JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

REVIEW PAPER ON DATA MINING TECHNIQUES, ALGORITHMS AND TOOLS

Ajay Kumar[1], Preeti Sondhi[2]
[1]M.tech Student,[2]Assistant Professor
Universal Group of Institutions Lallru Punjab,
Punjab Technical Universiy

Abstract: ***Data mining has made considerable growth in recent time but the main drawback is missing of data has remained a confront for data mining algorithms. The main task of data mining is to traverse the large amount of different types and categorize and Summarize it. In current world data mining is moreover accepted and make huget growth within the field of all kinds of application in software feild. Data mining is a process of extracting business useful data from large databases by using data mining tools and techniques. Data mining techniques have used to improve performance in Healthcare, Educational, Business areas by extracting unknown and applying data mining tools and algorithms techniques. Data mining is process used to extract data and discover knowledge from it and presenting it to humans by more understandable format and data mining is used to analyzing the data that stores in large data warehouses and discover last unknown unknown data structures and relations out of large data. This is a review paper that introduces the study of classification technique on the basis of the algorithms which is used to make prediction on large data set to analyze the future prediction in healthcare sector. This paper focus on various ensemble learning techniques like Boosting and Bagging and also comparison of ensemble learning techniques with each other by using base classifiers.***

Keywords: Data mining, Classification, Ensemble algorithms, Data mining tools

1.INTRODUCTION

Today's world there is advancement in database technologies and data collection techniques have collected large amount of data from large database. The volatile growth of data create the essential knowledge discovery from data which leads to optimistic transpire field, called data mining or knowledge discovery in database(KDD) [1]. Data mining is a process of extraction or mining the knowledge from large databases. Data mining is a negligible data extraction of unknown pattern and useful information from data. Data mining is used to discover knowledge and analyze data and introduce it in a easy form that human can understand easily [2]. The main functionalities of data mining to find pattern in mining tasks. Data mining tasks are often divided into two categories-descriptive and predictive. Descriptive mining task is identify general properties of data that stored in database, Predictive mining task achieve on the current data to make accurate prediction for future use [3]. Data mining is a field of computer science which involve the data extraction and from big databases and pattern discovery. The main goal of this process is to transform these patterns into the more understandable form. The data mining method used in the artificial intelligence, machine learning, business intelligence. With the use of data mining we can solve problems by analyzing data already stored in database.

1.1 DATA MINING PROCESS IN KDD

Knowledge Discovery is a important process in database contain some steps leading from data collected to some newly structure of new knowledgable data. It consist of the following steps as shown in Figure 1

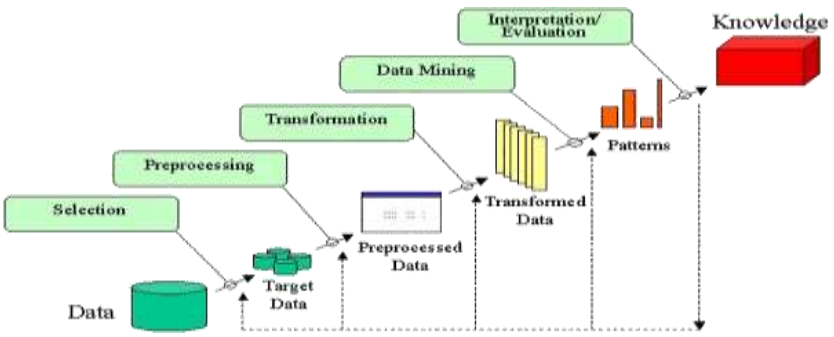

Figure 1. Knowledge Discovery in Database

1.1.1 DATA MINING STEPS IN THE KDD PROCESS

- **Data Cleaning:-** Also known as data cleansings , In this step the noise and inconsistent data is removed.
- **Data Integration:-** In this step, multiple data sources is combined**.**
- **Data Selection:-** During this step, appropriate data for the analysis purpose that are retrieve or extract from database.
- **Data Transformation:-** In this step, extracted data is transformed into appropriate form for mining to performing aggregation operations.
- **Data Mining:-** This is a important step where crucial intelligence steps are applied in order to extract pattern.
- **Pattern Evaluation:-** In this step, in previous step data pattern is extracted and in this step pattern is evaluated.
- **Knowledge Presentation:-** In the last step, visualization and representation knowledge techniques are applied.

