In *International ICSC Congress on Intelligent Systems and Applications (ISA'2000), Special Session on Knowledge Based Systems in Medicine, December 11-15 2000, Sydney.*

SPECIFICATION OF HEALTHCARE EXPERT SYSTEMS USING A MULTI-MECHANISM RULE-EXTRACTION PIPELINE

A Goh, SSR Abidi & KM Hoe
School of Computer Sciences
Universiti Sains Malaysia
11800 Penang, Malaysia
E-mail: alwyn@cs.usm.my

Abstract: The application of knowledge extraction methodologies in support of medical informatics promises interesting developments that could potentially improve many aspects of healthcare services. In this paper we outline a multi-stage rule extraction pipeline for rule-based knowledge discovery. The featured methodology would facilitate operationally straightforward extraction of symbolic rules from medical datasets, in particular those with unannotated ordinal or continuous-valued datavectors. The extracted rulesets will be used in the construction or enrichment of rule-based expert systems. Our pipeline incorporates well-established supervised and unsupervised machine learning methods used for data mining. The motivation for our work stems from the individual effectiveness of data mining methods available for datavector clustering, attribute discretisation and rule extraction. The featured knowledge extraction architecture will be tested and analysed using several well-known medical datasets.

Keywords: Rule extraction, clustering, discretisation, supervised learning, unsupervised learning

1. INTRODUCTION

Knowledge representation in the form of symbolic rulesets are appealing in a data mining context primarily due to their simplicity and unambiguity, and has long been central to research in artificial intelligence (AI) and knowledge engineering. Healthcare and medicine have traditionally provided a *problem-rich* environments for the evaluation of such advanced methodologies; and readily provides datasets for which analysis would be complicated by the occurrence of data heterogeneity and incompleteness—respectively resulting from datavectors being composed from combinations of symbolic, discrete and continuous-valued attributes—and degraded by the presence of noise, statistically anomalous values and missing attributes. Classification information pertaining to individual datavectors is also often unavailable—or at best available only in some imprecise *rule-of-thumb* form—in many healthcare problems, particular those which require highly specialised domain experts. The research described in this paper is motivated by the desirability of obtaining conceptually symbolic *clean* descriptions from real-life (by definition less than ideal) healthcare data, which would subsequently be useful in the construction of rule-based expert systems. Our featured methodology is a combinative application of various supervised and unsupervised learning mechanisms so as to enable rule-based knowledge extraction under some fairly generic application scenarios.

2. OVERVIEW AND ANALYSIS OF INDIVIDUAL PIPELINE MECHANISMS

2.1 Datavector Clustering: K-Means

We presume a data collection process resulting in unannotated datasets i.e. undifferentiated collection of multi-component datavectors $S = \{\mathbf{X}_i : \; i \in [1, n]\}$, for which the classification attribute $c(\mathbf{X}_i) = \alpha$ for $\alpha \in [1, k]$ is unknown. Our methodology assumes the possibility of deducing the value of classification attribute from information intrinsic within the datavector itself, which is usually a reasonable assumption for most datasets with ordinal or continuous-valued datavector components. A dataset satisfying this presumption would be divisible into k clusters, with cluster membership determined by datavector association i.e. k-means clustering. This well-known algorithm [1][17] allows for the division of the dataset into cluster-subsets i.e. $S = \bigcup_{\alpha} S_\alpha$, with each distinct cluster,

$$S_\alpha = \{\mathbf{X}_i^{(\alpha)} : \; c(\mathbf{X}_i^{(\alpha)}) = \alpha, \quad i \in [1, n_\alpha]\}$$

characterised by attributes:-

- Mean *(centre-of-mass)*: $\mu_\alpha = \frac{\sum_i x_i^{(\alpha)}}{n_\alpha}$
- Variance *(radius)* : $\sigma_\alpha = \sqrt{\frac{\sum_i \left| x_i^{(\alpha)} - \mu_\alpha \right|^2}{n_\alpha}}$
- Average pair-wise distance *(diameter)*: $\delta_\alpha = \sqrt{\frac{\sum_{i,j} \left| x_i^{(\alpha)} - x_j^{(\alpha)} \right|^2}{n_\alpha (n_\alpha - 1)}}$

with the geometric analogues provided by Zhang-Ramakrishnan-Livny [4].

