International Journal of Computer Sciences and Engineering
Vol.**11**, Issue **4**, pp.**46-54**, April **2023**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Review Paper

Literature Review on Tools & Applications of Data Mining

Anshu Shrivastava [1*], **Jay Kumar Jain** [2], **Dipti Chauhan** [3]

[1,2]SIRT, Bhopal, India
[3]CSE, PIEMR, Indore, India

**Corresponding Author: anshushrivastavasirt@gmail.com*

Received: 19/Feb/2023; **Accepted**: 20/Mar/2023; **Published**: 30/Apr/2023. | **DOI**: https://doi.org/10.26438/ijcse/v11i4.4654

Abstract: There are many new disciplines that have emerged as a result of technological advancement. Every day, enormous amounts of data are produced in many different areas, including science, engineering, health, and business. Data mining is a technique for gathering data from various sources and organizing it to produce insightful conclusions. Companies' today look to stay ahead of the competition by making it a priority to keep up with all new developments in data science and analytics. This paper explains data mining's applications in various areas as well as its methods and tools. This study concentrated on different data mining tools that are beneficial and identified as key fields of data mining technology. We are aware of the numerous domestic and international businesses, as well as small and big organizations.

Keywords: Data mining, Data set, KDD, Tools & Techniques, Rapid Miner, Weka, KNIME

1. Introduction

The process of organizing and extracting data from multiple sources in order to gain useful insights is known as data mining. In data science, one of these processes is data mining. Data mining, also known as Knowledge Discovery in Databases (KDD), is the process of finding patterns in a large set of data and data warehouses.[4] The process begins with providing the data mining tools with a certain amount of data, which then use statistics and algorithms to display the reports and patterns. These tools make it possible to see the results, which can be understood and used to make changes and improvements to businesses. One of the many analytical tools for analyzing data is data mining software. Data can be categorized, analyzed from a variety of perspectives, and relationships can be summarized using this tool.

Organizations use data mining broadly to develop marketing strategies, hospitals use it to develop diagnostic tools; e-commerce uses it to cross-sell products through websites, and many other applications. Information about how their current and potential customers act in the data they have gathered. Data mining is the process of analyzing data from various perspectives and shortening it into useful information that can be used to increase revenue, cut costs, or both [6]. In general, it is the process of discovering information within the data that queries and reports can't effectively reveal. Continue reading to learn about the numerous data mining applications that are transforming industries. A data cavity is no longer an option for modern businesses. To stay ahead of the competition, they must adapt and keep up with technological advancements and upcoming digital trends. [5]

Due to these challenges processing data and data mining techniques becomes a critically important role in development technology [2].Classification is one of the most useful techniques in data mining that classify data into structured class or groups. It helps the user in discovering the knowledge and future plan.

1.1 DATA MINING (Size 10 Bold)

DATA MINING is similar to extracting necessary data from deep data. Data mining is an important step in Knowledge Discovery in Databases (KDD) [1], [2]. KDD is the process of retrieving necessary information from large databases or data marts and converting the data into various patterns, summary reports, views, etc. Data mining automates the detection of relevant patterns in a database by using defined approaches and algorithms to look into current and historical data [7]. The stages of KDD are as follows:

1. Extraction of Data: We select data sets or data samples from large repositories or databases
2. Cleaning the Data: It's possible that the large databases' data set is incorrect. Using data transformation tools, we must simplify the data by removing inconsistent data and missing values.
3. Integration of Data: Use a variety of tools for data synchronization and migration to combine and store data from multiple sources in a single location.
4. Selection of Data: Utilizing methods such as Naive Bayes, Clustering, Neural-networks, and Regression, and

decision trees, and constructive data is extracted from the data source for the analysis.

5. Transformation of Data: During the mining procedure, summary and aggregation operations are used to transform the data into the appropriate format.
6. Exploiting data: There are a few ways that data can be turned into significant patterns.
7. Evaluation of Patterns: Interestingness measures are utilized to identify useful patterns. We employ visualization and summarization techniques in these steps.
8. Informational Presentation: Reports, tables, and other visualization tools are used to display the results of data mining.

The primary step in the data mining process is selecting a dataset from a vast repository [3, 7]. Dataset classification can be broken down into three categories: sequential data, graph-based data, and record data. Market-basket data, for instance, is graph-based data. Data with relationships between objects, such as linked web pages, are graph-based data. A sequence of events is called sequential data.

