Chapter 6

Summary and conclusion

Recall from Section 1.1 that this study investigated the use of a KDD environment in a real-world scenario, focusing on the following:

i. The development of an intelligent data analysis tool, modelled as a MAL system that combines more than one data mining technique into a unified framework for decision support.

ii. The evaluation of the capability of this intelligent data analysis tool, using co-operative inductive learning techniques, in analysing the context as embedded in qualitative data, to be used for decision-making.

This was addressed by the construction of an intelligent data analysis tool based on the philosophy of co-operative learning, applied to machine learning techniques as well as a human expert. The intelligent data analysis tool consisted of co-operative inductive learner teams operating in an environment modelled as a multi-agent learning system. Three different teams were constructed namely, a machine learner team, consisting of three inductive machine learners; a human learner team, consisting of a human expert and of a synthesis report written by experts; and lastly, a machine, human learner team combination, consisting of two inductive machine learners and a human expert.

During the study the human expert changed roles, from that of a traditional onlooker to an active participant within the system. This study focussed on the qualitative aspects of the data describing the problem domain compared to the traditional quantitative aspects. The human learner is an expert within this domain with background knowledge pertaining to the problem. To ensure that this knowledge is incorporated within the learning process the human learner had to actively participate in it. By this active participation the CILT-MAL system as a whole, as well as the individual machine learners benefited as follows. For the classification task, the human learner and machine learners participated in separate teams. The results of the classification task showed that the final knowledge base created by the fusion of the human and machine learner team's rule sets obtained the highest overall rule set accuracy. This highlighted the success of human-machine collaboration by the active participation of the human in the learning process. For the problem solving type task, the human and machine learners participated within the same team, hence they had access to each other's knowledge bases and were able to benefit from each other in a more direct way during the co-operative learning episode.

The success of the intelligent data analysis tool was determined by its ability to address two types of tasks within a real-world scenario namely, a classification task and a problem solving type task.

6.1 Summary

This thesis was organised as follows:

- Chapter 1, 2 and 3 formed part of the theoretical basis of the work presented. Chapter 1 introduced the aim of the study and outlined the research approach used during the study.
- Chapter 2 concerned multi-agent learning by means of co-operation. This chapter described co-operative inductive learning and hence proceeded to define co-operative inductive learner teams. Co-operative inductive learner teams were then modelled as a CILT-MAL system. Within the CILT-MAL system learning agents

were described using the inductive machine learner architecture. A new architecture, namely, Langley's machine learning framework, was introduced. This framework was used to better define the learning mechanisms of the learners that participated in this study.

- Chapter 3 introduced the case study i.e. the National Research and Technology Audit. This chapter provided background on the different surveys that made up the audit and gave a summation of the findings. The chapter concluded by introducing the subjects for intelligent data analysis.
- Chapter 4 and 5 presented the experimental work done during the study. Chapter 4 presented the execution of the classification task namely, grouping organisations into business-clusters using two learner teams in a CILT-MAL system. The chapter concluded with a discussion of the knowledge discovery that occurred.
- Chapter 5 presented the execution of the problem solving type task namely, discovering trends in the human resources for research and technology using a combined machine/human learner team in a CILT-MAL system and concluded with a discussion of the knowledge discovery that occurred.

6.2 Concluding remarks

Experimental results showed that the CILT-MAL system was able to successfully address both types of performance tasks, namely, a classification as well as problem solving type task. This study created a new framework for learning as it occurs during a problem solving type task, within the CILT-MAL system.

A number of the findings supported the conclusions of the human experts, as presented in the synthesis report. However, a number of findings contradicted their opinions, making it a valuable learning experience to both the intelligent data analysis tool as well as the human experts. This study showed how intelligent data analysis could enrich the conclusions made by the human experts, assisting them to gain confidence in their own areas of expertise, but also indicating their weaknesses.

Fusing the knowledge bases of the two inductive learner teams during the classification task showed the power of co-operation. This fused knowledge base obtained the highest level of overall rule set accuracy compared to all the other knowledge bases. Having a human learner participate in the co-operative learning environment proved to be very valuable. Each time a human learner or human learner team was added to the CILT-MAL system a significant improvement in performance occurred.

Comparing the findings of the CILT-MAL system, as contained in the knowledge bases, to those of the synthesis report, two dominant conclusions can be made. Firstly, for the classification task, the human experts involved in producing the synthesis report, grouped the companies surveyed into two broad business-clusters, for data analysis purposes. The two clusters were defined as continuous process industries and discrete product industries. To adhere to these definitions, companies had to be categorised according to the characteristics of their major product lines. However, the knowledge as contained in the knowledge base showed that the groupings were made according to SICs and there is no indication that the sophistication level of the products played a role in the classification, as one would have expected from the definition of the two clusters.

Secondly, the findings in the synthesis report, as related to the problem solving type task, state that there is a major mismatch between the needs of the economy and the human resource skills that the system is producing. It is crucial that the needs of the economy and the training and development of human resources be better matched. However, the knowledge as contained in the knowledge base indicates not a mismatch but more a communication gap between the human resource delivery system and the needs of the economy. The human resource delivery system is producing according to the trends of the future knowledge needs of a competitive market system. This is not communicated to the business sector therefore; the efforts of the delivery system are not being appreciated. For example, the opinion that there is an overproduction of social scientists compared to the technology related human resources might be a dangerous generalisation of the current situation that is not in line with the prediction of the complexity of the economic problems of the future. Success in the future will be dependent on the ability of intelligent technology to interact with knowledgeable people, the latter depending on a strong social science involvement, which the market demand will rely on.

A number of inadequacies of the intelligent data analysis tool constructed for this study came to light, which should be addressed in future work: Firstly, the inability of the CN2 machine learner to represent input as a combination of conjunctive and disjunctive attribute tests within a rule. This restriction on knowledge representation was a major drawback for representing the human learner's knowledge. Secondly, the inflexibility of the performance measures generated by the CN2 evaluation function had a negative effect on the efficiency of the system. Different types of tasks require different performance measures. Therefore, evaluation functions used to address real world problems must be able to accommodate a variety of measures as required by the specific performance tasks.

The NRT Audit Data Warehouse contains a wealth of useful information that could be used by decision makers, when creating a policy framework for research and technology in South Africa. The CILT-MAL system approach to knowledge discovery in data may be the key to unlock this source of information.