Chapter 4

Classification task: Grouping organisations into continuous process and discrete product business-clusters.

The survey of the technology base of the South African business sector, conducted as part of the NRT Audit, surveyed the technology base of large, medium and small business organisations in South Africa. The survey determined the extent to which private firms and public corporations within the 17 technology-driven industry sectors depend on technology and related competencies. To increase the coverage of the industry sectors the experts involved in producing the synthesis report grouped the companies surveyed into two broad business-clusters. This grouping became the classification task presented to the CILT-MAL system. The aim of the classification task presented in this chapter is to determine the applicability of a CILT-MAL system to a real-world scenario to classify organisations according to main commercial activities.

This chapter is organised as follows. Section 4.1 contains the task description, followed by Section 4.2 discussing the data pre-processing step. The initial exploratory results are introduced in Section 4.3. Section 4.4 introduces the experimental method and performance measures. This is followed by a description of the individual learning phase and co-operative learning episode of two of the co-operative inductive learning teams, as introduced in Section 2.4, namely the machine learner team and the human learner team. The individual and co-operative learning episodes of the machine learner team are described in Section 4.5 and that of the human learner team is described in Section 4.6. Episode five and six of the CILT learning process, namely, evaluation against the validation set and knowledge fusion

are discussed in Section 4.7, followed by a discussion of the results in Section 4.8. Section 4.9 concludes the chapter with a discussion of the success of the CILT-MAL system as a classifier in a real-world application.

4.1 Task description

Data supporting the survey were obtained from 313 economically significant companies by means of questionnaires and buy-in interviews. Representative sampling determined that 1 260 companies would constitute the broad sample for a fully representative picture. Only 313 of these companies were considered significant and were interviewed [DACST 1998]. The experts, involved in producing the synthesis report, grouped the 313 companies surveyed, within the 17 technology-driven business sectors, into two broader business-clusters, namely that of continuous processes and discrete products.

Continuous process companies are those organisations that produce commodities that are measured in units of volume, mass, length and area for example gold, water or chemicals. These products have well established international benchmark prices and clear product specifications. Discrete product clustered companies, on the other hand, are those organisations that sell their output as individual items e.g. computers, cars or televisions. Discrete products are complex to manufacture and are produced as an assembly of many different components that are frequently manufactured from different materials. The aim of this learning task was to classify the surveyed organisations into these two broadly defined business-clusters using the data collected during the survey, as stored in the NRT Audit Data Warehouse.

4.2 Data pre-processing

The data collected from the 313 organisations that participated in the survey were stored in the data structure presented in Figure 7.

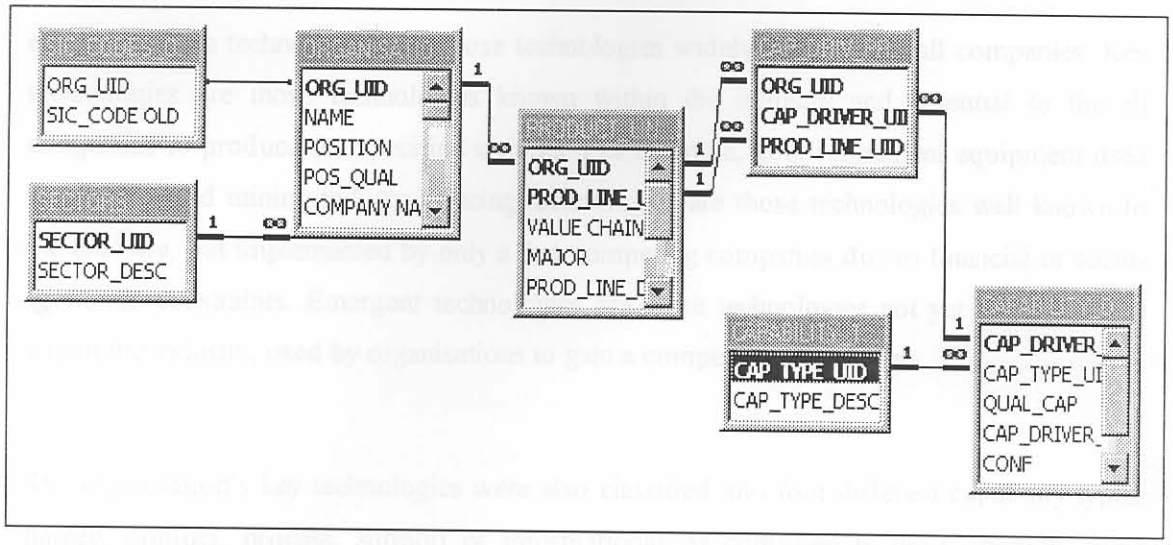

Figure 7 Business survey entity relationship diagram

The organisations as contained in the *Organisation* entity, were described in terms of two main aspects, as shown in Figure 7. Firstly, the key product lines produced by an organisation, as contained on the *Product_Line* entity, were considered with respect to their position in the value chain hierarchy. These positions denoted the level of sophistication of a product i.e. whether it was a "raw" material, component, sub-product, product, product system or a user system. For example, consider the value chain of a computer network. An organisation producing user systems, will supply customers with a network of computers including all the peripherals e.g. modems, printers, scanners, CD writers, etc, as well as the software and user-training. An organisation producing product systems sell packages, consisting of personal computers and the required software, which form part of the user system. Organisations producing products will, for example, produce printer cartridges used by the printers that form part of the user system. Organisations producing motherboards produce sub-products for personal computers. These sub-products are made up of components, which again are made of raw materials. The characteristics of each product produced by an organisation determine the product's position in the value chain hierarchy.