1. DATA MINING TECHNIQUES:-

There are several major data mining processing techniques have been developed and using in data mining or data processing projects including classification, clustering, prediction Descriptive and Predictive mining techniques. Descriptive approach includes some models for whole probability distribution of the data, partitioning of whole data into different parts and models describing the relationships between the variables. Predictive technique permits the value of attribute and variable is just too predicted from the known values of other attribute/variable. In this paper we studied the one descriptive technique i.e. clustering and one predictive technique of data mining i.e classfication.

2.1 PREDICTIVE TECHNIQUE

The main objective of a predictive approach is to allow the data mining tool to predict an not known value of a related variable; the target variable. If the target value is predefine number of discrete type labels, the data mining tasks is called classification.

2.1.1 CLASSIFICATION AND ALGORITHMS

Classification is the commonly applied data mining technique. Classification algorithm is applied to set of pre-classified data to develop a model that can classify the population records at large. This technique usually employs decision tree or Random tree based classification algorithms. The classification technique involves learning and classification. In the classification data is used to estimate the accuracy of population record [4]. Classification is a main techniques which developed as an crucial component of machine learning algorithms in order to extract data and data patterns from data that could be used for future data prediction. In classification a classifier is created from a data set and used to classify the unknown data. Classification has two steps: learning phase and classification phase. With the use of mapping functions we can classify any attribute in the classification phase. In the classification attribute is evaluated and accuracy is evaluated as the percentage of correct classification obtained. All algorithms are used in classification phase that have been applied to predict the performance of data set. They are Decision Tree, Bayesian Classifier, and Artificial Neural Network [5].Classification has the following classifier: Naïve Bayes, Decision Tree, and Decision Stump.

2.1.2 DECISION TREE

Decision tree is common classification model. In the decision tree each branch node represents a decision. Decision tree is mostly used for obtain important information for the objective of decision making. Decision tree have a root node that is used as starting node on which it is for users to take actions. From this start node user will divide each start point iterative according to the decision tree learning algorithms in data mining. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcomes.[6].

2.1.3 BAYESIAN CLASSIFIER

Bayesian classifier is statistical classifier that is represented as a graph structure. Naïve Bayes most important classifier that is used to predicts the data probabilities and statastics, such as the probability that give sample belongs to a particular class. Mostly Bayes algorithm is developed ,out of which Bayesian and naïve Bayes are the two most important techniques in data mining. Naïve Bayes algorithm conclude that the effect that an attribute plays on a given class is independent of the values of the other attributes in data mining [7].

2.1.4 DECISION STUMP:-

A decision Stump is a machine learning algorithm model consisting of one-level; decision tree. It is a base of decision tree in which one of the main internal node point which is mainly connected to the terminal nodes. Decision stump algorithm makes a exact prediction on base of value that a single input type. Sometimes they are also called as one-rule decision stump are

mainly used as components called weak learner and base learner in machine learning ensemble techniques like bagging and boosting/AddaBoost [8].

2.1.5 RANDOM FOREST:-

It is a type of supervised learning algorithm which consist of number of simple trees, which are used to determine the final outcome [9]

2.2 DESCRIPTIVE TECHNIQUE

A descriptive model in concise form, the main characteristics of the data set. It is crucial shortened of the data points, making it possible to study main aspects of the data set. Commonly a descriptive models is found through undirected data mining; that is a bottom-up approach [10].

2.2.1 CLUSTRING AND ALGORITHM

Clustering is mainly findings groups of objects that the objects in one group are going to be associated to one another and different from the objects in another group. Clustering are often considered the foremost important unsupervised learning technique. Clustering are often examined the foremost important unsupervised learning technique so as every other problem of this type. It deals with Descriptive approach includes models for overall probability distribution of the data, partitioning of whole data into groups and models describing the relationships between the variables [11]. Predictive technique permits the value of one attribute or variable is too predicted from the known data values of other attribute or variable. This paper repersent the studies for the one descriptive technique that is clustering and one predictive technique that is classification and tools used for better data analysis using data mining techniques.