1.2 What Can Data Mining Do?

Companies in a wide range of industries, including retail, finance, health care, manufacturing, transportation, and aerospace, are already utilizing data mining tools and methods to benefit from historical data, despite the fact that data mining is still in its infancy. Data mining helps analysts recognize significant facts, relationships, trends, patterns, exceptions, and anomalies that might otherwise go unnoticed[3][5]. Data mining can help spot trends in sales, develop smarter marketing campaigns, and accurately predict customer loyalty. It does this by sifting through warehoused information using pattern recognition technologies and statistical and mathematical techniques. . Data mining is used for the following specific purposes:

a) **Market-Segmentation**: Find out what makes customers who buy the same products from your company the same.
b) **Customer-Churn** - Determine which customers are most likely to switch to a competitor.
c) **Fraud-detection** - Identify the most likely fraudulent transactions.
d) **Direct-Marketing**: Determine which prospects ought to be added to a mailing list in order to get the most responses.
e) **Predict** what each visitor to a website will likely is interested in viewing through interactive marketing.
f) **Market Basket Analysis**: Recognize which goods or services are frequently purchased in combination; such as diapers and beer.
g) **Trend-Analysis**: Find out how this month's typical customers differ from last month's.

In view of the highlights of occasions, three sorts of successive information we have, to be specific Time-series information, Emblematic grouping information and Natural arrangement information.

This paper is organized as follows. Section 1 contains the Introduction of the Data Mining. Section 2 contains the different techniques of Data Mining. Section 3 contains the various application of Data mining. Next Section 4 contains the data mining tools and their comparison and section 5 shows the related work. And finally the last section the conclusion and future scope.

2. DATA MINING TECHNIQUES

Based on the type of data mining task, the best methods are used. Descriptive data mining tasks classify the characteristics of the data in the target dataset according to past or recent events. Based on past data, prediction tasks provide search results in the future. Some common data mining techniques include association rules, classification, clustering, prediction, regression, outlier detection, and sequential patterns. Predictive data mining techniques include outlier detection, regression and classification. Descriptive data mining techniques include association rules, clustering, and sequential pattern discovery. Data mining uses some central methods based on data mining and retrieval work.

2.1 Classification

Classification is used in our model to identify and assign a class to the new observation input data. By describing multiple attributes to identify a particular class, classification can help us determine the type of customer, item, or object. Data is divided into two sets: The training set, which was used to build the model, and Set of tests (used to verify the model)[10].

Diverse classes are separated using the data from the Training set. The model we create assigns a class to the test set data. This method makes use of Decision Trees, Bayesian Classifiers, Neural Networks, Support Vector Machines, K-Nearest Neighbour, Linear Regression, and Logistic Regression as classifiers. By identifying various attributes, such as the number of seats, car shape, and driven wheels, for instance, you can easily classify automobiles into various categories (sedan, 4x4, convertible). By comparing the features of a new car to our established definition, you can put it into a specific class. Customers can be categorized using the same principles, such as age and social group. It can be used for direct marketing, cataloguing sky surveys, detecting fraud and other things.

Furthermore, you can involve grouping as a feeder to, or the consequence of, different procedures. Decision trees, for instance, can be utilized to establish a classification. Utilizing common attributes across various classifications, you can identify clusters with clustering.

2.2. Clustering

The Clustering technique divides the data into groups or clusters, with objects in one cluster required to have features that are similar to those in the other, and objects in other clusters required to be less similar to one another. You can form a structure opinion by grouping different pieces of data together by looking at one or more attributes or classes. Simply put, clustering is the process of using one or more attributes to locate a group of related results.

Because it correlates with other examples, clustering makes it possible to see where the similarities and ranges agree. This makes it useful for identifying various pieces of information. We can assume that there is a cluster at a particular point and then use our identification criteria to determine if you are correct. Clustering can work in either direction.

An excellent illustration of various clustering representations can be seen in the graph in Figure. Different clustering techniques are utilized based on the application. A portion of the applications are Report bunching, Market division, Science, clinical imaging, Informal community investigation and so forth. Some of alliance technique is Dividing Strategy, Framework Based Technique, Thickness based Technique, Model-Based Technique, Progressive Technique, and Limitation based Strategy.

2.3. Sequential Patterns

The sequential patterns technique is used to predict sub- and sequential dependencies. GSP (Generalized Sequential Pattern), Free span, Prefix span, and SPADE (Sequential Pattern Discovery Using Equivalent Class) are some of the methods used to find sequential patterns. DNA sequences, blog click streams, telephone call patterns, stocks and markets, and other applications are among the applications.

