รายงานการวิจัย

การศึกษาเชิงการคำนวณซอฟต์เพื่อพัฒนา การวิจัยด้านระบบไฮบริดอัจฉริยะด้านการแพทย์โดยวิธีการรัฟเซตและสถิติ

Soft Computing Study for the Development of Research in Hybrid Medical Intelligent Systems by Rough Sets and Statistical Techniques

ชื่อผู้วิจัย นางสาวพรรณทิพย์ ภัทรอินทากร

ได้รับทุนสนับสนุนงานวิจัยจากกองทุนวิจัยสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ประจำปีงบประมาณ 2550

กิตติกรรมประกาศ

รายงานการวิจัยเรื่อง "การศึกษาเชิงการคำนวณซอฟต์เพื่อพัฒนา การวิจัยด้านระบบไฮบริดอัจฉริยะด้านการแพทย์โดยวิธีการรัฟเซตและสถิติ" (Soft Computing Study for the Development of Research in Hybrid Medical Intelligent Systems by Rough Sets and Statistical Techniques) เป็นโครงการที่จัดทำขึ้นโดยได้รับการสนับสนุนจาก กองทุนวิจัยสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง เพื่อนำข้อมูลไปใช้ประกอบการจัดทำแนวทางการศึกษาวิจัย เพื่อสร้างขีดความสามารถในเรื่องการพัฒนาด้านวิทยาศาสตร์เทคโนโลยีและอุตสาหกรรม ประเภทการวิจัยประยุกต์ สาขาวิจัยคณิตศาสตร์ประยุกต์

ผู้วิจัยขอขอบคุณ รศ.ดร.กิตติ ตีรเศรษฐ ประธานกรรมการ และขอขอบคุณกองทุนวิจัยสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ที่อนุมัติงบประมาณอุดหนุนวิจัย เป็นค่าใช้จ่ายของโครงการวิจัยนี้และขอขอบคุณฝ่ายเลขานุการและคณะทำงานที่ให้ความร่วมมือแก่นักวิจัยอย่างดียิ่ง

ดร.พรรณทิพย์ ภัทรอินทากร

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

2552

ชื่อโครงการ (ภาษาไทย) การศึกษาเชิงการคำนวณซอฟต์เพื่อพัฒนา การวิจัยด้านระบบไฮบริดอัจฉริยะด้านการแพทย์โดยวิธีการรัฟเซตและสถิติ

(ภาษาอังกฤษ) Soft Computing Study for the Development of Research in Hybrid Medical Intelligent Systems by Rough Sets and Statistical Techniques

ได้รับทุนอุดหนุนการวิจัยจาก กองทุนวิจัย สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ประจำปี 2550 **จำนวนเงิน** 240,000 บาท

ระยะเวลาทำการวิจัย 2 ปี **ตั้งแต่** 25 มิถุนายน 2550 ถึง 26 มิถุนายน 2552

หน่วยงานและผู้ดำเนินการวิจัยพร้อมหน่วยงานที่สังกัดและเลขหมายโทรศัพท์

ชื่อ-สกุล (ภาษาไทย) ดร.พรรณทิพย์ ภัทรอินทากร

ชื่อ-สกุล (ภาษาอังกฤษ) Dr. Puntip Pattaraintakorn

ตำแหน่งทางวิชาการ	อาจารย์	สัดส่วนการวิจัย	100%
สาขาวิชา	คณิตศาสตร์	คณะ	วิทยาศาสตร์
โทรศัพท์	02-326-4341-319, #319	โทรสาร	02-326-4354

บทคัดย่อ

ภาษาไทย

งานวิจัยระดับแนวหน้าทั้งในปัจจุบันและอนาคต จำเป็นต้องมีลักษณะสหวิทยาการ (multidisciplinary) และการบูรณาการของความรู้ความเชี่ยวชาญจากหลากหลายสาขาร่วมกัน เพื่อพัฒนาองค์ความรู้และเทคโนโลยีใหม่ ในอดีตงานวิจัยทางด้านฟิสิกส์และวิศวกรรมได้นำความรู้และทฤษฎีทางคณิตศาสตร์และสถิติไปประยุกต์ใช้ในการศึกษา และนำมาซึ่งการค้นพบใหม่ ๆ อย่างต่อเนื่องเป็นลำดับ ในขณะที่งานวิจัยทางด้านการแพทย์กลับมิได้มีการประสานความรู้ทางด้านคณิตศาสตร์และสถิติเพื่อวิเคราะห์และค้นพบองค์ความรู้ใหม่ ๆ เท่าใดนัก จนกระทั่งไม่กี่สิบปีที่ผ่านมา มีหัวข้อวิจัยทางด้านคณิตศาสตร์แขนงใหม่นั่นคือทฤษฎีรัฟเซต (rough set theory) รวมถึงวิธีการทางสถิติที่ประสานเข้ากับเทคโนโลยีด้านคอมพิวเตอร์แบบการคำนวณซอฟต์ (soft computing) เพื่อเพิ่มศักยภาพในการวิเคราะห์ข้อมูลทางการแพทย์ที่เพิ่มขนาดและความซับซ้อนขึ้นจนยากแก่การวิเคราะห์ด้วยศาสตร์แขนงใดแขนงหนึ่ง (stand-alone disciplinary) ได้รับความสนใจจากนักวิจัยทั่วโลกเพิ่มมากขึ้น โดยเฉพาะอย่างยิ่งเพื่อผลิตระบบไฮบริดอัจฉริยะด้านการแพทย์ (hybrid medical intelligent systems) อันเป็นเป็นทางเลือกใหม่ของวงการแพทย์ และมีบทบาทในการวิเคราะห์ค้นพบความสัมพันธ์ของปัจจัยความเสี่ยงของโรคและวินิจฉัย โรคควบคู่กันไป อันจะเป็นจุดเริ่มต้นไปสู่การวิจัยเพื่อค้นพบองค์ความรู้ใหม่ที่มีคุณภาพสูงมากขึ้น และเทียบเคียงกับนักวิจัยชั้นนำของโลก ซึ่งนั่นคือพันธกิจของโครงการวิจัยนี้ โดยใช้กลยุทธเริ่มต้นสร้างกลุ่มวิจัยที่เข้มแข็งในสหสาขาเพื่อดำเนินการวิจัยร่วมกัน ซึ่งจะได้การปฏิสัมพันธ์กับนักวิจัยที่มีความเชี่ยวชาญจากต่างสาขามาประสานกันให้เกิดผลงานวิจัยที่มีผลสัมฤทธิ์สูงสุด

ABSTRACT

The medical profession is challenged to find innovative approaches for discovering relevant disease information and using that information effectively. However, the amount and complexity of human medical data are the most difficult of all biological data to analyze and derive models. Previous researches have shown that multidisciplinary research i.e., mathematics: rough set theory, statistics: Cox proportional hazard and Kaplan-Meier estimate, computer science: soft computing and database management can cope with these difficulties. Thus, current research in medical science and multidisciplinary approaches tends to integrate several approaches together to construct hybrid medical intelligent systems. Hybrid medical intelligent system is a novel active research which has several open problems to conduct the study. For this reason, we propose soft computing study for the development of research in hybrid medical intelligent systems by rough sets and statistical techniques. We aim at develop innovative approach for eliciting and analyzing medical data and discover high level knowledge from the data e.g., decision rules, statistical models.

สารบัญเรื่อง

สารบัญเรื่อง

สารบัญตาราง

สารบัญรูป

Introduction

The medical profession is challenged to find innovative approaches for discovering relevant disease information and using that information effectively. Successful research has been conducted in this area [1], although, for the most part, this research has been performed on small, well understood data sets. Nonetheless, billions of people have at least some of their medical information collected in medical databases. The sheer volume such data highlights the need for further comprehensive and systematic analysis to improve overall and general health outcomes. To fulfill that need, it is essential to carefully analyze the data.

However, the amount and complexity of human medical data are the most difficult of all biological data to mine and analyze. Previous research have shown that the following approaches in mathematics, statistics and computer science can cope with these difficulties.

- Rough sets
- Statistics
- Statistical leanrning (e.g., survival analysis)
- Flow graphs
- Recommender system
- Distributed databases
- and other approaches in data mining.

Characteristics of medical data pose challenges to researchers. The uniqueness of medical data has been discussed in [1]; salient points include:

- The growing number of medical databases does not appear significant because they are not available in some edifying format. Generally, medical collections, diagnoses and treatments are subject to error rates, imprecision and uncertainty. We believe the application of rough sets combined with statistical learning to analyze this data can help overcome these problems.
- The structure of nearly all medical data resists analysis by any formulae, equations or theoretical model. Physical scientists measure data which can be analyzed with formulae or equations. Solutions to problems they solve can describe phenomenon and

relationships among those data. Usually, models are simplified by imposing constraints and using initial assumptions. In contrast, real world medical problems employ less structured data. Automatic clustering techniques can operate to great advantage in this case.

- Classical statistic experiments are designed theoretically. Once the hypotheses are set, all initial assumptions are predefined. Such procedures adapt poorly to change, even when the patient's conditions appear to change. We speculate that above techniques can alleviate this problem and increase reliability.

For these reasons, traditional manual data analysis is not adequate, and methods for efficient computer-based analysis, e.g., data mining, are indispensable. Successful data mining in medicine requires know-how and knowledge of medical data.

Survival analysis [2] is a branch of statistics that studies time-to-event data. Death or failure is called an *event* in the survival analysis literature. Survival analysis attempts to answer questions such as:

Is diabetes a significant risk factor for geriatric patients?

What is the fraction of patients who will survive past a certain time?

Survival analysis is called *reliability analysis* in engineering, and *duration analysis* in economics. Presently, survival data in existence worldwide highlights the need for further comprehensive and systematic analysis to improve overall health outcomes. Much data analysis research has been conducted in several areas [1, 3–5]. The aim of such data analysis techniques is to use the collected data for training in a learning process, and then to extract a hidden pattern by model construction. However, a successful technique involves far more than selecting a learning algorithm and running it over data sets. Successful data analysis requires know-how and in-depth knowledge of data. The challenges in real world problems are the complexity and unique properties of the survival data at hand. In many practical situations, survival data sets are vague and come with redundant and irrelevant attributes. The inclusion of these attributes in the data causes some difficulties in discovering the knowledge. To avoid these troubles, it is essential to precede the learning task with an at-

tribute selection process to delete redundancy records, uncertainty attributes and overwhelming data. To serve our first purpose, we create an attribute subset large enough to include all of the important attributes, but small enough for our learning system to handle easily.

Another issue in survival data analysis is the desire for automatic analysis processes [2]. Classical approaches are designed theoretically, automation is then increasingly challenging. Traditional data analysis is not adequate (e.g., Dempster-Shafer theory, grade of membership [33]), and methods for efficient mathematical and computer-based analysis, e.g., rough sets, are indispensable.

Rough set theory was developed by Zdzislaw Pawlak [6, 33, 8–10]. It provides system designers with the ability to compute with imperfect data. If a concept cannot be defined in a given knowledge base (*vagueness*), rough sets can approximate that knowledge efficiently. While logic is *deductive* and hardly applies to real situations, rough sets is in the form of *inductive reasoning* that widens the scope of the research to deal with real world data [8]. Rough sets do not require a specific model that can fit the data to be used in the analysis process. This ability provides flexibility in real situations. Rough sets provide a semi-automatic approach to data analysis and can be combined with other complementary techniques. Thus, current research tends to hybridize diverse methods of soft computing [1]. In this research, we offer a rough sets based approach with the capability to reason and to distil useful knowledge for survival data (e.g., risk factor, survival prediction model). We provide a more detail of principle, the system architecture and experiments using these concepts over a range of different data sets in Part I.

In Part II, four goals will be accomplished. Firstly, in some situations, smaller units of full data set are of great importance to analyze (e.g., men vs. women patients, young vs. old customers). Thus, the finer knowledge consideration is important in some aspects [33]. This research is intended to present mathematical proof and a case study of Pawlaks statement about distributed knowledge.

Second is the consideration of a case study on self-reported geriatric data (from Dalhousie Medical School, Canada). As stated in [23], the combination of attributes for predicting prolongation time might be different in men and women. Furthermore, when consider-

ing data which is self-reported, it usually contains optimistic bias. This research will illustrate that rough sets in combination with statistics, relational databases and systematic hypothesis tests is tolerant and robust in learning to extract significant risk factors and induce the rules for predicting survival time. The self-reported data which is usually optimistic can be reasoned by rough sets approach and overcome such bias successfully.

Third, manually evaluating important and interesting rules generated from data is generally infeasible due to the large number of rules extracted. Different approaches such as rule interestingness measures and rule quality measures have been proposed and explored previously to extract interesting and high quality association rules and classification rules. Rough sets theory was originally presented as an approach to approximate concepts under uncertainty. Alternatively, in this research, we explore rough sets based rule evaluation approaches in knowledge discovery. Rough set based rule evaluation approaches can be used in a straightforward way to rank the importance of the rules.

Finally, We complete our research by adding a recommender system. We propose a health recommendation system architecture using rough sets, survival analysis approaches and rule-based expert systems. Our main goal is to recommend clinical examinations for patients or physicians from patients self reported data. Such data will be treated as condition attributes, while survival time from a follow-up study will be treated as the target function. We have amalgamated rough set theory, relational databases, statistics, soft computing and several pertinent techniques to generate a hybrid intelligent system for survival analysis.

We introduce in Parts I - II our theoretical and/or basic concepts employed to analyze medical data. Next, details of our proposed methodological settings, architecture and/or theoretical proofs are described. In successive sections, the experiments, results, evaluation and/or comparison studies are reported. We provide summary at the end of every part.

Part I

1 Preliminaries and notations

1.1 The role of soft computing

One data analysis technique can generate very accurate results for one data set and poor results for another data set. Moreover, each technique has underlying advantages and disadvantages. The amount of real world data requires such techniques to have tractable time complexity, and simultaneously provide satisfactory outcome. Research in *soft computing* has demonstrated successes. Soft computing works synergistically with other data analysis methods to provide flexible analytical tools in real situations. Medsker [1] stated that soft computing differs from traditional computing in that it is tolerant of imprecision, uncertainty and partial truth. This guiding principle of soft computing can be used to achieve tractability, robustness and low cost solutions.

Rough sets is a leading soft computing approach. Works on hybrid rough sets based approaches have been conducted in [8–10, 63, 12–14] and in our previous studies to relational algebra [24, 16], to flow graphs [17], to Cox proportional hazard model [18] and to medical applications [19, 20]. However, the new generation of such research needs to understand the problem's nature to increase the intelligence of the system. This new generation of research can reach this objective by combining several related research areas.

We introduce the new perspective of hybrid rough sets based approach (Fig. 1). The components we integrate into our hybrid intelligent system are rough sets, relational algebra and other scientific areas. Afterwards, the reinforcement step increases the intelligence of the system to a high level hybrid intelligent system, such as optimization approaches.

1.2 Rough Sets

In the early 1980's, Pawlak [33] introduced rough sets theory. Rough set theory is the last and most important technique to turn our proposed system into a hybrid system. The purely statistical measurement gives reasonable evidence to support the hypothesis. When considering noisy real-world data, however, purely statistical measures can be less meaningful. Furthermore, the Cox proportional hazard

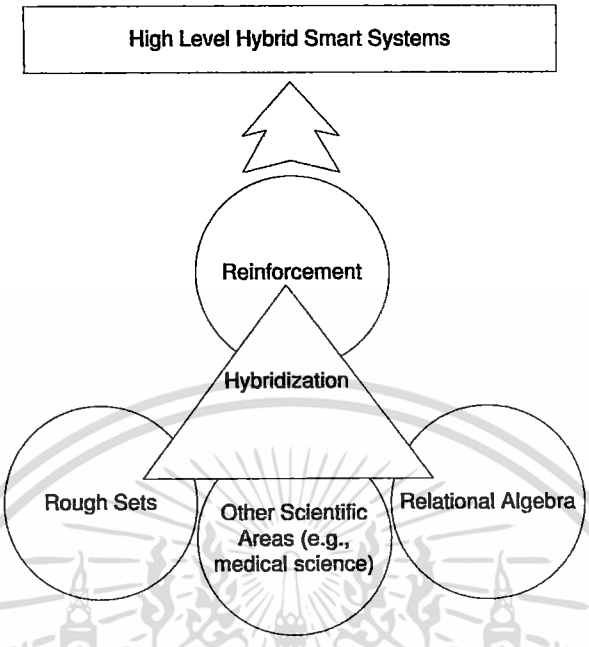


Fig. 1. A perspective of how to build a high level rough sets hybrid smart system.

model (for multiple attributes analysis) requires a relationship between the dependent and explanatory variables and uses fine-tuned tests (cf. [40]). Rough sets can perform this task efficiently [24, 25]. For these reasons, we propose the hybrid rough sets approach with the integration of rough sets and statistics.

A number of rough sets applications to medical survival data have been proposed. Primarily, these studies simply apply rough sets methods to data that happens to have a medical origin, without much regard for the underlying medical problem at hand. These studies contrast our approach that contains domain knowledge consideration with probe attributes and probe reducts [53]. Furthermore, our previous study [54] analyzed a number of survival data sets and captured the necessary semantic information embedded in the data.

