Selective Colligation and Selective Scrambling for Privacy Preservation in Data Mining

Ishwarya M.V[1], K.Ramesh Kumar[2]
[1] Hindusthan Institute of Technology and Science, Padur
[1] CSE Dept, Sri SaiRam Engineering College Tambaram, Chennai, TamilNadu, India
[2] Department of Information Technology, Hindusthan Institute of Technology and Science, Padur

Article Info

Article history:

Received Nov 25, 2017
Revised Jan 12, 2018
Accepted Feb 21, 2018

Keywords:

ANN
BayesianAlgorithm
Data Mining
PrivacyPreservation
SVM

ABSTRACT

The work is to enhance the time efficiency in retrieving the data from enormous bank database. The major drawback in retrieving data from large databases is time delay. This time hindrance is owed as the already existing method (SVM), Abstract Data Type (ADT) tree pursues some elongated Sequential steps. These techniques takes additional size and with a reduction of speed in training and testing. Another major negative aspect of these techniques is its Algorithmic complexity. The classification algorithms have five categories. They are ID3, k-nearest neighbour, Decision tree, ANN, and Naïve Bayes algorithm. To triumph over the drawbacks in SVM techniques, we worn a technique called Naïve Bayes Classification (NBC) Algorithm that works in parallel manner rather than sequential manner. For further enhancement we commenced a Naïve Bayes updatable algorithm which is the advanced version of Naïve Bayes classification algorithm. Thus the proposed technique Naïve bayes algorithm ensures that miner can mine more efficiently from the enormous database.

Corresponding Author:

Ishwarya M.V,
Hindusthan Institute of Technology and Science, Padur,
CSE Dept, Sri SaiRam Engineering College,
Tambaram, Chennai, TamilNadu, India
Email: ishwarya.cse@sairam.edu.in

1. INTRODUCTION

In the existing system, it works in a chronological progression. It will ensure only one process at a time either comparison or classification and extraction or prediction. So it takes additional time to relay the regression.The foremost phase of Support Vector Machine algorithm is to classify the relevant data and irrelevant data from the unstructured massive database. Then, it compares the given query with the relevant data. After carrying out the comparison procedure then it tends to precede the extraction process. In the extraction progression, it retrieves the requested data which is given by the user. Subsequently, it proceeds to complete the prediction process by rating. It takes additional size and additional time to compute. As in the existing system the process would be in sequential order, leads to increase in time and increase in memory.

Yang et al.[1] expressed design mechanisms, when given a preference profile submitted by a user that search a person with matching profile in decentralized multi-hop mobile social networks. The mechanisms are privacy-preserving: no participants 'profile and the submitted preference profile are exposed. Goga et al.[2] explored the reliably match profiles in practical knowledge, across real-world social networks, by exploiting public attributes, publicly provide about themselves. It also defined a set of properties for profile attributes–Availability, Consistency, non-Impressionability, and Discriminability (ACID)–that are both necessary and sufficient to determine the reliability of a matching scheme. Sun et al. [3] composed area based informal organization administrations (LBSNS) have been explored; this review constructs a model to analyze the security math, advantage structure, and sex contrasts.

Prakash et al. [4] implemented an approximation automated structure, called Filtered Wall (FW) and it filtered disposed of substance from OSN client substances. The goal is to utilize efficient classification procedure to stay away from overpowered by unsuccessful messages. In OSNs, content filtering can also be abused for a unique, more reactive. In [5] explained integration of Adaptive Weight Ranking Policy (AWRP) with intelligent classifiers (NB-AWRP-DA and J48-AWRP-DA) via dynamic aging factor to improve classifiers power of prediction. The methods are used to choose the best subset of features. In [6] introduced a new framework called Fuzzy based contextual recommendation system for classification of customer reviews. It extracts the information from the reviews based on the context given by users. In [7] studied to identify the best classifiers for class imbalanced health datasets through a cost-based comparison of classifier performance. The unequal misclassification costs were represented in a cost matrix, and cost-benefit.