Dataset clustering via k-means is initiated via the random assignment of all datavectors to the presumed set of constituent clusters, after which the cluster-specific attributes are computed. Each subsequent computational round would then reassign the cluster membership for each datavector $x_i \in S$ such that $\left| x_i^{(\alpha)} - \mu_\alpha \right|$ is minimised for the fresh α value. A set-wide error—which can be regarded as the merit criteria associated with the cluster representation of the dataset—based on differences between individual datavectors and the cluster-means is also computed after each round, thereby allowing for algorithm termination when the global error drops below some specified threshold or alternatively when stable cluster membership prevents further reduction in the global error.

The popularity of k-means in data mining applications is due (in large part) to the basic procedural simplicity, which frequently lends itself to many variations i.e. tree-like hierarchical clustering as suggested by [4]. The set-cluster representation is, on the other hand, non-deterministic (due to the random initial membership) and dependant on k i.e. the externally specified cluster multiplicity. There are two diametrically opposed strategies to deal with this issue i.e.:-

- Over-estimation in the initial k value followed by progressive amalgamation of clusters (thereby resulting in a reduced k) with closely-separated centres as defined by the comparison of the cluster-means with (σ, δ) values for both clusters
- Under-estimation in the initial k value followed by progressive partitioning of *loose* clusters (thereby resulting in an increased k) as defined by large (σ, δ) values

both of which (in common to basic k-means) would also require multiple iterations to select the *best* cluster representation i.e. resulting in the lowest global error. It should, however, be pointed out that cluster multiplicity can often be deduced for many data mining problems; as would justifiably be the case for both featured datasets i.e. the widely analysed *Wisconsin Breast-Cancer* and *New-Thyroid*, which respectively have two (benign and malignant) and three (normal, hyperthyroidal and hypothyroidal) classification values.

2.2 Attribute Discretisation: Chi-2 and MDL Partitioning

Discretisation of ordinal or continuous-valued datavector components is motivated by the desirability of a discrete-valued or binarised input attribute representation for the subsequent rule-extraction phase. Attribute discretisation procedures can be characterised as being *supervised* or *unsupervised*—respectively applied on datavectors for which the classification attribute is known and unknown—the former of which are considered to be far more compute-efficient. Supervised methods assume the successful computation of $c(\vec{x}_i^{(\alpha)}) = \alpha$—via k-means or an alternative clustering algorithm—following which statistical or information-theoretic discretisation can be executed on each component of multidimensional datavector $\vec{x}_i^{(\alpha)} = [\cdots, \ x_i^{(\alpha)}, \ \cdots]$. Statistical optimisation via Chi-2 [2] and entropy reduction via MDL partitioning [5] [6] can both be employed as pipeline mechanisms, and their respective bottom-up and top-down approaches provides for an interesting contrast.

Chi-2—which can be regarded as a *dynamic* refinement of its *static* predecessor i.e. Kerber's Chi-merge—generates a discretised attribute representation from the progressive pairwise merging of adjacent attribute-value intervals with the lowest significance level as indicated by the χ^2 statistical parameter. Interval merging would then commence from pairs of individual datavector attribute-values so as to maximise first vector-wise and subsequently component-specific χ^2 values. The resultant discretised vector would be of form $\vec{d}(\vec{x}_i) = [\cdots, \ d(x_i), \ \cdots]$ with individual components $d(x_i) \in \{\cdots, \ \mu_j^{(i)}, \ \cdots\}$ represented by the means of all attribute-values within the discretised interval. The

primary termination criteria for Chi-2 is motivated by the desirability of minimising class-erroneous discretisations i.e. identical representations $d(\vec{X}_i^{(\alpha)}) = d(\vec{X}_j^{(\beta)})$ for unequal classification values $\alpha \neq \beta$. Note the possibility of discretisation down to a single interval (encompassing all attribute-values) frequently occurs for at least some of the datavector components, the occurrence of which indicates the insignificance of that particular component for classification purposes. Chi-2 discretisation was applied in two previous works, rule extraction from neural network [8] and synthesis of maximal decision rules using rough sets [9].

