A Hybrid Approach for Recommendation System based on Web Mining

Gurleen Kaur
CSE & IT Deptt.
BBSBEC
FatehgarhSahib,Punjab,India
gurleenkour1122@gmail.com

Amanjot Kaur
Assistant Professor, CSE & IT Deptt.
BBSBEC
Fatehgarh Sahib, Punjab, India
amanjot.kaur@bbsbec.ac.in

Abstract **-** The significant issue of many on-line sites is the introduction of numerous decisions to the customer at once; this for the most part brings about strenuous and tedious in finding the correct item or data on the site. In the traditional methodologies, KNN based classification strategies were utilized which depended on suggestion handle. These have some real issue if the information differs. The arrangement approaches that were utilized as a part of customary work are fit just if the data variation is inside the cluster that they have. However, in the event that the data goes out of bound it is hard to perform classification. In this way, there is a need to include a classifier approach that can work in such conditions. For this, a hybrid approach comprising of Multi-Layer ANN and k-NN is proposed in order to take proper choices if there should be an occurrence of data variation. The proposed idea introduces an intelligent approach which captures the clients going out of bound and adds them into the cluster, so that they can be recommended to the user and no client is skipped.

Keywords-*Web Mining, Multi-Layer Artificial Neural networks (ANN), K-Nearest Neighbor(k-NN)*

*****

I. INTRODUCTION

Data mining methods give users a new energy to look into and control the current vast volume of information. Data mining process finds fascinating data from the concealed information which can either be used for future forecast or potentially insightful outlining the subtle elements of the information [1,10].

Web mining innovation is developing field of data mining for WWW based data and assets. The fundamental center of web mining is to use data mining ways and calculations to remove valuable and concealed patterns from unstructured and tremendous web information or assets [2,11].

Recommendation system uses the selected items as response and recommends the user a number of items that have chosen other users. This problem can have two interpretations: 1) the problem of item's recommendation for the user who is already working with the advisory system.

2) The problem of recommendations of objects for the new user that implements the first login [3,12,13].

In this work second problem is there when system can watch clients/users route conducted by following up on the client's click-stream information on a RSS reader site, to prescribe a proper arrangement of items that fulfills the need of a dynamic client in a Real-Time, online premise. The RSS (Really Simple Syndication) reader site is an example of online recommendation system where users can able to read daily news online across the globe.

In many cases K-Nearest Neighbor classification technique was used as it is extremely efficient and dependable strategy to know client's conduct, behavior and interest at a specific session. But in this paper the hybrid of K-Nearest Neighbor (k-NN) with Multi-Layer Artificial Neural Network is done under which two ways are analyzed, got more exact outcomes and that helps in increasing the effectiveness of the framework. This helps to give exact information to the users for a specific information. The MatLab software was used to interpret and present graphical results.

II. RELATED WORK

Adeniyi *et al.* [4]presented a study of automatic web usage data mining and recommendation system based on current user behavior through his/her click stream data on the newly developed RSS reader website, in order to provide relevant information to the individual without explicitly asking for it. The k-NN strategy has been prepared to use on-line and in Real-Time to distinguish customers/guests click stream information, coordinated it to a specific client gathering and suggested a customized perusing alternative that address the issue of the particular client at a specific time. To accomplish this, web clients RSS address document was extracted, scrubbed, arranged and gathered into significant session and information data mart was created. The outcomes demonstrated that the k-NN classifier was straightforward, steady, direct and basic as contrasted and different methods.

Bellary*et al.* [5] discussed various machine learning approaches used in data mining. Further they distinguished between symbolic and sub-symbolic data mining methods. After that, a hybrid method with the combination of Artificial Neural Network (ANN) and Cased Based Reasoning (CBR) in mining of data was proposed.

Bloggers are one of the powerful instruments of web which are considered as one of the significant tool of social and intuitive capacities in making IT world awesome. Two strategies were utilized by Farhad *et al.* [6] i. e. k-NN and ANNs. These strategies are grouped in light of Kohkiloye and Boyer-Ahmad province bloggers dataset. Considering the k-

NN and ANN strategies to classify bloggers, it can be seen that arrangement comes about enhanced and getting 90% of accuracy than that of k-NN with Decision trees.

