FACILITATING DECISION-MAKING PROCESS USING A HIERARCHICAL MULTI-ATTRIBUTE MODEL (HMAM) IN MODELING RELATIONAL DATA

Rayner Alfred

School of Engineering and Information Technology,
Universiti Malaysia Sabah, 88999 Kota Kinabalu, Sabah, Malaysia

ABSTRACT. *This work presents a novel approach that is capable of learning relational domain and generating automated hierarchical multi-attribute model (HMAM) to support the development of decision-making. In this paper, we describe the technique of generalizing data in relational domain using granularity computing as a means of data summarization to automate and support the construction of HMAM for decision-making. First, we introduce related works in relational data mining. Then, we introduce the concept of hierarchical multi-attribute model in decision modeling. We proceed by introducing our approach that uses the pattern-based aggregation approach to relational data mining and discuss the pre-processing procedure. Experimental results are presented based on the hepatitis dataset (KDD CUP 2005). The results of our analysis show that the proposed HMAM model is able to generate rules and the performance of classifier can be improved by adjusting the number of clusters generated.*

KEYWORDS. Granular computing, Relational data mining, Clustering, Decision-making, Data summarization, Hierarchical multi-attribute model

INTRODUCTION

The processing power to acquire and store large amount of data on documents has increased dramatically over the last few years. Despite the growing of computational power of modern computers, our abilities, to analyze these data for decision-making, are limited for data stored in relational model (multiple tables). We need to join these multiple tables in order to get more information about a specific record stored in target table that has *one-to-many* relationship with data stored in another table. However, most traditional data mining tools cannot handle relational dataset with high-dimensional of *one-to-many* relationship, unless pre-processing task is applied to the data for data conversion. Granular computing has begun to play important roles in bioinformatics, e-Business, security, machine learning, data mining, high-performance computing and wireless mobile computing in terms of efficiency, effectiveness, robustness and uncertainty in order to support the development of *decision-making* model.

This work presents a novel approach that is capable of learning relational domain and generating automated hierarchical multi-attribute model to support the development of decision-making. In this approach, we describe the technique of generalizing data with *one-to-many* relationship using granularity computing as a means of data summarization to automate and support the construction of hierarchical multi-attribute modeling (HMAM) for decision-making (Marko, 2001). We first introduce related works in relational data mining. Then, we introduce the concept of hierarchical multi-attribute model in decision modeling and the pattern-based aggregation approach to relational data mining and discuss the pre-processing procedure. Experimental results and conclusions are then presented to summarize this paper.

RELATED WORKS

Relational learning research is not a new research area and it has a long history. Muggleton and DeRaedt introduce the concept of Inductive Logic Programming (ILP) and its theory, methods and implementations in learning multi-relational domains (Muggleton and DeRaedt, 1994). ILP methods learn a set of existentially unified first-order Horn clauses that can be applied as a classifier (Dzeroski and Lavrac, 2001).

In a relational learner based on logic-based propositionalization (Kramer *et al.*, 2001), instead of searching the first-order hypothesis space directly, one uses a transformation module to compute a large number of propositional features and then uses a propositional learner. Both ILP and binary propositionalization are lack of support for numerical aggregation. In general, propositionalization approaches may outperform ILP or MRDM systems, as was suggested before in the literature (Dzeroski *et al.*, 1999; Srinivasan *et al.*, 1999). The choices of aggregation methods and parameters also have significant effects on the results on noisy real-world domains (Koller and Pfeffer, 1998). Krogel *et al.*, have conducted a comparative evaluation of approaches to Boolean and numeric aggregation in propositionalization; however their results are inconclusive (Krogel *et al.*, 2003). In contrast, Perlich and Provost have found that logic-based relational learning and logic-based (binary) propositionalization perform poorly on a noisy domain compare to numerical propositionalization (Perlich and Provost, 2003).

