IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

VISUALIZATION TECHNIQUES IN BIG DATA MINING FOR ENHANCING DATA EXPLORATION AND ANALYSIS

[1]A P Tamilselvan,
Research Scholar,
PG and Research Department of Computer Science,
Kaamadhenu Arts and Science College,
Sathyamanagalam, Tamilnadu, India.

[2]Dr.S.P.Priyadharshini,
Assistant Professor,
Department of Computer Applications,
Government Arts & Science College,
Sathyamanagalam, Tamilnadu, India.

Abstract - This paper provides an overview of big data mining techniques and their diverse applications across various domains. Big data mining involves extracting valuable insights from large datasets using advanced analytical methods such as machine learning and data mining algorithms. Applications span healthcare, finance, cybersecurity, and more, enhancing decision-making processes and improving operational efficiency. Key techniques include classification, clustering, association rule mining, and anomaly detection, tailored to handle the complexities of massive datasets. This overview aims to highlight the importance of big data mining in modern data-driven environments, emphasizing its role in transforming data into actionable knowledge.

Keywords: Big Data Mining, Machine Learning, Data Analytics, Applications, Transforming data;

1. Introduction

Big data mining has transformed the way organizations extract valuable insights from extensive and diverse datasets, tackling the challenges presented by today's massive data volumes, high velocity of data generation, and varied data types. This field encompasses sophisticated techniques and methodologies aimed at discovering patterns, trends, and actionable knowledge from large-scale data repositories. The advent of big data has been fueled by advancements in data collection technologies, such as sensors, social media platforms, and Internet of Things (IoT) devices, which consistently produce enormous volumes of both organized and unstructured data. In essence, big data mining goes beyond traditional data analysis by

leveraging scalable computational algorithms, machine learning, and statistical models to uncover implicit relationships. These techniques enable organizations to make informed decisions, optimize processes, and predict future outcomes based on empirical data rather than intuition alone.

Applications of big data mining span across various domains, including healthcare, finance, marketing, cybersecurity, environmental science. In healthcare, for example, big data mining aids in disease prediction, personalized medicine, and clinical decision support systems by analyzing large-scale patient data and medical records. Similarly, in finance, it facilitates fraud detection, risk assessment, and algorithmic trading strategies by analyzing market trends and financial transactions in real-time. Moreover, big data mining plays a crucial role in enhancing operational efficiencies and customer insights for businesses through sentiment analysis, recommendation systems, and market segmentation based on consumer behavior data. In cybersecurity, it helps in identifying anomalous activities and predicting potential security breaches by analyzing network traffic and user behavior patterns.

The methodologies employed in big data mining include clustering, classification, association rule mining, anomaly detection, and predictive modeling. These techniques are applied iteratively across massive datasets to extract meaningful information, reduce data redundancy, and improve decision-making processes. As big data continues to evolve, the challenges associated with data quality, scalability, privacy, and ethical considerations become increasingly pertinent. Addressing these challenges requires advancements in data management practices, algorithm development, and regulatory frameworks to ensure responsible and effective use of big data mining techniques. In this review, provide an overview of the fundamental principles, methodologies, and applications of big data mining. Explore the impact of big data mining across different sectors, highlighting its transformative potential and the opportunities it presents for innovation and societal advancement.

2. Literature Survey

1. Zhou X (2020) et.al proposed Deep correlation mining based on hierarchical hybrid networks for heterogeneous big data recommendations. Recent advancements in technologies like artificial intelligence, cyber intelligence, and machine learning have facilitated the pervasive integration of big data across various sectors, including industry, academia, and everyday life. In heterogeneous big data contexts, deep correlation mining techniques are the main topic of this paper. It introduces a Hierarchical Hybrid Network (HHN) model to depict diverse relationships among entities, employing measures to quantify internal correlations within each layer and external correlations between layers. An intelligent router, based on a deep reinforcement learning framework, is developed to optimize routing decisions across the HHN. Additionally, an enhanced random walk with restart algorithm utilizes the intelligent router to navigate the hierarchical influence and multiple correlations within the network. To support collaborative efforts in scholarly big data environments, an intelligent recommendation mechanism is devised. Experimental validation using data from DBLP and ResearchGate demonstrates the efficacy and practicality of the

proposed model and methods, highlighting their potential to enhance decision-making and collaboration in complex data-driven domains.