The primary purpose of our study is to explore individual attributes by statistics (univariate) while simultaneously exploring the effects of several attributes (multivariate) on survival by using a hybrid rough sets approach. The rough sets principle can perform

attribute selection of decision concepts that remains the same over all information. Finding a heuristic method for attribute selection that is feasible for large data sets is an open problem. Skowron et al. [48] showed that the lower and upper approximations, positive regions, short reducts, etc. can be computed in a straightforward manner from the discernibility matrix with $O(kn^2)$ time complexity where n is the number of examples and k is the number of attributes of the data set, which is not feasible for large data sets.

Applications of rough sets are widening and emerging and are continuously marked with advancements, for example, survival analysis. Some studies have been conducted using rough sets [31, 32]. These studies utilized inconsistent data that occurred in 246 records out of 557 records in throat cancer patients.

Nguyen et al. [35, 51] proposed several algorithms that do not require storing the discernibility matrix in the calculation step. Their algorithm for generating *short reducts* by using Johnson strategy has $O(k^2nlogn)$ time complexity. This algorithm is an efficient way to compute reducts without using a discernibility matrix.

Wroblewski [37] proposed a hybrid algorithm for generating reducts. His proposed approach is more efficient compared to a classical GA. Bazan et al. [38] reported a method to search for reducts that generates a minimal number of rules. The authors also introduced several measures for reduct quality. Fewer rules were generated from these reducts, occupied less memory and classified new examples faster.

In several studies, the effects of a certain attribute are the main goal of analysis. This attribute is not necessarily included in the reduct sets of the subtables. For example, the risk factor that will impact the progression of disease is the important candidate component in the reducts. Such risk factors should be further analyzed for some problems in the medical domain.

In [35, 51, 39], rough sets were redefined using database operations. The computing times were improved remarkably by using database operations and the database system directly. The straightforward approach using databases in the implementation is a promising approach for rough sets. Hence, benefits of database set operations `Count` (*Card*) and `Projection` ($\prod$) permit the computation to scale up in our study. The terms probe attribute and probe reducts were introduced in [24] as the following.

Let us assume that a *decision table* is denoted by $T(U, C, D)$, where C is the set of *condition attributes* and D is a singleton set of *target function*. For simplicity, we write C and D instead of $\{C\}$ and $\{D\}$.

Definition 1. *An attribute C_i is a core attribute if*

$$Card(\prod(C - C_i + D)) \neq Card(\prod(C - C_i)).$$

Definition 2. *An attribute $C_i \in C$ is a dispensable attribute with respect to D if*

$$Card(\prod(C - C_i + D)) = Card(\prod(C - C_i)).$$

Definition 3. *The degree of dependency, $K(R, D)$, between the attribute subset $R \subseteq C$ and attribute D in decision table $T(U, C, D)$ is*

$$K(R, D) = \frac{Card(\prod(R + D))}{Card(\prod(C + D))}.$$

Definition 4. *The subset of attributes $RED \subseteq C$ is a reduct of C with respect to D if*

$$K(RED, D) = K(C, D) \quad and$$

$$K(RED, D) \neq K(RED', D) \text{ for all } RED' \subset RED.$$

Definition 5. *A probe attribute $P \in C$ corresponding to $T(U, C, D)$ is defined as an attribute of concern in $T(U, C, D)$ for each domain by an expert.*

Definition 6. *A probe reduct corresponding to decision table $T(U, C, D)$ is defined as a reduct consisting of a selected before attribute of concern.*

Example 1. The notion of probe attribute and probe reducts can be described with the following example; in survival analysis, *survival time* is the decision attribute while *patient's symptoms*, *surgery type* and so on describe the condition attributes.

If we want to know about the survival time for each patient, the risk of radical surgery or mild surgery becomes the significant part of this determination. Hence, we consider the *surgery type* attribute as a probe attribute. The probe reducts are the reducts constructed from the probe attribute.

1.3 Survival Analysis

Survival analysis describes time-to-event analysis. Survival analysis is called *reliability analysis* in engineering, and *duration analysis* in economics. Survival analysis is the study of the time between entry to a study and a subsequent event (e.g., death, recurrence of cancer).

We accomplish this analysis by computing the probability that the event will occur within a specific time and include a comparison of several risk factors. Frequently however, the prediction of whether the event will eventually occur or not is of primary importance. It often happens that the study does not span enough time in order to observe the event for all patients.

Two extra factors we consider for survival analysis include:

- survival time (which is commonly misleading), the time patients are admitted to the study until the time to death as well as the time to particular events (e.g., recurrence of disease or time until metastatis to another organ),
- if any patient leaves the study for any reason, use of a censor variable is required, indicating the period of observation was cut off before the event of interest occurred. To properly address censoring, modeling techniques must take into account that for these patients the event does not occur during the follow-up period.

The following are some example questions of our study that will be answered:

"Is diabetes a significant risk factor for geriatric patients?"

"What are the rules for survival time predictions of geriatric patients?"

"What is the survival tendency of a geriatric patient?"

1.4 Kaplan-Meier Survival Analysis

In survival analysis it is highly recommended to use the Kaplan-Meier survival analysis method [41]. Kaplan-Meier survival analysis offers Kaplan-Meier survival curves, which provide insight into the survival function for each group. The proportion of the population of such patients who would survive a given length of time under the same circumstances is given by the Kaplan-Meier method or the product limit (PL) as shown in equation (1).

S is based on the probability that each patient survives at the end of a time interval, on the condition that the patient was present at the start of the time interval.

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{1}$$

where t_i is the period of study at point i, d_i is number of events up to point i and n_i is number of patients at risk prior to t_i.

The method produces a table and a graph, referred to as the life time table and survival curve. There are initial assumptions to make use of Kaplan-Meier method that appear in [41]. While the Kaplan-Meier method focused on a single risk factor or attribute, the Cox proportional hazard model is used for multiple attributes. This model assumes a relationship between the dependent and explanatory variables and uses fine-tuned tests (cf. [40]). In this research, we propose multiple attributes analysis with system hybridization using rough set theory that does not require any initial assumption (Sect. 1.2).

The authors set the threshold different for attribute values among such inconsistent records. They used the Kaplan-Meier survival function [41] to cluster groups of patients into approximately similar

Kaplan-Meier characteristics. The notion of using Kaplan-Meier with rough set theory is used successfully. However, the Kaplan-Meier curve is used under the condition that the resulting clusters are representative of the data by visualization. Indeed, the visual determination of such curves requires special treatment. Statistical hypothesis testing techniques are required to prove a significant difference between the curves. Our work provides an analysis schema to accomplish this as listed in the next section.

1.5 Log-rank Test

An important part of survival analysis is to analyze the risk factor on a plot of the survival curves (Sect. 1.4) for each group of interest. The comparison of the survival curves for two groups cannot based on visual impressions. Statistics yield useful survival analysis data and theoretical tests to provide solutions. Thus, we consider the log-rank [28], Brewslow [29] and Tarone-Ware tests [30] which explore whether or not to include the attribute (or risk factor) in the prediction survival model constructions. These tests calculate their *p-values* that test the

- *null hypothesis* (H_0–the survival curves has no significant difference in survival times in two groups of interest) against the
- *alternative hypothesis* (H_1–the survival curves has significant difference in survival times in two groups of interest).

For example, we can consider *diabetes* risk factor in geriatric data set. These three hypothesis testing approaches can answer the question: "Is diabetes a significant risk factor for geriatric patients?", under the example hypotheses:

H_0: No significant difference in survival times between diabetes groups.
H_1: Significant difference in survival times between diabetes groups.

The three statistical tests differ in how they weight the examples. The log-rank test weights all examples equally, the Breslow test weights earlier periods more heavily and the Tarone-Ware test weights earlier examples less heavily than the Breslow test.

The early studies applying rough set theory to survival analysis are [31,32]. They used rough sets to discover relevant patterns for complex decisions successfully. A case study considered is the postsurgery survival analysis for the head and neck cancer cases. Nonetheless, hypothesis testing for each risk factor was not included the analysis and is still an open problem.

1.6 Cox Model

Cox (1972) proposed a semi-parametric model for the hazard function that allows the addition of explanatory variables but keeps the baseline hazard as an arbitrary, unspecified, nonnegative function of time. While the Kaplan-Meier method focused on a single risk factor or attribute, the Cox proportional hazard model is used for multiple attributes. This model assumes a relationship between the dependent and explanatory variables and uses fine-tuned tests (cf. [40]). We analyze multiple attributes with system hybridization using rough set theory.

Using the method of maximum partial likelihood, we estimate the parameters in Cox's model. Partial likelihood is remarkable in that you can estimate the coefficients without having to specify the baseline hazard function h_0. The Cox hazard function for fixed-time covariates, X, is

$$h(t) = h_0(t)e^{b_1X_1+b_2X_2+...+b_kXk} \tag{2}$$

where $h(t)$ refers to the hazard function at time t, $h_0(t)$ refers to the baseline hazard or hazard for an individual when the value of all the independent variables equal zero.

The $X_1, X_2, ..., X_k$ refer to explanatory variables, $b_1, b_2, ..., b_k$ refer to Cox regression coefficients determined by partial likelihood estimation while k refers to the number of explanatory variables.

1.7 Rule Quality - Measure of Discrimination

In order to gauge the quality of the generated decision rules, *measure of discrimination* [55] will be used in our study.

Let Q_{MD} denote the measure of discrimination, R denotes rule or a query term in an information retrieval, D denotes the target

function or class of relevant documents and D' denotes the class of non-relevant documents. Q_{MD} can be expressed as follows:

$$Q_{MD} = log \frac{P(R|D)(1 - P(R|D'))}{P(R|D')(1 - P(R|D))}, \quad (3)$$

where P denotes probability.

1.8 Hybrid Reducts

The outcomes from rough sets and Cox methods, reducts and significant condition attributes, will be integrated to yield the new set. This new set contain the informative attributes from rough sets and significant attributes from statistics and is defined as follows:

Definition 7. ***(Hybrid Reducts)*** *Let* $REDU = \{redu_1, redu_2, ... redu_m\} \neq \emptyset$ *be a reducts set, where* m *is a number of attributes contained in the* $REDU$ *set. Let* $SIG = \{sig_1, sig_2, ...sig_n\} \neq \emptyset$ *be a significant condition attribute set, where* n *is a number of attributes contained in the* SIG *set. We define hybrid reducts as follows:*

$$\textit{hybrid reducts} = REDU \cup SIG$$

where $\cup$ *denotes set union operation.*

2 Methodology

2.1 Methodology

Our rough sets hybrid intelligent system architecture of rough sets, relational algebra, survival analysis, statistical learning and medical science is in Fig. 2.

It consists of three modules: attribute mining, model construction and evaluation. Our system serves to build a high level rough sets hybrid intelligent system by the inclusion of domain knowledge in the analysis process. Real world survival data and domain knowledge are input in *Attribute Mining* and are analyzed with the preprocessing step and new CDispro algorithm by the following steps.

- Preprocess survival data sets into a usable format for entry to the system e.g., discretization, inconsistency removal.

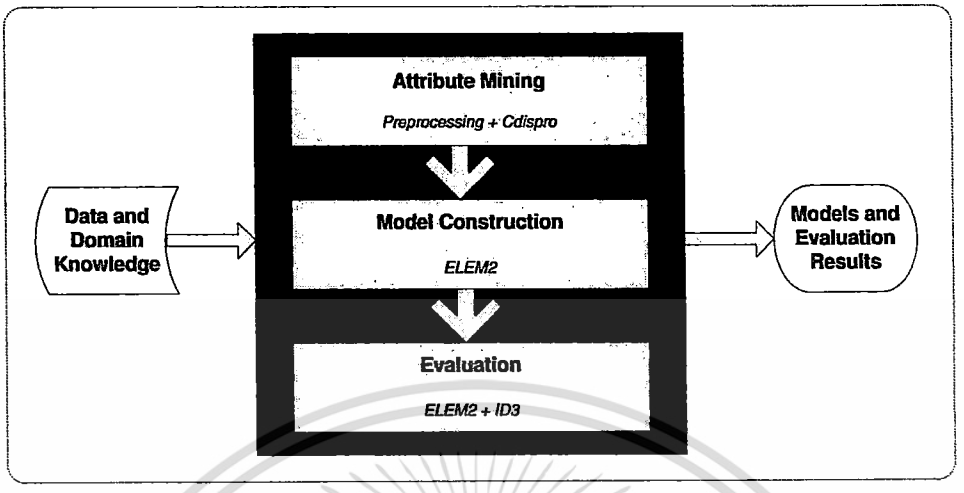


Fig. 2. A perspective of how to build a high level rough sets hybrid smart system.Our proposed rough sets hybrid intelligent system architecture.

- The Kaplan-Meier method analyzes the entire survival data and generates statistical summaries, e.g., life time table and survival curve of overall data. Subsequently, a particular risk factor is included in the Kaplan-Meier method to determine the survival curves with respect to the survival time attribute. Significance levels are tested by three statistical techniques. Overall outcomes are then considered and the probe attribute is identified for all survival data set and sent to next step.
- Rough set theory extracts core attributes, dispensable attributes and reducts. Due to uncertainty, survival data can have no core attributes or use all attributes as core. Our system will always complete despite this issue. The probe attribute from the previous step is used to guide generation of probe reducts.
- If the previous step returns the most distinguished selected attributes to predict survival time, our system sets reducts as the final attribute subset results. Otherwise, the probe attribute is employed to produce probe reducts that simultaneously extract the most informative information.

The outcome is the essential and informative attributes (significant risk factors). All of these acquired attributes will be input to *Model Construction.*

In *Model Construction* module, ELEM2 (Version 3) [55] derives a rough sets model in the form of decision rules for survival prediction. We also induce association rules in this step. Association rules are generated to explore the relative information for each risk factor. This component allows the degree of flexibility and generality that was designed, whereas most existing systems tend to be highly specialized toward a particular kind of rule generation technique. These rules are passed to the last module.

Furthermore, we develop and add other components to perform comprehensive survival analysis. The objective is to expand the utility of our system. The additional process is the multivariate analysis with the Cox method to generate Cox models. Furthermore, hybrid reducts that integrated the outcome from both rough sets and Cox method will be generated.

Then the ELEM2 and ID3 [58] (WEKA software [47]) are used to validate and obtain the evaluation results to guarantee the correctness of the rules. We compare performance outcomes from both the entire data and the reducts/probe reducts data. The final output are rules and their evaluation results. Rough sets has led to many interesting extensions to data mining [8–10] and feature selection [48].

2.2 CDispro algorithm

In this section, we present the main algorithm in the first module of our rough sets hybrid intelligent system. CDispro stands for "**C**ore-**D**ispensable Attributes and **Pro**be Reducts Extraction Algorithm". It was first proposed in [24] and is revised in this study.

CDispro is able to discover essential information from a data set using core attribute and reducts/probe reducts in Definitions 1-6. It comprises two main steps. CDispro discovers core attributes in the first step and provides two kinds of reducts in the following step;

- traditional reducts R, if no probe, and
- user defined probe reducts PR. CDispro takes domain knowledge into account by using an input probe attribute P.

Users can identify a probe attribute to produce the probe reducts PR. The probe attribute is an attribute known to be important for a

particular data set. Our previous study of probe reducts can be found in [16]. This revision of CDispro is also more efficient in computing reducts.

Another key feature of CDispro is to analyze both the relationship among condition attributes and the relationship between condition attributes and its target function.

Algorithm 1 New Core-Dispensable attributes and probe reducts extraction algorithm.

INPUT : A decision table: $T(U, C, D)$
A probe attribute: P
OUTPUT: Core attribute: $Core$
Dispensable attribute: Dis
Probe reducts/Reducts: PR/R
Set Core $= \emptyset$, Dis $= \emptyset$, PR/R $= \emptyset$.
//Construct Core, Dispensable attributes and Reducts or Probe reducts
For each attribute $c_i \in C$
{
IF $Card(\prod(C - c_i + D)) < Card(\prod(C - c_i))$
$Dis = Dis \cup c_i$
ELS IF $Card(\prod(C - c_i + D)) > Card(\prod(C - c_i))$
$Core = Core \cup c_i$ IF $P = \emptyset$
$R = C$ and $Reducts = R$
ELSE
PR = C and Core = Core ∪ Probe and Reducts = PR
//CDispro generates probe reducts if the user enters a predefined probe attribute. Otherwise, traditional reducts will be generated
}
For each $c_j \in \{C - Core\}$\{
//Measure the merit of each condition attribute and compare it to the other condition attributes to generate reducts or probe reducts
Find $merit(c_j, C, D)$
Sort c_j in decreasing order of $merit(c_j, C, D)$
}
Set $V = \{C - Core\}$
WHILE K(R,D) $\neq 1$
Select largest c_j by merit in the list
Reducts = Reducts ∪ c_j
$V = V - c_j$

3 Experiments

"... rough set theory it is not an alternative to classical set theory but it is embedded in it."
Zdzislaw Pawlak (2005), A Treatise on Rough Sets

3.1 Data and materials

We applied our rough sets hybrid intelligent system to the survival analysis data sets in Table 1. Melanoma data set comprises of 7 conditions attributes and 30 examples. The attribute's names are age, ini2, . . . , ini4a and trt. The survival time attribute is the time until a return to drug use and the censor attribute indicates the returning to drug use. While pbc (stands for Primary Biliary Cirrhosis) data set contains 17 condition attributes and 424 patient's records. Hepato, edema, bilir, . . . , alkal, and sgot are the attribute's names. Mayo Clinic trial in pbc of the liver conducted between 1974-1984. All patients referred to Mayo Clinic during that interval, met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine.