Dhivakar et al [8] elaborated recent approaches which are involved in privacy preservation like a randomization, Anonymization, perturbation and distributed privacy preservation methods. Janbandhu et al [9] expressed privacy preserving in data mining of many techniques along with their advantages and disadvantages. It also discussed about present limitations and scope for future research in privacy preserving data mining. Patel et al [10] introduced a certain transformation approach to deal with the privacy during mining. This approach main objective is to provide more accuracy of specific data and preserving privacy of original data.

To overcome these limitations we will introduce an algorithm called Naïve Bayes classification algorithm. This algorithm does the above process in a parallel manner. Thus the proposed technique Naïve bayes algorithm ensures that miner can mine more efficiently from the enormous database.Naïve Bayes Classification algorithm is used to perk up the time efficiency. It has worked quite well in many intricate real-world circumstances. Naïvebayes classification algorithm characterizes a lot of learning algorithms. Naive Bayes is an keen fast learning classifier. Thus, it could be used for making predictions in real time. It is easy to build and predominantly positive for very bulky classy classification methods.Naïve Bayes classifiers in learning problems requires lot of constraints linear in lot of variables. Naïve Bayes is represented in terms of probabilities. These probabilities are collected to form a file. For a learned naivesbayes model these files were utilized.Finally naïve bayes algorithm is easy to implement and it works in more beneficial way. It is preferred to choose this naïve bayes algorithm rather than other classification algorithms. This method is popularly known as “punching bag” for smarter algorithms.

2. RESEARCH METHOD

In bank database management system, we are going to apply Naïve Bayes classification algorithm especially in loan sector. If the person applies for a particular loan in a bank, the bank management checks the previous history of the person. Whether the person paid the previous loan balance or not and whether the person is able to pay the current loan based on the property of the person. The customer should provide the proper reason for acquiring the loan and then they should satisfy the loan criteria and followed by this, the loan will be afforded.If we apply naïve bayes algorithm in bank database the prediction will be accurate. The major use of naïve bayes algorithm in database management system is to increase the time efficiency because the NBC algorithm follows parallel processing. We are going to implement a tool called weka to run arff format of the bank database. The output of our project shows the retrieval time, true positive and false positive rate. The major use of the algorithm is to increase the time efficiency and accurate prediction of loan sector in bank database using naïve bayes classification algorithm. Using naïve bayes algorithm we can reduce the time consumption in large sectors. Figure 1 show the system architecture with data processing step-wise.

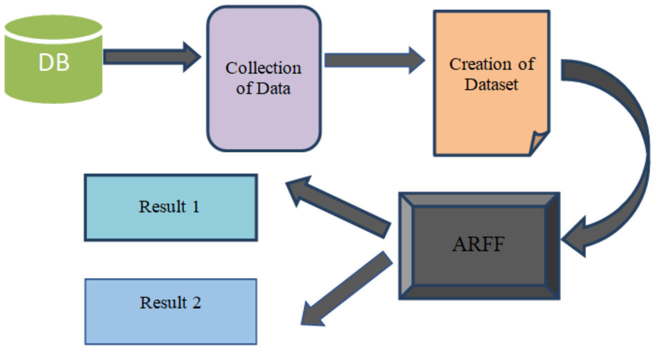


Figure 1. System Architecture Diagram

2.1. Proposed Technique

Let PT be the private table containing attributes A1,…, An where A1 is the first attribute and An is the last attribute. Let Ai,….,Aj be the set of quasi identifiers of PT such that (Ai,….,Aj) ⊆ (A1,…,An).Let the total number of tuples in PT be denoted as r .Hence let t1,…,trrepresent the tuples of PT. The algorithm is as follows:

a. Select the quasi identifier with the highest number of unique values say Am such that Am⊆ Ai,….,Aj.
b. Perform selective Colligation on Am as described in points 1 to 2.
   1) Let G1,…,Gnbegroups such that tuples in each group have same valueof Am. The tuples not in any group of G1,….,Gn are generalized.
   2) For the tuples in G1,….,Gn we consider the remaining quasi identifiers of Ai,….,Aj. For each group in G1 to Gn repeat step 2.2.1 For c in 1 to n in 2.2.1:
      a) For each tuple in Gc repeat steps 2.3.1.1 to 2.3.1.2.

