International Joint Conference on Rough Sets

Book of Abstracts

October 5-8, 2023 / Krakow, Poland

AGH University of Krakow

https://ijcrs2023.agh.edu.pl/

Book of Abstracts

International Joint Conference on Rough Sets

Kraków, Poland

2023

https://ijcrs2023.agh.edu.pl

International Joint Conference on Rough Sets

International Joint Conference on Rough Sets (IJCRS 2023)

Book of Abstracts

Editors

Andrea Campagner, IRCCS Istituto Ortopedico Galeazzi, Italy.

Oliver Urs Lenz, Ghent University, Belgium.

Shuyin Xia, Chongqing University of Posts and Telecommunications, China.

Dominik Ślęzak, University of Warsaw, Poland.

Jarosław Wąs, AGH University of Krakow, Poland.

JingTao Yao, University of Regina, Canada.

TEXnical Editing

Marcin Piekarczyk, AGH University of Krakow, Poland.

Tomasz Hachaj, AGH University of Krakow, Poland.

Front Cover Design

Tomasz Hachaj, AGH University of Krakow, Poland.

Contents

Welcome

This volume contains the extended abstracts selected for presentation at IJCRS 2023, the 2023 International Joint Conference on Rough Sets, held at AGH University of Kraków on October 5-8, 2023, in Kraków, Poland. Conferences in the IJCRS series, resulting from the merger of four separate conferences tying rough sets to various paradigms (RSCTC, data analysis; RSFDGrC, granular computing; RSKT, knowledge technology; and RSEISP, intelligent systems), are held annually: the first Joint Rough Set Symposium was held in Toronto, Canada, in 2007; followed by Symposiums in Chengdu, China in 2012; Halifax, Canada, 2013; Granada and Madrid, Spain, 2014; Tianjin, China, 2015, where the acronym IJCRS was proposed; continuing with the IJCRS 2016 conference in Santiago de Chile, IJCRS 2017 in Olsztyn, Poland, IJCRS 2018 in Quy Nhon, Vietnam, IJCRS 2019 in Debrecen, Hungary, IJCRS 2020 in La Habana, Cuba (held online), IJCRS 2021 in Bratislava, Slovakia (hybrid), and IJCRS 2022 in Suzhou, China (hybrid).

Following the success of the previous conferences, IJCRS 2023 continued the tradition of a very rigorous reviewing process. We would like to thank all the authors for contributing their papers. Without their contribution, this conference would not have been possible.

The IJCRS 2023 program was further enriched by eight Keynote Speeches, among them the one presented by Tsau Young Lin, the Founding President of the International Rough Set Society (IRSS), and the Anniversary Talk by Andrzej Skowron, IRSS Fellow and former President, who celebrated his 80th birthday during the conference. We are grateful to our Keynote Speakers, Weronika Adrian, Joel Holland, Andrzej Janusz, Tianrui Li, Tsau Young Lin, Pradipta Maji, Sheela Ramanna, and Andrzej Skowron.

The IJCRS 2023 program also hosted the Special Sessions on “Innovative Foundational Models for Rough Sets, Approximate Reasoning, and Granular Computing” and “Data Analytics in Cybersecurity and IoT Applications” , as well as the Panel on “Intelligent Informatics”. We are grateful to the Special Session Organizers Stefania Boffa, A Mani, Marcin Michalak, and Piotr Synak, to the Panelists Jimmy Huang, Duoqian Miao and Hung Son Nguyen, as well as to the Panel Moderator Pawan Lingras.

IJCRS 2023 would not have been successful without the support of many people and organizations. We are indebted to the Program Committee Members and external reviewers for their effort and engagement in providing a rich and rigorous scientific program. We greatly appreciate the co-operation, support, and sponsorship of various institutions, companies, and organizations, including the AGH University of Kraków, the Strategic Partners QED Software and DeepSeas, Honorary Patronage of the Polish Ministry of Science and Higher Education and of the Mayor of Kraków„ as well as the International Rough Set Society. We acknowledge the use of the Springer EquinOCS conference system for paper submission and review. We are also grateful to Springer for publishing the proceedings as a volume of LNCS/LNAI.

Last but not least, we would like to thank Anna Smyk, Tomasz Hachaj and the whole technical organization team at the AGH University of Kraków, for their great support and endless hours spent on the conference preparations.

Committees

Honorary Chairs

Andrzej Skowron, Systems Research Institute Polish Academy of Sciences, Poland.

Tomasz Szmuc, AGH University of Krakow, Poland.

Yiyu Yao, University of Regina, Canada.

General Chairs

Dominik Ślęzak, University of Warsaw, Poland.

Jarosław Wąs, AGH University of Krakow, Poland.

JingTao Yao, University of Regina, Canada.

Program Committee Chairs

Andrea Campagner, IRCCS Istituto Ortopedico Galeazzi, Italy.

Oliver Urs Lenz, Ghent University, Belgium.

Shuyin Xia, Chongqing University of Posts and Telecommunications, China.

Local Organizing Chairs

Tomasz Hachaj, AGH University of Krakow, Poland.

Soma Dutta, University of Warmia and Mazury in Olsztyn, Poland.

Marcin Piekarczyk, AGH University of Krakow, Poland.

Łukasz Rauch, AGH University of Krakow, Poland.

Łukasz Sosnowski, Systems Research Institute Polish Academy of Sciences, Poland.

Beata Zielosko, University of Silesia in Katowice, Poland.

Program Committee Members

Qiusheng An, Shanxi Normal University, China.

Piotr Artiemjew, University of Warmia and Mazury in Olsztyn, Poland.

Nouman Azam, National University of Computer and Emerging Sciences, Pakistan.

Mohua Banerjee, Indian Institute of Technology, India.

Jan Bazan, University of Rzeszow, Poland.

Urszula Bentkowska, University of Rzeszow, Poland.

Stefania Boffa, University of Milano-Bicocca, Italia.

Henri Bollaert, Ghent University, Belgium.

Joaquin Borrego-Diaz, Universidad de Sevilla, Spain.

Andrea Campagner, IRCCS Istituto Ortopedico Galeazzi, Italy.

Yuming Chen, Xiamen University of Technology, China.

Hongmei Chen, Southwest Jiaotong University, China.

Zehua Chen, Taiyuan University of Technology, China.

Davide Ciucci, University of Milano-Bicocca, Italy.

Chris Cornelis, Ghent University, Belgium.

Jianhua Dai, Hunan Normal University, China.

Tingquan Deng, Harbin Engineering University, China.

Thierry Denoeux, Universite de Technologie de Compiegne, France.

Murat Diker, Hacettepe University, Turkey.

Shifei Ding, China University of Mining and Technology, China.

Barbara Dunin-Képlicz, University of Warsaw, Poland.

Soma Dutta, University of Warmia and Mazury in Olsztyn, Poland.

Hamido Fujita, Iwate Prefectural University, Japan.

Can Gao, Shenzhen University, China.

Yang Gao, Nanjing University, China.

Anna Gomolinska, University of Białystok, Poland.

Salvatore Greco, University of Catania, Italia.

Jerzy Grzymala-Busse, University of Kansas, USA.

Shen-Ming Gu, Zhejiang Ocean University, China.

Christopher Henry, University of Winnipeg, Canada.

Mengjun Hu, University of Regina, Canada.

Qinghua Hu, Tianjin University, China.

Xuegang Hu, Hefei University of Technology, China.

Bing Huang, Nanjing Audit University, China.

Amir Hussain, Edinburgh Napier University, UK.

Masahiro Inuiguchi, Osaka University, Japan.

Ryszard Janicki, McMaster University, Canada.

Andrzej Janusz, University of Warsaw, Poland.

Richard Jensen, Aberystwyth University, United Kingdom.

Xiuyi Jia, Nanjing University of Science and Technology, China.

Chunmao Jiang, Fujian University of Technology, China.

Bin Jie, Hebei Normal University, China.

Olha Kaminska, Ghent University, Belgium.

Aquil Khan, Indian Institute of Technology, Indore.

Marzena Kryszkiewicz, Warsaw University of Technology, Poland.

Mihir Kumar Chakarborty, Jadavpur University, India.

Guangming Lang, Changsha University of Science and Technology, China.

Oliver Urs Lenz, Ghent University, Belgium.

Mikołaj Leszczuk, AGH University of Kraków, Poland.

Deyu Li, Shanxi University, China.

Fanchang Li, Soochow University, China.

Huaxiong Li, Nanjing University, China.

Jinhai Li, Kunming University of Science and Technology, China.

Jinjin Li, Minnan Normal University, China.

Kewen Li, China University of Petroleum, China.

Lei-Jun Li, Hebei Normal University, China.

Tianrui Li, Southwest Jiaotong University, China.

Tong-Jun Li, Zhejiang Ocean University, China.

Yuefeng Li, Queensland University of Technology, Australia.

Jiuzhen Liang, Changzhou University, China.

Shujiao Liao, Minnan Normal University, China.

Guoping Lin, Minnan Normal University, China.

Yaojin Lin, Hefei University of Technology, China.

Pawan Lingras, St. Mary's University, Canada.

Baoxiang Liu, North China University of Science and Technology, China.

Caihui Liu, Gannan Normal University, China.

Dun Liu, Southwest Jiaotong University, China.

Guilong Liu, Beijing Language and Culture University, China.

Wenqi Liu, Kunming University of Science and Technology, China.

Pradipta Maji, Indian Statistical Institute, India.

A Mani, Indian Statistical Institute, India.

Jesus Medina, University of Cadiz, Mathematics.

Jusheng Mi, Hebei Normal University, China.

Duoqian Miao, Tongji University, China.

Marcin Michalak, Łukasiewicz Research Network - Institute of Innovative Technologies EMAG, Poland.

Fan Min, Southwest Petroleum University, China.

Mikhail Moshkov, King Abdullah University of Science and Technology, Saudi Arabia.

Hung Son Nguyen, University of Warsaw, Poland.

Sinh Hoa Nguye, Polish-Japanese Academy of Information Technology, Poland.

Agnieszka Nowak-Brzezińska, University of Silesia in Katowice, Poland.

Krzysztof Pancerz, The John Paul II Catholic University of Lublin, Poland.

Witold Pedrycz, University of Alberta, Canada.

Shenglei Pei, Qinghai Nationalities University, China.

Daniel Peralta, Ghent University, Belgium.

Georg Peters, Munich University of Applied Sciences, Germany.

James Peters, University of Manitoba, Canada.

Lech Polkowski, Polish-Japanese Institute of Information Technology, Poland.

Jianjun Qi, Xidian University, China.

Jin Qian, Jiangsu University of Technology, China.

Yuhua Qian, Shanxi University, China.

Taorong Qiu, Nanchang University, China.

Sheela Ramanna, The University of Winnipeg, Canada.

Zbigniew Ras, The University of North Carolina at Charlotte, USA.

Sergio Ribeiro, Pontificia Universidade Catolica do Parana, Brazil.

Marek Reformat, University of Alberta, Canada.

Henryk Rybiński, Warsaw University of Technology, Poland.

Hiroshi Sakai, Kyushu Institute of Technology, Japan.

Lin Shang, Nanjing University, China.

Ming-Wen Shao, Chinese University of Petroleum, China.

Yanhong She, Xian Shiyou University, China.

Marek Sikora, Silesian University of Technology, Poland.

Andrzej Skowron, Systems Research Institute of Polish Academy of Sciences, Poland.

Dominik Ślęzak, University of Warsaw, Poland.

Roman Słowinski, Poznan University of Technology, Poland.

Jingjing Song, Macau University of Science and Technology, China.

Łukasz Sosnowski, Systems Research Institute of Polish Academy of Sciences, Poland.

Urszula Stańczyk, Silesian University of Technology, Poland.

Jaroslaw Stepaniuk, Bialystok University of Technology, Poland.

Lin Sun, Henan Normal University, China.

Zbigniew Suraj, University of Rzeszów, Poland.

Piotr Synak, DeepSeas, Switzerland.

Marcin Szczuka, University of Warsaw, Poland.

Marcin Szeląg, Poznań University of Technology, Poland.

Anhui Tan, Zhejiang Ocean University, China.

Adnan Theerens, Ghent University, Belgium.

Shusaku Tsumoto, Shimane University, Japan.

Alicja Wakulicz-Deja, University of Silesia in Katowice, Poland.

Renxia Wan, North Minzu University, China.

Baoli Wang, Yuncheng University, China.

Changzhong Wang, Bohai University, China.

Guoyin Wang, Chongqing University of Posts and Telecommunications, China.

Piotr Wasilewski, Systems Research Institute of Polish Academy of Sciences, Poland.

Lai Wei, Tongji University, China.

Ling Wei, Northwest University, China.

Wei Wei, Shanxi University, China.

Zhihua Wei, Tongji University, China.

Marcin Wolski, Maria Curie-Skłodowska University, Poland.

Wei-Zhi Wu, Zhejiang Ocean University, China.

Shuyin Xia, Chongqing University of Posts and Telecommunications, China.

Bin Xie, Hebei Normal University, China.

Jun Xie, Taiyuan University of Technology, China.

Jianfeng Xu, Nanchang University, China.

Jiucheng Xu, Henan Normal University, China.

Weihua Xu, Southwest University, China.

Zhan-Ao Xue, Henan Normal University, China.

Lin Xun, Henan Normal University, China.

Hailong Yang, Shaanxi Normal University, China.

Jilin Yang, Sichuan Normal University, China.

Tian Yang, Hunan Normal University, China.

Xibei Yang, Jiangsu University of Science and Technology, China.

Xin Yang, Southwestern University of Finance and Economics, China.

JingTao Yao, University of Regina, Computer Science.

Yiyu Yao, University of Regina, Canada.

Dongyi Ye, Fuzhou University, China.

Hong Yu, Chongqing University of Posts and Telecommunications, China.

Ying Yu, East China Jiaotong University, China.

Xiaodong Yue, Shanghai University, China.

Zhang Zehua, Taiyuan University of Technology, China.

Yanhui Zhai, Shanxi University, China.

Jianming Zhan, Hubei University for Nationalities, China.

Hongying Zhang, Xi'an Jiaotong University, China.

Hongyun Zhang, Tongji University, China.

Li Zhang, Soochow University, China.

Nan Zhang, Yantai University, China.

Qinghua Zhang, Chongqing University of Posts and Telecommunications, China.

Xianyong Zhang, Sichuan Normal University, China.

Xiaohong Zhang, Shaanxi University of Science and Technology, China.

Yan Zhang, Cal State University, San Bernardino.

Yanping Zhang, Anhui University, China.

Zehua Zhang, Taiyuan University of Technology, China.

Shu Zhao, Anhui University, China.

Huilai Zhi, Henan Polytechnic University, China.

Caiming Zhong, Ningbo University, China.

Bing Zhou, Sam Houston State University, USA.

Jie Zhou, Shenzhen University, China.

Xianzhong Zhou, Nanjing University, China.

Beata Zielosko, University of Silesia in Katowice, Poland.

Li Zou, Liaoning Normal University, China.

External Reviewers

Błażej Adamczyk, EFIGO, Poland.

Seiki Akama, C-Republic, Japan.

Tareq Alshami, Sana'a University, Yemen.

Janusz Borkowski, DeepSeas, Poland.

Duarte Folgado, Fraunhofer AICOS, Portugal.

Brunella Gerla, University of Insubria, Italy.

Arun Kumar, Indian Institute of Technology Kanpur, India.

Nicolas Madrid, University of Malaga, Spain.

Krishna Balajirao Manoorkar, Vrije Universiteit Amsterdam, Netherlands.

Petra Murinova, University of Ostrava, Czech Republic.

Yotaro Nakayama, BIPROGY, Japan.

Agnieszka Nowak-Brzezińska, University of Silesia, Poland.

Eloisa Ramirez Poussa, Universidad de Cadiz, Spain.

