Asian Journal of Information Technology 15 (21): 4150-4159, 2016
ISSN: 1682-3915

Support Vector Machine based Classification and Clustering for Identifying Unanimous Users

[1]J.S. Kanchana and [2]D. Sujatha
[1]Department of IT, Sivagangai, K.L.N. College of Engineering, 630612 Tamil Nadu, India
[2]Department of MCA, Anna University Regional Centre, BIT Campus, Trichy, India

Abstract: Social networks have become a rich and large repository of information about us as individuals. Due to the growth of social network usage, grouping of like-minded users for further processing is a major research issue. Some social networks even allow users to identify others based on their user interests and tags. Interest of the users can be found with few options in web known as tagging and rating. Rating is the method that finds the opinion of users about the items. Geographical location is becoming a major factor that influences the interest of users. This study describes the performance of employing unsupervised learning techniques such as Support Vector Machine (SVM) and naive bayes model for grouping of the unanimous users with and without location information. Location-based unanimous user identification provides better results in grouping the users. Naive bayes and SVM classifier results are compared through error rate and confusion matrix. The analysis proves that SVM provides better performance in terms of accuracy when compared with NaiveBayes classifier. The classified location based user data is clustered using self-organizing maps for better recommendations.

Key word: Index terms-classification, clustering, Naive Bayes (NB), Self-Organizing Map (SOM), Support Vector Machine (SVM), unanimous user identification

INTRODUCTION

The web is not only a space to search information, but also a tool to find many interesting things and it also enables users to share information (Adamic, 1999). Recommending people to each other with high performance is one of the major research areas. The tags are utilized to categorize the users. Rating is the other way of finding efficient resource through the web. Clustering the users with similar interest provides better results and directs to many applications such as friend suggestion, collaborative filtering, etc. (Jaffali *et al*., 2014). Tagging and rating are some of the facilities in the web. The tag is referred as people-powered metadata and it may be either descriptive or subjective (Smith, 2007). The tags are used to categorize the algorithm is used to check the location impact among the different users/people on the particular location. Classification and clustering are two different analytical approaches used in the web mining applications to identify the hidden patterns in data. Classification is a supervised learning technique that classifies data based on class label information. It is one of the important techniques used in web mining for efficient classification of the datasets.These classification techniques are used to build the models for future data trends prediction. There are various algorithms for classification such as NB classifiers and SVM based classifiers.

NB is a classification technique that is used to identify a target class. Based on the bnayes theorem, it computes the probabilities. Using these probabilities, it classifies the data. The resultant classifications are more accurate and effective and more sensitive to add new data to the dataset. Support vector machines are supervised learning techniqueswith associated algorithms that are used for data classification. Suppose a given set of training data, each data point belongs to one of the two classes. The goal of SVM is to identify the class of the new data point. In a given labeled class data, SVM outputs the optimal hyper planethat categorizes the new data points.

Clustering is an unsupervised learning techniquefor grouping the similar data items. During the intelligent grouping of the files, clustering process creates a more relevant set of search results. The clustering algorithms are used to predict the like minded users. These algorithms are used in the recommender systems to recommend new items based on the user's preferences. SOMis a very popular nonlinear, unsupervised and competitive learning algorithm used in the data

Corresponding Author: J.S. Kanchana, Department of IT, Sivagangai, K.L.N College of Engineering, 630612 Tamil Nadu, India

clustering applications. The objective of the SOM is to maximize the intracluster similarity and minimize the intercluster similarity, for projecting the results into the lower-dimensional space.

Literature work: This study illustrates the conventional research works related to the clustering and classification techniques. From the movie reviews, Liu *et al.* (2012) have obtained the user opinions by features extraction. To rate the particular resource, for instance a movie, sentiment classification was used to evaluate its rating. The rating and the comments about the movies were reviewed sentimentally. The opinions were classified based on the positive and negative sentiments and then, the rating and summarization were evaluated. Kanchana and Sujatha (2014) proposed a way for identifying the unanimous people by using their rating and signature. A user-signature was generated by establishing a set which represents resource tag and its rating. The signatures of various users were compared and the scale of unanimity was found. Then, a group of users who are like-minded was formed by applying the parameter based clustering with parametric group signature. Ting *et al.* (2011) has discussed about the NB classifier that allows independent usage of each attribute and also provides final results. So, it is more computationally efficient than any other classifiers. In his research, NB was employed for document classification.

Ahmed *et al.* (2014) explained NB model as the most effective learning algorithms for data mining and machine learning. For information retrieval, it has been used as a core classifier. Shieh and Liao (2012) has discussed about SOM which is a nonlinear unsupervised neural network mainly used for data clustering and visualization applications. SOM is a popular tool for data exploring. It can map the high dimensional data patterns to low dimensional space. Datta *et al.* (2015) proposed a scalable collaborative filtering framework for finding the similarity of the users, based on the rating of the target user and other users in a specific cluster. Clustering based recommendation has proposed to handle the scalability issue and speed up the recommendation system. Finley and Joachims (2005) presented an SVM algorithm for optimizing different clustering functions by training a clustering algorithm using the item-pair similarity measure. The algorithm was empirically evaluated for clustering the noun-phrase and news article.