Another variant of relational learning include distance-based methods (Knobbe *et al.*, 2001; Perlich and Provost, 2003). The central idea of distance-based methods is that it is possible to compute the mutual distance (Horvath *et al.*, 2001) for each pair of object for clustering (Dillon and Goldstein, 1984; McQueen, 1967). Probabilistic Relational Models (PRMs) provide another approach to relational data mining that is grounded in a sound statistical framework (Getoor *et al.*, 2001; Koller and Pfeffer, 1998). Getoor *et al.*, introduce a model that specifies, for each attributes of an object, its (probabilistic) dependence on other attributes of that object and on attributes of related objects. Propescul *et al.*, propose a combination approach called Structural Logistic Regression (SLR) that combines relational and statistical learning (Propescul *et al.*, 2002). Database numeric aggregation (Knobbe and De Haas, 2001) techniques propose a method in which aggregation is done by using some of the build-in functions of common relational database system such as *count*, *min*, *max*, *sum*, *avg* and *exist*. Another approach proposed by Perlich and Provost, where the main technique use vector distances for dimensionality reduction and is capable of aggregating high-dimensional categorical attributes that traditionally have posed a significant challenge in relational modelling (Perlich and Provost, 2005).

DECISION SUPPORT AND HIERARCHICAL MULTI-ATTRIBUTE MODEL

Decision Support

The term "*decision support*" has a variety of meanings depending on the context on how it is used. Marko has outlined the literature review of decision support in detail (Marko, 2001). Decision support can be categorized into *human decision science* (INSEAD, 2003) and *machine decision-making* (Power, 1999). *Human decision science* is defined as an interdisciplinary field which addresses three possibly overlapping aspects of human decision making: normative, descriptive and decision support itself.

There are some other definitions of decision support that focus on specialized disciplines (Figure 1), such as operations research and management science (Hillier and Lieberman, 2000), decision analysis (Clemen, 1996), decision support systems (Power, 1999), and others including data warehousing (**Han and Kamber, 2001**), group decision support systems (Power, 1999; Mallach, 1994) and computer-supported cooperative work. Decision analysis introduced by Clemen applied decision theory (Clemen, 1996). Decision analysis provides a framework for analyzing decision problems by structuring and breaking them down into more manageable parts, and explicitly considering the possible alternatives, available information, uncertainties involved and relevant preferences. Clemen introduces three models in decision-making, which are *influence diagram*, *decision tree* and *multi-attribute* models.

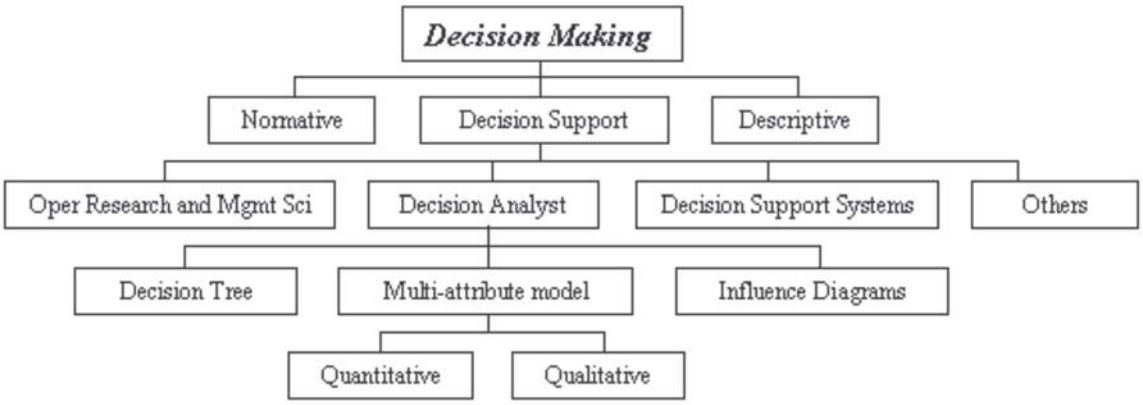

Figure 1. Decision making

Multi-Attribute Model

In principle, a multi-attribute model (MAM) (Clemen, 1996) represents a decomposition of a decision problem into smaller and less complex sub-problems. A model consists of *attributes* and *utility functions*, as shown in Figure 2. *Attributes* are variables corresponding to decision sub-problems and all attributes at the leaf are basic attributes and attribute at the node is aggregate attribute. *Utility functions* define the relationship between the attributes at different levels in the tree and they serve for the aggregation of partial sub-problems into the overall evaluation or classification of options.