Binbin Sang, Southwest Jiaotong University, China.

Anand Pratap Singh, Indian Institute of Information Technology, India.

Apostolos Tzimoulis, Vrije Universiteit Amsterdam, Netherlands.

Łukasz Wawrowski, Łukasiewicz Research Network - Institute of Innovative Technologies EMAG, Poland.

Jakub Wróblewski, DeepSeas, Poland.

Hao Wu, Sun Yat-Sen University, China.

Abstracts

A Python Toolkit for Information Systems over Ontological Graphs

Krzysztof Pancerz[1,2]

[1]Institute of Philosophy, The John Paul II Catholic University of Lublin, Poland

[2]MakoLab S.A., Lodz, Poland

Entended Abstract. We present the designed and implemented Python toolkit for dealing with information systems (understood as the Pawlak's knowledge representation systems) over ontological graphs. In such systems, the domain knowledge in a form of ontologies is enclosed. In this way, some valuable knowledge (especially in terms of semantic relations) can be added to knowledge discovery processes. In the first stage, the particular attention was focused on pre-processing methods as well as methods for determining indiscernibility relations and discernibility matrices, which are key notions in rough set approaches.

Introduction and Motivation

Our research concerns information systems over ontological graphs, where information systems are understood as Pawlak's knowledge representation systems [1]. Nowadays, one of the intensively developed trends of knowledge discovery is to process natural language statements (words, concepts, etc.). The main idea is that words or concepts are used in place of numbers, because the ability of a human is to perform many tasks without any number processing. To effectively carry out knowledge discovery processes, we have to consider the domain knowledge related to data semantics [2]. This knowledge can be represented by an ontology which describes a set of concepts together with the relationships which have been defined between them comprising the vocabulary from a given area (cf. [3]). In [4], we proposed to enclose the domain knowledge in a form of ontology in information systems. In this approach, attribute values are concepts that describe objects. It was assumed that the ontology is presented, in a simplified way, by means of the graph structure, called the ontological graph, in which, each node represents one concept from the ontology, whereas each edge represents a semantic relation between two concepts. Formally, a simple information system over ontological graphs consists of the nonempty, finite set U of objects, the nonempty, finite set A of attributes, the family of ontological graphs associated with attributes from A, the information function assigning concepts from ontological graphs associated with attributes to objects from U.

Main Features

In the developed and implemented Python toolkit for information systems over ontological graphs, an ontological graph in the form of the OWL ontology can be associated to each attribute in the underlying information system IS. An OWL ontology consists of three components: classes representing concepts, individuals being instances of classes, properties being binary relations on individuals [5]. Therefore, attribute values in IS can be either classes (concepts) or individuals. In the toolkit, we have used the following packages to process underlying data structures: *pandas* that delivers a data frame structure to represent an underlying information system IS in a tabular form, and *owlready2* that is used to process OWL structures. For information systems over ontological graphs, we can perform, among others, the following pre-processing procedures: *deinstantiation* replacing individuals being attribute values with the classes (concepts) whose instances they are, and *generalization* replacing classes (concepts) with more general classes (concepts). One can see that, in case of information systems over ontological graphs, we deal with more complex structures of sets of attribute values, i.e, ontological graphs, associated with attributes, delivering attribute values (individuals or classes). Therefore, key notions in rough set approaches (for example, indiscernibility relation [1], discernibility matrix [6]) can be defined in more complex way. In the presented toolkit, we have implemented methods of determining indiscernibility relations and discernibility matrices with respect to different levels of abstraction (related to individuals and hierarchies of classes in OWL ontologies). The presented toolkit will be a part of a larger Python module, called *OnDriML*, for ontologically-driven machine learning (cf. [7]).

References:

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

[2] Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann (2005).

[3] Neches, R., Fikes, R., Finin, T., Gruber, T., Patil, R., Senator, T., Swartout, W.: Enabling technology for knowledge sharing. AI Magazine **12**(3), 36–56 (1991).

[4] Pancerz, K.: Toward information systems over ontological graphs. In: Yao, J., et al. (eds.) Rough Sets and Current Trends in Computing, LNAI, vol. 7413, pp. 243–248. Springer-Verlag, Berlin Heidelberg (2012).

[5] Hitzler, P., KrÃűtzsch, M., Parsia, B., Patel-Schneider, P.F., Rudolph, S.: OWL 2 Web Ontology Language Primer. Tech. rep. (2009).

[6] Skowron, A., Rauszer, C.: The discernibility matrices and functions in information systems. In: Słowiński, R. (ed.) Intelligent Decision Support - Handbook of Applications and Advances of Rough Set Theory, pp. 331–362. Kluwer Academic Publishers, Dordrecht (1992).

[7] Bloehdorn, S., Hotho, A.: Ontologies for machine learning. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies, pp. 637–661. Springer, Berlin, Heidelberg (2009).

Keywords: Information systems over ontological graphs :: OWL ontology :: Python :: Software tool

Challenges in Descriptor-Based Information Systems and Machine Learning by Rule Generation

Hiroshi Sakai[1], Michinori Nakata[2]

[1]Graduate School of Engineering, Kyushu Institute of Technology, Tobata, Kitakyushu 804-8550, Japan

[2]Faculty of Management and Information Science, Josai International University, Gumyo, Togane, Chiba 283-0002, Japan

Extended Abstract. We combined the rough sets-based concepts by Pawlak and Skowron, a non-deterministic information system (NIS) by Orłowska and Lipski, missing values by Grzymała-Busse, and the Apriori algorithm by Agrawal to realize the NIS-Apriori based rule generator. Some actual execution videos are in http://www.mns.kyutech.ac.jp/~sakai/RNIA. We are now considering the following challenges.

(Challenge 1: DbIS and rule generation) We mainly handled tabular data sets and wanted to deal with non-tabular data sets. We propose the framework of a DbIS (Descriptor-based Information System) in Fig. 1. A rule is a logical implication consisting of descriptors, and we implicitly employed a descriptor [$attribute, val$] in tabular data sets until now. If we define descriptors in non-tabular data sets, we may generate rules from them. Rough

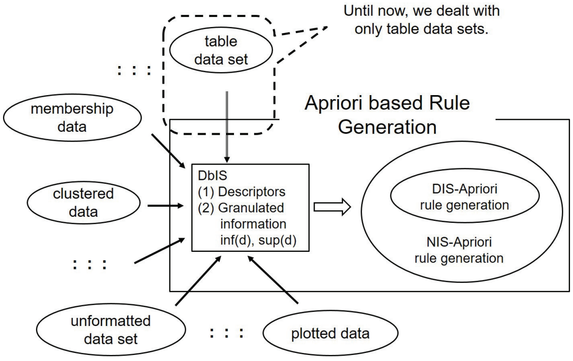
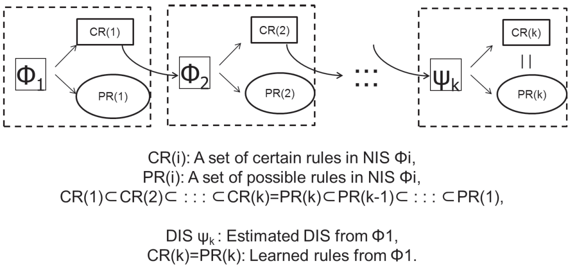

Until now, we dealt with only table data sets.
table data set
: : :
membership data
Apriori based Rule Generation
clustered data
DbIS
(1) Descriptors
(2) Granulated information
inf(d), sup(d)
DIS-Apriori rule generation
NIS-Apriori rule generation
: : :
unformatted data set
: : :
plotted data

Figure 1: The overview of rule generation from several types of data sets with uncertainty via DbIS [1].

set theory is not for only rule generation, and we see DbIS takes the role of the rough set theory specialized for rule generation. The extension from tabular data sets to DbIS will extend the research areas of rough sets much more.

(Challenge 2: Missing value estimation and machine learning by rule generation) The obtained rules from data sets are applied to decision-making. They take the similar role of regression lines in statistics. We are now considering the application of certain rules to missing value estimation in Fig. 2.



CR(i): A set of certain rules in NIS Φi,
PR(i): A set of possible rules in NIS Φi,
CR(1)⊂CR(2)⊂ : : : ⊂CR(k)=PR(k)⊂PR(k-1)⊂ : : : ⊂PR(1),

DIS ψ_k : Estimated DIS from Φ1,
CR(k)=PR(k): Learned rules from Φ1.

Figure 2: The overview of machine learning by rule generation [2].

In [2], we employed the given decision attribute *Dec* and estimated missing values so as to increase the *accuray* value of one selected certain rule (with the highest *accuracy* value) $Condition\_part \Rightarrow [Dec, val]$. However, we newly propose the following procedure.

1. We fix one attribute A in NIS Φ, where missing values exist.
2. We generate certain rules $Condition\_part \Rightarrow [A, val]$. Here, NIS-Apriori-based rule generation is essential because missing values usually exist in other attributes.
3. If the obtained certain rule matches the object with a missing value on the attribute A, we estimate it as the decision attribute value val in $[A, val]$.
4. We repeat the above 1, 2, and 3 procedures for the revised NIS Φ'.

This procedure detects hidden local dependency to the attribute A, and we may recover the missing value due to functional dependency in database theory. We are now investigating the validity of this procedure.

Acknowledgment: This work is supported by JSPS (Japan Society for the Promotion of Science) KAKENHI Grant Number JP20K11954.

References:

[1] Sakai, H., Nakata, M., Apriori-based rule generation with three-way decisions for heterogeneous and uncertain data, Proc. SCIS-ISIS2022.

[2] Sakai, H., Nakata, M., Watada, J., A proposal of machine learning by rule generation from tables with non-deterministic information and its prototype system, Proc. IJCRS2017.

Keywords: No keywords

Clusters Evolution Before, During and After the Pandemic

Raavee Kadam[1], Pawan Lingras[1]

[1]Saint Mary's University, Halifax, Canada

Extended Abstract. Clusters represent groups of similar data points. Once seen as static entities, clusters are now recognized as dynamic formations that shift and transform over time. The study of changing clusters over a period offers insights into the underlying patterns, trends, and shifts within data, enabling a deeper understanding of complex phenomena in various domains [1]. Clusters can change for a variety of reasons, reflecting the inherent dynamism underlying the data [1]:
setcounterfigure42

- Concept Drift: The fundamental concepts that define clusters may shift due to changing conditions, introducing new trends or behaviors.
- Seasonal Patterns: Some clusters might exhibit recurring patterns over time, which can be captured through dynamic clustering to discern cyclic trends.
- Emerging Anomalies: As data evolves, anomalies or outliers may arise, leading to the formation of new clusters or the modification of existing ones.
- Adaptive Systems: In applications like recommendation systems or personalized marketing, clusters can change based on user interactions and feedback.

Evolving clusters provide a dynamic lens to view complex datasets that transform over time. This evolving perspective not only enriches our understanding of data but also equips us with tools to make more informed decisions in an ever-changing environment. Along similar lines, this presentation reports changes in profiles of stores for a province-wide retailer before, during and after the pandemic. The data spans 7 years from January 2015 to June 2022. Point of sales information was available for around 106-109 stores of the organization over the entire study period. This data consisted of stores that served 50-51 of the forward sortation areas (FSA) within Canada. An FSA is a way to designate a geographical unit based on the first three characters in a Canadian postal code. Table 1 shows the annual sales information from 2015 to 2021. For privacy reasons, the actual values for sales are divided by a masked figure called X. The quantity reported is similarly masked using a factor Y.

Table 1: Annual Sales.

Year	**Sales*X**	**Quantity*Y**
2015	$1,666.67	100.00
2016	$1,933.33	100.00
2017	$2,000.00	116.67
2018	$2,000.00	116.67
2019	$2,000.00	116.67
2020	$2,333.33	133.33
2021	$2,333.33	133.33

We wanted to profile a store in a given year based on the percentage of products sold in various categories. The categories are labelled - A: Social products, B: Fashionable products, C: Connoisseur products, D: Emerging products. The percentage of revenues from each category was used to represent an annual pattern. The annual pattern for a store is treated as a separate object. Each store may have up to seven patterns for the years 2015-2021. The optimal number of clusters were determined using a sum of scatter within clusters. The sum of scatter started rising sharply after eight clusters. Therefore, eight clusters were deemed to be the optimal number of clusters. Eight distinct groups were identified along with the transition of a store from cluster to cluster over seven years. Furthermore, we also identified the transition of an FSA from cluster to cluster and changes in the number of patterns in each cluster over the seven years. Table 2 shows the centroids of each cluster. Each column represents a cluster. The clusters are labelled based on the revenues from each category A, B, C, D and N. For example, A41D22B20C17 means category A resulted in 41% of the revenue, category D provided 22% of the revenue, and so on. The categories in the cluster label are arranged based on decreasing revenue - e.g., cluster A41D22B20C17 has the highest revenue from category A, followed by D, B, and C.

Table 3 shows the distribution of cluster membership by year as well as total. Each row in the table shows the number of patterns belonging to the year in different clusters. The last

Table 2: Cluster Centroids.

	A41D22 **B20C17**	**A42B30** **C15D12**	**A44C24** **B17D14**	**A46D21** **C19B12**	**A46B21** **C17D15**	**A54C20** **D13B12**	**N**	**B60A19** **C08D06**
A : Social	41%	42%	44%	46%	46%	54%	0	19%
B : Chic	20%	30%	17%	12%	21%	12%	0	60%
C : Connoisseur	17%	15%	24%	19%	17%	20%	0	08%
D : Emerging	22%	12%	14%	21%	15%	13%	0	06%
N : Gifts	0	0	0	0	0	0	100	0

row shows the sizes of each cluster. It is interesting to note that clustering separated the last two clusters labelled N and B60A19C08D06 as outliers. These were special-purpose stores and correctly separated from other stores by the clustering process. The rest of the clusters were formed organically. One can see that there is a clear distinction between pre-pandemic and post-pandemic revenue patterns. Clusters A42B30C15D12, A44C24B17D14, A46B21C17D15, and A54C20D13B12 disappeared and were essentially replaced by A41D22B20C17 and A46D21C19B12. This suggested a decrease in previously considered social products with emerging products. The clustering managed to identify the effect of the pandemic without any explicit input regarding social behavior.

Table 3: Cluster membership by year.

Year	**A41D22** **B20C17**	**A42B30** **C15D12**	**A44C24** **B17D14**	**A46D21** **C19B12**	**A46B21** **C17D15**	**A54C20** **D13B12**	**N**	**B60A19** **C08D06**	**Total**
2015		18	24	2	31	28		2	105
2016	2	18	24	4	30	26		2	106
2017		17	21	4	37	25		1	105
2018	1	16	17	5	39	26		2	106
2019	4	15	15	6	41	25	1	1	108
2020	44	9	4	40	8	1	1	1	108
2021	63	4		34	4	1	1	1	108
Total	114	97	105	95	190	132	3	10	746

References:

[1] OpenAI (2023). ChatGPT (Mar 14 version) [Large language model]. https://chat.openai.com/chat.

