Jan A. Kors

Department of Medical Informatics,
Erasmus University, Rotterdam,
The Netherlands

Synopsis

Decision Support Systems and Knowledge Processing

This section of the Yearbook contains eight articles on Decision Support Systems and Knowledge Processing. Whereas in previous issues of the Yearbook separate sections were dedicated to Decision Support Systems and to Knowledge Processing, these sections have been merged in this year's edition. This is not unreasonable since the two fields of research have substantial overlap: knowledge processing techniques require construction of decision support systems to show their viability, and decision support systems require knowledge to be acquired, represented and processed to arrive at decisions. While knowledge processing issues are being addressed in all papers in this section, the main emphasis in most of them is on decision support aspects. These investigations are of an applied rather than a methodological nature.

Another observation is that four of the eight papers in this section have appeared in clinical journals, which may be taken as an indication of the continued interest of the medical community in decision support systems. It is a sobering thought, however, that none of the decision support systems presented here, and only a small minority of systems that have been reported on in the literature, are in routine clinical use today. Some of the factors that affect this unfortunate situation will be discussed later on.

The papers in this section encompass a variety of application areas, classification methods, and evaluation strategies. The decision support systems are concerned with diagnosis (of coronary heart disease, healed and acute myocardial infarction, oral and breast cancer) and prognosis (of kidney function and of survival for AIDS patients). To deliver their advice, they utilize case-based reasoning, statistical methods, artificial neural networks, and decision trees. The evaluation of system performance ranges from a simple assessment, using a limited set of cases and no comparison with human experts or other techniques, to more elaborate testing, in one case even assessing the performance of the system after its transfer to a clinical setting other than the one in which it had been developed.

Haddad et al. [1] describe an image processing system for the interpretation of myocardial perfusion scintigrams. The system is aimed at classifying a scintigram as being indicative for coronary heart disease (CHD) or not, in order to reduce interobserver variability and to enhance diagnostic accuracy. The authors employ a case-based reasoning method, comparing the scintigram at hand with those from a database of scintigraphic images that were validated by coronary angiography. A lot of effort is spent in finding the right input features, similarity metric, and adaptation strategy to adjust the likelihood of CAD. The authors consider the resultant system to be a prototype.

The method of case-based reasoning is closely connected to traditional nearest-neighbor methods. These methods have been shown to perform well in a wide variety of settings [2]. Apart from their performance, their simplicity and their ability to "explain" their decision by referring to similar cases are appealing. A potential drawback is the high computation time to classify a case when the image database is large, a problem the authors are well aware of but were not yet able to solve satisfactorily. The indexing scheme that the authors experimented with substantially degraded system performance.

The authors show that their system performs well as compared to two previous efforts in this field, but unfortunately do not compare the performance of their system with that of human experts. Whether their system offers a solution for the problems that motivated their study, the reduction of interobserver variability and enhancement of diagnostic accuracy in assessing CHD from myociardial scintigrams, thus remains unclear. In this respect, they show the feasibility of their approach, not its clinical utility.

To predict the kidney function of a patient in an intensive care unit for a period of several days, Schmidt et al. [3] also use a case-based reasoning approach. The difficulty of their application area is that the behavior of renal abnormalities over time is incompletely understood. Using a priori knowledge, the authors define a number of renal functional states and state transitions. The state is determined daily and a trend description over the past several days is generated. This trend description is then matched with a set of previously collected trends of longer duration, the continuation of the most similar trend serving as a prognosis for the trend under consideration. Since no adaptation is carried out after a similar trend has been retrieved, the approach essentially comes down to nearest neighbor classification. An interesting feature of the system is its ability to construct a set of prototypical trend descriptions by merging similar trends into prototypes. These prototypes in a sense substitute for the lack of medical knowledge in this field.

This work is an impressive endeavor to provide decision support for a very complex task. However, the practical usefulness of the system has still to be demonstrated. The abstracted renal functional states produced by the system were compared with those provided by human experts, but no mention is made of any validation of the system's trend prediction against the actual time course. Such an evaluation would seem essential to convince clinicians to use the system routinely.

Similar to the study of Haddad et al. [1], a substantial part of this study is devoted to proper definition and determination of features, underlining the maxim that good features are of paramount importance in the design of any classifier [4].

The remaining six papers in this section all deal with artificial neural networks (ANNs). The interest in this relatively new branch of classification methods is ever increasing, outside but also inside the medical world, as is reflected by introductory articles in major medical journals [5]. ANNs have proven to be accurate classifiers in a wide range of medical applications, but their value as a decision support tool is not unchallenged, mainly because of their inability to provide insight in their behavior [6].

