Applying Variable Precision Rough Set for Clustering Diabetics Dataset

Tutut Herawan, Wan Maseri Wan Mohd, A Noraziah

Faculty of Computer System and Software Engineering
Universiti Malaysia Pahang
Lebuh Raya Tun Razak, Gambang 26300, Kuantan, Pahang, Malaysia

{tutut,maseri,noraziah}@ump.edu.my

Abstract. Computational models of the artificial intelligence such as rough set theory have several applications. Rough set-based data clustering can be considered further as a technique for medical decision making. This paper presents the results of an experimental study of a rough-set based clustering technique using Variable Precision Rough Set (VPRS). Here, we employ our proposed clustering technique [12] through a medical dataset of patients suspected diabetic. Our results indicate that the VPRS-based technique is better than that the standard rough set-based techniques in the process of selecting a clustering attribute.
Keywords: Clustering; Rough set; Variable precision rough set model; Diabetic dataset.

1 Introduction

Clustering is a division of data into groups of similar objects. Each group, called a cluster, consists of objects that are similar to one another and dissimilar to objects of other groups. When representing data with fewer clusters necessarily loses certain fine details (akin to lossy data compression), but achieves simplification [1]. We refer the reader to [1,2] for a review of cluster analysis. Several cluster analysis techniques have been developed so far to group objects having similar characteristics. Clustering of categorical data is more challenging than that of numerical data. Most of the early cluster analysis techniques face problems due to the fact that much of the data contained in today's databases is categorical in nature. This necessitated the development of some algorithms for clustering categorical data [3]. For categorical data clustering, a new trend has become in algorithms which can handle uncertainty in the clustering process. One of the well-known techniques is based on rough set theory [4,5,6]. The main idea of the rough clustering is the clustering data set is mapped as the decision table. This can be done by introducing a decision attribute. Further, a divide-and-conquer method can be used to partition/cluster the objects. One of the successful pioneering rough clustering for categorical data techniques is Total Roughness [7] and Minimum-Minimum Roughness (MMR) [8]. These algorithms are based on the notions of lower and upper approximations of a set. Later a rough set-based attributes

dependency technique is proposed for handling the problem of computational complexity of TR and MMR, namely MDA [9,10]. However, in reviewing the above techniques, they are developed based on standard rough set model. To this, in facing noisy datasets as an integral part of databases, they fail to do. In order to overcome their drawback, an error parameter is introduced. Variable Precision Rough Set (VPRS) model proposed by Ziarko [11] is defined on the probabilistic space and will give us a new way to deal with the noisy data. In previous work, we have proposed an alternative technique for clustering noisy categorical data using Variable Precision Rough Set model [12,13]. It is shown that the proposed technique performs better than that the TR, MMR and MDA in handling noisy datasets.

Inspired by VPRS for handling noisy data, in this paper, we present the results of an experimental study of a rough-set based clustering technique using VPRS. Here, we employ our proposed clustering technique [11] through a medical survey dataset of patients suspected diabetic. Our results indicate that the VPRS-based technique is better than that the standard rough set-based techniques [7,8,9,10] in the process of selecting a clustering attribute.

The rest of this paper is organized as follows. Section 2 describes the notion of rough set theory which comprises an information system, set approximations and variable precision rough set model. Section 3 presents a short description on rough set-based techniques for selecting a clustering attribute, following by the proposed VPRS technique. Section 4 describes the experimental tests. Finally, the conclusion of this work is described in section 5.

2 Variable Precision Rough Set

2.1 Set Approximations

Motivation for rough set theory has come from the need to represent subsets of a universe in terms of equivalence classes of a partition of the universe. Here, we use the concept of rough set theory in term of data containing in an information system. An *information system* as in [5] is a 4-tuple (quadruple) $S = (U, A, V, f)$, where U is a non-empty finite set of objects, A is a non-empty finite set of attributes, $V = \bigcup_{a \in A} V_a$, V_a is the domain (value set) of attribute a, $f : U \times A \to V$ is a total function such that $f(u,a) \in V_a$, for every $(u,a) \in U \times A$, called information (knowledge) function.