Secondly, organisations were characterised in terms of the key technologies (capability drivers) that make their products sustainable in the market, as contained on the *Cap_Driv_Prod_Line* entity. These key technologies were categorised under their technological stages of development. The four stages utilised were base, key, pacing and

emergent. Base technologies, are those technologies widely used across all companies. Key technologies are those technologies known within the industry and essential to the all companies to produce the specified outputs. For example, gold refinement equipment used within the gold mining industry. Pacing technologies are those technologies well known to the industry, but implemented by only a few competing companies due to financial or socio-economic constraints. Emergent technologies are those technologies not yet widely known within the industry, used by organisations to gain a competitive advantage

The organisation's key technologies were also classified into four different capability types, namely product, process, support or informational, as contained in the *Capability_Type* entity. Product capabilities are those technologies directly related to the performance of a product, e.g. chipboard-covering technology used by the pulp and paper industry. Process capabilities are those technologies directly related to the manufacturing process, e.g. pocketing machines used by the textile industries. Support capabilities are indirectly related to the manufacturing of a product e.g. paper quality measuring machinery used by the paper and pulp industries. Lastly, informational capabilities are those technologies directly associated with the gathering, storage and access of information related to the product, process or within any support function, e.g. a payroll system.

Fayyad *et al* (1996) made the statement that data in real world data repositories are always incomplete due to the fact that the data are usually collected in an ad hoc manner, which causes, for example, missing values. Also, mistakes are made during data entry, which results in poor data being captured. As a result KDD cannot succeed unless the data is first cleaned during the data pre-processing phase [Fayyad *et al* 1996]. For this classification task 64 of the 313 organisations were excluded from the final data set, mainly due to missing values. The next section documents this reduction process.

4.3 Initial exploratory results

The 313 organisations surveyed included 928 key product lines and 1048 capability drivers. Since an organisation can be described in terms of its product lines and capability drivers,

the first training set had an instance for every product line, capability driver combination, giving a total of 1664 instances. Preliminary analysis by the learners showed that the learners were grouping product lines, into organisations instead of business-clusters. Hence, the organisation identifier was removed from the instances and the data were presented to the learners for preliminary analysis. Using the new data, the learners were grouping product lines into business-clusters, which implied that an organisation could be classified under more than one business-cluster depending on the product lines it carried. However, the aim of the classification task was to classify an organisation as a single unit into a business-cluster, therefore, it was decided to summarise all the data pertaining to an organisation into one instance. This was done in the following manner.

Every instance carried the organisation's details concerning:

- the number of product lines produced by the organisation classified under the four different capability types, as defined in Section 4.2,
- the number of product lines produced by the organisation classified under the four technological stages of development, as described in Section 4.2,
- the number of product lines produced by the organisation per value hierarchy chain position (only the first three stages were included),
- the standard industry code (SIC) the organisation is classified under, and
- the business-cluster the organisation belongs to.

Table 4 Sample data from the business sector profile data set

Process Cap	Product Cap	Information Cap	Support Cap	Base	Key	Pacing	Emergent	Product	Product System	User System	SIC	Business Cluster
0	2	0	0	2	0	0	0	0	0	0	3660	Product
1	1	0	0	0	2	0	0	2	0	0	3321	Process
7	5	0	0	0	12	0	0	4	0	0	3353	Product
4	1	0	0	5	0	0	0	0	0	0	?	Product

During preliminary analysis by the individual learners the individual inductive learning episode was repeated at least 10 times, with 10 randomised training sets. Finally a data set, which included 249 instances of the original 313 organisations surveyed, was used for the

experiment. During the preliminary data analysis 64 of the instances were eliminated due to missing values. Only 257 of the 313 organisations had SIC's, five of the organisations did not have any capability drivers, seven had no technological stages of development identified and two had no value chain hierarchies. SIC was indicated as the most significant attribute during preliminary analysis and it was therefore decided not to include the incomplete instances.

4.4 Experimental method and evaluation criteria

Following Theron (1993), the training set contained 70% of the available instances and the test and validation sets each 15% respectively. A total of 167 of the 249 randomly chosen instances were included in the training set, 42 in the test set and the remaining 40 formed the validation set.

Four of the learners introduced in Chapter 2, namely CN2, C4.5, BRAINNE and the human learner, participated in the CILT-MAL system. For the classification task the learners were grouped into two teams namely, the human learner team and the machine learner team. The machine learner team included the three machine learners CN2, C4.5 and BRAINNE. The human learner team consisted of the human learner that participated in the NRT Audit supplemented by the findings as contained in the synthesis report. Co-operative inductive learning, within the teams of this multi-agent system, occurred as described in Section 2.4.