MDL partitioning, on the other hand, proceeds via the identification of partition point p with which to bisect datavector component $X_i \in S = S_+ \cup S_-$ such that $X_i^{(\pm)} \in S_\pm = \{X_i : \begin{matrix} X_i > p \\ X_i < p \end{matrix}\}$. The basic idea is to select p so that the entropy of the bisected dataset i.e.

$$E(S_\pm, p) = \frac{|S_+|}{|S|} E(S_+) + \frac{|S_-|}{|S|} E(S_-)$$

is minimised. This procedure is applied repeatedly on a particular datavector component until further bisection no longer results in entropy reduction. Note that both statistical and information-theoretic methods result in dataset reduction from:-

- Dimensional: via identification of classification-insignificant components
- Numerical: due to $d(\vec{X}_i) = d(\vec{X}_j)$ for $i \neq j$,

viewpoints, with Chi-2 resulting in far more *aggressive* discretisation compared to MDL partitioning. This is empirically demonstrated by the approximate 90 % reduction—in both *Wisconsin Breast-Cancer* and *New-Thyroid* datasets—under Chi-2, as opposed the corresponding 40-70 % reduction under MDL partitioning.

2.3 Supervised Rule Extraction: LC Interpretation and Rough Sets

Generalised symbolic descriptions of annotated datasets can be obtained from a variety of supervised NN training and rule extraction methodologies—as systematically classified by the Andrews-Diederich-Tickle (ADT) [10] taxonomy—or alternatively using rough set analysis as outlined by Pawlak. The Rulex procedure formulated by Andrews-Geva [3] on LC networks with localised Radial Basis Function (RBF) like activation functions constructed using sigmoidal building blocks is employed within the pipeline due to the relatively fast LC parametric convergence during training, and also the *natural* manner in which rules are computed from trained LC network parameters. The latter feature provides for a notable contrast to the majority of NN-based rule extraction frameworks with distinct symbolic formulation processes executed after satisfactory completion of NN training. Usage of LC networks with localised activation—as opposed the more widely used Multi-Layer Perceptron (MLP) networks with globalised sigmoidal activation—can also be considered to be strongly motivated by the assignment of datavector class attributes via cluster membership, which is itself a localised process. LC network training followed by Rulex symbolic formulation can therefore be expected to perform adequately within the context of our data mining pipeline.

LC networks are composed of sigmoid pairs $p_i(x_i) = \left(1+e^{-k_i(x_i-c_i+b_i)}\right)^{-1} - \left(1+e^{-k_i(x_i-c_i-b_i)}\right)^{-1}$ restricted to a single dimension in a multi-dimensional datavector $\vec{x} = [\cdots, \ x_i, \ \cdots]$, with i used as the component dimension (as opposed to datavector) index. Note the localised functionality i.e. $p_i(x_i) \cong 0$ far away—as defined by the exponential suppression k_i—from the interval $[c_i - b_i, \ c_i + b_i]$ with centre c_i and breadth b_i. These unidimensional pairs are subsequently combined by a sigmoidally-activated output node of form $q(\vec{x}) = \left(1+e^{-\kappa\left(\sum_i p_i(x_i) - \pi\right)}\right)^{-1}$; with each individual p_i (retroactively interpretable as a hidden node) localised in the i-th component, but contributing a non-localised *ridge*-like projection in the other datavectors components $[\cdots, \ X_j, \ \cdots]$ (with $j \neq i$). The π parameter is intended to cancel-out all such ridges resulting from $\sum_i p_i$, thereby resulting in the sigmoidal superposition being localised *near* multi-dimensional interval $[\vec{c} - \vec{b}, \ \vec{c} + \vec{b}]$. Output node activation would therefore only occur in that interval, with π being dependent on datavector dimensionality and κ determining the *abruptness* of the activation interval.