2.2.2 K-MEAN ALGORITHM:-

K-mean is an reiteration clustering algorithm during which items are passed among sets of clusters unit the prefrenced set is reached. As such as it could be viewed as a kind of squared error algorithm, while the meeting criteria need not be defined based on the squared error. A high degree of similarity among elements in clusters is obtained, while a high degree of similarity among elements in clusters is obtained while a high degree of dissimilarity amidst elements in different types of clusters is achieved simultaneously[12].

Sets of algorithm

a) Firstly it would selects the initial k prototypes arbitrarily.

b) The squared error criterion is mainly used to determined the clustering algorithms quality.

c) In each iteration the prototype of every cluster is recomputed to be the cluster mean.

d) The basic version of k- means does not include any sampling techniques to scale to large databases.

2.2.3 HIERARCHICAL ALGORITHM:-

Hierarchical clustering algorithm differs in how the sets are created in it. A tree data structure called a dendrogram are often used to illustrate the hierarchical clustering technique and the sets of different clusters. The root is a dendrogram tree contains only one cluster where all the elements are together. The leaves within the dendrogram each contain one element cluster. Internal nodes within the dendrogram representing replaced or new clusters formed by merging various sort of clusters that emerge as its children within the tree. Each level in the tree is related with the distance measure that was used to merge the clusters. Every clusters created a selected level and particular level were combined because the child clusters had a distance between them but less than the distance value associated with this level in the tree [13].

3 ENSEMBLE LEARNING ALGORITHMS

Ensemble learning is additionally called learning multiple classifiers system. Ensemble method use multiple learning algorithms together for the same task with the aim to have better accuracy as compare to the individual learning model An ensemble contain number of learners called base learners. Base learners are generated by base learning algorithm such as Naive Bayes, Decision Stump or any other kind of learning algorithm from training dataset. This paper works on the following Ensemble learning algorithms:

3.1 BAGGING ENSEMBLE LEARNING ALGORITHM

This algorithm improves the accuracy and stability of learning algorithms used in statistical classification and regression techniques algorithm. This Algorithm is also known as bootstrap aggregation.

3.2 BOOSTING ENSEMBLE LEARNING ALGORITHM

Boosting algorithm creates strong classification tree because it forces new classifier to focus on the error produced by previous ones.

4 DATA MINING TOOLS

In current time there is various data mining tools are available to handle or manage the large number of datasets and also improve the quality of the data, such tools are Rapid Miner, Weka, Scikit-Learn, KNIME, orange, KEEL, Tanagra etc. These data miner tools make easy for data analyst to get the relevant information. Data mining tools are used to predict future trend, allowing to business more predictive, knowledge driven [14]. There are various Data mining techniques and algorithms have been implemented on these tools to extract the information and also to check their efficiency and accuracy. In this paper, we are going to discuss and compare only three tools i.e. Rapid Miner, Weka, and KNIME which are using the same platform.

4.1 WEKA

Weka is a well-known learning software written in Java developed at Waikato University in New Zealand. The Weka application let user a tool to identify hidden information from knowledge/data base and flat files system with simple to use option and visual interfaces. The Weka tool contains a set of visualization tools and algorithms for solving real-world data mining or anaysis data processing problems and predictive [15]. WEKA is an open source software issued undergoing the General Public License. WEKA come up with four application interface that is Explorer, Experimenter, Knowledge flow, and Simple Command line. It has four Graphical interfaces i.e. **Explorer** is an environment for exploring the data in WEKA with the help of the learning algorithm. **Experimenter** is an one of environment for perform and do statistical tests between learning algorithms schemes. **Knowledge Flow** is Java Beans for setting up and running machine learning experiments in data mining tools or other learning specific tools and algorithms. **Simple Command Line Interface** allows quickest execution of Weka commands and also provides simple command-line interfaces.

4.1.2 FEATURES

- WEKA is a Java based open source.
- It is easy to use for beginners level and has the ability of running several learning algorithms and comparing.
- It is a platform independent.
- It performs various data mining tasks including: Data pre-processing, Classification rules, regression, Clustering, association rules, visualization, feature selection and enhance the knowledge discovery.
- WEKA have a 49 Data pre-processing or mining tools, seventry six Classification and regression algorithms, Eight Clustering Algorithms, Three algorithm to finding association rules, Fifteen attribute or subset evaluator plus Ten searching Algorithm for feature selection [16].