2. El-Hasnony IM (2020) et.al proposed improved feature selection model for big data analytics. In addressing the challenge of feature selection in Big Data applications, this paper introduces a novel approach using a binary variant of wrapper feature selection GWO and PSO. The method leverages a K-nearest neighbor classifier with Euclidean separation matrices and incorporates a tent chaotic map to mitigate local optima issues. A sigmoid function transforms the continuous vector search space into a binary format suitable for feature selection. Comparative evaluations with established algorithms like PSO and GWO were conducted across twenty datasets, measuring criteria such as selected features ratio, classification accuracy, and computation time. Results show that the proposed model selected 196 features out of 773, outperforming GWO (393 features) and PSO (336 features). Moreover, the overall classification accuracy of 90% surpasses other algorithms' performance (81.6% for GWO and 86.8% for PSO), with a total processing time of 184.3 seconds, compared to 272 seconds for GWO and 245.6 seconds for PSO across all datasets.

3. Bhuyan HK (2021) et.al proposed Analysis of subfeature for classification in data mining. Feature selection is critical in data mining applications, particularly in diverse data analysis fields. Traditional feature selection methods often focus on analyzing features for classification, which may not sufficiently capture the complexity of the dataset. This limitation leads to redundant feature data and hampers further analysis. To address this, a novel approach proposes the identification of sub-features (SF) within traditional classes. These SFs are derived from a limited number of significant instances and form new classes or subclasses. The method employs an optimization model based on Lagrangian multipliers to identify and analyze SF data effectively. The theoretical methods of variance- and domain-based sub-features are applied to identify the most informative SFs. various classifiers and statistical methods, including local and global variance, evaluate the classification performance using SF data. Experimental results across different datasets demonstrate the effectiveness of the proposed model in generating novel classes based on selected SF data, highlighting its potential for enhancing classification accuracy in complex data environments.

4. Yin Y (2020) et.al proposed Dynamic data mining of sensor data. In the era of the Internet of Things (IoT), sensor-generated data has supplanted artificially compiled data, sparking significant interest in data mining research across academia and industry. In order to process sensor data, this research presents a dynamic data mining framework that emphasizes building a sensor data mining model that can adjust to dynamic changes. Each sensor network environment is treated as a distinct physical system, with its parameters trained and refined through historical sensor data collection and mining. The model explores associations between different sensor network environments by analyzing the relationships among their respective parameters. The experimental setup focused on variables such as transmission distance, transmission delay, sensor data characteristics, and data changes. Conducted on a designated experimental platform, the experiments demonstrated the model's efficacy in extracting dynamic data and identifying

stable patterns. Analysis of the results highlighted the model's potential for guiding dynamic sensor data mining practices, suggesting novel approaches for industrial big data analytics.

5. Rahman MM (2022) et.al proposed Educational data mining to support programming learning using problem-solving data. Online judge (OJ) systems play an important role in addressing the growing need for talented programmers in the ICT industry by providing real programming practice and evaluation in addition to traditional classroom instruction. These systems accumulate vast archives of problem-solving data, including solution codes, logs, and scores, which are invaluable for educational research in programming. This paper proposes an educational data mining framework aimed at enhancing programming learning through unsupervised algorithms. The framework begins with collecting and pre-processing problem-solving data from OJ systems, followed by applying the MK-means clustering algorithm to organize the data in Euclidean space. Statistical features are then extracted from each cluster, and the FP-growth algorithm is employed to mine data patterns and association rules within each cluster. Based on the extracted features, patterns, and rules, the framework generates recommendations to improve programming learning outcomes. Experimental validation using real-world data from 537 students in a programming course, along with synthetic data, demonstrates the framework's efficacy in uncovering actionable insights and areas for enhancement in programming education.