2.2.1.1. Fora tuple ensure that it has at least one more tuple in the same group which should have all the quasi identifier values (Ai,….,Aj) same as it. If so go to step 2.2.1.Else go to step
2.2.1.2. Generalize the tuple.

3. For each generalized tuple in PT repeat step 3.1 .

3.1. Select tuples which have unique quasi identifier set Ai,….,Aj.

4. Slice PT such that each sliced table contains highly correlated values. Let the sliced tables of PT be B1,….,Bk, such that k is the total number of sliced tables.

5. In the sliced tables select a table Bh in B1,….,Bk such that it has at least one quasi identifier.

6. Perform selective shuffling on the selected table Bh. This is done by shuffling the tuples selected in step 3.

2.2 Selective Colligation

Based on the above algorithm we perform selective Colligation to our table to show how it works. The selected quasi identifier (say in our table age) to generalize we perform selective Colligation. Firstly we try to identify the tuples that have the same age value. In the following table the same colored tuples have same age value.

Now the tuples in black color are unique tuples, each having unique age values. So, such tuples cannot be evicted from Colligation. Considering grouped tuples we first check their remaining quasi identifiers (sex, Bill Amount, Address).As per the proposed algorithm in a given group (same color) forevery Tuple in a group ensure that it has at least one more tuple in the same group which should have all the quasi identifier values same as it. For example considering red group tuples we can sXX that the tuples ZZZ and VVV have same quasi identifier values (23, M, 16000, ZZ) and the tuples UUU and WWW have same quasi identifier values (23, F, 20000, TT), so we nXXd not generalize it as it can’t be identified because of its commonness in all quasi identifier values with at least one more tuple. Considering the yellow group tuples, tuples XXX and QQQ have same quasi identifier values (27, F, 26000, XX), which nXXd not be generalized but the tuple ZZZ having different quasi identifier values (27, M, 31000, ZZ) from XXX and QQQ, nXXd to be generalized. Considering the grXXn group tuples since both of them have different values for the quasi identifier “Address” we generalize them. Table 1 expresses the selective colligation for patient.

Table 1: Sample Patient Dataset for Selective Colligation

Name	Age	Sex	Bill Amount	No of check ups	Address	Criticality rate of Disease(Out of 10)
ZZZ	23	M	16000	2	ZZ	7
YYY	35	M	20000	2	YY	5
XXX	27	F	26000	2	XX	9
WWW	31	M	20000	2	YY	6
ZZZ	27	M	31000	2	ZZ	10
VVV	23	M	16000	1	ZZ	8
XXX	30	M	20000	1	YY	8
UUU	23	F	20000	1	TT	7
TTT	35	M	20000	3	YY	7
QQQ	27	F	26000	2	XX	9
WWW	23	F	20000	3	TT	7
RRR	29	M	35000	1	ZZ	8
SSS	33	M	31000	2	ZZ	8

Table 2: Sample Patient Dataset for Colligation

Name	Age	Sex	Bill Amount	No of check ups	Address	Criticality rate of Disease(Out of 10)
ZZZ	23	M	16000	2	ZZ	7
YYY	30-40	M	20000	2	YY	5
XXX	27	F	26000	2	XX	9
WWW	30-40	M	20000	2	YY	6
ZZZ	20-30	M	31000	2	ZZ	10
VVV	23	M	16000	1	ZZ	8
XXX	30-40	M	20000	1	YY	8
UUU	23	F	20000	1	TT	7
TTT	30-40	M	20000	3	YY	7
QQQ	27	F	26000	2	XX	9
WWW	23	F	20000	3	TT	7
RRR	20-30	M	35000	1	ZZ	8
SSS	30-40	M	31000	2	ZZ	8

