International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

ANALYTICAL STUDY ON ROLE OF WEKA IN DATA MINING FOR PERFORMANCE MEASURE

RAKHI MATHUR
Research Scholar of Mewar University, Chittorgarh Rajasthan
DR. GANESH GOPAL VARSHNEY
Professor (Supervisor), Mewar University, Chittorgarh Rajasthan

Abstract
Data mining's limitless uses and strategies to deal with mining the data in an appropriate manner are making it more popular in a wide variety of academic disciplines. Given the state of affairs in the globe at the moment, this seems like a good method for making educated guesses about what will happen in the not-too-distant future. While a wealth of data has become available as healthcare research has advanced, the real challenge is in figuring out how to turn that data into actionable insights. This breakthrough can be unfurled with the use of data mining. Healthcare systems have a great deal of room to improve their data use via the application of data mining techniques. Therefore, it improves the intellect and reduces expenses. In this work, we take a broad look at the applications of Data Mining TOOLs including clustering, association analysis, and regression in the medical field. Data mining in healthcare is discussed, along with its current and potential uses, challenges, and developments.
Keywords: Data, Mining, Clustering, Healthcare

INTRODUCTION

Data Mining is the core innovation of KDD. In data mining, efficient algorithms are used to filter out interesting outliers and patterns. The quantity and variety of available data sources continues to grow, and data objects themselves have improved in many ways. As a result, innovative data mining techniques are required to fully capitalize on the data deluge. This section describes and illustrates the KDD process. Then, the main tasks involved in data mining are examined. Since the proposal oversees these tasks, clustering and organization are analyzed in depth. Not long after, the option of using compound data objects for the representation of complicated articles is introduced. Finally, a graphic sums up the plan, providing a visual summary of the points made.

Growing IT applications have shown IT's proven potential in many real-world contexts, including business, academic research, social and natural challenges. Due to the rapid computerization of daily interactions in these places, a massive amount of unrefined data with strong associations has been collected. One of the main benefits to the owner is this information. It's loaded with information on customers that can be used to improve things like marketing, politics, government policy, and product quality. Knowledge discovery in databases, or KDD, is the act of locating useful samples and information from a data stockroom. A Data Warehouse stores this information once it has been properly pre-processed. Data Mining is a sophisticated step forward in this direction, and it involves more than just the general use of algorithms for extracting patterns from data.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

Despite the fact that data mining has gained noteworthy ground amid recent years, most research exertion is committed to developing compelling and proficient algorithms that can remove knowledge from data and insufficient consideration has been paid to the part of domain expert in the discovery process. Data mining is the process of extracting useful information from large datasets in order to use it for specific purposes. Only the user can know for sure whether the acquired information is enough for the intended purpose. Furthermore, what one user may find useful is of little benefit to another. Instead than letting a machine-based data-mining process rely on trial-and-error methods, it is preferable to include human judgment wherever possible. The system's user is able to steer and filter incoming data without having to manually execute actions that should be automated. The user is freed from mundane, error-prone chores so they may instead concentrate on more complex matters like as making decisions, setting priorities, and adjusting to unexpected events. Furthermore, interactive data mining may motivate users to learn, improve their comprehension of the area, and enable them to explore creative possibilities. The system can only become better with feedback from users, thus the partnership is mutually beneficial. Therefore, it is important to develop an interface that allows for maximum user input.

Data Mining

When it comes to processing massive amounts of information, data mining is like a military tank. There are various motives to utilize data mining, for example, it may reduce expenses, augment revenue, upgrading customer and client experience and so forth.

In today's highly competitive business environment, companies need to use scientific and data mining developments to gain a competitive edge. In addition, the customer gets access to a greater amount of information about products on the web, increasing the likelihood that they will choose the superior product or service. To "use a variety of methods to recognize pieces of information or knowledge in data, and to separate these for decision support, expectation, figuring, and estimation" is a common definition of data mining (DM). The data is usually copious, but its structure is of little use since it cannot be put to coordinated use; the useful information is buried inside the data.