6. Ishaq A (2021) et.al proposed improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques. Cardiovascular disease remains a significant global health concern, posing challenges in predicting patient survival using clinical data analytics. Data mining techniques are crucial in transforming vast amounts of health data into actionable insights for informed decision-making. This study focuses on analyzing the survival of heart failure patients from a dataset of 299 hospitalized individuals. The goal is to identify key features and effective data mining methods that can enhance the accuracy of survival predictions in cardiovascular patients. Addressing the imbalance in class distribution is managed using the Synthetic Minority Oversampling Technique (SMOTE). Models are trained on features ranked highest by RF to improve predictive performance. Experimental results highlight ETC as the top-performing model, achieving an accuracy of 0.9262 with SMOTE. This study underscores the efficacy of advanced machine learning techniques in enhancing cardiovascular patient survival prediction through feature selection and model optimization.

7. Feng G (2022) et.al proposed Analysis and prediction of students' academic performance based on educational data mining. The adoption of intelligent technologies is gaining momentum in education, driven by the exponential growth of educational data that challenges traditional processing methods and introduces potential biases. This has underscored the need to revamp data mining research in education, particularly to mitigate biased evaluations and forecast student performance proactively. This study integrates clustering, discriminant analysis, and convolutional neural networks to analyze and predict academic outcomes. It introduces a novel statistical approach to optimize the determination of clustering numbers in the K-means algorithm, enhancing its accuracy and objectivity. Discriminant analysis evaluates the clustering efficacy, while convolutional neural networks handle labeled training and testing data,

facilitating robust performance prediction models. Validation employs two cross-validation methods to assess model effectiveness, confirming improved reliability and predictive accuracy. Experimental results highlight the statistical method's role in enhancing clustering determination and the overall predictive quality of educational data mining applications. This approach not only addresses longstanding challenges but also sets a benchmark for more reliable educational performance forecasting.

8. Haoxiang W (2021) et.al proposed big data analysis and perturbation using data mining algorithm. The rapid advancement of computing technologies has significantly increased data generation, particularly in fields like health informatics. However, concerns over data privacy remain paramount, with potential vulnerabilities that could lead to exploitation or unauthorized access. Existing methods for safeguarding data often face scalability and efficiency challenges, along with issues related to preserving privacy and data utility. This study proposes an innovative solution to address these challenges through an effective perturbation algorithm employing optimal geometric transformations for big data. The proposed method has undergone rigorous testing, evaluating its accuracy, resilience against attacks, scalability, and efficiency using five classification algorithms across nine datasets. Experimental results demonstrate that the proposed approach outperforms existing privacy preservation techniques in terms of attack resistance, scalability, execution speed, and accuracy. By leveraging optimal geometric transformations, the algorithm not only ensures robust privacy protection but also maintains data utility, making it a promising advancement for secure handling of large-scale data in sensitive domains like health informatics.

9. Abdelkader HE (2022) et.al proposed an efficient data mining technique for assessing satisfaction level with online learning for higher education students during the COVID-19. Educational institutions prioritize enhancing academic performance to elevate education quality. Educational Data Mining (EDM) leverages Data Mining (DM) concepts to analyze Student Satisfaction Levels (SSL) with Online Learning (OL) during COVID-19 lockdowns. Feature Selection (FS) is crucial in EDM to identify the most relevant features efficiently. The analysis focuses on the fitness and predictive accuracy of feature subsets of varying sizes. Results indicate that reducing feature dimensionality by up to 80% enhances classification accuracy by up to 100%. This underscores the critical role of FS in optimizing SSL prediction models, facilitating informed educational strategies. The findings emphasize the importance of selecting an optimal subset of features to achieve high predictive accuracy and effective educational outcomes in online learning environments.

10. X. Liu (2020) et.al proposed Big-Data-Based Intelligent Spectrum Sensing for Heterogeneous Spectrum Communications in 5G. In the context of contemporary wireless communications, spectrum sensing is essential for making effective use of available spectrum. However, the rapid expansion of wireless technologies has led to a surge in heterogeneous spectrum data, posing significant challenges to spectrum sensing complexity. To address this, machine-learning-assisted spectrum sensing has emerged as a promising approach. This article proposes an intelligent spectrum sensing method based on big data analytics to enhance heterogeneous spectrum sensing capabilities. The method involves establishing a cooperative spectrum sensing network to achieve wide-area broadband spectrum monitoring and gather

comprehensive big spectrum data. Simulation studies validate the network's efficacy in detecting spectrum availability. To enhance data reliability, correlations among big spectrum data in time, frequency, and spatial domains are analyzed to quantify spectrum similarities. Additionally, a clustering mechanism for big spectrum data aids in data matching and predicting heterogeneous spectrum states. Ultimately, these efforts culminate in the fusion of heterogeneous spectrum data to derive a comprehensive spectrum status, advancing spectrum management in dynamic wireless environments.