Lo et al. [7] made an ANN to discriminate between invasive and in situ breast cancer. The ANN consists of 10 input and 15 hidden nodes, and has been trained on a set of 96 malignant cases which is a surprisingly low number considering the number of connection weights that had to be established. The authors envisage their present ANN to be used in conjunction with another, previously devised ANN that distinguishes between benign and malignant cases. The clinical relevance of this two-stage classification procedure is the possible reduction of the number of surgical biopsies. The authors claim that the ANN is able to accurately classify invasion, but they do not substantiate this claim by comparing its performance with that of radiologists or of other classification methods. Another remark is that the ANN was trained and tested on a set of malignant cases only, whereas the cases it has to classify in its intended use as a second stage classifier will also contain benign cases incorrectly classified as malignant by the first stage. Due to these cases, the performance of the ANN may be different from the one that is reported. The authors are aware of this caveat, but have not accounted for it in their evaluation.

Brickley et al. [8] trained different ANNs to discriminate normal from abnormal, and premalignant from malignant oral smears. Obtaining cytological smears is a less invasive procedure than biopsy, and an ANN could be used to differentiate between those patients for whom a conventional biopsy is appropriate and those who can go without it. Using only five input variables, including age and sex, about 80% sensitivity and specificity was obtained in discriminating normal from abnormal smears. It is difficult to put these figures into perspective since the performance of the ANNs is not compared with that of human experts or other classification methods. While the authors remark that a comparison with linear discriminant techniques would be valuable, they did not do so although these techniques are readily available in statistical packages. Whether an ANN is the best choice for integration into an image analysis system that would fully automatically interpret oral smears, as envisaged by the authors, would require further analysis.

Heden et al. [9] compare the performance of an ANN and a cardiologist in classifying old anterior myocardial infarction from the 12-lead ECG, taking a diagnosis based on ECG-independent material as the reference. They clearly describe the procedures and parameter settings necessary to train and test the ANN. One difficulty is that a cardiologist is used to express the likelihood of infarction by a set of qualitative terms, such as "possible" or "probable", whereas the ANN renders a continuous output between 0 and 1. The authors map this output to the same set of qualifiers as used by cardiologists, and show that the ANN has similar specificity as the cardiologist, but higher sensitivity.

Interestingly, the authors scrutinize those cases where ANN and cardiologist showed substantial disagreement with the ANN being incorrect. Several reasons for the erroneous behavior of the ANN are suggested, but it proves difficult to explain its behavior exactly.

Since computer interpretation of ECGs has a long standing record of

research, it would have been interesting to compare the ANN's performance with that of established interpretation programs. Eventually, ANNs, like other classification methods, should be subjected to rigid evaluation such as the one carried out in the CSE study [10], where multiple interpretation programs were evaluated by an independent center on a large test set of ECGs containing various ECG abnormalities.

The paper by Kennedy et al. [11] gives an excellent description of the design and evaluation of an ANN for the diagnosis of acute myocardial infarction. The resultant ANN is compared with linear discriminant analysis and also prospectively tested against the classifications of admitting physicians from a different center to that in which it was developed, showing the good performance of the ANN. The authors also discuss the problems associated with the practical use of the decision aid on an accident and emergency department. Of all the studies in this section, the evaluation of the decision support system in this study comes closest to the point that a clinical trial could be carried out to demonstrate its beneficial impact. In this sense, it may serve as an example of comprehensive evaluation for any clinical decision support system.

In the paper of Ohno-Machado [12], an ANN is constructed for survival prediction of patients living with AIDS. The predictive performance of the ANN is compared with that of a Cox proportional hazards model: no significant differences emerged. The Cox model has the advantage of providing insight into the variables that are most important for prognosis. Thus, the use of an ANN for this task would not seem recommended, but the author still sees opportunities for prognostic ANNs when some of the assumptions that underly parametric methods, particularly about data distributions, cannot be verified.

In the last paper of this section, Silver and Hurwitz [13] address an important issue that is sometimes overlooked in research on decision support systems: the ability of a system to provide insight in the acquired knowledge and to explain its advice. They compared two types of classifiers, inductive decision trees and ANNs, in the noninvasive assessment of coronary artery disease (CAD) based on clinical characteristics, nonimaging stress components, and scintigraphic findings. For each of four levels of increasing severity of CAD, they constructed a decision tree and an ANN which discriminate more severe levels from less severe. Not only did they compare both systems with respect to diagnostic performance, which proved to be similar, but also with respect to explanatory capability. Not unexpectedly, the ease of interpretation of the decision trees is found to be far superior to that of the ANNs. The authors rightly remark that other statistical methods can suffer from the same explanatory weakness. A further analysis of the trees revealed the importance of scintigraphic attributes in decision making, confirming the results of previous research.