4.1.3 ADVANTAGES

- Easy to manipulate the data.
- Provide access to SQL databases.
- It provides two options for the user to interact through Explorer and Command line [17].
- Specially used for data mining.
- It provides severaal machine learning algorithms for tasks in data mining.
- It supports various standards for Data mining tasks that include the Data pre-processing or mining, Clustering and Classification, Regression, Feature selection and Visualization [18].

4.2 RAPID MINER

Rapid Miner is a tool for conducting data mining workflows for various tasks, ranging from different areas of data mining applications to different parameter optimization schemes [19]. One of the main traits of Rapid Miner is its advanced ability to program execution of complex workflows, all done within a visual user interface, without the need for traditional programming skills. Rapid Miner is the most powerful, easy to use and intuitive Graphical User Interface for the design of analytic process, that contain several “operators”. The operator functions as a single task in their process in which the input is produced by the existing output of the operator [20].

4.2.1 FEATURES

- It is platform independent.
- It is well suited with various databases like oracle, MySQL, Excel, SPSS, Microsoft SQL server etc.
- It provides Drag and Drop interface to design the data analytics or mining process.
- It supports and accepts new data drivers to perform task in Rapid Miner.
- It provides over and above Five hundered operators for all machine learning procedures or algorithms, and also combines learning schemes and attributes evaluators of the WEKA learning environment. [21].
- It allow user to work with different sizes and types of data sources.

4.2.2 ADVANTAGES

- It has enormous flexibility.
- It provides the combination of maximum algorithm of such data ming tools.
- Easy to debug the errors in rapid miner tool.

4.2 KNIME

KNIME (Konstanz Information Miner) is a widespread intended data mining tool based on the Java Developed platform and maintain by the Switzerland Company called KNIME. KNIME is a open-source software, though commercial licenses support. The tool adhere to the visual programming paradigm present in most DM tools, where building blocks are placed on a canvas and connected to obtain a visual program. In KNIME, these building blocks are called nodes [22]. In KNIME, representation of data sources, data mining algorithm, data transformations, visualizations techniques etc. defined by set of nodes that called workflow and each node has its particular inputs and outputs ports that depends on the functionality of the node [23]. For both simple and complex data types, KNIME allows analysis to search trends and predict future results. KNIME uses for teaching and research purpose which allows integrating the new algorithms and tools during a simpler manner.

4.3.1 FEATURES

- Available to everyone that allow users to use the comprehensive node API to add proprietary extensions. Intuitive user interface.
- It is easy to use and handle different functions.
- KNIME modules cover a wide variety of functionalities like, I/O, data manipulation, views, hilting etc. to better understand your data.
- It provides the users to create data flows or data pipeline visually, users can selectively execute some or all Data Analysis steps, study the results, prototypes modal, and collaborative interpretations [17].
- For cross validation and independent validation in KNINE tool, it come up with functionality to save parameters.

4.3.2 ADVANTAGES

- The major benefit of this is easy to use plug-in.
- It based on the node work which includes more than 100 nodes to examine the data [24].
- It come up with the data flow process by dragging and dropping new nodes.

4.4 R-PROGRAMMING TOOL:

R-Programming tool is written in C and FORTRAN language and allows the data miners or data anlyser to write the scripts just like a programming language platform. Hence, it is used to create statistical and analytical software for data mining and processing . It supports graphical analysis for both linear and nonlinear modelling, classification, clustering and time based data analysis.

4.4.1 FEATURES OF R TOOL

- R is quite just a domain-specific programming language aimed at data analysis.
- It performs multiple calculations with vectors.
- It is processing more than just statistics.
- It can running code without a compiler

4.4.2 ADVANTAGES

- R may be a programming language and environment developed for statistical analysis by practicing statisticians and researchers.
- R is freeware and opened sourced software, allowing anyone to use use modify it.
- R has no license restrictions [23].
- R has over 4800 packages available from multiple data repositories specializing in topics like econometrics data mining, data pre-processing, spatial analysis, and bio-informatics etc.
- R runs on different types of operating systems platform and different kind of hardware. It is popularly used to GNU/Linux, Macintosh, and Microsoft Windows, running on both 32 and 64 bit processors. R is a cross-platform [23].

