An Experiment to Determine Improvements in Automated Problem Solving in a Complex Problem Domain

M. Van Dyne
Dept. of Electrical Engr. and Computer Science
University of Kansas, Lawrence, KS 66045
vandyne@eecs.ukans.edu

C. Tsatsoulis
Dept. of Electrical Engr. and Computer Science
University of Kansas, Lawrence, KS 66045
tsatsoul@eecs.ukans.edu

Abstract

A previously constructed prototype expert system was extended to include case-based reasoning/learning, in order to determine if the automated problem solving behavior could be improved. The initial expert system was developed by using an inductive machine learning technique on 9,445 data records of pregnant women, providing production rules to predict preterm delivery. Its predictive accuracy was tested on a separate set of 9,445 data records. Next, the capability to reason from both production rules and input test cases was added to the system, in addition to the capability to internally modify its confidence in each piece of knowledge (rule or case) and the relative importance of patient attributes which appear to be predictive of preterm delivery. The system was structured such that the accuracy of either type of reasoning could be measured individually to determine how rule-based and case-based reasoning perform alone, and to determine how they perform together. Results show that the predictive accuracy of the system was improved, with different trends emerging, dependent on the bias of the learning data. Neither system performed as well alone as did both together.

Introduction

The current investigation focuses on studying the problem solving performance of induced rules versus case based reasoning in the complex problem domain of predicting preterm delivery in pregnant women. The predictive performance of rules alone was tested, then cases alone, and finally a combination of both rules and cases. Problem solving performance in each circumstance was measured and the conditions under which different types of reasoning performed better was established.

Problem Domain

Pregnancy is considered fullterm at 40 weeks gestation, however, 37 weeks is generally used as the criterion to determine whether a delivery is preterm or not. Accurate identification of pregnant women who are high-risk for preterm birth is important in determining which women will benefit from interventions designed to prolong gestation. Prolonging gestation can result in significant improvements in infant survival and reduced costs of neonatal intensive care (McLean, Walters, & Smith, 1993).

What makes the problem domain complex is that it is not clear what symptoms may be predictive of preterm delivery, and consequently, during the course of prenatal care, hundreds of data items may be collected. Furthermore, different data is collected at different facilities. From all this data, the healthcare provider must make decisions concerning preterm birth risk, and any plans for intervention (NIH Guide, 1992). Currently used manual screening tools have been estimated to be between 17 - 38 % predictive in determining preterm risk (McLean, Walters, & Smith, 1993).

Prototype Expert System

A prototype expert system was built in a previous effort (Van Dyne, et.al., 1994; Woolery, et.al., 1994) which used production rules generated by an inductive machine learning technique, LERS (Grzymala-Busse, 1988; 1989; 1991), in order to generate a knowledge base. A retrospective sample of high risk pregnant women was obtained from three databases containing 18,890 subjects from one local and two national sources. The databases were split into two equal halves, and one half of each database was used to generate production rules while the other half was used to test the accuracy of the prototype expert system. Accuracy was measured as the percentage of times the prototype's predicted outcome matched the actual patient outcome.

Each of the three databases contains a different, although overlapping, set of attributes for patients. The attributes from each database were combined into a single patient object so that rules generated from any of the learning databases could be run on examples from any of the test databases. Because of poor predictive performance of rules generated from the third database, only rules generated from the first two databases were included in the prototype expert system. This resulted in 520 rules in the prototype system. Using measurements of rule and rule base effectiveness generated by LERS, in addition to information on how many examples were used to learn the rule, a prioritization scheme, was developed for each rule in the system.

Current Investigation

The motivation for the current investigation is that the prototype expert system is static; that is, whenever a new patient record is entered, the prediction process will always follow the same reasoning, even if that record has been entered be-

fore and the prediction was incorrect. Unlike human performance, no learning takes place. If the predictive accuracy of the rule base was perfect, this would not be a concern, but the accuracy of the expert system ranges only from 51 - 88% correct. This is an improvement over current manual risk assessment techniques, but the question arises, can the system learn from its mistakes, and improve upon its predictive accuracy.