To learn complex functions, Renhou *et al.* [7] divided the sample input into a few subsets in which just a single yield extreme point exists by utilizing clustering technique. At that point multi-ANN was used to train the function. In the learning method each ANN prepares to approach a part of function with a subset of information tests. The results of all sub-ANNs assemble to get the entire arrangement. The multi-ANN model can approximate complex non-linear functions more adequately. The simulation results demonstrated that Multi-ANN connects to display and control complex dynamic frameworks. Contrasting and the single ANN strategy, the multi-ANN based technique has many merits, such as, higher learning speed, higher precision and more versatile for multivariate systems.

III. PROPOSED WORK

Literature review represents that many researchers done research on k-NN as it is one of the simplest methods for solving classification problems; however in the event if the data or information goes out of cluster, at that point that decision cannot be prescribed to the client. Although the classification approaches that used in traditional work are capable only if the data variation is within the cluster information that they have. Likewise the hybrid of Multi-Layer ANN with k-NN used as contrasting with Single-Layer ANN strategies the proposed approach enhances and adaptability greatly in learning processes of networks. Following is the description of proposed work:

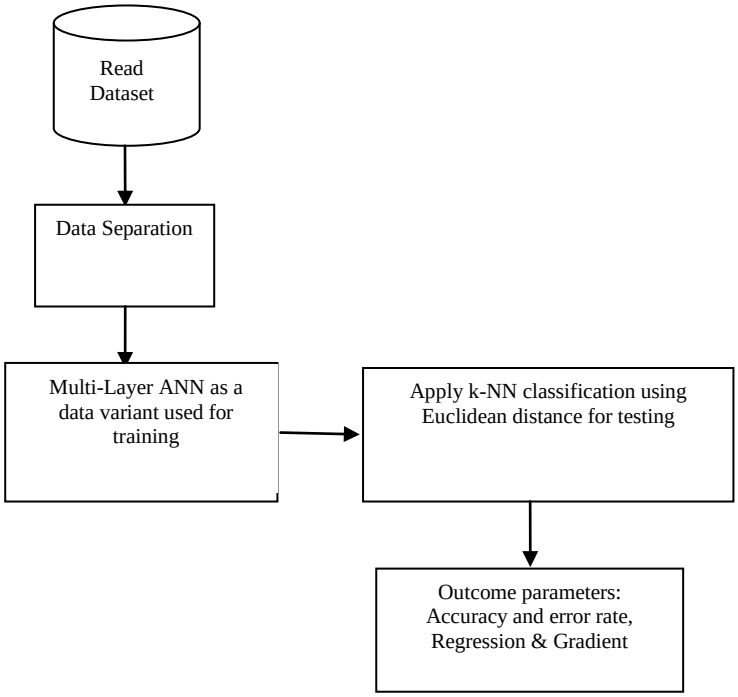

Figure.1. The Proposed Framework

3.1. Dataset Description:

Figure.2. shows RSS reader's dataset where 3 attributes in each record which consist of daily's name, news category and added required feed-type; the no. of clients/users is 13. This whole database is designed on MS-Excel sheet. Then it is fetched on MatLab tool by using the following method:

load Recomdataset [Data,Text]=xlsread('Dataset1.xlsx');

Users	Daily's Name	News Category	Added Required Feedtype
1	CNN News	World	www.*world
2	China Daily	Business	www.*business
3	Punch Ng	Politics	www.*politics
4	CNN news	Politics	www.*politics
5	Punch Ng	Entertainment	www.*entertainment
6	Thisday News	Politics	www.*politics
7	Vanguard News	Sports	www.*sports
8	Complete Football	Sports	www.*sports
9	Vanguard News	Politics	www.*politics
10	China Daily	Politics	www.*politics
11	Thisday News	World	www.*world
12	FOX	Sports	www.*sports
13	Sky Sports	Sports	www.*sports

Figure.2. RSS reader's dataset

3.2.Multi-Layer Artificial Neural networks

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system [15]. It includes two types of layer architecture:

a.) Single-layer ANN: By connecting various neurons, the genuine computing power of the neural networks comes, however even a single neuron can perform generous level of calculation [14]. The most widely recognized structure of connecting neurons into a system is by layers. The least complex type of layered system is appeared in Fig.3.

The input layer neurons are to just pass and distribute the inputs and perform no calculation. In this way, the main true layer of neurons is the one on the privilege [14]. Each of the inputs A_1, A_2, ...A_N is connected with each artificial neuron in the output layer through the connection weight. Since each estimation of output B_1, B_2, ..., B_N is figured from a similar arrangement of info esteems, each yield is fluctuated in light of the association weights.

