Available online at www.sciencedirect.com

ScienceDirect

Applied Mathematics Letters 21 (2008) 400–403

Applied Mathematics Letters

www.elsevier.com/locate/aml

# Integrating rough set theory and medical applications

Puntip Pattaraintakorn[a,*], Nick Cercone[b]

[a] Department of Mathematics and Computer Science, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Bangkok, 10520, Thailand

[b] Faculty of Science and Engineering, York University, Ontario, Canada M3J 1P3

Received 5 March 2007; accepted 30 May 2007

**Abstract**

Medical science is not an exact science in which processes can be easily analyzed and modeled. Rough set theory has proven well suited for accommodating such inexactness of the medical profession. As rough set theory matures and its theoretical perspective is extended, the theory has been also followed by development of innovative rough sets systems as a result of this maturation. Unique concerns in medical sciences as well as the need of integrated rough sets systems are discussed. We present a short survey of ongoing research and a case study on integrating rough set theory and medical application. Issues in the current state of rough sets in advancing medical technology and some of its challenges are also highlighted.

*Keywords:* Rough set theory; Medical applications; Survival analysis

## 1. Introduction

Pawlak [1] introduced mathematical rough set theory in the early 1980's. The theory was based on the discernibility of objects. Rough set theory provides systems designers with the ability to handle uncertainty. If a concept is 'not definable' in a given knowledge base, rough sets can 'approximate' with respect to that knowledge. From a medical point of view, the attribute-value boundaries are usually vague. In actual situations, physicians diagnose a patient and decide what is the best way to cure them. To apply rough sets to medical data and imitate this ability, many issues in rough set theory are raised [2]. For example, discretization is necessary, whether uncertainty is subjective or objective, and medical attribute values lead to difficult situations for rough set-based medical applications. These issues are also discussed by [3]. They pointed out that rough sets offer algorithms with polynomial time complexity and space complexity with respect to the number of attributes and examples. They also note that the advantages of the rough sets methodology consist of: (i) the basic tools are lower and upper approximations of the concept (which are well-defined sets) and (ii) rough sets methodology is computed directly from input data.

[*] Corresponding author. Tel.: +011 662 326 4341; fax: +011 662 326 4354.
*E-mail address:* kppuntip@kmitl.ac.th (P. Pattaraintakorn).

0893-9659/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aml.2007.05.010

## 2. Theoretical aspects of rough sets

We describe the fundamental theory of rough sets from [1,4]. Given a finite set $U \neq \emptyset$ (*universe*) of objects, any subset $X \subseteq U$ of the universe is called a *concept* in $U$ and any family of concepts in $U$ is referred to as *knowledge*. A family of classifications over $U$ is called a *knowledge base* over $U$. This formal foundation of rough set theory reveals that we consider the "universe" to be a finite set. Keeping this stability set in mind, all rough set theory in medical database or data warehousing applications is concerned with the meaningfulness of updating sets, for example, the insert, delete and join operations in database systems. Rough set methodology endeavors to discover the variety of data sources while requiring integration of other approaches to handle extensibility of data sets. Let $R \subseteq X \times X$ be an equivalence relation over $U$. Then $R$ is reflexive ($xRx$), symmetric (if $xRy$ then $yRx$) and transitive (if $xRy$ and $yRz$ then $xRz$). Define $U/R$ as the family of equivalence classes of $R$ and let $[x]_R$ denote a category in $R$ containing an element $x \in U$. Given a knowledge base $K = (U, \mathbf{R})$ if $\mathbf{P} \subseteq \mathbf{R}$ and $\mathbf{P} \neq \emptyset$, then there is an equivalence relation IND($\mathbf{P}$) called the *indiscernibility relation* over $\mathbf{P}$.