In using a case based approach in this situation, the existing rule base can be viewed as domain knowledge, albeit, faulty domain knowledge. Each rule can also be thought of as a generalized, or abstracted, case because it was generated inductively from a set of actual examples. Since each rule in the system already has a confidence associated with it, and the system will use the rule with the highest confidence that applies in a given situation, the confidence in that rule can be modified according to its success or failure on a case. Furthermore, cases can be added to the system to begin to fill in areas where the rule base is lacking. Each case may be thought of as a specific rule with values provided for each attribute in the antecedent.

Unlike a rule, however, the attribute/value pairs in an input case do not have to exactly match the attribute/value pairs in a case. In a rule, all antecedent attribute value pairs must match before the rule becomes eligible to fire. The same confidence rating scheme used for the rules can also be used for the cases. Furthermore, as attributes tend to become associated more with the success or failure of a prediction, the strength or predictive importance of these attributes can be modified so that future case matches pay more attention to those attributes that are more predictive.

Hybrid System Operation and Architecture

On starting the system, the user is allowed to choose the database from which to select records, and the number of records to be run. Depending on the database specified by the user, one or more database records is read into the system from one of three databases. Since each record structure is different, the first thing the system does is map the attributes and their values into a composite record structure. The attribute/value pairs in this record structure are used to match past cases in the system (if any) and to determine the rule with the highest confidence associated with it.

As records are entered into the system, they are added as cases to the case base with a confidence level just below that of the average rule. The system makes an outcome prediction using both the best matched case and the highest confidence rule. The piece of knowledge (rule or case) that has the higher confidence level is chosen for the overall system prediction, but predictive success of each candidate piece of knowledge is also measured so that comparisons between the techniques can be made. If the piece of knowledge failed to predict correctly, its confidence is decremented, and if successful, its confidence is incremented, in a scheme consistent with the initial confidence rating of the rules.

The attributes or features used in making the prediction are also evaluated. Each time a feature value matches in successfully predicting an outcome, its weight is incremented, and each time it matches, but is unsuccessful, its weight is

