iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** VIII **Month of publication:** August 2017

DOI: http://doi.org/10.22214/ijraset.2017.8022

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue VIII, August 2017- Available at www.ijraset.com

Recommendation System with Automated Web Usage Data Mining by Using k-Nearest Neighbour(KNN) Classification and Artificial Neural Network (ANN) Algorithm

Er. Jyoti[1], Er. Jagdeep Kaur[2]
[1,2]*Deptt.of CSE, Sant Baba Bhag Singh University, Jalandhar, India*

Abstract: ***Since the main problem of many on-line websites is the presentation of a lot of choices to the various clients at a time. This usually results in a time-consuming task in finding out the right product or information on the site. The user's current interest depends upon the navigational behavior which helps the organizations to guide users in their browsing activities and obtain some relevant information in a short span of time. Since the resulting patterns which are obtained through data mining techniques did not perform well in the prediction of future browsing patterns because of the low matching rate of resulting rules and of user's browsing behavior. This paper focuses on the study of the recommendation system and automatic web usage data mining which is based on current user behavior through his/her click stream data. The K-Nearest-Neighbor (KNN) algorithm has been trained to be used in real-time and on-line to identify clients and visitors clickstream data, matching it to a particular user group and recommends a tailored browsing option that meets the needs of the specific user at the particular time. But these have some major issue if the data is going to be varied, the clustering approach that was used in traditional work can only capable if the data variation was within the cluster information they are having, if data goes out of bound it was difficult to perform classification. So there is need to add a classifier approach so can work in these conditions too. In this paper, hybridization of traditional KNN and ANN is done which leads to the improved accuracy. In traditional KNN only distance between given two users is calculated.***

Here, newest accuracy of KNN is calculated by finding out the distances within all the users.

Keywords: Artificial Neural Network; Automated; Data Mining; K-Nearest Neighbor; Recommendation system ; Web Usage Mining.

I. INTRODUCTION

A. Data Mining

Data mining is the process which comes under the category of computer science in order to investigate large data sets which belong to the pattern. Here large data set stands for Big Data. Data mining is an automatic process which is used to extract meaningful information from the data storage and further use this information for various purposes[15,16].The extraction of meaningful data can be performed by matching patterns and it is achieved by cluster analysis, anomalies analysis, and dependencies analysis. Spatial indices are used to perform all above functions or processes. The matched pattern is a form of the brief summary of data stored in the data warehouse and these patterns are used for future prediction and various decision-making systems to take a right decision. [4]

For example, in the case of machine learning systems, this extracted information can be used for prediction analysis. Another example, data mining is a process which finds or investigates various groups of correlated data in the database which further can be used for predictive analysis in near future[17]. Data analysis, data collection, compilation of data is not a connected to the data mining but still included in the process of KDD i.e. Knowledge Discovery Database.[1]

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue VIII, August 2017- Available at www.ijraset.com

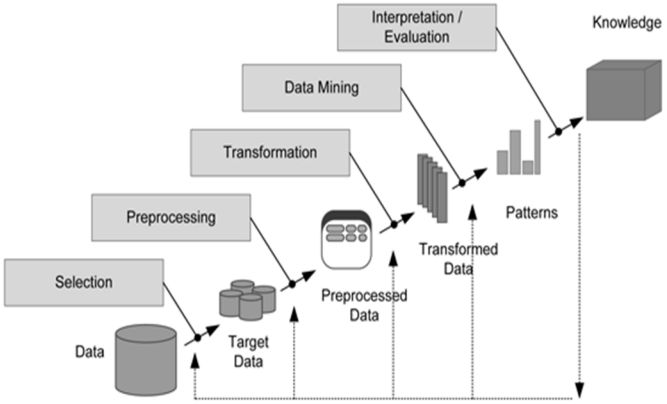


Figure 1 : Knowledge Discovery in Databases

Data mining is a process which is used to search a large amount of data in order to find the useful data. The goal of this technique is to find patterns that were previously unknown. Once the patterns are found they can further be used to make certain decisions for the development of their businesses. The iterative process consists of the following steps:

1) *Data cleaning:* It is a phase in which noisy data and irrelevant data are removed from the collection.
2) *Data integration:* At this stage, multiple data sources, mostly heterogeneous, may be combined in a common source.
3) *Data selection:* At this step, the data which is relevant to the analysis is decided on and retrieved from the data collection.
4) *Data transformation:* Data transformation is also known as Data Consolidation. It is a phase in which the selected data is transformed into appropriate forms for the mining procedure.
5) *Data mining:* It is the crucial step in which clever techniques are applied to extract patterns which are potentially useful.
6) *Pattern evaluation:* In this step, strictly interesting patterns representing knowledge are identified based on the given measures.
7) *Knowledge representation:* It is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help the users to understand and interpret the data mining results.[2]