Sequential patterns are a useful method for identifying trends, or regular occurrences of similar events, over longer-term data. With customer data, for instance, you can determine that customers buy a particular set of products together at various times of the year. You can use this information to automatically suggest that certain items be added to a shopping basket based on their frequency and previous purchasing history in an application for shopping baskets.

2.4. Association Rule Mining

The technique of mining association rules finds patterns in data and relationships between large data sets and correlations between them. The existence of an object can be predicted based on the existence of other objects. Association rules are if-then rules that calculate leverage, support, and reliability to identify regular patterns and relationships between objects. Some association rule algorithms are Apriori algorithm, FP growth algorithm, Eclat algorithm, market basket analysis, cross marketing, catalog planning use this technique.

An association rule is a rule that requires certain relationships between a set of database objects (e.g. "occur together") . Given a set of events, where each event is a set of labels, an association rule of the form “X->Y” is an expression where X and Y are sets of elements. The spontaneous meaning of such a rule is that transactions in the database containing X will tend to contain Y. An example of an association rule is: "30% of farmers who grow vegetables also grow fruits; 2% of all farmers grow both. Here we have 30% rule confidence and 2% rule support. The problem is to find all association rules that satisfy the user-specified minimum support and minimum confidence limits.

Association is probably the most renowned and the simplest data mining technique. Here you can create a simple correlation between two or more elements, often of the same type, to identify patterns. For example, by observing people's buying habits, it can be recognized that the customer always buys mobile cove when buying mobile phone, and therefore recommends buying mobile cove the next time he buys Mobile phone. Building connection or relation, data mining tools can be implemented easily with other different tools.

2.5. Outlier Detection

Value detection identifies and excludes outliers (sample data that behaves completely differently from other data) from a dataset. Some outlier detection methods include Z-score, DBSCAN, isolation forest, linear regression model (LMS, PCA), proximity-based models (non-parametric), and high-dimensional outlier detection methods. Some programs are for fraud detection shown in Fig.1.

Anomaly detection usually occurs in the investigative data analysis phase of the project data management process, and our decision to address them determines how well or badly the model performs for a given business problem. The presence of deviations greatly affects the model and thus the entire workflow.

They can be critical in data analysis for at least two reasons:

- Outliers can negatively distort the overall analysis result.
- Outsiders may be just what you're looking for, and that's when you need to talk to a domain expert.

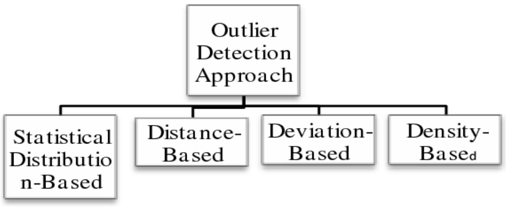

Fig.1. Outlier Detection Approach

2.6. Regression

Based on the response variable, the regression approach forecasts the value of a continuous variable termed the predictor variable (target-whose values are already known). The Simple Linear Regression Model, Lasso Regression, Logistic Regression, Support Vector Machines, Multivariate Regression method, and Multiple Regression Algorithm are a few of the regression algorithms used in data mining. For instance, there is a correlation between irresponsible driving and traffic accidents. Applications include predicting sales, finances, marketing, trend analysis, time series, calculating the age of fossils, etc.

Machine learning techniques that are used in supervised learning models include logistic regression and linear regression. Both employ labeled data to make predictions because they are both supervised models. Contrasting with logistic regression, which may be applied to both classification and regression issues but is more frequently employed as a classification procedure, linear regression is used for regression or to predict continuous values. In order to project value based on distinct features, regression models are used. Linear regression is used to fit a (linear) relationship between a continuous response variable and a set of predictor variables shown in Fig.2. However, if the response variable is binary (that is, yes/no), linear regression is not suitable.

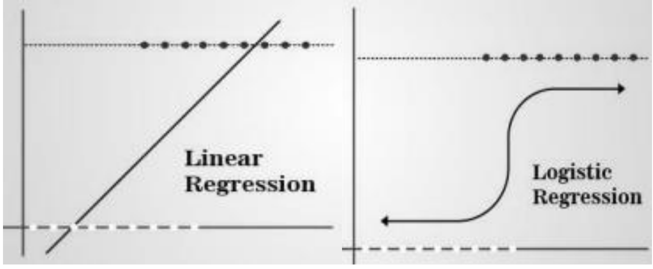

Fig.2. Machine Learning-Linear & Logistic Regression

2.7. Prediction

An estimation or classification can be applied to any forecast. The distinction is in the emphasis. We don't anticipate being able to go back and check if a data mining classification of a phone line as being predominantly utilized for internet access or a credit card transaction as being fraudulent was accurate. Our classification may be accurate or inaccurate, but there is no way to know for sure because the key events have already happened in the real world. The local ISP may or may not be called frequently on the phone. It is possible to check if you put in enough effort.