5 DATA MINING TREND

List of trends in data mining that reflects pursuit of the major challenges such as construction of integrated and interactive data mining environments, design of data mining languages [25]:

- Application Exploration
- Scalable and Interactive data mining methods
- Consolidation of data mining or processing with database systems, data warehouse systems and web database systems.
- Standardization of data mining query language
- Visual Data Mining
- Biological data mining
- Data mining and software engineering
- Web mining
- Distributed Data mining
- Real time data mining
- Multi Database data mining
- Information Security in data mining

6 RELATED WORK

Eibe Frank et al. (2004) did a study on WEKA tool and provides a brief introduction of the Weka tool [26]. **S.B. Kotsiantis al. (2006)** proposed a technique of boosting localized weak learners. In this paper boosting method used for classification and regression problems that works locally also performed a comparison with other well-known combining methods on standard classification and regression benchmark datasets using decision stump as base learner and therefore the proposed technique give the efficient result [27]. **Anshul Goyal, Rajni Mehta (2012)** relate performance evaluation of Naïve Bayes and J48 classification algorithms. The experimental results shows that the efficiency and accuracy of J48 is good as compare to NaiveBayes algorithm [28]. **Luis C. Borges et al. (2013)** compare four Data Mining and processing tools: KNIME, Orange, RapidMiner and Weka. The objective of the research is to find the most efficient tool and technique for the classification task. The experimental results show that there is no single tool or technique that always achieves the best result but some tool achieve better results more often than others [29]. **Narender Kumar Sabita Khatri (2017)** did an analysis on various data classification techniques i.e J48, NaiveBayes, RandomForest, SVM and k-NN on WEKA tool for early chronic kidney disease prediction. These are compared on the basis of evaluation parameters like Accuracy, ROC, kappa statistics, RMSE, MAE, TP rate, FP rate, precision, recall and f-measure. Experimental results shows that random forest classifier have better classification accuracy as compared to other classifier for chronic kidney disease data set [30].**E. Suriyapriya, M. Praveena (2017)** discussed a novel approach for developing a cluster and booster on the basis of data mining. Clustering with boosting improves the standard and quality of mining process. In this research paper, various boosting problems and their proposed solutions are discussed. In order to performance enhancement, integrate the boosting methodology with fuzzy cmeans (FCM) [31]. **Kuldeep Randhawa et al. (2018)** did a study on credit card fraud detection using AdaBoost and majority voting algorithm on Weka tool. In this research paper, machine learning algorithms are used to detect credit card fraud. To evaluate the model efficiency, basic algorithms are firstly used then AdaBoost and majority voting algorithms are applied. The experimental results shows that the bulkvoting method have good accuracy in detecting fraud cases in credit card and debit card.

7 CONCLUSION AND FUTURE SCOPE

This review paper firstly hold brief introduction about the concept of Data mining then by moving with regard to data processing in Data mining tools. This paper conferred a detailed description of data mining techniques, algorithms and Tools. It list the trends in Data mining. Data mining and techniques such as classification is used to test and train different learning schemes on the pre-processed data file and clustering used to apply different tools that identify clusters within the data file. This paper also focus on clustering algorithm such as K-means. The various algorithms and tools used for the mining of the knowledge are laid out in detail. The future scope provides enhancement and efficiency of knowledge within the system. They could foremost to better, faster and qualitative exaction of knowledge or data with better tools and techniques.