B.Web Usage Mining

It is one of the applications of the techniques of data mining to find out the interesting patterns from the Web data. Usage data captures the origin or identity of Web users along with their browsing behavior at the Web site. It generally focuses on techniques which predict the user behavior while the user is interacting with the Web. [9,11,13]The potential strategic aims in each domain into mining goal as the prediction of the user's behavior within the site, a comparison between expected and actual Web site usage, adjustment of the Web site to the interests of its users. There are no such definite distinctions between Web usage mining and other two categories. In the process of data preparation of Web usage mining, the Web content and Web site topology will be used as the information sources, which interacts Web usage mining with the Web content mining and Web structure mining.[10] Moreover, the clustering in the process of pattern discovery is a bridge to Web content and structure mining from usage mining.[3]

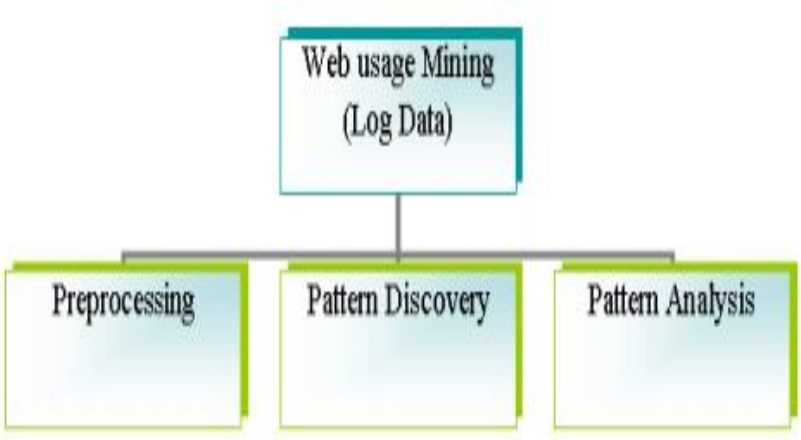


Figure 2: Contents of Web Usage Mining

II. LITERATURE REVIEW

Anand V. Saurkar present data mining and its various tasks related to the data mining. Data mining is vast research field for various researchers. The various tasks like rules learning, detection, classification, clustering etc have been demonstrated in this paper by the author. Data mining is basically introduced for the analyzing a large amount of data that is presented in the database. As a large amount of data is stored in the database for its maintenance a new field is emerging that is named as the data mining. Data mining is the process of the extraction of the useful data from the database. So here various tasks that are involved in the process of the data mining are demonstrated.[5]

Arno J. Knobbe, in his paper presents Data mining is the extraction of the useful and helpful information from the data set. The author of this paper has explained various tools for the data mining process that are based on the various machine learning algorithm. In this various algorithm have been studied like single-table is considered as one of the best algorithms but the major drawback is that the complex patterns present in the data are not expressible simply. Another algorithm is the ILP that work on the full relational database this algorithm is efficient and the scalability of the system is made. The framework and the architecture is designed in this system. The framework is the information of the data schema and the architecture designed will consist of the primitives that are used in defining the efficiency of the system. With the help of the WARMR algorithm the framework is demonstrated.[6]

Hüllermeier, Eyke. has presented the survey on various application and the contributions of the fuzzy sets that are used in the various fields like machine learning, data mining etc. Fuzzy sets are considered as the efficient methods of the extraction of the data from the database. From the experiment performed it is concluded that these systems are quite efficient and can be used for various future researches.[7]

Klose, A., et al. has presented that the data mining is a process which includes the various steps of KDD i.e. knowledge discovery in the database. In KDD modeling techniques are implemented. The research fields like machine learning, statistics, and AI use data mining. In this author defines that neuro-fuzzy methods are plays vital role in the field of data mining because they lead to the efficient results. But comprehensibility in performance is not easy to achieve it misguides the aim of the research or system.[8]

Xindong Wu, presents the top ten data mining algorithms which are identified by the IEEE International Conference on Data Mining (ICDM) in December 2006: C4.5, k-Means, SVM, Apriori, EM, PageRank, AdaBoost, kNN, Naive Bayes, and CART. These top 10 algorithms are among the most influential data mining algorithms in the research community. With each algorithm, they provide a description of the algorithm, discuss the impact of the algorithm, and review current and further research on the algorithm. These 10 algorithms cover classification. [12]