Cai *et al.* (2010) proposed a model for capturing the bilateral role of the user interactions within a social network and formulated CF methods to enable recommendation of the people. A neighbor-based CF algorithm was developed to predict the likeliness of the users to contact other users. Terveen and Hill (2001) presented a framework for understanding the recommender systems and surveyed a number of different approaches. The algorithms combining multiple types of information were developed to compute the recommendations of the users. Nayak (2014) discussed about the challenges and solutions related to the two-way recommendation methods in the social networks such as online dating networks. Braak *et al.* (2009) demonstrated a novel approach for determining the interest of the users by dividing the training data into the user-based profile clusters. The increase in the prediction speed of the novel approach without loss in the accuracy has shown by the experimental results. He *et al.* (2012) explored the classification issues on the uncertain data and proposed an algorithm for building NB classifier. A validation set was used to search an appropriate value for the user-specified parameter. Better classification of the uncertain data was achieved when compared to the traditional NB classifier.

Farid *et al.* (2014) introduced independent hybrid mining algorithms including the Decision Tree (DT) and NB for solving the classification problems in data mining applications. The performance of the proposed algorithm was evaluated using the benchmark dataset obtained from the University of California, Irvine (UCI) repository. The most valuable training datasets were extracted automatically and the highly effective attributes were identified for describing the instances from the noisy complex training databases. Valle *et al.* (2012) presented an approach for predicting the performance of the sales agents in a call center, based on the NB classifier. The operational records were used to predict the productivity of the agents. Training and testing of the classifier were performed using a 10-fold cross validation. The future performance of the sales agents was predicted efficiently based on their operational activities. Lin *et al* (2014) proposed a novel similarity measure between the documents with respect to the feature. The effectiveness of the proposed measure was evaluated on the real-time datasets for the text classification and clustering problems. The performance of the novel measure was better than the other similarity measures.

Melin and Castillo (2014) reviewed the type-2 fuzzy logic applications to handle higher degree of uncertainty in the pattern recognition, classification and clustering. The type-2 fuzzy logic has achieved better performance than the type-1 fuzzy logic. Zhang *et al.* (2014) constructed an object-similarity graph from the clustering results and propagated the labels on the graph to fulfill the uniformity of the prediction over the graph. The label

propagation algorithm was used for solving the convex learning problem. Ismail *et al.* (2011) proposed the combination of SOM and Least Square SVM (LSSVM) for time-series forecasting. The proposed approach has outperformed the single LSSVM model and provided a promising alternative technique for the time-series forecasting. Goyal *et al.* (2015) introduced a novel approach for predicting the priority of the software bugs to find out the improvement in the performance of the classifier. The title attribute of the bugs was clustered for grouping the similar bugs together. Then, the clusters were classified and priority was assigned to the bugs based on the severity of the attributes.

MATERIALS AND METHODS

Classification and clustering approach: This study explains the proposed SVM based classification and clustering for identifying the unanimous users. Figure 1 shows the overall flow diagram of the proposed approach. The main stages of the proposed approach are:

- Pre-processing
- User signature generation
- NB classifier
- Confusion matrix identification
- SVM classifier
- F-Measure comparison
- SOM

Pre-processing: Data pre-processing is a technique of removing unwanted, missing and noisy data and transforming the data into the desired format. The data source contains information about the books, users and their rating for each book and the desired location of the user. This data source is subjected to the preliminary processing known as pre-processing. Users who do not rate any book are eliminated for the next level processing, as the signature for the user cannot be created. Also data are separated into many chunks of data, for the feasible further processing.

User signature generation: The next process is to improve the security of the pre-processed data source, through the user signature generation. The user signature is created for the unique identification of the characteristics of the user. A signature is created for the valid users with the help of the category identification (ID) and the rating given by the user. Using this signature, the rating based algorithm uses the average rating given by the user for each category for the efficient identification of the user's attraction towards the particular category of the resources. The signature is the text string including category ID and average rating:

$$\text{Signature (s)} = 0_i(\text{categy ID C (j)} \ .\text{AvRating r (j,m)}) \quad (1)$$

$$\text{Avg\_Rating r (j,m)} = 0_i\text{rate (i) m} \quad (2)$$

Where:

'm' = The no of ratings done by user

'm' and 'j' = Denotes the category

NB classifier: Based on the user signature, average rating given by the user for each category is identified. Then it is applied to the two classifiers such as NB and SVM classifier. NB classifiers are a group of simple probabilistic classifiers based on the bayes theorem with strong (Naive) independence assumptions between the features. The first set is represented by a vector $X = (x_1,..., x_n)$ representing some'n' features (independent variables), it assigns to this instance probabilities $(C_k|x_1, ..., x_n)$. The above equation is written as:

$$\text{Posterior} = \frac{\text{Prior} \times \text{Likehood}}{\text{Evidence}} \quad (3)$$

The NB classifier is supervised learning algorithm. It is a probabilistic classification method.

