Hindawi
Complexity
Volume 2021, Article ID 7815418, 9 pages
https://doi.org/10.1155/2021/7815418

Research Article

Rough Set Approach toward Data Modelling and User Knowledge for Extracting Insights

Xiaoqun Liao,[1] Shah Nazir,[2] Junxin Shen,[3] Bingliang Shen,[3] and Sulaiman Khan[2]

[1]*Information and Network Center, Xi'an University of Science and Technology, Xi'an 710054, China*
[2]*Department of Computer Science, University of Swabi, Swabi, Pakistan*
[3]*Faculty of Management and Economics, Kunming University of Science and Technology, Kunming, Yunnan 650093, China*

Correspondence should be addressed to Shah Nazir; snshahnzr@gmail.com

Received 19 July 2020; Revised 18 August 2020; Accepted 27 August 2020; Published 27 January 2021

Academic Editor: M. Irfan Uddin

Information is considered to be the major part of an organization. With the enhancement of technology, the knowledge level is increasing with the passage of time. This increase of information is in volume, velocity, and variety. Extracting meaningful insights is the dire need of an individual from such information and knowledge. Visualization is a key tool and has become one of the most significant platforms for interpreting, extracting, and communicating information. The current study is an endeavour toward data modelling and user knowledge by using a rough set approach for extracting meaningful insights. The technique has used different rough set algorithms such as K-nearest neighbours (KNN), decision rules (DR), decomposition tree (DT), and local transfer function classifier (LTF-C) for an experimental setup. The approach has found its accuracy for the optimal use of data modelling and user knowledge. The experimental setup of the proposed method is validated by using the dataset available in the UCI web repository. Results of the proposed study show that the model is effective and efficient with an accuracy of 96% for KNN, 87% for decision rules, 91% for decision trees, 85.04% for cross validation architecture, and 94.3% for local transfer function classifier. The validity of the proposed classification algorithms is tested using different performance metrics such as F-score, precision, accuracy, recall, specificity, and misclassification rates. For all these performance metrics, the KNN classifier outperformed, and this high performance shows the applicability of the KNN classifier in the proposed problem.

1. Introduction

With the passage of time, the information and user knowledge become increasing. This is due to the advancements and rapid development in technology. Essential information has become the need of users in their daily life which requires the support of advanced tools like Hadoop, Tableau, Informatica PowerCenter, and so on. The data and knowledge exist in diverse shapes such as structured and unstructured. The structured data are mostly easily understandable and can be managed, while extracting meaningful insights from unstructured data has become a challenging task. According to the report of IDC [1], in late 2011 about 1.8 ZB of data were created. Globally, electronic data of approximately 1.2 ZB (10^{21}) are generated per year by diverse sources [2]. By 2020, 40 ZB data are expected [3]. Human beings are always interested to capture the knowledge in an easy and effective way. This easiness is due to the translation of data and knowledge through graphs or maps for user understanding.

The role of visual context is obvious through which the patterns are identified from huge bulk of data and can be transformed through graphics and visualizations. Conclusions are drawn from the data through collection of data, modelling of data, and processing of data finally to plot the derivations. From interrelated perspectives, the data, knowledge, and information are mostly used in visualization. The aim of visualization is to gain meaningful insights from the data [4]. Users can interact with the data using the techniques of visualizations and go for analysis of

the data and knowledge. One can communicate through data visualization in an effective and easy way for easy transferring of message and technical drawing for scientific purposes.

In order to support the data modelling and user knowledge, the proposed research contribution is to use a rough set approach toward user knowledge and data modelling for extracting insights. Different algorithms of rough set such as KNN, decision rules, decomposition tree, and LTF-C were used for the experimental setup. The dataset was used for the experimental setup of the proposed method which is available in the UCI web repository [5]. KNN is suggested in different problems such as text recognition [6].

The organization of the paper is as follows. Section 2 represents the related work to user knowledge, data modelling, and visualization with different literatures. Section 3 shows the research method and modelling of the proposed study with the detail of visualization of the dataset. Section 4 gives the results and discussion. The paper is concluded in section 5.

2. Related Work

Researchers are trying to use different approaches, tools, and techniques in order to analyze user knowledge, data modelling, and visualization. Table 1 shows the brief descriptions of the existing approaches available in the literature.