Table 1. Experimental data sets

Data sets	Number of condition attributes	Number of example
geriatric$_{nStatus}$	44	8547
geriatric$_{sTime}$	44	8546
melanoma [2]	7	30
PBC [52]	17	424

Table 2 shows description of geriatric data from Dalhousie Medical School (Canada) collected during 2002-2003[1]. We consider this geriatric data as two independent data sets followed by two separate target functions i.e., *geriatric*$_{nStatus}$ and *geriatric*$_{sTime}$, respectively.

(*geriatric*$_{nStatus}$) contains 8547 patient records with *notification status* as the target function. The objective is to develop a model to predict the *notification status* for new patient. *Geriatric*$_{nStatus}$ describes each patient with 44 condition attributes, e.g., *age at investigation, Parkinson's disease.* (*geriatric*$_{sTime}$) has the target function *survival time* (in months). *Geriatric*$_{sTime}$ has *notification status* attribute (dead or alive) as the censor attribute (c.f. [16]). The purpose of the study is to develop a model to predict the *survival time* for

[1] Collection of personal data creates privacy issues. Stronger privacy assurance anonymously requires specific model, and is outside the scope of this work.

Table 2. The geriatric data description

Attribute	Description	Attribute	Description
edulevel	Education level	*hbp*	High blood pressure
eyesight	Eyesight	*heart*	Heart
hearing	Hearing	*stroke*	Stroke
eat	Eat	*arthriti*	Arthritis or rheumatism
dress	Dress and undress yourself	*parkinso*	Parkinson's disease
takecare	Take care of your appearance	*eyetroub*	Eye trouble
walk	Walk	*eartroub*	Ear trouble
getbed	Get in and out of bed	*dental*	Dental
shower	Take a bath or shower	*chest*	Chest
bathroom	Go to the bathroom	*stomach*	Bladder
phoneuse	Use the telephone	*kidney*	Kidney
walkout	Get places out of walking distance	*bladder*	Stomach or digestive
shopping	Go shopping for groceries etc.	*bowels*	Bowels
meal	Prepare your own meals	*diabetes*	Diabetes
housewk	Do your housework	*feet*	Feet
takemed	Take your own medicine	*nerves*	Nerves
money	Handle your own money	*skin*	Skin
health	Health	*fracture*	Fractures
trouble	Trouble	*age*	Age group
livealon	Live alone	*studyage*	Age at investigation
cough	Cough	*sex*	Gender
tired	Tired	*livedead*	Notification status
sneeze	Sneeze	*surtime*	Survival time

each patient based on the training set, then cross-fold validate the model on the test set.

4 Experimental Results

4.1 Attribute mining

In the preprocessing step, since data inconsistency is an issue in rough sets and affects discernibility of the data, we performed a data cleaning step to obtain consistent data. A study of computing with inconsistency can be found in [21]. Subsequently, the consistent data is discretized [16]. Next, all preprocessed data sets are analyzed with the CDispro algorithm.

To determine the Kaplan-Meier estimate of the survival function, we took geriatric data and formed a series of time intervals. Each of these intervals is constructed in such a way that one observed death is contained in the interval. *Status = 0* indicates that the example has been censored and *status = 1* indicates death. Table 3 presents

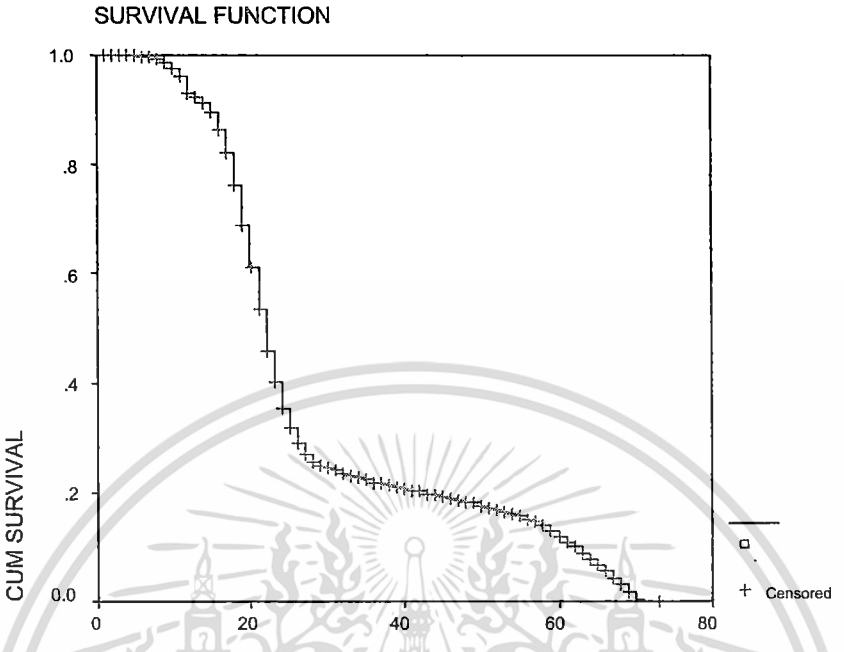
Table 3. Life time table (or the calculation of the Kaplan-Meier estimate) of the survival function of the geriatric data set.

Time (months)	Status	Cumulative survival	Standard error	Cumulative event	Number at risk
1	1			1	8,545
1	1	0.9998	.0002	2	8,544
			.		
			.		
			.		
1	0			2	8,526
2	1			3	8,525
2	1			4	8,524
2	1			5	8,523
2	1	0.9993	.0003	6	8,522
			.		
			.		
			.		
71	1			6,685	2
71	1	0.0010	.0010	6,686	1
73	0			6,686	0

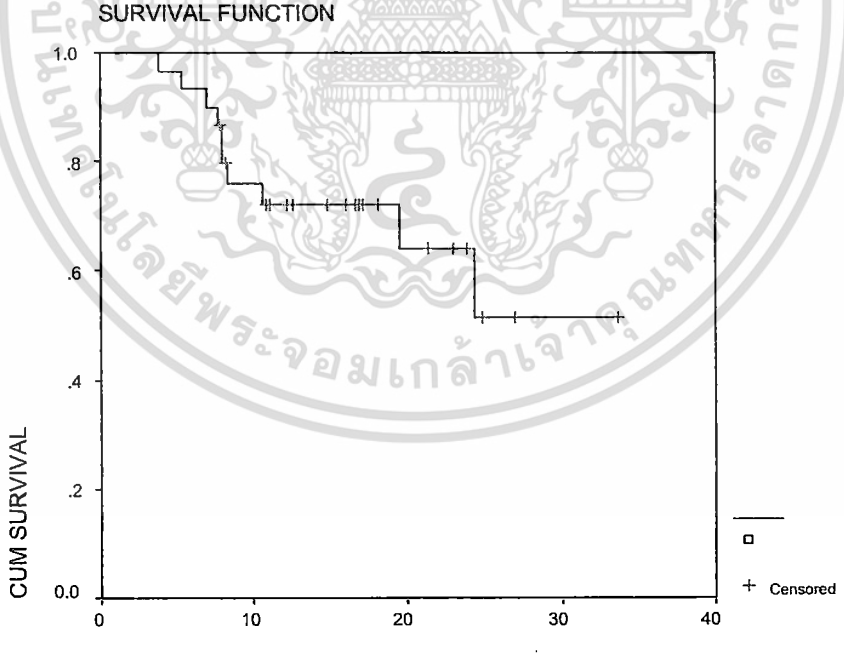
the life time table or the calculation of Kaplan-Meier estimate of the geriatric survival function.

We generate the Kaplan-Meier curves by calculating the Kaplan-Meier estimate of the survival function. A plot of this curve is a step function, in which the estimated survival probabilities are constant between adjacent death times and only decrease at each death. In Fig. 2(a), no risk factor is included, it displays one Kaplan-Meier survival curve in which all geriatric data are considered to belong to one group. The important aspect of the survival function is to understand how it influences survival time. Fig. 2(b) depicts the Kaplan-Meier survival curve in which all melanoma data belong to one group.

After considering all data belonging to one group, we examine the effect of all suspect attributes on survival time. All condition attributes are considered to be risk factors and are included in the Kaplan-Meier method, which is used to separate the data into several subgroups. Figs. 3(a) and (b) depict the analysis of *diabetes* and *Parkinson's* factors of geriatric data, respectively.

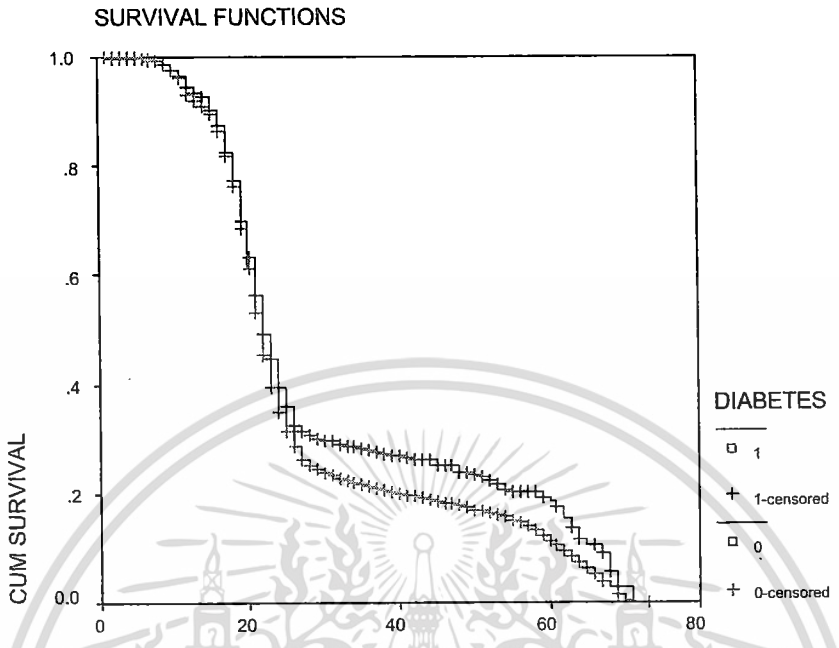


(a) Geriatric survival function: no risk factor is included, the geriatric data are considered to belong to one group.

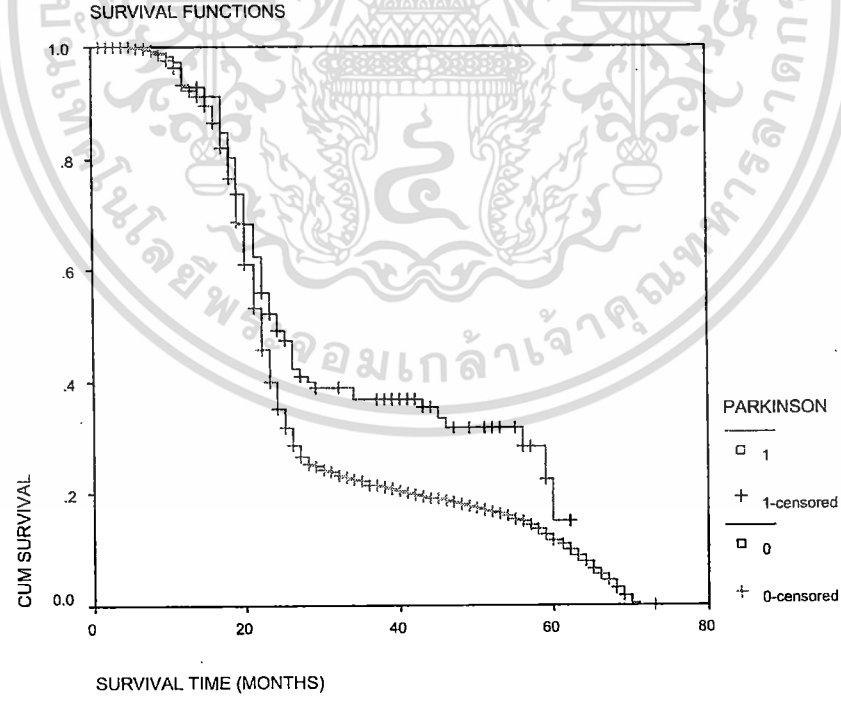


(b) Melanoma survival function: no risk factor is included, melanoma data are considered to belong to one group.

Fig. 3. Survival functions

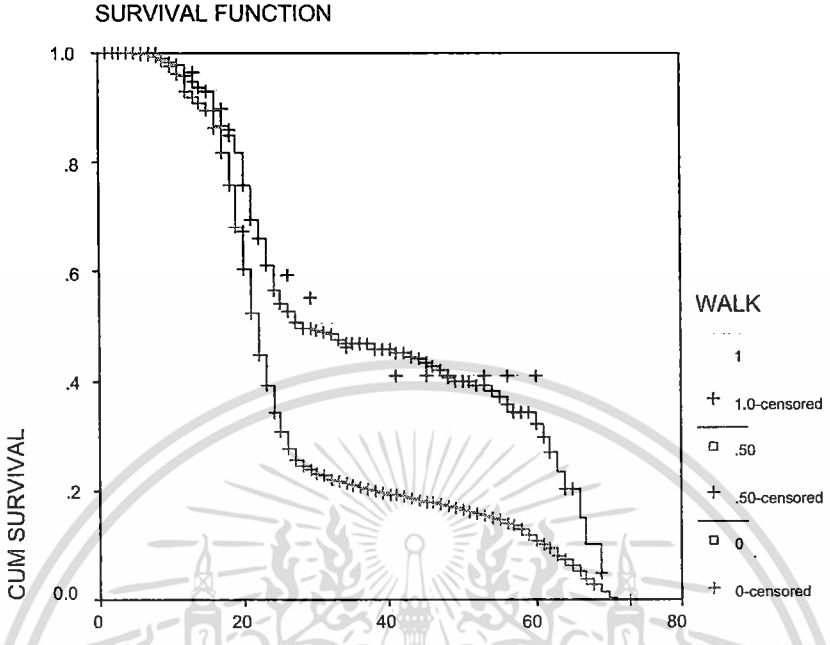


(a) Survival function of the *diabetes* factor from geriatric: two types of diabetes diagnosis {0,1} are compared.

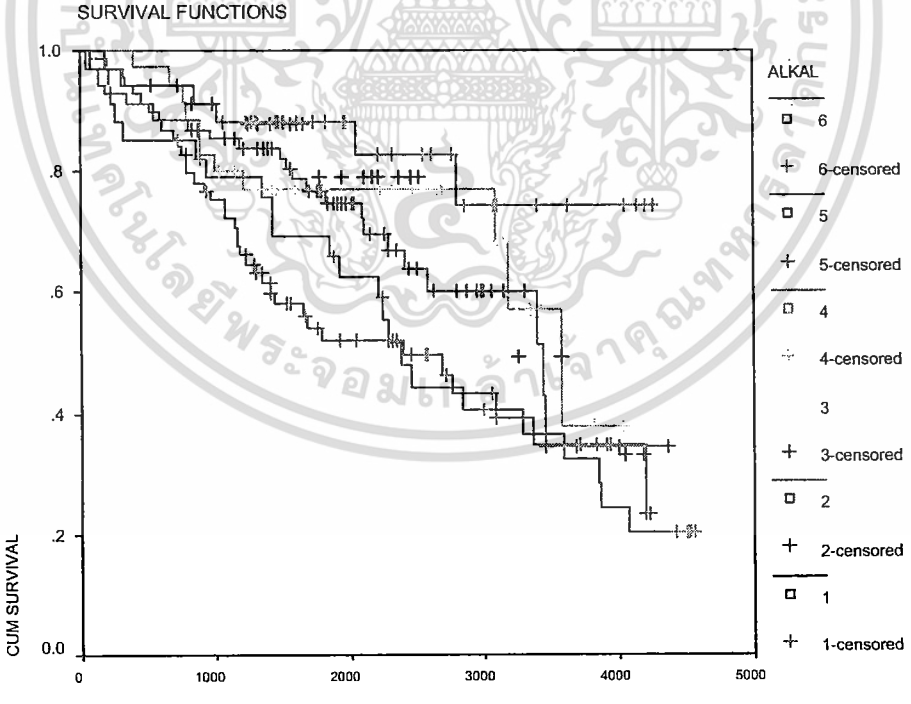


(b) Survival function of the *Parkinson's* factor from geriatric: two types of Parkinson's are compared.

Fig. 4. Survival functions



(a) Survival function of the *walk* factor from geriatric: 3 levels of whether patients can walk are compared.



(b) Survival function of the *akal* factor from PBC: 6 levels of alkaline phospatase in unit/liter are compared.

Fig. 5. Survival functions

In Fig. 3(a) we consider two possible types of diabetes diagnosis $\{0,1\}$ and the same for Fig. 3(b). One can see that the two curves from these groups of patients reveal different survival characteristics. We notice a slight difference between the two groups of diabetes patients and wider differences between two groups of Parkinson's. The patients who diagnosed diabetes and not Parkinson's seem to provide better results. However, the type without Parkinson's seem to have few censor cases left as time goes by. Initially at time zero, cumulative survival is 1. In both figures, the first 15 months after admission to our study reveals very little chance of dying and two groups are visually close together. During the 15-25 month period (the most steep part of the survival functions in Fig. 3(a)), the hazard of death has clearly increased. The patients start to have less risk of dying in the following 3 years.

