

Integrate Variable Precision Rough Sets and Modified PBMF Index Function for Partitioning and Classifying Complex Datasets

Kuang Yu Huang

Department of Information Management
Ling Tung University
#1 Ling Tung Road, Taichung City 408, Taiwan
kyhuang@mail.ltu.edu.tw

*Yu-Hsin Cheng

Department of Information Networking and System
Administration, Ling Tung University
#1 Ling Tung Road, Taichung City 408, Taiwan
hsin@teamail.ltu.edu.tw

Abstract—This study proposes a method for partitioning and classifying complex datasets using a hybrid method based on Fuzzy C-Means (FCM) method, Variable Precision Rough Set (VPRS) theory and a modified form of the PBMF index function (a cluster validity index function). The proposed VPRS index method partitions the attributes within the dataset rather than the data and achieves both the optimal number of clusters and the optimal classification accuracy. The validity of the proposed approach is confirmed by comparing the clustering results obtained from the VPRS method for a hypothetical function and a typical stock market system with those obtained from the conventional RS and PBMF methods, respectively. Overall, the results show that the VPRS index method not only has a better clustering performance than the PBMF method, but also achieves greater classification accuracy, and therefore provides a more reliable basis for the extraction of decision-making rules.

Keywords—Fuzzy C-Means; Variable Precision Rough Set; VPRS index method; PBMF-index method; Classification; Cluster

I. INTRODUCTION

Classification, which is a task of grouping data with multiple attributes into relevant categories, has become the most valuable process to learn the implicit knowledge among dataset. However, difficulties have been found in many classification approaches, and one of the major problems is related to clustering—partition the real-value attributes into discrete groups. In order to conduct a dataset classification, attribute clustering must be performed in prior, and accurate clustering is the premise to meaningful classifications.

In general, the problem of evaluating the optimality of the clustering results obtained for a particular dataset is referred to as the cluster validity problem [1]. Many methods have been proposed for assessing the validity of the clustering results obtained using fuzzy clustering schemes [2]. Amongst these methods, early indices such as the partition coefficient [3, 4] and classification entropy [5, 6] are based simply on the membership values of the instances within the dataset and are therefore computationally straightforward. In general, existing clustering methods cluster the dataset in accordance with the

norms of the instances rather than the values of the individual attributes. The results obtained using traditional clustering methods fail to take sufficient account of the complex interrelationships amongst the various attributes of the dataset. Therefore, in constructing data partitioning and classification algorithms, it is preferable to apply some form of classification-defining knowledge to the attribute values of the dataset such that the complex interrelationships amongst the various attributes can be taken more properly into account.

The literature contains many algorithms for automatic classification purposes, including decision-tree algorithms such as ID3 [7], rule-based algorithms such as CN2 [8], back-propagation networks [9], support vector machines [10], conformal predictors [11], Bayesian classifiers [12], and so forth. Whilst such algorithms have found widespread use in a diverse range of applications, including weather prediction, manufacturing process planning, medical diagnosis, and so on, they can not deal effectively with systems characterized by uncertain or missing information. Furthermore, to the best of the current authors' knowledge, only few attempts [13] have so far been made to integrate these classification algorithms with any form of cluster validity index in order to evaluate the optimality of the clustering results.

Rough Set (RS) theory was first introduced more than twenty years ago [14] and has emerged as a powerful technique for the automatic classification of instances [15]. However, performing a classification operation with a controlled degree of uncertainty or misclassification error is beyond the ability of the RS approach [16]. In an attempt to extend RS theory to such classification applications, Ziarko proposed a Variable Precision Rough Set (VPRS) methodology in which the instances within the dataset were analyzed and classified in terms of their statistical tendencies rather than their functional patterns [16, 17]. In VPRS theory, the uncertain nature of the information within the system of interest is handled using the concept of β -lower and β -upper approximate sets. However, as with the RS method, the classification performance of VPRS models is constrained by the quality of the original clustering results. Therefore, in meeting the needs of decision-makers for

* Corresponding author

sophisticated methods capable of extracting reliable decision rules from complex information systems, it is necessary to integrate the VPRS method with some form of optimized clustering scheme.

Accordingly, this study proposes a hybrid scheme for the partitioning and classification of complex datasets based upon the FCM clustering method and a VPRS model in which the optimal threshold parameter β is determined using the method presented by Huang [18]. In the proposed approach, the dataset is partitioned in terms of the attribute values of the instances rather than the norms of the instances in order to capture the complex interactions amongst the various instances.