3. Rough Set Approach toward Data Modelling and User Knowledge for Extracting Insights

Machine learning algorithms play an important role in different areas of research [16, 25–30]. In this paper, a rough set approach is used for data modelling and user knowledge to extract meaningful insights. The rough set approach works well in a situation of uncertainty by plotting the lower and upper approximations. The obtainable model or rough set consists of "IF THEN rules." The rough set was presented by Pawlak in 1982 [31]. It has a specific lower and upper approximation boundary area. Lowering the degree of precision in the data makes the data pattern more clearly. Rough sets and boundaries can be mathematically presented as follows [32]:

$$\begin{aligned} \overline{BX} &= \{X_i \in U \mid [X_i] \text{ Ind } (B) \cap x \neq 0\}, \\ \underline{BX} &= \{X_i \in U \mid [X_i] \text{ Ind } (B) \subset x\}. \end{aligned} \tag{1}$$

It shows two possibilities: the element belonging to the set and the element possibly belonging to the set. Figure 1 shows the concept of rough set.

Figure 2 represents the workflow of the rough set theory application. The main parts in the workflow are explained in this section.

The experimental process of the above flow shown in Figure 2 has been implemented using RSES [33]. Rough set and fuzzy rough set theories are based on some preliminary parts [34]. The reason behind the selection of the rough set approach for the proposed research is that it works very well in situations of uncertainty and vagueness. The following main parts were considered for the experimental setup:

(i) Decision/information table
(ii) Indiscernibility, reduct, and core
(iii) Cut and discretization
(iv) Rules generation

3.1. Classification Measurements. Various formulations have been performed for the classification measurements. The formulation of measures is given below:

$$\begin{aligned} &\text{sensitivity measure (recall)} = \frac{\text{TP}}{P}, \quad \text{where } P = \text{TP} + \text{FN}, \\ &\text{specificity} = \frac{\text{TN}}{N} \Longrightarrow 1 - \text{specificity} = \frac{\text{FP}}{N}, \\ &\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ &\text{accuracy} = \frac{\text{TP} + \text{TN}}{P + N}, \\ &\text{coverage} = \frac{\text{no. of cases satisfying condition and decision}}{\text{no. of cases satisfying decision}}. \end{aligned} \tag{2}$$

4. Results and Discussion

Different algorithms of rough set were applied for the experimental setup of the proposed research. These algorithms include KNN, decision rule, decomposition tree, and LTF-C. Figure 3 shows the knowledge level of the user along with the number of decision instances.

Figure 4 shows the algorithms along with the number of rules for the given decision instances.

Different performance metrics such as specificity, accuracy, precision, F-score, recall, and misclassification rates are followed to check the validity of the proposed model based on different classification algorithms. These algorithms include KNN, cross validation/k-fold mechanism, decision rules, decomposition trees, and local transfer function classifier. The accumulated results and discussion are discussed below in detail.

(i) KNN-based results: the results of the KNN classifier for four different keywords are depicted in Figure 5. From the figure, it is concluded that the KNN provides prominent results in terms of accuracy, misclassification rates, and other performance metrics selected. These high-performance results ultimately show the applicability of the proposed algorithm.
(ii) Cross validation method: the results of the cross validation architecture for four different keywords

Table 1: Techniques and their descriptions.

