British Journal of Computer, Networking and Information Technology
ISSN: 2689-5315
Volume 7, Issue 1, 2023 (pp. 18-24)
www.abjournals.org

# PROBLEM SPACES AND ALGORITHMS IN DATA MINING

**Ele B. I., Obono I. O. and Iwinosa A. A.**

Department of Computer Science, University of Calabar, Calabar, Cross River State, Nigeria

mydays2020@gmail.com, Csmpnd@gmail.com & iwinosaa@gmail.com,

**Cite this article:**

Ele B. I., Obono I. O., Iwinosa A. A. (2024), Problem Spaces and Algorithms in Data Mining. British Journal of Computer, Networking and Information Technology 6(1), 18-24. DOI: 10.52589/BJCNIT-ZRZ4EUKT

**Manuscript History**

Received: 30 Sept 2023

Accepted: 20 Nov 2023

Published: 2 Jan 2024



**ABSTRACT**: *Data mining has been described severally as the best thing to have happened to data and information management, especially these days that the cost of computing technologies and storage media are falling, data gathering tools becoming varied and very efficient and the boom in network computing becoming very rewarding. The challenges presented by the management and meaningful usage of large data sets have stimulated so much research in data mining. Consequently, the birth of a number of algorithms to provide insights to these big data has equally presented more complications in information processing computing. Therefore, this paper presents different problem spaces in data mining, available algorithms to mine these data and then mapping specific algorithms to specific problem spaces. Analysis of datasets from a typical financial institution suggests that no one algorithm is necessarily better than the other, but all have strengths and weaknesses depending on the particular problem spaces in use.*

**KEYWORDS**: Data mining, Problem spaces, Data analytics, Big data, Algorithms.



Article DOI: 10.52589/BJCNIT-ZRZ4EUKT
DOI URL: https://doi.org/10.52589/BJCNIT-ZRZ4EUKT

## INTRODUCTION

As our society becomes more complex and the factors supporting human activities are difficult to predict, we need knowledge to be able to adapt to this ever-changing situation. Knowledge from one source is not enough any longer, as we need knowledge from various sources, different environments and from different points in time; past present and the future to be able to make critical decisions that affect the very fabric of our everyday existence.

However, the authors in [2] opined that mining is the current hotspot, the most promising research area, through data mining research status, algorithms and applications of analysis to explore data mining problems and trends, which the development of data mining has certain reference value. They further observed that data mining is the advanced process which extracts the effective potential and comprehensive mode from the vast amounts of data in accordance with the established business goals.

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly vital, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. Therefore, Data Mining also popularly known as Knowledge Discovery in Databases (KDD) refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases [3].

The widespread use of databases from network point-of-sales transactions, web hits, credit card purchases, e-commerce to pixel by pixel satellite images of galaxies resulting from cheap data gathering and storage technologies has presented another set of challenges; How to infer meanings and usefulness from this large data sets (big data) so generated. Providing the capabilities to answer these questions as well as other functionalities like providing interface between human analysts and storage systems, data navigation and exploration, summarization and modeling of true business scenarios from large databases is the goal of data mining and more recently data analytics [4].

Therefore, this study is devoted to the analysis of various data mining algorithms in problem spaces and the identification of valuable knowledge provided by data mining techniques.

## RELATED WORK

When data mining started in the 1990s, interest was purely on research and academics. However, the authors in [2] showed clear need for new data analysis challenges posed by the overwhelming volume of data generated by modern data acquisition systems, which brings together a number of technologies enabling data mining intersecting several disciplines (see figure 1). There appear to have been an absence of a generally accepted definition of Data Mining since it was a new area of research; hence different interests saw data mining from the perspectives of their interest. The idea of the authors in [3] was to turn computers loose on the data and see what trends can be found, while the authors in [5] and [6] advocate the implicit extraction of useful knowledge from data using algorithmic means under acceptable computational efficiency limitations. The authors in [7] were more interested in the dredging of large geographical data for anomalies to initiate '**human in the loop**' analysis. Common to all interest groups however, is the understanding that data mining is a multi-discipline (figure


Article DOI: 10.52589/BJCNIT-ZRZ4EUKT
DOI URL: https://doi.org/10.52589/BJCNIT-ZRZ4EUKT

1) of applied science involved with the deployment of automated procedures or algorithms to analyze historical data in databases (not necessarily large databases) for hypothesis testing, rules generation and discoveries of hidden trends with potential useful insights for strategic decision making in any field. The authors in [1] and [3] summarized these definitions in figure 2. These divergent views also saw different data mining applications tailored with a bias towards each problem space of interests but with claims of a universal application. It was difficult accepting some of these claims since data management activities usually involve data classification, clustering, sequential analysis and sometimes a combination of all, which requires different approaches entirely to infer useful information.