Data mining allows market statisticians to extract useful information about individuals, patterns, and demographics from large datasets. DM comprises the application of quantifiable and numerical processes, for example, cluster analysis, automated interaction identification, predictive modeling and neural networking. If a business wants to get the most out of its database, it should hire the services of a data mining expert.

Data mining is the act of analyzing large amounts of data from new angles and drawing conclusions that may be used to increase revenue, decrease costs, or do both. Data mining is a technique for finding patterns or specific instances among several fields in a large database. The data mining tasks conducted determine the kind of examples that may be obtained.

Classification of Data Mining System

Data mining systems can be ordered by different criteria as takes after:

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

a) **According to the type of data sources mined**:
This categorization is based on the types of data that are processed, such as geographical information, multimedia files, schedules, texts, the Internet, and so on.

b) **According to the database involved**:

Relational databases, object-oriented databases, data warehouses, transactional databases, and so on were all included in this classification system based on their respective approaches to data presentation.

c) **According to the kind of knowledge discovered:**

This structure is based on the information discovered or the data mining functions used, such as depiction, separation, association, characterisation, clustering, etc. Some of these systems tend to be all-encompassing packages that provide a variety of data mining features.

d) **According to mining techniques used**:

Data analysis methods (such as machine learning, neural networks, genetic algorithms, statistics, the human senses, a focus on databases or data warehouses, etc.) determine this ranking. Data mining systems may be classified into many categories based on how involved the client is in the process: question-driven, interactive exploratory, and autonomous.

LITERATURE REVIEW

Delen and Demirkan (2013) There are three distinct types of data mining, and they are: expressive analysis, predictive analysis, and prescriptive analysis. The class is defined by its focus on data mining. An problem is discovered via visual analysis. Predictive analysis delves into the problem by looking forward to possible outcomes based on historical information. What should be done right now is the focus of prescriptive analysis. In the past, businesses relied mostly on recorded data for basic leadership. However, in the present day, they need solutions not only for breaking down obtained data, but also for doing projections of the available information and what kind of actions to make.

Rowe (2017) Predictive analysis is used to learn about potential actions in the future. It provides prompt explanations for what took transpired and what could happen next. Predictive research focuses on the future, whereas descriptive research is structured on the past. However, the information used in the forecasting study is historical in nature and is used to the prediction of future events. Predictive demonstrating involves making predictions based on current estimates of a number of parameters. Instances, emotions, and other relevant facts influence forecasts. Insights, machine adaptation, deep learning, data mining, replication, patterns, affiliations, affinities, and other scientific approaches are used to extract patterns from the data. Predictive research is a vital part of strategic planning for long-term success. Predictive research and the insights it provides should be required reading for all senior executives.

Minsker (2015) the concord is that descriptive examination is the primary stage before predictive and prescriptive inquiry, nonetheless whether predictive or prescriptive investigation ought to be lead initially separates emotions. Predictive analysis is the foundation of prescriptive research, which goes above and beyond assisting organizations in making decisions on the

following tasks. Some arguments support the idea that previous successes should be grasped and interpreted to actions with the use of prescriptive tools and expectations that will be used in the future when the outcomes are evaluated. The analysis originates with the prescriptive knowledge that gives contribution to predictive analysis.

Holzinger, (2012) The advancement of human insight via invention necessitates the combination of human PC interaction and KDD. It helps customers find obscure and hidden instances in a large data set that includes both useful and useless information. When taken together, they provide strong validation for the importance of such data samples for knowledge acquisition.

Borhade and Mulay (2015) developed a web-based interactive incremental data mining tool (OIIDM) for discovering crucial data samples in which to ground key insights. The suggested instrument gave many functions of incremental grouping, affiliation mining and order by complete engagement with the customers while stressing on client fulfillment. In order to bunch data, the researchers used the incremental k-implies technique and COBWEB; to sort the data, they used the Bayesian algorithm and C 5.0; and to mine affiliation data, they used the predictive apriority affiliation control mining algorithm. The gadget is useful for communicating with the customer and enables many data mining techniques.