Reference	Year	Method	Description
[7]	2020	Knowledge transfer by the domain-independent user latent factor	The study proposed an approach of knowledge transfer by the domain-independent user latent factor for cross domain recommender systems. The method has used tr-factorization.
[8]	2020	Assessment of linked data visualization tools	The study presented the analysis of the state-of-the-art tools for the visualization of linked data. List of 77 linked data visualization tools from the previous research and integrating new tools published are given. Based on usability and their features, the visualization tools are compared and described.
[9]	2020	Role of media in user participation	The study considered the usage of media effects in online commentaries on creating knowledge. The user groups were divided into three categories: passive participants, active participants, and bystanders. Their experimental results reveal that the active participants largely tend to use tablets, PCs, and smartphones for creation of knowledge in online space.
[10]	2020	Entrepreneurs' advantages from user knowledge to create innovation in the digital sector	The authors have focussed on the user knowledge value to entrepreneurs and tackled the gap in the literature associated to the activities of entrepreneurs and user knowledge in the digital services. The framework of innovation opportunity space is proposed and has been applied on a UK-based mobile telephony supplier giffgaff for the issues faced by the user knowledge application to digital services.
[11]	2019	Visualization of knowledge and nanocrystal modelling geometry	The study has extracted important insights from the crystal's geometry and physical properties for creation of new structuring according to the methodology of knowledge and visualization.
[12]	2019	User choice of interactive data visualization format	The authors have investigated cognitive style, task difficulty, spatial ability affect choice, and preference of visualization format and then how the visualization selected affects the confidence and decision accuracy.
[13]	2019	Architecture and optimization of data mining modelling for visualization of knowledge extraction	Gebremeskel and Biazen have designed a system capable of analyzing and handling the data which is in large scale.
[14]	2019	TrajAnalytics	The study presented TrajAnalytics, an open source software for modelling, transforming, and visualizing the urban trajectory data, for the study of urban and transportation. The approach allows practitioners to understand the data of the population mobility and find out knowledge. A conceptual model for data is presented which incorporates geostructures with trajectory data with the help of different access queries of data.
[15]	2019	Visualization and analysis of schemas and instances of ontologies for improving user tasks and knowledge discovery	The authors have proposed a solution of visual analytics based on the use of several coordinate views for the description of diverse aspects of ontology and the technique of degree of interest use for reduction of complexity in the visual representation of ontology.
[16]	2018	Interactive machine learning by visualization	The research presented an approach of visual analytics for the visual data mining and interactive machine learning. In the approach, techniques of multidimensional data visualization are applied for the facilitation of user interactions with machine learning and data mining process
[17]	2018	Making graph visualization a user-centred process	The study has explored a cognitive approach for following user-centred process in the visualization graph. A graph-based visualization model is proposed which is a two-stage conceptualized assessment cycle.
[18]	2018	A user-based taxonomy for deep learning visualization	Yu and Shi presented a minisurvey consisting of the user-based taxonomy that converts the works of state of the art in the field.

Table 1: Continued.

Reference	Year	Method	Description
[19]	2017	SemUI	The authors have proposed a SemUI tool-based solution as a multitiered method consisting of (a) a semantic layer which incorporates data through notion of entity of the real world and groups them based on their differences and similarities and (b) a layer of visualization which concurrently shows several views based on entities properties.
[20]	2017	Visualization of multidimensional resource space	The study proposed an interface of multidimension for adopting the resource space model and presented its advantages in property letting application.
[21]	2014	Model of knowledge generation for visual analytics	The authors proposed a model of visual analytics knowledge generation to tie different frameworks.
[22]	2014	CoDe modelling	The study presented a methodology for exploiting visual language CoDe based on a logic paradigm. The CoDe is giving a structure for organizing visualization by the CoDe model and represents graphically the relationships between items of the information.
[23]	2013	Visualizing the impact of time series data	Macek and Atzmueller presented a new concept of visualization for the user history interactions. Association rules are derived and visualized through heatmaps. The impact of the approach is demonstrated by real-world examples of data such as twitter dump of 2009.
[24]	2012	Graphical representation and exploratory visualization for decision trees in the KDD process	The authors presented an approach of representation and a scheme of investigative visualization for the decision tree in the knowledge discovery database process for data mining.

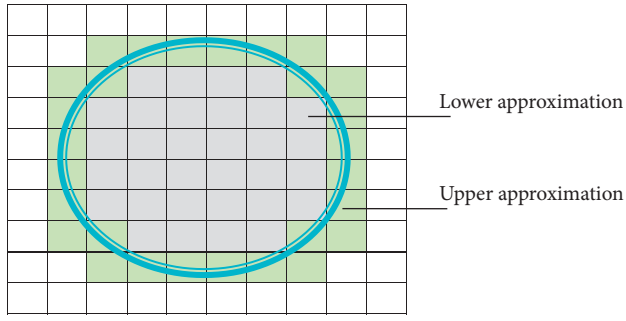

Figure 1: Concept of rough set.

are depicted in Figure 6. From the figure, it is concluded that the cross validation provides good results but its performance is not as good as the KNN-based model. It generates comparatively large values for the misclassifier than the KNN-based model. Also, it generates small % age values for other performance measures. These small accuracy values and high misclassification rates show the inability of the cross validation mechanism in the proposed field.

(iii) Decision rules: the results of the decision rules-based classification architecture are depicted in Figure 7. Compared to both the KNN and cross validation models, its accuracy results are too small and its misclassification rate is very high. This low performance reflects the inability of the decision rule-based architecture in the proposed field.

(iv) Decomposition tree: the results of the decomposition tree-based classification architecture are depicted in Figure 8. Compared to the prescribed KNN cross validation and decision rules-based models, its accuracy results are too small and its misclassification rate is very high. This low performance reflects the inability of the decision rule-based architecture in the proposed field.