Figures 4(a) and 4(b) display the analysis of risk factor *walk* (whether patient can walk?) of the geriatric data and risk factor *alkal* (alkaline phosphatase in unit/liter) of the PBC data respectively. Fig. 4(a) illustrates the walk risk factor is ambiguous between patients who cannot walk (group 1) and can walk but need help (group 0.5). We illustrated the walk risk factor to be a dispensable attribute. Alkal of PBC in the last figure is described by six possible groups and show the strong significant difference between each group.

The interpretation of Kaplan-Meier curves is only one significant part and it is not sufficient to design the most dangerous risk factor. In [31], Kaplan-Meier method and Prognostic Index (PI) are applied to head and neck cancer patients. Afterward, rough sets generate the decision rules. However, the hypothesis tests and *p-value* are not considered, thus the complete univariate analysis is required.

Since the comparison of the survival curves for two groups should be based on formal statistical tests, not visual impressions, we use the formal hypothesis tests (Sect. 1.5) to see if there is any statistical evidence for two or more survival curves being different and to complete univariate analysis. In practice, the log-rank tests are used because they do not assume any particular distribution of the survival function and are not bias to earlier period events. We then provide three statistical tests for the equality of the survival function *(degree of freedom = 1)*. The example for risk factor diabetes of geriatric data:

	Statistic	Significance
Log Rank	12.77	.0004
Breslow	5.78	.0162
Tarone-Ware	8.46	.0036

The effects of all risk factors on the survival curves are compared. This allows us to confirm which risk factor impacts survival time of patients significantly and should be considered as a probe attribute. An example series of tests can be found in [25]. We first consider all attributes to be potential candidates for the probe attribute if they have a *p-value* of less than 0.2.

By considering statistical test results and characteristics of survival curves, we can choose the probe attributes reasonably. Bazan et al. [31, 32] proposed an efficient approach for measuring distance between Kaplan-Meier curves. The conclusion for the geriatric data set is that diabetes has the most impact on survival time. In other words, H_1 is accepted and *diabetes* is a significant risk for geriatric patients (see hypotheses in Sect. 1.5). The probe attributes we generated are: {*diabetes*} for geriatric, {*ini2*} for melanoma and {*alkal*} for PBC.

We found core attributes and dispensable attributes by using CDispro. It provides preservation of classification when comparing its extracted attributes and original data while achieving high dimensionality reduction. The results in Table 4 illustrate the selection of core attributes and generation of reducts. {*diabetes*} is the core attribute for both $geriatric_{nStatus}$ and $geriatric_{sTime}$, which means that {*diabetes*} is the significant risk factor for notification status and survival time of our Canadians geriatric data. The previous studies, CDispro and ROSETTA [63], can be found in [24].

Our CDispro algorithm produces dispensable attributes, depicted in Table 5. The absence of these attributes does not decrease the predictive ability from the original data set. In the medical domain, the adoption of dispensable attributes can minimize the expensive series of laboratory tests or drop high risk treatments. Several factors, for example, {*eyesight*}, {*ear trouble*} are not significant risk factor for predicting notification status and survival time of the geriatric data.

We can distil traditional reducts or probe reducts as in (Table 6). Our system produces probe reducts using probe attributes if it

Table 4. Core attributes and reducts results generated from CDispro

Data sets	CDispro core attributes	CDispro reducts
$geriatric_{nStatu}$	edlev hear housw health livealo	edlev hear housw health livealo
	eyetro heart eartrou dental chest	eyetro heart eartrou dental chest
	diabetes study sex hbp	diabetes study sex hbp
$geriatric_{sTime}$	edlev eyesi hear shower phoneuse	edlev eyesi hear shower phoneuse
	meal shopping housew money tired	shopping housew money tired
	sneeze trouble livealo cough sex	sneeze trouble livealo cough sex
	arthriti eyetroub hbp heart bladder	arthriti eyetroub hbp heart bladder
	stroke dental stomac kidney age	stroke dental stomac kidney age
	chest bowels diabetes feet nerves	chest bowels diabetes feet nerves
	skin health fracture	skin health fracture
melanoma	age sex trt	age sex trt
PBC	none	none

Table 5. Dispensable attributes results generated from CDispro

Data sets	CDispro dispensable attributes
$geriatric_{nStatu}$	eyesi shopping trouble cough sneeze arthriti stomac bladder feet
	nerves skin fracture
$geriatric_{sTime}$	eartroub walk
melanoma	none
PBC	none

Table 6. Reducts and probe reducts from our system.

Data sets	Reducts	Probe reducts
geriatric	edulevel eyesi hear shower phoneuse shopping meal housew money health trouble livealo cough tired sneeze hbp heart stroke arthriti eyetroub dental chest stomac kidney bladder bowels diabetes feet nerves skin fracture age6 sex	N/A
melanoma	age, sex, trt	age, sex, trt, ini2, ini3a
PBC	none	alkali, drug

returns reducts that do not clarify pattern groups of survival. For example, the PBC data set use {*alkal*} as probe attribute and {*alkal, drug*} as probe reducts for handling situation with no reducts.

4.2 Model construction

All acquired attributes from the first module are passed to the second module, *Model Construction.*

ELEM2 generates decision rules in the form:

"If C_1 *is* c_1 *and* C_2 *is* c_2 *then* D *is* d_1*"*

where c_1, c_2 and d_1 are possible values corresponding to attribute C and D, respectively.

This rule can be used to predict the outcome in new data such as survival time of new elderly patient. Among over 800 rules of geriatric data, example rules of $geriatric_{sTime}$ are:

Decision Rule 1: *IF (health>0.25) and (hear=0) and (nerves=0) and (feet=0) and (heart=0) and (dental=0) and (stomac=0) and (hbp=0) and (diabet=0) and (age≤2) THEN (survival time = 7-18 months)*

Decision Rule 2: *IF (sex=0) and (edlevel=2) and (eyesi>0) and (0< health≤1) and (0<hear≤0.25) and (diabet=1) and (tired=0) and (feet=0) THEN (survival time = 56-73 months).*

The first medical diagnosis rule can be interpreted as:

- If the patient is unhealthy and
- has severe hearing damage and
- nerve problem and
- foot problem and
- heart disease and
- dental disease and
- stomach disease and
- high blood pressure and
- especially those who experience diabetes
- then the patient has a tendency of survival time around 7-18 months after being admitted to our study.

The rule quality (Sect. 1.7)of this rule is 1.8598.
The second rule is interpreted as:

- If a female patient has a low education level and
- an eyesight problem from low to serious type and
- a health problem from low to serious type and
- can hear quite well and
- does not have diabetes experience and
- is easily tired and
- has foot problems
- then the patient is likely to have a survival time between 56-73 months.

The rule quality is 1.5761.

As these rules show, the {*diabetes*} affects the survival time significantly. Our previous studies on univariate analysis have shown that {*diabetes*} is a very significant risk factor by using the Kaplan-Meier method and Log rank test [16]. From decision rule 1, we see that *combinations* of risk factors possibly affect the survival time. Thus, we perform multivariate analysis and {*diabetes, heart, trouble, getbed, walk, age, sex*} are the significant risk factors by using the Cox method [18]. Furthermore, these rules result in easy interpretation of survival prediction rules and can be read without prior expert knowledge.

The sample survival prediction rules from PBC data with the rule quality 2.0810 is:

Decision Rule 3: *IF (age>2) and (biliru≤3) and (albumi>3) and (alkal> 2) and (sgot>1) and (prothr>3) THEN (survival time = 1361-1781 days).*

We perform association rule generation to explore the strong relationship in the geriatric data set. The following are the strong relationship association rule examples we obtained from the geriatric data set.

Association rule 1: *If (getbed=0) and (takemed=0) then (eat=0).*

Association rule 2: *If (dress=0) and (takecare=0) and (getbed=0) and (parkinso=0) then (eat=0).*

Both association rules have support numbers 7921 and 7941, respectively. Both association rules have confidence equal to 1. The interpretation of these association rules are as follows.

First rule: if patients can get in and out of bed and take medicine by themselves very well then these patients can eat very well.

Second rule: if patients can dress and undress, take care of their appearances and can get in and out of bed by themselves very well, (even patients that experience Parkinson's) then they are likely to eat very well.

The mulitvariate analysis of our proposed method is Cox mothod. We took the significant variables mentioned above and treated them as the explanatory variables when constructing the Cox proportional hazard model. Our analysis results in the following.

Table 7. Case processing summary

Cases available in analysis	Number	Percent
Event(LIVMONTH)	6686	78.2%
Censored	1860	21.8%
Cases dropped	0	0.0%
Total	8546	100.0%

As Table 7 shows, the geriatric data set we used contains 8,546 records, with 1,860 censoring with no missing values, no cases with negative time and no cases before the earliest event in the stratum.

Table 8. Test of the coefficients; -2 Log likelihood of the null model ($-2LL_0$)

$-2LL_0$
110591.558

In Table 8, the $-2LL_0$ value of the model without any explanatory variable included is 110591.558.

The enter method is used and results in Table 9. *Diabetes* risk factor is entered as the first variable. The likelihood of the model with *diabetes* ($-2LL_1$) is 110579.417. The $-2LL$ decreases by 12.141.

Table 9. Test of the coefficients; -2 Log likelihood of the model with *diabetes* ($-2LL_1$)

$-2LL_1$	Change from the previous step		
	Chi-square	df	Sig.
110579.417	12.141	1	.000

This decline 12.141 is significant when considering the last column, $p - value = 0.0004932$.

Table 10. Explanatory variable *diabetes* in the equation with $df = 1$

Var.	b	SE	Wald	Sig	$Exp(b)$	95% CI L/U
DIABETES	-.148	.043	1.654	.001	.862	.792/.939

The interpretation of this risk factor in the Cox model is the hazard of a patient death is decreased 0.862 times if a patient has experienced diabetes diagnosis (Table 10). We concluded that *diabetes* is a significant risk factor.

The stepwise enter method is used until the last step when all explanatory variables are included as demonstrated in Tables 11-13.

Table 11. Test of the coefficients; -2 Log likelihood of the model with all explanatory variables

$-2LL$	Overall (score)		
	Chi-square	df	Sig.
110032.119	513.312	32	3.85e-088

In the last step, the $-2LL$ of the last step differ from the one in Fig. 8, 547.298 which is significant. In Table 11, the overall $-2LL$ is 110032.119.

The results in Table 13 demonstrate the utility and versatility of our system for analyzing survival data efficiently. Our system is also versatile with the comprehensive univariate and multivariate analysis processes. Nonetheless, the results from Tables 4 and 13 illustrate two different outcomes from two different aspects; discernibility relation from mathematics and population estimates from statistics.

Table 12. Explanatory variables *heart, sex* and *shower* in the equation with $df = 1$

Var.	b	SE	Wald	Sig	$Exp(b)$	95% CI L/U
HEART	-.072	.030	5.765	.016	.931	.878/.987
SEX	.165	.023	7.551	.000	1.179	1.119/1.243
SHOWER	.017	.094	.033	.856	1.017	.846/1.224

Table 13. Significant condition attributes produced from Cox model in the last step

Data sets	Significant attributes
geriatric	*diabetes heart trouble getbed walk age sex*

They are clearly not the disjoint sets or subset of each other, but they both are of inexpedient from both point of views.

One most important concept of this research is to integrate mathematics and statistics to generate the better outcomes that best suit for multidiscipline study. Thus, our new reducts (Definition 4) is constructed from the set operation as the following.

hybrid reducts $= \{$*edulevel eyesi hear shower phoneuse shopping meal housew money health trouble livealo cough tired sneeze hbp heart stroke arthriti eyetroub dental chest stomac kidney bladder bowels diabetes feet nerves skin fracture age sex getbed walk*$\}$

These hybrid reducts is constituted both mathematics and statistics importance. It should be considered to be the most informative and significant attributes for the survival data and for further analysis.

5 Validation and Comparison to Existing Results

5.1 Validation

The results illustrate a compact and easy interpretation of survival prediction rules with no medical expertise required. The improvements of all rule performance compared to rule constructions from entire data and from reducts/probe reducts are depicted in Fig. 6.

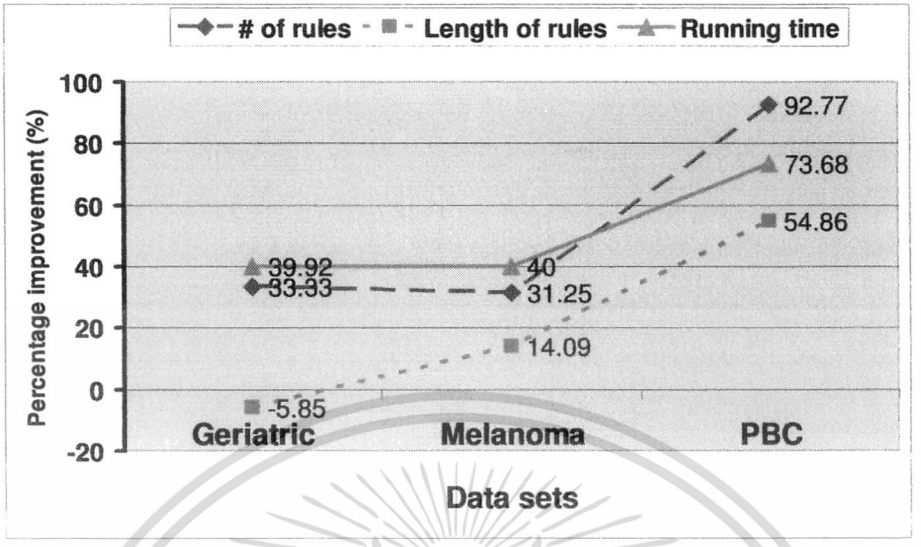


Fig. 6. Improved performance of the generated rules from the geriatric, melanoma and PBC data sets.

Almost all rule performance outcomes are improved (except the average number of geriatric survival prediction rules). The geriatric data is improved on average 24.47% for all outcomes. The melanoma and PBC data are improved average 28.45% and 73.77% for all outcomes respectively. Further, the average number, length and running time of the rules is improved an average of 52.45%, 21.03% and 51.20% for all data sets respectively.

After generating the survival prediction model with the previous paradigm, we will illustrate the quality of the rules by using the optional validation process. We run 10-fold cross validation with ID3 to illustrate the utility of derived rules. We use rule quality measurements: recall, precision, F-score and accuracy to gauge the quality of rules.

We then compare the improvement of rules generated from the entire data to those generated from reducts/probe reducts. Our validation process demonstrates the improvement for all measurements in Fig. 7. The validation results illustrate an average improvement of 2.82% while the running time improved on average 23.01%.

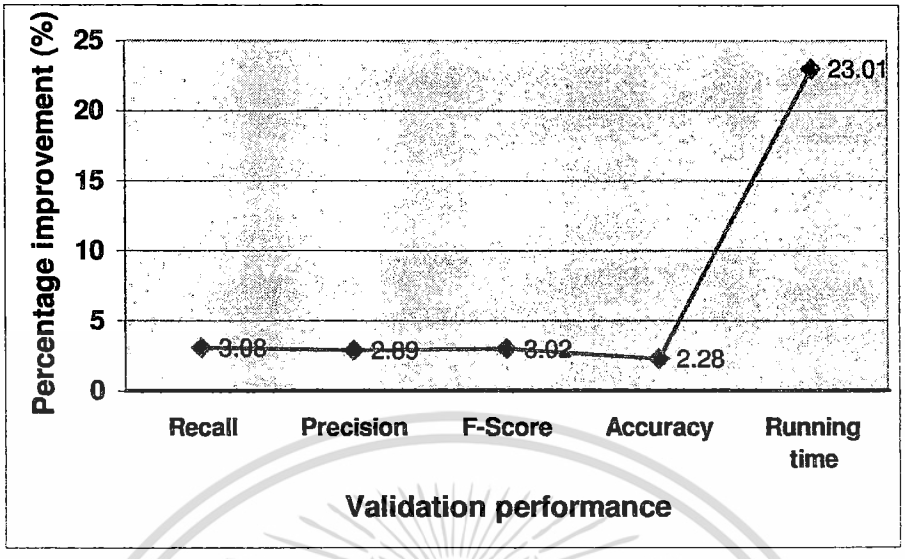


Fig. 7. Improved performance from 10-fold cross validation by ID3.

5.2 A case study: comparison to ANNs and frailty index

The same geriatric data was analyzed by other methods [22] to evaluate the potential of rough sets, artificial neural networks (ANNs) and the frailty index in predicting survival time. For the ANNs, randomly selected participants formed the training sample to derive relationships between the 40 variables and survival. An ANN's output was generated for each subject and a separate testing sample was used to evaluate the accuracy of prediction. An individual frailty index score was calculated as the proportion of deficits experienced (c.f. [22]). The output of the rough sets rules, an ANN's model and an unweighted frailty index in predicting survival patterns were measured using the accuracy rate. The accuracy rate of rough sets rules from validation was 83.79%–90.57% [16]. At the optimal receiver operating characteristic (ROC) value, the accuracy of the frailty index was 70.0%. The ANNs accuracy rate over 10 simulations in predicting the probability of individual survival was $79.2 \pm 0.8\%$.