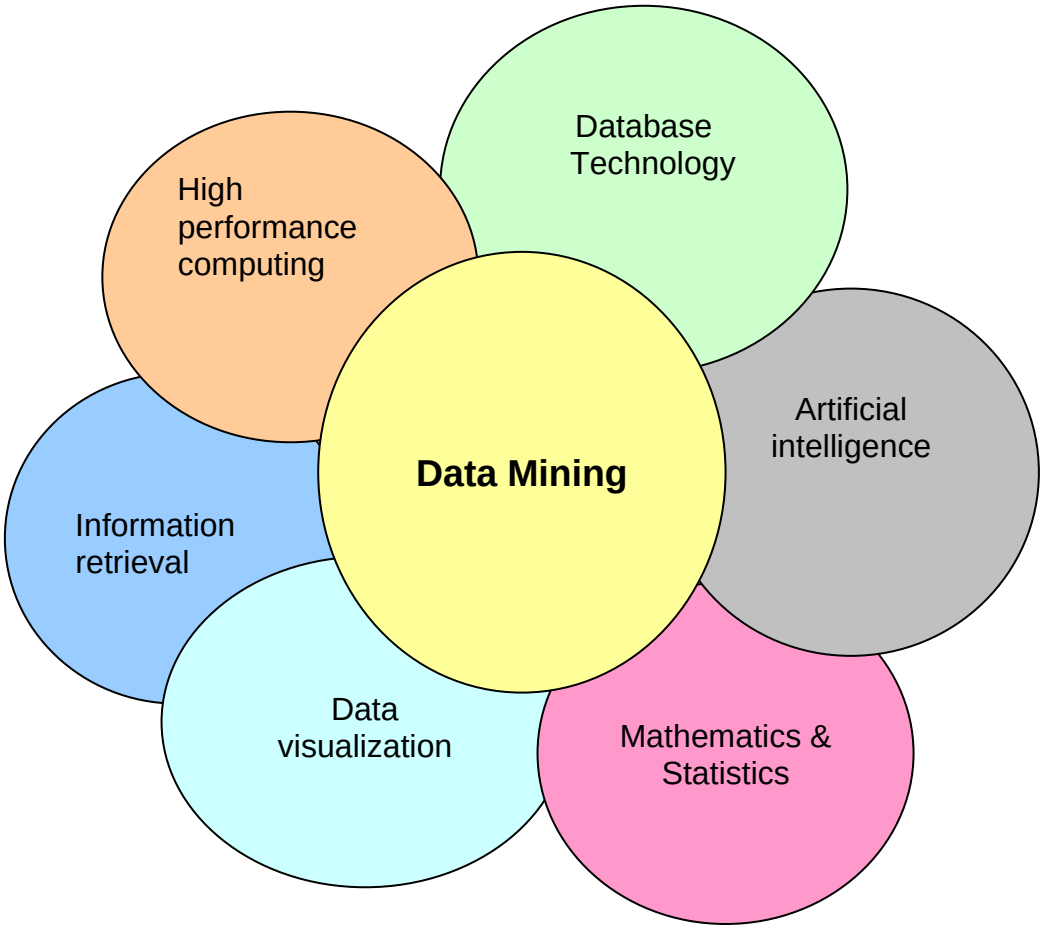


**Fig 1: Data Mining as a Multidiscipline**


Article DOI: 10.52589/BJCNIT-ZRZ4EUKT
DOI URL: https://doi.org/10.52589/BJCNIT-ZRZ4EUKT

British Journal of Computer, Networking and Information Technology
ISSN: 2689-5315
Volume 7, Issue 1, 2023 (pp. 18-24)
www.abjournals.org