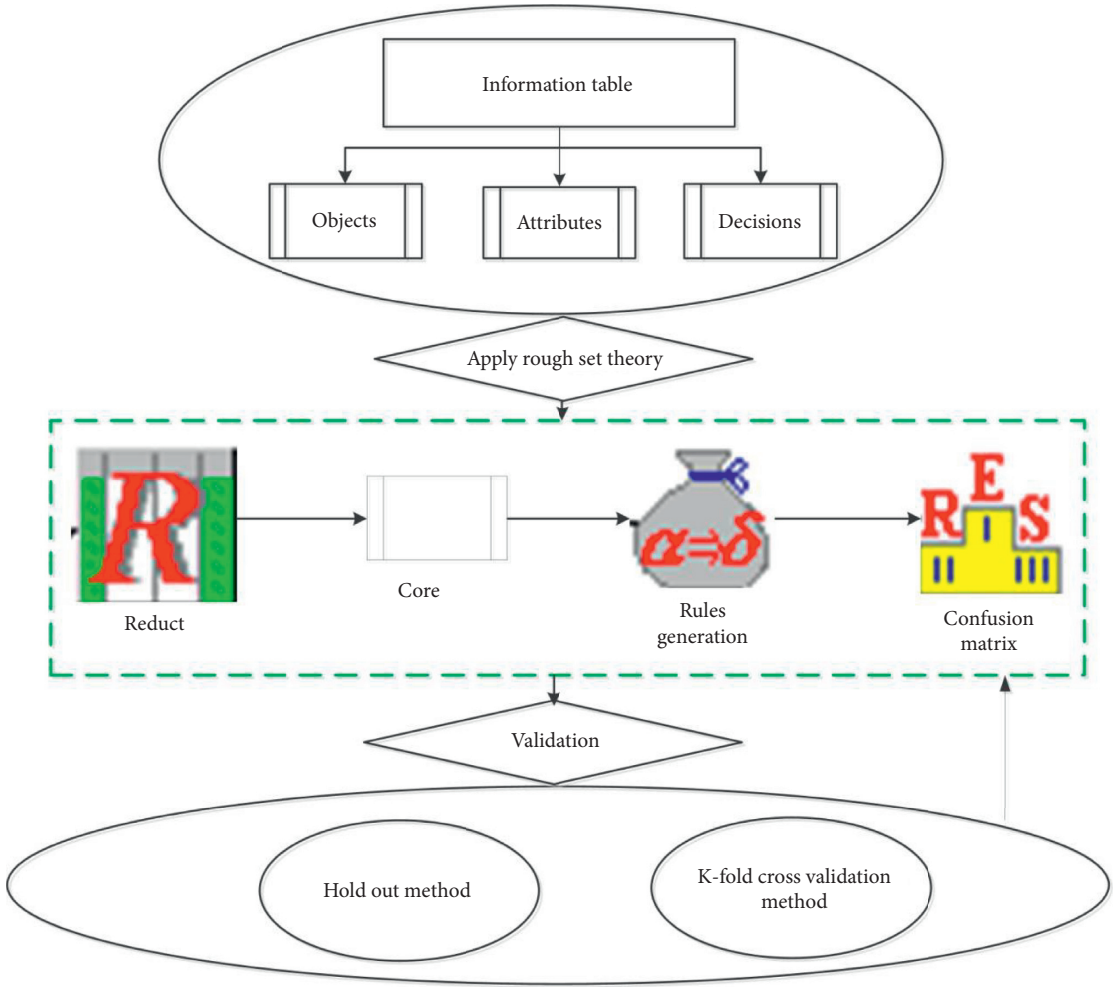

Figure 2: Generic rough set process.

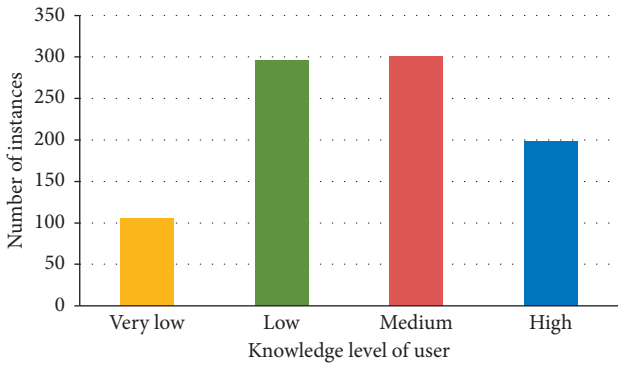

Figure 3: Knowledge level of the user along with the number of instances.