In Fig 2 depict Prediction tasks feel different because datasets are categorized according to predicted future behaviour or estimated future values. In prediction, the only way to validate classification accuracy is to wait and see. Some examples of prediction tasks are:

- Predict which customers will churn within the next 6 months.
- Predict which subscribers will order value-added services such as three-way calling and voicemail.

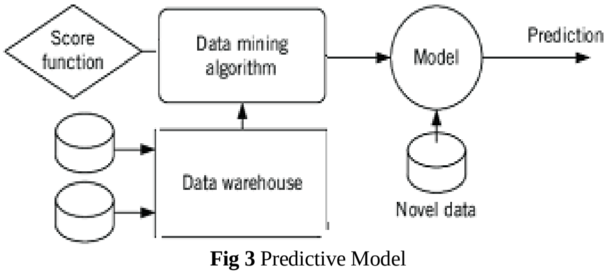

Fig 3 Predictive Model

Any of the techniques used for classification and estimation can be adopted for use in prediction by using training examples where the value of the variable to be predicted is already known, along with historical data for those examples. The historical data is used to build a model that explains the current observed behavior. When this model is applied to current inputs, the result is a prediction of future behavior depicted in fig 3.

2.8. Decision trees

Decision trees are commonly used in classification systems to relate type information, and prediction systems where different predictions are made based on past experience to help drive the structure and output of the decision tree depicted in fig 4.

Decision trees can be used as part of selection criteria or to support the use and selection of specific data within the forest. Within the decision tree, start with a simple question with two (and possibly more) answers. Each answer leads to different questions that help classify or identify the data. This allows you to classify your data and make predictions based on each response. The top decision node in the tree corresponding to the best predictor, called the root node. Decision trees can handle both categorical and numeric data.

2.9 Statistics

The problem of extracting knowledge from data was tackled by statisticians long before the first papers on artificial intelligence were published. For example, correlation analysis applies statistical tools to analyze correlations between two or more variables. Cluster analysis provides a way to find clusters in a large set of objects described by a vector of values. Factor analysis attempts to reveal the most important variables that explain clusters. Common techniques used for supervised classification tasks include Linear Discriminates, Quadratic Discriminates, K-nearest Neighbor, Naïve Bays, Logistic Regression and CART.

2.10 Machine Learning

Statistical methods have difficulty incorporating subjective, non-quantifiable information into models. Also, different distributions of parameters and independence of attributes should be assumed. Various studies have concluded that machine learning offers comparable (and often better) prediction accuracy. The superior performance compared to statistical methods may be attributed to the fact that they are freed from the underlying parametric and structural assumptions of statistical methods. Another weakness of statistical approaches to data analysis is the problem of interpreting results. Some of the machines learning techniques are listed below.

- Supervised Machine Learning.
- Unsupervised Machine Learning.
- Semi-Supervised Machine Learning.
- Reinforcement Learning.

2.11 Neural Networks

Artificial neural networks are computational models that consist of many nonlinear processing elements arranged in patterns similar to biological neural networks. A typical neural network has an activation value associated with each node and a weight value associated with each connection. Activation functions coordinate the triggering of nodes and the propagation of data through network connections with massive parallelism. The network can also be trained using examples by connection weight adjustment. A data mining neural network is a classification method that takes an input, trains it to recognize patterns in the input data, and predicts the output for new inputs of similar nature. Neural networks form the basis of deep learning, a subfield of machine learning that belongs to artificial intelligence.

3. Applications of Data Mining

Many different fields use data mining techniques for technical, commercial, and research purposes. Below is an overview of the most widely used data mining techniques and tools in various applications.

3.1 Financial Analysis

The banking and financial industry depends on high quality, reliable data. In the credit market, financial and user data can be used for a variety of purposes, including: B. for predicting loan payments and determining credit ratings. Data mining techniques also make such tasks more manageable.

Classification techniques facilitate the separation of important factors that influence a customer's banking decisions from irrelevant ones. In addition, multidimensional clustering methods can identify customers with similar loan payment behaviours. Data analysis and mining can also help detect money laundering and other financial crimes.

3.2 Telecommunication Industry

Telecom Industry is expanding and growing rapidly, especially with the advent of the Internet. Data mining enables key industry players to improve their service quality and gain an edge. Pattern analysis of Spatial-temporal databases can play an important role in mobile communications, mobile computing, and web and information services. Also, techniques such as outlier analysis can help you find fraudulent users. OLAP and visualization tools also help compare information such as user group behavior, profits, traffic, and system overload.

