods.

The neural logic network (7) is a finite directed graph. It usually consists of a set of input nodes and an output node. In its 3-valued form, the possible value for a node can be one of three ordered pair activation values (1,0) for 'true', (0,1) for 'false' and (0,0) for 'do not know'. Every synapse (edge) is assigned also an ordered pair weight (x,y) where x and y are real numbers. Different sets of weights enable the representation of different logical operations (i.e., conjunction, disjunction, implication, etc.). It is actually possible to map any rule of conventional knowledge into a neural logic network. Neural logic networks can be expanded into fuzzy neural logic networks, enabling this way the handling of real valued attributes (7). Even though powerful in their definition, neural logic networks are not widely applied. The main reason can be located in the fact that for the known training methodologies (7,8), the refinement of the edge weights reduces significantly the interpretability of these networks to expert rules, thus depriving these networks from their valuable feature. Some steps for the preservation of the interpretability have been performed by Chia and Tan (9), without however the ability to express arbitrarily large and connected neural logic networks.

The ability to construct functional trees of variable length is a major advantage of genetic programming over genetic algorithms. This property enables the search for very complex solutions that are usually in the form of a mathematical formula - an approach that is commonly known as symbolic regression. Later paradigms extended this concept to calculate any Boolean or programming expression. Consequently, complex intelligent structures, such as fuzzy rule-based systems or decision trees have already been used as the desirable target solution in genetic programming approaches (10-13). The main qualification of this solving procedure is that the feature selection, and the system configuration, derive in the searching process and do not require any human involvement. Moreover, genetic programming is capable of avoiding local minima. The potential gain of an automated feature selection and system configuration is obvious; no prior knowledge is required and, furthermore, not any human expertise is needed to construct an intelligent system. Nevertheless, the task of implementing complex intelligent structures into genetic programming functional sets is not straightforward. The function set that composes an intelligent system retains a specific hierarchy that must be traced in the GP tree permissible structures, thus: a) avoiding meaningless candidate solutions, and b) reducing the search space to valid solutions solely. This approach, known in literature as legal search space handling method (14), has been implemented in this work using context-free grammars.

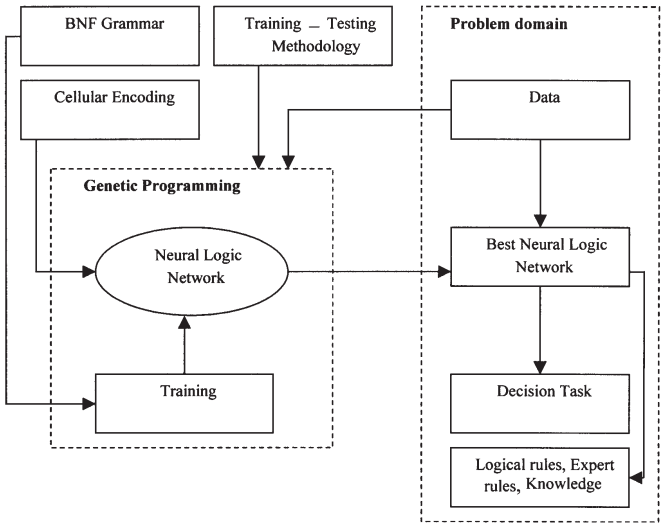

Figure 1. The ENLN methodology.

The genetic programming procedure might prove greedy in computational and time resources. Consequently, when the syntax form of the desired solution is already known, it is useful to restrain the genetic programming from searching solutions with different syntax forms (15,16). The most advantageous method to implement such restrictions among other approaches (17) is to apply syntax constraints to genetic programming trees, usually with the help of a context-free grammar declared in the Backus-Naur-Form (BNF) (18). The BNF-grammar generally consists of terminal nodes and non-terminal nodes. Although mapping decision trees or fuzzy rule-based systems to specific grammars can be relatively easy to implement, the execution of massively parallel processing intelligent systems - such as the neural logic networks - is not forthright. In order to explore variable sized solutions, we applied indirect encoding. The most common one is the cellular encoding (19,20), in which a genotype can be realized as a descriptive phenotype for the desired solution. More specifically, within such a function set, there are elementary functions that modify the system architecture together with functions that calculate tuning variables. Current implementations include encoding for feed forward and Kohonen neural networks (21,22) and fuzzy Petri-nets (22,23).