11. P. Pinoli (2018) et.al proposed DLA: a Distributed, Location-based and Apriori-based Algorithm for Biological Sequence Pattern Mining. The exponential expansion of genetic data has made scalable data mining technologies necessary. Frequent contiguous sequence mining, a technique crucial for understanding DNA function and structure, identifies common characteristics among related sequences. Although numerous sequence mining algorithms exist, many struggle with scaling issues or fail to ensure result completeness. A technique for distributed sequential pattern mining built on Apache Spark is presented in this study. Leveraging the Apriori Property and sequence location information, the algorithm significantly reduces candidate numbers per iteration. Experimental results on real-world datasets validate its performance, demonstrating superior scalability compared to other distributed solutions. This approach holds promise for advancing genomic research by enabling efficient mining of large-scale sequence data, facilitating deeper insights into biological systems.

12. L. Vaira (2016) et.al proposed A Big Data Analytics Framework for Supporting Multidimensional Mining over Big Healthcare Data. In today's healthcare landscape, there is significant focus on big data analytics, particularly in complex healthcare environments. Fetal growth curves, a prominent example of big healthcare data, play a crucial role in prenatal medicine for early detection of potential fetal growth problems, estimating prenatal outcomes, and promptly addressing complications. However, current curves and diagnostic techniques are criticized for their lack of precision. To address this, new techniques based on customized growth curves have been proposed in the literature. This paper discusses the challenge of building personalized fetal growth curves using big data techniques. The proposed framework introduces the concept of summarizing vast amounts of input data through multidimensional views, onto which well-established data mining methods such as clustering and classification are applied. This approach defines a multidimensional mining strategy tailored for complex healthcare environments. A preliminary analysis of the framework's effectiveness is also presented, highlighting its potential to enhance the precision and utility of fetal growth monitoring in prenatal care.

13. J. Z. Huang (2019) et.al proposed Random Sample Partition: A Distributed Data Model for Big Data Analysis. In response to the escalating volume of data, new approaches are essential for partitioning big data into statistically reliable data blocks that serve as representative samples taken from the big data analysis dataset as a whole. This paper introduces the Random Sample Partition (RSP) distributed data model, which divides a big dataset into disjoint data blocks known as RSP blocks. Every RSP block keeps a probability distribution that is comparable to the dataset as a whole. These RSP blocks can effectively

estimate the statistical properties of the data and facilitate the construction of predictive models without the need to process the entire dataset. The implications of the RSP model for big data sampling are illustrated, and a novel RSP-based approach to approximation big data analysis that may be used to a variety of industry scenarios is presented.

14. K. Ogohara (2017) et.al proposed Data Analysis Support by Combining Data Mining and Text Mining. Data mining and text mining approaches have become widely used in the analysis of questionnaire and review data in recent years. Data mining methods like association and cluster analysis are employed in marketing analysis to uncover hidden relationships and rules within vast numerical datasets. Conversely, text mining techniques such as keyword extraction and opinion mining are utilized for analyzing textual data from questionnaires or reviews, aiding in understanding consumer opinions. However, existing data mining and text mining tools often operate in separate environments, making it challenging to effectively analyze datasets containing both numerical and text data. This segregation hinders the ability to connect and interpret insights from both types of data comprehensively. To address this gap, this paper proposes a unified mining framework capable of handling both numerical and text data. The framework facilitates iterative data reduction and analysis using tools tailored for numerical and textual analysis within a cohesive environment. Experimental results demonstrate the effectiveness of this integrated system in analyzing review texts.