This geriatric$_{sTime}$ data was analyzed from different points of view: (i) The unweighted frailty index captured the relationship of the geriatric$_{sTime}$ data successfully, (ii) ANNs are able to automatically discover the non-linear characteristics in this data, (iii) rough sets offer the capability to handle vagueness and illustrated its use-

fulness on this data. Due to vagueness, redundancy and irrelevant attributes in the data, we conclude that rough sets and its discernibility relation can improve the analysis process efficiently and effectively.

6 Conclusion

Starting from mathematical rough set theory the central theme of this study is to invent the hybrid intelligent system. Our rough sets hybrid intelligent system is useful for survival analysis and extracting the most informative and useful knowledge. We created our system to have, the following features. Our system was designed to provide comprehensive survival data analysis tasks; preprocessing, analyzing process and postprocessing. We amalgamated rough sets and other techniques in soft computing to be able to make the analyzing process tolerant to imprecise and uncertain data. We guaranteed the correctness of rules by designing automatic validation processes. Furthermore, the computation times were improved significantly by using database operations. The experimental results show how our rough sets hybrid intelligent system could be employed to quickly process. Clinical diagnosis questions can be answered successfully e.g., is diabetes a significant factor for survival time of geriatric patients? Analysis results, show that it has significant impact on the survival time of geriatric patients. Decision rules described particular tendencies for survival outcomes of patients by using decision rules that are straightforward and simple to use.

Part II

7 Introduction

We solved one significant issue in survival data analysis that is the desire to automate the analysis processes in Part I. Traditional hard computing approaches do not perform automatic learning and thus they are less efficient for this data type. These hard computing approaches also require substantial space and time which can cause problems for large scale data. Moreover, as pointed out by Pawlak [59] precise reasoning from inexact data would be impossible. Thus, soft computing approaches are required. Several hybrid approaches in artificial intelligence provide a variety of soft computing methods that enhance the scalability significantly [1] e.g., rough sets [33] and relational database analysis as we illustrated in Part I.

We raise four research questions in this part. First, in some situations, smaller units of full data set are of great importance to analyze (e.g., men vs. women patients, young vs. old customers). Thus, the finer knowledge consideration is important in some aspects. This article is intended to present mathematical proof and a case study of Pawlak's statement about *distributed knowledge* [33] (as stated above).

Second is the consideration of a case study on self-reported geriatric data (from Dalhousie Medical School, Canada). As stated in [22], the combination of attributes for predicting prolongation time might be different in men and women. Furthermore, when considering data which is self-reported, it usually contains optimistic bias. This research will illustrate that rough sets in combination with statistics, relational databases and systematic hypothesis tests is tolerant and robust in learning to extract significant risk factors and induce the rules for predicting survival time. The self-reported data which is usually optimistic can be reasoned by rough sets approach and overcome such bias successfully.

Third, the general models of knowledge discovery in databases (KDD) contains processes including data preprocessing, knowledge discovery algorithms, rule generations and evaluations. Rule evaluation is a significant process in KDD. How to automatically extract important, representative rules to the human beings instead of selecting those useful rules manually are the main problems. Specific difficulties make the research of rule evaluation very challenging.

One of the difficulties is that real-world large data sets normally contain missing attribute values. They may come from the collecting process, or redundant scientific tests, change of the experimental design, privacy concerns, ethnic issues, unknown data and so on. Discarding all the data containing the missing attribute values cannot fully preserve the characteristics of the original data, and wastes part of the data collecting effort. Knowledge generated from missing data may not fully represent the original data set, thus the discovery may not be as sufficient. Understanding and utilizing of original context and background knowledge to assign the missing values seem to be an optimal approach for handling missing attribute values. In reality, it is difficult to know the original meaning for missing data from certain application domains. Another difficulty is that huge amount of rules are generated during the knowledge discovery process, and it is infeasible for humans to manually select useful and interesting knowledge from such rule sets.

Rough sets theory, originally proposed in the 1980's by Pawlak [33], was presented as an approach to approximate concepts under uncertainty. The theory has been widely used for attribute selection, data reduction, rule discovery and many knowledge discovery applications in the areas such as data mining, machine learning and medical diagnoses.

We are interested in tackling difficult problems in knowledge discovery from a rough sets perspective. In this research, we introduce how rough sets based rule evaluations are utilized in knowledge discovery systems. Three representative approaches based on rough sets theory are introduced. The first approach is to provide a rank of how important is each rule by rule importance measure (RIM) [61]. The second approach is to extract representative rules by considering rules as condition attributes in a decision table [62]. The third approach is applied to data containing missing values. This approach provides a prediction for all the missing values using frequent itemsets as a knowledge base. Rules generated from the complete data sets contain more useful information. The third approach can be used at the data preprocessing process, combining with the first or second approach at the rule evaluation process to enhance extracting more important rules. It can also be used alone as preprocessing of missing attribute values.

We address particular problems from real-world data sets, using recent missing attribute value techniques and rule evaluations based on rough sets theory to facilitate the tasks of knowledge discovery. The rule discovery algorithm focuses on association rule algorithms, although it can be classification algorithm, decision tree algorithm and other rule discovery algorithms from data mining and machine learning. We demonstrate the rule evaluation approaches using a real-world geriatric care medical data set.

Finally, given user profile information, recommender systems attempt to predict items (e.g., music, books, web pages) in which a user might be interested. Thus, general recommender systems were developed for e-commerce. Several studies have shown that systems predict the target user's requirement accurately (e.g., movies [81] and hardware retail [82]).

Our recommendation system was designed with the goal of providing accurate, low-cost medical recommendations. In countries where health care costs are prohibitively expensive, this system can provide a free alternative. While not seeking to be a drop-in replacement or perfect substitute for professional medical advice, there are many cases where some information is better than nothing. Consider the following examples. A patient can only afford a limited number of tests, but cannot determine which ones should take priority. There may be an inexperienced, or no doctor available and the patient would like a second opinion.

Regular check-ups allow doctors to diagnose diseases early, allowing wider options for treatment. Many patients cannot afford these regular visits, and instead only are examined when there is a problem. A free recommendation from our system may allow earlier diagnosis. Rural areas with limited access to medical professionals can also benefit from more frequent medical advice. To provide this service, we aim to eventually deploy this system in low-cost public kiosks. In this research, we describe the design of a web-based prototype.

8 Preliminaries and Notation

8.1 Rough Set Theory

The advantages of rough sets to survival analysis (adapted and extended from [60]) are outlined as follows.

1. Tolerant to vague and uncertain data which is commonly founded in actual data.
2. Discover relationships that would not be found while using traditional hard computing approaches.
3. Provide an efficient method for reasoning and extracting hidden patterns in data.
4. Discard unimportant attributes of data.
5. Generate decision rules.
6. Offer an easy to understand and straightforward interpretation of the rules.

Core Generation. Hu et al. [67] introduced a core generation algorithm based on rough sets theory and efficient database operations, without generating reducts. The algorithm is shown in Algorithm 2, where C is the set of condition attributes, and D is the set of decision attributes. $Card$ denotes the count operation in databases, and Π denotes the projection operation in databases.

Algorithm 2 Hu's Core Generating Algorithm [67]

Input Decision table $T(C, D)$,
C is the condition attributes set;
D is the decision attribute set.
Output $Core$,
Core attributes set.
begin
$Core \leftarrow \phi$
$Foreachconditionattribute$A $\in C${
If $Card(\Pi(C - A + D)) \neq Card(\Pi(C - A))$
{
$Core = Core \cup A$
}
}
$Return$Core

This algorithm is developed to consider the effect of each condition attribute on the decision attribute. The intuition is that, if the core attribute is removed from the decision table, the rest of the attributes will bring different information to the decision making. Theoretical proof of this algorithm is provided in [67]. The algorithm takes advantage of efficient database operations such as count and projection. This algorithm requires no inconsistency in the data set.

8.2 Rough Sets based KDD Systems

We briefly survey current rough sets based knowledge discovery systems. We discuss the individual functions of each system based on general characteristics, such as the input data sets, the preprocessing tasks, the related rough sets tasks, the rule generations and so on.

1. **ROSETTA** ROSETTA [63] software is a general purpose rough set toolkit for analyzing the tabular data, and is freely distributed. The downloadable versions for both the Windows and Linux operating systems are available. The software supports the complete data mining process, from data preprocessing, including processing incomplete data, data discretization, generating reduct sets which contain essential attributes for the given data set, to classification, rule generation, and cross validation evaluation. Some discretization and reducts generation packages are from RSES library [64].
2. **RSES2.2** RSES [64] stands for Rough Set Exploration System. There are downloadable versions for both the Windows and Linux operating systems. It is still maintained and being developed. The system supports data preprocessing, handling incomplete data, discretization, data decomposition into parts that share the same properties, reducts generation, classification, and cross validations and so on.
3. **ROSE2** ROSE [65] stands for Rough Sets Data Explorer. This software is designed to process data with large boundary regions. The software supports data preprocessing, data discretization, handling missing values, core and reducts generation, classifications and rule generation, as well as evaluations. This software

provides not only the classical rough set model, but also the variable precision model, which is not provided by [63] and [64].

4. **LERS** LERS [68] stands for Learning from Examples based on Rough Sets. It is not publicly available. The system was designed especially to process missing values of attributes and inconsistency in the data set. Certain rules and possible rules are both extracted based on the lower and upper approximations.

In addition to the rough sets based systems mentioned above, there are other available knowledge discovery systems based on the methodologies of rough sets such as GROBIAN [69] and DBROUGH [70].

8.3 Current Research on Rule Evaluations

Rule generation often brings a large amount of rules to analyze. However, only part of these rules are distinct, useful and interesting. How to select only useful, interesting rules among all the available rules to help people understand the knowledge in the data effectively has drawn the attention of many researchers. Research on designing effective measures to evaluate rules comes from statistic, machine learning, data mining and other fields. These measures fall into two categories of evaluation measures.

Rule Interestingness Measures. One category of evaluating rules is to rank the rules by rule interestingness measures. Rules with higher interestingness measures are considered more interesting. The rule interestingness measures, originated from a variety of sources, have been widely used to extract interesting rules. Different applications may have different interestingness measures emphasizing on different aspect of the applications. Hilderman provided an extensive survey on the current interestingness measures [71] for different data mining tasks. For example, *support* and *confidence* are the most common interestingness measures to evaluate the association rules.

Not all the interestingness measures generate the same rank of interestingness for the same set of rules. Depending on different application purpose, appropriate rule interestingness measures should be selected to extract proper rules. More than one measure can be

applied together to evaluate and explain the rules. Tan et. al. [72] evaluate twenty one measures in their comparative experiments and suggest different usage domains for these measures. They provide several properties of the interestingness measures so that one can choose a proper measure for certain applications. Their experiments also imply that not all the variables perform equally good at capturing the dependencies among the variables. Furthermore, there is no measure that can perform constantly better than the others in all application domains. Different measure is designed towards different domains.

Rule Quality Measures. The concept of rule quality measures was first proposed by Bruha [73]. The motivation for exploring this measure is that decision rules are different with different predicting abilities, different degrees to which people trust the rules and so on. Measures evaluating these different characteristics should be used to help people understand and use the rules more effectively. These measures have been known as rule quality measures.

The rule quality measures are often applied in the post-pruning step during the rule extraction procedure [74]. For example, some measures are used to evaluate whether the rules overfit the data. When removing an attribute-value pair, the quality measure does not decrease in value, this pair is considered to be redundant and will be pruned. As one of the applications, rule generation system uses rule quality measures to determine the stopping criteria for the rule generations and extract high quality rules. In [79] twelve different rule quality measures were studied and compared through the ELEM2 [74] system on their classification accuracies. The measures include empirical measures, statistical measures and measures from infomation theory.

8.4 Recommendation Rules

In general, a decision rule can have more than one antecedent (combined either by AND or OR logical operations). Similarly, a decision rule can have more than one consequent. Antecedents or consequents can describe unary relations, e.g.,

Bangkok is raining. Bangkok has a lot of traffic.

Antecedent or consequent can describe binary relations, e.g.,

Bangkok has more traffic than Toronto.

Decision rules can describe relations, e.g.,

IF Bangkok has a lot of traffic THEN travel on the subway.
IF Bangkok is raining THEN bring an umbrella.

Here, we analyzed the relationship within decision rules and added expert knowledge to generate a *recommendation rule* [83]. The following is an example of a recommendation rule that takes a set of inputs and gives advice as a result.

IF Bangkok has a lot of traffic AND Bangkok is raining
THEN travel on the subway AND bring an umbrella.

We introduce two measurements that will be used in our system with our recommendation rules. Given n, the total number of facts in the database, m the total number of rules, R_i is rule number i and p_{ij} is the priority of each fact j for rule i, the *rule priority* is calculated from the value-pair condition attribute that matched the fact as follows:

$$RULE\_PRIORITY(R_i) = \frac{1}{n}\sum_{j=1}^{n} p_{ij} \tag{4}$$

where i runs through $1, 2, \ldots, m$.

In some cases, a patient's input might match more than one recommendation rule and several rules are triggered. Our system will calculate the *recommendation score* from the rule quality [55], rule cover and rule priority (4) as follows:

$$RECOMMENDATION\_SCORE(R_i) = (RULE\_QUALITY(R_i) + RULE\_PRIORITY(R_i)) + \left(\frac{RULE\_COVER(R_i)}{n}\right). \tag{5}$$

Only one recommendation rule will be fired if the highest recommendation score exceeds the *recommendation threshold*, α. Otherwise, multiple recommendation rules will be fired.

8.5 Recommender Systems

A recommender system is a decision support system that provides a personalized solution in a brief and clear form from the user's

given information. In this study, we construct a recommender system based on rule-based systems. A recommender system consists of three components: (i) a database of rules, (ii) a database of facts and (iii) an inference engine [83]. First, the knowledge base contains a set of rules that represent the knowledge possessed by the system (e.g., from previous analysis). Second, the database of facts represents inputs to the system that are used to cause actions or derive recommendations. Finally, the inference engine is the part of the system that generates a recommendation. The inference engine uses the rules and facts when inferring recommendations.

Our system uses deduction to reach a recommendation from a set of antecedents, this is called *forward chaining*. This approach begins with a set of facts and rules, and tries to find a way of using those rules and facts to deduce a recommendation or a suitable action. To apply forward chaining, the first step is to take the facts from the fact database and check if any (combination) of these matches all the antecedents of the rules in the rule database. When all the antecedents of a rule are matched by facts in the database, then this rule is *triggered*. The rule is then *fired*, which means its conclusion is added to the facts database. If the conclusion of the rule that has fired is an action or a recommendation, then the system makes recommendations or actions take place. In our study, the only action is to provide recommendations.

9 Theoretical Discovery and Methodology

9.1 Rough Set Theory

The purely statistical measurement gives reasonable evidence to support the hypothesis. When considering noisy real-world data, however, purely statistical measures can be less meaningful. Rough sets approaches that can handle incomplete information are thus required. For these reasons, rough set theory is the last and most important technique to turn this study into a hybrid approach.

The followings are the rudiment of rough set theory. Any subset $X, Y \subseteq U$ of the *universe* U will be called a *concept* (or a *category*) in U. Let us assume a *decision table* denoted by $T(U, C, D)$, where C is the set of *condition attributes* and D is a singleton set of *target*

function. Let $[D] = \{\underline{D_1}, \underline{D_2}, \ldots, \underline{D_k}\}$ denote the equivalent classes induced by D, and $\forall A \subseteq C$, $[A] = \{\underline{A_1}, \underline{A_2}, \ldots, \underline{A_m}\}$ denotes the equivalent classes induced by A. If S is a subset of C or D and $t_i[S] = t_j[S]$ then tuples t_i and t_j are in the same equivalent class induced by attributes S [75].

Definition 8. *[75] For a given set $\underline{D_j}$, the* lower approximation, *$Lower_{[A]/\underline{D_j}}$, of $\underline{D_j}$ under $A \subseteq C$ is the union of all equivalent classes $\underline{A_i}$, each of which is contained by $\underline{D_j}$ and is defined as follows:*

$$Lower_{[A]/\underline{D_j}} = \bigcup\{\underline{A_i} | \underline{A_i} \subseteq \underline{D_j}, i = 1, 2, \ldots, m\}.$$

For any object $t_i \in Lower_{[A]/\underline{D_j}}$, t_i can be classified certainly to $\underline{D_j}$ and is defined as follows:

$$Lower_{[A]/\underline{D}} = \bigcup\{Lower_{[A]/\underline{D_j}} | \underline{D_j} \in [D], j = 1, 2, \ldots, k\}.$$

Definition 9. *[75] An attribute $C_j \in C$ is a* core attribute *in C with respect to D if*

$$Lower_{[C]/[D]} \neq Lower_{[C-C_j]/[D]}.$$

Definition 10. *Let $CORE(X)$ be a set of core attributes of set X. The cardinality of the set of core attributes is denoted by $n(CORE(X))$.*

Proposition 1 *[33]* $Lower(X) \cup Lower(Y) \subseteq Lower(X \cup Y)$.

Directly from Proposition 1, we can derive the following simple properties and propositions.

Property 1 $Lower(X) \subseteq Lower(X \cup Y)$.

Property 2 $Lower(Y) \subseteq Lower(X \cup Y)$.