John F. Roddick presents with the growth in the size of datasets, data mining has recently become an important research topic and is receiving substantial interest from both academia and industry. At the same time, a greater recognition of the value of temporal and spatial data has been evident and the first papers looking at the confluence of these two areas are starting to emerge. This short paper provides a few comments on this research and provides a bibliography of relevant research papers investigating temporal, spatial and spatiotemporal data mining.[14]

Parpinelli," proposed an algorithm for data mining which is known as Ant-Miner (ant-colony-based data miner). The goal of Ant-Miner is to extract classification rules from the data. This algorithm is inspired by both researches on the behavior of real ant colonies and some data mining concepts as well as principles. Performance of Ant-Miner with CN2 is compared. CN2 is a well-known data mining algorithm for classification, in six public domain data sets. The results provide evidence that: Ant-Miner is competitive with CN2 with respect to predictive accuracy.[18]

Byung-Hoon Park, This paper gives us a brief overview of the DDM algorithms, applications, systems and the emerging research directions. The structure of the paper is organized as follows. We first present the related research of DDM and illustrate data distribution scenarios. Then DDM algorithms are reviewed. Subsequently, the architectural issues in DDM systems and future directions are discussed.[19]

Michael J.Shaw presented that due to proliferation of information systems and technology, businesses increasingly have the capability to accumulate huge amounts of customer data in large databases. However, much of the useful marketing insights into customer characteristics and their purchase patterns are largely hidden and untapped. A current emphasis on customer relationship management makes the marketing function an ideal application area to greatly benefit from the use of data mining tools for the decision support. A systematic methodology that uses data mining and knowledge management techniques is proposed to manage the marketing knowledge and support marketing decisions. This is the methodology which can be the basis for enhancing customer relationship management. 20]

Hongiun Lu, proposed that Classification is one of the data mining problems receiving great attention recently in the database community. This paper presents an approach to discover symbolic classification rules using neural networks. Neural networks have not been thought suited for data mining because how the classifications were made is not explicitly stated as symbolic rules that are suitable for verification or interpretation by humans. With the proposed approach, concise symbolic rules with high accuracy can be extracted from a neural network. The network is first trained to achieve the required accuracy rate. Redundant connections of the network are then removed by a network pruning algorithm. The activation values of the hidden units in the network are analyzed, and classification rules are generated using the result of this analysis. The effectiveness of the proposed approach is clearly demonstrated by the experimental results on a set of standard data mining test problems.[21]

Raymond Chan, Traditional association rule mining algorithms only generate a large number of highly frequent rules, but these rules do not provide useful answers for what the high utility rules are. We develop a novel idea of top-K objective-directed data mining, which focuses on mining the top-K high utility closed patterns that directly support a given business objective. To association mining, we add the concept of utility to capture highly desirable statistical patterns and present a level-wise item-set mining algorithm. With both positive and negative utilities, the antimonotone pruning strategy in Apriority algorithm no longer holds. In response, we develop a new pruning strategy based on utilities that allow pruning of low utility itemsets to be done by means of a weaker but ant monotonic condition. Our experimental results show that our algorithm does not require a user-specified minimum utility and hence is effective in practice.[22]

III. KNN Approach

A. K-NN Algorithm

Out of the various algorithms used for classification, the KNN is the most commonly used algorithm. KNN algorithm is very easy to implement and give fairly good results. Also, KNN algorithm does not require any prior knowledge regarding dataset for classification. It performs classification purely on similarity basis. As the name of algorithm indicates, for classification of a novel tuple it looks out for 'k' nearest tuples to it. The procedure of classification in KNN starts with a data set. The data set is constituted of a certain number of attributes that define a data set. The data set is divided into two sets: the training set and test set. The training set is given as input to the algorithm while test set is used to calculate the accuracy of the algorithm. The division of the data set can be done using various methods such as hold-out method, random sampling, cross-validation etc. KNN classifies any new tuple by using training data tuples similar to it. Due to this KNN is also called local learner. There is no learning phase in KNN. It stores all the training tuples given to it as input without doing anything. All the computations are done at the time of classification of a test tuple. In KNN algorithm, the training tuples can be viewed as a set of data points in an n-dimensional space, where n dimensions are the set of n attributes describing the data set. When an unknown tuple comes for classification, we have to find out the k nearest data points to it in the n-dimensional space. To find the k nearest data points to the unknown tuple various distance metrics are used for example Euclidean distance, Minkowski distance, Manhattan distance.