RESEARCH DESIGN

Data mining technologies that allow for interactive examination of massive data volumes are the primary focus of this study. Data mining is characterized as an iterative procedure with several steps, such as gathering relevant data, cleaning it, using data mining algorithms, evaluating the results, etc. The customer asks an inquiry and immediately receives a response once everything is done. Results obtained at the end of the procedure might reveal a poor and incorrect selection of data sources. If the previous results are unsatisfactory, the customer specifies the next inquiry using his expertise in the field and the information from the previous result.

Research Problem

Data mining efforts are complex and often unsuccessful. In order to increase the success rate of data mining initiatives, it is crucial to manage human resources and their involvement in these endeavors, as well as to adhere to a set life cycle. In the case that the data mining process is let to continue in an automated method, the findings may not be useful in a specific application. Additionally, in such a scenario, the customer may choose to repeat the data mining process in its entirety for a different configuration of informational elements. Repeating this procedure yields enticing outcomes. The client is unaware of the decision-making process that led to a certain conclusion.

Source of Data

The data gathering process occurs before a research test. After reading a few comprehensive publications on the subject, I realized that many studies relied on confidential information obtained from hospitals and other medical facilities that was not publicly available. It took a lot of time and effort to get the useful information. This study period revealed that although there is fragment data available, for example on the internet, a large percentage of them are meaningless.

There is a lack of documentation of even the names of characteristics and the division into restriction and decision attributes in certain databases, which makes it impossible to use such databases for research. The UCI Medical Data Repository has made the decision to allow other institutions to conduct comparative studies and analyze the data. This was our primary motivation for settling on the datasets housed in UCI's data center. Each of the selected databases is unique. There are five main medical fields that they represent. This allows the performance of the algorithms to be evaluated over a wide range of real-world medical scenarios (attributes).

The UCI Repository of Machine Learning Databases and Domain Theories is a public online archive of informative data sets from a number of countries. All data sets are described briefly in text documents. Numerous academics have acknowledged these datasets, so you know they're a valuable resource. Five different medical datasets were used for the studies. The selected information pertains to five distinct areas of medicine.

DATA MINING IN WAIKATO ENVIRONMENT FOR KNOWLEDGE ANALYSIS (WEKA)

The WEKA software implements a wide variety of data mining techniques. The J48 decision tree is only one example of a tree-based method, but there are other rule-based methods like ZeroR and decision tables, as well as likelihood and backslide-based methods like the Naive bayes algorithm.

The WEKA (Waikato Environment for Knowledge Analysis) version 3.4.8 is chosen as the testing tool for medical data set analysis and performance evaluation of associated data mining systems. A clear illustration of the parameters used by the chosen data mining methods for exploration is provided. The model execution metrics used to establish a causal relationship between strategy success and accuracy are also presented. We can now see how each algorithm performs on medical datasets. The portion is built on claim participation with WEKA environment strengthened with information added.

Waikato Environment for Knowledge Analysis (WEKA)

Waikato Environment for Knowledge Analysis, or WEKA for short, is a collection of state-of-the-art data mining methods and tools for doing exhaustive research. This ecosystem was developed at New Zealand's University of Waikato. WEKA is distributed under the terms of the GNU General Public License and is written in the Java computer language. These complicated algorithms may be linked to a data collection for the purpose of definitive research and evaluation of data mining analysis. WEKA primarily makes use of three distinct approaches. Learning more about the data requires first looking at the results of data mining tactics; next, the age of the model is considered for the prediction of new instances. Data mining approach correlation for indicator selection, as in medical decision support systems, is the last but most important point of emphasis in the current ace's theory.

WEKA has three graphical user interfaces and one command line interface. Explorer is the default browser-based user interface. It's a graphical user interface consisting of a navigation bar and six panels representing various data mining methods. It facilitates attribute relationship

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

mining, data preprocessing, classification, and clustering. Moreover there is a likelihood to pick characteristics using the attribute evaluator and inquiry approach. The final option is perception charting the conditions among characteristics.