The remainder of this paper is organized as follows. Section II presents the fundamental principles of the FCM method, VPRS theory and the conventional PBMF clustering validity index, respectively. Section III describes the integration of these concepts to create the proposed VPRS partitioning and classification scheme. Section IV compares the performance of the proposed scheme with that of the conventional RS index function method and the PBMF (FCM-based) clustering method, respectively. Finally, Section V presents some brief concluding remarks and indicates the intended direction of future research.

II. REVIEW OF RELATED METHODOLOGIES

A. Index function $I_{\max}$

Assume that each instance x_i in the dataset has m conditional attributes and the l -th attribute a_l can be partitioned into p_l clusters, then $C_{a_l}(x_i)$ gives the index of the cluster to which the l -th attribute a_l of instance x_i belongs. Here $C_{a_l}(x_i)$ is given by $C_l(x_i) = I_{\max}(\mu_j(x_i(a_l))) = \text{Index}(\max(\mu_j(x_i(a_l))))$ for $1 \leq l \leq m, 1 \leq i \leq n$, where $I_{\max}(\mu_j(x_i(a_l)))$ returns the index of the cluster corresponding to the maximum value amongst all the membership functions of the l -th attribute of x_i .

B. Fundamental principles of VPRS theory

VPRS operates on what may be described as a knowledge-representation system, or information system [16]. The basic principles and notations of information systems (S) and the application of VPRS theory to the processing of such systems are described in the sections below.

1) β -lower and β -upper approximate sets: For a given dataset, any instances which are indistinguishable from one another when evaluated using a particular subset of all the attributes define an equivalence or indiscernibility relationship. In RS theory, this indiscernibility concept is handled using approximate sets. A typical information system has the form $S = (U, A, V_q, f_q)$, where U is a non-empty finite set of instances, A is a non-empty finite set of attributes describing these instances and $X \subseteq U$ and $R \subseteq A$. Generally speaking, the attributes in set A can be partitioned

into a set of conditional attributes $C \neq \{\}$ and a set of decision attributes $d \neq \{\}$, i.e. $A = C \cup d$ and $C \cap d = \{\}$. For each attribute, $q \in A$, V_q represents the domain of q , i.e. $V = \bigcup V_q$. Finally, $f_q : U \times A \rightarrow V$ is an information function defined such that $f(x, q) \in V_q$ for $\forall q \in A$ and $\forall x \in U$. The VPRS method proposed in this study utilizes the systematic method presented by the current author in [18] to determine an appropriate value of the threshold parameter β , i.e., the value of β at which a certain proportion of the instances in a particular conditional class are classified into the same decision class. When processing an information system using a VPRS model with $0.5 < \beta \leq 1$, the objective is to identify the β -lower and β -upper approximate sets associated with each cluster of the decision attribute. In general, the β -lower approximation of sets $X \subseteq U$ and $P \subseteq C$ can be expressed as

$$\underline{R}_P^\beta(X) = \{x \in U : P(X/[x]_P) \geq \beta\} = \bigcup \{[x]_P : P(X/[x]_P) \geq \beta\}$$

Similarly, the β -upper approximation of sets $X \subseteq U$ and $P \subseteq C$ is given by

$$\overline{R}_P^\beta(X) = \{x \in U : P(X/[x]_P) > 1 - \beta\} = \bigcup \{[x]_P : P(X/[x]_P) > 1 - \beta\}$$

Note that $P(X/Y) = \frac{|X \cap Y|}{|Y|}$ if $|Y| > 0$, and $P(X/Y) = 1$

otherwise. Note also that $|X|$ denotes the cardinality of set X .

In the particular case of $\beta = 1$, $\underline{R}_P^\beta(X)$ and $\overline{R}_P^\beta(X)$ are equivalent to the lower and upper approximate sets in RS theory. In other words, the VPRS model reverts to the traditional RS model.

2) *Accuracy of VPRS classification results*: The accuracy of the VPRS classification results can be quantified as follows:

$$\alpha_c^\beta = \left| \underline{R}_P^\beta(d | C_d(x) = c) \right| / \left| \overline{R}_P^\beta(d | C_d(x) = c) \right|,$$

where $\left| \underline{R}_P^\beta(d | C_d(x) = c) \right|$ and $\left| \overline{R}_P^\beta(d | C_d(x) = c) \right|$ are the cardinalities of the β -lower and β -upper approximate sets obtained when ranking the elements (x) in terms of the c -th cluster of the decision attribute d .