Keywords: No keywords

Rule-based action mining from survival data

Marek Hermansa[1,2], Marek Sikora[1,2], Beata Sikora[1], Łukasz Wróbel[1,2]

[1]Silesian University of Technology, 44-100 Gliwice, Poland

[2]Łukasiewicz Research Network - EMAG Institute, 40-189 Katowice, Poland

Extended Abstract. Logical rules ($\varphi \wedge \psi$) represent a simple and understandable form of knowledge representation. In rule-based reasoning, the premise φ determines which conditions must be satisfied for the conclusion ψ to be true. The rule premise has the form of a conjunction of elementary conditions $w_i \equiv a_1 \odot x_i$, where x_i is an element of the domain of the attribute a_i and $\odot$ is a relation symbol (e.g. $=, <, \leq, >, \geq, \in$). In the case of survival rules, the conclusion is the estimator of a survival function ($\hat{S}$), e.g., the Kaplan-Meier estimator [1]. Thus, a survival rule has the following form

$$\text{IF } w_1 \wedge w_2 \cdots \wedge w_n \text{ THEN } \hat{S}. \tag{1}$$

A survival action rule is a concatenation of two survival rules [2].

$$\text{IF } w_{1S} \to w_{1T} \wedge w_{2S} \to w_{2T} \cdots \wedge w_{nS} \to w_{nT} \text{ THEN } \hat{S}_S \to \hat{S}_T. \tag{2}$$

The elementary action $w_{iS} \to w_{iT}$, $i = 1, \ldots, n$, represents a change in the value of the attribute a_i. It consists of the premise w_{iS} of the elementary action, which specifies the source (S) range of values of this attribute, and the conclusion w_{iT} of the elementary action, which specifies the target (T) range of these values. In other words, the premise of a survival action rule defines the transition from the source to the target representation. This transition aims to change the survival curve contained in the conclusion of the rule. A survival action rule induction algorithm based on the sequential covering strategy is presented in [2]. In this paper, two strategies for applying the induced rules to recommend changes in attribute values are presented. The main objective of the recommendation is to change the attribute values of the considered example (e.g., a test example) such that its estimated survival time is significantly different from the estimated time before the changes.

We present two strategies for recommendation mining: a global strategy based on a set of induced action rules, and a local strategy based on any survival model [5] (e.g., Random Survival Forest). In the local strategy the survival model is called the arbiter model (AM).

In the global strategy – as in the standard action rule induction [6] – an example may be covered by several rules, to find recommendation it must be decided which elementary actions will be applied and to what extent (how large the change in attribute values should be). In the local strategy, it must be decided which attribute values should be changed to make the survival time estimated for the test example by AM significantly different before and after changing the attribute values.

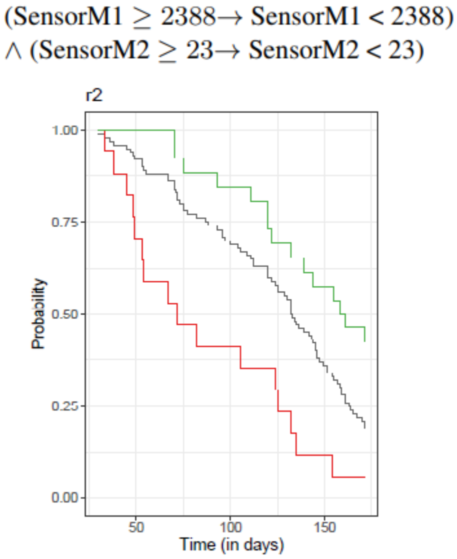


Figure 1: An exemplary survival rule. The gray, red, and green curves represent overall survival, the left rule, and the right rule, respectively. The rule illustrates how changes in the values monitored by the two sensors, SensorM1 and SensorM2, affect the estimated time of reliable operation of a device. The rule was induced based on the well-known benchmark data set FD001

Both the global and local strategies employ a method of computation of an approximate, quasi-shortest object-related decision reduct [3]. The reduct is searched in decision table with decision attribute reflecting discretized survival time. The recommendations are built based on the attributes included in the reduct. The presented strategies can be considered as actionable knowledge discovery techniques [4].

The experiments were performed on benchmark data sets. Since in most cases it is not possible to verify whether the recommended attribute value changes affected the change of the survival time of test examples, the independent survival model (*ISM*) for result verification was used. In the experiments, we verify whether the survival times - for examples with attribute values changing - estimated by our strategies and *ISM* do not differ statistically. 10-fold cross-validation was used as a method to validate the model.

Table 1: Method results for the selected 12 datasets obtained using 10-fold cross-validation and the local strategy. The columns indicate: the name of the dataset (Dataset); the average number of survival action rules for the dataset (n_R); the average number of actions in the rule (n_A); the right rule consistency score - the percentage of examples for which the p-value is less than 0. 05, testing the null hypothesis that the curve for the right side of the recommendation is identical to the curve derived from the independent survival model (Diff.); evaluation of the significance of the changes made - the percentage of examples for which the p-value is less than 0.05, testing the null hypothesis that the curves, before and after the recommendation, are identical (Sign.)

Dataset	n_R	n_A	Diff.	Sign.
bmt-ch	4.7	2.5	100.0	94.2
follic	7.2	1.9	100.0	89.6
GBSG2	13.8	3.3	100.0	92.9
hd	10.1	1.2	100.0	99.8
lung	7.1	2.8	100.0	92.1
Melanoma	3.9	2.6	100.0	94.7
mgus	4.1	3.2	99.2	90.0
pbc	6.8	2.7	100.0	92.1
std	15.1	4.2	100.0	80.4
uis	8.0	3.5	100.0	99.8
whas1	6.9	2.5	99.2	100.0
whas500	7.9	4.3	100.0	97.2

References:

[1] Kaplan, E. L., Meier, P.: Nonparametric estimation from incomplete observations. Journal of The American Statistical Association, vol. 53, pp. 457-481, 1958.

[2] Badura, J., Hermansa, M., Kozielski, M., Sikora, M., Wróbel: Separate-and-Conquer Survival Rule Learning. Knowledge Based Systems (accepted for publication; available online: https://www.sciencedirect.com/science/article/abs/pii/S0950705123007311).

[3] Nguyen, S., H., Nguyen, S.,H.: Some efficient algorithms for rough set methods. Proceedings IPMU, 96, pp. 1541–1457, 1996.

[4] Kalanat, N.: An overview of actionable knowledge discovery techniques. Journal of Intelligent Information Systems, 58, pp.1–21, 2022.

[5] Klein, J.P., Moeschberger, M.L.: Survival analysis: techniques for censored and truncated data. Springer Science & Business Media, New York (2005).

[6] Sikora, M., Mayszok, P., Wróbel, Ł.: SCARI: Separate and conquer algorithm for action rules and recommendations induction. Information Sciences, 607, pp. 849–868, 2022.

Keywords: Rule induction :: Action mining :: Survival analysis

Time Factor in Diachronic Embedding for Temporal Knowledge Graph Completion

Thuy-Anh Nguyen Thi[1,3], Hieu Cong Nguyen[2], Xuan-Hieu Phan[3], Quang-Thuy Ha[3]

[1]Banking Academy of Vietnam, Hanoi, Vietnam

[2]School of Applied Mathematics and Informatics, Hanoi university of Science and Technology, Hanoi, Vietnam

[3]VNU-University of Engineering and Technology (UET), Vietnam National University (VNU), Hanoi, Vietnam

Extended Abstract. Recently, one notable embedding-based temporal knowledge graph completion (TKGC) model is DE-SimplE (DE) which introduced by Goel et al. (2020). This model provides the characteristics of entities at any point in time. In this manuscript, we improve the DE-SimplE model with two components mentioned by Zhang et al. (2022) that can enhance the connections of facts over a period of time as well as combine the relation's meaning and temporal information.

Goel et al. [1] introduced an diachronic entity embedding (DE) function whose input is a pair of an entity and a timestamp and output is a hidden representation. The advantage of DE is that it could be combined with any KGC model to become a TKGC model. By experiments, the authors in [1] showed that relations might evolve at a very lower rate or only negligibly. Therefore, they only used a static representation for relations. As the same, the authors in [2] also used the static representation for relations. In this research, we show that the timestamps play important role for relations by combining the advance of two components which were mentioned by Zhang et al. [3]: The shared time window (STW) enhances the correlation between adjacent timestamps and the relation-timestamp composition (RTC) which integrates temporal features to the semantic features of relations.

By using STW, we modified the `DEEMB` function in [1] as follows:

$$\boldsymbol{z}_e^{\tau\prime} = \begin{cases} \boldsymbol{t}_c[n] + \boldsymbol{a}_e[n]\sigma(\boldsymbol{w}_e[n]\tau + \boldsymbol{b}_e[n]), & \text{if } 1 \leq n \leq \gamma d \\ \boldsymbol{a}_e[n], & \text{if } \gamma d \leq n \leq d \end{cases} \tag{1}$$

where $\boldsymbol{a}_e \in \mathbb{R}^d$ and $\boldsymbol{w}_e, \boldsymbol{b}_e \in \mathbb{R}^{\gamma d}$ are learnable representations of entity e, σ is an activation function, and $\boldsymbol{t}_c \in \mathbb{R}^{\gamma d}$ is the new component that indicates the features of cycle c which contains τ. For modeling relation, we adopt RTC which fuses both semantic information and temporal information through the $\circ$ element-wise product.

$$\boldsymbol{z}_r^{\tau\prime} = \boldsymbol{z}_r + \boldsymbol{z}_r \circ \boldsymbol{t}_{comp}^{\tau} \tag{2}$$

where $\boldsymbol{z}_r \in \mathbb{R}^d$ is the representation of relation r and $\boldsymbol{t}_{comp}^{\tau} \in \mathbb{R}^d$ is a learnable vector related to timestamp τ.

Our proposed model combines RTC and STW in the DE-SimplE model by using Equation (1) to model entities and Equation (2) to model relations. We compare the performance of our model with that of the DE-SimplE model, considered the best model in [1]. We also evaluate our proposed model using three datasets: ICEWS14, ICEWS05-15, and GDELT, with the hyperparameters from [1]. Since GDELT is very large, we run this dataset for 100 epochs instead of the 500 epochs mentioned in [1]

Table 1: Experimental results on ICEWS14, ICEWS05-15, and GDELT.

Model	ICEWS14					ICEWS05-15					GDELT				
	$MR\downarrow$	$MRR\uparrow$	$Hits@1\uparrow$	$Hits@3\uparrow$	$Hits@10\uparrow$	$MR\downarrow$	$MRR\uparrow$	$Hits@1\uparrow$	$Hits@3\uparrow$	$Hits@10\uparrow$	$MR\downarrow$	$MRR\uparrow$	$Hits@1\uparrow1$	$Hits@3\uparrow$	$Hits@10\uparrow$
DE-SimplE [1]	-	0.526	41.8	59.2	**72.5**	-	0.513	39.2	57.8	74.8	-	**0.230**	**14.1**	**24.8**	**40.3**
Proposed Model	226	**0.549**	**45.3**	**61.0**	72.3	111	**0.543**	**42.2**	**61.3**	**77.6**	60.35	0.222	13.7	23.7	38.6

Our final results are reported in Table 1. Best results are represented in **bold** font. Based on experimental results from Table 1, we demonstrate that the timestamps factor plays a crucial role in enhancing the performance of the DE-SimplE model for TKGC.

References:

[1] R. Goel, S. M. Kazemi, M. Brubaker, P. Poupart, "Diachronic embedding for temporal knowledge graph completion", *Proc. of the AAAI conference on artificial intelligence*, vol. 34, no. 04, pp 3988–3995, 2020.

[2] T.-A. Nguyen Thi, V.-P Ta, X.-H. Phan, Q.-T. Ha, "An Improvement of Diachronic Embedding for Temporal Knowledge Graph Completion", *15th Asian Conference on Intelligent Information and Database Systems*, in press.

[3] F. Zhang, Z. Zhang, X. Ao, F. Zhuang, Y. Xu, and Q. He, "Along the Time: Timeline-traced Embedding for Temporal Knowledge Graph Completion", *Proceedings of the*

31st ACM International Conference on Information & Knowledge Management, pp 2529–2538, 2022.

Keywords: Temporal knowledge graph :: Diachronic embedding

Mining Inter-sequence Patterns with Itemset Constraints

Anh Nguyen[1], Ngoc-Thanh Nguyen[1], Loan T.T. Nguyen[2,3], Bay Vo[4]

[1]Department of Applied Informatics, Wroclaw University of Science and Technology, Poland

[2]School of Computer Science and Engineering, International University, Ho Chi Minh City, Vietnam

[3]Vietnam National University, Ho Chi Minh City, Vietnam

[4]Faculty of Information Technology, HUTECH University, Ho Chi Minh City, Vietnam

Extended Abstract.

1. Introduction

Mining inter-sequence patterns represents a novel and crucial research avenue within data mining. However, this research model has limitations, notably in generating a large number of frequent patterns. This places significant demands on both storage space and processing time. Given the overwhelming amount of information, pro-cessing it becomes a formidable challenge. Consequently, our research has concen-trated on generating frequent patterns based on user-defined criteria, thereby curtail-ing the proliferation of such patterns. In light of this approach, we have introduced an algorithm named DBV-ISPMIC. Furthermore, we have developed a property and leveraged it to propose an enhanced algorithm to reduce the need for candidate check-ing. Lastly, we put forth the pDBV-ISPMIC algorithm, representing an optimized parallelization approach for the DBV-ISPMIC algorithm. The pDBV-ISPMIC algorithm demonstrates that employing a parallel system yields notable improvements in algorithm execution time. Furthermore, it underscores the superior performance of the DBV-ISPMIC algorithm compared to other methods in mining inter-sequence pat-terns with itemset constraints, as substantiated by our experimental results.

2. Contributions

In this study, we aim to solve the task of inter-sequence pattern mining with itemset constraints. Unlike using itemset constraints for mining sequence patterns, this mining task requires more complex processing because many candidates are generated during the mining process. Our significant contributions are as follows.

1. Based on the EISP-Miner algorithm [1] and a using dynamic bit vector data structure [2], we state the problem of inter-sequence pat-tern mining in combination with itemset constraints.

2. We then suggest a proposition to help reduce candidate checking during sequence expansion according to the EISP-Miner algorithm, thus reducing the search space for inter-sequence pattern mining with itemset constraints.

3. Next, an algorithm named DBV-ISPMIC is developed to discover constraints-based inter-sequence patterns. A parallel version of the DBV-ISPMIC algorithm, named the pDBV-ISPMIC algorithm, was presented.

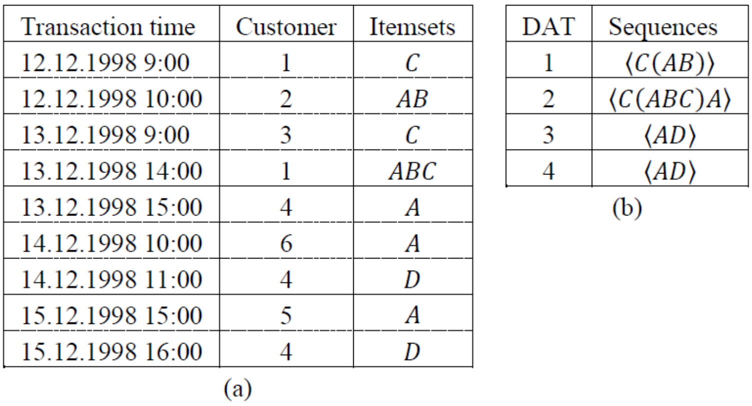
4. Finally, we conduct experiments with various databases to evaluate the proposed method.

3. Basic concepts and problem statement

Definition 1 (*items*, *itemsets*, *sequences*, *sequence database*). Let t be the set of items, $t = u_1, u_2, ..., u_m$ where u_i is an item $(1 \leq i \leq m)$. A sequence $s = \langle t_1, t_2, t_3, ..., t_n \rangle$ is an ordered list of itemsets where $t_i \subseteq t$ $(1 \leq i \leq n)$ is an itemset. A sequential database $D = s_1, s_2, s_3, ..., s_w$ where $w = |D|$ is the number of sequences in D and s_i $(1 \leq i \leq w)$ is a pair of values $\langle Dat, Sequence \rangle$, in which Dat is the property of s_i used to describe contextual information based on the time of the transaction.