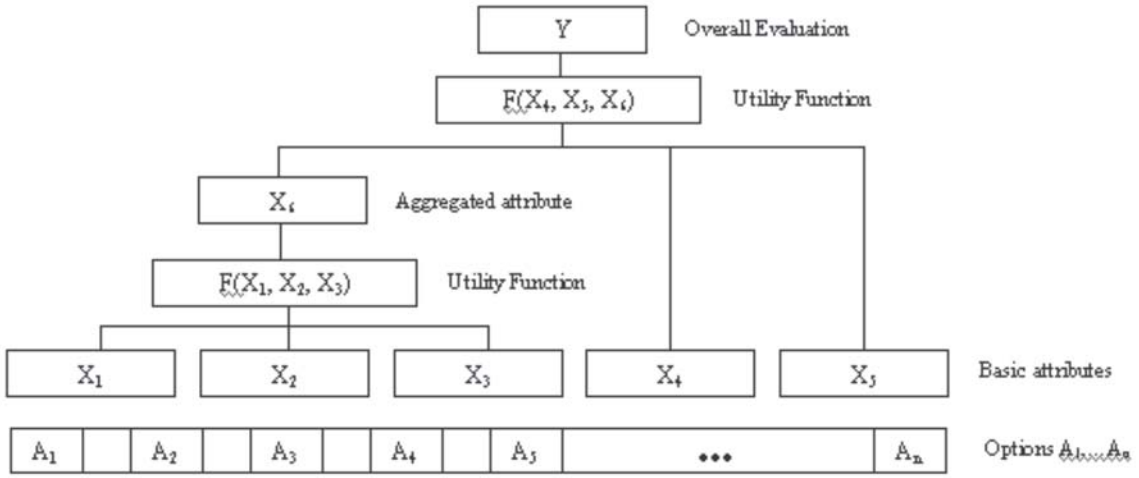

Figure 2. Components of a multi-attribute model.

The overall evaluation (Y) of an option is finally obtained as the value of one or more root attributes in Figure 2. There are two types of MAM: *quantitative decision model* and *qualitative decision model*. In *Quantitative decision model*, all attributes are continuous and the utility functions are typically defined in term of attributes' weights, such as a weighted average of lower-level attributes (DAS, 2001; Younes, 2001; Parmigiani, 2002). In contrast, in *qualitative decision model*, all attributes are either nominal or ordinal (Marko, 2001), whose values are usually string values rather than numbers and the utility functions use clustering functions to summarize data. This paper emphasizes the *qualitative decision model* in constructing decision making model. Figure 3 illustrates how learning relational domains using dynamic aggregation based on patterns' distance (DARA) algorithm provides a concrete foundation for bridging relational data mining and MAM. DARA algorithm (shown in Figure 4) uses data summarization as the utility function to automate the construction of multi-attribute model to support decision-making.