Proposition 2 *If $X, Y \neq \emptyset$ and $X \cap Y = \emptyset$, then*

$$Lower(X) \cup Lower(Y) \subset Lower(X \cup Y),$$
$$Lower(X) \subset Lower(X \cup Y) \quad and$$
$$Lower(Y) \subset Lower(X \cup Y).$$

Proof. Since $X \subset X \cup Y$ and $Y \subset X \cup Y$, we thus have $Lower(X) \subset Lower(X \cup Y)$ and $Lower(Y) \subset Lower(X \cup Y)$, which gives $Lower(X) \cup Lower(Y) \subset Lower(X \cup Y)$. Properties 1, 2 yield $Lower(X) \subset Lower(X \cup Y)$ and $Lower(Y) \subset Lower(X \cup Y)$, respectively.

Next, we come to main contribution of mathematical proof of distributed decision table.

Proposition 3 *If we are given $X, Y \subseteq U$, and $X \cup Y = C$, where $X, Y \neq \emptyset$ then*

$$n(CORE(X)) < n(CORE(X \cup Y)) \quad and$$
$$n(CORE(Y)) < n(CORE(X \cup Y)).$$

Proof. Let $X, Y \subseteq U$ and $X \cup Y = C$ are in a decision table $T(U, C, D)$. For a subset $X \subset U$, let A be a core attribute of X, then by Definitions 8, 9, $\exists t$ such that $t \in Lower_{[X]/[D]}$ but $t \notin Lower_{[X-A]/[D]}$ which gives

$$Lower_{[X]/[D]} \neq Lower_{[X-A]/[D]}.$$

Next, let B be a core attribute of C, then by Definitions 8, 9, $\exists t_0$ such that $t_0 \in Lower_{[C]/[D]}$ but $t_0 \notin Lower_{[C-B]/[D]}$ which gives

$$Lower_{[C]/[D]} \neq Lower_{[C-B]/[D]}.$$

Because $[X]/[D] \subset [C]/[D]$ by Proposition 2, there is $t \neq t_0$ such that $t \in Lower_{[X]/[D]}$ but $t \notin Lower_{[X-A]/[D]}$ which gives $Lower_{[X]/[D]} \neq Lower_{[X-A]/[D]}$ whereas $t_0 \notin Lower_{[C]/[D]}$ and $t_0 \notin Lower_{[C-B]/[D]}$.

Hence, $n(CORE(X)) < n(CORE(X \cup Y))$.

Similarly, for a subset $Y \subset U$, we obtain $n(CORE(Y)) < n(CORE(X \cup Y))$.

Let us briefly comment on Proposition 3. As Pawlak pointed out in [33] (Proposition 1) that in general the knowledge included in a distributed knowledge base is less than in the integrated one. Thus, dividing a decision table to smaller units causes loss of some information. In this research, Proposition 3 says that the number of core attributes of distributed decision table or database is less than the integrated one.

For example, Song et al. [22] analyzed the same geriatric data set that we use in the present research. They remarked that the differences between genders should be determined. Hence, dividing this geriatric data set to decision tables of male and female patients will be considered in next section. The experimental results to support our results in this section will be presented in Tables 16, 17, respectively.

9.2 Rule Evaluation on Knowledge Discovery

In this section, we first examine a current rough set knowledge discovery system, and suggest the importance of rule evaluations. We propose rule evaluation approaches and their functions in knowledge discovery systems.

Analyzing RSES – Rough Set Exploration System We take the RSES [64] system as an example system, and study in more detail of the role of rule evaluations. We show that current systems are limited with regard to rule evaluation, and we emphasize the importance of rule evaluation in current knowledge discovery systems.

RSES (Rough Set Exploration System) is a well developed knowledge discovery system focusing on data analysis and classification tasks, which is currently under development. Figure 8 shows a use of the system on a heart disease data set for classification rule generation.

The data input to RSES is in the form of decision table $T = (C, D)$, where C is the condition attribute set and D is the decision attribute set. Preprocessing is conducted once the data is imported to the system, during which stage the missing attribute values are handled and discretization is performed if necessary as well. Reducts

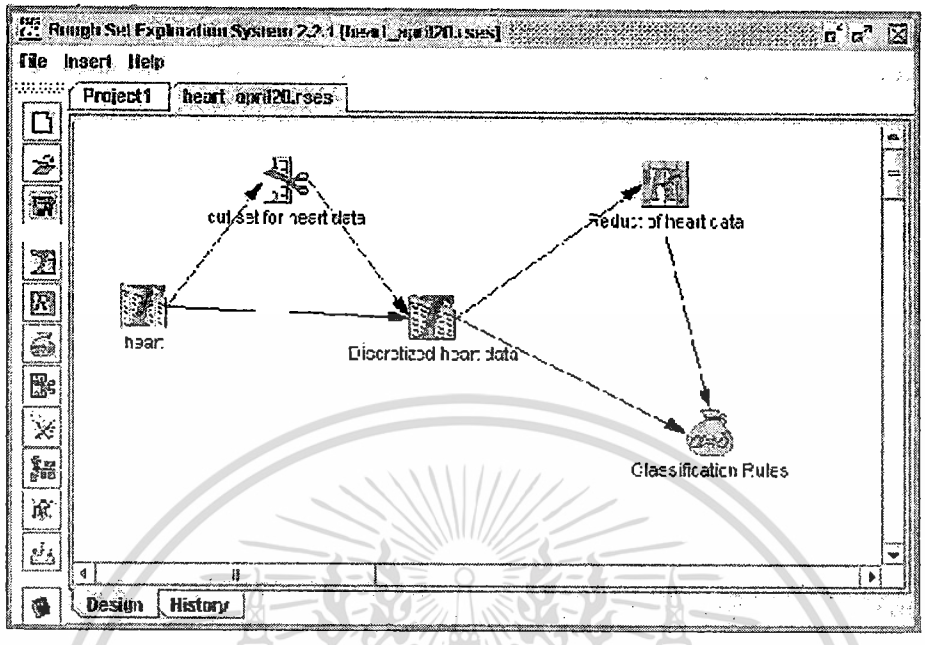


Fig. 8. Using Rough Set Exploration System on heart data

are then generated, classification rules based on the reducts are extracted.

RSES provides four approaches on processing missing attribute values, such as removing data records with missing values, assigning the most common values of the missing attribute within the same decision class and without the same decision class, and considering missing attribute values as a special value of the attribute [64]. These approaches are used during the data preprocessing stage in the system. Although these approaches are fast and can be directly applied in the data, they lack the ability of preserving the semantic meanings of the original data set. Missing values may be assigned, however, the filled values may not be able to fully represent what is missing in the data.

RSES provides rule postprocessing, which are "rule filter", "rule shorten" and "rule generalize". "Rule filter" removes from the rule set rules that do not satisfy certain support. "Rule shorten" shortens the length of the rules according to certain parameters [64]. "Rule generalization" generalizes rules according to a system provided pa-

rameter on the precision level. Although these rule postprocessing approaches provide an easier presentation of all the rule sets, these approaches do not provide ways to evaluate which rules are more interesting, and which rules have higher quality. These functions cannot provide a rank of rules according to a rule's significance to the users.

Enhanced Knowledge Discovery System based on Rough Sets We present a rough set based knowledge discovery system, as shown in Figure 9.

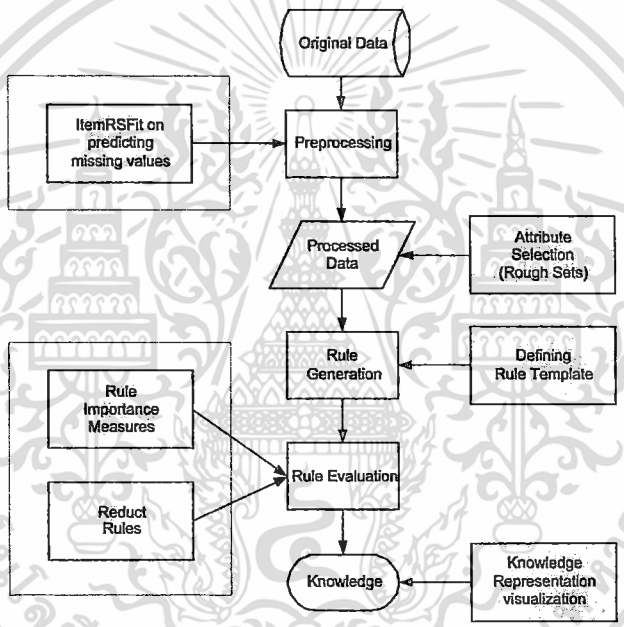


Fig. 9. The Knowledge Discovery Based on Rough Sets Theory

In this general purpose knowledge discovery system, data from different application domains are first imported into the system. Preprocessings including missing attribute values processing, discretization, are conducted in this stage. After the data is preprocessed attribute selections are conducted. Depending on the output, different attribute selection approaches can be applied here. Rule generation algorithms extract rules. After the rule sets are obtained, the impor-

tant postprocessing - rule evaluations are performed in this stage. Rules are finally represented, possibly visualized in a certain format, as knowledge to the end users.

We introduce three approaches integrated into this general purpose KDD system as shown in Figure 9. The first approach *ItemRSFit* is used in the data preprocessing stage. The second approach, *rule importance measure* is used to rank rules during the rule evaluation process. The third approach of extracting *reduct rules* is also used during the rule evaluation stage. We will elaborate these approaches in the following.

I. Predicting missing attribute values based on Frequent Itemset. ItemRSFit approach is a recently developed approach on predicting missing attribute values based on association rules algorithm and rough sets theory. It has been shown on both large scale real world data set and UCI machine learning data sets on the improved prediction accuracies.

ItemRSFit approach is an integration of two other approaches from association rule algorithm and rough sets theory. Priori to the association rule generation, frequent itemsets are generated based on the item-item relations from the large data set according to a certain *support*. Thus the frequent itemsets of a data set represent strong correlations between different items, and the itemsets represent probabilities for one or more items existing together in the current transaction. When considering a certain data set as a transaction data set, the implications from frequent itemsets can be used to find to which attribute value the missing attribute is strongly connected. Thus the frequent itemset can be used for predicting the missing values. We call this approach "itemset-approach" for prediction. The larger the frequent itemsets used for the prediction, the more information from the data set itself will be available for prediction, hence the higher the accuracy will be obtained. However, generating frequent itemset for large data set is time-consuming. Although itemsets with higher support need less computation time, they restrict item-item relationships, therefore not all the missing values can be predicted. In order to balance the tradeoff between computation time and the percentage of the applicable prediction, another approach must be taken into consideration.

A reduct contains a subset of condition attributes that are sufficient enough to represent the whole data set. The intersection of all the possible reduct is the core. Therefore the attributes contained in the reduct or core are more important and representative than the rest of the attributes. Thus by examining only attributes within the same core or reduct to find the similar attribute value pairs for the data instance containing the missing attribute values, we can assign the most relevant value for the missing attribute. Since this method only considers a subset of the data set, which is either the core or the reduct, the prediction is quite fast. This approach "RSFit" is recently proposed in [76], and it is an alternative approach designed for fast prediction. It can be used to predict missing attributes that cannot be predicted by the frequent itemset.

We integrate the prediction based on frequent itemset and RSFit approach into a new approach **ItemRSFit** to predict missing attribute values. Frequent itemsets are used to predict missing values first, and RSFit approach is used to predict the rest of the missing values that cannot be predicted by the frequent itemsets. This integrated approach can predict missing values from the data itself, therefore less noise is brought into the original data. The details on the ItemRSFit approach is presented in [77].

Properly processed data can improve the quality of the generated knowledge. Therefore the ItemRSFit approach is used in this system at the preprocessing stage. It helps to preserve the qualities of the original input data to this system, thus facilitate the rule evaluation process.

II. Rule Importance Measures. Rule importance measure [61] is developed to provide a diverse rank of how important the association rules are, although this approach can also be applied to rules generated by other rule discovery algorithms.

Association rules algorithm can be applied on this transaction data set to generate rules, which have condition attributes on the antecedent part and decision attributes on the consequent part of the rules. Rules generated from different reduct sets can contain different representative information. If only one reduct set is being considered to generate rules, other important information might be omitted. Using multiple reducts, some rules will be generated more

frequently than other rules. We consider the rules that are generated more frequently more important.

The *Rule Importance* is defined to be important by the following definition.

Definition 11. *If a rule is generated more frequently across different rule sets, we say this rule is* more important *than rules generated less frequently across those same rule sets.*

Rule importance measure is defined as follows,

Definition 12.

$$\textit{Rule Importance Measure} = \frac{\textit{Number of times a rule appears in all the generated rules from the reduct sets}}{\textit{Number of reduct sets}}.$$

The definition of the rule importance measure can be elaborated by Eq. 6. Let n be the number of reducts generated from the decision table $T(C, D)$. Let *RuleSets* be the n rule sets generated based on the n reducts. $ruleset_j \in RuleSets$ $(1 \leq j \leq n)$ denotes individual rule sets containing rules generated based on reducts. $rule_i$ $(1 \leq i \leq m)$ denotes the individual rule from *RuleSets*. RIM_i represents the rule importance measure for the individual rule. Thus the rule importance measures can be computed by the following

$$RIM_i = \frac{|\{ruleset_j \in RuleSets | rule_i \in ruleset_j\}|}{n}. \quad (6)$$

The details of how to use rule importance measures can be found in [61]. Rule importance measure can be integrated into the current rough sets based knowledge discovery system to be used during the rule evaluation process. A list of ranked important rules can therefore be presented with their rule importance measures to facilitate the understanding of the extracted knowledge.

III. Extracting Reduct Rules. In [62] a method of discovering and ranking important rules by considering rules as attributes was introduced. The motivation comes from the concept of reduct. A reduct of a decision table contains attributes that can fully represent the original knowledge. If a reduct is given, rules extracted based on this reduct are representative of the original decision table. Can we

take advantage of the concept of a reduct to discover important rules?

We construct a new decision table $A_{m\times(n+1)}$, where each record from the original decision table $u_0, u_1, ..., u_{m-1}$ are the rows, and the columns of this new table consists of $Rule_0, Rule_1, ..., Rule_{n-1}$ and the decision attribute. We say a rule can be applied to a record in the decision table if both the antecedent and the consequent of the rule appear together in the record. For each $Rule_j$ ($j \in [0, ..., n-1]$), we assign 1 to cell $A[i, j]$ ($i \in [0, ..., m-1]$) if the rule $Rule_j$ can be applied to the record u_i. We set 0 to $A[i, j]$ otherwise. The decision attribute $A[i, n]$ ($i \in [0, ..., m-1]$) remains the same as the original values of the decision attribute in the original decision table. Eq. 7 shows the conditions for the value assignments of the new decision table.

$$A[i,j] = \begin{cases} 1, & \text{if } j < n \text{ and } Rule_j \text{ can be applied to } u_i \\ 0, & \text{if } j < n \text{ and } Rule_j \text{ cannot be applied to } u_i \\ d_i, & \text{if } j = n \text{ and } d_i \text{ is the corresponding decision} \\ & attributes for u_i \end{cases} \quad (7)$$

where $i \in [0, ..., m-1], j \in [0, ..., n-1]$.

We further define *Reduct Rule Set* and *Core Rule Set*.

Definition 13. ***Reduct Rule Set.*** *We define a reduct generated from the new decision table A as* ***Reduct Rule Set.*** *A* Reduct Rule Set *contains* Reduct Rules.

The *Reduct Rules* are representative rules that can fully describe the decision attribute.

Definition 14. ***Core Rule Set.*** *We define the intersection of all the* Reduct Rule Sets *generated from this new decision table A as* Core Rule Set. *A* Core Rule Set *contains* ***Core Rules.***

The *Core Rules* are contained in every *Reduct Rule Set.*

By considering rules as attributes, reducts generated from the new decision table contain all the important attributes, which represent the important rules generated from the original data set; and it excludes the less important attributes. Core attributes from the new decision table A contain the most important attributes, which represent the most important rules.

Other Enhancements. The three approaches discussed in our research have shown to effectively evaluate rules. There are other techniques that can be used along with these approaches in Figure 9. For example, during the rule generation process, properly defined rule templates can not only reduce the computation of rule generations, but it also ensures high quality rules, or interesting rules generated according to the application purposes. Important attributes, such as probe attributes can be defined in the data preprocessing stage for generating rules containing such attributes for generating expected rules.

Our motivation is, proposing approaches to enhance the current knowledge discovery system, to facilitate the knowledge discovery process on discovering more interesting and higher quality rules.

10 Experiments

10.1 Distributed Databases via Rough Set Theory

We conduct experiment to extract core attributes of full geriatric data set, female and male patients to show a computational evidence of Proposition 3.

In this study, after removing inconsistent records from the data, the distributed self-reported geriatric data of male and female patients are provided in the Tables 14, 15, respectively.

Table 14. The portion of distributed self-reported geriatric data of female patients.

Patients	EYESI	HEALT	...	AGE	TIME
1	.25	0	...	5	5
2	.25	.5	...	1	2
...	...	...	...	...	...
975	0	.25	...	2	5

Dispensable attributes are excessive attributes in the data set which can be removed, whereas core attributes are an important portion of the data that affect the prediction of the decision, as shown in Table 16.

Table 15. The portion of distributed self-reported geriatric data of male patients.

Patients	EYESI	HEALT	...	AGE	TIME
1	.25	.75	...	3	2
2	0.25	0	...	4	4
...	...	...	...	...	...
890	0.25	0.25	...	4	2

Table 16. Dispensable and core attributes results.