3.3 Intrusion Detection

Global connectivity in today's technology-driven economy poses security challenges to network management. Network resources can be subject to threats and actions that compromise their confidentiality and integrity. Intrusion detection has therefore become an important data mining practice.

This includes association and correlation analysis, aggregation techniques, visualizations, and query tools that can effectively detect anomalies and deviations from normal behavior.

3.4 Retail Industry

Organized retail has a plethora of data points covering sales, purchase history, product delivery, consumption, and customer service. The database is even bigger with the rise of e-commerce marketplaces. In modern retail, data warehouses are designed and built to take full advantage of data mining. Multidimensional data analytics help you process data related to different types of customers, products, regions, and time zones. Online retailers can also recommend products to increase sales and analyze the effectiveness of advertising campaigns. From identifying purchasing patterns to improving customer service and satisfaction, data mining opens many doors in this area.

3.5 Higher Education

As the demand for higher education increases worldwide, educational institutions are looking for innovative solutions to meet the increasing demand. Institutions can use data mining to predict which students will enroll in a particular program and who will need additional graduation support, improving overall enrolment management.

Effective analytics also make it more convenient to predict student paths and present data. Data mining techniques can thus help uncover hidden patterns in the vast databases of higher education.

3.6 Energy Industry

Today, big data is also available in the energy sector, demonstrating the need for suitable data mining techniques. Decision tree models and support vector machine learning are among the most popular approaches in the industry, providing viable solutions for decision-making and management. In addition, data mining can also generate productive profits by predicting power generation and reimbursing electricity bills.

3.7 Spatial Data Mining

Geographic information systems (GIS) and several other navigation applications use data mining to protect and make sense of important information. This new trend involves the extraction of geographic, environmental, and astronomical data, including imagery from space. Spatial data mining can typically reveal aspects such as topology and distance.

3.8 Biological Data Analysis

The practice of biological data mining is common in genomics, proteomics, and biomedical research. From characterizing patient behaviour and predicting doctor visits to identifying treatments for patient ailments, data science techniques have many benefits.

- Some of the data mining applications in bioinformatics are listed below.
- Semantic integration of heterogeneous and distributed databases
- Association and path analysis
- Using visualization tools
- Structural pattern recognition
- Analysis of gene networks and protein pathways

3.9 Other Scientific Applications

Rapid numerical simulations in scientific fields such as chemical engineering, fluid dynamics, climate and ecosystem modelling generate huge data sets. Data mining brings capabilities such as data warehousing, data pre-processing, visualization, and graph-based mining.

3.10 Manufacturing Engineering

System-level design uses data mining to extract relationships between portfolio and product architectures. In addition, this method is also suitable for predicting product cost and development time.

3.11 Criminal Investigation

Data mining activities are also used in criminology to study the characteristics of crime. First, we need to convert the text-based crime report into a word processing file. The process of identification and crime handling then takes place by discovering patterns in vast data stores.

3.12 Counter-Terrorism

Sophisticated mathematical algorithms can indicate which intelligence unit should lead counter-terrorism operations. Data mining is also useful for police management tasks such as: B. When locating employees and marking searches at border crossings.

3.13 Market Basket Analysis

Shopping basket analysis [5], [6] is used to predict customer behaviour in retail. It's based on the theory that if you buy one set of products, your customers are likely to buy another set of products. Here, the association rule mining technique is applied. We will contribute to sales increase and design store layout according to customer's purchasing behavior. Data mining tools used in this area include R, SAS (Statistical Analysis System), MEXL, and XLMINER.

3.14 Education

Educational data mining is an emerging field focused on developing methods to find desired information from various educational domains [12], [16]. Data mining applications in this area predict student outcomes, student learning behavior, find weak students, and more. Student learning patterns are used to develop teaching methods. Records Records are used in educational applications. Data mining tools used in education include SPSS, KEEL, Weak, and Spark MLLib.

3.15 Web Mining

Web mining uses data mining techniques to discover related web documents and website patterns [13], [14].

Classification, clustering, and regression techniques are web content mining (to extract useful information from web documents), web structure mining (to discover structural information from web sites), and web usage mining (log mining). Used in applications such as The data mining tools used here are SAS (Statistical Analysis System), Scrapy, Page Rank, etc.