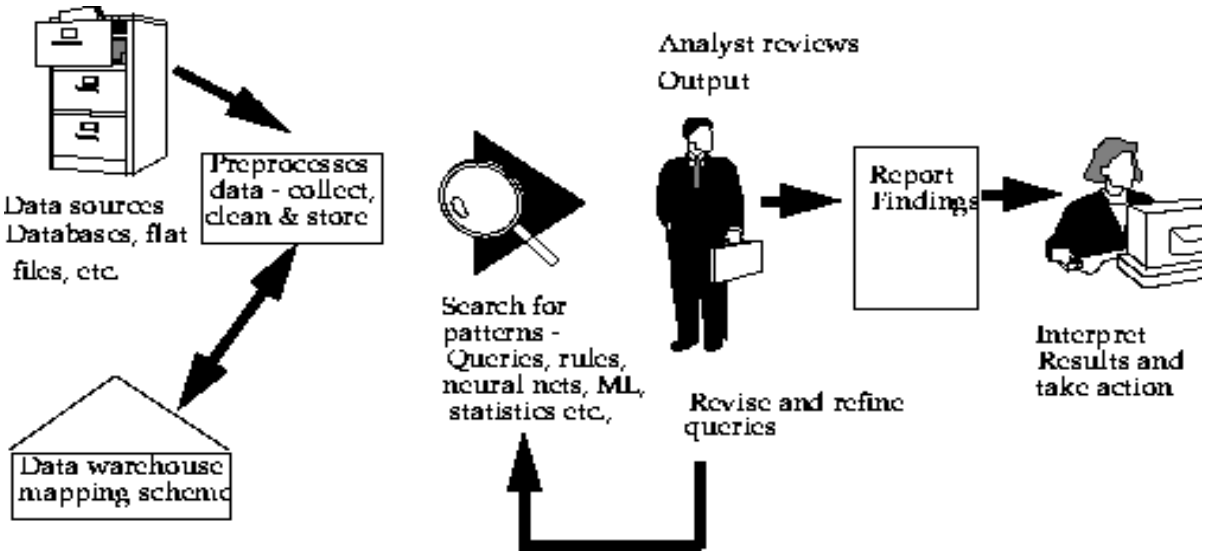


**Figure 2: Data mining process [1]**

### Problem Spaces

It is important to understand that data mining problems usually fall under two broad categories of problem spaces namely prediction or supervised problems (classification and regressions) and description or unsupervised problems (association, clustering and time series). So a data-mining algorithm is usually tailored to address problems in each problem space or domain. Classification and regression problem spaces are usually modeled using decision trees, neural networks and Naïve/simple Bayes algorithms. The authors in [5] and [8] identified metric distance based methods like k-nearest neighbor algorithms, model based methods, partition based methods and sequential algorithms to provide solutions in descriptive problem spaces that require clustering and segmentation of records.

## MATERIALS AND METHOD

To identify different problem spaces, and determine if data mining algorithms outperform each other, we source data sets from Fortune Bank PLC, Nigeria. Though, it was not possible to obtain large data sets in the magnitude of terabytes and pecobytes. Also due to privacy and security issues, sensitive attributes of the records were omitted from this study, the records were pruned to contain only five attributes; name, gender, age, income and account type. The source data, a prune record of 2047 rows and five attributes in excel format was uploaded to an IBM Compatible Pentium 4 running an IBM DB2 Universal Database Enterprise. IBM also provided a complimentary copy of the IBM Intelligent Miner Version 2. This software application was ideal for this study due to the number of data mining algorithms, all woven into this one application.

This study aims to use the different data-mining algorithms on these data and see if any hidden trend is revealed about the customers using different algorithms. Also the time complexities


Article DOI: 10.52589/BJCNIT-ZRZ4EUKT
DOI URL: https://doi.org/10.52589/BJCNIT-ZRZ4EUKT

required for generating these rules or classifications, clarity of results was compared to determine which perform better.

## RESULTS AND DISCUSSIONS

The IBM intelligent miner has a beautiful array of visualizers to display results. The miner displays the output in colorful histograms, bar charts and pie charts.

### Clustering Problem Spaces

Figures 4.1 through to figures 4.4 provide the results when clustering algorithm is applied on the sample data. The mining algorithm generated nine clusters by which the sample records are distributed. Within each cluster, the bar charts and pie charts represent the active attributes that were used to generate the clusters. The attributes with the greatest influence are displayed on the left and least influence on the right. The cluster size is represented in percentage and is shown as the numbers on the left while the numbers on the right represent the cluster ID.

The largest cluster in figure 4.1 is the top most cluster with 61% of customers with a current account. Figure 4.2 provides details of this cluster and further analysis will provide insight about those categories of customers with current accounts in the bank.

Using this algorithm on the sample data provided the following advantage.

(a) Run time: the observed run time of this algorithm on sample data was 0.01 seconds

(b) Discoveries and insights: by identifying cluster three with 61% population, will be easier for Fortune bank to tailor a particular product for these customers.

(c) The algorithm is easier to interpret.

### Neural Network

It was observed **that the k nearest neighborhood algorithm** is best suited for problems of association and cluster building where there was no previous criteria or information for comparison. Hence, it is best suited for the data available for this project.