2.3 Scrambling and Selective Colligation

In the above Table 2 after performing selective Colligation, we can sXX that some generalized tuples still have unique quasi identifier set which is a threat to privacy. For example tuples like ZZZ (yellow group) and RRR both have age in the range 20-30, but they differ in the quasi identifier Bill Amount which makes them unique and hence identifiable. Similarly SSS also differs in both Bill Amount and location with the similar ranged tuples YYY and WWW. So before slicing we select such tuples as per the algorithm as in table 3. After selection we slice the table using one of the existing slicing algorithms that has the best time efficiency and it shown in table 4. In the sliced tables we select any table as per our wish (with the constraint that it should have at least one quasi identifier) and shuffle the tuples that we selected before slicing process. By doing selective shuffling we have eliminated the possibility of privacy brXXch to certain records that the possibility of being identified (eg records like SSS, RRR) even after the Colligation process. Moreover selective Colligation consumes less time as compared to full Colligation as no existing shuffling algorithm can guarantXX a time efficiency of O(1) and hence the time efficiency of shuffling process depends on input size. Table 3 explains the tuples to be shuffled after applying proposed algorithm.

Table 3: Selection of tuples to be shuffled.

Name	Age	Sex	Bill Amount	No of check ups	Address	Criticality rate of Disease(Out of 10)
ZZZ	**23**	**M**	**16000**	**2**	**ZZ**	**7**
YYY	**30-40**	**M**	**20000**	**2**	**YY**	**5**
XXX	**27**	**F**	**26000**	**2**	**XX**	**9**
WWW	**30-40**	**M**	**20000**	**2**	**YY**	**6**
***ZZZ**	**20-30**	**M**	**31000**	**2**	**ZZ**	**10**
VVV	**23**	**M**	**16000**	**1**	**ZZ**	**8**
XXX	**30-40**	**M**	**20000**	**1**	**YY**	**8**
UUU	**23**	**F**	**20000**	**1**	**TT**	**7**
TTT	**30-40**	**M**	**20000**	**3**	**YY**	**7**
QQQ	**27**	**F**	**26000**	**2**	**XX**	**9**
WWW	**23**	**F**	**20000**	**3**	**TT**	**7**
***RRR**	**20-30**	**M**	**35000**	**1**	**ZZ**	**8**
***SSS**	**30-40**	**M**	**31000**	**2**	**ZZ**	**8**

Tuples with asterisk are selected.

Table 4 explained the sliced method for patient dataset for age, sex, bileing details with Disease. The privacy preservation technique can be appilied for selected arttibutes of dataset. Table 5 explores the suffling process of selected credential arrtibutes.

Table 4: Sliced Tables for Patient Dataset

{Age,Sex,Bill Amount}	{No of check ups,Address,Criticality rate of Disease(out of 10)}
{23 ,M,16000}	{2,ZZ,7}
{30-40,M,20000}	{2,YY,5}
{27,F,26000}	{2,XX,9}
{30-40,M,20000}	{2,YY,6}
*{20-30,M,31000}	{2,ZZ,10}
{23,M,16000}	{1,ZZ,8}
{30-40,M,20000}	{1,YY,8}
{23,F,20000}	{1,TT,7}
{30-40,M,20000}	{3,YY,7}
{27,F,26000}	{2,XX,9}
{23,F,20000}	{3,TT,7}
*{20-30,M,35000}	{1,ZZ,8}
*{30-40,M,31000}	{2,ZZ,8}