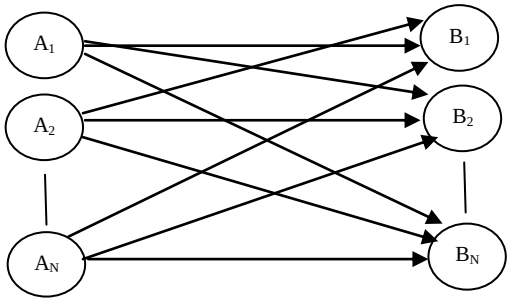

Figure.3. Single-Layer ANN

b.) Multi-Layer ANN: To accomplish more elevated amount of computational capabilities, a more complex structure of neural system is required. Fig. 4 demonstrates the multi-layer neural system which separates itself from the single-layer arrange by having one or more hidden layers [7,8].

In this multi-layer structure, the input nodes pass the data to the units in the first hidden layer, and then the outputs from the first hidden layer are passed to the next layer, and so on[8].

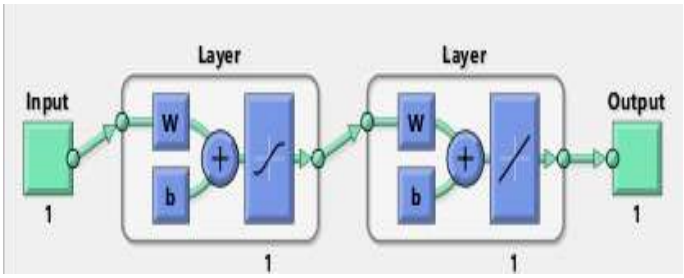

Figure.4. Multi-Layer ANN

In our work, Fig.4. architecture is used where 1 neuron in input layer passes the data to first hidden layer and the output of first hidden layer takes as an input for second hidden layer and the output of second hidden layer passes as an output of whole data.

Pseudo-code 1: Pseudo-Code of Multi-Layer ANN

1. Begin
2. Initialization Parameters: ANN object creation
inputs:{*11,0,4,6,0,4,8,7,4,4,1,7,7*}
targets:{1,2,3,4,5,6,7,8,9,10,11,12,13}
3. Net. Perform Fcn: Multi-Layer ANN method
Net = newff(minmax(inputs),[11],{'tansig','purelin'},'traingd','learngd');
4. Define Iterations For Training:
Net.trainParam.epochs = 100;
5. Train the Network
Net=train(Net,inputs,targets);
6. Compute performance of network & Simulate Results
Yf = sim(Net,inputs);
7. End

In other words, its objective is to find data variation using following method:

Net= newff(minmax(inputs),[11],{'tansig','purelin'},'traingd','learngd'); (3.1)

where newff: creates a feed-forward backpropagation network,
tansig: Tan-sigmoid transfer functions to calculate output,
purelin: Linear transfer function to calculate output,
traingd: Gradient Descent backpropagation(a training fn.)
learngd: Gradient descent weight and bias learning function

3.3.K-Nearest Neighbor (k-NN)

Many researchers have attempted to use K-Nearest Neighbor classifier for pattern recognition and classification in which a specific test tuple is compared with a set of training tuples that are similar to it [9]. The K-Nearest Neighbor algorithm was used alongside with five other classification methods to combine mining of web server logs and web contents for classifying user navigation pattern and predicts user future request [4].

The K-Nearest Neighbor classifier usually applies the Euclidean distance after classification process. In this research work, the Euclidean distance approach will be applied to verify and sort the clients/users [9]. To calculate the distance, we simply compare the corresponding values of the attributes of client 'a1' with that of 'a2'

a.) if the values are the same, then the difference is taken to be zero(0)
b.) otherwise the difference is taken to be one (1).

In our experiment, suppose our dataset have three attributes as Daily Name, Daily Type and News category and that W is a client with Daisy as username and 123 as password.

The Euclidean distance between a training tuple and a test tuple can be derived as follows:

Let Wi be an input tuple with p features ($w_{i1}, w_{i2}, \ldots, w_{ia}$)
Let n be the total number of input tuples (i =1,2,. . .,n)
Let a be the total number of features (j =1,2,. . .,a)
The Euclidean distance between Tuple Wi and Wt(t =1,2,. . .,n) can be defined as:

$$d(w_i, w_t)=\sqrt{(w_{i1}- w_{t1})^2 + (w_{i2}- w_{t2})^2 + \ldots..+ (w_{ia}- w_{ta})^2} \quad (3.2)$$