Restricted LC training would entail the progressive re-estimation of free parameters $[\vec{c},\ \vec{b},\ \vec{k}]$ so as to maximise gradient descent with respect a set-wide error based on differences between computed output, $q(\vec{x})$ and the previously assigned (via k-means) classification attribute. Rulex subsequently employs the post-training LC parameters to establish class membership for datavectors in $[\vec{X}^{(-)},\ \vec{X}^{(+)}]$, with $x_i^{(\pm)} = c_i \pm r_i^{(\pm)}$ the upper and lower activation thresholds. Effective LC training was found to require a reasonably sizeable training dataset, hence the appropriateness of MDL discretisation prior to Rulex. For the *Wisconsin Breast-Cancer* and *New-Thyroid* datasets, this algorithmic combination results in symbolic rules of significantly greater accuracy compared to Rulex preceded by Chi-2 discretisation. There would also seem to exist the non-negligible eventuality of LC training failure using Chi-2 discretised datasets, as arose for the *New-Thyroid* dataset.

Rough set theory generalises the notion of conventional set membership so that individual datavectors can be assigned to lower and upper approximations, respectively denoted $\underline{S_\alpha}$ and $\overline{S_\alpha}$ with $\overline{S_\alpha} = \underline{S_\alpha} = S_\alpha$ allowing recovery of *crisp* conventional sets. Class-uncertainty—resulting from ambiguity in the classification attribute after discretisation—would result in datavector membership in the boundary region $\overline{S_\alpha} - \underline{S_\alpha}$. Rule extraction from rough sets is based on the construction of reducts i.e. minimal attribute-sets that allow for the identification of datavectors in a particular subset from the rest of the dataset. Reducts are essentially realisations of discernability relations specifying class membership, and are calculable from equivalence classes which are usually generated using genetic algorithm (GA) based methods. Reducts and the extracted rules are characterised by:-

- *Support-level*: indicative of their applicability with respect the dataset
- *Length*: number of attributes required for class-equivalence discernibility,

with the major operational complication arising from rule redundancy and questionable utility of reducts which only apply towards relatively small datavector subsets. Rule extraction should therefore be preceeded by the elimination of reducts with low support-level and short length, with the latter criteria necessary so as to avoid the emergence of over-specific rules. For the *Wisconsin Breast-Cancer* and *New-Thyroid* datasets, it was determined that reduct elimination (exceeding 90 % in some cases) can be safely executed without significantly jeopardising the accuracy (exceeding 70 % in all cases) of the extracted ruleset.

3. DESCRIPTION OF HYBRIDISED FRAMEWORK

3.1 Multi-Mechanism Architecture

The proposed rule-extraction framework would feature sequential application of the following processes:

(1) Unsupervised cluster formation: k-means
(2) Supervised attribute discretisation: Chi-2 or MDL partitioning
(3) Supervised training: restricted LC networks or reduct formation from rough sets
(4) Rule extraction: Rulex on LC networks or reduct-based rule formation

as illustrated in Fig 1 above, which would (in an actual operational environment) be preceded by a filter stage for the execution of routine preprocessing tasks i.e. attribute scaling (to mitigate against numerically large datavector components being disproportionately influential), and the elimination of incomplete (from a descriptive viewpoint) or unlikely (via identification of statistical outliers) datavectors. Note the usage of the computed classification attribute from the first processing stage in the next two, eventually resulting in a symbolic ruleset for each distinct classification attribute $\alpha \in [1, k]$. This enables the usage of the featured framework in support of data mining applications where the classification is a priori unknown.