4. Data Mining Tools

4.1 RapidMiner

- RapidMiner is a free, open source data science platform that offers hundreds of algorithms for data preparation, machine learning, deep learning, text mining, and predictive analytics.
- Drag-and-drop interface and pre-built models allow non-coders to intuitively create predictive workflows for specific use cases such as fraud detection and customer churn. In the meantime, programmers can use her Rapid Miner’s R and Python extensions to customize data mining.
- After creating a workflow to analyze your data, visualize your results in RapidMiner Studio to find patterns, outliers, and trends in your data.
- Last but not least, the platform has a large and enthusiastic community of users who are always ready to help.

4.2 IBM SPSS Modeler

- IBM SPSS Modeler is a data mining solution that enables data scientists to accelerate and visualize the data mining process. Advanced algorithms can be used by users with little or no programming experience to build predictive models with a drag-and-drop interface.
- IBM's SPSS Modeler enables data science teams to import large amounts of data from multiple sources and rearrange it to uncover trends and patterns. The standard version of this tool handles numeric data in spreadsheets and relational databases. To add text analytics functionality, you need to install the premium version.

4.3 Weka

- Weka is open source machine learning software with a huge collection of algorithms for data mining. Developed by the University of Waikato, New Zealand and written in JavaScript. Easy-to-use graphical interfaces Easy-to-use graphical interface supporting various data mining tasks such as preprocessing, classification, regression, clustering and visualization various data mining tasks such as preprocessing, classification, regression, clustering and visualization. For each of these tasks, Weka provides built-in machine learning algorithms. This allows you to quickly test ideas and deploy models without writing code. Maximizing these benefits requires in-depth knowledge of the various algorithms available so that you can choose the right algorithm for your particular use case. Weka was originally developed to analyze data in the agricultural sector. Today, it is mainly used by researchers, industrial scientists, and for educational purposes**.**

4.4 KNIME

- KNIME is a free, open source data mining and machine learning platform. An intuitive user interface lets you create end-to-end data science workflows, from modeling to production. And various pre-built components allow you to quickly model without typing a single line of code. A number of powerful extensions and integrations make KNIME a versatile and scalable platform for handling complex data types and using advanced algorithms. KNIME enables data scientists to create applications and services for analytics or business intelligence. For example, common use cases in the financial industry include credit scoring, fraud detection, and credit risk assessment.

4.5 Orange

- Orange is a free, open-source data science toolbox for developing, testing, and visualizing data mining workflows. It is component-based software with a large collection of off-the-shelf machine learning algorithms and text mining add-ons. It also has advanced features for bioinformaticians and molecular biologists. Orange also allows for interactive data visualization, offering a number of graphs such as silhouette plots and sieve graphs. And even non-programmers can perform data mining tasks using visual programming with a drag-and-drop interface. Developers, on the other hand, can choose to mine data with her Python**.**

4.6 Apache Mahout

- Apache Mahout is an open source platform for building scalable machine learning applications. Its purpose is to allow data scientists and researchers to implement their own algorithms. Written in JavaScript and based on Apache Hadoop, the framework focuses on three main areas: Recommender engine, clustering, and classification. This is suitable for complex, large-scale data mining projects dealing with huge amounts of data. In fact, it's used by major web companies like LinkedIn and Yahoo. Apache Mahout is free to use under the Apache license and supported by a large user community.

4.7 SAS Enterprise Mining

- SAS Enterprise Miner is an analytics and data management platform. The goal is to simplify the data mining process and enable analysts to transform big data into insights. Through an interactive graphical user interface (GUI), users can quickly generate data mining models and use them to solve critical business problems. SAS offers a variety of algorithms for data preparation and exploration, and for creating advanced predictive and descriptive models. Businesses can use SAS Enterprise Mining for fraud detection, resource planning, improving marketing campaign response rates, and more.

5. Related Work

Jindal and Liu (2007) [1] propose mining opinions from product reviews, forum posts, and blogs as an important research topic with many applications. Existing research has focused on extracting, classifying, and summarizing opinions from these sources. Although no research has been published on this topic yet, website spam and email spam have been extensively studied. Review spam is very different from website spam or email spam and requires different detection techniques.

Horse et al. (2009) [2] suggest that detection and filtering are still the most viable ways to combat spam email. There are many spam mail filters in operation that are quite successful. Proactively detecting new types of spam without prior knowledge remains a major challenge. Negative selection is a branch of the artificial immune system. It has strong temporal properties and is especially useful for discovering unknown temporal patterns. This property makes it an excellent candidate for quickly detecting and detecting new types of spam emails.