Table 5: After Selective Shuffling

{Age,Sex,Bill Amount}	{No of check ups,Address,Criticality rate of Disease(out of 10)}
{23 ,M,16000}	{2,ZZ,7}
{30-40,M,20000}	{2,YY,5}
{27,F,26000}	{2,XX,9}
{30-40,M,20000}	{2,YY,6}
*{20-30,M,31000}	{2,ZZ,8}
{23,M,16000}	{1,ZZ,8}
{30-40,M,20000}	{1,YY,8}
{23,F,20000}	{1,TT,7}
{30-40,M,20000}	{3,YY,7}
{27,F,26000}	{2,XX,9}
{23,F,20000}	{3,TT,7}
*{20-30,M,35000}	{2,ZZ,10}
*{30-40,M,31000}	{1,ZZ,8}

2.4 Enhancing efficiency in data mining using classification algorithm:

2.4.1. Data collection:

Data collection is a means for gathering facts, statistics and details from different sources. In this stage, data set consists of large number of files 1000 data from distributed data out of which 100 are from particular instances. This hosts information about different types of loans and their criteria information. The related data are collected based on various surveys, records, feedbacks, customer information, and loan details in different branches of a bank details in Figure 2.

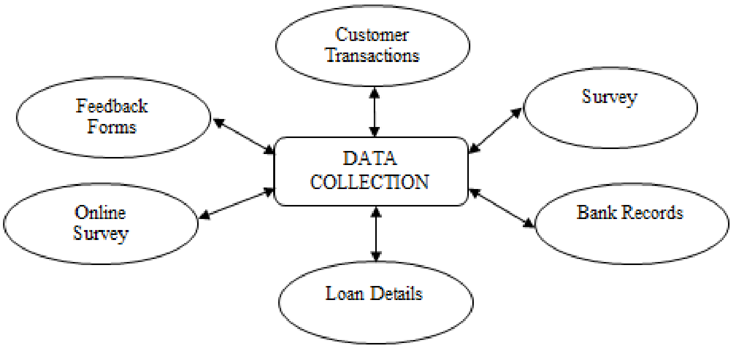


Figure 2. Data Collection

2.4.2 Dataset creation:
The data set preprocessing steps are explained details in Figure 3.

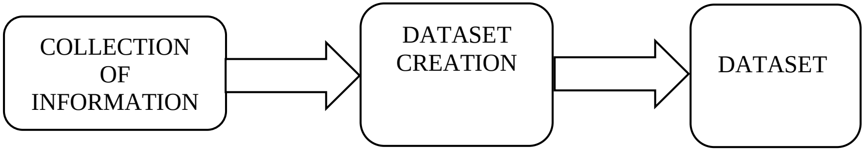


Figure 3. Dataset Creation

2.4.3. Feature extraction
The feature extraction method is applied to extract the fearture from preprocessed dataset which details are explained Figure 4.Hence; the proposed classifier will predict the dataset for visualizations.

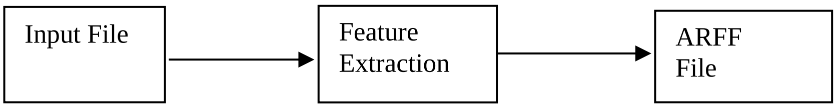


Figure 4. Feature Extraction

2.4.4. Support vector machine:
After the featurte extraction, Support vector machine process the prediction process to visulaize data with privacy. The SVM predict the privacy preserved data based on slected attributes that process details are explained in Figure 5.

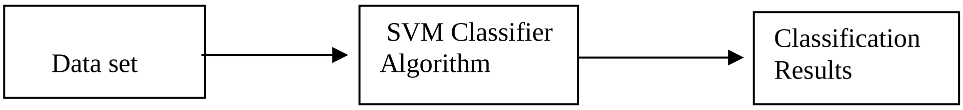


Figure 5. Support Vector Machine

2.4.5. Naive bayes classifier:
We use Naïve Bayes Classification algorithm to perk up the time efficiency. It has worked quite well in many intricate real-world circumstances. Naive Bayes is a keen fast learning classifier. Thus, it could be used for making predictions in real time that detail shows in Figure 6. It is easy to build and predominantly positive for very bulky datasets and it is known to surpass even exceedingly classy classification methods.