The general procedure followed by the suggested methodology, is shown in Fig. 1. The decision task for the given medical problem and the available data, are properly formulated and encoded, with the BNF grammars and cellular encoding principles, in order to be used by the genetic programming generalization mechanism. The training phase then initiates for the construction of an efficient neural logic network. The process evolves over time, until the best possible neural logic network architecture is found for the given training and test data. Then the output is transformed into specific logical rules and related medical knowledge. A more detailed description of the ENLN methodology can be found in refs. 1,24. As mentioned above already, the programming functions implemented by the ENLN correspond to known logical operators of first or higher order logics, such as conjunction, disjunction, equivalence, majority, at least-k property, etc. Specifically, a variety of functions have been defined and implemented, like PROG, S1, S2, P1, P2, IN, E, LNK, CNR, NUM, CUT, K, CNRSEL, etc., whose operation in the phase of design and operation of an ENLN, is briefly explained in Table II.

Results

To avoid over-fitting during the ENLN training phase, a validation set was used. The extracted solution achieved accuracy 94.25% (164/174) in unknown data (test set). The accuracy in the training set and in the validation set was 97.99% (341/348) and 97.12% (169/174) correspondingly. The extracted solution in prefix notation is shown in the last row of Table II and corresponds to a sequence of interdependent logical rules connected with brackets as logic programming expressions. Note that the final decision result is formed during three phases: first the ENLN-output is produced in prefix notation, then the graphical representation of the ENLN takes place, and finally the following logical rules are formed: $Q_1 \leftarrow$ conjunction (T2, T6, T8); $Q_2 \leftarrow$ priority (Q_1); $Q_3 \leftarrow$ k-majority (T5); $Q \leftarrow$ conjunction (Q_1, Q_2, Q_3, T1, T2, T3, T5, T6, T7).

The above solution was extracted after 26000 iterations of the entire algorithmic scheme. The first logical operation named conjunction denotes simultaneous satisfaction of conditions existing for the involved attributes. The second operation checks priority among the specific attributes involved. The operation k-majority examines specific conditions for the involved attributes and becomes true when the majority of them are true. In the above solution priority and k-majority work as operations that de-amplify the processed input. The final conjunction is very important for the production of the final decision. The attributes included in this conjunction (T1, T2, T3, T5, T6, T7) are considered necessary for the final decision, the same as attribute T8 appeared in the first rule. Attributes T4 and T9 do not appear at all in the final solution. A detailed study of the graphical representation of the ENLN in relation to the acquired rules, leads to the conclusion that T3 is the most important attribute for forming the final diagnosis. The detailed information for reading the above result can be found in refs. 1,24,25.

Discussion

Note that the proposed ENLN system obtains a very high diagnostic accuracy compared to other approaches found in literature, Thus, it can be directly used by medical experts as a 'black-box' diagnostic engine for performing an assisting second opinion diagnosis, through a proper interface that presents this result in a more readable and ready-to-use way. As the reader can observe, the total outcome and thus, also the extracted decision rules are rather complicated to be analyzed and discussed in deep detail. Nevertheless, a number of interesting points can be observed.

Only 15 complete diagnostic paths, consisting of one to five premise parts, are adequate to describe the whole set of cases (700 past diagnoses of individuals) and produce a generalized network of inference that can diagnose correctly

Table II. Brief description of the ENLN-function set used and presentation of the final solution obtained for the breast cancer diagnosis problem.