Data	Dispensable attribute	Core attribute
full geriatric data	EARTROU, WALK	EDULEVEL, EYESI, AGE, HEAR, SKIN, SHOWER, PHONE, SHOP, MEAL, MONEY, CHEST, TROUB, COUGH, TIRED, HOUSEW, HEALTH, LIVEALO, SNEEZE, HBP, HEART, STROKE, ARTHRI, EYETROU, DENTAL, STOMAC, KIDNEY, BLADDER, BOWELS, DIABETES, NERVES, FEET, FRACTURE, GENDER
female patients	EYETROU	EYESI, HEAR, DIABETES, AGE, EDULEVEL
male patients	SHOWER, COUGH,	EYESI, CHEST, HEAR, HEALTH, SNEEZE, EYETROU, STROKE, ARTHRI, AGE

As one can see number of core attributes of female and male patients are 5 and 9, respectively, which are less than number core attributes of full geriatric data. Reducts in Table 17 were generated from our CDispro algorithm presented in Part I. They are the reduced number of attributes that can be used to predict decision as well as the entire data set. One can see that dispensable attributes and reducts are different when compare full geriatric data set to distributed data sets and thus satisfy Pawlak's remark [33]. These results also reveal differences between genders which also satisfy the remark in [22].

Table 17. Reduct results of geriatric data.

Data	Reduct
full geriatric data	EDULEVEL, EYESI, HEAR, SHOWER, PHONE, SHOP, MEAL, HOUSEW, MONEY,HEALTH, TROUB, LIVEALO, AGE, COUGH, TIRED, SNEEZE, HBP, HEART, STROKE, KIDNEY, BLADDER, BOWELS, DIABETES, FEET, NERVES, SKIN, FRACTURE, GENDER
female patients	EYESI, HEAR, DIABETES, AGE
male patients	EYESI, HEAR, HEALTH, EDULEVEL, STROKE, ARTHRI, CHEST, AGE

10.2 Experiments with Missing Values and Comparison Studies

We demonstrate, through a series of experiments, that systems improved by the proposed rule evaluation approaches can help humans discover and understand more important rules.

Specifying Rule Templates Apriori association rules algorithm is used to generate rules. Because our interest is to make decisions or recommendations based on the condition attributes, we are looking for rules with only decision attributes on the consequent part. Therefore, we specify the following 2 rule templates to extract rules

we want as shown by Template 8, and to subsume rules as shown by Template 9.

$$\langle Attribute_1, Attribute_2, \ldots, Attribute_n \rangle \rightarrow \langle DecisionAttribute \rangle \tag{8}$$

Template 8 specifies only decision attributes can be on the consequent part of a rule, and $Attribute_1$, $Attribute_2$, ..., $Attribute_n$ lead to a decision of $DecisionAttribute$.

We specify the rules to be removed or subsumed using Template 9. For example, given rule

$$\langle Attribute_1, Attribute_2 \rangle \rightarrow \langle DecisionAttribute \rangle \tag{9}$$

the following rules

$$\langle Attribute_1, Attribute_2, Attribute_3 \rangle \rightarrow \langle DecisionAttribute \rangle \tag{10}$$

$$\langle Attribute_1, Attribute_2, Attribute_6 \rangle \rightarrow \langle DecisionAttribute \rangle \tag{11}$$

can be removed because they are subsumed by Template 9. Take the geriatric care data as an example, in the rule set, a rule shown as Eq. 12 exists

$$SeriousChestProblem \rightarrow Death \tag{12}$$

the following rule is removed because it is subsumed.

$$SeriousChestProblem, TakeMedicineProblem \rightarrow Death \tag{13}$$

Geriatric Care Data Set We perform experiments on a geriatric care data set as described in Part I.

The ItemRSFit approach is implemented by Perl and the experiments are conducted on Sun Fire V880, four 900Mhz UltraSPARC III processors. We use apriori frequent itemset generation [80] to generate frequent 5-itemset. The core generation in RSFit approach is implemented with Perl combining the SQL queries accessing MySQL (version 4.0.12). ROSETTA software [63] is used for reduct generation.

10.3 Experiments on Predicting Missing Attribute Values

In order to show the ItemRSFit approach obtains better prediction accuracy than the existing approach (i.e., RSFit), we perform the experiments on the geriatric care data set by randomly selecting 150 missing values from the original data. We then apply both RSFit approach and ItemRSFit approach on predicting missing values, and compare the accuracy of the prediction. Figure 10 demonstrates the comparison predicting abilities between RSFit and ItemRSFit approaches. We can see from the figure that the smaller the support is, the more accurate the prediction of the missing attribute values for the ItemRSFit approach obtains; whereas for the RSFit approach, the accuracy remains the same as the value of support gets smaller; and the accuracy obtained by RSFit is always lower than the ItemRSFit approach. This result demonstrates that frequent itemsets as knowledge base can be effectively applied for predicting missing attribute values.

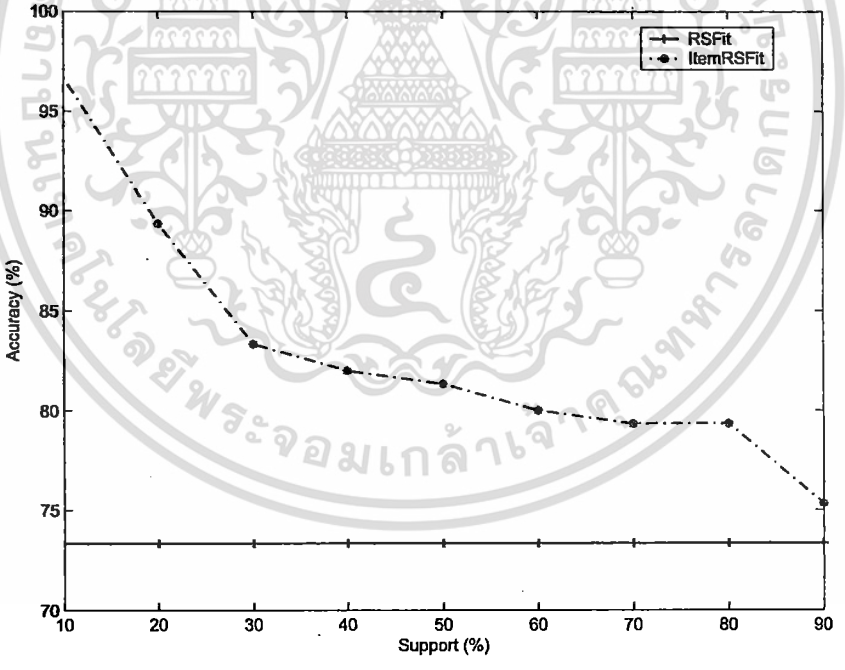


Fig. 10. Accuracy Comparisons for Geriatric Care Data with 150 Missing Attribute Values

10.4 Experiments on Rule Importance Measure

In our experiment, we use the genetic algorithm to generate multiple reduct sets with the option of full discernibility. The apriori algorithm [80] for large item sets generation.

The core attributes for this data set are *eartrouble, livealone, heart, highbloodpressure, eyetrouble, hearing, sex, health, educationlevel, chest, housework, diabetes, dental, studyage.*

Table 18. Reduct Sets for the Geriatric Care Data Set after Preprocessing

No.	Reduct Sets
1	{edulevel,eyesight,hearing,shopping,housewk,health,trouble,livealone, cough,sneeze,hbp,heart,arthriti,eyetroub,eartroub,dental, chest,kidney,diabetes,feet,nerves,skin,studyage,sex}
2	{edulevel,eyesight,hearing,phoneuse,meal,housewk,health,trouble,livealon, cough,sneeze,hbp,heart,arthriti,evetroub,eartroub,dental, chest,bladder,diabetes,feet,nerves,skin,studyage,sex}
...	...
86	{edulevel,eyesight,hearing,shopping,meal,housewk,takemed,health, trouble,livealone,cough,tired,sneeze,hbp,heart,stroke,arthriti, eyetroub,eartroub,dental,chest,stomach,kidney,bladder,diabetes, feet,fracture,studyage,sex}

Table 18 shows selected reduct sets among the 86 reducts generated by ROSETTA. All of these reducts contain the core attributes. For each reduct set, association rules are generated with $support = 30\%, confidence = 80\%$. [2] 218 unique rules are generated over these 86 reducts. These rules as well as their rule importance are shown in Table 19. Among these 218 rules, 87 rules have rule importance of no less than 50% , 8 of which have rule importance of 100%. All the rules with rule importance of 100% contain only core attributes.

10.5 Experiments on Generating Reduct Rules

The new decision table $A_{8535\times219}$ is constructed by using the 218 rules [3] as condition attributes, and the original decision attribute

[2] Note that the value of support and confidence can be adjusted to generate as many or as few rules as required.

[3] There are 1615 rules generated by apriori algorithm from the original data set with $support = 30\%, confidence = 80\%$, after applying the rule template. We can cir-

Table 19. Rule Importance for the Geriatric Care Data

No.	Selected Rules	Rule Importance
0	SeriousHeartProblem → Death	100%
1	SeriousChestProblem → Death	100%
2	SeriousHearingProblem, HavingDiabetes → Death	100%
3	SeriousEarTrouble → Death	100%
4	SeriousEyeTrouble → Death	100%
5	Sex_Female → Death	100%
...	...	...
10	Livealone, HavingDiabetes, NerveProblem → Death	95.35%
...	...	...
216	SeriousHearingProblem, ProblemUsePhone → Death	1.16%
217	TakeMedicineProblem, NerveProblem → Death	1.16%

as the decision attribute. Note that after reconstructing the decision table, we must check for inconsistency again before generating reduct rules for this table. After removing the inconsistent data records, there are 5709 records left in the new decision table. The core rule set is empty. We use Johnson's reduct generation algorithm on this table $A'_{5709\times219}$ and the reduct rule set is $\{Rule_0, Rule_1, Rule_3, Rule_5, Rule_{19}, Rule_{173}\}$. We show these rules in Table 20. From Table 20 we can see that the reduct rule sets con-

Table 20. Reduct Rules for the Geriatric Care Data

No. in Table 19	Reduct Rules	Rule Importance
0	SeriousHeartProblem → Death	100%
1	SeriousChestProblem → Death	100%
3	SeriousEarTrouble → Death	100%
5	Sex_Female → Death	100%
19	Livealon, OftenSneeze, DentalProblems, HavingDiabetes → Death	82.56%
173	ProblemHandleYourOwnMoney → Death	27.91%

tain 6 rules. There are 4 rules judged to be the most important. The rule importance for $Rule_0$, $Rule_1$, $Rule_3$ and $Rule_5$ are all 100%. The $Rule_{19}$ has the importance of 82.56%, which is more important among the 218 rules.

cumvent problems inherent in considering all 1615 generated rules using the 218 unique rules that are derived from the 86 reducts obtained by ROSETTA's genetic algorithm.

10.6 Recommender Systems

In this section, we provide a case study based on a geriatric data set as described in Part I for recommender system applications.

Table 21. Geriatric survival prediction rules database

Decision Rules	Rule Quality	Rule Cover
IF (edlevel!=2 or 4) and (0<shopping≤0.5) and (meal≤0)and (trouble>0) and (livealo>0) and (sneeze≤0) and (hbp≤0) and (eyetrou≤0) and (feet≤0) and (nerves≤0) and (sex>1) THEN (survival time = 7–18 months)	1.969635	13
IF (housew>0) and (cough≤0) and (tired≤0) and (hbp≤0) and (eyetrou≤0) and (kidney>0) and (bowels≤0) and (nerves≤0) and (2<age6≤4)and (sex>1) THEN (survival time = 7–18 months)	1.935979	12
...	...	...
IF (edlevel!=2) and (eyesi≤0) and (health>0) and (trouble≤0) and (sneeze≤0) and (heart≤0) and (arthriti≤0) and (eyetrou>0) and (dental≤0) and (chest≤0) and (kidney≤0) and (bladder≤0) and (feet≤0) and (skin≤0) and (age6≤1) THEN (survival time = 7–18 months)	1.012614	1

Table 22. Fact database of geriatric

Facts	Priority	Facts	Priority	Facts	Priority	Facts	Priority
eyesi < 0.25	0.3	hear < 0.25	0.3	eat = 0	0.1	cough = 0	0.2
tired = 0	0.1	sneeze = 0	0.2	hbp = 0	0.5	heart = 0	1.0
arthriti = 0	1.0	stroke = 0	0.8	parkinso = 0	1.0	eyetrou = 0	0.2
eartrou = 0	0.2	dental = 0	0.4	chest = 0	1.0	stomac = 0	0.9
kidney = 0	0.9	bladder = 0	0.8	bowels = 0	0.8	diabet = 0	1.0
feet = 0	0.1	nerves = 0	0.9	skin = 0	0.6	fracture = 0	0.9
age6 > 3	0.7						

Table 22 shows our fact database. The priority range is $[0, 1]$ (where 1 is the highest priority). Please note that, {age} is included in the fact database but we will not recommend any test for this fact. We then use the facts from Table 22 to calculate the rule priority (Sect. 8.4) and add it to the rules in Table 21. The rules in Table 21 are transformed to decision rules to recommend tests and stored in our knowledge base (Table 23).

For an example of how the rule priority is calculated, take the first rule in Table 21. Its rule priority is equal to $(0.2 + 0.5 + 0.2 + 0.1 + 0.9)/25 = 0.076$ where $n = 25$. The conclusion (action) of the rule are the tests: *(test sneeze), (test hbp), (test eyetrou), (test feet)*

Table 23. Recommendation rules of geriatric in the knowledge base

Recommendation Rules	Rule Quality	Rule Cover	Rule Priority
IF (edlevel!=2 or 4) and (0<shopping≤0.5) and (meal≤0) and (trouble>0) and (livealo>0) and (sneeze≤0) and (hbp≤0) and (eyetrou≤0) and (feet ≤0) and (nerves≤0) and (sex>1) THEN (test sneeze) and (test hbp) and (test eyetrou) and (test feet) and (test nerves)	1.969635	13	0.076
IF (housew>0) and (cough≤0) and (tired≤0) and (hbp≤0) and (eyetrou ≤0) and (kidney>0) and (bowels≤0) and (nerves≤0) and (2<age6≤4) and (sex>1) THEN (test cough) and (test tired) and (test hbp) and (test eyetrou) and (test bowel) and (test nerve)	1.935979	12	0.136
...	...	...	...
IF (edlevel!=2) and (eyesi≤0) and (health>0) and (trouble≤0) and (sneeze ≤0) and (heart≤0) and (arthriti≤0) and (eyetrou>0) and (dental≤0) and (chest≤0) and (kidney≤0) and (bladder≤0) and (feet≤0) and (skin≤0) and (age6≤1) THEN (test eyesi) and (test sneeze) and (test heart) and (test arthriti) and (test dental) and (test chest) and (test kidney) and (test bladder) and (test feet) and (test skin)	1.012614	1	0.252

Table 24. Example input and output.

Example input	Example output
IF (edlevel!=2 or 4) and (0 < shopping < 0.5) and (meal≤0) and (trouble≥0) and (livealo ≥0) and (sneeze≤0) and (hbp≤0) and (eyetrou≤0) and (feet≤0) and (nerves≤0) and (sex>1) THEN (survival time = 7-18 months)	Recommended clinical examinations: sneeze, high blood pressure, eye trouble, feet, nerves

and *(test nerves)* that match the facts in Table 22. When a user inputs their data into our system, if the prediction is a critical case (survival time 7–18 months), our recommendation system will start its analysis. For example, suppose input was as in Table 24. This patient's input matches the first recommendation rule and does not match any other rule. We trigger the first recommendation rule and fire the action of first recommendation rule. The tests recommended to the user are shown in Table 24.

11 Conclusion

In this research, mathematical proofs of Pawlak's rough set theory about core attributes in distributed decision table based on rough sets and relational databases are proposed. We devise a rough sets hybrid approach with statistics and direct use of relational database operations to survival analysis.

A case study on actual self-reported geriatric data for survival analysis is presented. Risk factors, prolongation time prediction rules and validation are performed and discussed. We illustrate that the present approach can be applied to bias within self-reported data efficiently.

We also study the work of rough sets based rule evaluations on knowledge discovery system. We propose solutions to the challenging problems brought by large real world data sets, such as the existence of missing values and analyzing huge amount of generated rules manually. Three rough set based approaches to enhance the current KDD systems on rule evaluations are introduced. The *ItemRSFit* approach is used to predict missing attribute values using frequent itemset as a knowledge base. Complete data can be obtained using this approach. The *rule importance measure* provides a ranking of how important is a rule. Finally, the *reduct rules* are extracted using the concept of reduct by considering rules as condition attributes in a decision table. Experimental results on a real world geriatric care data set demonstrate the utilities of applying rough sets based rule evaluations to enhance current KDD systems.

Finally, we have proposed a health recommendation system architecture using rough sets, survival analysis and rule-based expert systems. Our system was designed with the goal of providing accurate,

low-cost clinical examination recommendations given patients' self reported data. In countries where health care costs are prohibitively expensive, this system can provide a free alternative. Our system generates not only decision rules but also applicable recommendations for patients.