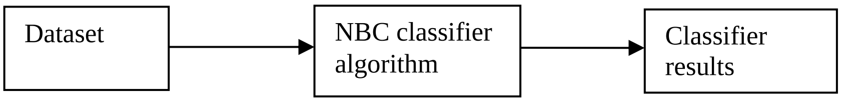
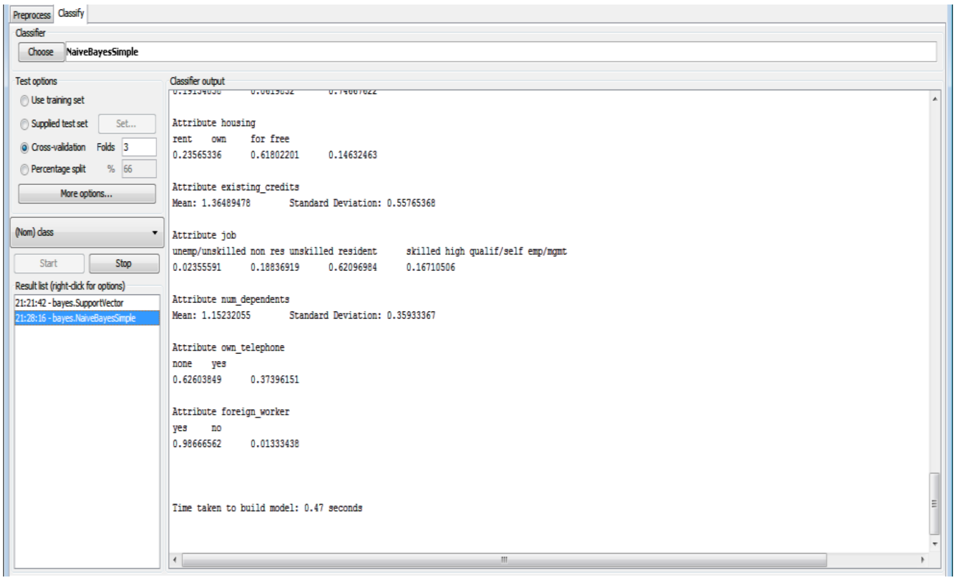


Figure 6. Naïve Bayes Classifier



NaiveBayesSimple

Folds 3

21:21:42 - bayes.SupportVector
21:28:16 - bayes.NaiveBayesSimple

```
Attribute housing
rent    own     for free
0.23565336      0.61802201      0.14632463

Attribute existing_credits
Mean: 1.36489478        Standard Deviation: 0.55765368

Attribute job
unemp/unskilled non res unskilled resident      skilled high qualif/self emp/mgmt
0.02355591      0.18836919      0.62096984      0.16710506

Attribute num_dependents
Mean: 1.15232055        Standard Deviation: 0.35933367

Attribute own_telephone
none    yes
0.62603849      0.37396151

Attribute foreign_worker
yes     no
0.98666562      0.01333438



Time taken to build model: 0.47 seconds
```

Figure 7. Predicted Result of Naïve Bayes Classifier

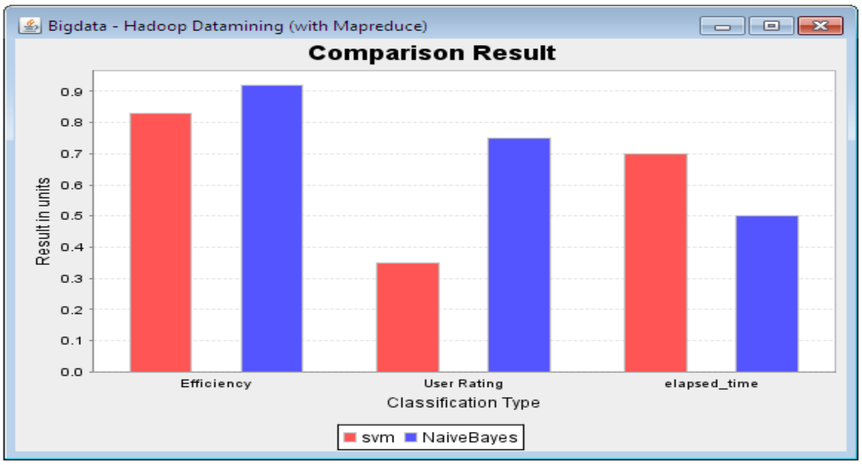


Figure 8.Comparative Result of NB(Naïve Bayes) & SVM Support Vector Machine)

