EFFICIENCY IMPROVEMENT AND OPTIMIZATION USING ADAPTIVE APPROACH FOR PSYCHOPATHOLOGY

Dissertation submitted in partial fulfilment of the requirements for the Degree of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

RUCHIKA

11615837

Supervisor

MANEET KAUR

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

November 2017

TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering

Program : P172::M.Tech. (Computer Science and Engineering) [Full Time]

COURSE CODE : CSE548 **REGULAR/BACKLOG :** Regular **GROUP NUMBER :** CSERGDC

Supervisor Name : Maneet Kaur **UID :** 15709 **Designation :** Assistant Profe
Qualification : _____ **Research Experience :**
_____

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Ruchika	11615837	2016	K1637	9041885953

SPECIALIZATION AREA : Database Systems **Supervisor Signature:** _____

PROPOSED TOPIC : Efficiency Improvement and optimization using Model approach for Psychopathology

Qualitative Assessment of Proposed Topic by PAC		
Sr.No.	**Parameter**	**Rating (out of 10)**
1	Project Novelty: Potential of the project to create new knowledge	7.33
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	7.33
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	7.00
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	7.33
5	Social Applicability: Project work intends to solve a practical problem.	7.33
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	6.67

PAC Co ımittee Members		
PAC Member 1 Name: Kewal Krishan	UID: 11179	Recommended (Y/N): Yes
PAC Member 2 Name: Raj Karan Singh	UID: 14307	Recommended (Y/N): NA
PAC Member 3 Name: Sawal Tandon	UID: 14770	Recommended (Y/N): NA
PAC Member 4 Name: Dr. Pooja Gupta	UID: 19580	Recommended (Y/N): Yes
PAC Member 5 Name: Kamlesh Lakhwani	UID: 20980	Recommended (Y/N): NA
PAC Member 6 Name: Dr.Priyanka Chawla	UID: 22046	Recommended (Y/N): NA
DAA Nominee Name: Kuldeep Kumar Kushwaha	UID: 17118	Recommended (Y/N): Yes

Final Topic Approved by PAC: **Efficiency Improvement and optimization using Model approach for Psychopathology** **Overall Remarks:** Approved **PAC CHAIRPERSON Name:** 11024:: Amandeep Nagpal

Approval Date: 04 Nov 2017 11/28/2017 12:05:04 PM

ABSTRACT

The data mining is the method in which it can abstract the useful data from the raw statistics. The k-mean is the clustering algorithm which can cluster similar and dissimilar type of data. The output of clustering will be given as input to classification which can classify data into two classes. In the research, the SVM (Support Vector Machine) classifier is used to categorize the records. The k-mean algorithm can be improved using back propagation algorithm which can increase accuracy of clustering and reduce execution time. Also comparing the result of back propagation algorithm, forward propagation algorithm, without propagation and then compare all these result with best classification algorithm.

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation/dissertation proposal entitled "EFFICIENCY IMPROVEMENT AND OTIMIZATION USING ADAPTIVE APPROACH FOR PSYCHOPATHOLOGY" in partial fulfilment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Mrs. Maneet Kaur. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Signature of Candidate

Ruchika

Reg .No: 11615837

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.Tech dissertation proposal entitled "**EFFICIENCY IMPROVEMENT AND OPTIMIZATION USING ADAPTIVE APPROACH FOR PSYCHOPATHOLOGY"**, submitted by **Ruchika** at **Lovely Professional University, Phagwara, India** is a bonafide record of her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor
Ms. Maneet Kaur
Date: Nov 30, 2017

Counter Signed by:

1) **Concerned HOD:**
   HoD's Signature: _____
   HoD Name: _____
   Date: _____

2) **Neutral Examiners:**

   External Examiner

   Signature: _____

   Name: _____

   Affiliation: _____

   Date: _____

   Internal Examiner

   Signature: _____

   Name: _____

   Date: _____

ACKNOWLEDGEMENT

I am using this opportunity to express my gratitude to everyone who supported me throughout the course of dissertation. I am thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during this thesis work. I am sincerely grateful to them for their truthful and illuminating views on many issues related to this research.

I express my sincere thanks to my guide **Ms. Maneet Kaur** for his invaluable assistance, motivation, guidance and encouragement without which this research work will be dream. In spite of his busy schedule, he was always there to iron out difficulties which kept o aspiring at regular intervals.

I am really thankful to our **Lovely Professional University** for providing me with an opportunity to undertake this research topic in this university and providing us with all the facilities.

I am highly thankful to my friends and family for their active moral support, valuable time and advice. I am thankful to all of those, particularly the various friends, who have been instrument in cresting proper healthy and constructive environment and including new and fresh innovative ideas during project, without their help, it would have been difficult to complete dissertation within time.

TABLE OF CONTENTS

LIST OF ACRONYMS / ABBREVIATIONS

AI	Artificial Intelligence
NLP	Natural Language Processing
SVM	Support Vector Machine
MLP	Multilayer Perceptron
DM	Data Mining
ML	Machine Learning
ACO	Ant Colony Optimization
MDMP	Medical Decision Model with Punning
GA	Genetic Algorithm

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 DATA MINING

This age is referred as information age, because information leads to power and success. Now a day's people are able to collect remarkable amounts of information with the help of computers, satellites, etc., types of sophisticated technologies [1]. Initially, all the information used to store in computers that sort that combination of information that becomes overwhelming. To handle this great amount of data a new concept of data mining has been added. Due to extensive availability of large amount of data, there is requirement for revolving that data into useful and meaningful information, knowledge is the main reason of attraction of data mining in information industry [2]. The business management organizations, production control organization and market analysis department, to engineering design department, science consideration types of applications can use that information and knowledge. A natural evolution of information technology results viewed as a data mining. In data mining knowledge is given as a output by taking data in input.