Definition 2 (*Megasequence*). Given a list of k sequences $\langle d_1, s_1 \rangle, \langle d_2, s_2 \rangle, ..., \langle d_k, s_k \rangle$ in the sequential database. A megasequence with $k \geq 1$ is denoted as $\Psi = s_1[0] \cup s_2[d_2 - d_1] \cup ... \cup s_k[d_k - d_1]$. From the example database shown Table 1, with $maxspan = 1$ and $DAT = 1$ as the reference point, we have a list of megasequences shown in Table 2.

Table 1: An example of creating a sequential database (b) from a customer dataset (a).



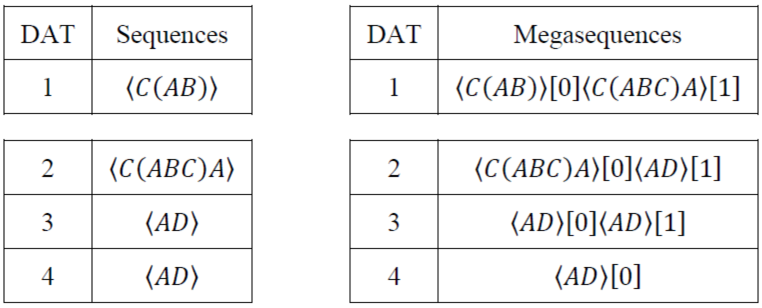
Transaction time	Customer	Itemsets
12.12.1998 9:00	1	C
12.12.1998 10:00	2	AB
13.12.1998 9:00	3	C
13.12.1998 14:00	1	ABC
13.12.1998 15:00	4	A
14.12.1998 10:00	6	A
14.12.1998 11:00	4	D
15.12.1998 15:00	5	A
15.12.1998 16:00	4	D

(a)

DAT	Sequences
1	$\langle C(AB)\rangle$
2	$\langle C(ABC)A\rangle$
3	$\langle AD\rangle$
4	$\langle AD\rangle$

(b)

Table 2: Converting a sequential database to megasequences.



DAT	Sequences
1	$\langle C(AB)\rangle$
2	$\langle C(ABC)A\rangle$
3	$\langle AD\rangle$
4	$\langle AD\rangle$

DAT	Megasequences
1	$\langle C(AB)\rangle[0]\langle C(ABC)A\rangle[1]$
2	$\langle C(ABC)A\rangle[0]\langle AD\rangle[1]$
3	$\langle AD\rangle[0]\langle AD\rangle[1]$
4	$\langle AD\rangle[0]$

Definition 3 (*1-patterns extension*). Given two frequent 1-patterns $\alpha = \langle u\rangle[0]$ and $\beta = \langle v\rangle[0]$. α is joinable to β in any instance and there are three types of join opera-tion: (1) itemset-join: $\alpha \cup_i \beta = \langle(u, v)\rangle[0]|\langle(u, v)\rangle[0]$; (2) sequence-join: $\alpha \cup_s \beta = \langle(u, v)\rangle[0]$; (3) inter-join: $\alpha \cup_t \beta = \langle u\rangle[0]\langle v\rangle[x]|1 \leq x \leq maxspan$.

For example, given $maxspan = 2$, $\langle C\rangle[0]\cup_i\langle D\rangle[0] = \langle(CD)\rangle[0]$; $\langle C\rangle[0]\cup_s\langle D\rangle[0] = \langle CD\rangle[0]$; and $\langle C\rangle[0] \cup_t \langle D\rangle[0] = \langle C\rangle[0]\langle D\rangle[1], \langle C\rangle[0]\langle D\rangle[2]$.

Definition 4 (*k-patterns extension*). Given 2 frequent k-patterns α and β , $k > 1$, then $sub_{k,k}(\alpha) = (u)[i]$, and $sub_{k,k}(\beta) = (v)[j]$. α is joinable to β if $sub_{1,k-1}(\alpha = sub_{1,k-1}(\beta)$ and $i \leq j$, which yields three types of join operation: (1) itemset-join: $\alpha \cup_i \beta = \alpha +_i (v)[j]|(i = j) \wedge (u < v)$; (2) sequence-join: $\alpha \cup_s \beta = \alpha +_s (v)[j]|(i = j)$; (3) inter-join: $\alpha \cup_t \beta = \alpha +_t (v)[j]|(i < j)$.

For example, $\langle BC\rangle[0] \cup_i \langle BD\rangle[0] = \langle B(CD)\rangle[0]$, $\langle BC\rangle[0] \cup_s \langle BD\rangle[0] = \langle BCD\rangle[0]$, and $\langle BC\rangle[0] \cup_t \langle B\rangle[0]\langle D\rangle[2] = \langle BC\rangle[0]\langle D\rangle[2]$.

Definition 5 (*Prefix*). A pattern $\beta = \langle b_1, b_2, ..., b_m\rangle$ is called a prefix of pattern $\alpha = \langle \alpha_1, \alpha_2, ..., \alpha_n\rangle$ if and only if $b_i = \alpha_i$ for all $1 \leq i \leq m-1$, $b_m \subseteq \alpha_m$, $m < n$.

For instance, the prefixes of pattern $\langle D(BC)B\rangle[0]$ are: $\langle D\rangle[0]$, $\langle DB\rangle[0]$ and $\langle D(BC)\rangle[0]$. Thus, based on this definition, any sequence would be the prefix of its extended sequences.

Definition 6 (*Problem statement*). Given a sequence database D, the minimum support ($minsup$), and a set of constraint itemsets $IC = c_1, c_2, c_3, ..., c_k$. The task of inter-sequence pattern mining with an itemset constraint is to discover all frequent sequences $\alpha = \alpha_1[w_1], \alpha_2[w_2], ..., \alpha_m[w_m]$ such that $\exists \alpha_i[w_i] \in \alpha$, $\exists b_j \in IC : b_j \subseteq \alpha_i$.

For instance, let $IC = (C), (E)$, the sequence $\langle C(AB)\rangle[0]\langle C(ABC)A\rangle[1]$ satisfies the constraint whereas the sequence $\langle AD\rangle[0]\langle A\rangle[1]$ does not.

4. Proposed algorithm

4.1 DBV-ISPMIC algorithm

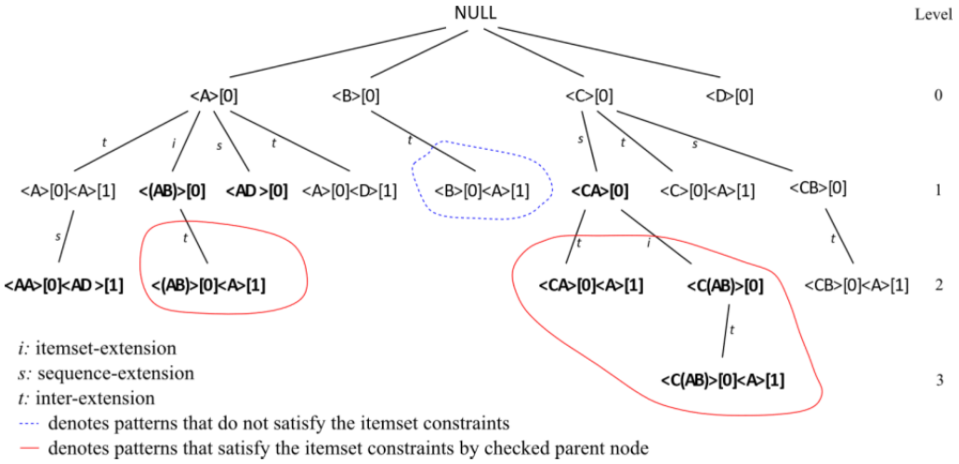


Figure 1: The extended tree of patterns corresponding to the example database.

4.2 Parallel DBV-ISPMIC algorithm

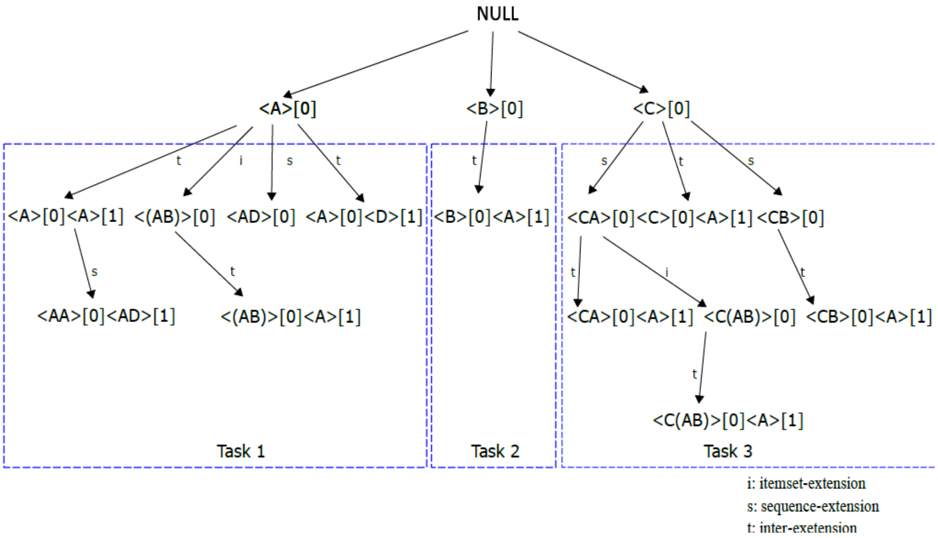


Figure 2: Example of using parallel processing for ISP-tree extension.

5. Experimental databases

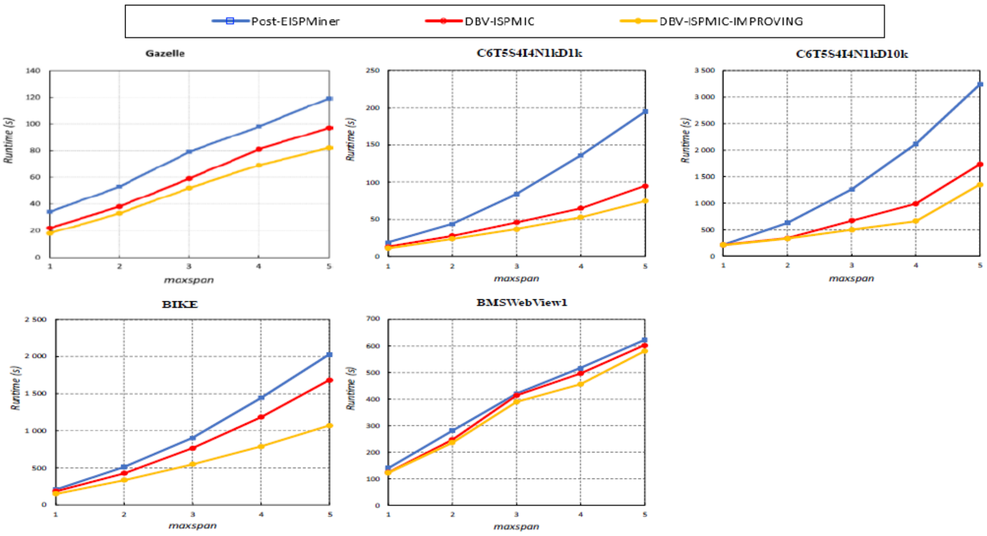


Figure 3: Runtime.

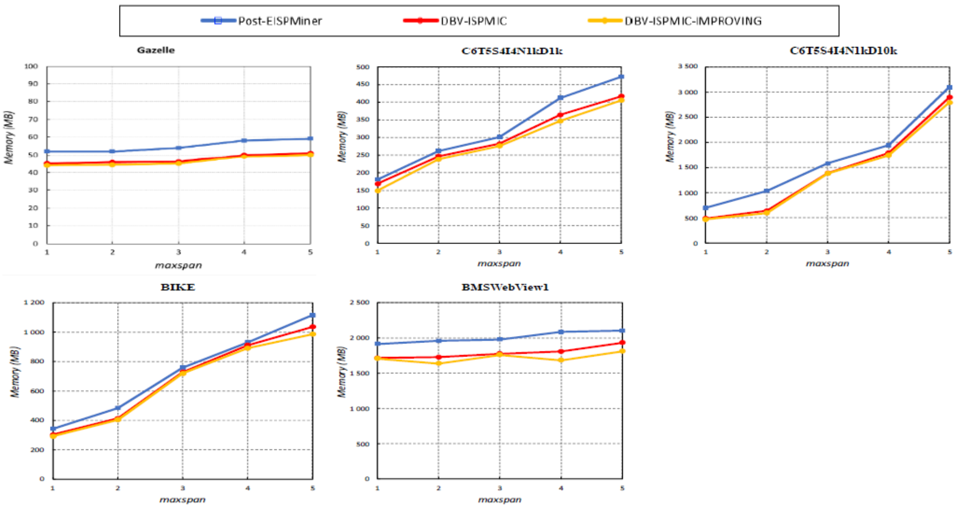


Figure 4: Memory usage.

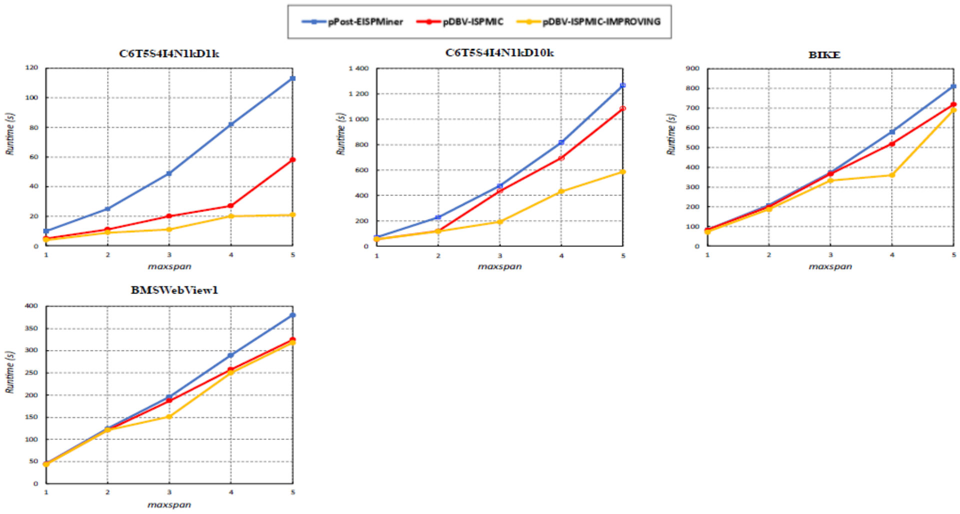


Figure 5: Parallel method for efficient mining of inter-sequence patterns with itemset constraints.

References:

[1] Wang, C. S., Lee, A. J. T. (2009). Mining inter-sequence patterns. Expert Systems with Applications, 36(4), 8649-8658.

[2] Vo, B., Tran, M. T., Hong, T. P., Nguyen, H., Le, B. (2012). A dynamic bit-vector

approach for efficiently mining inter-sequence patterns. Proceedings - 3rd International Conference on Innovations in Bio-Inspired Computing and Applications, IBICA 2012, 51-56.