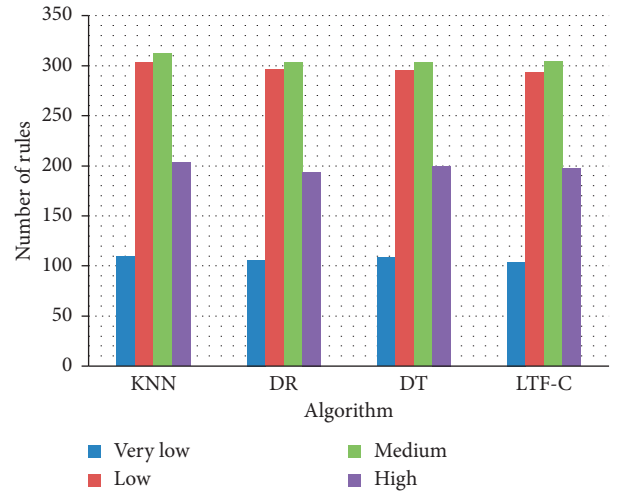

Figure 4: Algorithms along with the number of rules for the given decision instances.

(v) LTF-C-based results: the results of the LTF-C-based classification architecture are depicted in Figure 9. For some keywords, it generates the optimum results, but for some instances, it generates high misclassification rates. For two objects, it generates a misclassification rate greater than 60% and 17% that can generate vague results. In recognition task, vague results are never acceptable, and this ultimately reflects the non-applicability of the LTF-C-based architecture in the proposed model.

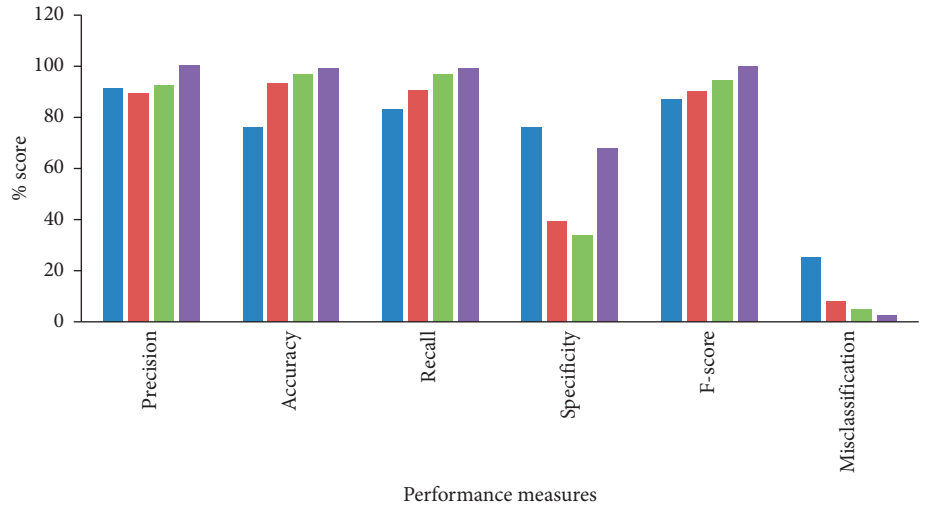

Figure 5: KNN-based performance results.

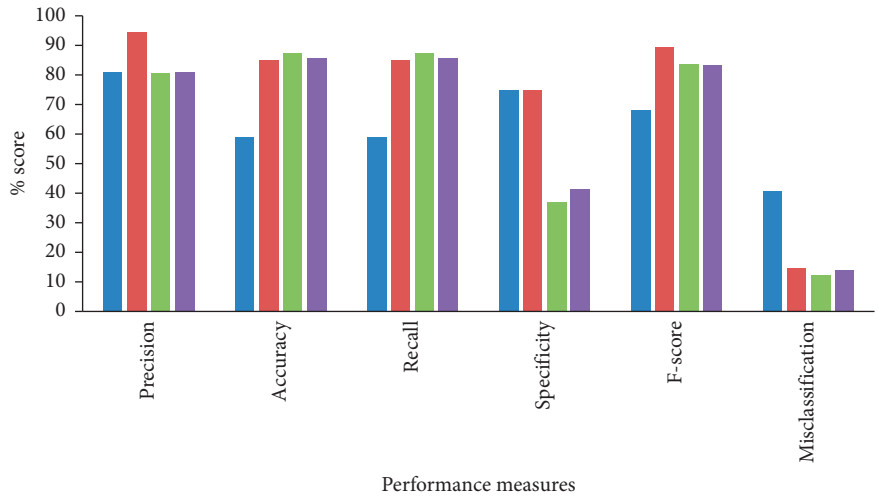

Figure 6: Cross validation-based performance results.