In 1996 Fayyad, Piatetsky-Shapiro, Smyth has highlighting some of the distinct characteristics and given a definition of data mining process [3]. They defined data mining as "it is nontrivial procedure of recognizing effective, unique, hypothetically useful, and eventually logical patterns in the data". In data mining, process used it must be non-trivial, modest computations and statistical procedures are not considered here. So, data mining is not just forecasting which vendor will make in the most of future sales by computing which product make the most sales in the preceding year and how much profit earned by an organization.

The understanding of types of problems, tasks addressed by data mining helps in understanding it. The data mining responsibilities can be characterized into either having to do with forecast or explanation is given at high level. The prediction of value can be based on other existing information with the help of predictive tasks.

Predicting when a customer will leave a company predicting, whether a transaction is fraudulent or not [4] and identifying the best customers to receive direct marketing offers [5] are the examples of predictive task. The data is summarized in some manner is the task of descriptive.

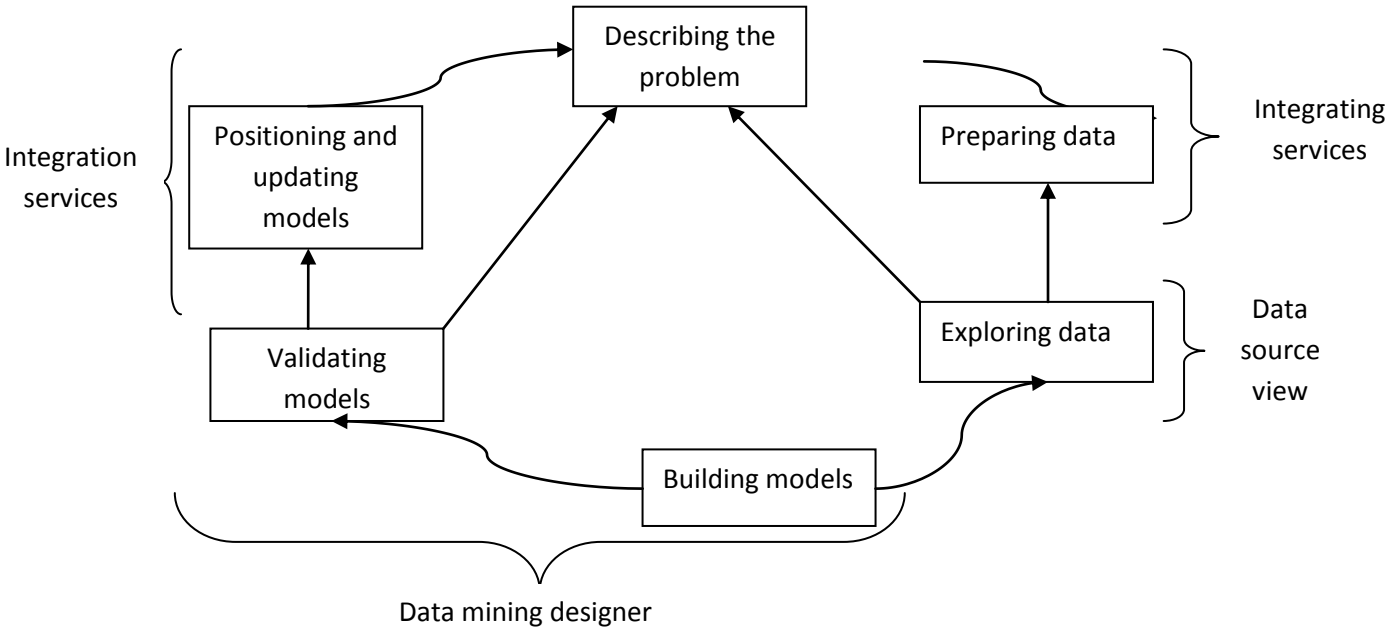

Fig. 1.1: Data mining concept

Automatically segmenting consumers created on their likenesses and dissimilarities [6] and finding relations between products in market basket data are the examples of descriptive task.

The data mining model used to define and collects the patterns and trends. Mining models can be applied to specific scenarios, such as:

- **Forecasting**: Estimating sales of product, predicting server system loads or server system downtime , their number of users
- **Risk and probability**: Choosing the best consumers for targeted mailings, defining the possible solutions for risk scenarios and vulnerability, assigning probabilities of solutions to analyzes or other conclusions
- **Recommendations**: Determining which products are likely to be sold together, which product can't sell together , then generating recommendations for it
- **Finding sequences**: Analyzing customer shopping selections in a shopping cart, predicting next likely or unlikely products

- **Grouping**: Separating consumers or products into cluster of related items, analyzing and predicting similarities that exist between that products

The mining model construction is a part of superior procedure that includes number of things. This process is defined using mostly 6 steps given below:

1. Formulate the Problem
2. Preparing the Data
3. Exploring Data
4. Building Models
5. Exploring and Validating Models
6. Deploying and Updating Models

The Fig. 1 shows relationship between each steps required for mining model. The creation of data mining model is an iterative process and dynamic process of mining the data. After exploring the data this may be happen that user will feel that collected data is insufficient. So, several models have been built in order to find the right model the first main step is the formation of problem. Then that according to that problem data is prepared and explored. This data is used to construct a model to validate, explore the model and deploy it.

1.2 CLUSTERING

The data mining is the process of investigating data from different perceptions and summarizing it into useful and meaningful information. Data mining may consists of anomaly detection, association rule learning, regression, classification summarization and clustering [7]. A cluster consists of large number of data objects, similar are clustered in the same cluster and dissimilar are in other clusters. The clusters can be identified irrespective of their size using best clustering algorithm. Scalability, ability are mostly used to handle noisy data, insensitivity to the order of input records entry, etc., are some requirements that are required while using the clustering algorithms. Data mining is a multi-step procedure that needs accessing of data, preparing data for a data mining algorithm, mining the data or hidden pattern, analyzing results and taking appropriate action. The accessed data can be stored in one or more operational

databases for example in data warehouse or a flat file. The data can be mined using two supervised learning or unsupervised clustering approaches [8].