In general term, the Euclidean distance between two clients/users

W1 = ($w_{11}, w_{12}, \ldots, w_{1n}$) and W2 = ($w_{21}, w_{22}, \ldots, w_{2n}$) will be,

$$dist(w_1 , w_2)=\sqrt{(w_{1i} - w_{2i})} \quad (3.3)$$

IV. RESULTS

The experimental results are evaluated from the proposed framework in Fig.1 on RSS reader's dataset. The hardware requirements used by proposed system are 2.8 GHz processor, 8 GB RAM and 200 GB Hard disk and implementation tool is MATLAB R2014a. Table.2. and Fig.7&8.shows that the accuracy of the proposed system is much more as compared to traditional technique.

a.) Training measures of proposed system when number of iterations are 100:

Regression plot shows the network output with respect to targets for training
The gradient displays the slope of the tangent of the graph of the function.

Table.1. Training measures of proposed system are:

Training Parameters	Hybrid Multi-Layer ANN & k-NN
Regression	0.38461
Gradient	0.61803
Validation Checks	0

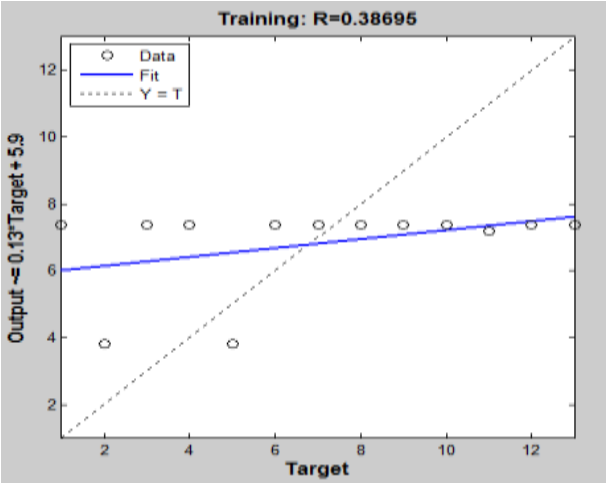

Figure.5. Regression

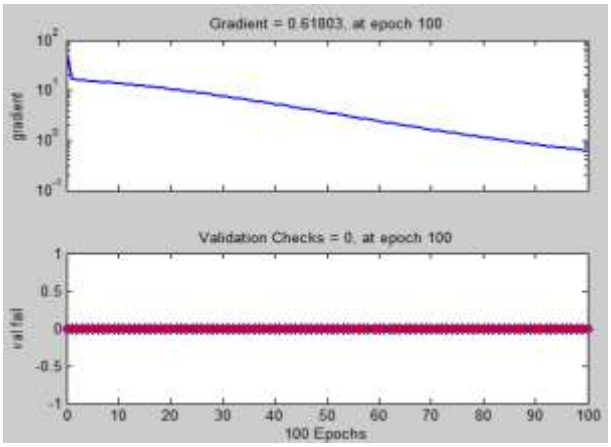

Figure.6. Gradient and Validation checks

b.) Performance measures of proposed system are:

Accuracy= (fR/Match)*100

wherefR = Final no. of recommendations,

Match= total no. of observed values corresponding to the inputs to the function which generated the recommendations.

Table.1. Performance measures of proposed system are:

Performance Parameters	Approaches	
	k-NN	Hybrid Multi-Layer ANN & k-NN
Accuracy	69.2308	92.3077
Error Rate	30.7692	7.6923