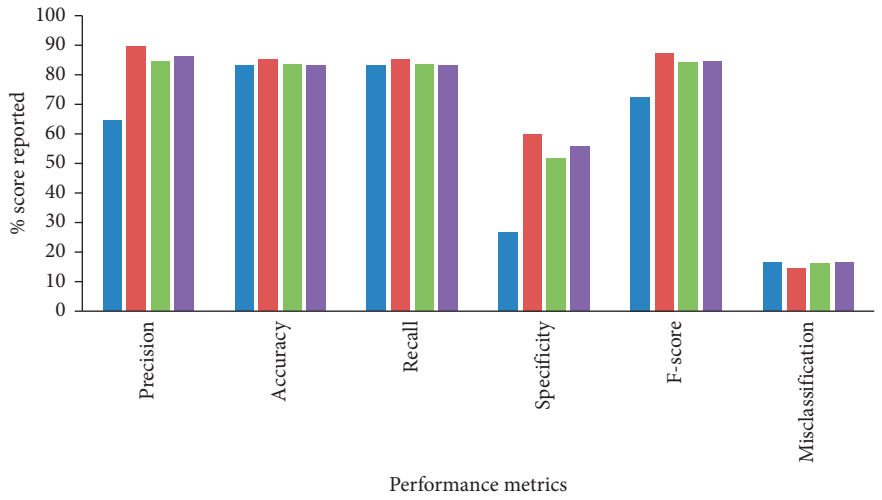

Figure 7: Decision rules-based performance results.

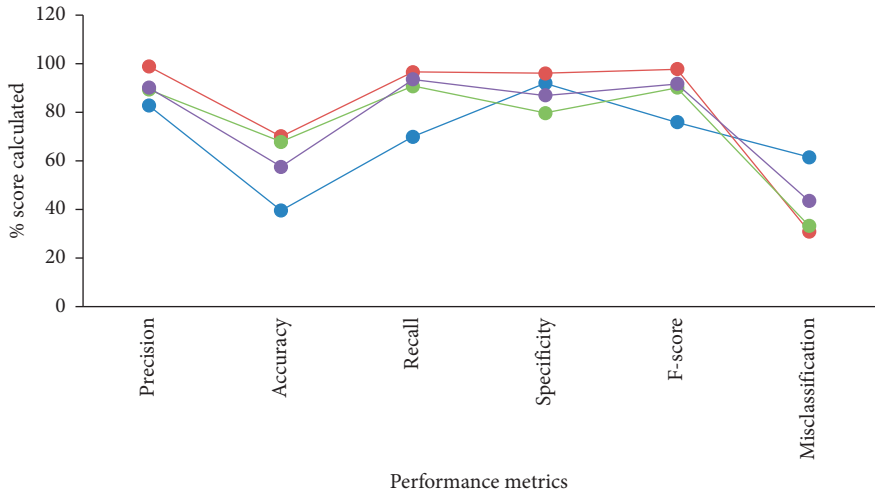

Figure 8: Decomposition tree based-performance results.

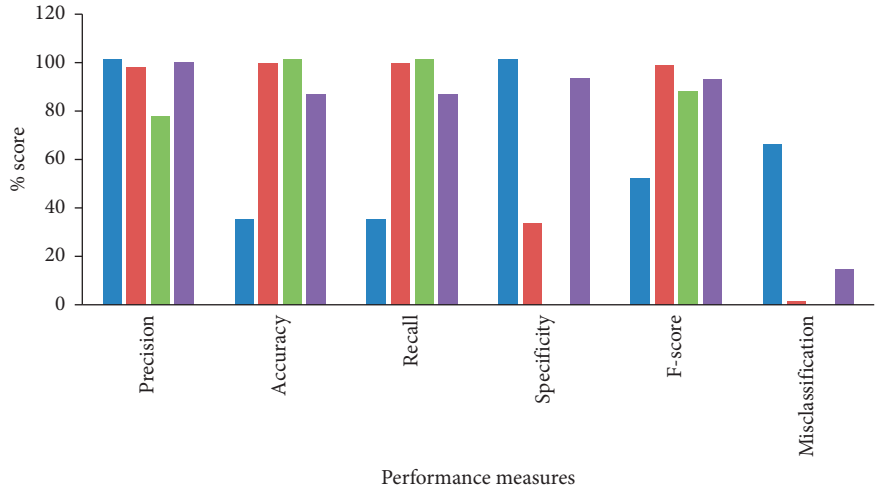

Figure 9: LTF-C-based performance results.