The K-Nearest Neighbor classifier usually applies the Euclidean distance between the training tuples and the test tuple.

$$d(x_i, x_t) = \sqrt{(x_{i1} - x_{t1})^2 + (x_{t2} - x_{t2})^2 + \ldots + (x_{ip} - x_{tp})^2}$$

In general term, the Euclidean distance between two Tuples for instance

X1 = (x11, x12, ... x1n) and X2 = (x21, x22, ...x2n) will be

$$\text{dist}(x_1, x_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}$$

The pseudo code for the KNN algorithm is given below:

B. Algorithm

Input Parameters: Dataset, k

Output: Classified test tuples

1) Store all the training tuples.

2) For each unseen tuple which is to be classified

a) Compute distance of it with all the training tuples.

b) Find the k nearest training tuples to the unseen tuple.

c) Assign the class which is most common in the k nearest training tuples to the unseen tuple.

End for The K-Nearest Neighbor (K-NN) algorithm is one of the simplest methods for solving classification problems; it often yields competitive results and has significant advantages over several other data mining methods. [18]

d) Providing a faster and more accurate recommendation to the client with desirable qualities as a result of a straightforward application of similarity or distance for the purpose of classification.

e) Our recommendation engine collects the active users' clickstream data, match it to a particular user's group in order to generate a set of recommendation to the client at a faster rate.

IV. ANN Approach

With the rise of modern electronics, it was only natural to try to harness this thinking process. The first step toward artificial neural networks came in 1943 when Warren McCulloch, a neurophysiologist, and a young mathematician, Walter Pitts, wrote a paper on how neurons might work. They modeled a simple neural network with electrical circuits. Neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. A trained neural network can be thought of as an "expert" in the category of information it has been given to analyze.

In electronics engineering and related fields, artificial neural networks (ANNs) are computational or mathematical models that are inspired by a human's central nervous system (in particular the brain) which is capable of machine learning as well as pattern recognition. Whereas animal's nervous system is more complex than the human so the system designed like this will be able to solve more complex problems. Artificial neural networks are generally presented as systems of highly interconnected "neurons" which can compute values from inputs. [24]

Neural Network is just like a website network of interconnected neurons which can be millions in number. With the help of these interconnected neurons, all the parallel processing is being done in a body and the best example of Parallel Processing is human or animal's body.

Currently, artificial neural networks are the clustering of the primitive artificial neurons. This clustering occurs by creating layers which are then connected to one another. How these layers connect is the other part of the "art" of engineering networks to resolve the complex problems of the real world.

So neural networks, with their stronger ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques to be noticed by either humans or other computer techniques.

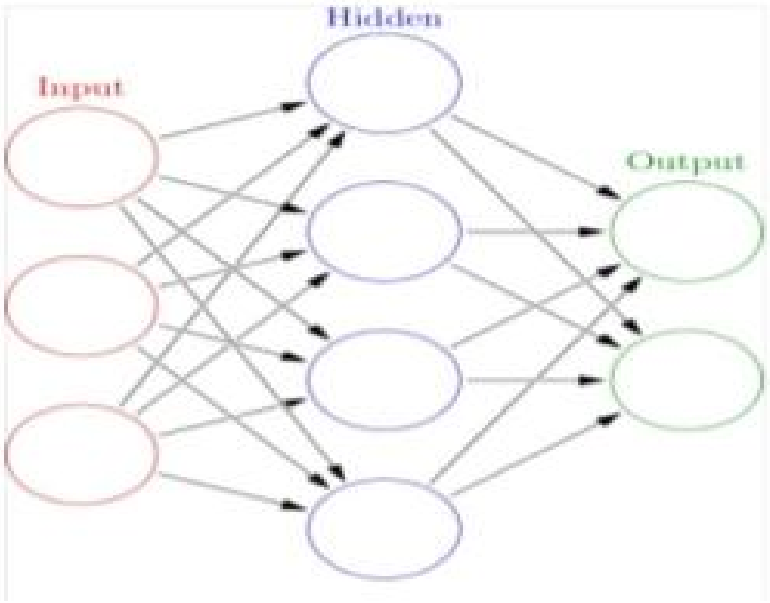


Figure 3: A Simple Neural Network

The word network in the term 'artificial neural network' refers to the interconnections between the neurons in the different layers of each system. An example system has three layers. The first layer has input neurons which send data via synapses to the second layer of neurons, and then via more synapses to the third layer of output neurons. More complex systems will have more layers of

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue VIII, August 2017- Available at www.ijraset.com

neurons with some having increased layers of input neurons and output neurons. The synapses store parameters called "weights" that manipulate the data in the calculations. An ANN is typically defined by three types of parameters: [23]

A. The interconnection pattern between the different layers of neurons
B. The learning process for updating the weights of the interconnections
C. The activation function that converts a neuron's weighted input to its output activation.