Function PROG/CNLN	Consists always the initial node of a tree. Creates the embryonic network, later used by S1, S2, P1, P2, to be expanded
Function S1	Enters a node in serial to the node that is applied, and is applied to input nodes
Function P1	Enters a node in parallel to the node that is applied, and is applied to input nodes
Function S2	Enters a node in serial to the node that is applied - is used for hidden layer nodes
Function P2	Enters a node in parallel to the node that is applied, and is also used for hidden layer nodes. This mechanism is used to ensure that population individuals will include at least one input node
Function IN	Assigns a variable to the input node that it is applied
Function E	The operation of function E is to mark the end of the expansion of the network
Function LNK	Provides the framework for the application of cut function. It enables the non-full connectivity of the network, a feature that offers larger solution search space
Function CNR/rule	Performs the node inference. Based on the first parameter, the corresponding calculation is performed. The second parameter assists the calculation for the at-least-k and majority-of-k operators. It can process any real valued variables
Function NUM	Returns an integer in the interval [1, 256] to be used by the calling LNK function
Function CUT	Returns an integer in the interval [0, 1] to be used by the calling LNK function. If the returned value is 1, then the link will be ignored in the calculations
Function CNRSEL	Returns an integer in the interval [0, 8] to be used as first parameter of the CNR
Function K	Returns an integer in the interval [1, 256] to be used by the CNR function, if CNRSEL returns 3, 4 or 6 (corresponding to the calculation of the at least k-true, at least k-false and majority of k functions)
Solution obtained for breast cancer diagnosis	CNLN (P1 (P1 (IN T3) (S1 (IN T7) (RULE 0 0) E)) (P1 (P1 (P1 (P1 (P1 (IN T1) (S1 (IN T5) (LINK 257 2 (RULE 6 6)) E)) (S1 (S1 (IN T2) (LINK 115 0 (LINK 179 0 (RULE 0 0))) E) (LINK 115 0 (RULE 0 0)) E)) (S1 (IN T6) (RULE 0 0) E)) (P1 (S1 (IN T1) (RULE 0 0) (S2 E (RULE 0 0) (S2 E (RULE 0 0) E))) (S1 (P1 (S1 (IN T2) (RULE 0 0) (P2 (S2 E (RULE 0 0) E) (RULE 0 0) E)) (P1 (IN T8) (S1 (IN T6) (RULE 6 6) E))) (RULE 0 0) (P2 E (RULE 0 0) (S2 (S2 E (RULE 2 4) E) (RULE 0 0) E))))) (P1 (P1 (IN T3) (S1 (IN T7) (RULE 0 0) E)) (P1 (P1 (IN T1) (IN T7)) (S1 (IN T2) (LINK 115 0 (RULE 0 0)) E))))) (RULE 6 6))

almost 19 out of 20 new cases arriving to a hospital with the suspicion of breast cancer. Input information for diagnosing each case, will have to be the cell measurements T1 to T9. Knowing that the diagnostic capability of a medical expert exceeds 95%, the second opinion provided by the ENLN system, if agrees with the opinion of the expert, in fact ensures a secure diagnosis. On the other hand, when disagreement exists between the expert and the system, more attention should be paid in the case of under diagnosis.

Cell characteristics T1, T2 and T7 (clump thickness, uniformity of cell size and bland chromatin) appear more often in the entire rule set and thus, seem to be the very important for differentiating among healthy and cancerous cases, the same as T3 and T6 (uniformity of cell shape and bare nuclei). Characteristics T5 and T8 (single epithelial cell size and normal nucleoli) appear only in one decision path of the produced rule set and seem to play a secondary role in the diagnostic process. Nevertheless, after analysing more carefully the entire rule set, we arrived at the conclusion that attribute T3 (i.e. cell characteristic denoted as 'uniformity of cell shape'), proves to be the most important of all in extracting the diagnosis, due to its position and influence in most diagnostic rule paths (for more details in reading and understanding the produced ENLN decision rules, see ref. 24). The finding agrees with the related literature, as lesions that consist of uniform or relatively uniform cells are lesions with good differentiation and generally display a low invasive potential. Such examples are cribriform and micropappilary DCIS. Lesions which contain cells that are not uniform are more poorly differentiated such as comedo-carcinoma.