The performances of the learners were evaluated by the CN2 evaluation function. The performance measure was the accuracy by which a learner's rule set was able to predict the class of unseen instances i.e. the overall rule set accuracy, as defined in Section 2.5.4.1. The machine learners applied a reduced error-pruning scheme (REP), as discussed in Section 2.4.2 to prune their rule sets.

4.5 Learning of the classification task by the machine learner team

This section describes the individual and co-operative learning phases, as introduced in Section 2.4, of the machine learner team. Three inductive machine learners, CN2, C4.5 and BRAINNE participated in the machine learner team.

4.5.1 Individual learning phase

The individual learning phase introduced in Section 2.4.1, consisted of three episodes, the individual inductive learning episode, followed by the evaluation episode against the training set and finally, the evaluation episode against the test set. The three learners, CN2, C4.5 and BRAINNE received the full training set of 167 instances, as described in Section 4.3, and individually proceeded to execute their learning episodes independently of each other.

The parameters of the CN2, C4.5 and BRAINNE algorithms, as introduced in Section 2.5.4, 2.5.5 and 2.5.6, were assigned the following values according to the recommendations made by Theron (1993).

- CN2 algorithm: threshold = 0
  star size = 5
- C4.5 algorithm: weight = 2
  confidence = 5%
  redundancy = 0.5
- BRAINNE algorithm: epochs = 400
  step = 0.2
  hidden units = 6
  momentum = 0.7

After the individual inductive learning episodes were completed, the rule sets, generated by the learners, were evaluated using the CN2 evaluation function and the rules, together with their performance measures, were subsequently placed into their individual knowledge bases. Table 5 shows the overall rule set accuracies obtained by the learners during their individual learning phase. Step one lists the learners' overall rule set accuracies, as evaluated by their own unique evaluation functions, against the training set. Step two lists the rule set accuracies of the same rule sets, as evaluated by the CN2 evaluation function, against the training set. Lastly, step three lists the rule set accuracies of the rule sets, as evaluated by the CN2 evaluation function, against the test set.

Table 5 Accuracies and rule sets after the individual machine learner episodes

Step	Overall Rule Set Accuracy			Number of rules in KB		
	CN2	C4.5	BRAINNE	CN2	C4.5	BRAINNE
1	98.2%	97%	81.9%	14	11	5
2	98.2%	94.6%	73.7%	14	11	5
3	81%	83.3%	71.4%	14	11	5

The rule set produced by the C4.5 learner had the highest overall accuracy (83.3%), followed by CN2 (81%). The BRAINNE learner failed to find a highly accurate set of rules, and obtained the lowest overall rule set accuracy of 71.4%. The average rule set accuracy was 80.1% and the average accuracy of the individual rules were 61.2%. These averages were used as the performance threshold values. The evaluation of the individual rule sets showed that the learners learned accurate rules for the discrete product class, but failed to find accurate rules for the continuous process class. This was due to the fact that the training set contained more than twice the number of discrete product instances compared to the number of continuous process instances. This implies that the coverage of the discrete product class was twice as good as that of the continuous process class. The aim of the co-operative learning episode was therefore to improve the individual results and to produce sets of informative rules that describe the continuous process class.

4.5.2 Co-operative learning episode

Next the three learners initiated the full co-operative learning episode, as described in Section 2.4.2. Each team member queried the knowledge base of the other team members to find high quality rules that are related to their own low quality rules. These high quality rules were then used to produce a NewRule list.

4.5.2.1 The CN2 co-operative learning episode

Recall from Section 2.4.2 that low quality rules are the individual rules in a rule set with performance measures below some predetermined threshold value. For this experiment the threshold value was determined by the average accuracy of an individual rule calculated over the accuracies of all the individual rules as contained in the different learners' rule sets. The average individual rule accuracy of the machine learner team was 61.2%. Therefore, any rule with accuracy lower than 61.2% was considered to be of a low quality. CN2 produced 5 low quality rules during the individual inductive learning episode. The CN2 learner engaged in co-operative learning by searching the knowledge bases of the other two team members looking for high quality rules that relates to the five low quality rules. One high quality rule was found on the C4.5 team member's knowledge base. This rule overlapped one of the low quality rules and subsumed another. The CN2 learner also obtained a high quality rule from the BRAINNE team member's knowledge base. This rule was in conflict with a low quality rule produced by CN2. The remaining two low quality rules were classified as misconceptions due to their low coverage. The two new high quality rules were placed on CN2's NewRule list and the five low quality rules were removed from the CN2 knowledge base.