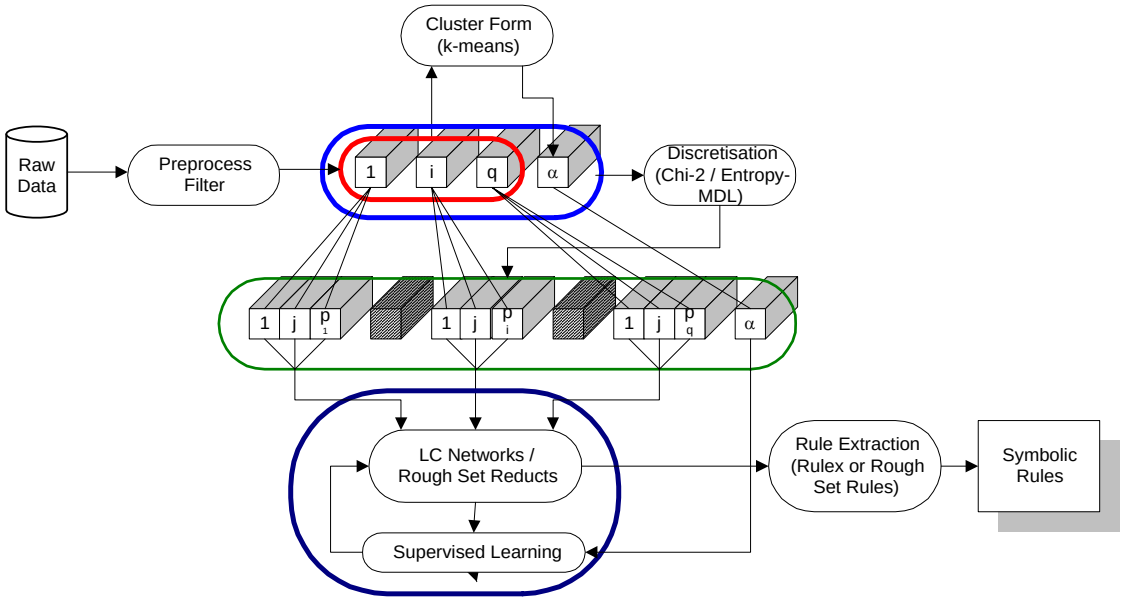

Figure 1: Architecture of rule extraction pipeline

We are currently investigating the effectiveness of the proposed framework on datasets with a mixture of categorical and ordinal/continuous-valued datavector attributes, which would constitute an important generalisation over purely ordinal/continuous-valued datavectors. The categorical datavector components can be handled separately and introduced directly into the NN training stage, however this assumes (without much justification) that computation of classification attribute primarily depends on the ordinal/continuous components. A more well-founded approach would entail class computation from the categorical components (for instance using methods based on rough set [7] analysis), in parallel with the cluster formation process for the ordinal/continuous components. Classification information obtained from independant analysis of categorical and ordinal/continuous attributes would subsequently have to be integrated, following which various downstream analytic processes (as previously indicated) can be executed. Sensitivity analysis [12]—based on computations of first and second derivatives of the classification error in trained NNs—has been demonstrated as being effective in determining the relative importance (with respect classification) of individual attributes, and would conceivably be useful in such a capacity.

3.2 Experimental Data

The *Wisconsin Breast-Cancer* and *New-Thyroid* datasets (both obtained from UCI machine learning repository) were chosen due to all their datavector components being ordinal/continuous-valued, with the respective characteristics indicated in Table 1 below.

The provided classification information is required for an evaluation of the k-means process, which is highly accurate for both datasets. Both class-subsets in *Wisconsin Breast-Cancer* and all three in *New-Thyroid* are also well-separated—i.e. with inter-mean distances fairly large compared to the radii or diameters—thereby allowing for a reasonable degree of optimism with respect the accuracy of the extracted symbolic rulesets. Following this Chi-2 and MDL partitioning is executed, thereby resulting in dataset reduction as indicated in Table 2 below.