Definition 1. *Let $S = (U, A, V, f)$ be an information system and let B be any subset of A. Two elements $x, y \in U$ is said to be B-indiscernible (indiscernible by the set of attribute $B \subseteq A$ in S) if and only if $f(x,a) = f(y,a)$, for every $a \in B$.*

Obviously, every subset of A induces unique indiscernibility relation. Notice that, an indiscernibility relation induced by the set of attribute B, denoted by $IND(B)$, is an equivalence relation. It is well known that, an equivalence relation induces unique

partition. The partition of U induced by $IND(B)$ in $S=(U,A,V,f)$ denoted by U/B and the equivalence class in the partition U/B containing $x\in U$, denoted by $[x]_B$.

The notions of lower and upper approximations of a set can be defined as follows:

Definition 2. *Let $S=(U,A,V,f)$ be an information system, let B be any subset of A and let X be any subset of U. The B-lower approximation of X, denoted by $\underline{B}(X)$ and B-upper approximations, denoted by $\overline{B}(X)$ of X, respectively, are defined by*

$$\underline{B}(X)=\{x\in U \mid [x]_B \subseteq X\} \text{ and } \overline{B}(X)=\{x\in U \mid [x]_B \cap X \neq \phi\}.$$

2.2 Variable Precision Rough Set

Variable precision rough set (VPRS) extends standard rough set theory by the relaxation of the subset operator [11]. It was proposed to analyze and identify data patterns which represent statistical trends rather than functional. The main idea of VPRS is to allow objects to be classified with an error smaller than a certain pre-defined level. This introduced threshold relaxes the rough set notion of requiring no information outside the dataset itself.

Definition 3. *Let a set U as a universe and $X,Y\subseteq U$, where $X,Y\neq\phi$. The error classification rate of X relative to Y is denoted by $e(X,Y)$, is defined by*

$$e(X,Y)=\begin{cases}1-\dfrac{|X\cap Y|}{|X|}, & |X|>0\\ 0, & |X|=0\end{cases}.$$

Definiton 4. *Let U be a finite set and let β be a real number within the range $0\le\beta<0.5$ and a set $X\subseteq U$. The B_β-lower approximation of X, denoted by $\underline{B}_\beta(X)$ and B_β-upper approximation of X, denoted by $\overline{B}_\beta(X)$, respectively, are defined by*

$$\underline{B}_\beta(X)=\{x\in U: e([x]_B,X)\le\beta\} \text{ and } \overline{B}_\beta(X)=\{x\in U: e([x]_B,X)<1-\beta\}.$$

The set $\underline{B}_\beta(X)$ is called the positive region of *X*. It's the set of object of *U* that can be classified into *X* with error classification rate not greater than β. Then we have $\underline{B}_\beta(X)\subseteq\overline{B}_\beta(X)$ if only if $0\le\beta<0.5$, which means that β be restricted in an interval $[0,0.5)$ in order to keep the meaning of the "upper" and "lower" approximations.

3 Rough set-based techniques for selecting a clustering attribute

In our previous work [12], we analyze two establish rough-set based techniques for selecting a clustering attribute. The first technique called Total Roughness (TR) is proposed by Mazlack *et al.* [7]. With the TR technique, the attribute selected as the clustering (partitioning) attribute is based on the maximum value of mean roughness. The second technique called Min-Min Roughness (MMR) is proposed by Parmar *et al.* [8]. The least of mean roughness of an attribute is the best attribute to be selected. Later, in [9,10], we showed that in almost all type of non-noisy database, the techniques of TR, MMR and MDA have the same result in selecting a clustering attribute. However, for noisy datasets they fail to handle, since no recommendation for selection a clustering attribute.

In this section, we present the proposed technique, which we refer to as Maximum Accuracy of Variable Precision Rough Set [12]. The technique uses the accuracy of approximation using variable precision of attributes in the rough set theory. The higher accuracy of approximation using variable precision of attributes is the more accurate (higher of accuracy of stranded approximation) for selecting clustering attribute. The justification that the accuracy of approximation using variable precision of attributes, the more accurate for selecting clustering attribute is stated in Proposition 1.