The C4.5 high quality rule described the class discrete products and had a rule accuracy of 76.2%. The rule included the following tests:

```
Process Capability < 7

SIC < 3520.
```

The BRAINNE high quality rule, with a rule accuracy of 69%, described the continuous processes class. The rule included the following tests:

```
Product Capability < 3
Support Capability < 5
Product  Value Hierarchy < 5
Key Technology Stage < 9
Product  System Value Hierarchy < 5
Base Technology Stage < 8
302 < SIC < 2521.
```

Next, the data generation process, as introduced in Section 2.4.2, was used to generate two new sets of 167 training instances, one for each rule in the NewRule list. The two new training sets contained the same class distribution as the original training set, with 110 instances belonging to the discrete product class and 57 to the continuous process class.

The first 167 instances generated by the data generator contained 110 examples belonging to the discrete products class, with values:

```
0 <= Process Capability < 7,
SIC < 3520
```

The SICs were randomly chosen, while still representing the distribution of SICs (less than 3520) as contained in the original training set. All the attributes not included in the tests were assigned values distributed similarly to the original training set by ensuring that the mean value and variance of the attributes remained the same.

The 57 examples, belonging to the continuous process class, were generated with values distributed similarly to the original training set, by ensuring that the mean value and variance of the attributes remained the same.

The resultant 167 training instances produced by the data generator were added to the original 167 instance in the training set. In this way, a training set biased to the high quality rule was created, while maintaining the original distribution of the attributes that were not included in the rule. Similarly, a second, new training set was generated, biased towards the BRAINNE high quality rule on the NewRule list.

The CN2 learner then completed two iterations of the individual learning phase to extract new sets of rules from the new training sets. After two iterations the CN2 learner generated two new high quality rules, namely:

```
If    Process Capability < 6
and   3551.5 < SIC < 4380.5
then     business sector  = Product ,
with a 73.8% accuracy
```

and

```
If    744.5 < SIC < 3082
then  business sector = Process,
with a 76.2% accuracy.
```

CN2 produced a rule set of 9 high quality rules with an overall rule set accuracy of 85.7%, improving its original overall rule accuracy by 4.7%. The resulting high quality rule set was pruned and then tested against the test set, giving an overall rule set accuracy of 88.1% and average rule accuracy of 66.4%.

At the end of the individual learning phase, described in Section 4.5.1, the CN2 learner's discrete product concept description was 100% accurate compared to that of the continuous process concept description which was 50% accurate. During the co-operative learning episode, the overall rule set accuracy of the discrete product concept description could not be improved, since it was already 100%. However, the individual rule accuracies of the discrete product concept description improved, which in turn improved the average rule accuracy. The continuous process concept description's accuracy improved from 50% to 68.8%. The highest accuracy of a continuous process concept description generated by the machine-learner team, during their individual learning phase, was 68.8%. The CN2 learner was able to improve its continuous process concept description accuracy to 68.8%, but not higher, confirming Viktor's (1999) finding that co-operative learning cannot negate the effect of poor quality rules. This implied that the overall rule set accuracy of 88.1% could not be improved.

4.5.2.2 The C4.5 co-operative learning episode

During the individual learning phase, C4.5 produced four low quality rules describing the discrete products class but no low quality rules describing the continuous process class.

The C4.5 learner engaged in co-operative learning and obtained two high quality rules from its CN2 team member describing the discrete product class. These rules overlapped C4.5's four low quality rules. The two new high quality rules were placed in C4.5's NewRule list and the four low quality rules were removed from C4.5's knowledge base.

The CN2 high quality rules describing the discrete products class that were included on the NewRule list were:

```
If    Process Capability < 9.5
and   3551.50 < SIC < 4380
then  business sector = Product,
```

with a rule accuracy of 73.8%

and

```
If    Process Capability < 0.5
and   Key < 3
and   Emergent < 1
then  business sector = Product,
```

with a rule accuracy of 64.3%

The C4.5 data generator proceeded by using the rules in the NewRule list to generate two new sets of 167 training instances each. The two new training sets contained the same class distribution as the original training set, with 110 instances belonging to the discrete product class and 57 to the continuous process class.

The first 167 instances generated by the C4.5 data generator contained 110 examples belonging to the discrete product class, with values:

```
0 <=  Process Capability < 9.5,
3551.5 <  SIC < 4380
```

where the SICs were randomly chosen, but still representing the distribution of SICs between 3552 and 4380 as contained in the original training set. All of the attributes not included in the attribute value tests received values distributed in exactly the same way as in the original training set, by ensuring that the original mean value and variance of the attributes remained the same.

A total of 57 new instances belonging to the continuous process class were generated. These values were added to the original training set and the data generator produced a new training set with 334 instances. Similarly, a second new training set was generated, biased towards the second CN2 high quality rule on the NewRule list.

After two individual learning phases, the C4.5 learner produced one new high quality rule, with 64.3% accuracy, namely:

```
If    Process Capability < 1
and   Key < 4
and   Emergent < 1
then     business sector  = Product ,
```

C4.5 produced a new rule set with eight high quality rules and an overall rule set accuracy of 83.3%, as measured against the test set, therefore not improving its overall rule set accuracy. The resulting high quality rule set was pruned by applying the REP rule-pruning algorithm. The pruned rule set, consisting of four high quality rules, was tested against the test set, resulting in an overall rule set accuracy of 90.5% and an average rule accuracy of 75.6%, thus, improving its overall rule set accuracy by 7.2%.