Wang and Liu (2010) [3] presented various approaches to solve the problem of spam propagation. Most of these approaches cannot be flexibly and dynamically adapted to spam. A new approach to combat spam is proposed, based on the detection of reliable behaviour during transmission sessions. Behavioural detection of email delivery patterns to allow regular servers to detect malicious connections before the email body is delivered. An integrated anti-spam framework was developed that combines reliable behaviour detection and Bayesian analysis. The effectiveness of both trusted behaviour detection and built-in filters were evaluated.
Algar et al. (2010) [4] states that inference mining from product reviews, forum posts, and blogs is an important research topic with many applications today. Previous research has focused on classifying and summarizing these online opinions. An important issue related to the reliability of online opinions was largely ignored. No studies have been reported that assessed the reliability of the reviews. This is important for all opinion-based applications, but web spam and email spam have been extensively studied, and both duplicate and near-duplicate reviews are classified as spam reviews, partially related, and classified as proprietary.

A review rating that was classified as a non-spam review was suggested. Chen et al. (2010) [5] has discussed Email as a kind of semi-structured document, and using spam-specific features could improve the email classification results. The decision tree data mining technique to dig out the potential association rules among these attributes of email, and then to identify unknown email's category based on these rules has been applied. The efficiency of the method is not lower than that of other existing methods of checking whole email content text. Salama et al. (2012) [6] has presented a comparison among the different classifiers decision tree (J48), Multi- Layer Perception (MLP), Naive Bayes (NB), Sequential Minimal Optimization (SMO), and Instance Based for K-Nearest neighbour (IBK) on three different databases of breast cancer (Wisconsin Breast Cancer (WBC), Wisconsin Diagnosis Breast Cancer (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC)) by using classification accuracy and confusion matrix based on 10-fold cross validation method.

A fusion at classification level between these classifiers to get the most suitable multi-classifier approach for each data set has been introduced. The experimental results have shown that in the classification using fusion of MLP and J48 with the PCA is superior to the other classifiers using WBC data set. The PCA has been used in WBC dataset as a features reduction transformation method which combines a set of correlated features. All experiments have been conducted in WEKA data mining tool.

Liu and Yang (2012) [7] has proposed that the spam messages in mobile phone were flooded and the management to it was not effective. The characteristics of spam messages, its forming reason and its harm has been analysed. It has been discussed that the classification method of filtering spam messages, and points out it is the key work of the researchers to develop the more effective classification method.

6. Conclusion and Future work

This article introduced various algorithms for data mining. The overall goal is to evaluate the use of data mining algorithms for various types of systems. Each algorithm has a different purpose and purpose of classifying the dataset in a different way. As with clustering-based algorithms, clusters can evolve from homogeneous data items in a given data set. B. A gender attribute that divides the dataset into males and females. Also, a priori evaluation of relationships between attributes, such as bonus attributes, has a significant impact on an individual's monthly income. In the near future, we plan to use various data mining algorithms to efficiently detect email spam from specific email sets. Implementation of specific behaviours is done using popular data mining tools such as Weka, MATLAB, and Web-Miner.

Every area is now digitized, generating massive amounts of data every day. Data mining plays a key role in managing, analyzing, and extracting the required information from these large databases. It provides an overview of various applications of data mining and the techniques and tools used in each application.

References

[1] Aarti Sharma, Rahul Sharma, Vivek Kr. Sharma, Vishal Shrivastava, ―Application of Data mining-A Survey Paper‖ in International Journal of Computer Science and Information Technologies, Vol. 5 (2), 2023-2025 2014, ISSN: 0975-9646.

[2] Bharati M. Ramageri, ―Data Mining Techniques and Applications‖ in Indian Journal of Computer Science and Engineering, Vol. 1 No. 4 301-305, Dec 2014.

[3] Saima Anwar Lashari, Rosziati Ibrahim, Norhalina Senan, N. S. A. M. Taujuddin,‖ Application of Data Mining Techniques for Medical Data Classification: A Review‖ in MATEC Web of Conferences 150, 06003, (2018),UCET2017.

[4] D.Usha Rani, ―A Survey on Data Mining Tools and Techniques in Medical Field‖ in International Journal of Advanced Networking & Applications (IJANA), Volume: 08, Issue: 05 Pages: 51-54 (2017) Special Issue.

[5] Manpreet Kaura, Shivani Kanga, ―Market Basket Analysis: Identify the changing trends of market data using association rule mining‖ in International Conference on Computational Modeling and Security (CMS 2016), Procedia Computer Science 85 (2016) 78 – 85.