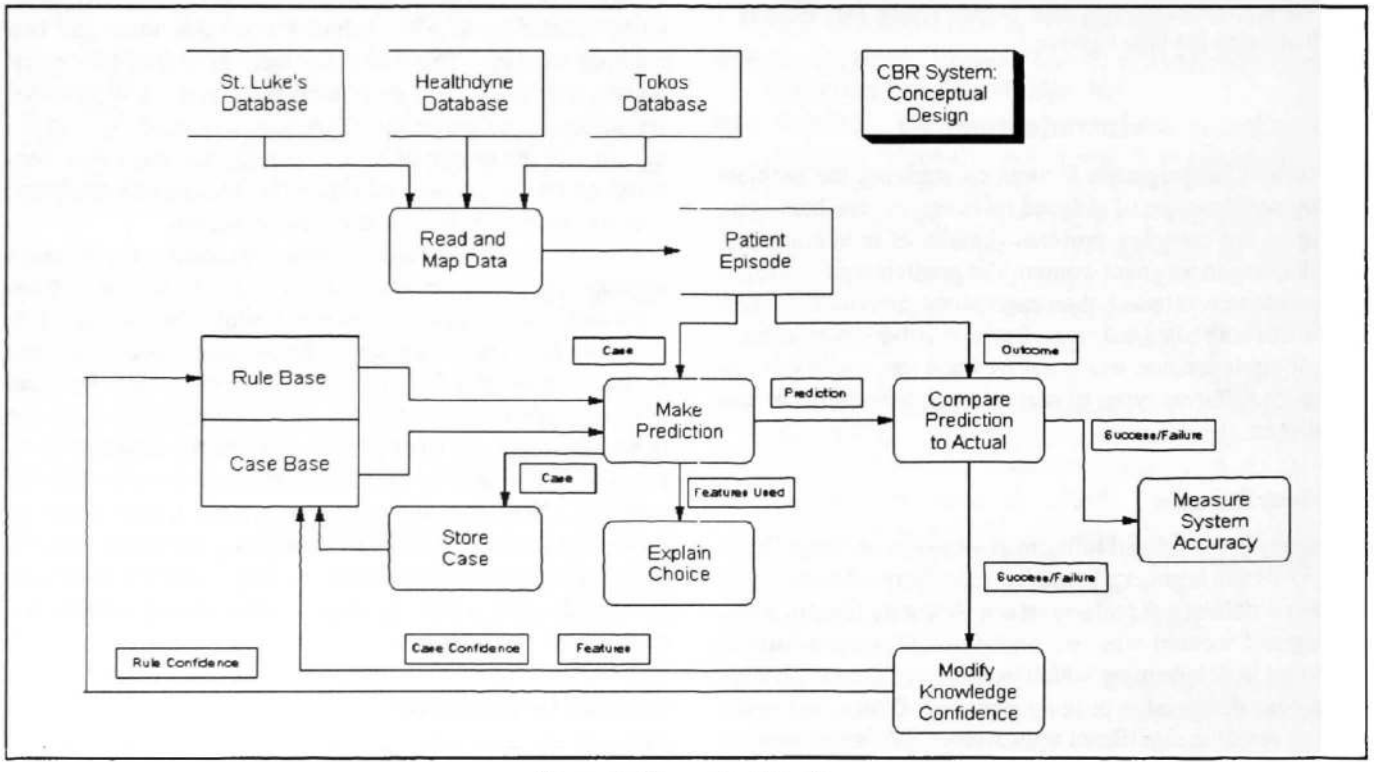

Figure 1: Conceptual Design

decremented. Features used by both rules and cases are treated in the same manner. The intention behind this scheme is to begin to weight each attribute according to its predictive value of the outcome, and thus allow the system to build an indexing scheme as its experience increases. Case matching is done by determining which features of the input case have the same value as features of cases in the case base, and producing a match score according to the weight of each matching feature. Presumably, then, case matches will become more appropriate as predictive features are weighted more strongly than if a simple feature count match is used.

Figure 1 shows the conceptual design of the experimental hybrid reasoning system.

Case Base

The structure of the case base in this system is flat; that is, no generalization or abstraction of cases is performed. The composite object structure into which the input record was mapped is the same structure used to store cases. There are 48 attributes in the object, and 7 of these allow multiple values. Also, an input record will not have values to fill all of the attributes, or object slots, since the structure embodies different attributes from different databases.

Case Retrieval

Cases are chosen according to the best match score between the input record and cases in the case base. Every case in the case base is considered each time a new case is entered. The best match score is determined by the number of feature values in a case that match the input record values, and the weight of those features. Therefore, if a few "important" features match, it will be considered to be a better match than if many "unimportant" features match. Because the different databases include different attributes, attributes in either the case or the input record may not have values, and the case matching proceeds despite these missing values.

Indexing

The variable weighting of features depending on their association with successful predictions can be viewed as building indices over time. However, all features are still considered in the matching process, so that the feature weighting can be continually refined.

Learning

There are three areas in which the system learns to improve its own performance. The first of these is modification of rule confidence depending on the success or failure of a rule in a given case. The second is the modification of case confidence, again depending on it predictive success. These two types of learning occur whenever a rule or case is considered, not just when the actual prediction from that rule or case is used as the system prediction. The third area of learning is discriminating the more important or predictive features from the less important features so that more appropriate case matches can be made as the system's experience level increases.

Procedure

Tests were run using both the original expert system prototype and the current hybrid system to determine differences in problem solving performance. Because each database represents a different population, the systems were tested separately on 300 records from each of the three test databases. Results of the expert system were first recorded, then results of running the hybrid system were recorded as rules only, cases only, and the combination of rules and cases.