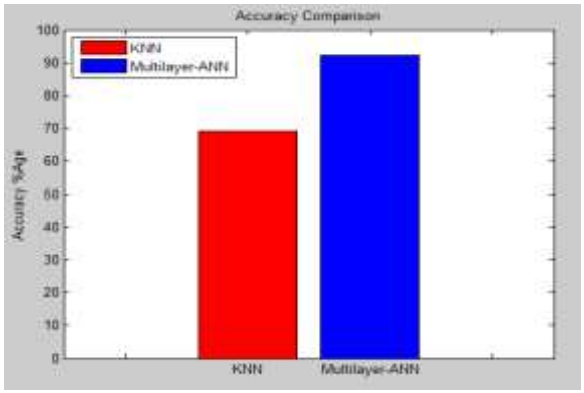

Fig.ure7. Accuracy Comparison

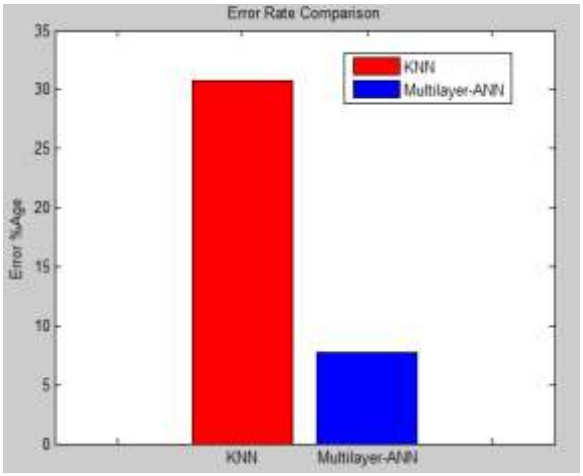

Figure.8. Error Rate Comparison

V. CONCLUSION

As presentation of information at a time according to the need of the specific user is important, so there will be an efficient approach to improve the performance. So we propose a hybrid approach comprising of k-NN and Multi-layer ANN to present proper choice to the user if data goes out of bound or cluster. The RSS data set from the traditional work used to evaluate the results. The Multi-Layer ANN technique used for training and k-NN for testing purpose. The results show that the comparison of hybrid approach to traditional approach concludes that this hybrid approach is better than traditional approach.

REFERENCES

[1] H. Jiawei, K. Micheline, "Data mining concept and Techniques", Second Ed., Morgan Kaufmann Publishers. pp. 285–350. , Elsevier. 2006.

[2] Dhandi, M., &Chakrawarti, R. K. "A comprehensive study of web usage mining". In *Colossal Data Analysis and Networking (CDAN)*, pp. 1-5. IEEE. 2016.

[3] Stekh, Y., Lobur, M., Artsibasov, V., &Chystyak, V, "Methods and tools for building recommender systems". In *Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), 13th International Conference*. pp. 300-305. IEEE. 2015

[4] Adeniyi, D. A., Wei, Z., &Yongquan, Y. "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method". *Applied Computing and Informatics*, Volume 12, Issue 1, pp. 90-108. 2016.

[5] Bellary, J., Peyakunta, B., &Konetigari, S. "Hybrid Machine Learning Approach In Data Mining". *Machine Learning and Computing (ICMLC), Second International Conference.* pp. 305-308. IEEE. 2010.

[6] Gharehchopogh, F. S., Khaze, S. R., &Maleki, "A new approach in Bloggers Classification with Hybrid of K-Nearest Neighbor and Artificial Neural Network algorithms. *Indian Journal of Science and Technology*, Volume 8, Issue 3, pp. 237-246. 2015.

[7] Renhou, L., &Feng, G, "Complex function approximation based on multi-ANN approach". *Journal of Systems*

Engineering and Electronics, Volume 6, Issue 2, pp. 22-31. 1995.

[8] Mehta, A. J., Mehta, H. A., Manjunath, T. C., &Ardil, C. "A Multi-Layer Artificial Neural Network architecture design for Load Forecasting in Power Systems". *International Journal of Applied Mathematics and Computer Sciences*, Volume 4, Issue 4, pp. 227-240. 2008.

[9] Vaarandi, R., &Pihelgas, M, "Logcluster - A Data Clustering and Pattern Mining algorithm for Event Logs". In *Network and Service Management (CNSM), 11th International Conference.*pp. 1-7. IEEE. 2015.

[10] Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A. V., &Rong, X, "Data mining for the internet of things: literature review and challenges". *International Journal of Distributed Sensor Networks*. 2016.

[11] Upadhyay, A. &Purswani, B. "Web usage mining has pattern discovery". *International Journal of Scientific and Research Publications*, Volume 3, Issue 2, pp. 1-4. 2013.

[12] Sumathi, C. P., Valli, R. P., &Santhanam, T "Automatic recommendation of web pages in web usage mining". *International Journal on Computer Science and Engineering*, Volume 2, Issue 9. pp. 3046-3052. IJCSE.2010.

[13] Zhao, X., &Ji, K, " Tourism e-commerce recommender system based on web data mining". In *Computer Science & Education (ICCSE), 8th International Conference.* pp. 1485-1488. IEEE. 2013

[14] Singh, V., &Lal, S. P, "Digit recognition using single layer neural network with principal component analysis". In *Computer Science and Engineering, 2014 Asia-Pacific World Congress,* pp. 1-7. IEEE. 2014

[15] Hüllermeier, E. "Fuzzy sets in machine learning and data mining". *Applied Soft Computing*, Volume 11, Issue 2, pp. 1493-1505. 2011.