Definition 5. *The accuracy of approximation variable precision (accuracy of variable precision roughness) of any subset* $X \subseteq U$ *with respect to* $B \subseteq A$ *is denoted by* $\alpha_{B_\beta}(X)$. *It presented as*

$$\alpha_{B_\beta}(X) = \frac{\left|\underline{B}_\beta(X)\right|}{\left|\overline{B}_\beta(X)\right|}$$

where $|X|$ denotes cardinality of *X*. If $\beta = 0$, it is the traditional rough set model of Pawlak.

Proposition 1. *Let* $S = (U, A, V, f)$ *be an information system,* $\alpha_B(X)$ *be an accuracy of roughness,* $\alpha_{B_\beta}(X)$ *is an accuracy of variable precision roughness and given* β *the error factor of variable precision. I f* $(0 \le \beta < 0.5)$, *then* $\alpha_B(X) \le \alpha_{B_\beta}(X)$.

Proof. From Definition 4, if $\beta \ge 0.5$, then $\underline{B}_\beta(X) \not\subset \overline{B}_\beta(X)$. Thus, for $0 \le \beta < 0.5$, we have $\underline{B}_0(X) \supseteq \underline{B}_\beta(X)$ and $\overline{B}_0(X) \subseteq \overline{B}_\beta(X)$. Consequently, $\left|\underline{B}_0(X)\right| \le \left|\underline{B}_\beta(X)\right|$ and $\left|\overline{B}_0(X)\right| \ge \left|\overline{B}_\beta(X)\right|$. There are two cases for the value of β i.e. for $\beta = 0$, from Definition 4, we have $\alpha_B(X) = \alpha_{B_\beta}(X)$. While, for $0 < \beta < 0.5$, we have

$$\left|\underline{B}(X)\right| \le \left|\underline{B}_\beta(X)\right| \text{ and } \left|\overline{B}_\beta(X)\right| \le \left|\overline{B}(X)\right|.$$

Hence

$$\frac{\left|\underline{B}(X)\right|}{\left|\overline{B}(X)\right|} \le \frac{\left|\underline{B}_{\beta}(X)\right|}{\left|\overline{B}_{\beta}(X)\right|}.$$

Therefore, $\alpha_{B}(X) \le \alpha_{B_{\beta}}(X)$.

Definition 6. *Suppose* $a_i \in A$, $V(a_i)$ *has k-different values, say* γ_k, $k = 1,2,\cdots,n$. *Let* $X(a_i = \gamma_k)$, $k = 1,2,\cdots,n$ *be a subset of the objects having k-different values of attribute* a_i. *The accuracy of the set* $X(a_i = \gamma_k)$, $k = 1,2,\cdots,n$ *for given* β *error factor, with respect to* a_j, *where* $i \neq j$, *denoted* $\alpha_{\beta a_j}(X \mid a_i = \gamma_k)$, *is defined by*

$$\alpha_{\beta a_j}(X \mid a_i = \gamma_k) = \frac{\left|\underline{B}_{\beta} X_{a_j}(a_i = \gamma_k)\right|}{\left|\overline{B}_{\beta} X_{a_j}(a_i = \gamma_k)\right|},\ k = 1,2,\cdots,n.$$

The mean accuracy of attribute $a_i \in A$ with respect to $a_j \in A$, where $i \neq j$, denoted by $\mathrm{MAC}_{a_j}(a_i)$, is evaluated as follows

$$\mathrm{MAC}_{a_j}(a_i) = \frac{\sum_{k=1}^{|V(a_i)|} \alpha_{\beta a_j}(X \mid a_i = \gamma_k)}{|V(a_i)|},$$

where $V(a_i)$ is the set of values of attribute $a_i \in A$.

Definition 7. *Given n attributes, maximal accuracy of attribute* $a_i \in A$ *with respect to* $a_j \in A$, *where* $i \neq j$, *refers to the maximum of* $\mathrm{MAC}_{a_j}(a_i)$, *denoted* $\mathrm{MA}(a_i)$, *is obtained by the following formula*

$$\mathrm{MA}(a_i) = \mathrm{mean}\left(\mathrm{MAC}_{a_j}(a_i)\right),\quad 1 \le i, j \le n.$$