Deriving naivebayes classifier: The Bayes rule is defined as:

$$P(B/A) = \frac{P(B/A)\,P(B)}{P(A)} \quad (4)$$

Given two classes c_1, c_2 and the document d':

$$P(c1/d) = \frac{P(c1).P(d/c1)}{P(d)} \quad (5)$$

$$P(c2/d) = \frac{P(c2).P(d/c2)}{P(d)} \quad (6)$$

We are looking for a c_i that maximizes the posterior probability p (c_i|d). P (d) is the same in both cases. Thus:

$$\text{Camp} = \text{argams} \times P(c)\ P(d/c) \quad (7)$$

Estimating parameters for the target function: We are looking for the estimates P(c) and p(d/c), P(c) is the fraction of the possible true cases. This is expressed as:

$$P(c) = \frac{Nc}{N} \quad (8)$$

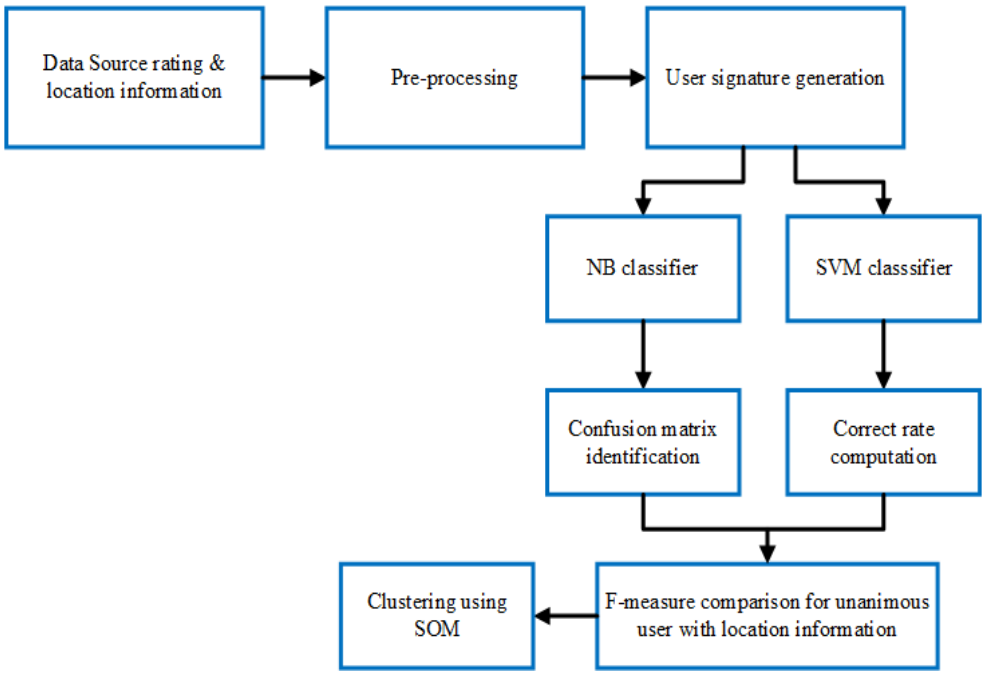

Fig.1: Overall flow diagram of the proposed approach

Where:
N = The number of all documents
N_c = The number of documents in the class
c = d-vector in the space X

$$P(d/c) = P\left(<t_1, t_2, t_3 ... tn | c\right) \quad (9)$$

By using the chain rule $P(A \hat{} B) = P(A/B)P(B)$, we have:

$$P\left(<t_1, t_2, t_3, ..., t_n > | c\right) = P\left(<t_1 / t_2, ..., t_n, c\right) \quad (10)$$
$$P\left(t_2, t_3, ..., t_n, c\right)$$

NB assumptions of independence:
- The values of all attributes are independent of each other
- The conditional probabilities are the same, irrespective of the position in the document
- We assume the document is a "bag-of-words"

The conditional probability is given as:

$$P\left(d/c\right) = P\left(<t_1, t_2, t_3, ..., t_n > | c\right) = \pi\, P\left(t_k | c\right) \quad (11)$$

The target function is:

$$Cmap = \arg\max P(c/d) = \arg\max P(c)\pi\, P\left(t_k | c\right) \quad (12)$$

NB estimation: For each term 't', P (t|c) is estimated:

$$P(t/c) = \frac{Tct}{\sum t' \varepsilon V Tct} \quad (13)$$

The T_{ct} is the count of the term 't' in all documents of the class 'c'. The estimate will be '0', if a term does not appear with a class in the training data, then smoothing is required.