The most interesting result drawn from the application of ENLN on the breast tumor data-set, seems to be the exception of two features, the 'marginal adhesion' (T4) and the 'mitoses'

(T9), which are not included in the list of features having highly diagnostic value. Regarding the exclusion of the 'marginal adhesion' feature, this corresponds to the already known medical knowledge that there is a macromolecule on the surface of the cells, which is only found on intraductal carcinoma cells (IDC). Given the fact that 90% of intralobular carcinoma cells (ILC) and also all normal cells do not exhibit this feature, marginal adhesion cannot be used to differentiate between benign and malignant cells, since most of the lobular cancers do not exhibit this feature. Regarding the exclusion of the 'mitoses' feature, both, benign and malignant tumors display a degree of cell proliferation, which is higher and atypical in malignant lesions as it is known in literature. However, low-grade malignant lesions such as DCIS grade I and II, display a similar rate of cell proliferation to certain benign conditions, such as sclerosing adenosis and atypical ductal hyperplasia, thus making it difficult to discriminate between benign and malignant tumors using this feature.

Collaborative expert medical staff that studied and analyzed the ENLN results from the medical viewpoint, stated that the phase of preprocessing the data might be of major importance, as it involves human intelligence and expertise in the proper 'case by case' modeling of the problem under consideration. Thus, such a method that combines human and machine attitudes might be superior to other competitive automated techniques for image analysis, signal processing, etc.

Concluding, the ENLN methodology can effectively model and automate the diagnostic mechanism for large, vague, complicated domains of medical interest such as breast cancer diagnosis, based on data extracted from available past medical records. ENLN consist an efficient generalization methodology as they are trained, validated and tested in both, known and unknown cases, avoiding the danger of over-fitting on the training data set. Furthermore, the proposed methodology can uncover possible relations existing within the data set, performing as an advanced pre-statistic knowledge discovery process. Future work includes an additional breast cancer data collection, for the further statistical investigation of the hypothesis for the major importance: a) primarily of clump thickness, uniformity of cell size and bland chromatin, and b) secondarily of uniformity of cell shape and bare nuclei, regarding breast cancer diagnosis from cell characteristics. Data regarding breast tumor malignancy diagnosis used in this study were taken from Hettich *et al* (26).