The current trend in rough set theory explores the complementary mathematical properties with other mathematics disciplines. In [5], the author studied the ordered set of rough set theory and proved that the relations are not necessarily reflexive, symmetric or transitive. Next, as defined, with $X \subseteq U$ and $R \in \text{IND}(K)$,

$$x = \underline{R}X \quad \text{if and only if } [x]_R \subseteq X \tag{1}$$

$$x = \overline{R}X \quad \text{if and only if } [x]_R \cap X \neq \emptyset \tag{2}$$

called the *R-lower approximation* and *R-upper approximation* of $X$ respectively. Also let $\text{POS}_R(X) = \underline{R}X$ denote the $R$-positive region of $X$, $NEG_R(X) = U - \overline{R}X$ denote the $R$-negative region of $X$ and $\text{BN}_R(X) = \overline{R}X - \underline{R}X$ denote the $R$-borderline region of $X$.

The degree of completeness can also be characterized by the *accuracy measure*, in which *card R* represents the cardinality of set $R$ as follows:

$$\alpha_R(X) = \frac{\text{card}\underline{R}}{\text{card}\overline{R}} \quad \text{where } X \neq \emptyset. \tag{3}$$

Accuracy measures try to express the degree of completeness of knowledge. Eq. (3) is able to capture how large the boundary region of the data sets is; however, we cannot easily capture the structure of the knowledge. A fundamental advantage of rough set theory is the ability to handle a category that cannot be sharply defined given a knowledge base. Characteristics of the potential data sets can be measured through the rough sets framework. We can measure inexactness and express topological characterization of imprecision with:

(1) If $\underline{R}X \neq \emptyset$ and $\overline{R}X \neq U$, then $X$ is *roughly R-definable*.
(2) If $\underline{R}X = \emptyset$ and $\overline{R}X \neq U$, then $X$ is *internally R-undefinable*.
(3) If $\underline{R}X \neq \emptyset$ and $\overline{R}X = U$, then $X$ is *externally R-undefinable*.
(4) If $\underline{R}X = \emptyset$ and $\overline{R}X = U$, then $X$ is *totally R-undefinable*.

With Eq. (3) and classifications above we can characterize rough sets by the size of the boundary region and structure. Rough sets are treated as a special case of relative sets and integrated with the notion of Belnap's logic [6].

## 3. Medical science

Traditional medical data analysis tends to employ analysts who are familiar with particular data and use statistical techniques to provide reports. This approach is no longer viable. We extend some uniqueness properties in medical data in [7]. The salient points for rough sets are:

*Sensitivity and specificity analysis*: Most diagnoses and treatments in medical science are imprecise and accompanied by rates of error. The authors reveal the meaninglessness of sensitivity, specificity and accuracy measures used to evaluate data mining applications. Use of rough sets is able to circumvent this limitation with its ability to handle imprecise and uncertain data.

*Poor physical formulae or equations for characterizing medical data*: Other physical sciences mainly observe and collect data that can be fit into formulae reasonably and solved for the characteristics or relationship of that data.

However, medical data is less amenable to such formulae unless we employ a number of initial assumptions and constrain settings. Rough set theory offers a schematic approach for analyzing data without initial assumptions.

*Prior setting*: The rough sets approach takes advantage of a correct proven philosophy to work with medical data without a strong a priori reasoning. Theoretically, statistical clinical trials use predetermined hypotheses, and a priori assumptions in initial hypotheses test design. These also usually require a perfectly random statistical distribution.

The subject of rough set theory has evolved, and continues to evolve. However, stand-alone rough set theory hardly applies for dynamic medical data nor does it possess the ability to process a variety of data such as images, graphs and physician's notes. Rough sets provide a semi-automatic approach to medical data analysis and can combine with other complementary techniques, e.g., soft-computing, data mining, statistics, intelligent systems, machine learning, pattern recognition. For a good reference on data mining, soft-computing and knowledge discovery software tools please refer to [8,9].