The C4.5 learner generated the most accurate continuous process concept description during the individual learning phase, with an accuracy of 68.8% with no low quality rules, when measured against the test data set. When there is a limited amount of data available inductive learners tend to overfit data, especially decision tree generation algorithms, due to their greedy search method as described in Section 2.5.5.3. Overfitting is a phenomenon during which an algorithm lacks the ability to distinguish between data trends that a rule set should be modelled on and random outliers that should be ignored [Fayyad *et al* 1996]. As mentioned in Section 4.5.1, the training set had limited data covering the continuous process class. Hence, C4.5, a decision tree algorithm, overfitted the instances describing the continuous process class and generated the most accurate rule set as evaluated against the training set.

Co-operative learning therefore only affected the discrete product class. During the co-operative learning episode, the learner was not able to improve the rule set accuracy for the discrete product concept description. This explains why the overall rule set accuracy did not

change. However, the new rule set contained eight high quality rules, compared to the 11 original rules consisting of four low quality and seven high quality rules. After rule pruning, the overall rule set accuracy increased by 7.2%, resulting in an overall rule set accuracy of 90.5%. The discrete product concept description was 100% accurate when applied to the test set whereas the continuous process concept description was only 75% accurate against the test set. This increase in accuracy can be explained by the fact that C4.5, a decision tree generation algorithm, overfitted the data in the training set, which implies that the model, i.e. rule sets, generated by C4.5 were not general enough. Where a general model means that the rule sets, derived from the training set, apply equally well to new sets of data from the same problem not included in the training set [Berthold *et al* 1999]. Pruning is a technique used to generalise concept descriptions by preventing recursive splitting on attributes that are not clearly relevant [Russell *et al* 1995]. Hence, the pruning algorithm generalised the concept descriptions, generated by the C4.5 learner, by removing irrelevant attribute tests. The increase in accuracy, after the rule set was pruned, implies that the concept descriptions generated by C4.5 overfitted the training data and therefore was too specific. By pruning the rule set, the individual rules became more general and therefore the rule set performed significantly better.

4.5.2.3 The BRAINNE co-operative learning episode

BRAINNE produced two low quality rules during the individual inductive learning episode, one describing each class. The BRAINNE learner engaged in co-operative learning and obtained two high quality rules, one from the CN2 team member, which overlapped a low quality rule, and the other from the C4.5 learner that subsumed a low quality rule. The two new high quality rules were placed in BRAINNE's NewRule list and the two low quality rules were removed from BRAINNE's knowledge base.

The CN2 high quality rule describing the continuous processes class that was included in the NewRule list was:

```
If    Process Capability > 8.5
```

```
and    Emergent < 5
and    SIC > 3180
then   business sector = Process,
```

with a rule accuracy of 66.7%

The C4.5 high quality rule describing the discrete product class that was placed in the NewRule list was:

```
If     Process Capability < 7
and    SIC > 3520
then   business sector = Product
```

with a rule accuracy of 76.2%.

The BRAINNE data generator used the rules contained in the NewRule list to generate two new sets of 167 training instances each. The two new training sets contained the same class distribution as the original training set, with 110 instances belonging to the discrete product class and 57 to the continuous process class.

The first 167 instances generated by the BRAINNE data generator contained 57 examples belonging to the class continuous processes, with values:

```
8.5 <= Process Capability <= 28,
0 <= Emergent < 5
3180 < SIC <= 4911
```

with the SICs randomly chosen, but still representing the distribution of SICs between 3180 and 4911 as contained in the original training set. All the attributes not included in the tests received values distributed in exactly the same way as in the original training set by ensuring that the mean value and variance of the attributes remained the same.

The 110 examples belonging to the discrete product class were generated with values distributed in the same way as in the original training set, by ensuring that the original mean value and variance of the attributes remained the same.

The resultant 334 training instances included the 167 produced by the data generator and the original 167 from the training set. In this way, a training set biased to the high quality rule was created. Similarly a second new training set was generated, biased towards the second C4.5 high quality rule in the NewRule list.

The individual inductive learning episode was re-iterated twice and the BRAINNE learner produced two new high quality rules, namely:

```
If    17 < Process Capability < 29
and   7 < Base < 29
and   Emergent < 5
and   2511 < SIC < 4888
then     business sector  = Process ,
```

with a 61.9% accuracy and

and

```
If    8 < Process Capability < 29
and   0 < Base < 13
and   Emergent < 5
and   3020 < SIC < 4897
then     business sector  = Process ,
```

with a 61.9% accuracy and

The BRAINNE learner produced a rule set containing five high quality rules with an overall rule set accuracy of 71.4%, therefore not improving its overall rule accuracy. The resulting

high quality rule set was pruned by applying the REP rule-pruning algorithm. The pruned rule set, consisting of three high quality rules was tested against the test set, resulting in an overall rule set accuracy of 78.6%. The BRAINNE learner improved its overall rule set accuracy by 7.2%.