References

1. Tsakonas A, Aggelis V, Karkazis I and Dounias G: An evolutionary system for neural logic networks using genetic programming and indirect encoding. J Appl Logic 2: 349-379, 2004.
2. Mangasarian OL and Wolberg WH: Cancer diagnosis via linear programming. SIAM News 23: 1-18, 1990.
3. Mangasarian OL, Setiono R and Wolberg WH: Pattern recognition via linear programming: theory and application to medical diagnosis. In: Large-Scale Numerical Optimization. Coleman TF and Li Y (eds). SIAM Publ., Philadelphia, pp22-30, 1999.
4. Wolberg WH and Mangasarian OL: Multi-surface method of pattern separation for medical diagnosis applied to breast cytology. Proc Natl Acad Sci USA 87: 9193-9196, 1990.
5. Bennett KP and Mangasarian OL: Robust linear programming discrimination of two linearly inseparable sets. Optimization Methods and Software 1: 23-34, 1992.
6. Zhang J: Selecting typical instances in instance-based learning. In: Proceedings of the 9th International Machine Learning Conference, Aberdeen, Scotland. Morgan Kaufmann, pp470-479, 1992.
7. Teh HH: Neural Logic Networks: A New Class of Neural Networks. World Scientific, 1995.
8. Tan AH and Teow LN: Inductive neural logic network and the SCM algorithm. Neurocomp 14: 157-176, 1997.
9. Chia HWK and Tan CL: Neural logic network learning using genetic programming. Int J Comput Intell Applications 1: 357-368, 2001.
10. Alba E, Cotta C and Troya JM: Evolutionary design of fuzzy logic controllers using strongly-typed GP. In: Proceedings IEEE Int. Symposium on Intelligent Control, NY, USA, pp127-132, 1996.
11. Tsakonas A and Dounias G: Hierarchical classification trees using type-constrained genetic programming. In: Proceedinfs of First Int. IEEE Symposium in Intelligent Systems, Varna, Bulgaria. IEEE Publications, 2002.
12. Tsakonas A, Dounias G, Axer H and von Keyserlingk DG: Data Classification using Fuzzy Rule-Based Systems represented as Genetic Programming Type-Constrained Trees. Proc UKCI-01, Edinbourgh, UK, pp162-168, 2001.
13. Tsakonas A and Dounias G: A scheme for the evolution of feed-forward neural networks using BNF-grammar driven genetic programming. Proceedings of Eunite-2002 Symposium, European Network of Excellence for Intelligent Technologies, Algarve, Portugal. Verlag-Mainz Publications, 2002.
14. Yu T and Bentley P: Methods to evolve legal phenotypes. Comp Sci 1498: 280-291, 1998.
15. Gruau F, Whitley D and Pyeatt L: A Comparison between cellular encoding and direct encoding for genetic neural networks. In: Genetic Programming 1996: Proceedings of the First Annual Conference, Cambridge, MA. Koza JR, Goldberg DE, Fogel DB and Riolo RL (eds). MIT Press, pp81-89, 1996.
16. Montana DJ: Strongly typed genetic programming. Evol Comput 3: 199-230, 1995.
17. Paterson N and Livesey M: Evolving caching algorithms in C by GP. In: Genetic Programming 1997. MIT Press, pp262-267, 1997.
18. Naur P: Revised report on the algorithmic language ALGOL 60. Communications ACM 6: 1-17, 1963.
19. Whigham P: Search bias, language bias and genetic programming. In: Genetic Programming 1996. MIT Press, pp230-237, 1996.
20. Gruau F: Neural Network Synthesis using Cellular Encoding and the Genetic Algorithm. PhD Thesis, Ecole Normale Superieure de Lyon anonymous ftp:lip.ens-lyon.fr (140.77.1.11) pub/Rapports/PhD PhD94-01-E.ps.Z.
21. Gruau F: On using syntactic constraints with genetic programming. In: Advances in Genetic Programming. Angeline PJ and Jinnear KE Jr (eds). MIT Press, 1996.
22. Tsakonas A and Dounias G: Decision making in the medical domain: comparing the effectiveness of GP-generated fuzzy intelligent structures. In: Proceedings of Eunite-2003 Symposium, European Network of Excellence for Intelligent Technologies, Oulu, Finland. Verlag-Mainz Publications, 2003.
23. Wong MI: A flexible knowledge discovery system using genetic programming and logic grammars. Decision Support Systems 31: 405-428, 2001.
24. Tsakonas A: Intelligent Methodologies for Decision Making in Complex Managerial and Financial Environments: Evolutionary Neural Logic Networks, PhD Thesis. University of the Aegean, Department of Financial and Management Engineering, Chios, Greece (in Greek), 2004.
25. Tsakonas A and Dounias G: Evolutionary neural logic networks in two medical decision tasks. In: Proceedings of Eunite-04 Symposium, European Network of Excellence for Intelligent Technologies, Aachen, Germany. Verlag-Mainz Publications, 2004.
26. Hettich, S, Blake CL and Merz CJ: UCI Repository of Machine Learning Databases. Department of Information and Computer Science, University of California, Irvine, CA (http://www.ics.uci.edu/~mlearn/MLRepository.html), 1998.