15. J. Zhang (2018) et.al proposed PPSF: An Open-Source Privacy-Preserving and Security Mining Framework. In recent decades, ensuring data privacy and security has become increasingly critical, as powerful data mining tools can potentially reveal confidential or private information. While numerous frameworks and tools have been developed to address these concerns, they primarily rely on data anonymization techniques. On the other hand, this study presents a novel framework for privacy-preserving and security data mining, known as the Privacy-Preserving and Security Mining Framework (PPSF). A range of methods, such as data anonymity, privacy-preserving data mining (PPDM), and privacy-preserving utility mining (PPUM), is available in the open-source data mining library PPSF. The framework is equipped with a user-friendly interface, enabling users to execute algorithms and visualize results effectively. It is an actively maintained project with regular updates that include new algorithms, optimizations, and comprehensive documentation. This innovative framework aims to enhance data security and privacy while supporting the effective execution of privacy-preserving data mining operations, making it a valuable tool for researchers and practitioners working in sensitive data environments.

3. Literature Survey Comparison Table

Study	Proposed Method	Application	Techniques Used	Key Findings
Zhou X (2020)	Deep correlation mining using HHN model	Heterogeneous big data recommendations	Hierarchical Hybrid Networks, Deep Reinforcement Learning, Random Walk with Restart	Enhanced decision-making and collaboration in scholarly environments; validated with DBLP and ResearchGate data
El-Hasnony IM (2020)	Improved feature selection using binary GWO and PSO	Big data analytics	K-nearest neighbor classifier, Euclidean separation matrices, Tent chaotic map	Selected fewer features with higher accuracy and faster computation time compared to traditional PSO and GWO
Bhuyan HK (2021)	Analysis of subfeatures for classification	Big data mining in diverse fields	Lagrangian multipliers, variance- and domain-based sub-features, various classifiers	Effective generation of novel classes, enhancing classification accuracy
Yin Y (2020)	Dynamic data mining framework for sensor data	IoT and sensor networks	Historical data collection, parameter training, association analysis	Demonstrated efficacy in extracting dynamic data and identifying stable patterns in sensor environments
Rahman MM (2022)	Educational data mining for programming learning	Big data ICT industry, programming education	MK-means clustering, FP-growth algorithm	Effective in improving programming learning outcomes with insights from problem-solving data
Ishaq A (2021)	Prediction of heart failure survival using SMOTE	Cardiovascular patient survival	SMOTE, Random Forest, Extra Trees Classifier	Achieved high accuracy (0.9262) in survival prediction

				with effective feature selection and model optimization
Feng G (2022)	Analysis and prediction of academic performance	Education, student performance	Clustering, discriminant analysis, convolutional neural networks	Improved reliability and predictive accuracy in educational data mining applications
Haoxiang W (2021)	Data mining and perturbation algorithm	Health informatics	Optimal geometric transformations, privacy preservation techniques	Outperformed existing privacy techniques in accuracy, scalability, and data utility
Abdelkader HE (2022)	Assessing online learning satisfaction	Higher education during COVID-19	Feature Selection, predictive models	Enhanced classification accuracy and effective educational strategies with reduced feature dimensionality
X. Liu (2020)	Intelligent spectrum sensing for 5G	Wireless communications	Big data analytics, cooperative spectrum sensing	Improved spectrum management through comprehensive spectrum status derived from heterogeneous data
P. Pinoli (2018)	Distributed, location-based Apriori algorithm	Biological sequence pattern mining	Apache Spark, Apriori Property	Superior scalability and performance in genomic research with efficient mining of large-scale sequence data
L. Vaira (2016)	Multidimensional mining framework	Big healthcare data	Clustering, classification, multidimensional views	Enhanced precision in fetal growth monitoring and prenatal care
J. Z. Huang (2019)	Random Sample Partition model	Big data analysis	Distributed data blocks, probability	Effective in estimating statistical

			distribution	properties and constructing predictive models without processing entire datasets
K. Ogohara (2017)	Combining data and text mining	Questionnaire and review data analysis	Association analysis, keyword extraction, opinion mining	Effective integrated analysis of numerical and text data in a cohesive environment
J. Zhang (2018)	Privacy-Preserving and Security Mining Framework (PPSF)	Data privacy and security	Data anonymity, PPDM, PPUM	Enhanced data security and privacy with a user-friendly interface and effective execution of privacy-preserving operations

This table summarizes the contributions of each study, highlighting their proposed methods, contexts, techniques used, and Key Findings.