- **Supervised:** In case of supervised learning both input and desired results are included that makes it fast and more accurate. While process of learning the known correct results are given in inputs to model. Neural network, Multilayer perceptron, Decision trees are the examples of supervised models.
- **Unsupervised:** In this case correct results are not provided to the model during training set. It can be used to clusters the input data into various classes on the basis of their statistical assets only. Example of unsupervised clustering are: Different types of clustering, distances and normalization, k-means, self-organizing maps [9].

Different steps involved in data clustering

- **Well-separated clusters**: A cluster is a set of data items, it cluster the nearest (or more similar) to every other data items in the same cluster as compared to any other data items that is not belong to that cluster.
- **Center-based clusters:** A cluster is a set of data objects if any data object in a cluster is nearest (more similar) to the "center" of a cluster, than to the center of any other data object cluster. The center of a cluster is frequently a centroid [10].
- **Contiguous clusters:** A cluster is a set of data items so that a data item in a cluster is nearest (or more similar) to one or more other data items in the cluster as compared to any other data items that is not in that cluster.
- **Density-based clusters:** A cluster is a collection of data items with dense region of points, which is separated by according to the low-density regions, from other regions that is of high density regions of the data sets.
- **Shared Property or Conceptual Clusters:** Finds a data item clusters that share some common property or characteristics to represent a particular concept.

1.3 MACHINE LEARNING

The machine learning provides ability to systems that enable it to automatically learn and improve it from experience. It is an application of artificial intelligence (AI) that

not required ML (Machine Learning) to be programmed explicitly [11]. Their main goals for developing a programs that are able to access data sets and use it to learn and find new sets form them. The direct experience or instructions are the examples of learning process data observations techniques. The data observed from that are used to analyze the patterns in data and make better decisions based on future results [12]. The main purpose of ML is to allow computers to learn automatically without involvement and assistance of human and accordingly correct required actions.

When there is need to use Machine Learning:

There are some cases when there is need of machine learning other than directly or simply programed computers system to bring out the task are problem complication and requirements for adaptively.

Tasks which are more complex and difficult to programs are given below:

- **Tasks Performed by Humans beings:** There are number of such responsibilities that are performed by human creatures on routine basis but the main point here is how a programming can be done to perform them [13]. The car driving, speech acknowledgement and image accepting are the examples of such type of tasks. In all of these activities, various phases of the art machine learning programs, it basically "learn from their past capability," attain reasonably results are given, once they are showing to suitably many training sets examples.
- **Responsibilities away from Human Abilities:** A large variety of tasks that are advantage from machine learning methods is associated to the examination of wide and complex data sets. The various sets of data, revolving medical records into medical information, weather forecast, and analysis of genomic data, Web search engines, and electronic commerce are the examples of the data sets. With more and more existence of electrical or digitally detailed data, it becomes observable that they are various resources of important information hidden in data records that are really too much large and too much difficult for persons to create logic.

  Learning to find the significant patterns in bulky and difficult data sets is a capable with domain in which the arrangement of programs that learn with the

practically boundless memory volume and ever growing processing speed for computer system for providing the new possibilities to the users [14].

Adaptively: Some drawback of programmed tools is their inflexibility once the user writes the program and installed, after implementation it will remain unchanged. However, many different responsibilities change time to time and user to user. Machine learning tools programs which performance accept the input data sets and suggest a solution to such problems [15]. These are adaptive to their environment changes according to the requirements of users. Various successful applications for machine learning to these types of problems include programs that decrypt the handwritten text, where a static program can adjust to changes between the writing of various users; include spam recognition programs, it automatically to alteration the nature of junk e-mails; and speech recognition platforms.

1.4 HOW MACHINE LEARNING USED IN DATA MINING

Machine learning and data mining are research areas of computer science whose quick development is due to the advances in data analysis research, growth in the database industry and the resulting market needs for methods that are capable of extracting valuable knowledge from large data stores [16]. This chapter gives an informal introduction to machine learning and information mining, and describes selected machine learning sets and information mining methods explained by examples. In the simplest case, data mining techniques operate on a single data table. Different rows in the data table with respect to their training sets to be analyzed in terms of their attributes and the classes to which they fit. There are two main methods:

- **Supervised learning**: In supervised learning all training set and test data sets are already classified into various classes.
- **Unsupervised learning:** In unsupervised learning data sets are not classified into classes and not categorized. In both cases, the goal is to induce a model for the entire dataset, or to discover one or more patterns that hold for some part of the dataset.

Number of researchers are working on using Machine learning in data mining for getting rid of data mining issues and improving its working efficiency [17]. There are number of existing techniques that can be used for the same purpose.

Machine learning is a different and stimulating field [18], and there are various methods of describing it:

- **The Artificial Intelligence View:** Learning is central to human information and talent and it is also critical for constructing intelligent machines. Years of effort in AI have shown that attempting to build clever computers via programming all the policies can't be done; automated gaining knowledge of is crucial. For example, we human beings are now not born with the potential to recognize language then we analyze it and it makes experience to attempt to have computer systems learn language as a substitute of making an attempt to program it all it.
- **The Software Engineering View:** Machine learning lets in us to application computer systems by using example, which can be simpler than writing code the typical way
- **The Stats View:** Machine gaining knowledge for computer science and statistics: computational methods are applied to statistical problems. Machine learning has been applied to a sizable range of troubles in many contexts, beyond the regular records problems. Machine studying is regularly designed with unique concerns than data (e.g., velocity is often more necessary than accuracy).