C. PBMF index function

The PBMF cluster validity index function [19] ensures the formation of a small number of compact clusters within the dataset and maximizes the separation distance between at least two of these clusters. The PBMF index function is formulated

as $PBMF(K) = \left(\frac{1}{K} \times \frac{\overline{E}_1}{J_{m'}} \times D_K \right)$, where K is the number of

clusters, $J_{m'} = \sum_{k=1}^K \sum_{j=1}^n \mu_{kj}^{m'} \|x_j - z_k\|$, $\overline{E}_1$ is constant for a given

dataset and is used to prevent the value of the second term from approaching zero, and $D_K = \max_{i,j=1}^K \|z_i - z_j\|$. In addition, n is

the total number of instances in the dataset, $U(X) = \lfloor \mu_{kj} \rfloor_{K \times n}$

is the matrix used to partition the data, m' is the fuzzification parameter, and z_k is the centroid of the k -th cluster. In utilizing the PBMF index function for data clustering purposes, the objective is to find the value of K (i.e. the number of clusters) which maximizes the index value.

III. VPRS INDEX METHOD

This study proposes a hybrid scheme comprising a Variable Precision Rough Set (VPRS) model and a modified form of the PBMF index function for optimizing both the number of clusters within the dataset and the corresponding classification accuracy. Assuming that U is the domain of discourse and R is the set of equivalences of U , the VPRS classification problem can be formulated as

$$X \subseteq U \text{ is : } (\underline{R}_P^\beta(X), \overline{R}_P^\beta(X)), BND_P^\beta(X),$$

where X is the set of instances; U/I_P is the quotient set of U ; I_P is the indiscernibility of R ; $\{\}$ is the empty set; R is the attribute set of X and includes both the conditional attribute set (C) and the decision attribute set (d), $P \subseteq C$; $\underline{R}_P^\beta(X)$ is the β -lower approximate set of X ; $\overline{R}_P^\beta(X)$ is the β -upper approximate set of X ; and $BND_P^\beta(X)$ is the boundary set of X .

Significantly, the VPRS index method proposed in this study clusters the attribute values of the instances within the dataset rather than the norms of the instances. Thus, in contrast to the PBMF index function method, the proposed approach takes better account of the intrinsic interrelationships amongst the various independent parameters of the dataset. The detailed parameters of the VPRS index function are discussed in the following section.

A. Parameters of VPRS index function

The VPRS index function proposed in this study has the form

$$VPRS(N, \alpha_c^\beta) = \left(\frac{1}{N} \times \frac{\overline{E}_1}{F'_N} \times D'_N \right)$$

where N is the number of clusters assigned to the conditional and decision attributes and α_c^β is the classification accuracy when evaluated in terms of the c -th cluster of the decision attribute. In addition, F'_N is obtained by aggregating the value of E'_c for each cluster of the decision attribute (d),

where E'_c is given by $E'_c = \sum_{j=1}^n \overline{\mu}_{cj}^{m'}(x_j(d)) \|x_j - z'_c\| / \alpha_c^\beta$, in

which $\overline{\mu}_{cj}(x_j(d))$ is the membership function of instance x_j in the c -th cluster of the decision attribute d and z'_c is the multi-dimensional centroid of the β -lower approximate sets associated with the c -th cluster of the decision attribute d and is obtained by computing the mean values of the conditional

and decision attributes of each instance within the corresponding sets. Furthermore, m' is the fuzzification parameter and n is the total number of instances in the dataset. Finally, the value of D'_N is equal to the maximum separation distance amongst the centroids of all the β -lower approximate sets associated with the different clusters of the decision attribute, i.e. $D'_N = \max_{i,j=1}^N \|z'_i - z'_j\|$. Note that the value of D'_N is upper bounded by the maximum separation distance amongst all possible pairs of data points in the dataset.

Note that parameter F'_N in the VPRS index function differs slightly from parameter $J_{m'}$ in the PBMF index function in that its value depends on E'_c and therefore takes the classification accuracy into account.