Confusion matrix identification: The results of the NB classifier are then arranged in the form of confusion matrix. Confusion matrix is a specific tabular structure that allows evaluation of the performance of the classification algorithm. Each column of the matrix represents the instances in a predicted class and each row represents the instances in an actual class.

SVM classifier: SVM is an optimally defined surface.Itis linear or nonlinear in the input space. Linear in a higher dimensional feature space.SVM is implicitly defined by a kernel function:

$$K(X, Y) \rightarrow Z \quad (14)$$

They are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis.

Given a set of training examples, such that each example belongs to one of the two categories. An SVM training algorithm is a non-probabilistic binary linear classifier for building a model that assigns new examples into the respective category.The linear SVM process is as below.Given some training data 'D', a set of 'n' points of the form:

$$D = \{(Xi, Yi) | xi \in Rp, Yi \in \{-1, 1\}\} \quad (15)$$

Where:

i = 1-n

y_i = Either

1 or-1 = The class to whichthepointbelongs

Each X_i is a P-dimensional real vector. The maximum-marginhyper plane that divides the points having $y_I = 1$ from those having $y_i = -1$ is found out. Any hyper plane is written as the set of points X satisfying:

$$W.X - B = 0 \quad (16)$$

where, . denotesthe dotproduct.W is the normal vector tothehyper plane.The parameter B/‖w‖ determines the offset of the hyper plane from the origin along the normal vector 'W'.If the training data are linearly separable, two hyper planesare selected in a way that they separate the data and there are no points between them and then try to maximize their distance. The region bounded by them is called "the margin". These hyper planes can be described by the equations:

$$W \cdot X - B = 1 \quad (17)$$

and:

$$W \cdot X - B = -1 \quad (18)$$

Geometrically, the distance between these two hyper planes is 2/‖w‖, so to maximize the distance between the hyper planesand minimize ‖W‖. As falling of the data points into the margin is prevented, the following constraint is added for each 'i' either:

$$W \cdot X_i - B >= 1 \text{ for Xi of the first class} \quad (19)$$

$$W \cdot X_i - B < 1 \text{ FOR Xi of the second class} \quad (20)$$

This can be rewritten as: $Y_i(W \cdot X) >= 1$ for all 1<i<n. This is put together to get the optimization problem: Minimize (in W, B) ‖W‖ subject to (for any i = 1-n):

$$Y_i(W \cdot X) >= 1 \quad (21)$$

The support machine applied data is then applied with correct data computation. F-measure comparison for unanimous users with location information. The traditional F-measure or balanced F-score (F_1 score) is the harmonic mean of precision and recall:

$$F\_Measure = \frac{Precision \times Recall}{Precision + Recall} \quad (22)$$

The general formula for the real positive β is:

$$F\beta = (1+\beta 2) = \frac{Precision \times Recall}{(\beta 2\ precision) + Recall} \quad (23)$$

The formula in terms of Type 1 and type 2 errors:

$$F\beta = \frac{(1+\beta 2).TP}{(1+\beta 2).TP + \beta 2.FN + FP} \quad (24)$$

Precision (P): Fraction of retrieved docs that are relevant = P (relevant|retrieved):

$$P = \frac{TP}{TP + FN} \quad (25)$$

Recall (R): Fraction of relevant docs that are retrieved = P(retrieved|relevant):

$$R = \frac{TP}{TP + FP} \quad (26)$$

Where:

'TP' = True Positive value, TN denotes True Negative value

FP = False Positive value and FN denotes False Negative value

Unanimous user identification procedure:

Input: DATASET (D) Corpus with User ratings of books, user location

Output: Performance evaluation of SVM and NB Classifier

Step 1: Extract Location Based rating datasets from the Book review corpus for SVM classifier

Step 2: Extract Book based rating datasets from the Book review corpus for NaiveBayes classifier

Step 3: Setting Threshold values for rating to calculate the performance metrics

Step 4: Setting of T=0, 0>T<5 and T>5 for count the TP, TN,FP,FN values

Step 5: For each

TP = Count Positive Comments in which User Location when T>5

FN = Count Negative Comments in which User Location when T=0

FP= Count Positive Comments in which User Location when T>0 and T<=5

End for

Step 6 :For each

Calculate Precision, recall and F-measure values

End for

After this F-Measure comparison, For further optimization self organizing maps are used.

SOM: SOMs are mainly used for data visualization. To construct a SOM, first step is initialization of the weight vectors. Then, a sample vector is selected randomly and the map of the weight vectors is searched, to find the optimal weight vector that provides best representation of the sample. The neighboring vectors are located proximate to each other. The chosen weight vector is rewarded to become more similar to the randomly selected sample vector. The neighbors of the chosen weight are also rewarded to become more similar to the chosen sample vector. The whole process is repeated for a large number of times by default it is repeated >1000 times.