REFERENCES

[1] U.M. Fayyad and P. Smyth. Image Database Exploration: Progress and Challenges. In Proc. 1993 Knowledge Discovery in Database.
[2] Prajapati. D, Prajapat. J, "Handling missing values: Application to University Data Set", August, 2011.
[3] Zhao, Kaidi and Liu, Bing, Tirpark, Thomas M. and Weimin, Xiao,"A Visual Data Mining Framework for Convenient Identification of Useful Knowledge", ICDM '05 Proceedings of the Fifth IEEE International Conference on Data Mining, vol.-1, no. 1,pp.- 530- 537,Dec 2005.
[4] Mrs. Bharati M. Ramageri Data mining technqiues and application Mrs. Bharati M. Ramageri
[5] A.Dinesh Kumar1 , R.Pandi Selvam2, K.Sathesh Review on Prediction Algorithms in Educational Data Mining
[6] A.Dinesh Kumar1 , R.Pandi Selvam2, K.Sathesh Kumar Review on Prediction Algorithms in Educational Data Mining
[7] Dorina Kabakchieva, "Predicting Student Performance by Using Data Mining Methods for Classification", Cybernatics and nformation Technologies, Vol 13, 2013.
[8] Margaret H. Dunham, "Data Mining Introductory and Advanced Topics",Dorling KindersleyvPvt.Ltd.India,Sixth Edition,2013
[9] Available: http://www.statsoft.com/textbook/random-forest
[10] R. Andrews, J. Diederich, A. B. Tickle," A survey andcritique of techniques for extracting rules from trained artificial neural networks", Knowledge-Based Systems, vol.- 8,no.-6, pp.-378-389,1995
[11] V.P. Muthukumar, Dr.S.Subbaiah , S. Srinivasan A review paper on data mining techniques algorithms and tools
[12**]** Margaret H. Dunham, "Data Mining Introductory and Advanced Topics",Dorling Kindersley Pvt.Ltd.India,Sixth Edition,2013
[13] Phyu, Thair Nu. "Survey of classification techniques in data mining."International MultiConference of Engineers and Computer Scientists, 2009.
[14] K. Rangra, K.L. Bansal , Comparative Study of Data Mining Tools, International Journal of Advanced Research in Computer Science and Software Engineering, 4(6), June 2014.
[15] http://www.iasri.res.in/ebook/win_school_aa/notes/Data_ Preprocessing.pdf
[16] S.K. David, Amr T.M. Saeb, K.A. Rubeaan, Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics, Computer Engineering and Intelligent System, 4(13), 2013.
[17] K. Saravanapriya, A Study on Free Open Source Data Mining Tools, International Journal of Engineering and Computer Science, 3(12), December 2014.
[18] S. Singhal, M. Jena, A study on WEKA tool for Data Preprocessing, Classification and Clustering, International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2(6), May 2013.
[19] Ingo Mierswa, MichaelWurst, Ralf Klinkenberg, Martin Scholz, and Timm Euler. Yale: Rapid prototyping for complex data mining tasks. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 935{940. ACM, 2006.
[20] S. Sarumathi, N. Shanthi, S. Vidhya, M. Sharmila, A Review: Comparative Study of Diverse Collection
[21] M. Vijayakamal, M. Narendhar, A Novel Approach for WEKA & Study On Data Mining Tools, International Journal of Engineering and Innovative Technology (IJEIT), 2(2), August 2012.
[22] A. Jović*, K. Brkić* and N. Bogunović An overview of free software tools for general data mining
[23] S. Gunnemann, H. Kremer, R. Musiol, R.Haag, T. Seidl, A Subspace Clustering Extension For the
[24] KNIME Data Mining Framework, 2012 IEEE 12th International Conference on Data Mining Workshops.
[25]Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques",Second Edition 2006.
[26] Eibe Frank et al., "Data Mining in Bio-informatics Using WEKA, Bioinformatics," Vol 20(15), pp. 2479-2481, 2004
[27] S.B. Kotsiantis et al., "Local Boosting of Decision Stump for Regression and Classification Problem," Journal of Computer, 2006.
[28] Anshul Goyal and Rajni Mehta, "Performance Comparison of Naïve Bayes and J48 Classification Algorithms," International Journal of Applied Engineering Research, vol. 7(11), 2012.
[29] Luis C. Borges et al., "Comparison of Data Mining Techniques and Tools for Data Classification," International Conference on Computer Science and Software Engineering, pp-113-116, 2013
[30]Narander Kumar and Sabita Khatri, "Implementing WEKA for medical data classification and early disease prediction," 3rd IEEE International Conference on Computational Intelligence and Communication Technology" 2017.
[31] E. Suriyapriya and M. Praveena, "Clustering and Boosting in Data Mining," International Journal of Engineering Science and Computing, vol. 7(8), 2017.