B. Tendencies of terms within VPRS index function

As discussed in Section III.A, the VPRS index function has the form $VPRS(N, \alpha_c^\beta) = \left(\frac{1}{N} \times \frac{\overline{E}_1}{F'_N} \times D'_N \right)$. In other words, the index function comprises three terms, namely, $1/C$, $\overline{E}_1/F'_N$ and D'_N . Clearly, the value of the first term decreases as the number of clusters assigned to the conditional and decision attributes, N , increases. In other words, the value of the index function decreases with an increasing value of N . Regarding the second term in the index function, the value of $\overline{E}_1$ is constant for a given dataset and is equal to E_1 in the conventional PBMF index function. Meanwhile, as discussed in the previous section, F'_N is equal to the sum of all E'_c , where each E'_c includes the classification accuracy α_c^β associated with the c -th cluster of the decision attribute d . Hence, the VPRS index function increases as E'_c decreases. The third term, D'_N , measures the maximum distance between the centroids of the β -lower approximate sets associated with the different clusters of the decision attribute, and increases as N increases. Thus, the contribution of D'_N to the value of the VPRS index function increases as the number of decision attribute clusters increases.

C. Detailed processing steps in proposed VPRS index function method

The details of each processing step are summarized in the following paragraphs.

Step 1: Specify number of clusters per attribute in interval $[2, N_{\max}]$

As described in Section I, the proposed VPRS index method utilizes an iterative procedure to optimize the number of clusters assigned to each conditional and decision attribute

within the system of interest. (Note that the number of conditional attributes to be considered in the optimization process is specified in advance by the user and the number of decision attributes is set by default to one.) As described in Section III, the conditional and decision attributes are partitioned into an equal number of clusters, N , where N is bounded by the interval $[2, N_{\max}]$, in which 2 represents the minimal number of clusters per attribute (and is the default setting) and $N_{\max}$ represents the maximum permissible number of clusters per attribute and is set on a case-by-case basis in such a way as to achieve a satisfactory trade-off between the classification performance and the computational complexity of the solution procedure.

Step 2: Fuzzify attributes of information system using FCM method.

In general, a continuous-value information system can only be converted into an equivalent fuzzy information system once a classified fuzzy set has been obtained. In the FCM clustering procedure performed in the VPRS index function method, the interval values $[\alpha, \beta]$ of all the conditional and decision attributes are assigned to p_l fuzzy clusters. The continuous-value information system (U, A, V_q, f_q) is then converted into the fuzzy information system $(U, \tilde{A}, \Phi, d)$, where $\Phi = \{\tilde{A}_{lj} \mid l \leq m, j \leq p_l\}$, in which $\tilde{A}_{lj} = \mu_j(x_i(a_l))$ and indicates the values of the membership functions associated with the l -th conditional attribute a_l of the i -th instance.

Step 3: Assign each instance attribute to appropriate conditional or decision attribute cluster.

Utilizing the index function $C_{a_l}(x_i) = I_{\max}(\mu_j(x_i(a_l))) = \text{Index}(\max(\mu_j(x_i)))$ for $1 \leq l \leq m, 1 \leq i \leq n$, the membership functions of each attribute of each instance are processed in order to determine the cluster to which the attribute value belongs.

Step 4: Identify VPRS sets and compute corresponding classification accuracy.

Having mapped the attribute values of all the instances to the appropriate conditional or decision attribute clusters, the procedure described in [18] is applied to determine an appropriate value of the threshold parameter β in the VPRS model. The β -lower approximate sets and boundary sets associated with each cluster c of the decision attribute d are then extracted in accordance with the definitions presented in Section II.C. Finally, the classification accuracy associated with each cluster of the decision attribute is evaluated by computing the cardinality ratio of the corresponding β -lower approximate sets to the β -upper approximate sets.

Step 5: Calculate centroids of β -lower approximate sets associated with each cluster of decision attribute.

The multi-dimensional centroids of the β -lower approximate sets associated with each cluster of the decision attribute d are obtained by calculating the mean values of the conditional and decision attributes of all the instances within the corresponding sets.

Step 6: Determine the value of the VPRS cluster validity index.

Having determined the classification accuracy and centroids of the β -lower approximate sets, the optimality of the clustering solution is evaluated using the VPRS cluster validity index given in Section III.A.

Step 7: Check termination criterion.

Having computed the value of the index function for the current number of clusters per attribute N , a check is made to see whether or not N is equal to the upper bound value ($N = N_{\max}$). In the event that N is not equal to $N_{\max}$, the value of N is incremented by 1, and the FCM, VPRS and cluster validity index computation procedures are repeated. Once the termination criterion is satisfied, the iteration procedure ceases and the computational process moves to the final step.

Step 8: Identify value of VPRS cluster validity index.