RESULTS AND DISCUSSION

Performance analysis: The data corpus consists of 3 table such as user information table, books information table and rating with location information table. The user information table comprises of 2,00,000 users. The books information table consists of 2,00,000 books and the rating with location information table consists of 10,54,552 entries. After preprocessing the books information other than the category and ISBN are eliminated, since they are not used for current processing. Like this, after generating the location factor, users who do not possess any location factor are also eliminated, since they cannot be compared with other users. To verify the effectiveness of the algorithm initial location based analysis was done with only 22,000 rating entries and clusters were created for few 1000's of users. Also few users remain unrated. If 200,000 entries are considered, more users are clustered and so it becomes complex to find highly like-minded users. The performance of SVM and naivebayes are compared in terms of confusion matrix and correct data rate computation for unanimous users with location information and without location information. The unanimous users with location information provide better performance by the computation of confusion matrix and correct data rate. In order to improve the accuracy, unanimous users with location information are further applied to naivebayes and SVM classifier. The experimental results shows better performance in SVM classifier for location based unanimous users in terms of F-Measure computation. Table 1 gives the comparison with and without location based information for the two classifiers. It can be deduced from the table that the confusion matrix results are better for the data set without location information. As location information is added the NB classifier is not able to report proper classification. The SVM correct rate is substantially better compared to NB in both the cases. Table 1 shows the comparative analysis of the NB and SVM classification based on the location information. Table 2 shows the F-measure calculation using precision and recall values. Table 3 shows the comparative analysis of F-measure for NB and SVM. Figure 2 shows the comparison of the NB and SVM classifier for the location based unanimous users.

Table 1: Comparative analysis of NB and SVM classification based on location information

Method	NaiveBayes confusion matrix	SVM correct rate
Clustering of unanimous u without location information	cMat1=	0.8200
	7 0 0 0	
	1 27 1 0	
	0 0 43 1	
	0 0 0 17	
Clustering of unanimous users with location information	cMat1 =	0.9000
	8 1 2 7 1	
	1 15 1 6 1	
	1 7 10 7 0	
	1 6 1 12 1	
	1 2 3 4 7	

Table 2: Calculation of F-measure using precision and recall values

User location	Naive bayesclassifier			SVM classifier		
	Precision	Recall	F-measure	Precision	Recall	F-measure
CA	0.65	0.7	0.67	0.96	0.67	0.79
EA	0.76	0.76	0.76	0.94	0.97	0.96
FL	0.41	0.44	0.42	0.67	0.61	0.64
NY	0.65	0.7	0.67	0.96	0.54	0.69
TX	0.65	0.72	0.68	0.96	0.67	0.78

SOM hits: Plotsomhits plots a SOM layer with the hexagonal-shaped default SOM topology. Each neuron in the SOM shows the number of the classified input vectors. The relative number of the vectors for each neuron is shown based on the size of a colored patch. Figure 3 shows the SOM clusters and Fig. 4. shows the SOM hits. The location of the neurons in the topology and the number of training data associated with each neuron are shown in Fig. 3 and 4. There are 100 neurons in each topology. The maximum number of hits associated with any neuron is 22. Thus, there are 22 input vectors in that cluster.