Knowledge Flow is a visual interface for selecting components from the toolbar, placing them on a dedicated canvas, and then associating them into a coordinated diagram for further processing and investigation. This interface allows for the planning and execution of data processing operations outside of the data stream.

Experimenter, a third graphical interface, is useful for considering the execution of data mining algorithms. The effectiveness of various data mining methods on certain datasets may be evaluated with the help of this module. Mechanization of the procedure allows for foregone insights. This unit is an essential part of the test. In the case of medical datasets, it generates useful internal and external insights. Measurements useful in the event of medical diagnostic assistance may be prepared after the selection of various methodologies, their parameters, and datasets.

During the practical aspects of a master's thesis proposal, Experimenter and Explorer are two common interfaces. What is easily performed by converting.txt files along the path shown in Figure 8.1, WEKA enables studying the data sets stored in.arff documents. The data record is organized like a decision table, with the name of the table coming first, followed by the names and types of attributes, and finally the values of the attributes being monitored. This straightforward archival structure makes it possible to move the claimed dataset to a setting organized in this fashion and to break it down.

```
@relation diabetes

@attribute pregnant real
@attribute plasma real
@attribute diastolic real
@attribute triceps real
@attribute insuline real
@attribute mass real
@attribute pedigree real
@attribute age real
@attribute diabetes {1,0}

@data
6,148,72,35,0,33.6,0.627,50,1
1,85,66,29,0,26.6,0.351,31,0
8,183,64,0,0,23.3,0.672,32,1
1,89,66,23,94,28.1,0.167,21,0
0,137,40,35,168,43.1,2.288,33,1
```

Figure 1 Sample .arff file for WEKA

Using the WEKA environment, in-depth and comprehensive research is feasible. It's why we choose to use it to delve into databases of medical records. Because WEKA and its documentation are freely available, it is possible to do the kind of in-depth comparisons that are shown in the ace's postulate and contrasted with her outcomes and presented here. The use of cutting-edge technology ensures that probes are thorough and accurate.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
IJARST
ISSN: 2457-0362

As with the diabetic database, the validation of the tree's execution follows a similar path. Experiments showed that a dataset split at 66% yielded the highest number of True Positives (around 69.7%). The most extraordinarily bad True Positive result is found for 30% split of dataset. For the rest of the setups, the True Positives figures we gathered were quite close to 63%. When it relates to the True Negatives the greatest results were produced for the 10-overlay cross-approval. However, its advantage over the other setups was negligible (about 1-5%). Misclassifications are uncommon, as shown by the low False Negative and False Positive rates.

Models were successfully transmitted in all test settings that organized over 90% of events. Except for the Root Mean Squared Error and the Root Relative Squared Error, most of the errors were rather modest (less than 20%). Most configurations achieved a True Positive rate of better than 91% while maintaining a False Positive rate of less than 9%. In addition, remarkable outcomes were seen in terms of Precision, Recall, F-measure, and Area Under the Curve. Most of the resulting models have impressively high values (more than 90%) for these criteria. With such promising outcomes, pinpointing the optimal settings for the C4.5 algorithm is a challenge. However, the 30% split was the one that picked up the most obviously negative outcomes. The error rate was highest and the classification quality was the worst (poor Precision, F-measure, Recall, AUC, and TP rate) with this setup.

Table 1: Performance of the C4.5 with the respect to a testing configuration for the breast cancer database

testing method	**training set**	**10-fold cross-validation**	**5- fold cross-validation**	**15- fold cross-validation**	**30% split**	**50% split**	**66%split**
correctly classified instances	98.5%	97.1%	96.4%	96.5%	94.4%	94.9%	96.8%
Mean absolute error	4.8%	6.7%	6.9%	6.8%	9.2%	8.4%	5.9%
Root mean squad error	14.8%	19.9%	21.5%	21.7%	26.2%	25.3%	19.3%
Relative absolute error	9.3%	13.3%	13.7%	14.1%	18.7%	17.1%	11.8%
Root relative squad error	29.9%	41.5%	43.9%	44.5%	54.2%	52.3%	42.6%
TP	64.6%	63.7%	63.9%	63.9%	62.6%	63.5%	69.7%
FN	2.5%	3.4%	3.2%	3.2%	5.4%	6.6%	4.0%
FP	1.6%	2.7%	3.4%	3.5%	3.3%	1.6%	2.3%
TN	35.5%	34.4%	33.7%	33.6%	32.8%	32.3%	28.1%
TP rate	38.9%	97.5%	97.7%	97.7%	94.5%	92.9%	96.9%