When overviewing this year's selection of papers in the section Decision Support Systems and Knowlege Processing, it is striking that most systems are prototypes and far from being in routine clinical use. The system described by Kennedy et al. [11] comes close, but even in this case the authors deem a large-scale evaluation study necessary before its widespread use can be recommended. The reasons for this slow dissemination are probably manifold [14]. Here, just a few points are raised related to the present studies. All employ techniques that allow the construction of a classifier from a database of labelled cases, obviating a laborious and time-consuming knowledge acquisition process involving clinicians. This advantage is counteracted, however, if clinicians are only called in after the system has been developed. The clinical need for decision support may then turn out not to be as high as the system developers had expected, or the advice given by the system may not meet user requirements of performance and accountability. This latter aspect is of particular concern in view of the ubiquitous adoption of ANNs. If clinicians want insight into the advice provided by a decision support system, an ANN may not be the first choice, as was nicely illustrated by Silver and Hurwitz [13]. Moreover, the decision to use an ANN often seems more inspired by fashion than by careful consideration or comparison with other approaches. Especially when the system performance is not gauged against human experts or a gold standard, it will be very difficult to convince clinicians that they will benefit from the system's advice. Such a comparison, while only being a first step on the long way to user acceptance, appears to be a *sine qua non.*

References

1. Haddad M, Adlassnig KP, Porenta G. Feasibility analysis of a case-based reasoning system for automated detection of coronary heart disease from myocardial scintigrams. Artif Intell Med 1997;9:61-78.
2. Weiss SM, Kulikowski CA. Computer systems that learn. San Mateo: Morgan Kaufmann; 1991.
3. Schmidt R, Heindl B, Pollwein B, Gierl L. Multiparametric time course prognoses by means of case-based reasoning and abstractions of data and time. Med Inform 1997;22:237-50.
4. Duda R, Hart P. Pattern classification and scene analysis. New York: Wiley; 1973.
5. Cross SS, Harrison RF, Kennedy RL. Introduction to neural networks. Lancet 1995;346:1075-9.
6. Wyatt J. Nervous about artificial neural networks? Lancet 1995;346:1175-7.
7. Lo JY, Baker JA, Kornguth PJ, Iglehart JD,

Floyd CE. Predicting breast cancer invasion with artificial neural networks on the basis of mammographic features. Radiology 1997;203:159-63.
8. Brickley MR, Cowpe JG, Shepherd JP. Performance of a computer simulated neural network trained to categorise normal, premalignant and malignant oral smears. J Oral Pathol Med 1996;25:424-8.
9. Heden B, Ohlsson M, Rittner R, Pahlm O, Haisty WK, Peterson C, Edenbrandt L. Agreement between artificial neural networks and experienced electrocardiographer on electrocardiographic diagnosis of healed myocardial infarction. J Am Coll Cardiol 1996;28:1012-6.
10. Willems JL, Abreu-Lima C, Arnaud P, Van Bemmel JH, Brohet C, Degani R, et al. The diagnostic performance of computer programs for the interpretation of electrocardiograms. New Engl J Med 1991;325:1767-73.
11. Kennedy RL, Harrison RF, Burton AM, Fraser HS, Hamer WG, MacArthur D, McAllum R, Steedman DJ. An artificial neural network system for diagnosis of acute myocardial infarction (AMI) in the accident and emergency department: evaluation and comparison with serum myoglobin measurements. Comput Methods Prog Biomed 1997;52:93-103.
12. Ohno-Machado L. A comparison of Cox proportional hazards and artificial neural network models for medical prognosis. Comput Biol Med 1997;22:55-65.
13. Silver DL, Hurwitz GA. The predictive and explanatory power of inductive decision trees: a comparison with artificial neural networks learning as applied to the noninvasive diagnosis of coronary heart disease. J Investig Med 1997;45:99-108.
14. Heathfield HA, Wyatt J. Philosophies for the design and development of clinical decision-support systems. Meth Inform Med 1993;32:1-8.

Address of the author:
Jan A. Kors,
Department of Medical Informatics,
Faculty of Medicine and Health Sciences,
Erasmus University,
3000 DR Rotterdam, The Netherlands
e-mail: kors@mi.fgg.eur.nl