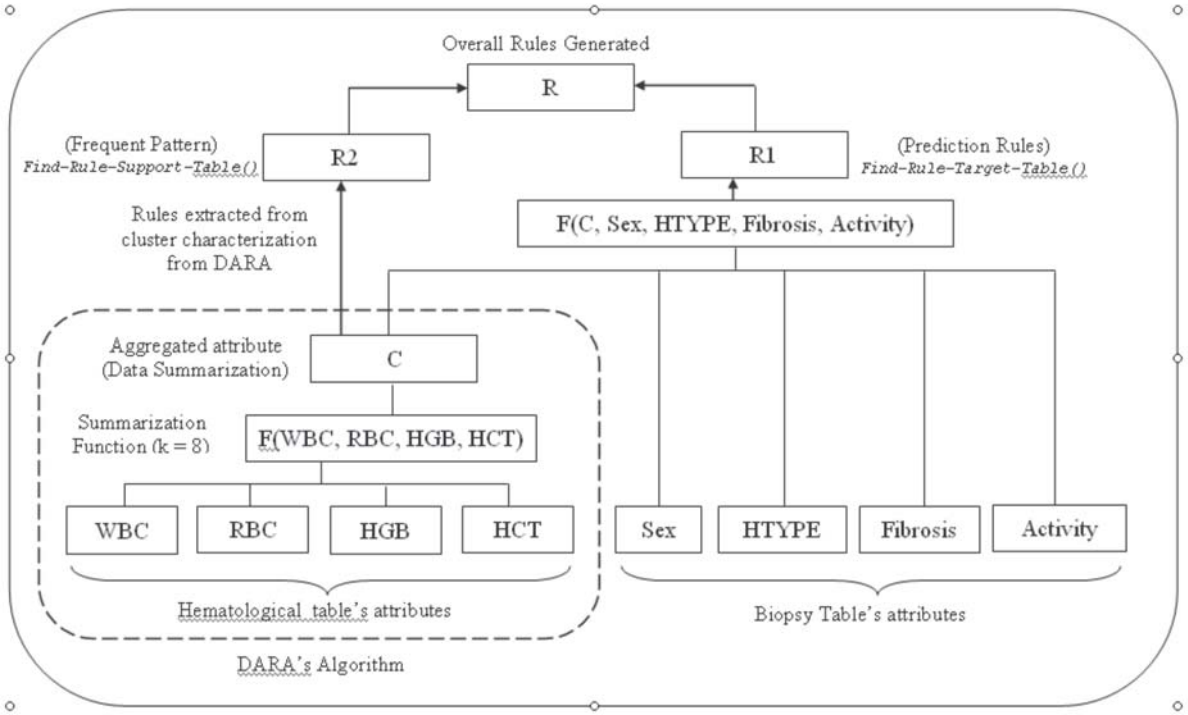

Figure 3. Components of multi-attribute decision model for hepatitis datasets.

PATTERN-BASED FEATURE AGGREGATION

A common method to aggregate a single categorical attribute with numerous patterns is the selection of a subset of pattern that appears most often or based on the distribution. In this concept, each record (row) is viewed as a vector whose dimensions correspond to patterns occur in the target table stored in relational domain; the component magnitudes are the *pf-irf* weights, as describes in Equation 1, of the patterns which is adapted from *tf-idf* weights (Salton, 1986).

$$pf\text{-}irf = pf(p, r) \cdot irf(p) \quad (1)$$

$$irf(p) = \log \frac{|R|}{rf(p)} \quad (2)$$

$$sim(r_i, r_j) = \frac{r_i \cdot r_j}{||r_i|| \cdot ||r_j||} \quad (3)$$

Pf-ipf is the product of *pattern frequency Pf(p, r)*, and the *inverse record frequency* (Equation 2). *Pf(p, r)* refers to the number of times pattern *p* occurs in the corresponding record *r*. In Equation 2, $|R|$ is the number of records in the table and *rf(p)* is the number of records in which pattern *p* occurs at least once. Therefore, given two vectors (records) with component magnitudes described in Equation 1, the similarity between two records is then computed in Equation 3, where r_i and r_j are vectors with *pf-irf* coordinates as described above. Aggregation can be defined as a summarization of the underlying pattern or distribution from which the related objects were sampled. Once we compute the *pf-irf* weights, then we can compute the distance between each record and cluster them based on their eights. By grouping them into clusters or segments (Bezdek, 1998), we are generalizing or aggregating them based on the underlying pattern or distribution from which the related objects were sample.

The process of data summarization is done using DARA's algorithm (Figure 4) through the data generalization. The generalization task is done by converting each record's measurement into patterns, (in Create-Pattern() from Figure 4). Then, data summarization is done for each record, by first computing the *pattern-frequency* and *inverse-record frequency* (Equation 2) and then grouping them based on the distance between records (Equation 3). This individual-centered concept, in which all rows belonging to a specific record is considered as a pattern that characterizes the individualism of each record. For instance, Figure 5 depicts the summarization process for each record using Equation 1 and 3. Firstly, each record is characterized by patterns of WBC, RBC, HGB and HCT measurements and they are converted into binary codes (01 = below normal, 10 = normal, 11 = above normal). Then, using Equation 1, we compute the magnitude weight for each pattern. For example, given p = 10101010 and r = 1, *pf-irf*(10101010,1) = 4 x log (5/4) = 0.387. All records are then clustered (using Compute-Similarity-And-Transform() in Figure 4), based on the records' component magnitudes in Equation 1. The component magnitude for each pattern is computed repeatedly for all records.

```
Input: A relational database
Output: A set of Rules
Procedure:
      Rule set R = Empty
      Create-Pattern()
      Compute-Similarity-And-Transform()
      Update-Target-Table()
      Rule R1 = Find-Rule-Target-Table()
      Add R1 to R
      Rule R2 = Find-Rule-Support-Table()
      Add R2 to R
      Return R
End Procedure
```

Figure 4. Dynamic Aggregations of Relational Attributes Algorithm (DARA).