1.5 TYPES OF CLASSIFIERS

k-Nearest Neighbor: This type of classifier, a patter data is classified by allocating class label to it which are most commonly characterized the k nearest patterns. Class with minimum k-mean are used to distribute a test pattern that shows that this method is complex to distance gathering. The Euclidean distance metric is employed for getting minimum average distance [19]. All features are normalized into same range this is the main requirement of this metric approach. The k-nearest neighbor classifier is predictable non-parametric classifiers which are used for better performance for optimum values of k.

- **Bayesian Classifier:** In supervised parametric classifiers theory, most general approach used is quadratic discrimination. When dealing with d-dimensions the obtained decision boundaries by these classifiers can become very complicated. Most of the discriminant function generation computation has been done off-line. This approach can be more affected by curse of dimensionality as in this quadratic discriminant a huge amount of factors need to be considered. In case of small training samples its performance is affected drastically.
- **Multi-layer Perceptron (MLP):** The multi-layer perceptron classifier is a primary feed ahead synthetic neural network. They have used a single hidden layer at the beginning for simplicity (simplifies deciding on the range of neurons) and then went for two hidden layers for better classification performance. The hidden units had been chosen in a different way for every records set. The range of hidden neurons was once observed out experimentally over a variety of trials [20]. A rule of thumb two is to pick the number of hidden neurons such that the complete quantity of weights in the internet is roughly n/10, n being the complete number of education points. The neural network was educated the use of the back-propagation algorithm, According to the multi-layer perceptron skilled the use of the back-propagation learning algorithm approximates the most effective discriminant function defined by means of Bayesian theory.
- **SVM Classification:** SVM is a classification algorithm based on optimization theory and initially developed by [21]. Here, an object is considered as an n-dimensional vector and it separates such objects with an n-1 dimensional hyper plane. This is referred to as a linear classifier. There are many hyper planes that are used to classify data.

CHAPTER 2

LITERATURE SURVEY

M. Sharma, et.al, (2017), Background have observed medical informatics as an unrestrained growth in database. Nowadays medical industry have huge amount of data sources that are of use only if these are analyzed on time effectively. The known and unknown available patterns in medical databases are investigated using artificially intelligent data mining techniques. According to research paper, author explored the practices of diverse data mining techniques, the role of used dataset, effect of preprocessing and performance of different data mining techniques. The aim of authors is to give different medical science used data mining techniques. The review shows that significant effort been made for mining the data allied to the Cardiology and Diabetes. Their survey shows that the number of papers published in cardio, diabetes, digestive, dentistry and ophthalmology disease diagnosis using data mining are 42%, 26%, 18%, 10% and 4% respectively. They have given a focus on making a new model for ophthalmology, dentistry and digestive disorders type's diseases [22]. The rate of usage of preprocessing in diagnosis of different disorders related to cardio, diabetes, digestive, dentistry and ophthalmology lies between 10.65%–17.75%, 8.48%–14.80%, 4.58–8.93%, 2.96%–7.73% and 5.83%–12.93% respectively. In the end they have develop smart diagnostic system to aware and save human masses from wide critical spectrum of diseases related to ophthalmology, oral and digestive systems.

Yu-Xuan Wang, et.al, (2017), have analyzed that the data mining and machine learning significance has been highlighted in different application scenarios. To analyze the huge amount of data different data mining and machine learning techniques are used to create more commercial values in high end enterprise systems. The advancement in technology has made it possible to use data mining and machine learning on personal computers or embedded systems type low end systems. In this paper [23], authors have proposed research on the management de-signs of different components of the system; most of the work are built upon the characteristics of the system, which may change from time to time. The proposed approach makes it impossible to optimize the system performance with static, or statically adaptive,

system designs. So, authors have proposed an idea to embed the supports of data mining and machine learning to the design of operating system. This makes it able to discover a new, automatized way to adaptively optimize the systems without using complex algorithms. They have used cache design as a data set in order to authenticate the planned idea in this decision maker are used to automatically controlled the replacement of cached contents. The system monitor collected data are analyzed after decision maker replied on a data miner. A sequence of trials has been performed in order to verify efficiency of the considered case that shows improved results.

Zhiqiang Ge, et.al, (2017), have recommended that over a past several decades an important role is played by analytics and data mining in information discovery and choice making/supports in the technique industry. Machine learning serves as primary tools for statistics extraction, computational engine, information pattern consciousness and predictions. In this paper, authors have provided an overview on current data mining and analytics functions in the system enterprise over the previous countless decades. They have considered eight unsupervised and ten supervised learning algorithms in state-of-art of data mining and analytics. Authors have also given an application status of semi-supervised learning algorithms [24]. Both unsupervised and supervised laptop gaining knowledge of methods have already been widely used in the manner industry, which about bills for 90%-95% of all applications. The semi-supervised machine studying has been introduced in latest years , accordingly its application will grow to be extra popular in the close to future. It can be expected that records mining and analytics will play increasingly vital roles in the process industry, with the development of new desktop studying technologies. In this respect, the role of statistics mining and analytics is to carry researchers and practitioners from one of kind cultures to work together.

Anna L. Buczak, et.al, (2016), have given a evaluation on exceptional machine mastering (ML) and records mining (DM) methods for cyber analytics in assist of intrusion detection. The strategies that are the most wonderful for cyber purposes have now not been established. Existing techniques are very complex so, it is not possible to make one suggestion for every method, based on the kind of assault the system is supposed to detect. Based on the quantity of citations or the relevance of rising method [25], papers demonstrating every approach were identified, read, and

summarized. Because information is so important in ML/DM approaches, some frequent cyber information sets used in ML/DM are described. The difficulty of ML/DM algorithms is addressed, dialogue of challenges for the use of ML/DM for cyber security is presented, and some hints on when to use a given approach are provided. There are number of standards that need to be taken into account while responsible the effectiveness of the methods. In one section of paper accuracy, complexity, time for classifying and unknown illustration with a trained model has been given in feature for each ML or DM method. At the end, they have given the main reason of this paper is to examine the approaches for fast incremental learning which is used for regular updates of models for misapplication and irregularity uncovering.