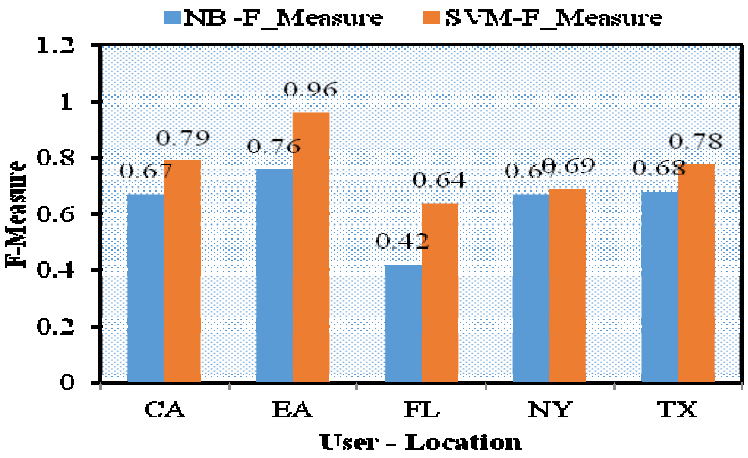

Fig. 2: Comparison of NB and SVM classifier for location based unanimous users

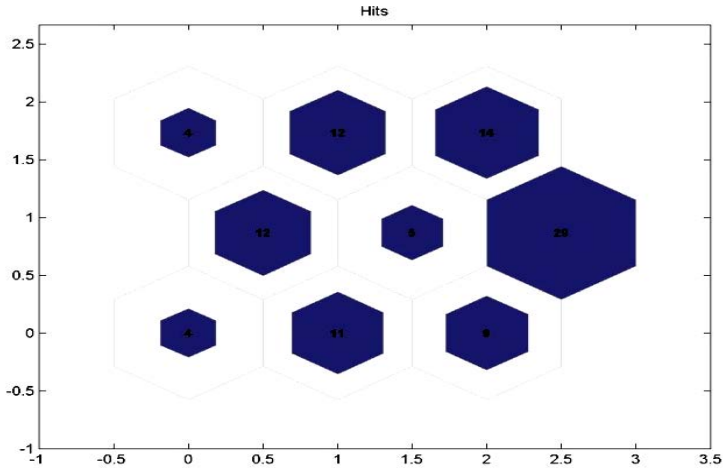

Fig. 3: SOM clusters

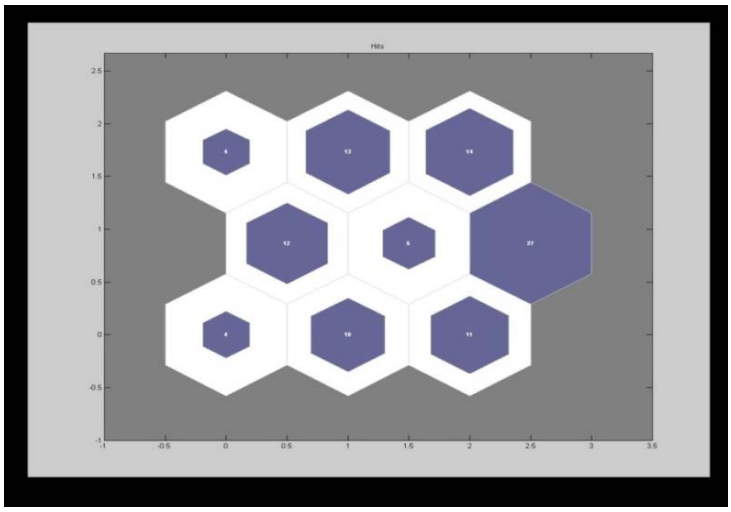

Fig. 4: SOM hits

SOM neighbor weight distances: Figure 5 shows the SOM neighbor weight distances. The blue colored hexagons in Fig. 5 denote the neurons and the red-colored regions connect the neighboring neurons. The distance between the neurons is indicated by the different colors along with the red lines. The darker colors represent the larger distances between the neurons and lighter colors represent the smaller distances.

Table 3: Comparative analysis of F-measure for NB and SVM

User Location	NB-F_Measure	SVM-F_Measure
CA	0.67	0.79
EA	0.76	0.96
FL	0.42	0.64
NY	0.67	0.69
TX	0.68	0.78

SOM weight plane: The visualization of the weights that connect each input vector to each neuron is shown in Fig. 6 Larger weights are represented using the darker colors. It is assumed that the input vectors are highly correlated with each other, if the connection patterns of the two input vectors are more similar. In such case input vector 1 has different connections than the input vector

SOM plane weight position: Figure 7 shows the SOM plane weight position. The input vectors are plotted as green dots. The classification of the input space by the SOM is depicted in Fig. 7. The blue-gray dots are used to represent the weight vector for each neuron and the red lines are used for connecting the neighboring neurons.