Keywords: No keywords

Variable Precision Rough Set Model and Generalized γ-Decision Valuation

Soma Dutta[1], Dominik Ślęzak[2,3,4]

[1]University of Warmia and Mazury in Olsztyn, ul. Słoneczna 54, 10-710 Olsztyn, Poland

[2]Institute of Informatics, University of Warsaw, ul. Banacha, 02-097 Warsaw, Poland

[3]QED Software Sp. z o.o., ul. Miedziana 3A/18, 00-814 Warsaw, Poland

[4]DeepSeas, USA/Poland, https://www.deepseas.com/

Extended Abstract. Given an information system $\mathbb{A} = (U, \mathcal{A})$ [1,2], using the classical rough set model a set of objects from U (or a concept) can be described in terms of three regions, namely positive region, negative region and boundary region of the concept. However, in practice going beyond such approximation more finer tuning techniques are required so that some important data about the so-called boundary objects or negative examples can also be restored. *Variable Precision Rough Set* (VPRS) model [3] provides a generalization of the classical rough set model by incorporating a probabilistic measure as a threshold for tuning the inclusion of an object as a positive example. Accordingly different aspects of data reduction [4] come up. In the context of a decision system, i.e. $(U, A \cup \{d\})$ where d is a designated attribute denoting decision, the aspect of data reduction is to restore the significant character of the decision classes (i.e., sets of objects $D_1, D_2, \ldots D_r$ having specific decision values from a value set, say $V_d = \{v_1, v_2, \ldots, v_r\}$) as much as possible encoded in the granules, created with respect to the information signatures of the objects. In this regard, the first step is to design a decision valuation in a way that the information about the decision classes obtained from a given decision system, say $(U, \mathcal{A} \cup \{d\})$, can be encoded in that decision valuation, say ν_d, in a compressed way. Then the next step is to look for a subset of attributes which can preserve the information compressed in ν_d with respect to the whole set of attributes. Following this line of thought there can be many decision valuations focusing on different aspects of decision making and consequently different notions of decision reduct obtained based on them.

A decision valuation is a function which assigns a value from a decision space, say $\mathcal{D}$, to the objects of a decision system based on their information signature. The meaning of $\mathcal{D}$ may depend on the way of designing a decision model to infer about decision values. That is, if $\mathcal{V}_A$ denotes the set of all information signatures (i.e., the vectors of values describing objects with respect to the conditional attributes) available in $\mathbb{A}$ for any $X \subseteq A$, then $\nu : \mathcal{V}_A \mapsto \mathcal{D}_\nu$ represents a decision valuation [5] having decisions from the subspace $\mathcal{D}_\nu$ of $\mathcal{D}$.

For example γ-decision reduct looks for a reduced set of attributes focusing only on the consistent objects of a given decision system. Thus, the information regarding the inconsistent object of a data table is completely ignored while making decision. In contrary to that, ∂-decision reduct stores decision related information of all the objects belonging to a particular equivalence class. So, the respective decision valuations can be presented as follows.

Example 1. Given $\mathbb{A} = (U, A \cup \{d\})$, the generalized decision valuation (or ∂-decision valuation in short) $\nu_\partial : \mathcal{V}_A \mapsto \mathcal{D}_\partial$ is defined as follows; for any vector $\overrightarrow{x} \in \mathcal{V}_A$ of values on attributes $X \subseteq A$:

$$\nu_\partial(\overrightarrow{x}) = \{d(u') : Inf_X(u') = \overrightarrow{x}\} \tag{1}$$

whereby $\mathcal{D}_\partial$ denotes the space of all non-empty subsets of decision values.

Example 2. Given $\mathbb{A} = (U, A \cup \{d\})$, the positive decision valuation (γ-decision valuation in short) $\nu_\gamma : \mathcal{V}_A \mapsto \mathcal{D}_\gamma$ is defined as:

$$\nu_\gamma(\overrightarrow{x}) = \begin{cases} v_k & \text{if } \nu_\partial(\overrightarrow{x}) = \{v_k\} \\ ? & \text{otherwise} \end{cases} \tag{2}$$

whereby $\mathcal{D}_\gamma$ is the set of all decision values along with the dummy value "?". Example 2 models a scenario in which decision can be made only in the case of full certainty/consistency; otherwise we may attach a dummy value "?" where the value "?" may be compared to *do not care* following [6]. The decision valuation ν_γ is the base function for classical rough set model for positive region, whereas we will discuss that a generalized version of ν_γ, called ν_p, is the base function for VPRS model. ν_p model in one way helps to overcome the limitation of classical ν_γ model by not throwing out all inconsistent cases from further steps of decision making. However, there still remains the possibility of improving the model and we focus on generalizing and combining ν_γ with other decision valuations in a way so that we get rid of some limitations of VPRS model as well.

References:

[1] Pawlak, Z. (1982). Rough Sets. International Journal of Computer and Information Sciences, 11(5), 341-356.

[2] Pawlak, Z., Skowron, A. (2007). Rudiments of Rough Sets. Information Sciences, 177(1), 3-27.

[3] Ziarko, W. (1993). Variable Precision Rough Set Model. Journal of Computer and System Sciences, 46(1), 39-59.

[4] Ślęzak, D., Dutta, S. (2018). Dynamic and Discernibility Characteristics of Different Attribute Reduction Criteria. Proceedings of IJCRS 2018, 628-643.

[5] Dutta, S., Ślęzak, D. (2024). Nature of Decision Valuations in Elimination of Redundant Attributes. International Journal of Approximate Reasoning.

[6] Clark, P. G., Gao, C., GrzymaÅĆa-Busse, J. W. (2017). A Comparison of Mining Incomplete and Inconsistent Data. Information Technology and Control, 46(2), 183-193.

Keywords: No keywords

Rough-Fuzzy Hybrid Approach to Attribute Importance and Ranking

Sinh Hoa Nguyen[1], Hung Son Nguyen[2]

[1]Polish-Japanese Academy of Information Technology, ul. Koszykowa 86, 02-008 Warsaw

[2]Institute of Computer Science, University of Warsaw, ul. Banacha 2, Warsaw, 02-097, Poland

Extended Abstract. Attribute importance refers to a vector of weights that are assigned to attributes and describe the effect on the overall performance or outcome of a model, system, or process. The concept is commonly used in a variety of fields, including machine learning, statistics, data analysis, and decision making. In machine learning, the importance of attributes helps determine which features have the greatest impact on model predictions. Attribute importance and ranking can have several benefits including feature selection, model interpretability and dimensional reduction. Attribute importance can be calculated by using many different techniques including *Random Forest Classifier*[1] and *Permutation Feature Importance.* Using the Random forest algorithm, the feature importance can be measured as the average impurity decrease computed from all decision trees in the forest. The second method focuses on observing how predictions of the ML model change when we change the order of variables. One of the applications of Rough set theory in machine learning is the so-called feature selection by means of finding a reduct set of attributes [2]. The concept of reducts can be used for feature ranking in natural way [3]. In [4], the RAFAR (Rough-fuzzy Algorithm For Attribute Ranking) methodology has been presented. This is a hybrid approach that combines discernibility relation of the rough set theory and the ranking method from intuitionistics fuzzy [5] theory. The RAFAR methodology consists of two main steps: (1) construction of a fuzzy pairwise comparison matrix (called Intuitionistic Fuzzy Preference Relation (IFPR)) for the set of attributes and (2) converting this matrix into the optimal consistent *weight vector*, which is the the resulting vector of attribute importance values. The experiment result on benchmark data sets were very promising. However, it exposes certain shortcomings that will improve in this paper. The new research results include:

1. **Uncertain Preference Relation and Ranking models:** The uncertainty

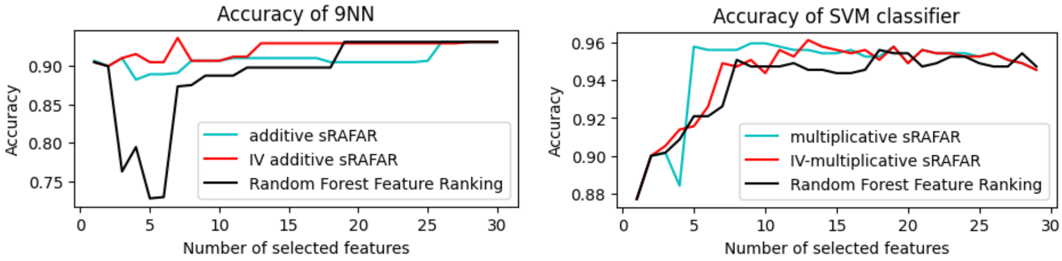


Figure 1: The accuracy comparison between five attributes ranking methods using 9NN and the SVM classifiers on WDBC dataset.

of the preference relation can be represented by either fuzzy or intuitionistic fuzzy or interval-value fuzzy set theory. The key difficulty of this research lies in modeling the pseudo-transitivity of the uncertain preference relations. This paper presents a new ranking method based on interval-valued fuzzy vectors, i.e., $\mathbf{w} = ([l_1, r_1], \cdots, [l_n, r_n])$, where $0 \leq l_i \leq r_i \leq 1$ for $i = 1, \cdots, n$. This method operates on the concepts of lower and upper bounds of the fuzzy preference relation, which is very close to the main philosophy of rough sets.

2. **Scalability of the RAFAR methodology:** One of the disadvantages of the RAFAR method in [4] is the time complexity of the matrix calculation step, which is $O(n^2 \cdot m \log m)$, where n is the number of attributes and m is the number of objects. In this paper, a new methods of generating a comparison matrix that takes into account not only the discernibility strength of a single attribute, but also its potential to be combined with other attributes is developed. This paper presents several randomized techniques that can be applied for data sets with large number of attributes as well as for data sets with large number of objects. These modifications make RAFAR more scalable, but still maintain the high level of accuracy.

The research results in this paper are both theoretical and empirical. Some interesting properties of the class of all uncertain preference matrices, like convexity or positive definiteness, will be proved, therefore the corresponding convex optimization problems can be efficiently solved. The experiment results on benchmark data sets (see Figure

1) are showing that, in many cases, the proposed method outperforms other existing ranking methods.

References:

[1] Breiman, L. (2001). Random Forests. Machine Learning, 45, 5-32.

[2] Pawlak, Z., Skowron, A. (2007). Rudiments of rough sets. Information Sciences, 177(1), 3-27.

[3] Zielosko, B., Stańczyk, U. (2020). Reduct-based ranking of attributes. Procedia Computer Science, 176, 2576-2585.

[4] Vo, B. K., Nguyen, H. S. (2022). Feature Selection and Ranking Method based on Intuitionistic Fuzzy Matrix and Rough Sets. Proceedings of the 17th Conference on Computer Science and Intelligence Systems, 30, 279-288.

[5] Atanassov, K. T. (1986). Intuitionistic Fuzzy Sets. Fuzzy Sets Syst., 20(1), 87-96.

Keywords: No keywords

A Probabilistic Rough Set Model Based on Bayesian Learning

Yoshifumi Kusunoki

Graduate School of Informatics, Osaka Metropolitan University

Gakuen-cho 1-1, Naka-ku, Sakai, Osaka 599-8531, Japan

Extended Abstract. In this study, we propose a probabilistic rough set model based on Bayesian learning. Reflecting granularity of information, we define a probabilistic model for classification of a target set. Then, we train a rough membership function by Bayesian learning, namely the function is defined by the predictive distribution of the classification under the condition that training data are known. Furthermore, we use a model selection technique of Bayesian learning in order to search the best rough membership function. After the learning, an algorithm of attribute reduction or rule induction is performed using approximation regions derived from the obtained rough membership function.

Rough set models [1] can be used as a data analysis tool to deal with uncertainty and/or inconsistency induced by indiscernibility and/or granularity of information. If attributes describing data are incomplete, a set of objects can be described clearly. In that case, the theory of rough sets provides lower and upper approximations to express uncertainty of the classification. Furthermore, rough set models are used to solve tasks of machine learning and data mining. Attribute reduction and rule induction are two major solutions for machine learning. Those algorithms are derived by reducing redundant attributes with preserving the positive region or structure of approximations [2]. Recently, explainability of machine learning methods has been attracting attention. Rough set approaches can be a solution to explainability.

The original rough set model does not tolerate to errors and/or noise in data, because of its strict application of set-operations. There are several probabilistic extensions of the rough set model to handle such statistical matters. The variable precision rough set model [3] and its generalization are the most popular probabilistic models, in which a quantity of uncertainty is measured by a rough membership function. This function can be interpreted as the probability that an object x is in the target set Y under the condition that the values of x are known. The Bayesian rough set model [4] is an another rough set model based on a probabilistic framework. in which x is assigned to the lower approximation when the posterior of $x \in Y$ given the description of x is greater than the prior.

Consider a situation that we want to apply attribute reduction to a data set. In that case, there are few objects in every equivalence (indiscernibility) class. Hence, the rough membership function is not reliable for classification of objects, especially prediction of unseen objects. We consider it is a drawback of the rough set approach. In the machine learning literature, classification models are regularized to avoid overfitting and improve its generalization capability. Bayesian learning [5] is one approach of regularization, in which a mechanism of classification is expressed by a probabilistic model and it is regularized by a prior distribution of parameters. In general, probabilistic rough set models does not have the functionality of regularization.

In this study, we will propose a probabilistic rough set model based on Bayesian learning. We define a probabilistic model reflecting granularity of information. Then, a rough membership function is defined the probability of $x \in Y$ with the fixed model parameters based on the maximum a posterior probability, namely parameters are determined by maximizing the posterior probability. Our approach is similar to the naive Bayesian rough set model [6], however, we use a more sophisticated probabilistic model. We remark that our model is not a replacement of the decision-theoretic rough set model [6]. We can combine these two models in the view of the statistical decision theory.

Furthermore, to search the best rough set model (rough membership function), we use a model selection technique of Bayesian learning, where models are varied according to attribute subsets. After the learning, we will perform an attribute reduction algorithm using the obtained rough membership function, and enumerate attribute subsets meaningful for the classification. In the presentation, we will clarify the advantage of the proposed approach by using numerical examples.

References:

[1] Pawlak, Z. (1982). Rough Sets. International Journal of Computer and Information Sciences, 11(5), 341-356.

[2] Kusunoki, Y., Inuiguchi, M. (2015). Structure-Based Attribute Reduction: A Rough Set Approach. In U. Stańczyk and L. C. Jain (Eds.), Feature Selection for Data and Pattern Recognition (pp. 113-160). Springer Berlin Heidelberg.

[3] Ziarko, W. (1993). Variable Precision Rough Set Model. Journal of Computer and System Sciences, 46(1), 39-59.

[4] Ślęzak, D., Ziarko, W. (2005). The investigation of the Bayesian rough set model. International Journal of Approximate Reasoning, 40(1), 81-91.

[5] Nakajima, S., Watanabe, K., Sugiyama, M. (2019). Variational Bayesian Learning Theory. Cambridge University Press.

[6] Yao, Y., Zhou, B. (2016). Two Bayesian approaches to rough sets. European Journal of Operational Research, 251(3), 904-917.

Keywords: Rough membership function :: Bayesian learning :: Probabilistic model

Monte Carlo Feature Filtering

Krzysztof Mnich[1], Witold R. Rudnicki[1]

[1]University of Białystok, Białystok, Poland

Extended Abstract. We investigate non-parametric and parametric Monte Carlo methods used to determine the statistical significance of scores of variables in information systems for which the null distribution is unknown.

The power of Monte Carlo tests was evaluated, which allowed us to estimate the computational complexity of the test for a desired power. The parametric approach, assuming a specific form of the tail of the null distribution, proved to be much more computationally efficient than the commonly used non-parametric one.

Introduction

Data sets that describe real-life phenomena often consist of thousands of variables involved in millions of potential interactions. Thus, feature selection and dimensionality reduction are required to make the results of analysis explainable. To this end, various methods are used, which yield importance scores of the variables under scrutiny. However, the significance of the scores is hard to determine since the null hypothesis distribution is often unknown. To this end, one can apply Monte Carlo method and compare scores of the original variables with those computed for random (contrast) variables.

Non-parametric and parametric Monte Carlo tests

In the non-parametric Monte Carlo method the test statistic is defined as a rank k of the score x among N_c scores $\{x_c\}$ of the contrast variables [1]. The p-value, defined as a probability that the rank is at least k for x drawn from the same distribution as $\{x_c\}$, is $p(k) = k/(N_c + 1)$. For a significance level $\alpha = 0.05$ with FWER correction for multiple tests, $N_c \geq 20N$ is needed, where N is the number of original variables. This can be unacceptably computationally expensive.