About 2/3 of the rules in the expert system originate from the first database. It represents a local population at higher risk for preterm delivery than the general population because it operates as a referral center for patients experiencing problems in pregnancy (not necessarily preterm delivery problems). The percentage of women who deliver preterm in this population is approximately 25%. The second set of data originated from a home uterine monitoring company, and the additional 1/3 of the rules in the system were induced from it. This data represents a nationwide sample of very high risk women, as only high risk patients (all preterm delivery) are referred for this service. The population represented in this data delivers preterm approximately 73% of the time. The final database was provided by another home uterine monitoring company. Again, the population is considered very high risk for preterm delivery, although the preterm delivery rate is lower than the previous database; approximately 66% of the patients deliver preterm in this population.

False positive and false negative predictions were measured in addition to overall system predictive accuracy to determine which problem solving method performed better under each condition. False positive predictions occur when the prototype system predicts preterm delivery and the patient actually delivered fullterm. False negative predictions occur when the prototype system predicts fullterm delivery and the patient actually delivered preterm.

Results

The first test condition was with the prototype expert system, using only rules as a means of prediction. These results establish the baseline accuracy to be exceeded by the hybrid system.

Table 1: Prototype expert system; database 1

	Number of Cases	Correct	Mis-classified	Un-classified
Overall	300	274 (91%)	25 (8%)	1 (0%)
Fullterm	255	242 (94%)	12 (4%)	1 (0%)
Preterm	45	32 (71%)	13 (28%)	0 (0%)

Table 2: Prototype expert system; database 2

	Number of Cases	Correct	Mis-classified	Un-classified
Overall	300	180 (60%)	107 (35%)	13 (4%)
Fullterm	87	48 (55%)	37 (42%)	2 (2%)
Preterm	213	132 (61%)	70 (32%)	11 (5%)

Table 3: Prototype expert system; database 3

	Number of Cases	Correct	Mis-classified	Un-classified
Overall	300	149 (49%)	91 (30%)	60 (20%)
Fullterm	191	102 (53%)	45 (23%)	44 (23%)
Preterm	109	47 (43%)	46 (42%)	16 (14%)

In all three databases, the prototype expert system accuracy is better than the performance of manual risk scoring techniques. Of particular interest is the performance of the expert system on the third database. Because of the poor predictive performance of rules originally induced from this data, those rules were not included in the expert system. Nevertheless, the rules generated from the other two databases were able to provide better accuracy on this data than manual techniques provide.

The next set of results are from running the hybrid system, measuring the predictive results of rules only and cases only. Note that the accuracy rates of rule performance is expected to differ from the prototype expert system because the hybrid system modifies rule confidence, thus influencing which rules will fire for a given record.

Table 4: Rule-only vs. case only performance; database 1

	Correct	Fullterm Misclassified	Preterm Misclassified
Rules Only	274 (91%)	4 (2%)	22 (49%)
Cases Only	240 (80%)	35 (14%)	25 (56%)

Table 5: Rule-only vs. case only performance; database 2

	Correct	Fullterm Misclassified	Preterm Misclassified
Rules Only	192 (64%)	52 (60%)	56 (26%)
Cases Only	199 (66%)	70 (80%)	31 (15%)

Table 6: Rule-only vs. case only performance; database 3

	Correct	Fullterm Misclassified	Preterm Misclassified
Rules Only	183 (61%)	63 (33%)	54 (50%)
Cases Only	197 (66%)	40 (21%)	63 (58%)

In all three database tests, the rule performance met or exceeded its original accuracy. This increase in accuracy can only be due to the dynamic confidence modification of each rule based on its predictive success. The proportion of incorrect predictions between false positives and false negatives did not remain the same, however. In general, predictive accuracy increased for the majority portion of the database population; for example, there were more fullterm than preterm records in the first database, and the number of incorrect predictions in this category decreased, at the expense of predictive accuracy for preterm records. This trend was reversed in the second database, where the population was represented by more preterm than fullterm patients, and accuracy improved in both categories in the third database.