## 4. A case study

Currently [10], illustrated rough sets offer a useful mechanism for analyzing and distilling essential attributes and rules from survival data. HYRIS is able to handle censor variable and survival time attributes that are a speciality for survival analysis. The Kaplan–Meier method, hazard function, log-rank, Brewslow, Tarone–Ware tests and CDispro [11] incorporate the rough sets framework to generate core, dispensable attributes, probe, reducts, and probe reducts. The authors demonstrated the utility of HYRIS by investigating a particular problem using geriatric data, melanoma data, pneumonia data and primary biliary cirrhosis data. The experimental results were validated with ELEM2 [12], and illustrated efficient data analysis and informative symbolic rule induction. HYRIS presented an ad hoc tool for highlighting specific survival domain knowledge. The previous study can be found in [13].

## 5. Current issues

### 5.1. Ability of rough sets to handle images and color images

In [14], the authors introduced the rough sets direct representation of the region of interest (ROI). Rough sets provided reasonable structures for the overlap boundary given domain knowledge. The case study for images of the heart on cardiovascular magnetic resonance (MR) images also extends to handling multiple types of knowledge including: myocardial motion, location and signal intensity. A study concerned with distinguishing different picture types of the central nervous system is introduced in [15]. Research involving color images appears in [16]. *Histon* are used as primary measure. The basic idea of a histon is to build a histogram on top of the histograms of the primary color components red, green, and blue. The authors show that the base histogram correlates with the lower approximation, whereas the encrustation correlates with the upper approximation. The problem of a machine vision application where an object is imaged by a camera system is considered in [17]. The object space can be modeled as a finite subset of the Euclidean space when the object's image is captured via an imaging system. Rough sets can bound such sets and provide a mechanism for modeling the spatial uncertainty in the image of the object. This work introduced a rough sets approach for building pattern matching systems that can be applicable with a wide range of images in medical sciences.

### 5.2. Ability of rough sets to handle signals/graphs

A hybrid rough sets system for sparse coding is constructed in [18]. The classification task endeavors to classify the decomposed cortical evoked potentials (EPs). [19] presented electrocardiogram (ECG) analysis based on rough set theory. The authors discussed feasibility of the automatic ECG recognition over MIT-BIH data. Another interesting line of research that deals with graph theory is in [20].

### 5.3. Rough sets in bioinformatics

Since 2000, rough sets have been surprising useful for bioinformatics. Tools for knowledge discovery that support gene expression analysis, annotation and visualization have also been created [21]. [22] applied the ROSETTA framework for learning from time series of gene expression data. The experimental results illustrated reduce the number of genes to the number of discernible genes. An evolutionary rough *c*-mean clustering algorithm has been performed over microarray gene expression colon cancer data [23]. This work represented clusters in terms of upper and lower approximations in rough set theory; the Davies–Bouldin index is used as the fitness function to be minimized.

## 6. Conclusion

The other challenging open problems for rough sets are: hypothesis building, how to partition the data into appropriate distributed subsets, transformation, updating the number of examples, and attribute and system extensibility.

## Acknowledgements

The first author was supported by KMITL research fund, Thailand. The second author was supported by NSERC, Canada.

## References

[1] Z. Pawlak, Rough sets, in: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Dordrecht, 1991.

[2] W.W. Koczkodaj, M. Orlowski, V.W. Marek, Myths about rough set theory, Commun. ACM 41 (11) (1998) 102–103.

[3] J.W. Grzymala-Busse, W. Ziarko, Data mining and rough set theory, Commun. ACM 43 (4) (2000) 108–109.

[4] J. Komorowski, L. Polkowski, A. Skowron, Rough sets: A tutorial, in: S.K. Pal, A. Skowron (Eds.), Rough Fuzzy Hybridization — A New Trend in Decision Making, Springer, 1999, pp. 3–98.

[5] J. Jarvinen, The ordered set of rough sets, in: S. Tsumoto, et al. (Eds.), RSCTC, in: Proceedings LNAI, vol. 3066, Springer, 2004, pp. 49–58.