Figure 7 eshows Naïve Bayes Classification techniques result efficiency, includes CPU processing time, retrieval time, computation time. Naïve bayes algorithm ensures that miner can secure mine more efficiently from the enormous database. Hence, Figure 8 depicts the proposed Naïve Bayes Classification techniques best result compare the other techniques.

3. Conclusion

By Naïve Bayes Classification algorithm, the whole time which includes CPU processing time, retrieval time, computation time will be reduced. Because of the parallel processing, the speed of retrieving data from large datasets or enormous database is increased. Naïve Bayes Algorithm will also predict more accurately. The prediction will base on the criteria given by the management system. It is very simple representation and doesn't allow for rich hypotheses. It needs a very small amount of training data. For further enhancement we commenced a Naïve Bayes updatable algorithm which is the advanced version of Naïve Bayes classification algorithm.Thus the proposed technique Naïve bayes algorithm ensures that miner can mine more efficiently from the enormous database. Finally naïve bayes algorithm is easy to implement and it works in more beneficial way. It is preferred to choose this naïve bayes algorithm rather than other classification algorithms. This method is popularly known as "punching bag" for smarter algorithms.

REFERENCES

[1] Zhang, L., Li, X. Y., & Liu, Y. *Message in a sealed bottle: Privacy preserving friending in social networks*. In Distributed Computing Systems (ICDCS), 2013 IEEE 33rd International Conference on IEEE, 2013; 327-336.

[2] Goga, O., Loiseau, P., Sommer, R., Teixeira, R., & Gummadi, K. P. *On the reliability of profile matching across large online social networks*. In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015; 1799-1808.

[3] Sun, Y., Wang, N., Shen, X. L., & Zhang, J. X. Location information disclosure in location-based social network services: Privacy calculus, benefits structure, and gender differences. *Computers in Human Behavior,* 2015; 52: 278-292

[4] Prakash, G., Saurav, N., & Kethu, V. R., "An Effective Undesired Content Filtration and Predictions Framework in Online Social Network", *International Journal of Advances in Signal and Image Sciences,* vol. 2, no. 2, pp. 1-8, 2016.

[5] Olanrewaju, R. F., & Azman, A. W., "Intelligent Cooperative Adaptive Weight Ranking Policy via dynamic aging based on NB and J48 classifiers", *Indonesian Journal of Electrical Engineering and Informatics (IJEEI),* vol. *5, no.* 4, pp. 357-365, 2017.

[6] Sulthana, R., & Ramasamy, S., "Context Based Classification of Reviews Using Association Rule Mining, Fuzzy Logics and Ontology", *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. *6, no.*3, pp. 250-255, 2017.

[7] Rao, R. R., & Makkithaya, K., "Learning from a Class Imbalanced Public Health Dataset: a Cost-based Comparison of Classifier Performance", *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 7, no. 4, pp. 2215-2222, 2017.

[8] Dhivakar K., Mohana S., "A Survey on Privacy Preservation Recent Approaches and Techniques", *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 2, issue 11, 2014, pp. 6559-6566.

[9] Janbandhu S., Chaware S.M, "Survey on Data Mining with Privacy Preservation", *International Journal of Computer Science and Information Technologies*, Vol. 5, No.4, 2014, pp. 5279-5283.

[10] Patel J. D., Patel S., "A Survey on Data Perturbation Techniques for Privacy Preserving in Data Mining", *International Journal for Scientific Research & Development*, vol. 3, issue 01, pp. 52-54, 2015