Once the termination criterion has been satisfied, the values of the VPRS index function obtained for $N = 2 \sim N_{\max}$ are compared. The maximum value of the index function amongst all $2 \sim N_{\max}$ solutions is taken as the VPRS cluster validity index and corresponds to the clustering solution which optimizes both the number of clusters per attribute and the overall classification accuracy of the dataset.

IV. PERFORMANCE EVALUATION OF VPRS INDEX METHOD

In this section, the validity and effectiveness of the proposed VPRS index method is evaluated by considering three illustrative examples, namely one example relating to a dataset in which the conditional and decision attributes are related via a hypothetical function and two examples relating to electronic stock data extracted from the financial database maintained by the Taiwan Economic Journal (TEJ) [18]. In performing the evaluations, the effectiveness of the proposed method is explored by comparing the partitioning and classification results obtained for the three illustrative examples with those obtained from a Rough Set (RS) method and the conventional PBMF index (FCM-based) method, respectively. The PBMF index method clusters the norms of the instances within the dataset, whereas the VPRS and RS index methods cluster the attribute values. Note that the RS index method [13] is a particular case of the VPRS index method in which the cluster indices obtained from the FCM procedure are processed using a RS model rather than the VPRS model.

A. An illustrative example

The performance of the VPRS index method in partitioning and classifying real-world complex datasets was evaluated using electronic stock data extracted from the financial database maintained by the Taiwan Economic Journal (TEJ) [18] for the first quarter of 2008. The database maintains a total

of 53 financial indices for each company. However, for simplicity, the performance evaluations considered just one decision attribute, i.e. EPSNet Income, and 12 conditional attributes, i.e. (i) Business Profit Rate, (ii) PreTax Income %, (iii) Net Nonop.Inc./Rev, (iv) PSPre_Tax Income, (v) Oper.Income/Capital, (vi) PreTax Income/Capital, (vii) DaysA/R Turnover, (viii) DaysInventory Turn, (ix) DaysA/P Turnover, (x) Net Operating Cycle, (xi) PSOperating Income, and (xii) D/E Ratio. Having deleted the records for which one or more fields were incomplete, a total of 327 records were obtained (see Table 1 for representative values of each index for a selected subset of these 327 records). This illustrative example considers the performance improvement obtained in the VPRS method when increasing both the number of attributes considered within the clustering process and the number of clusters assigned to each attribute. Accordingly, the stock market data was processed repeatedly by the FCM clustering process with the number of conditional attributes specified in the range [2,12], respectively, and the number of clusters specified in the range [2,39].

The results presented in Table 2 show that the cardinality of the β -lower approximate sets obtained in the VPRS index method is larger than or equal to that of the lower approximate sets obtained in the RS index method. In other words, the VPRS β -lower approximate sets contain a greater number of instances than the RS lower approximate sets. For the case in which the clustering process is based on a small number of conditional attributes, i.e. {2, 3, 4, 5}, the approximate sets obtained using the RS index method contain a very small number of instances, while those obtained using the VPRS index method contain virtually all of the instances within the dataset. However, some of the instances within the VPRS approximate sets are subject to uncertainty. For example, when the number of conditional attributes and the number of clusters are both specified as 2, the RS index method yields only 3 instances, whereas the VPRS index method clusters all 327 instances. In other words, neither method accurately captures the relationship between the conditional attributes and the decision attribute. Thus, when classifying complex datasets using either method, a greater number of conditional attributes should be considered in the clustering process in order to ensure the completeness and reliability of the clustering results.

V. CONCLUSIONS

This paper has presented a cluster validity index method based on a Variable Precision Rough Set (VPRS) model and a modified form of the PBMF index function for the partitioning and classification of complex, real-world datasets. In the proposed approach, designated as the VPRS index method, the dataset is partitioned in accordance with the specified number of attributes and attribute clusters by the Fuzzy C-Means (FCM) clustering method and the resulting cluster indices are then processed by a VPRS model in order to identify the β -lower approximate sets associated with each cluster of the decision attribute and to evaluate the corresponding classification accuracy. Finally, the quality of the clustering solution is evaluated using a modified form of the PBMF index function. The effectiveness of the proposed approach has been

confirmed by comparing the results obtained for the partitioning and classification of a real-world database containing stock market information with the corresponding results obtained using the RS index method. The evaluation results support the following major conclusions:

- (1) In the PBMF index method, the optimality of the clustering results is evaluated in accordance with (a) the distance between each instance and the cluster centroids, and (b) the maximum separation distance between the cluster centroids. By contrast, in the VPRS index method, the quality of the clustering solution is evaluated in terms of (a) the distance between the instances and the centroids of the β -lower approximate sets associated with each cluster of the decision attribute, (b) the maximum distance between the centroids of the β -lower approximate sets associated with the different clusters of the decision attribute, and (c) the classification accuracy of the clustering results as evaluated by the cardinality ratio of the β -lower approximate sets to the β -upper approximate sets for each cluster of the decision attribute. In other words, in contrast to the conventional PBMF index method which only determines the optimal clustering of the data within the dataset given a specified number of clusters, the VPRS index method determines both the optimal number of attribute clusters within the dataset and the optimal classification accuracy.
- (2) The PBMF index method clusters the norms of the instances within the dataset, whereas the VPRS and RS index methods cluster the attribute values. As a result, the VPRS and RS index methods take better account of the intrinsic interrelationships amongst the various conditional and decision attributes in the dataset when partitioning the instances and prevent the clustering results from being dominated by the attribute(s) with a higher order of magnitude.
- (3) When using the RS and VPRS index methods to classify complex datasets based on a limited number of attributes, the cardinality of the β -lower approximate sets obtained in the VPRS index method is larger than or equal to that of the lower approximate sets obtained in the RS index method. As the number of attributes considered in the clustering process is increased, the difference between the cardinality of the β -lower approximate sets and that of the lower approximate sets classified by the two methods reduces. However, the number of instances classified by the VPRS method is still greater than that classified by the RS method.
- (4) Although the VPRS index method classifies a far larger subset of the instances than the RS method when used to cluster complex datasets on the basis of a limited number of attributes, some of the

instances are characterized by uncertainty. In other words, the VPRS model fails to capture the true relationship between the conditional attributes and the decision attribute. Therefore, a greater number of conditional attributes are required in order to accurately model the interdependencies between the conditional attributes and decision attribute of the dataset.

Overall, the results presented in this study have confirmed that the proposed VPRS index method provides an effective tool for optimizing both the number of attribute clusters and the classification accuracy when applied to the partitioning and classification of complex, real-world datasets. However, the quality of the clustering results is degraded as the number of attributes considered in the clustering process is reduced. Therefore, reducing the number of uncertain instances in the β -lower approximate sets of the dataset is essential in improving the performance of the proposed method. Thus, a future study will explore the feasibility of improving the performance of the proposed VPRS classification method by processing (simplifying) the dataset utilizing a β -reducts method when applying the FCM, VPRS and cluster validity index function computation procedures.

REFERENCES

- [1] J.C. Bezdek, "Cluster validity with fuzzy sets," J. Cybernet., vol. 3, pp. 58-74, 1974.
- [2] W. Wang, and Y. Zhang, "On fuzzy cluster validity indices," Fuzzy Sets and Systems, vol. 158, pp. 2095-2117, 2007.
- [3] E. Trauwaert, "On the meaning of Dunn's partition coefficient for fuzzy clusters," Fuzzy Sets and Systems, vol. 25, pp. 217-242, 1988.
- [4] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.
- [5] J. C. Bezdek, "Cluster validity with fuzzy sets," J. Cybernet., vol. 3, pp. 58-73, 1974.
- [6] J. C. Bezdek, "Numerical taxonomy with fuzzy sets," J. Math. Biol., vol. 1, pp. 57-71, 1974.
- [7] J. Ross Quinlan, Induction of decision trees. Machine Learning, Kluwer Academic Publishers, vol.1, 1986, pp. 85-106.
- [8] P. Clark, and T. Niblett, The CN2 induction algorithm. Machine Learning, Kluwer Academic Publishers, vol. 3(4), 1989, pp. 261-283.
- [9] D.E. Rumelhart, G.E. Hinton, and R.J. Williams, "Learning internal representations by error propagation," In: Rumelhart DE, McClelland JL, editors. Parallel distributed processing: explorations in the microstructure of cognition. Cambridge: MIT Press, pp. 318-363, 1986.
- [10] V.N. Vapnik, The nature of statistical learning theory, New York: Springer, 2000.
- [11] V. Vovk, A. Gammerman, and G. Shafer, Algorithmic learning in a random world, New York: Springer, 2005.
- [12] N. Friedman, D. Geiger, and M. Goldsmidt, Bayesian network classifiers. Machine Learning, vol. 29(2), 1997, pp. 131-163.
- [13] K. Y. Huang, "Applications of an Enhanced Cluster Validity Index method based on the Fuzzy C-means and Rough Set Theories to Partition and Classification," Expert Systems With Applications, vol. 37, pp. 8757-8769, 2010.
- [14] Z. Pawlak, "Rough sets," International Journal of Information and Computer Sciences, vol. 11(5), pp. 341-356, 1982.
- [15] K. Y. Huang, and C. J. Jane, "A Hybrid Model for Stock Market Forecasting and Portfolio Selection Based on ARX, Grey System and RS Theories," Expert Systems With Applications, vol. 36, pp. 5387-5392, 2009.
- [16] W. Ziarko, "Variable precision rough set model," Journal of Computer and System Sciences, vol. 46, pp. 39-59, 1993.
- [17] W. Ziarko, "Probabilistic decision tables in the variable precision rough set model," Computational Intelligence, vol. 17, pp. 593-603, 2001.
- [18] K. Y. Huang, "Application of VPRS model with enhanced threshold parameter selection mechanism to automatic stock market forecasting and portfolio selection," Expert Systems With Applications, vol. 36, pp. 11652-11661, 2009.
- [19] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, "A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification," Fuzzy Sets and Systems, vol. 155, pp. 191-214, 2005.