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
IJARST
ISSN: 2457-0362

FP rate	2.8%	5.7%	4.8%	8.2%	7.8%	2.9%	5.6%
Precision	99.8%	98.6%	97.5%	97.3%	97.5%	99.2%	99.3%
Recall	98.7%	97.5%	97.7%	97.3%	94.5%	99.9%	96.7%
F-measure	99.5%	97.9%	97.6%	97.7%	95.8%	96.2%	98.1%
AUC	99.9%	97.8%	97.9%	97.2%	93.8%	95.8%	98.8%

Figure 2 displays the same data that can be found in Table 1. In contrast to the diabetes database, the breast cancer database had far less categorization mistakes, as seen in the image.

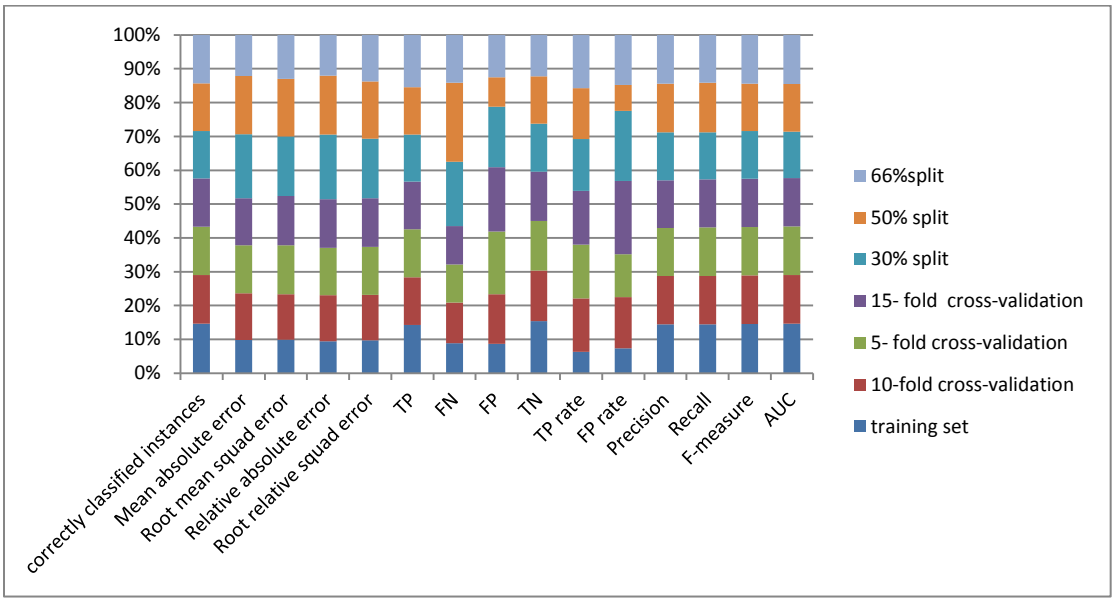

Figure 2 Relation between the performance measures and the testing configurations of the C4.5 for the breast cancer database

CONCLUSION

Data mining is a cutting-edge, emerging technology that has great promise for helping businesses zero in on the most important data stored in their databases. One definition is "the computerized analysis of large or complex data sets in order to discover interesting examples or patterns that would otherwise go unnoticed."

Data mining is a technique that allows for this'smart' use of technology by sifting through large datasets in search of interesting and previously hidden information. This technique builds on extensive work in areas like as statistics, machine learning, design recognition, databases, and improved registration.