4 Experiment results

4.1 Material

Diabetes mellitus is now a major global public health problem. The incidence and prevalence of diabetes are escalating especially developing and newly industrialized nations. The estimated number of 80 million sufferers in 1990 is expected to double by the year 2000. In Asia alone, it is estimated that the total number of diabetes could reach more than 138 million [14]. In Malaysia, hundreds of thousands of people are

afflicted with this chronic disease. In this work, the data of patients suspected diabetic was collected from a survey in Kuantan Pahang Malaysia. There are 252 suspects of diabetic and ten symptoms founded; Often Thirst (81 suspects), Excessive Hunger (81 suspects), Frequent Urination (81 suspects), Tiredness and Fatigue (81 suspects), Rapid and/or Sudden Weight Loss (81 suspects), Blurred Vision (45 suspects), Numbness and/or Tingling in the Hands and Feet (45 suspects), Slow healing of Minor-to-treat Yeast Infection in Women (45 suspects), Recurrent or Hard-o-treat Yeast Infection in Women (45 suspects), and Dry or Itchy Skin (43 suspects). All attributes has two values (yes/no), signifying the absence or presence of some feature. If the attribute value is yes, then the symptom is occur and otherwise is not. The summary dataset is summarized in Table 1.

4.2 Clustering problem

The problem of the VPRS-based data clustering trough the above dataset is how to clusters the suspects of diabetic in the same group with similar characteristics (symptoms). It can be done, firstly among all attributes we want to select a clustering attribute. To select the candidate, we employ the proposed technique described in sub-section 3. Lastly, the divide and conquer technique is used to group the suspects having the same symptoms.

4.3 Result

In this experiment we choose the value of $\beta = 0.4$. The algorithm of TR, MMR, MDA and VPRS are implemented in MATLAB version 7.6.0.324 (R2008a). They are executed sequentially on a processor Intel Core 2 Duo CPUs. The total main memory is 4G and the operating system is Windows XP Professional SP3. The results are described in Tables 2.

Table 2. The mean roughness value using TR technique

Technique	Selected Clustering Attribute
TR	None
MMR	None
MDA	None
VPRS	Dry or Itchy Skin (with mean roughness: 0.0276)

From Table 2, the selected attributes is Dry or Itchy Skin. Therefore we have two initial clusters of patients (P), i.e.:

Cluster 1:
P1,P2,P3,P4,P5,P7,P8,P9,P10,P11,P12,P13,P14,P15,P16,P17,P18,P19,P20,P21,P22,P27,P28,P2,,P31,P32,P34,P37,P38,P39,P41,P42,P44,P47,P48,P49,P51,P52,P54,P57,P58,P59,P61,P62,P64,P67,P68,P69,P71,P72,P74,P77,P78,P79,P81,P82,P84,P86,P87,P88,P89,P91,P92,P94,P97,P98,P99,P101,P102,P103,P104,P107,P108,P109,P111,P112,P114,P117,P118,P

119,P121,P122,P124,P128,P133,P138,P143,P148,P153,P154,P155,P156,P157,P158,P159,P160,P161,P162,P193,P194,P195,P196,P197,P198,P199,P200,P201,P202,P232,P233,P234,P235,P236,P237,P238,P239,P240,P241,P242,P243,P244,P245,P246,P247,P248,P249,P250,P251,P252
Cluster 2:
P101,P102,P103,P104,P107,P108,P109,P111,P112,P114,P117,P118,P119,P121,P122,P124,P12,,P133,P138,P143,P148,P153,P154,P155,P156,P157,P158,P159,P160,P161,P162,P193,P194,P195,P196,P197,P198,P199,P200,P201,P202,P232,P233,P234,P235,P236,P237,P238,P239,P240,P241,P242,P243,P244,P245,P246,P247,P248,P249,P250,P251,P252,P6,P23,P24,P25,P26,P30,P33,P35,P36,P40,P43,P45,P46,P50,P53,P55,P56,P60,,P63,P65,P66,P70,P73,P75,P76,P80,P83,P85,P90,P93,P95,P96,P100,P105,P106,P110,P113,P115,P116,P120,P123,P125,P126,P127,P129,P13,P131,P132,P134,P135,P136,P137,P139,P140,P141,P142,P144,P145,P146,P147,P149,P150,P1,1,P152,P163,P164,P165,P166,P167,P168,P169,P170,P171,P172,P173,P174,P175,P176,P177,P178,P179,P180,P181,P182,P183,P184,P185,P186,P187,P188,P189,P190,P191,P192,P203,P204,P205,P206,P207,P208,P209,P210,P211,P212,P213,P214,P215,P216,P217,P218,P219,P220,P22,,P222,P223,P224,P225,P226,P227,P228,P229,P230,P231