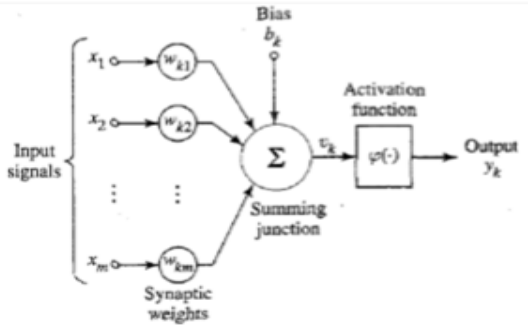


Figure 4: Nonlinear Model of Neuron

V. RESEARCH METHODOLOGY

Firstly, data set is taken to work with for the recommendation. Then, KNN feature extraction process is applied. After that training and testing of KNN and ANN based hybrid proposed model is done for taking decisions. Then, at last, comparison of the traditional approach and proposed work is taken place.

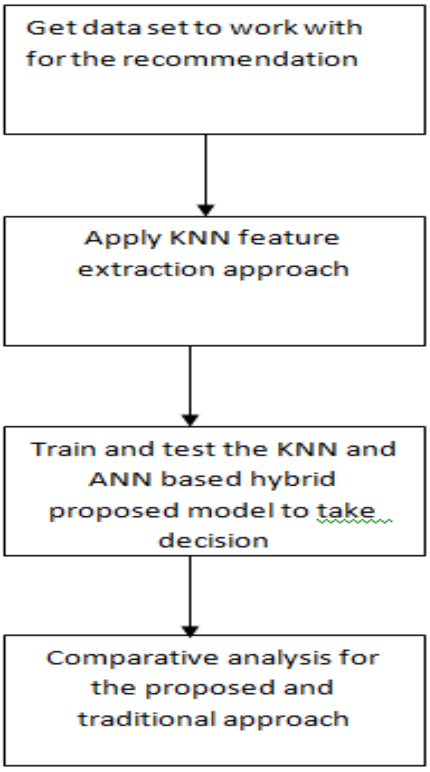


Figure 5: Proposed Work

VI. RESULTS AND DISCUSSION

The proposed and existing algorithm is implemented in MATLAB to test on the desired dataset.

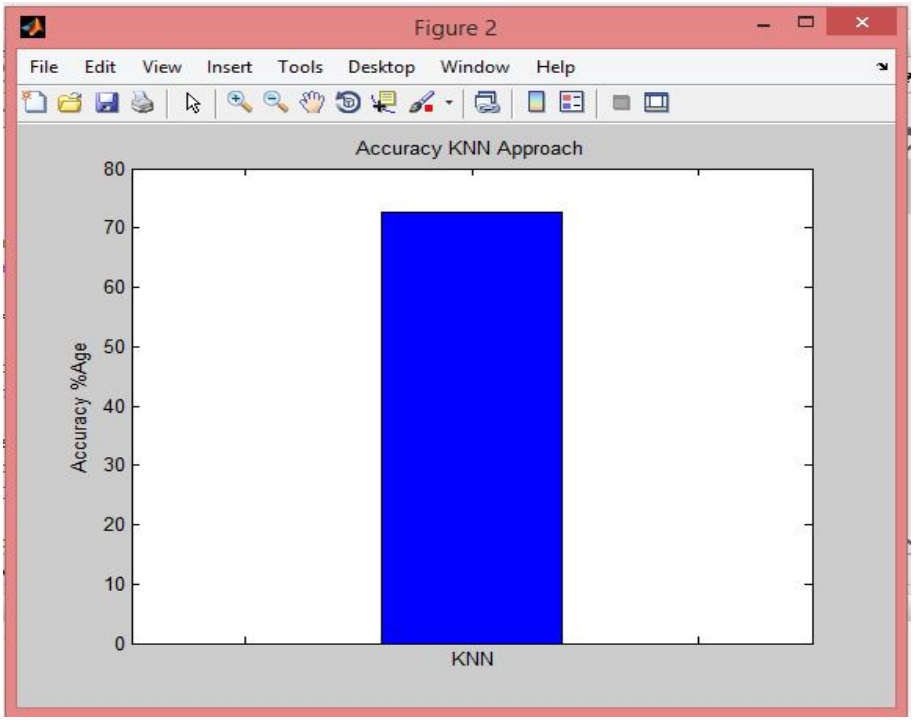


Figure 6: Accuracy KNN Approach

As shown in the figure , the percentage accuracy of KNN is calculated. It is found to be 72.72%. The existing work only calculates the Euclidean distance between the given users. Here, the distance of each and every user is taken with each other.