How, therefore, can data mining provide insight into phenomena over which one has no control and hence cannot immediately a priori muse? Modeling is the strategy utilized to act out these successes. Demonstrating systems have been around for quite some time, but it wasn't until registering innovations came along that we could store massive amounts of data and use automated displaying procedures to anticipate and understand the 'hidden' examples within data.

These days, medical databases collect an enormous quantity of information. Information on symptoms and diagnoses for serious illnesses may be stored in such databases. The task of uncovering such linkages in historical data is much simplified when medical frameworks are used. This information may be used in the diagnosis of similar situations in the future.

The study's major goal was to identify the most popular data mining methods currently used in MDSS and to evaluate their efficacy on a small selection of medical datasets. There were three algorithms chosen: C4.5, MLP, and NB are the models used. Heart disease, skin disease, hepatitis, breast cancer, and diabetes data sets were used in the evaluation of five UCI databases. Several efficiency measures were used, including correct classification percentage, True Positive and False Positive rates, Area Under the Curve (AUC), Precision, Recall, F-measure, and error configuration. The true motivation for this study lay in the fact that no prior work was found that compared and contrasted these three algorithms under identical circumstances.

In light of the intriguing nature of medical data, medical data mining is unique. There are strict limits on the range of possible values for the characteristics in medical data. These characteristics always exist in just two states. They refer to the presence or absence of a few key features, such as symptoms or an analysis. In contrast, specified interims (such as circulatory strain or body temperature) are often used to assign values to multi-valued properties. These tend to be optimistic. Attributes speaking to screening outcomes, such as ECG, RTG, etc., tend to have negative values.

REFERNCES

1. Abe H., Yokoi H., Ohsaki M. and Yamaguchi, T. (2007). Developing an Integrated Time-Series Data Mining Environment for Medical Data Mining. Seventh IEEE International Conference on Data Mining, 28-31 Oct. 2007, 127-132.
2. Turban, Efraim; Sharda, Ramesh; Delen, Dursun 2011. Decision Support and Business Intelligence Systems. Pearson Education Inc. New Jersey, USA.
3. EmilyRowe, Enhancing horizon scanning by utilizing pre-developed scenarios: Analysis of current practice and specification of a process improvement to aid the identification of important ‘weak signals’, Technological Forecasting and Social Change, Volume 125, December 2017, Pages 224-235
4. Holzinger, A., Dehmer, M., & Jurisica, I. (2014). Knowledge Discovery and interactive Data Mining in Bioinformatics - State-of-the-Art, future challenges and research directions. BMC Bioinformatics, 15(Suppl 6), I1. http://dx.doi.org/10.1186/1471-2105-15-s6-i1
5. Borhade, M. & Mulay, P. (2015). Online Interactive Data Mining Tool. Procedia Computer Science, 50, 335-340. http://dx.doi.org/10.1016/j.procs.2015.04.039
6. Chau, D. (2012). Data Mining Meets HCI: Making Sense of Large Graphs. Carnegie Mellon University, Pittsburgh
7. Dehuri, S. & Ghosh, A. (2013). Revisiting evolutionary algorithms in feature selection and nonfuzzy/fuzzy rule based classification. Wiley Interdisciplinary Reviews: Data Mining And Knowledge Discovery, 3(2), 83-108.

8. Sasan, H. & Sharma, M. (2016). Intrusion Detection Using Feature Selection and Machine Learning Algorithm with Misuse Detection. International Journal Of Computer Science And Information Technology, 8(1), 17-25. http://dx.doi.org/10.5121/ijcsit.2016.8102
9. van Gerven M. A.J., Jurgelenaite R., Taal B. G., Heskes T., Lucas P. J.F., Predicting carcinoid heart disease with the noisy-threshold classifier. Artificial Intelligence in Medicine, 2007, vol. 40, 45-55.
10. MGH Laboratory of Computer Science – projects – dxplain, Laboratory of Computer Science, Massachusetts General Hospital. 2007. http://lcs.mgh.harvard.edu/projects/dxplain.html retrieved in 1.05.2007