Chamath Malinda Ariyawansa, et.al, (2016), have analyzed that airport is considered as a main opinion of the state that produces a lasting impression on its users. Currently main challenge tackled by airports is the intricacy of players, processes and the inability of the schemes to share and examine that data. An isolated solution is implemented by many airports in order to face above mentioned challenges. They are not very holistic as only specific processes and functions have been improved by that solution. In order to adjust source chain, share real time information, forecasting definite results, track, achieve and locate all of its assets should be taken care by airport ecosystem. So, there is need to make a integrated, unified, creative and arranged to use platform to make smart decisions and support airports to reach its succeeding level. In paper [26], authors have given a review on certain data mining methods that can be combined in to such organization. Airlines, airport trades area and airport individuals itself is measured for the cause along with data mining methods that can be applied to these individuals. This will help in improving light postponement forecast, customer profiling, segmentation, association rule mining. At the end they have given different better methods for an intelligent airport system.

Jahin Majumdar, et.al, (2016), have recommended Data Mining and Machine Learning is a most current research field in computer science which is appropriate in today's world of immeasurable data. The size of data is getting increased so, there is need of rapidly abstract information from data foundations to aid data investigation research and expand industry and market requirements. The k-means, Apriori,

PageRank etc., are the primary data mining algorithms that are in today use but this can be even more enhanced by machine learning that learns from complex patterns. The SFS and SBS methods are the finest with Forward Selection method. The experimental model suggested uses a SVM because of its accuracy although being a bit computationally heavy [27]. The dataset will be capable to define the level of correctness of SVM. In this paper, authors have focused on the numerous existing methods where Machine Learning algorithms have been used to expand records classification and pattern credit in Data Mining especially for Feature Selection. At the end the existing techniques are compared and best one if founded from them. In order to overcome the theoretical limitations of existing algorithms they have planned a new heuristic approach.

Priyanka Dhaka et al. (2016), do survey that Mental health fundamentally degree with high rate of depression, disorder and type of disorder that the humanoid is affected. They use genetic algorithm and big data tool MongoDB for analysis purpose [28]. Genetic algorithm yield optimized result by applying random operations on data and MongoDB is used for together processing, searching and storing the data. Genetic algorithm give optimum solution .Consequence of these applied technique recommend the superior cure of Mental disorder and support doctors to give healthier treatment to their patient with more polished information, in a lesser amount of time and cost.

Lijun Hao,Shumin et al. (2016), they work on open stack and cloud computing, Hadoop technology used for better improvement in the medical fields. Some of most popular data mining tools are used like R language, SQL, Excel, SAS etc. [29]. They also normalized the data and remove incomplete, redundant data and make various classes of it. By using these tools and various mining algorithm they integrate the various most famous analytical mining tools and compare their results. They also configure the personal working platform and reduce cost for device. They configure the virtual machine concept for processing big data related to medical field.

Miroslav Bursa et al. (2011), they use Novel nature inspired techniques for mining the loosely structured medical data. They deploy Ant colony optimization (ACO) approach for loosely structured data records and for clustering they use DTree Method [30]. This approach discovered the shortest path between the large dataset. Output that structure show the further processing. By using this approach it automatically grouped the relevant literals and makes their clusters. It will increase the speed of loosely structured texted attribute and construct a lexical analysis grammar for comparing the classical methods. Speedup can allow performing more and more iteration loosely structured data for making it better and efficient for processing.

Guiduo Duan et al. (2016), they work on FP-Growth algorithm (Extended Prefix-Tree Structure) for correlating the medical insurance data cost and relevant factors [31]. They use pre –pruning and post- pruning process and classification of data based on three criteria like accuracy, stability and complexity. FT-Growth algorithm work on the concept of Divide- and –Conquer method which compress the frequent data items into smaller data structure called FT-Tree. MDMP (Medical Decision Model with Pruning) use pruning concept in decision tree construction. They use FT-Growth algorithm for extracting the dataset and then calculate their information gain value. According to that value they decide whether to the discard dataset or not. For future work, they will focus on how to improve the accuracy of classification.

B.V. Kiranmayee et al. (2016), they focus on brain tumour detection. They purposed a methodology which detects the brain tumour on training and testing phases. They purposed an algorithm that classifies the images as tumour prone or not [32]. They work on ID3 (Decision Tree Builder) algorithm for detection purpose and also classification prototype application datasets for health care. They implement it in Java for making decision tree classification prior to diagnosis of brain tumour deceases. The purposed algorithm can also check the quality of image that is used as for detecting the brain tumour. ID3 algorithm can integrated with real world health care software products.

Hengyi Hu et al. (2017), they work on modular ontologies, authoritative medical ontologies (AMOs), association rules and apriori algorithm. Ontologies represent the highly detailed related to knowledge domain [33]. Its main goal is to describe a method for capture the existing symptom for depression and their related drugs associated with their successful recovery. This approach has two benefits: A) Make assumption regarding to existing patient data. B) reuse the domain knowledge for discovering the new knowledge. They work on similarity functions and Semantic Web Rule Language (SWRL) rules in Protégé. Apriori algorithm is used to mine the frequent data set and their association rules. For further, electrical medical record can be mined by using multi-agent system. Those automatically update the ontologies and automate the data entry.

Razieh Asgarnezhad et al. (2017), they work on efficient preprocessing techniques for replacing the missing and selecting the well-known data set for diabetes mellitus [34]. They firstly remove the missing values with Mean, Median, and KNN. Then selection method can include forward selection and backward eliminate brute force and evolutionary techniques. They work on SVM (Support Vector Machines) for predicting the classification and for optimization they use Genetic algorithm (GA). This will increase the accuracy and precision result for predictive model. GA also improves the performance.