Table1 Illustrative financial data records extracted from TEJ database for first quarter in 2008

Company codes	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)	(m)
1	-2.17	-3.78	-1.61	-0.01	-0.04	-0.07	96.85	252.7	6.8	342.7	0	97.51	0.16
2	-1.58	0.34	1.92	0.01	-0.38	0.08	90.42	56.26	38.85	107.8	-0.04	37.07	0.73
3	-1.58	82.01	80.67	2.42	0.38	23.33	60.36	91.89	38.72	113.5	0.04	141.5	-0.07
4	37.16	31.04	-6.12	2.6	31.07	25.69	85.92	84.46	57.44	112.9	3.11	55.07	9.16
...	...	...	...	...	...	...	...	...	...	...	...	...	...
324	8.17	9.37	1.2	1.35	11.68	13.4	102.0	26.22	44.16	84.02	1.21	53.21	6.18
325	7.38	7.37	-0.01	1.05	10.56	10.54	43.8	21.31	29.23	35.87	1.06	20.97	3.73
326	0.08	-0.97	-1.05	-0.07	0.06	-0.69	102.0	15.24	88.44	28.77	0.01	33.46	3.13
327	1.99	3.41	1.42	0.32	1.84	3.16	46.88	129.5	63.08	113.3	0.18	192.8	2.24

Column headings: (a) Business Profit Rate (b) Pre-Tax Income % (c) Net Non-op.Inc./Rev. (d) PS-Pre_Tax Income (e) Oper.Income/Capital (f) Pre Tax Income/Capital (g) Days-A/R Turnover (h) Days-Inventory Turn. (i) Days-A/P Turnover (j) Net Operating Cycle (k) PS-Operating Income (l) D/E Ratio (m) EPS-Net Income

Table 2 Cardinalities of VPRS β -lower approximate sets and RS lower approximate sets for financial data extracted from TEJ database as function of number of clusters per attribute and number of conditional attributes. (Note that entries in parentheses indicate the cardinality of the RS lower approximate sets.)