The provided classification information is required for an evaluation of the k-means process, which is highly accurate for both datasets. Both class-subsets in *Wisconsin Breast-Cancer* and all three in *New-Thyroid* are also well-separated—i.e. with inter-mean distances fairly large compared to the radii or diameters—thereby allowing for a reasonable degree of optimism with respect the accuracy of the extracted symbolic rulesets. Following

Dataset	*Dataset size*	*Datavector Attributes*	*Classifications*	*Clustering accuracy*
Wisconsin Breast-Cancer	683	9	2	96 %
New-Thyroid	215	4	3	86 %

Table 1: Dataset characteristics

this Chi-2 and MDL partitioning is executed, thereby resulting in dataset reduction as indicated in Table 2 below.

	Discretisation method	
Dataset	*Chi-2*	*MDL*
Wisconsin Breast-Cancer	91 %	72 %
New-Thyroid	88 %	41 %

Table 2: Dataset reduction after attribute discretisation

Chi-2 also allows for the outright elimination (as being classification-irrelevant) of datavector attributes i.e. 2 out of 9 for *Wisconsin Breast-Cancer* and 2 out of 5 for *New-Thyroid*. The relative inter-class datavector distribution can also be demonstrated to be qualitatively preserved, with the notable exception of *Wisconsin Breast-Cancer* under MDL discretisation. This did not, however, seem to affect the overall effectiveness of the rule extraction pipeline; as will be demonstrated by subsequently presented experimental data. The discretised datavectors and the cluster-assigned classification attributes are then divided into training and test datasets.

LC network training followed by Rulex seems to result in a relatively compact ruleset, in contrast to the rough set analysis which would generate an overlarge ruleset without reduct elimination as previously discussed. The level of ruleset reduction possible using dataset-specific criteria is indicated in Table 3 below, with the end result being a reasonably small (and therefore conceptually useful) symbolic ruleset. The effectiveness (with respect the test dataset) of all possible algorithm combinations—i.e. Chi-2 and MDL discretisations with LC network and rough set rule extraction—is presented in Table 4 below, with the only problem being the LC network non-trainability using the Chi-2 discretised *New-Thyroid* dataset. Note that the featured algorithm combinations would for the most part result in acceptable (exceeding 70%) classification performance. We are currently evaluating the effectiveness of the presented framework on other larger and more complex datasets, and hope to report the results thereof in a subsequent publication.

4. CONCLUDING REMARKS

The proposed sequential application of:-

(1) Classification via datavector clustering
(2) Feature selection and data simplification via discretisation
(3) Knowledge extraction via supervised learning and symbolic rule generation

appears to be a fundamentally sound methodology for the analysis of unannotated datavectors with ordinal or continuous-valued attributes. Such attribute values would be a natural consequence of instrumentalised data collection, and as such would constitute an important sub-category of healthcare problems. Expert systems constructed from the extracted rules would provide a useful operational tool, particularly in a decision-support capacity when specialist expertise is not readily available. Knowledge extraction from a more generic data analytic framework—i.e. one able to handle both categorical and ordinal/continuous attributes on an equivalent basis—would be even more useful, and we anticipate this being an interesting line of research to undertake.

	Discreti -sation	*Reducts before elimination*			*Elimination criteria*	*Reducts after elimination*			*Ruleset size*
Dataset		*Number*	*Support*	*Length*		*Number*	*Support*	*Length*	
Wiscon-sin Breast-Cancer	*Chi-2*	9	1-12	5-6	Support > 2 Length < 6	1	3	5	5
	MDL	27	1-18	4-8	Support > 2 Length < 6	2	3	5	4
New-Thyroid	*Chi-2*	1	12	2	n/a	1	12	2	8
	MDL	8	1-5	3-4	Support > 4 Length < 4	1	5	3	11

Table 3: Reduct elimination prior to rule extraction

	Algorithmic methods			
Dataset	*Chi-2/ LC network*	*Chi-2/ Rough set*	*MDL/ LC network*	*MDL/ Rough set*
Wisconsin Breast-Cancer	69 %	75 %	92 %	71 %
New-Thyroid	n/a	77 %	83 %	70 %

Table 4: Test classification accuracy of extracted rulesets

ACKNOWLEDGEMENTS

LC networks and the RULEX algorithm implementations used in our experiments was by Robert Andrews and Shlomo Geva, and were obtained from [15]. MDL discretization and rough sets rule generation was performed on the ROSETTA [13] software system for data analysis obtained from [14]. Reducts are generated using the in-built Dynamic Reducts by genetic algorithm method implemented in the RSES library, developed at the Group of Logic, University of Warsaw, Poland.