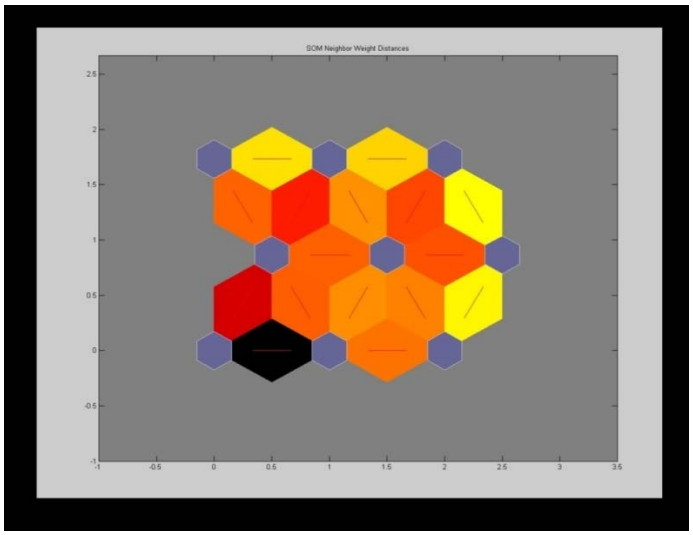

Fig. 5: SOM neighbor weight distances

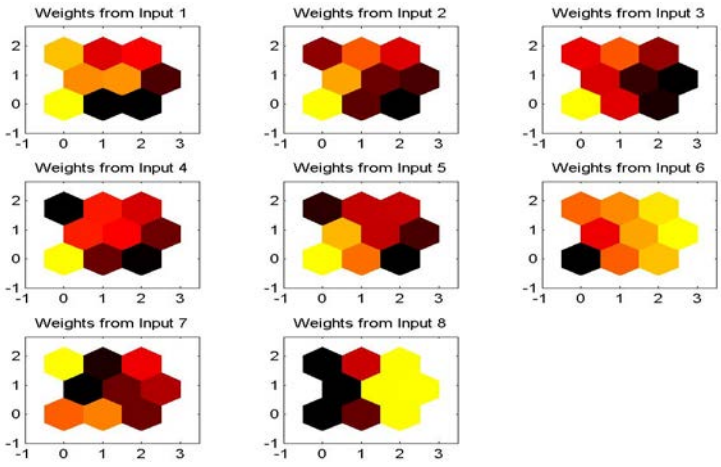

Fig. 6: SOM plane with weights

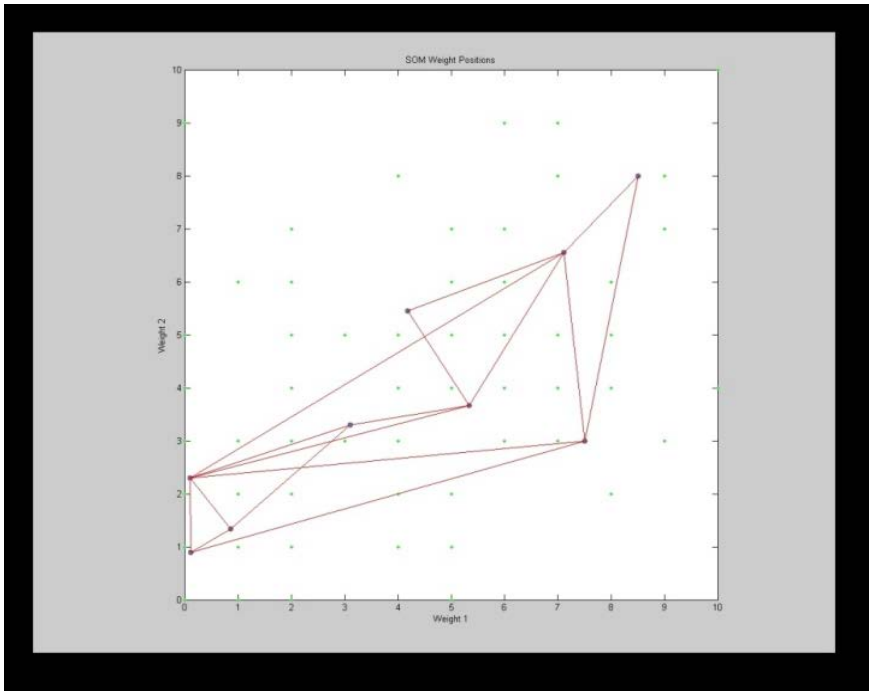

Fig. 7: SOM plane weight positions

CONCLUSION

The perception of the web has increased due to the introduction of new social platform which are in need of methods and tools to support user's and search for other user groups which communicates their own interests. The advent of social networks, web user groups and other user groups has changed the ways of sharing information among the users. In this study, SVM classifier plays a major role for grouping of unanimous users with the location information. Rating is a process that attracts the users to estimate the content in the web. Users, those who are involved in rating can rate the resources available in the web, based on their perceptions. Users can represent their interests using rating. This has created an opportunity for location factor to represent the users based on rating, since the web is a place where people search for people who have similar interests. To find the location impact, SVM and NaiveBayes classifier are used and the performance parameters such as confusion matrix and correct data rate are used to evaluate the methods. After finding the location impact in order to improve the performance of location based unanimous users NaiveBayes and SVM classifier results are compared in terms of F-measure, The experimental results shows SVM shows better performance in terms of accuracy when compared with NaiveBayes classifier. Further, the data are clustered using SOM for improved recommendations. Self-organizing map is a data clustering and a visualization technique which is used to visualize the location based unanimous user's clusters. The influence of the location can be applied to various recommendation systems in order to improve their efficiency of recommendations.

REFERENCES

Adamic, L.A., 1999. The Small World Web. In: Theory and Practice of Digital Libraries. Abiteboul, S. and A.M. Vercoustre (Eds.). Springer Berlin Heidelberg, Berlin, Germany, ISBN: 978-3-540-66558-8, pp: 443-452.

Ahmed, I., D. Guan and T.C. Chung, 2014. SMS classification based on naive bayes classifier and apriori algorithm frequent itemset. Int. J. Mach. Learn. Comput., 4: 183-187.