As an alternative, a parametric method can be used, that assumes a specific form of the null distribution, or its far tail only, up to unknown parameters. In many cases these parameters can be estimated with a sufficient precision using a relatively small number of contrast variables.

Here, we investigate parametric Monte Carlo tests and estimate the number of contrast variables needed to achieve a reasonable power of the test.

A general framework for Monte Carlo test

The p-value of x given $\{x_c\}$, for both non-parametric and parametric approach, can be found as the predictive posterior distribution of x. We first define p-value $\hat{p}(x|\theta)$ as a function of parameters θ, then find the posterior distribution of the parameters $f(\theta|\{x_c\})$. The eventual p-value is obtained by integration of $\hat{p}$ over θ, according to the law of total probability.

The exponential and power tail

Hill in [2] proposed the following null distribution of x above some point of x_0:

$$\hat{p}(x|\beta,\gamma) = \beta e^{-\gamma[\phi(x)-\phi(x_0)]},$$

where $\phi(x)$ is a known function of x. In particular, $\phi(x) = x$ corresponds to the exponential tail; $\phi(x) = \log(x)$ yields the power tail; $\phi(x|\kappa) = x^\kappa$, with a third parameter κ, approximates other cases. The posterior distributions of β, γ are Beta and Gamma distributions, respectively. With the parameter κ, the joint distribution of κ, γ can be approximated as a bivariate normal distribution.

The power of Monte Carlo tests

The loss of power of a Monte Carlo statistical test with respect to the case of a known null distribution is a probability of confirming the null hypothesis when the exact p-value is evenly distributed below the significance level α:

$$\Delta_B(\alpha) \equiv \mathcal{P}\left(p_{MC} > \alpha \mid p_{exact} \sim U(0,\alpha)\right)$$

The power loss can be calculated exactly or approximated as:

$$\Delta_B(\alpha) \approx \frac{1}{\sqrt{2\pi}} \left. \frac{\sigma(x)}{\alpha} \right|_{x:\mu(x)=\alpha}$$

where μ and σ are the expectation and standard deviation of the posterior distribution of hypothetical p-value $\hat{p}(x)$, defined above. This allows us to estimate N_c needed for the desired power of the test, see Table 1. The parametric methods allow us to reduce significantly the number of contrast variables, and hence the computational complexity.

Table 1: The number of contrast variables required to achieve 10% test power loss for $\alpha = 0.05$ with Bonferroni correction.

N	100	1 000	10 000	100 000
Non-parametric	32 000	320 000	3 200 000	32 000 000
Exponential	920	1 561	2 372	3 351
Normal	1 188	2 322	3 857	5 801
Exp. tail 10%	4 627	9 354	15 769	23 871
Three-par. tail 10%	8 834	23 762	48 957	86 187

References:

[1] Kursa, M. B., Jankowski, A., Rudnicki, W. R. (2010). Boruta - A System for Feature Selection. Fundamenta Informaticae, 101(4), 271-285.

[2] Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. The Annals of Statistics, 1163-1174.

Keywords: Feature selection :: Statistical test :: Null distribution :: Monte Carlo method :: Contrast variables

Resilient feature subsets selection with approximate decision reducts and clustering for ρ-resilient ϵ-constraint ensemble blending

Marek Grzegorowski[1], Eyad Kannout[1], Michał Grodzki[1], Hung Son Nguyen[1]

[1]Institute of Informatics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland

Extended Abstract. Machine learning (ML) is gaining popularity due to the proven accuracy of ML models in many domains. However, in practical applications, many factors may affect the quality of the developed model. Therefore, besides high accuracy, it is expected to assure more quality criteria, such as stability and resilience. The level of resilience depends primarily on the information contained in the data.

Research in this area leads to resilient feature selection, such as r-ℂ-reducts, that preserve their properties even in case of loss of up to 'r' attributes [1]. However, in some situations, r-ℂ-reducts cannot be calculated. For instance, in predicting increased seismic activity in hard coal mines [2], there is a strong dependency on a few conditional attributes, the loss of which cannot be compensated by other features, and the only feasible solution of assuring prediction quality is to ensure this data is always available. In this study, we present the algorithm searching for ρ-resilient reducts - a less rigid approach for resilient feature selection that overcomes this limitation.

In the developed algorithm, we derive N approximate decision reducts. We represent each reduct as a feature vector $R \in \{0,1\}^{|A|}$ where '1' indicates that a particular feature belongs to a reduct R. To provide an expected resilience level and derive possibly dissimilar subsets of attributes [3], we cluster the extracted N reducts (encoded as $\{0,1\}^{|A|}$ vectors) with K-means. Subsequently, we merge all reducts within each cluster with the logic 'OR' operator on their $\{0,1\}^{|A|}$ representation. Having $K = \rho$, we end up with ρ dissimilar subsets of attributes, each containing some surplus yet similar features. Let us present this concept in the following toy example. Let $\mathbb{S} = (U, A \cup \{d\})$ be a decision table, where $A = \{a,b,c,d,e,f,g\}$. Let the expected resilience level $\rho = 2$. Let's assume the triples correspond to reducts: $\{a,b,c\}, \{a,b,d\}, ...$ We may notice that attribute sets $\{a,b,c\}, \{a,b,d\}$ are relatively similar. Furthermore, we may infer that

attributes ‘c’ and ‘d’ provide similar information, complement to $\{a, b\}$. When we cluster those attribute sets together we end up with a superreduct $R_1 = \{a, b, c\} \cup \{a, b, d\} \subseteq A$, with a slight redundancy of information (e.g., in terms of discernibility). Similarly, for with $R_2 = \{d, e, f\} \cup \{d, e, g\}$. Furthermore, we may notice that R_1 and R_2 are dissimilar (only one common attribute ‘d’). Therefore, fitting two models on R_1 and R_2 attribute subsets should provide us with a certain level of resilience and diversity. We may notice that, unlike r-$\mathbb{C}$-reducts, the resilience level is not guaranteed, yet only expected.

Ultimately, we aim to construct the complete model training and ensemble blending procedure that allows for optimizing several quality criteria and taking into account the robustness and resilience of the ensemble blended. Here, we adapt ε-constrained scalarization for the investigated criteria, but instead of choosing a single model, we blend the ensemble of several Pareto-optimal solutions. Our method enables training models on possibly dissimilar subsets of attributes increasing the variety of different models within an ensemble and leading to a more robust, stable, and resilient ensemble. To show the particular qualities of our solution, we performed a set of experiments on data from the logistics industry [4]. The preliminary experimental study showed that our method yields great results and significantly improves the resilience of the ensemble.

In future research, we plan to apply more advanced feature and instance selection techniques, incorporating experts’ knowledge into the machine learning processes, or ensuring the ensemble diversity more explicitly by various feature space granulations [3, 5]. It is important to also show more theoretical properties of the developed construct and to conduct experimental analysis on a more significant number of real data sets from various industries.

Acknowledgement: Research co-funded by Polish National Science Centre (NCN) grant no. 2018/31/N/ST6/00610.

References:

[1] Grzegorowski, M., Ślęzak, D. (2019). On resilient feature selection: Computational foundations of r-C-reducts. Information Sciences, 499, 25-44.

[2] Janusz, A., Grzegorowski, M., Michalak, M., Łukasz Wróbel, Sikora, M., Ślęzak,

D. (2017). Predicting Seismic Events in Coal Mines Based on Underground Sensor Measurements. Engineering Applications of Artificial Intelligence, 64, 83-94.

[3] Grzegorowski, M., Janusz, A., Ślęzak, D., Szczuka, M. S. (2017). On the Role of Feature Space Granulation in Feature Selection Processes. In J.-Y. Nie, Z. Obradovic, T. Suzumura, R. Ghosh, R. Nambiar, C. Wang, H. Zang, R. Baeza-Yates, X. Hu, J. Kepner, A. Cuzzocrea, J. Tang, M. Toyoda (Eds.), 2017 IEEE International Conference on Big Data, BigData 2017, Boston, MA, USA, December 11-14, 2017 (pp. 1806-1815). IEEE Computer Society.

[4] Kannout, E., Grodzki, M., Grzegorowski, M. (2022). Considering various aspects of models' quality in the ML pipeline - application in the logistics sector. In M. Ganzha, L. A. Maciaszek, M. Paprzycki, D. Slezak (Eds.), Proceedings of FedCSIS'22, Sofia, Bulgaria, September 4-7, 2022 (Vol. 30, pp. 403-412).

[5] Grzegorowski, M., Litwin, J., Wnuk, M., Pabis, M., Marcinowski, L. (2023). Survival-Based Feature Extraction - Application in Supply Management for Dispersed Vending Machines. IEEE Trans. Ind. Informatics, 19(3), 3331-3340.

Keywords: No keywords

Robust Aggregative Feature Selection to gain insight into the data

Radosław Piliszek[1], Witold R. Rudnicki[2,1]

[1]Computational Centre, University of Białystok, Białystok, Poland

[2]Institute of Computer Science, University of Białystok, Białystok, Poland

Extended Abstract.

Introduction

Rough sets present a very useful, formal mechanism to derive knowledge from data. Presented on synthetic data, they clearly show the desired properties. However, real-world data, especially biomedical data, require proper curation before analysis. It is known that rough sets work better when the data is free from noise, both irrelevant and redundant features. Irrelevant only blur the view as they can only add the noise. Redundant features, while not irrelevant in principle, bring information that is already present in other features. Their inclusion may reduce the effects of the random noise, however, at the price of adding complexity. Thus, pre-filtering is of utmost importance. Moreover, to be able to gain direct insight into the data, the original features must be preserved - thus, feature selection is chosen as opposed to feature extraction. However, well-established feature selection methods may not offer enough stability in their results [3]. Moreover, the application of minimal-optimal methods to eliminate redundant variables contributes to another problem - due to their construction, they often introduce significant overfit.

Proposed solution

To mitigate the issues described above, we propose a novel method of feature selection we call Robust Aggregative Feature Selection (RAFS) and an accompanying feature dissimilarity measure - Symmetric Target Information Gain (STIG) - that is rooted in information theory and accounts for both synergy and redundancy effects. The STIG measure is defined using conditional entropies with respect to the decision variable D and the examined variables X and Y:

$$STIG(D|X,Y) = \frac{H(D|X) + H(D|Y) - 2H(D|X,Y)}{2}$$

The RAFS method depends on hierarchical clustering with the chosen measure (such as STIG) to reduce the dimensionality. It ensures robustness by employing an internal cross-validation scheme and result aggregation based on cluster representatives' popularity.

Table 1: Jaccard Score (JS) and Consistency Score (CS) for RAFS with STIG versus other established methods. FDR pre-filtering correction (level 0.10) has been applied and $n = 8$ variables were chosen. Best scores are bolded.

Method	JS	CS
RAFS STIG single-linkage	**0.47**	24616
RAFS STIG (pre-computed 1D) single-linkage	0.35	**32139**
mRMR	0.29	16538
RFE	0.15	1820
top	0.30	6499

The proposed method has been applied to various real-world datasets, including the BLCA dataset, for which results are shown here. The BLCA dataset has a binary decision variable and contains 38404 continuous variables describing RNA-seq gene expression for 476 patients. The method has been validated in rigorous external cross-validation. The stability results based on well-established metrics [1, 2] are presented in Table 1.

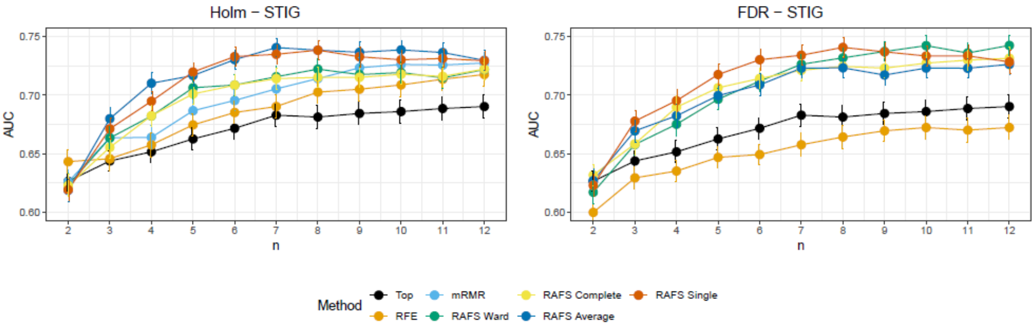


Figure 1: Random forest AUC results for different numbers (n) of variables taken for RAFS with STIG and other established methods. The upper plot is for Holm correction (level 0.05) in the pre-filtering stage, and the lower - FDR (level 0.10).

RAFS outperforms the alternatives by a considerable margin. Similarly, it outperforms the alternatives in terms of classification accuracy. We postulate both aspects are necessary to gain the correct insight. Noticeably, the mRMR method, while being the best alternative in terms of stability, is worst in terms of classification accuracy with FDR pre-filtering (its line is below the plot threshold of 0.6).

References:

[1] Lustgarten, J.L., et al.: Measuring stability of feature selection in biomedical datasets. In: AMIA annual symposium proceedings. vol. 2009, p. 406. Am. Med. Inf. Assoc. (2009).

[2] Wang, X., et al.: Optimal consistency in microrna expression analysis using reference-gene-based normalization. Molecular BioSystems 11(5), 1235-1240 (2015)

[3] Yang, Q., et al.: Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data. Briefings in Bioinformatics 21(3), 1058-1068 (2020).

Keywords: Insightfulness :: Feature selection :: Robustness :: Stability :: Information theory :: Hierarchical clustering :: Cross-validation :: Result aggregation

Granular Spectrum of Covering Rough Sets

Piotr Wasilewski[1,2], Dominik Ślęzak[3,4,5]

[1]Systems Research Institute, Polish Academy of Sciences, ul. Newelska 6, 01-447 Warsaw, Poland

[2]Faculty of Computer Science, Dalhousie University, 6050 University Avenue, Halifax, Nova Scotia, Canada

[3]Institute of Informatics, University of Warsaw, ul. Banacha 2, 02-097 Warsaw, Poland

[4]QED Software Sp. z o.o., ul. Miedziana 3A m. 18, 00-814 Warsaw, Poland

[5]DeepSeas USA / Poland, ul. Aleje Jerozolimskie 123A, 02-017 Warsaw, Poland

Extended Abstract. We introduce granular spectrum of the object space presented in Figure 1 which is determined by granular approximation operators based on arbitrary coverings of the object space. These operators were introduced in [1] and are presented below:

$$G_{\forall}(X) := \{a \in U : \forall A \in Gr(a) A \subseteq X\} \quad G^{\exists}(X) := \{a \in U : \exists A \in Gr(a) A \cap X \neq \emptyset\}.$$

$$G_{\exists}(X) := \{a \in U : \exists A \in Gr(a) A \subseteq X\} \quad G^{\forall}(X) := \{a \in U : \forall A \in Gr(a)\ A \cap X \neq \emptyset\}.$$

where $Gr(a) := \{A \in Gr(U) : a \in A\}$ and $Gr(U)$ is a family of granules covering space U. One can note that pairs of operators presented above are pairs of dual operators, for example $G_{\forall}(X)^{\complement} = G^{\exists}(X^{\complement})$. First two operators were investigated in [2] where operator $G^{\exists}$ was presented in a new equivalent form based on biting procedure. All of four operators presented above were reintroduced and developed in [3].

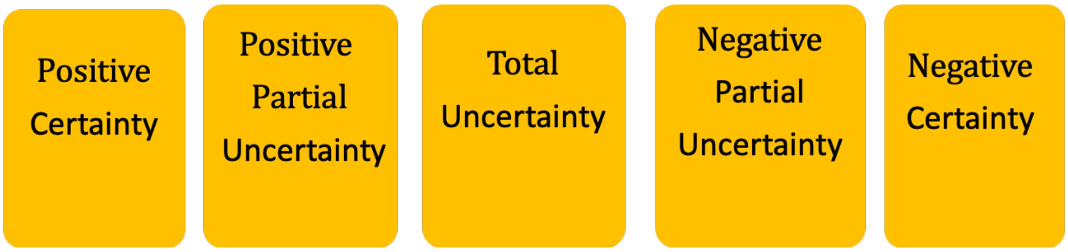


Figure 1: Granular spectrum of space $(U, Gr(U))$ determined by subset $X \subseteq U$.