[6] A. Mousavi, P. Jabedar-Maralani, Relative sets and rough sets, Int. J. Appl. Math. Comput. Sci. 11 (3) (2001) 637–654.

[7] K.J. Cios, G.W. Moore, Uniqueness of medical data mining, Artif. Intell. Med. 26 (1–2) (2002) 1–24.

[8] M. Goebel, L. Gruenwald, A survey of data mining and knowledge discovery software tools, ACM SIGKDD Explorations 1 (1) (1999) 20–23.

[9] S. Mitra, S.K. Pal, P. Mitra, Data mining in soft computing framework: A survey, IEEE Trans. Neural Networks 13 (1) (2002) 3–14.

[10] P. Pattaraintakorn, N. Cercone, K. Naruedomkul, Selecting attributes for soft-computing analysis in hybrid intelligent systems, in: RSFDGrC. Proceedings LNCS, vol. 3642, 2005, pp. 698–708.

[11] P. Pattaraintakorn, N. Cercone, K. Naruedomkul, Hybrid intelligent systems: Selecting attributes for soft-computing analysis, in: Proc. of the 29th COMPSAC, 2005, pp. 319–325.

[12] A. An, N. Cercone, ELEM2: A learning system for more accurate classifications, in: 12th CSCSI, in: Proceedings LNCS, vol. 1418, 1998, pp. 426–441.

[13] J. Bazan, A. Skowron, D. Slezak, J. Wroblewski, Searching for the Complex Decision Reducts: The Case Study of the Survival Analysis, in: LNAI, vol. 2871, Springer-Verlag, Berlin, Heidelberg, 2003, pp. 160–168.

[14] S. Hirano, S. Tsumoto, Rough representation of a region of interest in medical images, Int. J. Approx. Reason. 40 (1–2) (2005) 23–34.

[15] J. Jelonek, K. Krawiec, R. Slowinski, J. Stefanowski, R. Slowinski, Rough set reduction of features for picture-based reasoning, in: T.Y. Lin (Ed.), Soft Computing, Simulation Councils, San Diego, 1995, pp. 89–92.

[16] A. Mohabey, A.K. Ray, Rough set theory based segmentation of color images, in: Proc. of the 19th NAFIPS, 2000, pp. 338–342.

[17] D. Sinha, P. Laplante, A rough set-based approach to handling spatial uncertainty in binary images, Eng. Appl. Artif. Intell. 17 (2004) 97–110.

[18] G.M. Boratyn, T.G. Smolinski, M. Milanova, A. Wrobel, Sparse coding and rough set theory-based hybrid approach to the classificatory decomposition of cortical evoked potentials, in: L. Wang, J.C. Rajapakse, K. Fukushima, S. Lee, X. Yao (Eds.), Proc. of the 9th ICONIPR, vol. 5, 2002.

[19] X. Huang, Y. Zhang, A new application of rough set to ecg recognition, in: Proc. of the Int. Conf. on Machine Learning and Cybernetics, 2003, pp. 1729–1734.

[20] P. Mitra, S.K. Pal, A. Siddiqi, Non-convex clustering using expectation maximization algorithm with rough set initialization, Pattern Recogn. Lett. 24 (2003) 863–873.

[21] T. Jenssen, A. Laegreid, J. Komorowski, E. Hovig, PubGene: Discovering and visualising gene–gene relations, in: S. Miyano, R. Shamir, T. Takagi (Eds.), RECOMB, Universal Academy Press, Inc., Japan, 2000, pp. 48–49.

[22] T.R. Hvidsten, A. Laegreid, J. Komorowski, Learning rule-based models of biological processes from gene expression time profiles using gene ontology, in: Microarrays, Bioinform. J. 19 (9) (2003) 1116–1123 (special issue).

[23] S. Mitra, An evolutionary rough partitive clustering, Pattern Recogn. Lett. 25 (2004) 1439–1449.