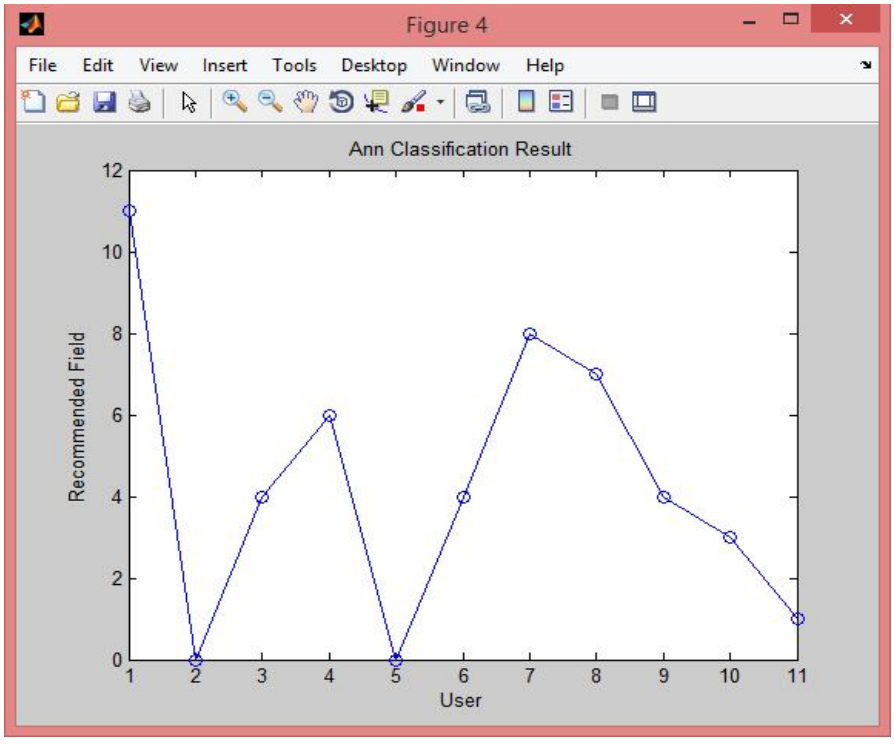


Figure 7: ANN Classification Result

As shown in the figure, here classification of ANN is defined between the user and recommended field.

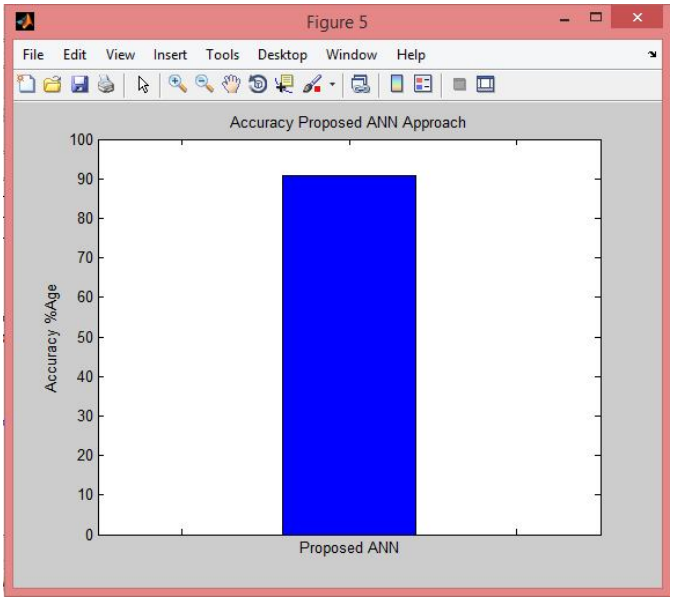


Figure 8: Accuracy proposed ANN Approach

As shown in the figure, here accuracy percentage of proposed ANN approach is calculated which comes out better then the KNN approach. Here, the calculated accuracy is 90.90%.