MID	WBC	RBC	HGB	HCT
1	7.4	4.76	16.2	46.3
1	7.8	4.80	16.0	46.5
1	8.1	4.78	15.6	46.2
1	8.4	4.85	16.2	47.2
2	6.4	5.63	17.0	51.8
2	6.9	5.58	16.9	50.6

Hematological

Encoded Patterns Representing Each Record

1, 10101010, 10101010, 10101010, 10101010,
→ 2, 10111011, 10111011

Figure 5. Data generalization for *one-to-many* relationship.

The set of overall rules, R, (in Figure 3) is obtained from the combination of two set of rules, R1 and R2. In Figure 4, R1 is obtained by using the function *Find-Rule-Target-Table()*, where it uses existing attribute-value classifiers such as C4.5, Conjunctive Rules and Naïve Bayes. We use the *Weka* software (Witten and Frank, 1999) to extract R1. On the other hand, R2 is induced by using *Find-Rule-Support-Table()* as shown in Figure 4 that describes the characteristics of each cluster by finding pattern that has the maximum component magnitudes for each cluster. The result of our experiment on Hepatitis dataset is discussed in the next section, in which rules generated from the HMAM using DARA is more efficient in terms of the percentage of correctly classified instances. In Figure 3, we integrate the relational learning algorithm, DARA, with HMAM in supporting the construction of decision support system.

EXPERIMENTAL RESULTS ON HEPATITIS DATASET

The database collected at Chiba University hospital contains information on patients' exam dating from 1982 to 2001. Among the topics suggested by the Hepatitis dataset, we proposed to evaluate whether the level of biopsy activities and the type of hepatitis can be estimated based on laboratory tests namely WBC, RBC, HGB and HCT. These laboratory tests were chosen based on the work reported by Watanabe *et al.* (Watanabe *et al.*, 2003). The approach adopted here consisted of analyzing blood tests together with the biopsy results, seeking patterns that might indicate a correlation between the patients' exam results and the degree of their activities and also the type of their hepatitis. We also evaluate the performance of the classifiers by adjusting the number of clusters to get the most improved result from this experiment.

The accuracy estimation for three classifiers obtained from the *10-fold cross validation* results are shown in Table 1. In this experiment, we the clustering technique used is a k-means clustering technique. In Table 1, the percentage of correctly classified instances for type of hepatitis increases significantly by 1.16% using *DARA algorithm using k-means clustering* when number of clusters is 8 or 45. In contrast, in Table 2, the percentage of correctly classified instances for Biopsy Activities increases significantly by 1.45% when k = 40.

Table 1. Percentage of correctly classified instances for type of hepatitis.

Clusters	0	2	4	6	8	10	15	20	25	30	35	40	45	50	55	60	65
C4.5	70.4	69.9	69.9	69.9	**71.6**	70.4	70.5	69.8	69.9	70.7	71.4	70.5	**71.3**	70.1	69.7	69.5	70.3
N. Bayes	70.4	69.9	70.4	70.4	**70.5**	70.8	68.9	69.8	68.8	70.4	69.9	71.1	**70.4**	69.9	70.1	70.1	70.4
Con Rules	70.4	70.4	70.4	70.4	**70.4**	70.4	70.4	70.4	70.4	70.2	70.4	70.4	**70.4**	70.4	70.4	70.4	70.4

Table 2. Percentage of correctly classified instances for biopsy activities.