Case only performance was not as good as rule only performance in the first database, where the predictive accuracy from rules was already very high. In both the second and third databases, however, case only performance exceeded that of rule only performance. Performance by category echoed the population distribution, as it did with the rule-only condition. Note that the rule condition was not always capable of providing a prediction for a given record, and these are listed as "unclassified" in the prototype system results table. Cases were always able to provide a prediction, so there were no unclassified records. For comparison to case performance, unclassified and misclassified records are included together in the rule only condition.

Finally, hybrid system results in which the system itself decided whether to use a case or a rule as the basis for prediction are shown below.

Table 7: Hybrid reasoning system; database 1

	Number of Cases	Correct	Mis-classified	Un-classified
Overall	300	275 (91%)	25 (8%)	0 (0%)
Fullterm	255	251 (98%)	4 (1%)	0 (0%)
Preterm	45	24 (53%)	21 (46%)	0 (0%)

Table 8: Hybrid reasoning system; database 2

	Number of Cases	Correct	Mis-classified	Un-classified
Overall	300	212 (70%)	88 (29%)	0 (0%)
Fullterm	87	35 (40%)	52 (59%)	0 (0%)
Preterm	213	177 (83%)	36 (16%)	0 (0%)

Table 9: Hybrid reasoning system; database 3

	Number of Cases	Correct	Mis-classified	Un-classified
Overall	300	215 (71%)	85 (28%)	0 (0%)
Fullterm	191	167 (87%)	24 (12%)	0 (0%)
Preterm	109	48 (44%)	61 (55%)	0 (0%)

From these results, it is apparent that using a combination of both rules and cases increased system accuracy over using either one individually, and over using the original prototype expert system. In the first database, only one additional case was correctly predicted, but in the second database, the hybrid reasoning system performed 10 percentage points better than did the expert system. Furthermore, while in both runs, the predictive performance was better on preterm deliveries, it is much better in the hybrid system. The hybrid system produced no unclassified predictions. On the third database, the hybrid reasoning system again outperformed the expert system, this time by 22 percentage points. Note that the sample of data used for testing in this case contained more full term examples than preterm, although the overall population in this database has the reverse bias. Interestingly enough, the hybrid system had an overall accuracy rate on this database that was higher than the accuracy rate on the second database. The performance of the hybrid reasoning system on this data may be explained by the fact that the expert system rules were not generated from this database, therefore, the cases provided a better source of prediction than did the rules.

Conclusions

The most prominent result of this experiment is that using cases can improve problem solving performance over the use of rules. This was evident in both the cases-only condition and the hybrid system condition. In a study of how people use analogues to make decisions, Klein and Calderwood (1988) reported that people found specific instances to be more helpful than generalized knowledge. Perhaps this is because the specificity of a case can match the current situation better, and thus decision making can be more accurate, providing one explanation for the improvement in performance after case inclusion. In her summary of a panel discussion on analogical reasoning, Seifert (1989) pointed out that while we tend to think that goals may be important in episode retrieval, people tend to rely more on surface features. The hybrid system did not address goals in selecting cases or rules, but still exhibited very good accuracy in prediction.

Riesbeck and Schank (1989) describe two categorizations of experience that are relevant here: ossified cases and paradigmatic cases. Ossified cases look like rules because they have been abstracted from a number of cases, much like the inductive generation of the rules for this system. Paradigmatic cases are those that are complete memories and serve as unique examples in the domain. Through the use of confidence modification in this system, those cases that are paradigmatic of predictive circumstances tend to rise to the top of the knowledge base.

The dynamic knowledge confidence modification of the system also appears to contribute to the improved performance, as indicated by the accuracy improvement of the rules-only condition versus the original expert system. According to Schank (1982), memory is dynamic, and the same input experiences at two different times can result in different remindings simply because memory may have reorganized itself in the interim. While the hybrid system does not reorganize itself in the same manner as Schank's description of MOPs and TOPs, it does reorganize its confidence in pieces of knowledge, and thus which knowledge will be used to make the current prediction.