Parisa Naraei et al. (2016), they compare two different algorithms i.e. multilayer perception neural networks and SVM for heart disease detection [35]. SVM is a statistical learning theory which is used for classification. They use WEKA tool for

classification and replacing the missing value by using filter in WAKE tool. Back propagation algorithm used in neural network. It boosts the accuracy and preprocessing techniques process. The results are comparable with other studied algorithms that are: Neural Networks, Decision Tree, and Naïve Bayes for same data sets. Measure their accuracy performance for those data sets.

Houssem Turki et al. (2014), they work on Knowledge Discovery Data (KDD). They deploy the Bayesian Network (BN) and Dynamic Bayesian Network (DBN) for temporal Knowledge Discovery Data (KDD) [36]. BN represent the prior distribution of random variables. A previous purposed method has limitation of initialization problem. But use of BN can decrease the amount of calculations which authority only valuable variables for prediction. They work to achieve the main objective to develop the incremental algorithm for DBN (Dynamic Bayesian Network) structure for:
a) Huge amount of pragmatic data.
b) Heterogeneous skilled of information.
c) State art for contributing in the algorithm developed for learning structure.
Purposed algorithm main objective is to provide the incremental organization learning algorithm for a system to predict the mental impedance in DS children. It also increases the accuracy of diagnosis.

Sneha Chandra et al. (2015), they deploy the enhancement in classification accuracy by using adaptive classifier using various image processing and existing classification algorithms. They work on Bayesian Classification, Decision Tree classification, Ensemble classification, Laplacian correction, Euclidean distance, K-Mean clustering, Mean Gray level algorithm and Rule-Based classification [37][38]. Their main objective is to produce the extra definite, detailed and precise result. Classifier used for prediction purpose and for increasing the classification accuracy. They use Bagging as the collaborative method for enlightening the classification accuracy. Higher accuracy can be achieved by AC (Adaptive Classifier but it is still insufficient to achieve 100% accuracy.

Shubpreet Kaur et al. (2015), they work on KDD and WEKA tool is used for measuring the Drug addiction. Their main objectives are:

a) Generate the efficient way to extract the meaningful data
b) Predict the diseases with higher accuracy and lower cost
c) Simultaneously retrieve the information and minimize the effort.

They compare the various data mining techniques like ANN, Decision Tree, Logistic Regression, KNN, NB, SVM and apriori algorithm and various data mining tools are compared like WEKA, TANAGRA, ORANGE, R, RAPID MINER, KNIME etc., [39] but most used data mining tool is WEKA and technique is D Tree, NB and ANN. Decision Tree give maximum accuracy and less human efforts are needed in this algorithm.

CHAPTER 3

PRESENT WORK

3.1 PROBLEM FORMULATION

The data mining is the approach which extracts the useful information from the raw data. The clustering is the approach which can cluster similar and dissimilar type of data. The k-mean clustering is the partitioned type of clustering which works in three phases. In the first phase, the dataset is taken as input which gets pre-processed. In the second phase, the arithmetic mean is whole dataset is calculated which define centroid point . In the last phase, the Euclidian distance will be calculated from the central point and points which have similar distance will be clustered in first cluster and other in the second cluster. To predict the human disorders the technique if classification will be applied which can mark the clusters which have human disorders and which don't have disorders. In the k-mean clustering the Euclidian distance is calculated statically which reduce accuracy of classification. In this research work, k-mean algorithm will be improved to increase accuracy of classification

3.2 OBJECTIVES

Following are the various research objectives: -

1. To Study as well as Analyze the various prediction algorithms in data mining.

2. To improve performance of k-mean and SVM classification based prediction analysis algorithms in data mining

3. The proposed improvement will be based on the neural networks to improve accuracy of classification

4. Implement proposed technique and compare with existing in terms of various parameters.

3.3 RESEARCH METHODOLOGY

The clustering is efficient approach which can cluster similar and dissimilar type of data. The k-means is the efficient clustering algorithm. The k-means algorithm can pre-process the data, arithmetic mean of the whole dataset will be calculated which define centroid point. In the last phase, the Euclidian distance will be calculated from the central point which clusters similar and dissimilar data. The SVM classifier will be applied on the cluster data to classify dataset in two classes of humans with brain disorder and humans with brain disorder. In this research, the back propagation will be applied in the k-means clustering algorithm which can calculate Euclidian distance in dynamic manner. When the Euclidian distance will be calculated in dynamic manner it directly increases accuracy of clustering. . The back propagation algorithm is the algorithm which studies from the preceding involvements and drive new values. The formulation given below is used to drive values from the input dataset. In the formula given the x is the each point in the dataset and w is the value of the data point from which the actual output is taken and biases the value which is used to change the final value of output. The output of each iteration is compared with the output of next iteration and iteration at which error is minimum is the final value of Euclidian distance. When the error is reduced , the accuracy of clustering is increased and execution time is reduced.