5. Conclusion

With the enhancement of technology, the level of user knowledge is increasing day by day. This increase of information is in volume, velocity, and variety. Extracting meaningful insights is the dire need of an individual from such information and knowledge. Visualization is a key tool and has become one of the most significant platforms for interpreting, extracting, and communicating information. The current study is an endeavour toward data modelling and user knowledge by using the rough set approach for extracting meaningful insights. The technique has used different rough set algorithms such as KNN, decision rules, decomposition tree, and LTF-C for the experimental setup. The approach has found its accuracy for the optimal use of data modelling and user knowledge. The experimental setup of the proposed method is validated by using the dataset available in the UCI web repository. The KNN algorithm shows good accuracy among the algorithms used for the experimental setup of the proposed research. The results have an accuracy of 96% for KNN, 87% for decision rules, 91% for decision trees, 85.04% for cross validation architecture, and 94.3% for LTF-C. The validity of the proposed classification algorithms is tested using different performance metrics such as F-score, precision, accuracy, recall, specificity, and misclassification rates.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Science and Technology Project of State Grid Xizang Electric Power Co., Ltd. (SGXZJY00JHJS2000007), Influence of Energy Storage

Technology Application on Power Grid, and Science and Technology Project of State Grid Zizang Electric Power Co., Ltd (SGXZJY00JHJS2000008), Research Technology Service of Multi Energy Complementary Demonstration Application.

References

[1] IDC, "Analyze the future," 2014, http://www.idc.com/.

[2] M. Hilbert and P. López, "The world's technological capacity to store, communicate, and compute information," *Science*, vol. 332, no. 6025, pp. 60–65, 2011.

[3] S. Sagiroglu and D. Sinanc, "Big data: a review," in *Proceedings of the International Conference on Collaboration Technologies and Systems (CTS'13)*, pp. 42–47, San Diego, CA, USA, May2013.

[4] M. Chen, D. Ebert, H. Hagen et al., "Data, information, and knowledge in visualization," *IEEE Computer Graphics and Applications*, vol. 29, no. 1, pp. 12–19, 2009.

[5] H. T. Kahraman, I. Colak, and S. Sagiroglu, "Developing intuitive knowledge classifier and modeling of users' domain dependent data in web, knowledge based systems," 2013, https://archive.ics.uci.edu/ml/datasets/User+Knowledge+Modeling.

[6] S. Khan, H. Ali, Z. Ullah, N. Minallah, S. Maqsood, and A. Hafeez, "KNN and ANN-based recognition of handwritten Pashto letters using zoning features," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 10, pp. 570–577, 2018.

[7] A. K. Sahu and P. Dwivedi, "Knowledge transfer by domain-independent user latent factor for cross-domain recommender systems," *Future Generation Computer Systems*, vol. 108, pp. 320–333, 2020.

[8] F. Desimoni and L. Po, "Empirical evaluation of linked data visualization tools," *Future Generation Computer Systems*, vol. 112, pp. 258–282, 2020.

[9] J. Chang and J. Hwang, "The role of media in user participation: focusing on the knowledge activity in online space," *Telematics and Informatics*, vol. 51, Article ID 101407, 2020.

[10] S. Flowers and M. Meyer, "How can entrepreneurs benefit from user knowledge to create innovation in the digital services sector?" *Journal of Business Research*, 2020.

[11] J. Constant, "Knowledge visualization and nano-crystal modeling geometry," *Applied Surface Science*, vol. 473, pp. 668–672, 2019.

[12] W. Luo, "User choice of interactive data visualization format: the effects of cognitive style and spatial ability," *Decision Support Systems*, vol. 122, Article ID 113061, 2019.

[13] G. B. Gebremeskel and B. H. B. Biazen, "Architecture and optimization of data mining modeling for visualization of knowledge extraction: patient safety care," *Journal of King Saud University - Computer and Information Sciences*, 2019.