Clusters	0	2	4	6	8	10	15	20	25	30	35	40	45	50	55	60	65
C4.5	61.5	61.5	60.5	60.5	60.5	60.7	60.9	60.9	61.4	60.5	**62.3**	**62.9**	61.4	60.9	61.7	61.7	**62.3**
N. Bayes	59.7	60.5	61.3	60.2	60.9	61.4	61.5	61.5	61.5	61.4	**61.3**	**63.1**	61.4	60.9	61.5	62.1	**61.4**
Con Rules	61.5	59.7	59.7	59.7	59.7	59.7	59.7	59.7	59.7	59.4	**59.7**	**59.7**	59.7	59.7	59.7	59.7	**59.7**

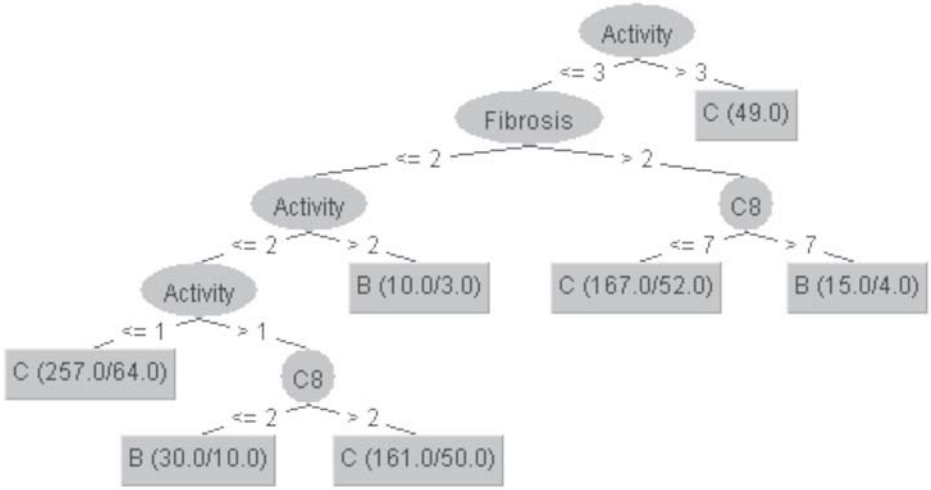

Figure 6. R1 obtained using C4.5 for type of hepatitis (k = 8).

Figure 6 and 7 depicts set of rules, R1, induced using C4.5 (Witten and Frank. 1999) for classifying the type of hepatitis and also the biopsy activity. For each case, we also get the data summarization (R2) for each cluster as shown in Table 3 and 4 where N = Normal, AB = Above Normal, and BN = Below Normal. This data summarization can be considered as rules extracted from the dataset. For instance, in Table3, for the number of clusters which is greater than 2, we have patients who have normal level of WBC, RBC, HGB and HCT. Table 5 and 6 summarize the findings based on R1 and R2 for classifying the type of hepatitis and classifying the activities of virus.

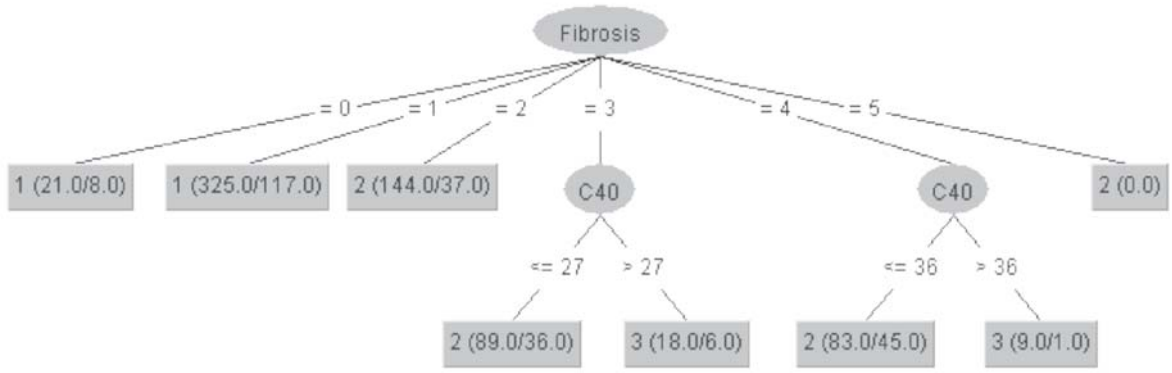

Figure 7. R1 obtained (C4.5) for fibrosis activity (k = 40).

Table 3. Characterization of cluster for type of hepatitis.