$$Error = Desired\ Ou$$

Output: $\sum_{\substack{w=0 \\ x=0}}^{\substack{w=n \\ x=n}} x_n w_n + bias$

$$Error = Desired\ Output - Actual\ Output$$

Fig 3.1: Euclidian Distance Formula

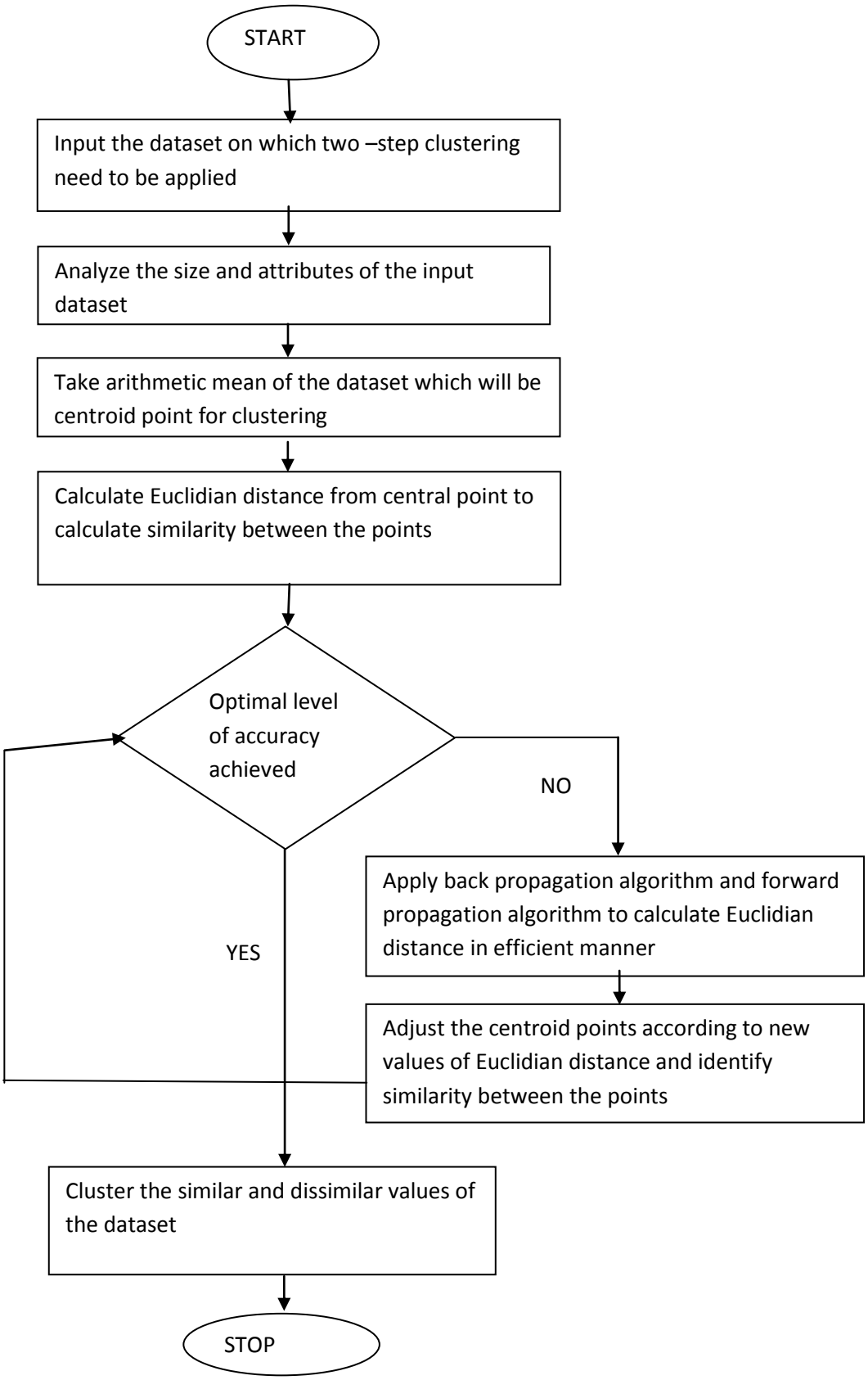

Fig 3.2: Proposed Flowchart

3.4 EXPECTED OUTCOMES

The various expected outcomes from our research work are as described bellow :

1. In this research work, improvement in the k-mean algorithm can be done which increase accuracy of classification. This leads to increase prediction values for human disorder with data mining

2. The proposed improvement can also reduce execution time.

CHAPTER 4

CONCLUSION

4.1 CONCLUSION

In this research work, it has been concluded that clustering is approach to cluster similar and dissimilar type of data. The k-means algorithm works in the three phases in which data is pre-processed, the in second phase central point is defined and in the last phase Euclidian distance is calculated on the basis of which data get clustered. The SVM classifier will be applied which will classify similar and dissimilar type of data. In this research, the k-mean algorithm will be improved using back propagation algorithm which increase accuracy and reduce execution period.

REFERENCES

[1] Jiawei Han, Micheline Kamber. "Data Mining: Concepts and Techniques." Vol. 3, PP. 1-31, 2000.

[2] M.Gary Weiss, D Brian. Davison. "Data Mining: To appear in the Handbook of Technology Management." H. Bidgoli (Ed.), John Wiley and Sons, Vol. 3, PP. 121-140, 2010.

[3] Osmar R. Zaïane. "Chapter I: Introduction to Data Mining." CMPUT690 Principles of Knowledge Discovery in Databases. Vol. 2, pp. 1-19, 1999.

[4] T. Fawcett, F Provost. "Adaptive fraud detection Data Mining and Knowledge Discovery." Vol. 3, pp. 291-316, 1997.

[5] C. X. Ling, C. Li. "Applying Data Mining to Direct Marketing." In W. Kou and Y. Yesha (eds.), Electronic Commerce Technology Trends: Challenges and Opportunities, IBM Press, Vol. 3, pp. 185-198, 2000.

[6] Y. Chen, G Zhang, D. Hu, S. Wang. "Customer segmentation in customer relationship management based on data mining", In Knowledge Enterprise: Intelligent Strategies in Product Design, Manufacturing and Management, Boston: Springer, vol. 3, pp. 288-293. 2006.

[7] Jyoti, Neha Kaushik, Rekha. "Review paper on Clustering and Validation Techniques", International Journal for Research in Applied Science and Engineering Technology." Vol. 2, pp. 182-186, 2014.

[8] Dr. S. Rajagopal. "Customer data clustering using data mining technique",2011 International Journal of Database Management Systems (IJDMS), Vol. 3, PP. 21- 32.

[9] Fraley, Andrew, and Thearting, Kurt. "Increasing customer value by integrating data mining and campaign management software in Data Management." Vol. 2, pp. 49–53, 1999.