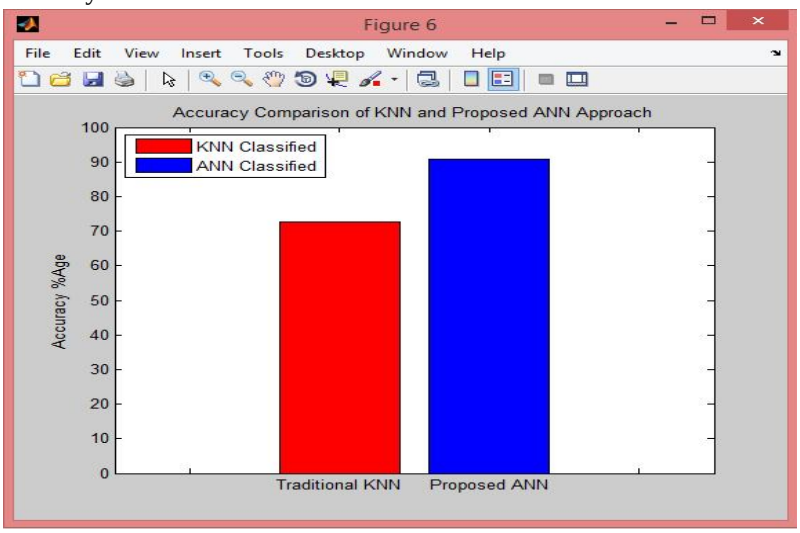


Figure 9: Accuracy comparison of KNN and Proposed ANN Approach

As shown in the figure, accuracy comparison of traditional KNN and proposed ANN is done which is the hybridisation result and it is seen that proposed ANN gives more accurate results then traditional KNN.

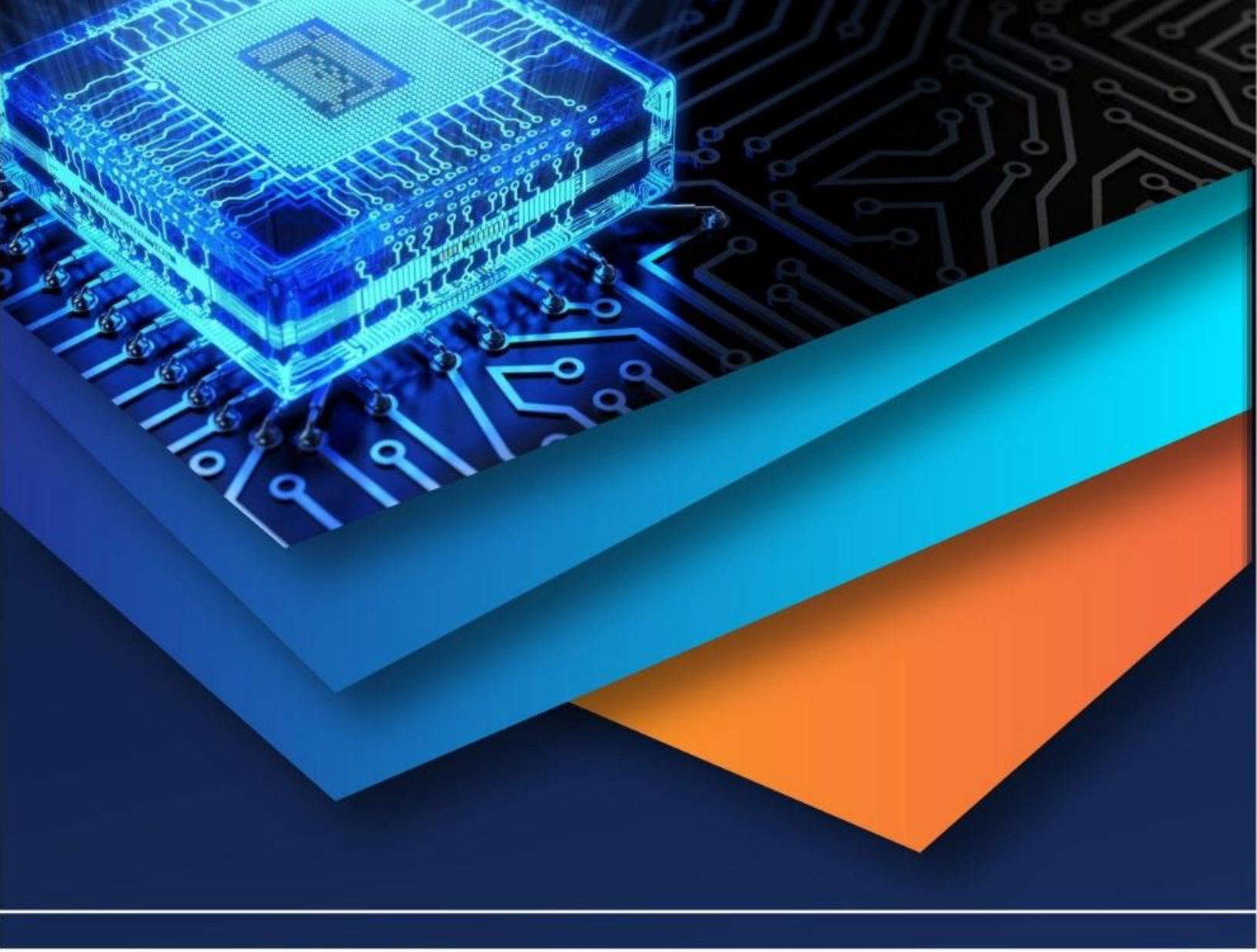
VII. CONCLUSION

This work is carried out by using MATLAB. Thus, the above results shows that when hybridisation of KNN(K-Nearest Neighbour) and ANN(Artificial Neural Network) is done then more accuracy is achieved. Here not only hybridisation achieves good results but KNN when considered with each and every users gives us better results. It also shows better accuracy then the traditional KNN algorithm. When traditional KNN and proposed ANN is compared then accuracy of proposed ANN comes out the best which is 90.90% and that of traditional KNN was 72.72% which helps to improve the recommendation system. Thus, in this way users will able to get accurate desired data.