Clusters	WBC, RBC, HGB, HCT	Weight
C > 2	N, N, N, N	10391.2
C ≤ 2	BN, N, BN, BN	2149.4
C ≤ 2	N, AN, N, N	2445.4
C > 7	N, N, BN, BN	1395.7
C ≤ 7	N, N, N, N	10391.2

Table 4. Characterization of level of biopsy activities.

Clusters	WBC, RBC, HGB, HCT	Weight
C ≤ 36	N, AN, N, N	1947.9
C > 36	N, N, N, N	715.1
C ≤ 27	N, AN, N, N	1947.9
C > 27	N, N, N, N	715.0

Table 5. Finding for classifying type of hepatitis.

TYPE	FINDINGS
Hepatitis C	a) Fibrosis level is F2 or lower and the activity of virus is A1, b) Fibrosis level is F2 or lower and the activity of virus is A2 with WBC, RBC, HGB and HCT at normal level c) Fibrosis level is greater than F2 and the activity of virus is A3 or lower with WBC, RBC, HGB and HCT at normal level d) Fibrosis level is greater than F2 with WBC, RBC, HGB and HCT at normal level
Hepatitis B	a) Fibrosis level is F2 or lower and the activity of virus is A3 or greater, b) Fibrosis level is F2 or lower and the activity of virus is A2 with WBC, HGB and HCT at below normal and RBC at normal level c) Fibrosis level is greater than F2 with WBC, RBC at normal level but HGB and HCT at below normal level

Table 6. Finding for classifying level of virus activities.

TYPE	FINDINGS
1	a) Fibrosis level is F0 and F1
2	a) Fibrosis level is F2
	b) Fibrosis level is F3 or F4 with WBC, HGB and HCT at normal level and RBC at above normal level
3	a) Fibrosis level is F3 or F4 with WBC, RBC, HGB and HCT at normal level

In short, by varying the number of clusters, we obtain a few different accuracy estimations for the three classifiers that include C4.5, Naive Bayes and Conjunctive Rules classifiers. The results with the highest accuracy estimations are taken into consideration when extracting rules to model the relational datasets.

CONCLUSION AND FUTURE WORKS

In this paper, we propose Dynamic Aggregation of Relational Attributes (DARA), an efficient approach to learning relational domain and we integrate DARA with the HMAM to support modeling of decision support for Hepatitis dataset. The results revealed that *DARA algorithm* generates rules and the performance of the classifiers can be improved by adjusting the number of clusters used. There are some other techniques that can be considered in the transformation process, such as Self Organizing Map (SOM) technique. SOM is very effective to be used when we have a lot of missing data and this could improve the transformation-based approach in multi-relational domain. In the future, we would proceed to validate the clinical reasonability of the results and validate the usefulness of the system on other datasets. We would also apply one of the optimization techniques, e.g., genetic algorithm, in order to find the best number of clusters used to model the relational database.

REFERENCES

Agrawal, R. & Srikant, R. 1994. Fast Algorithms for Mining Association Rules. In *Proc. of the International Conference on Very Large Databases*, Santiago de Chile, Chile, 1994

Bezdek, J.C. 1998. Some New Indexes of Cluster Validity, *IEEE Trans. Syst*., Man, Cybern. B, **28**:301-315.

Clemen, R. T. 1996. Making Hard Decisions: An introduction to Decision Analysis, Duxbury Press.

DAS. 2001. *Decision Analysis Software*. http://faculty.fuqua.duke.edu/daweb/dasw.htm

Dillon, W. & Goldstein, M. 1984. *Multivariate analysis*, John Wiley and Sons, Chichester, 157-208.

Dzeroski, S., Blockeel, H., Kompare, B., Kramer, S., Pfahringer, B. & Van Laer, W. 1999. Experiments in Predicting Biodegradability, In *Proceedings of Inductive Logic Programming* '99.

Dzeroski, S. & Lavrac, M. 2001. *Relational Data mining*. Springer-Verlag.

Getoor, L. Friedman, N. Koller, D. & Pfeffer, A. 2001. Learning Probabilistic relational models. In Dzeroski, S. and Lavrac, N. editors. *Relational Data mining*. Springer-Verlag.

Han, J. & Kamber, M. 2001. Data Mining: Concept and Techniques, Morgan Kaufman.