BRAINNE was the weakest performer in the machine learner team, with an overall rule set accuracy significantly lower than that of the other members. Co-operative learning enabled the BRAINNE learner to improve its performance significantly. However, the learner was unable to improve its performance to the same accuracy level as its team members. The artificial neural network trained by BRAINNE, using the training set, had an overall rule set accuracy of 82%. The rule set extracted by the BRAINNE rule extraction algorithm had an overall rule set accuracy of 73% on the same training set. According to Craven *et al* (1993), the aim of a rule extractor can be described as "*given a trained neural network and the examples used to train it, the rule extractor should produce a concise and accurate description of the network*". The performance of a rule extractor is measured in terms of a degree of fidelity, where fidelity is measured by comparing the classification performance of the rule set to that of the trained neural network from which the rules were extracted. The accuracy of the set of rules was 9% lower than that of the trained neural network, indicating that the rule set does not model the trained neural network to a comparable degree of fidelity. This indicates that the rule extractor did not model the trained neural network accurately.

Table 6 summarises the results of the co-operative learning episode. Step one lists the results before pruning and step two lists the results after pruning, as tested against the test set using the CN2 evaluation function.

Table 6 Accuracies and rule sets after the co-operative learning episode

Step	Overall Rule Set Accuracy			Number of rules in KB		
	CN2	C4.5	BRAINNE	CN2	C4.5	BRAINNE
1	85.7%	83.3%	71.4%	9	8	5
2	88.1%	90.5%	78.6%	9	4	3

4.6 Learning of the classification task by the human learner team

The human learner team consisted of a human expert that participated in NRT Audit supplemented with reference material i.e. the synthesis report. A knowledge engineer, as defined in Section 2.5.7.2, played the role of the performance element, as defined in Section 2.3. The knowledge engineer presented the learning element i.e. human learner, with the training set. The knowledge engineer then scheduled an interview with the human learner for a later date. A week later the knowledge engineer met with the human learner for a structured interview during which the relevant knowledge pertaining to the classification task was acquired from the learner.

4.6.1 Individual learning phase

The human learner presented the knowledge engineer with two concept descriptions, one for each class. The human learner's concept descriptions were expressed explicitly in terms of the attributes and their values that were present in the training set. During the interview it became clear that the human learner based his decisions on which attribute and attribute value to use for describing a class often on memory of specific experiences, rather than the actual data as represented in the training set. With the assistance of the knowledge engineer, the human learner transferred these concept descriptions into decision lists i.e. rule sets that can be interpreted by the critic, the CN2 evaluation function.

The major obstacle the knowledge engineer encountered during the knowledge acquisition process was that of knowledge transfer, as discussed in Section 2.5.7.1. Sestito *et al* (1996) describes this mismatch in knowledge representation as *the difference between the structures of human expert's knowledge compared to the representation of knowledge by a program.* The human learner that participated in the learning episode created concept descriptions by means of a set of disjunctive rules. The CN2 concept description language, as defined in Section 2.5.4.2, only supports conjunctive rules. However, the CN2 evaluation function evaluates the overall rule set as one disjunctive rule. Therefore, the human expert's concept descriptions were represented as disjunctive rule sets consisting of individual conjunctive rules expressed in terms of the CN2 concept description language. Each individual rule had

low individual rule accuracy, because of the limited coverage of each section of the disjunctive rule set. But the overall rule set, representing the disjunctive rule, had a high overall rule set accuracy, since the overall rule set accuracy measurement took the disjunctive nature of the rule set into consideration.

This can be explained as follows:

A training set T, consists of 50 instances describing class Y and a learner extracts the following rules describing class Y:

R_1 : (x > a) $\Rightarrow$ Y

R_2 : (z < b) $\Rightarrow$ Y

If R_1 covered 10 of the 50 instances describing class Y then R_1's correct positive coverage will be 20%. If R_2 covered 5 of the 50 instances describing class Y then R_2's correct positive coverage will be 10%. However, the overall rule set accuracy, as calculated by the CN2 evaluation function will be (10 + 5)/ 50*100 = 30%. Therefore, the following disjunctive rule, R_3, which is represented by the above rule set, is:

R_3 : (x > a) $\cup$ (z < b) $\Rightarrow$ Y

= R_4 : $R_1 \cup R_2$ $\Rightarrow$ Y

has an individual rule set accuracy of 30%, compared to the 10% and 20% accuracies of the individual rules, R_1 and R_2.

This implies that a disjunction of low accuracy rules could be as accurate as a single disjunctive high accuracy rule. This explains why the overall rule set accuracy of the concept description generated by the human learner was high, even though the average individual rule accuracies were low. Table 7 compares the individual learning phases of the

four learners in the two teams, as evaluated by the CN2 evaluator. First, the rule sets were evaluated against the training set, followed by the evaluation against the test set.