Braak, P.T., N. Abdullah and Y. Xu, 2009. Improving the performance of collaborative filtering recommender systems through user profile clustering. Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, September 15-18, 2009, IEEE Computer Society, Washington, DC, USA., ISBN: 978-0-7695-3801-3, pp: 147-150.

Cai, X., M. Bain, A. Krzywicki, W. Wobcke and Y.S. Kim *et al*., 2010. Collaborative Filtering for People to People Recommendation in Social Networks. In: Advances in Artificial Intelligence. Jiuyong, L. (Ed.). Springer Berlin Heidelberg, Berlin, Germany, ISBN: 978-3-642-17431-5, pp: 476-485.

Datta, S., J. Das, P. Gupta and S. Majumder, 2015. SCARS: A scalable context-aware recommendation system. Proceedings of the 2015 Third International Conference on Computer, Communication, Control and Information Technology (C3IT), February 7-8, 2015, IEEE, Hooghly, India, ISBN: 978-1-4799-4446-0, pp: 1-6.

Farid, D.M., L. Zhang, C.M. Rahman, M.A. Hossain and R. Strachan, 2014. Hybrid decision tree and naive bayes classifiers for multi-class classification tasks. Expert Syst. Appl., 41: 1937-1946.

Finley, T. and T. Joachims, 2005. Supervised clustering with support vector machines. Proceedings of the 22nd International Conference on Machine Learning, August 7-11, 2005, ACM, New York, USA, ISBN: 1-59593-180-5, pp: 217-224.

Goyal, N., N. Aggarwal and M. Dutta, 2015. A Novel Way of Assigning Software Bug Priority Using Supervised Classification on Clustered Bugs Data. In: Advances in Intelligent Informatics. Sayed, E.M.E.A. S.M. Thampi, T. Hideyuki, S. Piramuthu and T. Hanne (Eds.). Springer International Publishing, Berlin, Germany, ISBN: 978-3-319-11217-6, pp: 493-501.

He, J., Y. Zhang, X. Li and P. Shi, 2012. Learning naive bayes classifiers from positive and unlabelled examples with uncertainty. Int. J. Syst. Sci., 43: 1805-1825.

Ismail, S., A. Shabri and R. Samsudin, 2011. A hybrid model of Self-Organizing Maps (SOM) and Least Square Support Vector Machine (LSSVM) for time-series forecasting. Exp. Syst. Appl., 38: 10574-10578.

Jaffali, S., S. Jamoussi, A.B. Hamadou and K. Smaili, 2014. Clustering and classification of like-minded people from their Tweets. Proceedings of the 2014 IEEE International Conference on Data Mining Workshop, December 14, 2014, IEEE, Shenzhen, China, ISBN: 978-1-4799-4275-6, pp: 921-927.

Kanchana, J.S. and S. Sujatha, 2014. Clustering unanimous web users based on rating and user-signature. J. Emerg. Technol. Web Intell., 6: 359-363.

Lin, Y.S., J.Y. Jiang and S.J. Lee, 2014. A similarity measure for text classification and clustering. IEEE. Trans. Knowl. Data Eng., 26: 1575-1590.

Liu, C.L., W.H. Hsaio, C.H. Lee, G.C. Lu and E. Jou, 2012. Movie rating and review summarization in mobile environment. IEEE. Trans. Syst. Man Cybern. Part C. Appl. Rev., 42: 397-407.

Melin, P. and O. Castillo, 2014. A review on type-2 fuzzy logic applications in clustering, classification and pattern recognition. Appl. Soft Comput., 21: 568-577.

Nayak, R., 2014. Two-way recommendation methods for social networks. Proceedings of the 7th Workshop on Ph.D Students, November 3-7, 2014, ACM, Shanghai, China, ISBN: 978-1-4503-1481-7, pp: 33-34.

Shieh, S.L. and I.E. Liao, 2012. A new approach for data clustering and visualization using self-organizing maps. Expert Syst. Appl., 39: 11924-11933.

Smith, G., 2007. Tagging: People-Powered Metadata for the Social Web. New Riders, San Francisco, California, ISBN: 978-0-321-52917-6, Pages: 203.

Terveen, L. and W. Hill, 2001. Beyond recommender systems: Helping people help each other. HCI. New Millennium, 1: 487-509.

Ting, S.L., W.H. Ip and A.H. Tsang, 2011. Is naive bayes a good classifier for document classification?. Int. J. Software Eng. Its Appl., 5: 37-46.

Valle, M.A., S. Varas and G.A. Ruz, 2012. Job performance prediction in a call center using a naive Bayes classifier. Expert Syst. Appl., 39: 9939-9945.

Zhang, X.Y., P. Yang, Y.M. Zhang, K. Huang and C.L. Liu, 2014. Combination of classification and clustering results with label propagation. IEEE. Signal Process. Lett., 21: 610-614.