The granular spectrum presented in this paper consists of five elements which take the

specific forms for every subset X of the object space in the following way: the positive certainty region takes the form of $G_{\forall}(X)$, the positive partial uncertainty takes the form of $G_{\exists}(X) \setminus G_{\forall}(X)$. Analogously the negative certainty region and the negative partial uncertainty region take the forms of $G_{\forall}(X^{\mathsf{C}})$ and $G_{\exists}(X^{\mathsf{C}}) \setminus G_{\forall}(X^{\mathsf{C}})$ respectively and finally the total uncertainty region takes the form of $U \setminus [G_{\forall}(X) \cup G_{\forall}(X^{\mathsf{C}})]$.
We investigate the presented approximation operators without posing any conditions on the nature of granules. We also discuss some earlier definitions of rough set approximations known from the literature [4-8] and we show in what sense they are special cases of our framework.

References:

[1] Yao, Y. Y.: On Generalizing Rough Set Theory. In: Lecture Notes in Artificial Intelligence 2639, 44–51. Springer, Heidelberg (2003).

[2] Ślęzak, D., Wasilewski, P.: Granular sets - foundations and case study of tolerance spaces. In: Lecture Notes in Artificial Intelligence 4482, 435–442. Springer, Heidelberg (2007)

[3] Yao, Y. Y., Yao, B.: Covering based rough set approximations. Information Sciences **200**, 91–107 (2012)

[4] Pawlak, Z.: Rough sets. International Journal of Computing and Information Sciences. **18**, 341–356 (1982).

[5] Lin, T. Y.: Topological and Fuzzy Rough Sets. In: Roman Słowiński (Ed.), Intelligent Decision Support, pp. 287-304, Springer (1992).

[6] Lin, T. Y., Liu, Q.: Rough Approximate Operators: Axiomatic Rough Set Theory. In: Ziarko, W. P. (ed.), Rough Sets, Fuzzy Sets and Knowledge Discovery. pp 256–260, Springer (1994).

[7] Skowron, A., Stepaniuk, J.: Tolerance approximation spaces. Fundamenta Informaticae **27**, 245–253 (1996).

[8] Z. Pawlak, Some Issues on Rough Sets, Transactions on Rough Sets I, Journal Subline, Lectures Notes in Computer Science 3100, 1–58 (2004)

Keywords: Rough sets :: Tolerance rough sets :: Granular sets :: Covering-based rough sets :: Coverings :: Tolerance relations

Symmetry of Attribute Decompositions Based on Generalized Decision Functions

Dominik Ślęzak[1,2,3]

[1]Institute of Informatics, University of Warsaw, ul. Banacha 2, 02-097 Warsaw, Poland

[2]QED Software Sp. z o.o., ul. Miedziana 3A, 00-814 Warsaw, Poland, www.qed.pl

[3]DeepSeas USA / Poland, ul. Aleje Jerozolimskie 123A, 02-017 Warsaw, Poland, www.deepseas.com

Extended Abstract. We continue our research on generalized decision functions [1]. Our previous findings in this area were summarized in [2]. One of the topics recalled therein was the task of building the ensembles of complementary reducts. This task can be also referred as the attribute decomposition problem and it was analyzed earlier in [3]. Generalized decision functions were originally introduced for decision tables $\mathbb{A} = (U, A \cup \{d\})$ consisting of categorical attributes $A \cup \{d\}$, where $d \notin A$ is a distinguished decision attribute. Formally, for a given subset of attributes $B \subseteq A$, the generalized decision function ∂_B assigns each object $u \in U$ with a set of decision values occurring for all objects which are indistinguishable (indiscernible) from u subject to its values on B. Let us denote by $B(u)$ a vector of values of attributes $a \in B$ occurring for u. Then, for each $u \in U$, we can express $\partial_B(u)$ as $\partial_B(u) = \{d(u') : u' \in U, B(u') = B(u)\}$.

Generalized decision functions are closely related to rough set approximations of decision classes in decision tables. They can be used to express non-determinism of the rough-set-based decision models and they are helpful to compare the theory of rough sets with other approaches to uncertainty management [4]. To emphasize the meaning of the distinguished attribute d, we can write $\partial_{d|B}$ instead of ∂_B. We can also consider decision tables $\mathbb{A} = (U, A \cup D)$ with multiple decision attributes $d \in D$ and redefine generalized decision functions as collecting the sets of possible (for objects indistinguishable from u) vectors of decision values, i.e. $\partial_{D|B}(u) = \{D(u') : u' \in U, B(u') = B(u)\}$.

Now, let us consider two subsets of attributes $B1, B2 \subseteq A$, such that $B1 \cup B2 = A$. When we consider generalized decision functions induced by each of those subsets separately, we can observe inclusions $\partial_{D|B1}(u) \supseteq \partial_{D|A}(u)$ and $\partial_{D|B2}(u) \supseteq \partial_{D|A}(u)$. We can surely observe also inclusions of the form $\partial_{D|B1}(u) \cap \partial_{D|B2}(u) \supseteq \partial_{D|A}(u)$. Any strict inclusion of this kind means that we lose (some part of) ability to reason deterministically about objects $u \in U$ when using smaller subsets of attributes. However, if we choose the subsets $B1, B2 \subseteq A$ in such a way that the equality $\partial_{D|B1}(u) \cap \partial_{D|B2}(u) = \partial_{D|A}(u)$ is satisfied for each $u \in U$, then we can say that the generalized decision model based on $\partial_{D|A}$ can be *decomposed* onto two simpler models based on $\partial_{D|B1}$ and $\partial_{D|B2}$, which are complementary to each other, i.e., those simpler models may lose some information locally but they provide the same information as $\partial_{D|A}$ when combined together.

In this short paper, we go a bit further and we study generalized decision functions for information tables (information systems) $\mathbb{A} = (U, A)$ without distinguished decision attributes, whereby for any $X, Y \subseteq A$ we can put:

$$\partial_{X|Y}(u) = \{X(u') : u' \in U, Y(u') = Y(u)\} \tag{1}$$

Accordingly, for any $X, Y, Z \subseteq A$, we can define the criterion

$$\partial_{X|Y \cup Z}(u) = \partial_{X|Y}(u) \cap \partial_{X|Z}(u) \quad \forall u \in U \tag{2}$$

which means that the generalized decision model that reasons about the vectors of values of attributes in X based on the values of attributes in $Y \cup Z$ can be *decomposed* without a loss of information (determinism) onto two simpler models based on Y and Z separately. In particular, let us notice that the subsets $X, Y, Z \subseteq A$ can overlap with each other. For instance, if $X \cap Y \neq \emptyset$, then all vectors of values of attributes in X which belong to $\partial_{X|Y}(u)$ are naturally the same when projected onto $X \cap Y$.

The goal of this paper is now to draw the reader's attention to a surprising (at least for the author) mathematical property of such understood criterion for generalized decision function decomposition. Namely, the following is satisfied:

Proposition 1. For any $\mathbb{A} = (U, A)$, for any $X, Y, Z \subseteq A$, the following statements are equivalent to each other:

$$\begin{aligned}
\partial_{X|Y\cup Z}(u) &= \partial_{X|Y}(u) \cap \partial_{X|Z}(u) \quad \forall u \in U\\
\partial_{Y|X\cup Z}(u) &= \partial_{Y|X}(u) \cap \partial_{Y|Z}(u) \quad \forall u \in U\\
\partial_{Z|X\cup Y}(u) &= \partial_{Z|X}(u) \cap \partial_{Z|Y}(u) \quad \forall u \in U
\end{aligned}$$

The proof of this kind of symmetry will be shown at the conference presentation.

References:

[1] Pawlak, Z., Skowron, A. (2007). Rudiments of Rough Sets. Information Sciences, 177(1), 3-27.

[2] Ślęzak, D. (2015). On Generalized Decision Functions: Reducts, Networks and Ensembles. In Y. Yao, Q. Hu, H. Yu, J. W. Grzymała-Busse (Eds.), Proceedings of RSFDGrC 2015 (Vol. 9437, pp. 13-23). Springer.

[3] Ślęzak, D. (1999). Decomposition and Synthesis of Decision Tables with Respect to Generalized Decision Functions. In S. K. Pal, A. Skowron (Eds.), Rough Fuzzy Hybridization - A New Trend in Decision Making (pp. 110-135). Springer.

[4] Skowron, A., Grzymała-Busse, J. W. (1994). From Rough Set Theory to Evidence Theory. In R. R. Yager, J. Kacprzyk, M. Fedrizzi (Eds.), Advances in the Dempster-Shafer Theory of Evidence (pp. 193-236). Wiley.

Keywords: No keywords

Rough Set Theory as Guiding Heuristics in IT System Diagnosis

András Földvári[1],András Pataricza[1]

[1]Department of Measurement and Information Systems

Budapest University of Technology and Economics, Budapest, Hungary

Extended Abstract. Requirements against modern IT-based services increasingly demand high reliability, integrity, and high availability of the systems. At the same time, the ever-increasing structural and functional complexity poses complicated challenges in diagnostics, i.e., the detection, logical, and physical localization of accidental or intentional faults.

A significant difference between the diagnostic classical industrial systems and IT systems originates in how faults occur and manifest as failures. In classical industrial systems, the propagation mechanism of fault effects is relatively simple, so a single or few abstraction-level deep system model is typically sufficient to underpin the diagnostic process. In IT systems, however, faults occur at low levels (e.g., hidden software bugs, possible faults in external services, or transient hardware faults).

- Their *error propagation path* is long due to the system's complexity and observable manifestation as failures occur at the end of the path in services delivered to the user (in monitored systems, a few monitoring agents improve observability). The aspects corrupted by the propagation of errors may change along the error propagation path (missing data input leads to an output integrity error).

- In addition, many components' internal structure and operation are *unknown* or cannot be monitored with a realistic effort.

- Due to the high frequency of operation, *non-determinism* due to data dependency of operations, etc., a significant part of faults occur as *rare events*. Hence, the diagnostic problem becomes a rare event root-cause analysis task in *big data* streams to be addressed at several levels of abstraction.

- On the other hand, limiting the *diagnostic resolution* to the level of repair/replacement units (FRU) is as restoring system operability takes priority over full-detail fault isolation.

Rough Set Theory (RST) offers a mathematical paradigm for system diagnosis (SD) by meaningful insights from complex data sets. RST operates on the principle of identifying discernible patterns within decision tables, wherein each column corresponds to an observed *syndrome* in the system, and the decision variable represents an FRU-level fault mode. RST is a favorite candidate to manage diagnostic uncertainties from the abovementioned factors.

Model-based RST-based granular SD aims to achieve diagnostic resolution down to the FRU level (discernibility of FRU fault modes) to simplify fault mitigation. Applying the RST approximation under an anticipated fault mode(s) consistent with the syndrome, the following outcomes can be obtained: i) The boundary region is empty as only objects related to faults are present within the positive region. The diagnosis is complete and perfect for single or multiple (indistinguishable) faults in the set; or ii) the boundary region is nonempty, indicating uncertainty in the fault mode to syndrome mapping, thus necessitating a more precise evaluation after model refinement. This refinement can occur in two ways: through value refinement (more accurate acquisition for representation of the observations) and by introducing new attributes (state refinement). Model refinement raises the question of data representation in the system model. RST-based SD involves discretizing attributes [1] to work with discrete models by aggregating continuous attributes to discrete values, while the discretization must preserve the essential characteristics for the diagnostic process.

While phenomenological discretization (commonly used in RST) is only based on actual observations (e.g., measuring the speed of a car), discretization for technical diagnosis must preserve system and requirement-specific domain knowledge (e.g., overspeeding limit). We propose using qualitative modeling, common in the engineering practice, for discretization by distinguishing the different value domains in the system operations and requirements (e.g., Safety Integrity Levels [2] in critical applications). Still, it can introduce non-determinism in the faulty case, resulting in modeling ambiguity, which leads to diagnostic uncertainty.

RST, by its very nature, is a perfect guiding paradigm for managing diagnostic uncertainty. Highlighting and measuring indiscernibility is an exact counterpart of diagnostic uncertainty. This way, methods for reducing it in RST exactly map to model refinement-based sequential diagnosis approaches. Moreover, the core and derived notions in RST have a straightforward interpretation and explanation in terms of technical diagnosis, thus making this powerful mathematical theory close to engineering thinking.

We propose using iterative RST-supported diagnostic model refinement, which assists in determining the appropriate level of abstraction and handling modeling uncertainties using approximations and quality metrics (e.g., accuracy). It is further supported by the qualitative modeling perspective to aid the discretization process, enhancing diagnosis accuracy.

Acknowledgement: This work was funded by the PrOTectME (EIT Digital 21293), and the DigitalTech EDIH (101083965) 2021-1.2.1-EIT-KIC-2021-00006 projects.

References:

[1] Shen, L., Tay, F. E., Qu, L., Shen, Y. (2000). Fault diagnosis using Rough Sets Theory. Computers in Industry, 43(1), 61-72.

[2] IEC SC 65A. Functional safety of electrical/electronic/programmable electronic safety-related systems (Techreport IEC 61508). The International Electrotechnical Commission.

Keywords: No keywords

A note on NP-hardness of deriving the simplest ensembles of rough-set-based decision models from the data – the cases of decision bireducts and generalized decisions

Dominik Ślęzak[1,2,3]

[1]Institute of Informatics, University of Warsaw, ul. Banacha 2, 02-097 Warsaw, Poland

[2]QED Software Sp. z o.o., ul. Miedziana 3A, 00-814 Warsaw, Poland, www.qed.pl

[3]DeepSeas USA / Poland, ul. Aleje Jerozolimskie 123A, 02-017 Warsaw, Poland, www.deepseas.com

Extended Abstract. *Decision bireducts* provide the means for learning simple decision models from the data; as such, they are often considered as an alternative to better known *decision reducts* originating from the theory of rough sets [1]. Moreover, the ensembles of decision bireducts turn out to be useful in many real-world application areas, as providing easily interpretable classification models and insightful attribute rankings [2].

Let us recall that, for a decision table $\mathbb{A} = (U, A \cup \{d\})$, the pair (X, B), $X \subseteq U$, $B \subseteq A$, is called a decision bireduct, if and only if the values of attributes in B determine the values of the decision attribute d when $\mathbb{A}$ is limited to X, whereby B cannot be reduced and X cannot be extended without losing that kind of determination [3]. It is also known that such pairs (X, B) correspond to the collections of decision rules with their left parts based on the values of attributes in B and their supports summing up to X [1].

In [4], we studied the problem of deriving the simplest ensembles of bireducts from the data. We considered an ensemble of m decision bireducts $(X_1, B_1), ..., (X_m, B_m)$ to be *valid* for $\mathbb{A} = (U, A \cup \{d\})$, if and only if the decision value of each object $u \in U$ was correctly recognized by decision rules corresponding to more than $m/2$ out of bireducts. Then, we showed that it is NP-hard to search for such *valid* ensembles of decision bireducts that minimize the cardinality of the biggest B_i out of $B_1, ..., B_m \subseteq A$.