It is worthy to note that there is no universal application of these algorithms on any data sources; rather each algorithm is better suited for a specific or particular problem or domain. Presented in the table below is a summary of the observations and results of the application of intelligent miner on the database


Article DOI: 10.52589/BJCNIT-ZRZ4EUKT
DOI URL: https://doi.org/10.52589/BJCNIT-ZRZ4EUKT

**Table 1: Summary of Results from Sample Data using IBM Intelligent Miner**

| Problem space | Algorithms | Observed results | Limitations |
|---|---|---|---|
| Prediction or supervised problems (classification and regressions) | Decision trees | The depth of the tree generated at depth 32 was too bushy to yield any meaningful result. When trees were pruned to seven depth, the results were too scanty and the rules generated too general for any meaningful interpretation. | Clearly not suitable for problems in this domain |
| Description or unsupervised problems (association, clustering and time series) | Neural networks, Genetic algorithms | Was able to partition data into nine clusters but no interestingness found. Was very exhaustive in analyzing data eliminating any bias. Good for data with non-linear relationships. Can work well with incomplete data. 0.02(nx2) run time was observed. | Complex and difficult to understand. |
| | K-nearest neighbor | Was able to segment the data into nine clusters of customers, three of the clusters with interestingness characteristics of more than 60%. Run time was 2 seconds (order n). Results were easy to interpret. | Suspicion that the runtime will increase exponentially on a large database. |

## CONCLUSION

The fact that data mining provides significant insight for business organizations and research institutions through knowledge discovery and hypothesis testing is not in doubt, however if more research effort is channeled into developing a coherent and unified approach to data mining, more benefits will be derived from this field. Presently research in this area is dominated by business owners, hence the reason for different techniques. Academia is encouraged to show more interest. There exist some serious challenges for further research like developing an effective means for data sampling, data reduction, and dimensionality reduction and developing algorithms that allow for proper trade offs between complexity and understanding of models for the purpose of visualization and reporting. The authors in [10] summarized some of the challenges that will provide data mining with tremendous benefits if properly researched.


Article DOI: 10.52589/BJCNIT-ZRZ4EUKT
DOI URL: https://doi.org/10.52589/BJCNIT-ZRZ4EUKT

## REFERENCES

[1] Han, J., Kamber, M., & Pei, J. (2011). Data mining concepts and techniques. Morgan Kaufmann, San Francisco.
[2] Fang, W. & Wang, Y. (2013). The Development of Data Mining. *International Journal of Business and Social Science*, 4(16), 45 - 58.
[3] Zaki, M. J. & Meira, M. J. (2017). Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press, New York, USA.
[4] Gaber, M. M. (2010). Scientific Data Mining and Knowledge Discovery: Principles and Foundations. Springer, New York, USA.
[5] Sethunya, R. J., Hlomani, H., and Keletso, L. (2016). Data Mining Algorithms: An Overview. International Journal of Computers and Technology. 15(6), 68-75.
[6] Tan, P. N., Steinbach, M., & Kumar, V. (2006). Introduction to Data Mining. Pearson Addison Wesley, Boston.
[7] Adebowale, A., Idowu, S. A., & Amarachi, A. (2013). Comparative Study of Selected Data Mining Algorithms Used for Intrusion Detection, *IJSCE*, , 3(3), 234-244.
[8] Ye, N. (2013). Handbook of data mining. Mahwah, NJ: Lawrence Erlbaum.
[9] Sushmita, M., Sankar, P. K. and Pabitra, M. (2012). Data Mining in Soft Computing Framework: A Survey. IEEE Transaction on Neural Networks, 13(1), 3-14.
[10] Zafarani, R., Abbasi, M. A. and Liu, H.(2014). Social Media Mining. Cambridge University Press, New York, USA.
[11] Kesavulu, E., Reddy, V. N. and Rajulu, P. G. (2011). "A Study of Intrusion Detection in Data Mining". Proceedings of the World Congress on Engineering WCE, July 6 - 8, 2011, London, UK.
[12] Edelstein, H. (2009). Introduction to data mining and knowledge discovery. Two Crow corporation. www.twocrow.com.
[13] Lappas, T., and Pelechrinis, K. (2006). Data Mining Techniques for Network Intrusion Detection Systems, Department of Computer Science and Engineering Riverside, Riverside CA.


Article DOI: 10.52589/BJCNIT-ZRZ4EUKT
DOI URL: https://doi.org/10.52589/BJCNIT-ZRZ4EUKT