3. Conclusion

In this paper, the field of big data mining encompasses a diverse array of techniques and applications that are revolutionizing various sectors. Methods like natural language processing, data visualization and machine learning are essential for gleaning insightful information from large datasets. Applications span across healthcare, finance, cybersecurity, and beyond, where big data mining enhances decision-making, improves operational efficiency, and drives innovation. As the volume and complexity of data continue to grow, ongoing research and advancements in big data mining methodologies will further unlock hidden patterns and information, providing businesses with never-before-seen chances to get a competitive edge and successfully handle societal issues.

References

1. Zhou X, Liang W, Kevin I, Wang K, Yang LT. Deep correlation mining based on hierarchical hybrid networks for heterogeneous big data recommendations. IEEE Transactions on Computational Social Systems. 2020 May 8;8(1):171-8.
2. El-Hasnony IM, Barakat SI, Elhoseny M, Mostafa RR. Improved feature selection model for big data analytics. IEEE Access. 2020 Apr 7;8:66989-7004.
3. Bhuyan HK, Ravi V. Analysis of subfeature for classification in data mining. IEEE Transactions on Engineering Management. 2021 Aug 4;70(8):2732-46.
4. Yin Y, Long L, Deng X. Dynamic data mining of sensor data. IEEE Access. 2020 Feb 27;8:41637-48.
5. Rahman MM, Watanobe Y, Matsumoto T, Kiran RU, Nakamura K. Educational data mining to support programming learning using problem-solving data. IEEE Access. 2022 Mar 8;10:26186-202.
6. Ishaq A, Sadiq S, Umer M, Ullah S, Mirjalili S, Rupapara V, Nappi M. Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques. IEEE access. 2021 Mar 4;9:39707-16.
7. Feng G, Fan M, Chen Y. Analysis and prediction of students' academic performance based on educational data mining. IEEE Access. 2022 Feb 15;10:19558-71.
8. Haoxiang W, Smys S. Big data analysis and perturbation using data mining algorithm. Journal of Soft Computing Paradigm (JSCP). 2021 Apr 19;3(01):19-28.
9. Abdelkader HE, Gad AG, Abohany AA, Sorour SE. An efficient data mining technique for assessing satisfaction level with online learning for higher education students during the COVID-19. IEEE Access. 2022 Jan 13;10:6286-303.
10. X. Liu, Q. Sun, W. Lu, C. Wu and H. Ding, "Big-Data-Based Intelligent Spectrum Sensing for Heterogeneous Spectrum Communications in 5G," in IEEE Wireless Communications, vol. 27, no. 5, pp. 67-73, October 2020, doi: 10.1109/MWC.001.1900493.
11. E. Stamoulakatou, A. Gulino and P. Pinoli, "DLA: a Distributed, Location-based and Apriori-based Algorithm for Biological Sequence Pattern Mining," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 1121-1126, doi: 10.1109/BigData.2018.8622007.
12. M. Bochicchio, A. Cuzzocrea and L. Vaira, "A Big Data Analytics Framework for Supporting Multidimensional Mining over Big Healthcare Data," 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, USA, 2016, pp. 508-513, doi: 10.1109/ICMLA.2016.0090.
13. S. Salloum, J. Z. Huang and Y. He, "Random Sample Partition: A Distributed Data Model for Big Data Analysis," in IEEE Transactions on Industrial Informatics, vol. 15, no. 11, pp. 5846-5854, Nov. 2019, doi: 10.1109/TII.2019.2912723.

14. T. Matsumoto, W. Sunayama, Y. Hatanaka and K. Ogohara, "Data Analysis Support by Combining Data Mining and Text Mining," 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), Hamamatsu, Japan, 2017, pp. 313-318, doi: 10.1109/IIAI-AAI.2017.165.
15. J. C. -W. Lin, P. Fournier-Viger, L. Wu, W. Gan, Y. Djenouri and J. Zhang, "PPSF: An Open-Source Privacy-Preserving and Security Mining Framework," 2018 IEEE International Conference on Data Mining Workshops (ICDMW), Singapore, 2018, pp. 1459-1463, doi: 10.1109/ICDMW.2018.00208.