Table 7 Accuracies and rule sets after the individual learning phase of both teams

Set	Overall Rule Set Accuracy				Number of rules in KB			
	CN2	C4.5	BRAINNE	Human	CN2	C4.5	BRAINNE	Human
Training	98.2%	94.6%	73.7%	91%	14	11	5	9
Test	81%	83.3%	71.4%	88.1%	14	11	5	9

During the evaluation of the four learners' rule sets against the training set, as generated during the individual inductive learning episodes, CN2 scored the highest overall rule set accuracy, followed by C4.5 and then the human learner. However, when the same rule sets were evaluated against unseen instances the human learner scored the highest overall rule set accuracy followed by C4.5 and CN2. CN2 and C4.5's performance measure dropped by 17% and 11% respectively, but the human learner had a fluctuation of only 3%. This result emphasises the value of the background knowledge the human learner possesses within this problem domain. With the help of this knowledge, the human learner was able to construct concept descriptions that included knowledge extracted from prior experience, not necessarily reflected by the instances in the training set. However, the machine learners were restricted to the knowledge embedded in the training set and generated concept description reflecting that. Hence, when the data set changed the new knowledge in the test set is not reflected in the machine learner's concept descriptions and therefore the decrease in the machine learners' performance.

4.7 Validation evaluation episode and knowledge fusion

Full co-operation involves the active participation of all the team members during co-operative learning. The individual and team results are evaluated by considering the improvement over past performances and/or the actual final mark achieved. In a machine learning environment, the individual learner's final rule sets are evaluated against a validation set and the accuracy's are placed in each learner's knowledge base. Next, these knowledge bases of the individual team members are fused together into a team knowledge

base. The purpose of this step was to produce, for each full co-operative learner team, an integrated knowledge base that contains the results of the team effort. These team knowledge bases were subsequently pruned and validated against the validation set. Table 8 compares the overall rule set accuracies of the individual learners knowledge bases after the co-operative learning episode, to that of the team knowledge bases after knowledge fusion, as tested against the validation set. Lastly, the two team's knowledge bases were fused together into a final knowledge base. The result of the knowledge fusion episode is summarised in Table 9.

Table 8 Accuracies and rule sets after validation and knowledge fusion

Pruned	**Overall Rule Set Accuracy (Number of rules in KB)**				
	CN2	C4.5	BRAINNE	Machine learner team	Human learner team
No				77.5% (16)	80% (9)
Yes	85% (9)	72.5% (4)	67.5% (3)	82.5% (5)	

Table 9 Final accuracies and rule sets of fused knowledge bases

Overall Rule Set Accuracy (Number of rules)			
	Machine learner team	Human learner team	Machine and human learner team results fused together
Un-pruned	77.5% (16)	80% (9)	87.5% (14)
Pruned	82.5% (5)		87.5% (11)

The continuous process concept description generated by the machine learning team during knowledge fusion had an accuracy of 83.3% when evaluated against the validation set. The next best performing description is that of the CN2 learner, with an accuracy of 58.3%. This implies that, by means of full co-operation, the machine learner team was able to improve its predictive accuracy of the continuous process concept description by at least 25%. This team achieved the goal of the co-operative learning episode namely, to produce sets of informative rules that describe the continuous process class. The discrete product concept description accuracy also showed a slight improvement of 6.8%, from 82.5% to 89.3%. The final knowledge base created by the fusion of the human and machine learner team's pruned rule sets had the highest overall rule set accuracy of 87.5%. This result highlights the

success of human-computer collaboration. It emphasises the importance of determining what the human learner is good at and what the machine learners are good at, so that a collaborative strategy focussing on the strengths of the teams can emerge. Most importantly, these results indicated that in a real-world scenario, full co-operation is preferred over individual learning.

4.8 Discussion

This section highlighted the knowledge acquired during the learning of a classification task. It discussed how the rule sets, obtained through co-operative learning, in some instances supported the findings of the human experts, while in others contradicted the findings of the human experts. It went on to describe new additional findings that were made during the learning process, which could have been considered when the human experts compiled the synthesis report.

The following knowledge acquired by the CILT-MAL system supported some of the findings of the human experts. The human experts were of the opinion that technologies in the key stage of sophistication are more dominant in the discrete product companies compared to the continuous process companies. On the other hand, technologies in the base or emergent stage of sophistication are more dominant in the continuous process companies. Continuous process companies, who are competing for an increased market share internationally, use emergent technologies in order to minimise production costs and increase profit margins. Therefore, these companies have to combine basic technologies with "cutting-edge" technologies in a well-balanced manner to maximise their profit margin. The international prices and manufacturing specifications of the products produced by the continuous process companies are usually internationally determined, e.g. gold and steel. However, according to the data, a total of 62% of the discrete product companies used technologies in the key technological stage of sophistication, compared to only 46% of the continuous process companies. In addition, 26% of the continuous process companies employed technologies in the emergent technological stage of sophistication, compared to only 6% of the discrete product companies. This indicates that the rules, as produced during co-operative learning, supported the human expert's opinions as stated above.