[6] Dr. M. Dhanabhakyam, Dr. M. Punithavalli, ―A Survey on Data Mining Algorithm for Market Basket Analysis‖ in Global Journal of Computer Science and Technology Volume 11 Issue 11 Version 1.0 July 2011, Publisher: Global Journals Inc. (USA) Online ISSN: 0975-4172 & Print ISSN: 0975-4350.

[7] MohammadReza Keyvanpoura, Mostafa Javidehb, Mohammad Reza Ebrahimia,‖ Detecting and investigating crime by means of data mining: a general crime matching framework‖ in Procedia Computer Science 3 (2011) 872–880. Available:

[8] Chauhan, Dipti, Jay Kumar Jain, and Sanjay Sharma. "An end-to-end header compression for multihop IPv6 tunnels with varying bandwidth." 2016 Fifth international conference on eco-friendly computing and communication systems (ICECCS). IEEE, 2016.

[9] Jain, Jay Kumar, Devendra Kumar Jain, and Anuradha Gupta. "Performance analysis of node-disjoint multipath routing for mobile ad-hoc networks based on QOS." International Journal of Computer Science and Information Technologies 3.5 (2012): 5000-5004

[10] Pushpesh Pant, SriramPandey, ―Application of Data Mining Tools and Techniques in Material Selection‖ in International Journal of Scientific & Engineering Research, Volume 8, Issue 4, April-2017.

[11] V.K. Jha, R.K. Singh ―Application of Data Mining in Manufacturing Industry‖ in International Journal of Information Sciences and Application.. Seoul, Korea, Apr 8, 2014. In: PAPADOPOULOS, Symeon, ed. and others. Proceedings of the SNOW 2014 Data Challenge co-located with 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, pp. 57-63, April 8, 2014 ISSN 0974- 2255 Volume 3, Number 2 (2011), pp. 59-64.

[12] Ashish Dutti, Maizatul Akmar Ismaili, Tutut Herawan, ― A Systematic Review on Educational Data Mining‖ in Digital Object Identifier 10.1109/ACCESS.2017.2654247, Volume 5, 2017.

[13] Dr. S. Vijiyarani, Ms. E. Suganya, ―Research Issues in Web Mining‖ in International Journal of Computer- Aided Technologies (IJCAx) Vol.2, No.3, July 2015.Brijendra Singh, Hemant Kumar Singh, ―Web Data Mining Research: A Survey‖ in IEEE International conference.

[14] Abdullah H. Wahbeh, Qasem A. Al-Radaideh, Mohammed N. Al - Kabi, and Emad M. Al.

[15] Jain, Jay Kumar, and Sanjay Sharma. "Performance Evaluation of Hybrid Multipath Progressive Routing Protocol for MANETs." International Journal of Computer Applications 71.18 (2013).

[16] Smitha, T., & Kumar, V. S. (2013). Applications of big data in data mining. International journal of emerging technology and advanced engineering, 7(3).

AUTHORS PROFILE

Anshu Shrivastava is an Assistant Professor in Department of Computer Applications at Sagar Institute of Research &Technology, Bhopal. She is currently pursuing PhD. She has published several research papers in National & International Conferences and having 15 years of teaching experience.

Dr. Jay Kumar Jain is currently working as an Associate Professor in CSE Department & MCA (Head), Sagar Institute of Research & Technology, Bhopal. He did his Ph. D. from Maulana Azad National Institute of Technology, Bhopal in 2015. He was awarded with Research Fellowship by MHRD for completing his Ph.D. He has research as well as teaching experience of about 15 years. He has published around 50+ research papers and book chapters in SCI/SCOPUS/Referred International/National Journal and Conferences. He has also granted 2 International Patents and published 2 Indian patents and also registered 2 copyrights in IPR, India. He has been a reviewer in many international journals/conferences including Elsevier, IEEE Access, and Springer. He has lifetime membership of various professional societies such as CSI, Franklin, IDES, IAENG, SDIWC, and many more. He has also working as an innovation Coordinator and Ambassador in Institution's Innovation Council (IIC), SIRT Bhopal, under Ministry of Education (MoE), Govt. of India. His research interests include Wireless Sensor Networks, Internet of Things, and Mobile Ad hoc Networks.

Dr. Dipti Chauhan is working as Associate Professor in Computer Science & Engineering Department at Prestige Institute of Engineering & Research., Indore. She did her Ph. D. from Maulana Azad National Institute of Technology Bhopal in 2016. She is having an overall teaching and research experience of about 15 years. She has published papers in International/National journals and International/National conferences of repute. Her research interests include wireless networks, advanced computer networks & the Internet of Things. She is the Certified Network Engineer from University Sains Malaysia, NAV 6.