REFERENCES

[1] Léon Bottou and Yoshua Bengio, "Convergence Properties of the K-Means Algorithms", Proc. of 7th Conf. on Neural Information Processing Systems, Denver, USA, 1994.

[2] Huan Liu and Rudy Setiono, "Chi2: Feature Selection and Discretization of Numeric Attributes", *Proc. 7th International Conference on Tools with Artificial Intelligence*, Washington D.C., 1995.

[3] Robert Andrews and Shlomo Geva, "Rule Extraction from Local Cluster Neural Nets", submitted to Neurocomputing, February, 2000.

[4] Tian Zhang, Raghu Ramakrishnan, and Miron Livny, "BIRCH: An Efficient Data Clustering Method For Large Databases", Proc. of 1996 ACM-SIGMOD Int. Conf. on Management of Data, Montreal, Quebec, 1996.

[5] James Dougherty, Ron Kohavi and Mehran Sahami, "Supervised and Unsupervised Discretization of Continuous Features", Proc. Machine Learning the 12th Int. Conf., 1995.

[6] Ron Kohavi and Mehran Sahami, "Error-based and Entropy-based Discretization of Continuous Features", Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, pages 114-119, AAAI Press, August 1996.

[7] Jan Komorowski and Aleksander Øhrn, "An Introduction to Rough Sets", Proc. of ECAI 98 Workshop on Synthesis of Intelligent Systems from Experimental Data, 1998.

[8] Ismail A. Taha and Joydeep Ghosh, "Symbolic Interpretation of Artificial Neural Networks", IEEE Trans. Knowledge and Data Engineering, Vol.11, No.3, pp. 448-463, May/June 1999.

[9] Xiaohua Hu and Nick Cercone, "Learning Maximal Generalized Decision Rules via Discretization, Generalization and Rough Set Feature Selection", Proc. 9th Int. Conf. on Tools with Artificial Intelligence (TAI '97), 1997.

[10] Alan B. Tickle, Robert Andrews, Mostefa Golea and Joachim Diederich, "The Truth Will Come To Light: Directions and Challenges in Extracting the Knowledge Embedded Within Trained Artificial Neural Networks", IEEE Trans Neural Networks Vol.9, No.6, pp.1057-1068, 1998.

[11] Alwyn Goh, Hoe Kok Meng and Jagdesh Singh, "Hybrid Kohonen Multi-Layer Perceptron Neural Network System for Extracting Symbolic Classification Rules from Unannotated Datasets", Proc. Comp Sc and Info Tech Conf, Confed of Scientific and Technological Associations in Malaysia (COSTAM)., Penang, Malaysia, 1998.

[12] Jingtao Yao, Nicholas Teng, Hean-Lee Poh, and Chew Lim Tan, "Forecasting and Analysis of Marketing Data Using Neural Networks", Journal of Information Science and Engineering, Vol. 14, pp. 843-862, 1998.

[13] Aleksander Øhrn, "Discernibility and Rough Sets in Medicine: Tools and Applications", Ph.D. Thesis, Norwegian Uni. of Sc. and Tech., Trondheim, Norway, 1999.

[14] The ROSETTA Software System Homepage, http://www.idi.ntnu.no/~aleks/rosetta/rosetta.html

[15] Local cluster neural networks, RBP and RULEX rule extraction software, http://www.fit.qut.edu.au/~robert/rulexsoftware.html

[16] Dorian Pyle, Data Preparation for Data Mining, San Francisco, CA: Morgan Kaufmann. 1999.

[17] John A. Hartigan, Clustering Algorithms, New York: John Wiley & Sons. 1975.