References

1. Larry, M.R.: Hybrid Intelligent System. Kluwer Academic Publishers, Boston (1995).
2. L.T. Elisa, W.W. John, Statistical Methods for Survival Data Analysis, 3rd ed., John Wiley and Sons, New York, 2003.
3. I. Cohen, A. Garg, T.S. Huang, N. Sebe, Machine Learning in Computer Version, Springer, Berlin, 2005.
4. R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, 5th ed., Prentice Hall, New Jersey, 2002.
5. S.J. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, 2nd ed., Prentice Hall, New York, 2002.
6. P. Zdzislaw, Rough sets, Int. J. Inform. Comput. Sc. 11 (5) (1982) 341–356.
7. P. Zdzislaw, Rough Sets. Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Dordrecht, 1991.
8. Z. Pawlak, A. Skowron, Rudiments of rough sets, Inform. Sciences 177 (1) (2007) 3–27.
9. Z. Pawlak, A. Skowron, Rough sets: Some extensions, Inform. Sciences 177 (1) (2007) 28–40.
10. Z. Pawlak, A. Skowron, Rough sets and Boolean reasoning, Inform. Sciences 177 (1) (2007) 41–73.
11. O. Alexander, Discernibility and Rough Sets in Medicine: Tools and Applications, Dissertation, Norwegian University of Science and Technology, Norway, 1999.
12. J. Komorowski, L. Polkowski, A. Skowron, Rough sets: A tutorial, in: S.K. Pal, A. Showorn (Eds.), Rough Fuzzy Hybridization: A New Trend in Decision-Making, Springer-Verlag, Berlin, 1998, pp. 3–98.
13. J.F. Peters et al. (Eds.), Transactions on Rough Sets VI: Journal Subline, Lect. Notes Comp. Sci. 4374 Springer, Heidelberg, 2007.
14. J.F. Peters et al. (Eds.), Transactions on Rough Sets VII: Journal Subline, Lect. Notes Comp. Sci. 4400 Springer, Heidelberg, 2007.
15. P. Pattaraintakorn, N. Cercone, K. Naruedomkul, Hybrid intelligent systems: selecting attributes for soft-computing analysis, in: Proceedings of the 29th Annual International Computer Software and Applications Conference, Edinburgh, Scotland, IEEE Computer Society, 2005, pp. 319–325.
16. P. Pattaraintakorn, N. Cercone, K. Naruedomkul, Selecting attributes for soft–computing analysis in hybrid intelligent systems, in: D. Slezak et al. (Eds.), Lect. Notes. Artif. Int. 3642, Proceedings of the 10th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, Regina, Canada, Springer-Verlag, Berlin, 2005, pp. 698–708.
17. P. Pattaraintakorn, N. Cercone, K. Naruedomkul, Rule learning: ordinal prediction based on rough set and soft-computing, Appl. Math. Lett. 19 (12) (2006) 1300–1307.

18. P. Pattaraintakorn, N. Cercone, Hybrid rough sets-population based system, in: J.F. Peters (Eds.), Transactions on Rough Sets VII: Journal Subline, Lect. Notes Comp. Sci. 4400 Springer, Heidelberg, 2007, pp. 190–205.
19. P. Pattaraintakorn, N. Cercone, K. Naruedomkul, Hybrid rough sets intelligent system architecture for survival analysis, in: J.F. Peters (Eds.), Transactions on Rough Sets VII: Journal Subline, Lect. Notes Comp. Sci. 4400 Springer, Heidelberg, 2007, pp. 206–224.
20. P. Pattaraintakorn, N. Cercone, Integrating rough set theory and medical applications, Appl. Math. Lett. (in press) DOI: 10.1016/j.aml.2007.05.010.
21. J. Bazan, A. Skowron, D. Slezak, J. Wroblewski, Searching for the complex decision reducts: The case study of the survival analysis, in: N. Zhong et al. (Eds.), Lect. Notes. Artif. Int. 2871, Proceedings of the International Syposium on Methodologies for Intelligent Systems, Maebashi City, Japan, Springer-Verlag, Berlin, 2003, pp. 160–168.
22. X. Song, A. Mitnitski, C. MacKnight, K. Rockwood, Assessment of individual risk of death using self-report data: An artificial neural network compared to a frailty index, J. Am. Geriatr. Soc. 52 (2004) 1180–1184.
23. X. Song, A. Mitnitski, C. MacKnight, K. Rockwood, Assessment of individual risk of death using self-report data: An artificial neural network compared to a frailty index, J. Am. Geriatr. Soc., Vol. 52, pp. 11801184, 2004.
24. Pattaraintakorn, P., Cercone, N., Naruedomkul, K.: Hybrid intelligent systems: selecting attributes for soft-computing analysis. In: Proceedings of the 29th Annual International Computer Software and Applications Conference (COMPSAC2005), Edinburgh, Scotland, UK, vol. 1, IEEE Computer Society (2005) 319–325.
25. Pattaraintakorn, P., Cercone, N., Naruedomkul, K.: Selecting attributes for soft-computing analysis in hybrid intelligent systems. In: Slezak, D., Yao, J.T., Peters, J.F., Ziarko, W., Hu, X. (Eds.), Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing (RSFDGrC2005), Regina, Canada, Part II, Lecture Notes in Artificial Intelligence, vol. 3642. Springer-Verlag, Heidelberg (2005) 698–708.
26. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 53 (1958) 457–481.
27. Cox, D.R.: The analysis of exponentially distributed life-times with two types of failure. Journal of the Royal Statistical Society 21 (1959) 411–412.
28. Peto, R., Peto, J.: Asymptotically efficient rank invariant procedures. Journal of the Royal Statistical Society 135 (1972) 185–207.
29. Gehan, E.A.: A Generalized Wilcoxon test for comparing arbitrarily singly-censored data. Biometrika 52 (1965) 203–223.
30. Tarone, R.E., Ware, J.: On distribution-free tests for equality of survival distributions. Biometrika 64 (1977) 156–160.
31. Bazan, J., Osmolski, A., Skowron, A., Slezak, D., Szczuka, M., Wroblewski, J.: Rough set approach to the survival analysis. In: Bazan, J., Osmolski, A., Skowron, A., Slezak, D., Szczuka, M., Wroblewski, J. (Eds.), Rough Sets and Current Trends in Computing (RSCTC 2002), Malvern, PA, USA, Lecture Notes in Artificial Intelligence, vol. 2475, Springer-Verlag, Berlin, Heidelberg (2002) 522–529.
32. Bazan, J., Skowron, A., Slezak, D., Wroblewski, J.: Searching for the complex decision reducts: the case study of the survival analysis. In: Bazan, J., Skowron, A., Slezak, D., Wroblewski, J. (Eds.), Foundations of Intelligent Systems, Lecture Notes in Artificial Intelligence, vol. 2871, Springer-Verlag, Berlin, Heidelberg (2003) 160–168.

33. Pawlak, Z.: Rough sets. Theoretical aspects of reasoning about data. Kluwer Academic Publishers, Dordrecht (1991).
34. Skowron, A., Rauszer, C.: The discernibility matrices and functions in information systems. In: Slowinski, R. (Ed.), Intelligent Decision Suppport - Handbook of Applications and Advances of the Rough Sets Theory. Dordrecht, Kluwer (1992) 331–362.
35. Nguyen, S.H., Nguyen, H.S.: Some efficient algorithms for rough set methods. In: Proceedings of the Sixth International Conference on Information Processing and Management of Uncertainty Knowledge Based Systems (IPMU1996), Granada, Spain, vol. 3, (1996) 1451–1456.
36. Nguyen, H.S.: Approximate Boolean reasoning approach to rough sets and data mining. In: Slezak, D., Yao, J.T., Peters, J.F., Ziarko, W., Hu, X. (Eds.), Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing (RSFDGrC2005), Regina, Canada, Part II, Lecture Notes in Artificial Intelligence, vol. 3642, Springer-Verlag, Heidelberg (2005) 12–22.
37. Wroblewski, J.: Theoretical foundations of order-based genetic algorithms. Fundamenta Informaticae 28 (1996) 423–430.
38. Bazan, J., Nguyen, H.S., Nguyen, S.H., Synak, P., Wroblewski, J.: Rough set algorithms in classification problems. In: Polkowski, L., Lin, T.Y., Tsumoto, S. (Eds.), Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems, Studies in Fuzziness and Soft Computing, vol. 56, Springer-Verlag/Physica-Verlag, Heidelberg (2000) 49–88.
39. Hu, X., Han, J., Lin, T.Y.: A new rough sets models based on database systems. Fundamenta Informaticae 59(2-3) (2004) 1–18.
40. D. R. Cox: The Analysis of Exponentially Distributed Life-times with Two Types of Failure, J. of the Royal Statistical Society, vol. 21, 1959, 411–422.
41. E. L. Kaplan, P. Meier: Nonparametric Estimation from Incomplete Observations, J. of the Amer. Stat. Asso., vol. 53, 457–481, 1958.
42. A. Kusiak, B. Dixon, S. Shah: Predicting Survival Time for kidney Dialysis Patients: A Data Mining Approach, Computers in Biology and Medicine 35, 2005, 311–327.
43. S.H. Nguyen, H.S., Nguyen, Some efficient algorithms for rough set methods, in: Proceedings of the Sixth International Conference on Information Processing and Management of Uncertainty Knowledge Based Systems, Granada, Spain, 1996, pp. 1451–1456.
44. X. Hu, T.Y. Lin, J. Han, A new rough sets models based on database systems, Fund. Inform. 59 (2–3) (2004) 1–18.
45. A. An, N. Cercone, ELEM2: A learning system for more accurate classifications, in: R.E. Mercer, E. Neufeld (Eds.), Advances in Artificial Intelligence, Lect. Notes. Artif. Int. 1418, Proceedings of the 12th Biennial Conference of the Canadian Society for Computational Studies of Intelligence on Advances in Artificial Intelligence, Vancouver, BC, Canada, Springer-Verlag, Berlin, 1998, pp. 426–441.
46. J.R. Quinlan, Induction of decision trees, *Mach. Learn.* 1 (1986) 81-106.
47. I. Witten, E. Frank, Practical machine learning tools and techniques with JAVA implementations, Morgan Kaufmann, San Francisco, 2000.
48. R. Swiniarski, A. Skowron, Rough set methods in feature selection and recognition, *Pattern Recogn. Lett.* 24 (6) (2003) 833–849.
49. X. Hu, N. Cercone, Discovery of decision rules in relational databases: a rough set approach, Proceedings of the International Conference on Information and Knowledge Management, ACM (1994) 392–400.

50. A. Skowron, C. Rauszer, The discernibility matrices and functions in information systems, in: R. Slowinski (Ed.), Intelligent Decision Suppport - Handbook of Applications and Advances of the Rough Sets Theory, Dordrecht, Kluwer, 1992, pp. 331–362.
51. H.S. Nguyen, Approximate Boolean reasoning approach to rough sets and data mining, in: D. Slezak et al. (Eds.), Lect. Notes. Artif. Int. 3642, Proceedings of the 10th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, Regina, Canada, Springer-Verlag, Berlin, 2005, pp. 12–22.
52. C. Blake, E. Keogh, C. Merz, *UCI Repository of Machine Learning Databases*, http://www.ics.uci.edu, University of California, 2005.
53. P. Pattaraintakorn, N. Cercone, K. Naruedomkul: Hybrid Intelligent Systems: Selecting Attributes for Soft-Computing Analysis, in Proc. of COMPSAC, 2005, 319–325.
54. P. Pattaraintakorn, N. Cercone, K. Naruedomkul: Selecting Attributes for Soft-Computing Analysis in Hybrid Intelligent Systems, in Lect. Notes. Artif. Int., vol. 3642, D. Slezak et al. Eds. Springer-Verlag, Berlin, Heidelberg, 2005, 698–708.
55. An, A., Cercone, N.: ELEM2: a learning system for more accurate classifications. In: Proceedings of the 12th Biennial Conference of the Canadian Society for Computational Studies of Intelligence on Advances in Artificial Intelligence, Lecture Notes in Computer Science, vol. 1418, Springer-Verlag, Heidelberg (1998) 426–441.
56. Pattaraintakorn, P., Cercone, N., Naruedomkul, K.: Rule analysis with rough sets theory. In: Proceedings of the IEEE International Conference on Granular Computing (IEEEGrC2006), Atlanta, USA, IEEE Computer Society (2006) 582–585.
57. Pattaraintakorn, P., Cercone, N., Naruedomkul, K.: Rule learning: ordinal prediction based on rough set and soft–computing. Applied Mathematics Letters 19 (2006) 1300–1307.
58. Quinlan, J.R.: Induction of decision trees. Machine Learning 1 (1986) 81-106.
59. Z. Pawlak, A treatise on rough sets, in J.F. Peters et al. (Eds.) *Transactions on rough sets IV, LNCS 3700*, Berlin: Springer-Verlag, 2005, pp. 1–17.
60. Z. Pawlak, "Rough set theory for intelligent industrial applications," in *Proc. 2nd Intelligent Processing and Manufacturing of Materials, Vol. 1*, Hawaii, USA, 1999, pp. 37–44.
61. Li, J. and Cercone, N.: Introducing A Rule Importance Measure. Transactions on Rough Sets, Springer LNCS, vol 5 (2006)
62. Li, J., Cercone, N.: Discovering and Ranking Important Rules. In Proceedings of IEEE International Conference on Granular Computing, vol 2, Beijing China 25-27 July (2005) 506–511
63. Øhrn, A.: Discernibility and Rough Sets in Medicine: Tools and Applications. PhD Thesis, Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Norway. (1999)
64. RSES 2.2 User's Guide. Warsaw University. http://logic.mimuw.edu.pl/~rses/
65. Predki, B., Wilk, Sz.: Rough Set Based Data Exploration Using ROSE System. In: Foundations of Intelligent Systems. Ras, Z. W., Skowron, A., Eds, LNAI 1609, Springer-Verlag, Berlin (1999) 172-180
66. Chouchoulas, A. and Shen, Q.: Rough Set-Aided Keyword Reduction For Text Categorization. Applied Artificial Intelligence, vol 15 (2001) 843–873
67. Hu, X., Lin, T., Han, J.: A New Rough Sets Model Based on Database Systems. Fundamenta Informaticae 59 no.2-3 (2004) 135–152

68. Freeman, R. L., Grzymala-Busse, J. W., Riffel, L. A., Schroeder, S. R.: Analyzing the Relation Between Heart Rate, Problem Behavior, and Environmental Events Using Data Mining System LERS. In 14th IEEE Symposium on Computer-Based Medical Systems (CBMS'01) (2001)
69. Ivo, D., Gunther, G.: The Rough Set Engine GROBIAN. In Proc. of the 15th IMACS World Congress, vol 4, Berlin, August (1997)
70. Hu, T., Shan, N., Cercone, N. and Ziarko, W.: DBROUGH: A Rough Set Based Knowledge Discovery System, Proc. of the 8th International Symposium on Methodologies for Intelligent System, LNAI **869**, Spring Verlag (1994) 386–395
71. Hilderman, R. and Hamilton, H.: Knowledge discovery and interestingness measures: A survey. Technical Report 99-04, Department of Computer Science, University of Regina, October (1999)
72. Pang-Ning Tan and Vipin Kumar and Jaideep Srivastava: Selecting the right interestingness measure for association patterns. Processings of SIGKDD. (2002) 32–41
73. Bruha, Ivan: Quality of Decision Rules: Definitions and Classification Schemes for Multiple Rules. In Machine Learning and Statistics, The Interface, Edited by G. Nakh aeizadeh and C. C. Taylor. John Wiley & Sons, Inc. (1997) 107–131
74. An, A. and Cercone, N.: ELEM2: A Learning System for More Accurate Classifications. In: Proceedings of Canadian Conference on AI (1998) 426–441
75. X. Hu, J. Han, T.Y. Lin, "A new rough sets models based on database systems," *Fund. Inform.*, Vol. 59(2–3), pp. 1–18, 2004.
76. Li, J. and Cercone, N.: Assigning Missing Attribute Values Based on Rough Sets Theory. In Proceedings of IEEE Granular Computing, Atlanta, USA. (2006)
77. Li, J. and Cercone, N.: Predicting Missing Attribute Values based on Frequent Itemset and RSFit. Technical Report, CS-2006-13, School of Computer Science, University of Waterloo (2006)
78. Li, J. and Cercone, N.: Empirical Analysis on the Geriatric Care Data Set Using Rough Sets Theory. Technical Report, CS-2005-05, School of Computer Science, University of Waterloo (2005)
79. An, A. and Cercone, N.: Rule Quality Measures for Rule Induction Systems: Description and Evaluation. Computational Intelligence. **17-3** (2001) 409–424.
80. Borgelt, C.: Efficient Implementations of Apriori and Eclat. Proceedings of the FIMI'03 Workshop on Frequent Itemset Mining Implementations. In: CEUR Workshop Proceedings (2003) 1613-0073 http://CEUR-WS.org/Vol-90/borgelt.pdf
81. Blatter, M., Zhang, Y., Maslow, S.: Exploring an Opinion Network for Taste Prediction: An Empirical Study. Physica A **373** (2007) 753–758
82. Liu, D., Shih, Y.: Hybrid Approaches to Product Recommendation Based on Customer Lifetime Value and Purchase Preferences. J. Syst. Software **77** (2005) 181–191
83. Coppin, B.: Artificial Intelligence Illuminated. Jones and Bartlett Publishers, Inc., Sudbury, Mass (2004)