In some instances the results of the CILT-MAL system contradicted the opinions of the human experts. For example, knowledge acquired by the CILT-MAL system showed that 97.5% of the continuous process companies employ process type technologies. Process type technologies are those technologies and competencies directly related to the manufacturing process. On the other hand, only 42% of the discrete product industry employs process type technologies. The human experts were of the opinion that, when it comes to process type technologies, there should be no significant difference in their use within the two business-clusters. The finding by the CILT-MAL system indicates that discrete product companies do not rely on process type technologies during the final stages of production. This approach leads to job creation, an important issue in South Africa and one of the many challenges facing the country in the 21st century, as discussed in Chapter 3. However, this approach leads to an increase in the production cost of goods, which has a negative influence on the selling price of the product and eventually the competitiveness in the market.

Co-operative learning highlighted new, additional findings. For example, the companies within the textile and footwear industry are unique compared to the other discrete product companies. During all stages of learning a dedicated rule was generated for the leather and footwear industry companies. This rule indicated that these companies did not adhere to the general concept description of the discrete product cluster. Further investigation discovered that more than 20% of these companies were not able to adequately describe their technologies in the survey [AMI 1998]. This indicated a low level of technological orientation in this sector. Only 12% of the companies could specify their research and technology outputs [AMI 1998], which indicated a lack of technology management in these organisations. The major technology types identified within this industry were of the product and process capability type. Even within these major technology types, the companies identified less than the average number of technologies identified across the business-cluster [AMI 1998]. The textile and footwear industry sector indicate that the biggest threat to their competitive environment is illegal imports into South Africa [AMI 1998]. This, in turn pressurises their pricing structure. To overcome this hurdle, productivity must be increased, by investing in new technologies, enabling them to reduce overall cost and lower the price of merchandise. The CILT-MAL system identified the textile and footwear sector companies as being different compared to the other discrete product companies. After

further investigation of the data the following problem was identified. Currently, many of the companies in this sector are struggling to survive because of their high production costs and therefore the inability to price their products competitively. One cannot help wondering if this could have been prevented if the cost trends were identified at an earlier stage.

Lastly, it was established that the most significant attribute enabling learners to distinguish between the two business-clusters was the SIC attribute. The SIC's, identifying discrete product companies, were grouped into three ranges. These ranges were imbedded within the six ranges that identify the continuous process companies. According to Langley's machine learning framework, one of the key aspects of the environment is the representation of knowledge, of both input to learning and output of learning. The combination of the input to learning not being in continuous ranges and CN2's inability to formulate disjunctive rules i.e. not being able to represent the three ranges as a disjunctive rule restricted the learners ability to learn.

4.9 Conclusion

This chapter presented a classification type task in a real-world application. This task concerned the grouping of organisations into business-clusters by means of co-operative learning in a multi-agent learning environment, as described in Section 4.1. In Sections 4.2 and 4.3 the data selection process that divided the data into three sets, i.e. a training, test and validation set, was discussed. The experimental method and evaluation criteria were presented in Section 4.4. Section 4.5 and 4.6 introduces the teams and showed how co-operative inductive learning manifested within the MAL system. The next section, Section 4.7, presented the way in which the knowledge bases were fused together and discussed the success of the knowledge fusion episode. The chapter concludes with a discussion of the major trends discovered during the co-operative learning episode.

When comparing the major trends discovered by the CILT-MAL system to the findings of the synthesis report, regarding this specific classification task, one comes to the following conclusion. That, the human experts involved in producing the synthesis report grouped the

companies surveyed into two broad business-clusters, the two clusters were defined as continuous process industries and discrete product industries. To adhere to these definitions, companies should have been categorised according to the characteristics of their major product lines. Knowledge, as contained in the knowledge base of the CILT-MAL system, showed that the groupings were made instead according to the organisation's SIC. There is no indication that the level of sophistication of the products played a role in the classification, as one would have expected from the definition of the two clusters.

The results indicated that a KDD environment, modelled as a MAL system, could be used successfully for a real-world application. In this case, through the construction of a knowledge base, a valuable tool in developing a technology policy framework was provided. This can be attributed to the following: [Viktor *et al* 2000]:

- Human experts have seen that they can successfully verify their results against the data. This helps building experts confidence in their own pre-empted ideas and also changing their ideas where being proved to be incorrect.
- The results as contained in the team knowledge base confirmed certain pre-empted ideas from experts, but also showed where the pre-empted ideas were wrongfully made. The approach can therefore be used to ensure that decisions are taken according to correct assumptions.
- The KDD process provided a tool that should enable government to make sense of the large amount of data produced by the NRT Audit. The socio-economic threats of the incorrect application of a Science and Technology policy in South Africa can, if care is not taken, widen the gap between the economic 'haves' and 'have-nots'. With this in mind, it is very important to consider the way in which policy will be formulated in South Africa. These inputs can sensibly be used to ensure that the policy forming is not top-down driven, but rather a bottom-up approach.

The next chapter presents the application of the CILT-MAL system to a problem solving type task concerning the quality of human resources, in scarce disciplines of specialisation, as introduced in Section 3.3.