Fig 1. The clusters obtained

From Figure 1, the divided and conquer technique can be applied recursively to obtain further clusters. At subsequent iterations, the leaf node having more objects is selected for further splitting. The splitting processes terminate when it reaches a pre-defined number of clusters. This is subjective and is pre-decided based either on user requirement or domain knowledge.

4.4 Cluster purity and its visualization

The purity of clusters was used as a measure to test the quality of the clusters [5]. According to the measurement, a higher value of overall purity indicates a better clustering result, with perfect clustering yielding a value of 1 [5].
For attribute Dry or Itchy Skin (IS), we have the following clusters purity.

Table 6. The clusters purity

Cluster Number	Class 1	Class 2	Purity
1	83	46	0.643411
2	43	80	0.650407
Overall Purity			0.646909

5 Conclusion

This paper presented an alternative technique for categorical data clustering using Variable Precision Rough Set model. We have shown that the proposed technique able for handling noisy data, as baseline standard rough set-based techniques fail to do. For selecting the clustering attribute, it is based on the maximal of accuracy of approximation using variable precision of attributes in the rough set theory. The results

of an experimental study through a medical survey diabetic dataset are described and it is shown that our technique provides better performance in selecting the clustering attribute.

Acknowledgement. The authors thank to Tengku Nurdania Tengku Majani from FSKKP UMP for providing a survey data of patients suspected diabetic. Fundamental Research Grant Scheme (FRGS) from Ministry of Higher Education of Malaysia No. Vote RDU 110104.

References

1. P. Berkhin. (2001) Survey of clustering data mining techniques. [On-line].Available: http://www.accrue.com/products/rp_cluster_review.pdf http://citeseer.nj.nec.com/berkhin02survey.html
2. Xu, R., Wunsch, D.: Survey of Clustering Algorithms. IEEE Transaction on Neural Networks, 16(3), 645-677 (2005)
3. P. Kumar and B.K. Tripathy. MMeR: an algorithm for clustering heterogeneous data using rough set theory. Int. J. of Rapid Manufacturing, 1 (2), 189–207 (2009)
4. Pawlak, Z.: Rough sets. International Journal of Computer and Information Science 11, 341–356 (1982)
5. Pawlak, Z.: Rough sets: A theoretical aspect of reasoning about data. Kluwer Academic Publisher, (1991)
6. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Information Sciences 177 (1), 3–27, (2007)
7. Mazlack, L.J., He, A., Zhu, Y., Coppock, S.: A rough set approach in choosing partitioning attributes. In the proceeding of the ISCA 13th, International Conference, CAINE-2000, 1–6 (2000)
8. Parmar, D., Wu, T., Blackhurst, J.: MMR: An algorithm for clustering categorical data using rough set theory, Data and Knowledge Engineering 63, 879–893 (2007)
9. Herawan, T., Deris, M.M., Abawajy, J.H.: Rough set approach for selecting clustering attribute. Knowledge Based Systems, 23(3), 220–231 (2010)
10. Herawan, T., Deris, M.M.: A framework on rough set-based partitioning attribute selection. In D.S. Huang et al. (Eds.): ICIC 2009, Lecture Notes in Artificial Intelligence 5755, 91–100, 2009. © Springer-Verlag (2009)
11. Ziarko, W.: Variable precision rough set model. Journal of computer and system science 46, 39–59 (1991)
12. Yanto, I.T.R., Herawan, T., Deris, M.M.: Data clustering using Variable Precision Rough Set Model. Intelligent Data Analysis, 15 (4), 465–482 (2011).
13. Yanto, I.T.R., Vitasari, P., Herawan, T., and Deris, M.M.: Applying Variable Precision Rough Set Model for Clustering Student Suffering Study's Anxiety. Expert System with Applications, 39 (1), 452–459, (2012).
14. http://www.diabetes.org.my/article.php?aid=63 retrieved on August 9th 2012.