Hillier, F.S. & Lieberman, G.J. 2000. *Introduction to Operation Research*, McGraw Hill.

Horvath, T., Wrobel, S. & Bohnebeck, U. 2001. Relational Instance-based Learning with Lists and Terms. *Machine Learning*, 43(1/2): 53-80.

INSEAD, 2003. *Decision Sciences*. PhD Program Description, http://www.insead.edu/phd/program/decision.htm

Kirsten, M., Wrobel, S. & Horvath, T., 2001. Distance Based Approaches to Relational Learning and Clustering. In Dzeroski, S., and Lavrac, N., editors. *Relational Data mining*. Springer-Verlag.

Knobbe, A., De Haas, M. & Siebes, A. 2001. Propositionalization and Aggregates. In *LNAI*, 2168:277-288.

Koller, D. & Pfeffer, A. 1998. Probabilistic Frame-based Systems. In *AAAI/IAAI*, 580-587.

Kramer, S., Lavrac, N. & Flach, P. 2001. Propositionalization Approaches to Relational Data Mining. In Dzeroski, S. and Lavrac, N., editors. *Relational Data mining*. Springer-Verlag.

Krogel, M.A., Rawles, S., Zelezny, F., Flach, P.A., Lavrac, N. & Wrobel, S. 2003. Comparative Evaluation of Approaches to Propositionalization. In *13th International Conference on Inductive Logic Programming* (ILP), 197-214.

Mallach, E.G. 1994. Understanding Decision Support Systems and Expert Systems, Irwin, Burr Ridge.

Marko, B. 2001. *Decision Support*. In Mladenic, D., Lavrač, N., Bohanec, M. & Moyle, S. 2003. *Data Mining and Decision Support: Integration and Collaboration*, Kluwer Aca Pub.

McQueen, J. 1967. Some Methods of Classification and Analysis of Multivariate Observations. In *Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 281-293.

Muggleton, S.H. & DeRaedt, L. 1994. Inductive Logic Programming: Theory and Methods. *The Journal of Logic Programming*, 19&20:629-680.

Muggleton, S.H. 1995. Inverse Entailment and Progol. *New Generation Computing*, 13:245-286.

Parmigiani, G. 2002. *Modelling in Medical Decision Making: A Bayesian Approach.* John Wiley & Sons, Ltd.

Perlich, C. & Provost, F. 2003. Aggregation-based Feature Invention and Relational Concept Classes. In Proceedings *of the Ninth ACM International Conference on Knowledge Discovery and Data Mining (KDD)*.

Perlich, C. & Provost, F. 2005. ACORA: Distribution-based Aggregation for Relational Learning from Identifier Attributes. *Journal of Machine Learning*.

Power, D.J. 1999. *Decision Support Systems Glossary*, http://DSSResources.COM/glossary/

Propescul, A., Ungar, L.H., Lawrence, S. & Pennock, D.M. 2002. Structural Logistic Regression: Combining Relational and Statistical Learning. In *Proceedings of the workshop on Multi-Relational Data Mining* (MRDM-2002). University of Alberta, Edmonton, Canada, 130-141.

Salton, G.M. & McGill, M.J., 1986. *Introduction to Modern Information Retrieval*, McGraw-Hill, Inc., New York, NY.

Srinivasan, A. & King, R.D. 1999. Feature Construction with Inductive Logic Programming: A Study of Quantitative Predictions of Biological Activity Aided by Structural Attributes. *Data Mining and Knowledge Discovery*, **3(1)**:37-57.

Srinivasan, A., King, R.D. & Bristol, D.W. 1999. An Assessment of ILP-Assisted Models for Toxicology and the PTE-3 Experiment, In *Proceedings of Inductive Logic Programming* '99.

Watanabe. T., Suzuki. H. & Takabayashi. K., 2003. Application of Prototypeline to Chronic Hepatitis Data. In *Working Core of ECML/PKDD 2003 Discovery Challenge*, 166-177.

Witten, I. & E. Frank. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufman.

Younes, H.L.S. 2001. *Current Tools for Assisting Intelligent Agents in Real-time Decision Making,* MSc Thesis, http://www-2.cs.cmu.edu/~lorens/papers/mscthesis.html