[10] P. Bhargavi, S. Jyothi. "Applying Naïve Bayes Data Mining Technique for Classification of Agricultural land Soils." International Journal of computer Science and Network Security IJCSNS, Vol. 9, PP. 117-122, 2009.

[11] Shai Ben-David Shai Shalev-Shwartz. "Understanding Machine Learning: From Theory to Algorithms." PP. 1-499, 2014.

[12] Alex Smola, S.V.N. Vishwanathan. "Introduction to Machine Learning", the press syndicate of the university of Cambridge. PP. 1-2, 2008.

[13] Ackerman, M. & Ben-David. “Measures of clustering quality: A working set of axioms for clustering.” In Proceedings of Neural Information Processing Systems (NIPS). Vol. 4, PP. 121–128, 2008.

[14] P. Bartlett, O. Bousquet, S. Mendelson. “Local rademacher complexities, Annals of Statistics.” Vol. 4, PP 1497–1537, 2005.

[15] Bartlett, P. L. & S. Ben-David. “Hardness results for neural network approximation problems.” Theor. Comput. Sci., vol. 1, pp. 53–66, 2002.

[16] P. Stone, M. Veloso. “Multiagent systems: A survey from a machine learning perspective.” Autonomous Robots. Vol. 3, pp. 345–383, 2000.

[17] L. Saitta, F. Neri. “Learning in The Real Machine Learning.” Volume. 3, pp. 133–163, 1998.

[18] S Salzberg . “A nearest hyperrectangle learning method. Machine Learning”, Vol. 3, pp. 251–276, 1991.

[19] R.O. Duda. “Pattern Classification 2nd Edition.” 2000 John Wiley & Sons Inc..

[20] D.W. Ruck. “The Multi-Layer Perceptron as an Approximation to A Bayes Optimal Discriminant Function.” IEEE Transactions On Neural Networks, Volume 1. 1990.

[21] C. Cortes et.al. “Support Vector Network And Machine Learning”, Vol. 20, pp. 273-297, 1995.

[22] M. Sharma, G. Singh, R. Singh, “Stark Assessment of Lifestyle Based Human Disorders Using Data Mining Based Learning Techniques”, Volume 5, PP. 202-222, 2017.

[23] Yu-Xuan Wang, Sun QiHui , Ting-Ying ChiNn PC Huang. “Using Data Mining and Machine Learning Techniques for System Design Space Exploration and Automatized Optimization.” Proceedings of the 2017 IEEE International Conference on Applied System Innovation, Vol. 15, PP. 1079-1082, 2017.

[24] Ge Zhejiang, Zhihuan Song, X. Steven Ding. “Data Mining and Analytics in the Process Industry: The Role of Machine Learning.” Translations and content mining are permitted for academic research only, Vol. 5, pp. 20590-20616, IEEE 2017.

[25] Anna Buczak, Er Guven. “A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection”, IEEE Communications Surveys And Tutorials, Vol. 19, PP. 1153-1176, 2016.

[26] Ariyawansa Malinda, Achala Aponso, “Review On State of Art Data Mining and Machine Learning Techniques for Intelligent Airport Systems.” International Conference of Information Management (ICIM), 2nd International Conference, vol. 6, pp. 122-128, IEEE 2016.

[27] Shruti Gupta, Jahin Majumdar, Anwesha Mal, “Heuristic Model to Improve Feature Selection Based on Machine Learning in Data Mining”, 2016 6th International Conference Cloud System and Big Data Engineering (Confluence), Vol. 3, PP. 73-77, 2016.

[28] P. Dhaka and R. Johari, “Big Data Application : Study and Archival of Mental Health Data , using MongoDB,” pp. 3228–3232, 2016.

[29] L. Hao, S. Jiang, B. Si, and B. Bai, “Design of the Research Platform for Medical Information Analysis and Data Mining,” pp. 1985–1989, 2016.

[30] M. Bursa and L. Lhotska, “Novel Nature Inspired Techniques in Medical Data Mining,” vol. 7, pp. 286–288, 2011.

[31] G. Duan, D. Ding, and A. F. P. G. Algorithm, “An Improved Medical Decision Model Based on Decision Tree Algorithms,” no. 2014, pp. 151–156, 2016.

[32] B. V Kiranmayee, “A Novel Data Mining Approach for Brain Tumour Detection.”

[33] H. Hu, L. Kerschberg, A. A. Medical, and O. Amos, “Standardizing the Crowdsourcing of Healthcare Data Using Modular Ontologies,” pp. 107–112, 2017.

[34] R. Asgarnezhad, M. Shekofteh, and F. Z. Boroujeni, “IMPROVING DIAGNOSIS OF DIABETES MELLITUS USING COMBINATION OF PREPROCESSING TECHNIQUES,” vol. 95, no. 13, pp. 2889–2895, 2017.

[35] P. Naraei, V. Street, V. Street, and V. Street, “Application of Multilayer Perceptron Neural Networks and Support Vector Machines in Classification of Healthcare Data,” no. December, pp. 848–852, 2016.

[36] H. Turki and M. Ben Ayed, “Using Dynamic Bayesian Networks for the Down Syndrome,” pp. 163–167, 2014.

[37] S. Chandra, “Creation of an Adaptive Classifier to Enhance the Classification Accuracy of Existing Classification Algorithms in the Field of Medical Data Mining,” pp. 188–193, 2015.

[38] S. Chandra and M. Kaur, “Enhancement of Classification Accuracy of our Adaptive Classifier using Image Processing Techniques in the Field of Medical Data Mining,” pp. 948–954, 2015.

[39] S. Kaur and R. K. Bawa, “Future Trends of Data Mining in Predicting the Various Diseases in Medical Healthcare System,” vol. 6, no. 4, pp. 17–34, 2015.