A basic tenet of Schank's is that "we understand in terms of what we already understood" (Schank, 1982). While the initial rule base in the system contained approximately 50% preterm delivery rules and 50% fullterm delivery rules, when cases were added as pieces of knowledge, the system's knowledge began to reflect population biases. This may account for the preferential accuracy improvement exhibited by the system toward majority categories as represented in the popu-

lation.

Dynamically modifying the predictive weight of features in the system provided a means of focusing on the more important aspects of a given case. Barsalou also provides support for this concept, via the idea of "selective attention" (Barsalou, 1993). He states, "The basic idea is that the attentional system is capable of focusing strategic processing on various aspects of a perceptual experience and extracting them as individual components, while simultaneously tuning out other components to a large extent." Furthermore, Barsalou indicates that memories are likely to be retrieved based on three factors: frequency, recency, and context. The confidence modification of rules and cases addresses the frequency aspect of case retrieval, while the feature weighting partially addresses context. An interesting extension to the system would address the recency aspect, in which pieces of knowledge that are not used gradually "atrophy" in confidence, and if not used for a sufficiently long time, are eventually forgotten.

Overall, then, the results of this study indicate that the problem solving performance of the system was dramatically improved by the inclusion of cases with which to reason rather than the only rules.

Acknowledgments

The authors of this paper would like to thank St. Luke's Perinatal Center, Healthdyne, and Tokos for providing data for the study. The initial prototype expert system development was funded by a Small Business Innovation Research (SBIR) grant to IntelliDyne, Inc., through the National Institutes of Health, National Center for Nursing Research, grant number 1 R43 NR02899-01A1. The case based learning extension was completed through the University of Kansas, Department of Electrical Engineering and Computer Science.

References

Barsalou, L.W. (1993). Flexibility, structure, and linguistic vagary in concepts: Manifestations of a compositional system of perceptual symbols. In A.F. Collins, S.E. Gathercole, M.A. Conway, & P.E. Morris (Eds.), *Theories of Memory* (pp. 29-101). Hillsdale, NJ: Lawrence Erlbaum Associates.

Grzymala-Busse, J. (1988). Knowledge acquisition under uncertainty - a rough set approach. *Journal of Intelligent and Robotic Systems*. 3-16.

Grzymala-Busse, J. (1989). An overview of the LERS1 learning system. In *Proceedings of the Second International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, 2, (pp. 838-844).

Grzymala-Busse, J. (1991). *Managing Uncertainty in Expert Systems*. Boston: Kluwer Academic Publishing.

Klein, G.A. & Calderwood, R. (1988). How do people use analogues to make decisions?. In *Proceedings of the DARPA Case-Based Reasoning Workshop*, (pp. 209-223).

McLean, M., Walters, W., & Smith, R. (1993). Prediction and early diagnosis of preterm labor: a critical review. *Obstetrical and Gynecological Survey*. 48 (4), 209-225.

NIH (National Institutes of Health) (1992). *Guide to Grants and Contracts*. July 10, 8.

Riesbeck, C. & Schank, R. (1989). *Inside Case-Based Reasoning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schank, R. (1982). *Dynamic Memory: A Theory of Learning in Computers and People*. Cambridge University Press.

Seifert, C. M. (1989). Analogy and case-based reasoning. In *Proceedings of the DARPA Case-Based Reasoning Workshop*, (pp. 125-129).

Van Dyne, M., Woolery, L., Grzymala-Busse, J. & Tsatsoulis, C. (1994). Using machine learning and expert systems to predict preterm delivery in pregnant women. In *Proceedings of the Tenth IEEE Conference on Artificial Intelligence Applications '94*, (pp. 344-350), Los Alamitas, CA: IEEE Computer Society Press.

Woolery, L., Van Dyne, M., Grzymala-Busse, J. & Tsatsoulis, C. (in press). Machine learning for development of an expert system to support assessment of preterm labor risk. *Nursing Informatics '94*.