An analogous problem can be specified for so-called *generalized decision reducts* [5] – the irreducible subsets of attributes $B \subseteq A$ which induce the same generalized decision functions $\partial_B : U \to 2^{V_d}$ (where V_d denotes the set of all values of decision attribute d

which occur in $\mathbb{A}$), $\partial_B(u) = \{d(u') : B(u') = B(u)\}$ (where $d(\cdot)$ and $B(\cdot)$ denote the values and vectors of values occurring on objects in U, respectively) as the whole sets of attributes, i.e. $\partial_B(u) = \partial_A(u)$ for each $u \in U$. Herein, a generalized decision reduct $B \subseteq A$ can be interpreted as corresponding to a collection of non-deterministic decision rules which – for each combination of attributes in B – point at a set of decision values which are possible for objects for which we can observe that combination.

A *generalized decision ensemble* $B_1, ..., B_m \subseteq A$ can be regarded as *valid*, if and only if the equality $\partial_{B_1}(u) \cap ... \cap \partial_{B_m}(u) = \partial_A(u)$ holds for each $u \in U$. This kind of validity means that although decision rules induced by combinations of values observed on particular subsets B_i, $i = 1, ..., m$, can point at the decision value sets that are bigger than in the case of ∂_A (i.e. each B_i – when considered separately – may not satisfy the conditions for being a generalized decision reduct), they become to work exactly like ∂_A when combined together (i.e. even if some $\partial_{B_i}(u)$ points at a possible decision value outside $\partial_A(u)$, then the other $\partial_{B_j}(u)$, $i \neq j$, eliminates that unwanted possibility).

In [6], we stated that – just like in the case of decision bireducts in [4] – the problem of deriving the simplest valid generalized decision ensembles from the data is NP-hard. Again, we used the same interpretation of simplicity, i.e., for a given ensemble of subsets $B_1, ..., B_m \subseteq A$, we looked at the maximum cardinality $|B_i|$, $i = 1, ..., m$, and we tried to minimize it. However, in this short paper, we study a different version of simplicity which refers to the overall *size* of decision rules induced by the generalized decision ensembles, as well as the ensembles of decision bireducts. That size can be measured as the sum of lengths of all decision rules that correspond to the given ensemble.

It turns out that the problems of deriving the simplest – which now means the minimum size – generalized decision ensembles and ensembles of decision bireducts are NP-hard as well, just like in the case of analogous problems investigated in [4,6]. The proofs of this kind of NP-hardness will be shown at the conference presentation.

References:

[1] Stawicki, S., Ślęzak, D., Janusz, A., Widz, S. (2017). Decision Bireducts and Decision Reducts - A Comparison. International Journal of Approximate Reasoning,

84, 75-109.

[2] Janusz, A., Ślęzak, D., Stawicki, S., Stencel, K. (2023). A Practical Study of Methods for Deriving Insightful Attribute Importance Rankings Using Decision Bireducts. Information Sciences, 645, 119354.

[3] Ślęzak, D., Janusz, A. (2011). Ensembles of Bireducts: Towards Robust Classification and Simple Representation. In T.-H. Kim, H. Adeli, D. Slezak, F. E. Sandnes, X. Song, K.-I. Chung, K. P. Arnett (Eds.), Proceedings of FGIT 2011 (Vol. 7105, pp. 64-77). Springer.

[4] Ślęzak, D., Stawicki, S. (2020). The Problem of Finding the Simplest Classifier Ensemble is NP-Hard - A Rough-Set-Inspired Formulation Based on Decision Bireducts. In R. Bello, D. Miao, R. Falcon, M. Nakata, A. Rosete, D. Ciucci (Eds.), Proceedings of IJCRS 2020 (Vol. 12179, pp. 204-212). Springer.

[5] Pawlak, Z., Skowron, A. (2007). Rudiments of Rough Sets. Information Sciences, 177(1), 3-27.

[6] Ślęzak, D. (2015). On Generalized Decision Functions: Reducts, Networks and Ensembles. In Y. Yao, Q. Hu, H. Yu, J. W. GrzymaÅĆa-Busse (Eds.), Proceedings of RSFDGrC 2015 (Vol. 9437, pp. 13-23). Springer.

Keywords: No keywords

Interrelationship between Discernibility and Monotonicity in Attribute Reduction

Dominik Ślęzak[1,2,3], Soma Dutta[4]

[1]Institute of Informatics, University of Warsaw, Poland, ul. Banacha, 02-097 Warsaw, Poland

[2]QED Software Sp. z o.o., ul. Miedziana 3A/18, 00-814 Warsaw, Poland, https://qed.pl

[3]DeepSeas, USA/Poland, https://www.deepseas.com/

[4]Department of Mathematics and Informatics, University of Warmia and Mazury

ul. Słoneczna 54, 10-710 Olsztyn, Poland

Extended Abstract. We aim to explore two very important properties of rough-set-inspired [1,2] approaches dealing with elimination of redundant attributes during the feature selection processes [3]. Usually, these approaches assume that information about decision values, as present in the original decision table, should be kept unchanged during such elimination processes. The information about the decision attribute from a decision system can be represented by various decision valuations [4,5], e.g., rough membership function, generalized decision valuation etc. Such representations of decision information are induced based on the particular clusters of objects having the same values on the considered set of conditional attributes; these are the equivalence classes of objects, i.e., the so-called indiscernibility classes.

When an attribute is removed from the considered set of attributes, then some of those classes, generated based on the whole set of attributes, are merged with each other. Given that a decision information about a decision table $\mathbb{A} = (U, A \cup \{d\})$ is modeled by a decision valuation ϕ, a subset B of attributes is said to keep the same information about decision as the whole set A ($B \subseteq A$), if for each object $u \in U$ in the considered decision table, its indiscernibility class with respect to B, denoted by $[u]_B$, induces the same decision information $\phi([u]_B)$ as in case of A (i.e., $\phi([u]_B) = \phi([u]_A)$). Following the rough set terminology, such subsets $B \subseteq A$ can be called ϕ-superreducts. Now while looking for such reduced subsets of attributes there are two properties that are commonly checked.

The first of them states that if B is not a ϕ-superreduct, then neither of its subsets is. This property – often called the monotonicity – is important to design efficient algorithms for finding ϕ-superreducts in large data sets. However, one needs to remember that not all decision valuations ϕ satisfies it. This property is equivalent to the weak union property from the theory of semigraphoid [6]. For instance, for a decision valuation ϕ the conditional independence statement, denoted by $I_\phi(d|B|A \setminus B)$, means $\phi([u]_B) = \phi([u]_A)$ for all $u \in U$. The weak union property imposes that $I_\phi(d|X|Y \cup W) = I_\phi(d|X \cup W|Y)$ where $X, Y, W \subseteq A$. That is, the claim that the decision attribute d is independent of a set of attributes $Y \cup W$ in presence of another set of attributes X is the same as ensuring that d remains independent of any subset of $Y \cup W$ when the removed attributes are added to the set X.

The second property refers to the idea of replacing the condition (*) $\forall_{u \in U}\ \phi([u]_B) = \phi([u]_A)$ with its alternative form (**) $\forall_{u,u' \in U}\ \phi([u]_A) \neq \phi([u']_A) \Rightarrow [u]_B \neq [u']_B$. Such form lets us redefine the original decision attribute d as the new one – interpreted as ϕ_d – and utilize powerful Boolean-reasoning-based algorithms to search for ϕ-superreducts. For example, if we consider the generalized decision valuation ∂ then the original decision attribute d can be translated to a new decision attribute ∂_d where $\partial_d([u]_A)$ collects all decision values incurred in the equivalence class.

However, as in case of the monotonicity, only for some of the decision valuations ϕ criteria (*) and (**) are equivalent to each other. We refer to this kind of equivalence as the discernibility property. In this paper, we show that for all decision valuations ϕ, which satisfy the discernibility property, the monotonicity property is satisfied as well. Moreover, we show that these two general properties are not equivalent to each other, i.e., there are functions ϕ which satisfy the monotonicity property but do not satisfy the discernibility property.

References:

[1] Pawlak, Z., Skowron, A. (2007). Rudiments of Rough Sets. Information Sciences, 177(1), 3-27.

[2] Pawlak, Z. (1982). Rough Sets. International Journal of Computer and Information

Sciences, 11(5), 341-356.

[3] Chandrashekar, G., Sahin, F. (2014). A Survey on Feature Selection Methods. Computers & Electrical Engineering, 40(1), 16-28.

[4] Ślęzak, D., Dutta, S. (2018). Dynamic and Discernibility Characteristics of Different Attribute Reduction Criteria. Proceedings of IJCRS 2018, 628-643.

[5] Dutta, S., Ślęzak, D. (2024). Nature of Decision Valuations in Elimination of Redundant Attributes. International Journal of Approximate Reasoning.

[6] Pearl, J., Paz, A. (1987). Graphoids: Graph-based Logic for Reasoning about Relevance Relations. In B. D. Boulay, D. Hogg, L. Steele (Eds.), Advances in Artificial Intelligence-II (pp. 357-367). North-Holland.

Keywords: Feature Selection :: Attribute Reduction :: Monotonicity of Attribute Reduction :: Discernibility in Attribute Reduction :: Valuations :: Conditional Independence :: Rough Membership Function :: Generalized Decision Function

Deriving Volumetric Anomalies from Huge Log Event Data Sets using Approximate SQL Engine

Janusz Borkowski[1], Agnieszka Chądzyńska-Krasowska[2,1], Joel Holland[1], Marcin Kowalski[1], Dominik Ślęzak[3,4,1], Piotr Synak[1], Arkadiusz Wojna[1], Jakub Wróblewski[1]

[1]DeepSeas, USA / Poland / Switzerland

[2]Polish-Japanese Academy of Information Technology, Warsaw, Poland

[3]Institute of Informatics, University of Warsaw, Warsaw, Poland

[4]QED Software, Warsaw, Poland

Extended Abstract. Volumetric anomalies are one of the most important indicators in the cybersecurity analytics [1,2]. In the practical scenario whereby the network event logs are parsed prior to inserting them into a relational database, volumetric anomalies usually take a form of conjunctions of attribute-value conditions over the subsets of data columns, wherein it is expected that the volumes of events matching with those conjunctions of conditions are much higher *currently* than it could be observed *in the past.*

The algorithms designed to search for such volumetric anomalies face a number of interesting challenges, e.g.: How to explore a "lattice" of possible conjunctions of conditions? How to assess that the *current* volume is "much higher" than the volume *in the past*? How to specify the *current* and the *past* time periods? In particular, there is also a computational challenge related to calculation of those volumes against huge data sets, especially when it comes to the historical event logs.

In this presentation, we report our experiences with real-time detection of volumetric anomalies within the data sets of event logs of the size of tens of terabytes. For this purpose we utilize our approximate relational database engine, which enables us to store the original data in a compacted / compressed / summarized form, so the SQL operations are potentially inexact but very fast [3]. In particular, it is interesting to note that the design of this engine is based on the theory of rough sets. Moreover, it is interesting to discuss to what extent some potential inaccuracies in the SQL query results can really influence the process of volumentric anomaly detection.

References:

[1] Kołodziej, J., Krzysztoń, M., Szynkiewicz, P. (2023). Anomaly Detection in TCP/IP Networks. Proceedings of ECMS 2023, 542-548.

[2] Siris, V. A., Papagalou, F. (2006). Application of Anomaly Detection Algorithms for Detecting SYN Flooding Attacks. Computer Communications, 29(9), 1433-1442.

[3] Ślęzak, D., Glick, R., Betliński, P., Synak, P. (2018). A New Approximate Query Engine Based on Intelligent Capture and Fast Transformations of Granulated Data Summaries. Journal of Intelligent Information Systems, 50(2), 385-414.

Keywords: No keywords

Using Rough Sets for Assessment of Uncertain Properties of Ring Gears after High Pressure Gas Quenching

Łukasz Rauch[1], Krzysztof Bzowski[1], Maciej Pietrzyk[1], Jakub Łazarski[1], Ivan Milenin[1]

[1]AGH University of Krakow, Poland

Extended Abstract. The main objective of the work was to solve the problem of long last-ing computationally intensive multiscale numerical models for simulations of industrial processes. The presentation will contain the industrial process idea and parameters, complex fully coupled numerical model and propositions of further improvements.

1. Introduction and research methodology

In the aerospace industry, processes are defined by stringent requirements for product quality and reliability. In the gear manufacturing process, it is important to control the phase composition and microstructure of the surface layer. This layer is crucial for mechanical properties such as wear resistance and fatigue strength of gear rings. Thermochemical treatment processes, especially carburising, play a fundamental role in manufacturing. These processes lead to an evolution of the microstructure, which significantly affects the final properties of the material. Advances in technology, new steels such as Pyrowear 53, and vacuum carburisation, allow gas quenching to be carried out in a single furnace chamber immediately after the carburisation cycle. Pyrowear 53 steel was used for the experimental part of the study for identification of model parameters. The study used the Johnson-Mehl-Avrami-Kolmogorov (JMAK) model, which was upgraded for the simulation of transformations in Pyrowear 53 steel. A numerical cooling model based on Computational Fluid Dynamics (CFD) analysis was developed using Abaqus software and fully integrated with microscale models. Improvement of computational complexity of a fully coupled CFD model was the main objective of the work.

2. Results

The most important result for the described approach is the effect of phase transformations on the obtained temperature distributions (see Fig. 1) and final distribution of the material properties in the product, especially stress distribution influencing cracking of the produced part during exploitation.

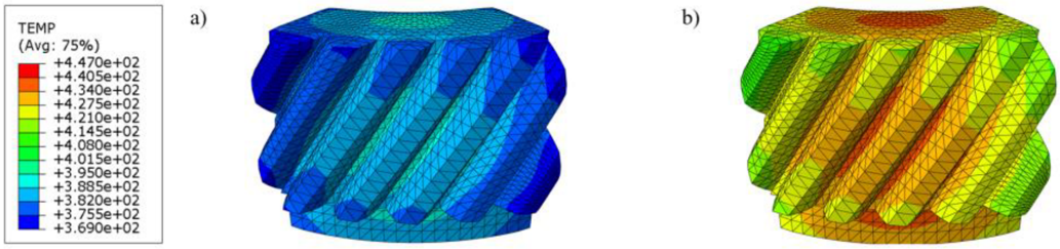


Figure 1: Temperature distribution on the surface of a gear wheel section at time $t = 360s$: without consideration of phase transformations (a), with consideration of phase transformations (b).

3. Discussion and conclusions

The critical cooling rate for core samples with nominal chemical composition was approximately $3°C/s$. For cooling rates below $0.02°C/s$, a purely ferritic microstructure was predicted. The critical cooling rate for the samples after carburising was about $0.1°C/s$. At slow cooling, pearlite appeared in the microstructure, but martensite was observed over the entire range of the investigated cooling rates. Multiscale simulations showed that cooling at a gas inlet velocity of $9m/s$ produced a purely martensitic microstructure, which was crucial for obtaining assumed product properties. During the presentation of the work influence of the numerical assumptions, especially meshing procedure, on the reliability of the final results will be discussed. One of the most important issues in this case was meshing of the carburized layer which in comparison to the core material is very thin and influences the computational cost the most. Application of the rough sets for modelling of material properties, which dependently on the meshing density can be assessed with uncertainty, can improve computational efficiency and quality of the assessment. The possibility of rough sets application will be discussed during the presentation. Additionally, the replacement of computationally intensive CFD model with Finite Element Modelling (FEM) of heat transfer with convection will be also discussed.

Acknowledgement: The financial support of National Centre of Research and Development, project no. TECHMATSTRATEG2/406725/NCBR/2020, is acknowledged.

References:

[1] Rauch, Ł., Zalecki, W., Kuziak, R., Garbarz, B., Raga, K., Bzowski, K., Pietrzyk, M.: Numerical simulations of aircraft engine ring gears quenching by using mean field model of phase transformations in Pyroware Steel 53. Journal of Aerospace Engineering 36(6), (2023)

Keywords: Numerical simulations :: Modelling with uncertainty

Authors Index

Keywords Index