No. of clusters	No. of conditional attributes										
	2	3	4	5	6	7	8	9	10	11	12
2	327(3)	327(62)	327(62)	303(77)	303(131)	325(164)	327(176)	325(188)	325(191)	325(191)	325(191)
3	306(3)	325(35)	325(38)	315(46)	315(49)	319(91)	317(128)	304(200)	304(223)	304(223)	312(215)
4	255(12)	242(16)	288(40)	314(74)	312(80)	319(134)	304(212)	309(269)	303(281)	303(281)	313(304)
5	288(12)	187(12)	318(37)	315(53)	306(80)	303(150)	305(238)	301(268)	305(305)	307(291)	311(307)
6	61(11)	102(19)	288(45)	287(76)	276(73)	275(174)	283(240)	303(294)	315(305)	317(315)	323(321)
7	81(12)	129(14)	299(71)	292(105)	295(98)	274(210)	315(294)	321(315)	321(318)	323(320)	325(325)
8	105(15)	97(28)	255(85)	253(126)	233(134)	273(239)	313(313)	319(319)	319(319)	321(319)	327(327)
9	60(16)	80(38)	185(105)	251(148)	262(191)	280(271)	317(317)	321(323)	325(323)	325(325)	327(327)
10	35(17)	85(37)	215(111)	214(168)	227(177)	290(278)	315(310)	325(325)	323(323)	325(325)	327(327)
11	47(27)	88(43)	203(128)	220(171)	228(194)	307(291)	325(320)	327(327)	327(327)	327(327)	327(327)
12	58(34)	82(53)	227(174)	234(210)	258(231)	307(303)	323(321)	325(325)	325(325)	327(325)	327(327)
13	58(28)	107(71)	198(164)	230(219)	261(230)	305(303)	323(323)	327(327)	327(327)	327(327)	327(327)
14	73(29)	118(79)	240(194)	272(240)	270(244)	320(314)	325(325)	327(327)	327(327)	327(325)	327(327)
15	67(35)	112(74)	221(200)	269(255)	275(256)	319(319)	327(327)	327(327)	327(327)	327(327)	327(327)
16	57(39)	135(84)	246(219)	288(275)	281(264)	321(321)	327(325)	327(327)	327(327)	327(327)	327(327)
17	84(39)	132(100)	258(246)	293(292)	286(282)	327(327)	327(327)	327(327)	327(327)	327(327)	327(327)
18	67(55)	154(119)	256(245)	296(294)	304(297)	327(327)	327(327)	327(327)	327(327)	327(327)	327(327)
19	70(60)	165(134)	265(268)	298(293)	300(293)	325(325)	327(327)	327(327)	327(327)	327(327)	327(327)
20	88(52)	185(155)	288(269)	305(305)	312(302)	325(325)	327(327)	327(327)	327(327)	327(327)	327(327)
21	97(74)	182(158)	287(275)	311(305)	315(310)	325(325)	327(327)	327(327)	327(327)	327(327)	327(327)
22	92(75)	169(168)	288(294)	316(303)	312(306)	327(327)	327(327)	327(327)	327(327)	327(327)	327(327)
23	83(78)	193(182)	292(295)	316(312)	317(316)	327(327)	327(327)	327(327)	327(327)	327(327)	327(327)
24	99(89)	200(177)	309(304)	312(309)	320(317)	327(327)	327(327)	327(327)	327(327)	327(327)	327(327)
25	99(92)	196(191)	310(306)	320(320)	320(318)	327(327)	327(327)	327(327)	327(327)	327(327)	327(327)
26	126(106)	208(212)	308(303)	321(321)	321(321)	327(327)	327(327)	327(327)	327(327)	327(327)	327(327)
27	134(113)	214(225)	313(311)	323(319)	323(321)	327(327)	327(327)	327(327)	327(327)	327(327)	327(327)
28	120(113)	232(231)	315(315)	325(323)	323(323)	327(327)	327(327)	327(327)	327(327)	327(327)	327(327)
29	130(127)	238(226)	307(319)	327(321)	323(323)	327(327)	327(327)	327(327)	327(327)	327(327)	327(327)
30	128(134)	234(240)	321(317)	325(325)	325(323)	327(327)	327(327)	327(327)	327(327)	327(327)	327(327)
31	146(144)	232(248)	319(313)	323(323)	325(325)	327(327)	327(327)	327(327)	327(327)	327(327)	327(327)
32	150(145)	253(250)	319(317)	325(325)	325(325)	327(327)	327(327)	327(327)	327(327)	327(327)	327(327)
33	156(155)	242(253)	319(319)	327(327)	325(321)	327(327)	327(327)	327(327)	327(327)	327(327)	327(327)
34	154(146)	249(252)	321(321)	323(323)	323(325)	327(327)	327(327)	327(327)	327(327)	327(327)	327(327)
35	176(160)	248(253)	319(317)	325(325)	325(323)	327(327)	327(327)	327(327)	327(327)	327(327)	327(327)
36	167(160)	256(251)	325(321)	325(323)	325(323)	327(327)	327(327)	327(327)	327(327)	327(327)	327(327)
37	178(157)	265(261)	319(323)	325(325)	325(323)	327(327)	327(327)	327(327)	327(327)	327(327)	327(327)
38	183(170)	263(273)	321(321)	325(319)	323(323)	327(327)	327(327)	327(327)	327(327)	327(327)	327(327)
39	184(176)	267(261)	325(323)	327(325)	323(325)	327(327)	327(327)	327(327)	327(327)	327(327)	327(327)