[14] S. Al-Dohuki, F. Kamw, Y. Zhao, X. Ye, J. Yang, and S. Jamonnak, "An open source TrajAnalytics software for modeling, transformation and visualization of urban trajectory data," in *Proceedings of the IEEE Intelligent Transportation Systems Conference (ITSC)*, pp. 150–155, Auckland, New Zealand, October 2019.

[15] I. C. S. Silva, G. Santucci, and C. M. D. S. Freitas, "Visualization and analysis of schema and instances of ontologies for improving user tasks and knowledge discovery," *Journal of Computer Languages*, vol. 51, pp. 28–47, 2019.

[16] H. Li, S. Fang, S. Mukhopadhyay, A. J. Saykin, and L. Shen, "Interactive machine learning by visualization: a small data solution," in *Proceedings of the IEEE International Conference on Big Data*, pp. 3513–3521, Seattle, WA, USA, August 2018.

[17] W. Huang, J. Luo, T. Bednarz, and H. Duh, "Making graph visualization a user-centered process," *Journal of Visual Languages & Computing*, vol. 48, pp. 1–8, 2018.

[18] R. Yu and L. Shi, "A user-based taxonomy for deep learning visualization," *Visual Informatics*, vol. 2, no. 3, pp. 147–154, 2018.

[19] F. Giunchiglia, S. R. Ojha, and S. Das, "SemUI: a knowledge driven visualization of diversified data," in *Proceedings of the IEEE 11th International Conference on Semantic Computing (ICSC)*, pp. 234–241, San Diego, CA, USA, February 2017.

[20] M. A. Rafi, "Visualization of multi-dimensional resource space," in *Proceedings of the 13th International Conference on Semantics, Knowledge and Grids (SKG)*, pp. 182–187, Beijing, China, August 2017.

[21] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim, "Knowledge generation model for visual analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1604–1613, 2014.

[22] M. Risi, M. I. Sessa, M. Tucci, and G. Tortora, "CoDe modeling of graph composition for data warehouse report visualization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 3, pp. 563–576, 2014.

[23] B. Macek and M. Atzmueller, "Visualizing the impact of time series data for predicting user interactions," in *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pp. 1477-1478, Ontario Canada, August 2013.

[24] W. A. C. Rojas and C. M. Villegas, "Graphical representation and exploratory visualization for decision trees in the KDD process," in *Proceedings of the 2012 XXXVIII Conferencia Latinoamericana En Informatica (CLEI)*, pp. 1–10, Medellin, Colombia, October 2012.

[25] J. Zhang, S. Nazir, A. Huang, and A. Alharbi, "Multicriteria decision and machine learning algorithms for component security evaluation: library-based overview," *Security and Communication Networks*, vol. 2020, 2020.

[26] A. U. Haq, J. P. Li, J. Khan et al., "Intelligent Machine Learning approach for effective recognition of diabetes in the E-healthcare using clinical data," *Sensors*, vol. 20, no. 9, p. 2649, 2020.

[27] L. Guangjun, S. Nazir, H. U. Khan, and A. U. Haq, "Spam detection approach for secure mobile messages communication using machine learning algorithms," *Security and Communication Networks*, vol. 2020, pp. 1–6, Article ID 8873639, 2020.

[28] A. Ahmad, C. Feng, M. Khan et al., "A systematic literature review on using machine learning algorithms for software requirements identification on stack overflow," *Security and Communication Networks*, vol. 2020, pp. 1–19, Article ID 8830683, 2020.

[29] A. U. Haq, J. Li, M. H. Memon et al., "Feature selection based on L1-norm support vector machine and effective recognition system for Parkinson's disease using voice recordings," *IEEE Access*, vol. 7, pp. 37718–37734, 2019.

[30] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mobile Information Systems*, vol. 2018, pp. 1–21, Article ID 3860146, 2018.

[31] Z. Pawlak, "Rough sets," *International Journal of Computer & Information Sciences*, vol. 11, no. 5, pp. 341–356, 1982.

[32] Z. Pawlak, *Rough Set: Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publishers, Norwell, MA, USA, 1992.

[33] RSES2, 2020, http://www.mimuw.edu.pl/~szczuka/rses/start.html.

[34] L. S. Riza, A. Janusz, C. Bergmeir et al., "Implementing algorithms of rough set theory and fuzzy rough set theory in the R package "RoughSet"" *Information Sciences*, vol. 287, pp. 68–89, 2014.