REFERENCES

[1] Jiawei Han and Micheline Kamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd ed.

[2] Fayyad, Usama; Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996). "From Data Mining to Knowledge Discovery in Databases". http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996Fayyad.pdf Retrieved 2008-12-17.
[3] Dr.R.Lakshmipathy, V.Mohanraj, J.Senthilkumar, Y.Suresh, "Capturing Intuition of Online Users using a Web Usage Mining" Proceedings of 2009 IEEE International Advance Computing Conference (IACC 2009)Patiala, India, 6-7 March 2009
[4] PhridviRaj MSB., GuruRao CV (2013) Data mining – past, present and future – a typical survey on data streams. INTER-ENG Procedia Technology 12:255 – 263
[5] Anand V. Saurkar,(2014) , "A Review Paper on Various Data Mining Techniques", IJARCSSE, vol 4(4), Pp 98-101, 2014
[6] Arno J. Knobbe," Multi-Relational Data Mining", SIKS, Pp 1-130, 2015
[7] Hüllermeier, Eyke. "Fuzzy sets in machine learning and data mining." Applied Soft Computing 11.2, Pp 1493-1505,2011
[8] Klose, A., et al. "Data mining with neuro-fuzzy models." Data mining and computational intelligence", Springer, Pp 1-5, 2001
[9] Hitesh Hasija and Deepak Chaurasia;" Recommender System with Web Usage Mining basedon Fuzzy C Means and Neural Networks"; NGCT-2015
[10] Prajyoti Lopes and Bidisha Roy;"Dynamic recommendation system using web usage mining for E-commerce users";ICACTA-2015
[11] D.A. Adeniyi, Z. Wei,and Y. Yongquan;" Automated web usage data miningand recommendation system using K-Nearest Neighbor (KNN)classification method";Applied Computing and Informatics (2016).
[12] Xindong Wu,"Top 10 algorithm in data mining", Springer, vol 14, Pp 1-37, 2008
[13] Himangni Rathore, Hemant Verma, "Analysis on Recommended System for Web Information Retrieval Using HMM", International Journal of Engineering Research and Applications ISSN : 2248-9622, Vol. 4, November 2014.
[14] John F. Roddick,"A bibliography of Temporal, Spatial and Spatio-temporal data mining research", SIGKDD, vol 1(1), Pp 34-38, 1999
[15] S.Kaviarasan,K.Hemapriya,K.Gopinath,Semantic Web Usage Mining Techniques for Predicting Users' Navigation Requests, International Journal of Innovative Research in Computer Science and Communication Engineering,Vol. 3,Issue 5,[ISSN:2320- 9801],2015.
[16] B.Lalithadevi, A. Mary Ida,W.Ancy Breen, A New Approach for Improving World Wide Web Techniques in Data Mining, International Journal of Advanced Research in Computer Science and Software Engineering,Vol. 3,Issue 1,[ISSN:2277 128X],201.
[17] Shu-Hsien, C. Pei-Hui, H. Pei-Yuan, Data mining techniques and applications- A decade review from 2000 to 2011, Journal of expert system with applications 39 (2012) (2012).
[18] Parpinelli,"Data mining with an ant colony optimization algorithm", IEEE, vol 6(4), Pp 321-332, 2002
[19] Byung-Hoon park, "Distributed Data Minning", citeseer, 2002.
[20] Michael J.Shaw,"Knowledge management and data mining for marketing", Elsevier, vol 32, Pp 127-137, 2001.
[21] Hongiun Lu,"Effective Data Mining Using Neural Networks", IEEE, vol 8(6), Pp 957-962, 1996.
[22] Raymond chan,"Mining high utility itemset", IEEE, 2003
[23] Prof. Leslie Smith, " An Introduction to Neural Networks", University of Stirling., 1996,98,2001,2003.
[24] About Artificial Neural Network from website http:// en.wikipedia.org/wiki/Artificial_neural_network.
[25] "A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm," God, Your Book Is Great !!, 18-May-2010.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY