A Comparison on Performance of Data Mining Algorithms in Classification of Social Network Data

Mrs.P.Nancy
Ph.D Research Scholar,
Department of Computer science and Engineering,
Rajalakshmi Engineering College,
Affiliated to Anna University,
Chennai, Tamilnadu, India.

Dr.R.Geetha Ramani
Professor & Head,
Department of Computer science and Engineering,
Rajalakshmi Engineering College,
Affiliated to Anna University,
Chennai, Tamilnadu, India.

ABSTRACT
Data Mining (the analysis step of the Knowledge Discovery in Databases process or KDD), a relatively young and interdisciplinary field of computer science, is the process of discovering or extracting new patterns from large data sets involving methods from statistics and artificial intelligence. It is commonly used in marketing, surveillance, fraud detection, scientific discovery and now gaining wide way in social networking. Anything and everything on the Internet is fair game for extreme data mining practices. Social media covers all aspects of the social side of the internet that allow us to get contact and carve up information with others as well as intermingle with any number of people in any place in the world. This paper uses the dataset "Social side of the Internet" from Pew Research Center. The focus of the research is towards exploration on impact of the internet on social group activities using Data Mining Techniques. The original dataset contains 162 attributes which is very large and hence the essential attributes required for the analysis are selected by feature reduction method. The selected attributes were applied to Data Mining Classification Algorithms such as RndTree, ID3, K-NN, C-RT, CS-CRT, C4.5 and CS-MC4. The Error rates of various classification Algorithms were compared to bring out the best and effective Algorithm suitable for this dataset.

General Terms
Feature Selection, Algorithm, error rates.

Keywords
Knowledge discovery in databases, data mining, surveys.

1. INTRODUCTION
Knowledge discovery in databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [1] [10] [11]. Data mining is the process of automatic classification of cases based on data patterns obtained from a dataset [5]. Data Mining involves an incorporation of techniques from multiple disciplines such as database and data warehouse technologies, statistics, machine learning, pattern recognition, neural networks and data visualization. A number of Algorithms have been developed and implemented to dig out information and discern knowledge patterns that may be constructive for decision support. Once these patterns are extracted they can be used for automatic classification of case mixes [1]. Classification and prediction [12] [13] are the techniques used to make out important data classes and predict probable trends. Anything and everything on the Internet is fare game for extreme data mining practices. Social media covers all aspects of the social side of the internet that allow us to get contact and carve up information with others as well as interact with any number of people in any place in the world.

D. E. Brown, V. Corruble, and C. L. Pittard [6] compared decision tree classifiers with back propagation neural networks for multimodal classification problems. J. Catlett [7] has explained how knowledge patterns can be generated from large databases. M. James [8] in his work describes the various classification algorithms. T. Cover and P. Hart [9] performed classification using K-NN and proved its accuracy.

The dataset used in this paper is from "Social side of the Internet" obtained from a new national survey by the Pew Research Center. This report is based on the findings of a survey on Americans' use of the Internet. The results in this report are based on data from telephone interviews conducted by Princeton Survey Research Associates International from November 23 to December 21, 2010, among a sample of 2,303 adults, age 18 and older. Telephone interviews were conducted in English and Spanish by landline (1,555) and cell phone (748, including 310 without a landline phone). In this survey, Pew Internet asked about 162 questions to 27 different kinds of groups and found great diversity in group membership and participation using traditional and new technologies. It becomes clear as people are asked about their activities that their use of the internet is having a wide ranging impact on their engagement with civic, social, and religious groups. [2].

It was found from the survey conducted by Pew Research Center that 75% of all American adults are active in some kind of voluntary group or organization and internet users are more likely than others to be active: 80% of internet users participate in groups, compared with 56% of non-internet users. And social media users are even more likely to be active: 82% of social network users and 85% of Twitter users are group participants. This dataset is used for the first time in comparing the Data Mining Classification Algorithms.

1.1 Organization of the Paper
The paper is organized as follows: Section 2 gives the portrayal of the dataset and its categorization which is under consideration for this research and Section 3 defines the proposed system and

its phases. Analysis and results are presented in Section 4 and finally, Section 5 gives the conclusion of the research.

2. DATA SET DESCRIPTION

The dataset used in this paper is from "Social side of the Internet" obtained from a new national survey by the Pew Research Center. This report is based on the findings of a survey on Americans' use of the Internet. The Dataset includes 162 attributes with 2303 records. The attributes were based on the questions posed towards the people. Some of the Sample questions in the survey include:

- When you, personally, are deciding whether to join a new social, civic, professional, religious or spiritual group or organization, how important is -- Whether you think the group can accomplish its goals?
- When you, personally, are deciding whether to join a new social, civic, professional, religious or spiritual group or organization, how important is -- How much it costs to participate in the group?
- There are different things that might keep a person from participating in groups. Is -- You can't find groups or organizations with people who share your interests and beliefs -- a reason for you, or not? IF YES: A MAJOR or MINOR reason?
- There are many different ways people can participate in social, civic, professional, religious or spiritual groups today. In the past 30 days, have you -- Attended meetings or events for a group you are active in?
- Thinking about the different social, civic, professional, religious or spiritual groups in which you are currently active do any of these groups -- Have their own blog?
- Overall, would you say the internet has a MAJOR, MINOR or NO impact at all on your ability to -- Find social, civic, professional, religious or spiritual groups that match your interests?
- Thinking again about all of the different groups in which you are currently active. Did you discover any of these groups ON THE INTERNET that you otherwise would not have known about, or not?

The original dataset is very vast with 162 attributes. To begin with, it is categorized into subsets for analysis of s Algorithms in Data Mining which is shown in the Table 1.

Table 1. Description of subsets

Subset Number	Focus
1	Participation in Face book
2	Participation in Twitter
3	Impact of Internet on Social Group activities & accomplishment

In subset 3, the focus is on impact of internet on the ability of social, civic, professional, religious or spiritual groups towards various aspects which is depicted in the Table 2.

Table 2. Survey questions to identify the impact of internet on social groups

Question Number	Description
1	Overall, do you think the internet has a MAJOR, MINOR or NO impact at all on the ability of social, civic, professional, religious or spiritual groups to -- Recruit new members?
2	Overall, do you think the internet has a MAJOR, MINOR or NO impact at all on the ability of social, civic, professional, religious or spiritual groups to -- Impact local communities?
3	Overall, do you think the internet has a MAJOR, MINOR or NO impact at all on the ability of social, civic, professional, religious or spiritual groups to -- Impact society at large?
4	Overall, do you think the internet has a MAJOR, MINOR or NO impact at all on the ability of social, civic, professional, religious or spiritual groups to -- Communicate with members?
5	Overall, do you think the internet has a MAJOR, MINOR or NO impact at all on the ability of social, civic, professional, religious or spiritual groups to -- Find people to take leadership roles?
6	Overall, do you think the internet has a MAJOR, MINOR or NO impact at all on the ability of social, civic, professional, religious or spiritual groups to -- Organize activities?
7	Overall, do you think the internet has a MAJOR, MINOR or NO impact at all on the ability of social, civic, professional, religious or spiritual groups to -- Raise money?
8	Overall, do you think the internet has a MAJOR, MINOR or NO impact at all on the ability of social, civic, professional, religious or spiritual groups to -- Draw attention to an issue?
9	Overall, do you think the internet has a MAJOR, MINOR or NO impact at all on the ability of social, civic, professional, religious or spiritual groups to -- Connect with other groups?

3. PROPOSED SYSTEM MODEL

This section deals with the architecture of the proposed system model which is shown in Figure 1. The subsets of the original dataset as described in Table 1 are considered for further analysis of Classification Algorithms.

It includes the following phases:

- Data Cleaning(replacement of missing Values)
- Data Pre-processing
  Feature Reduction (relevant attributes required to perform are selected)
- Data Mining Classification Algorithms
- Analysis of error rates produced by Algorithms
- Identifying the best Algorithm for the dataset

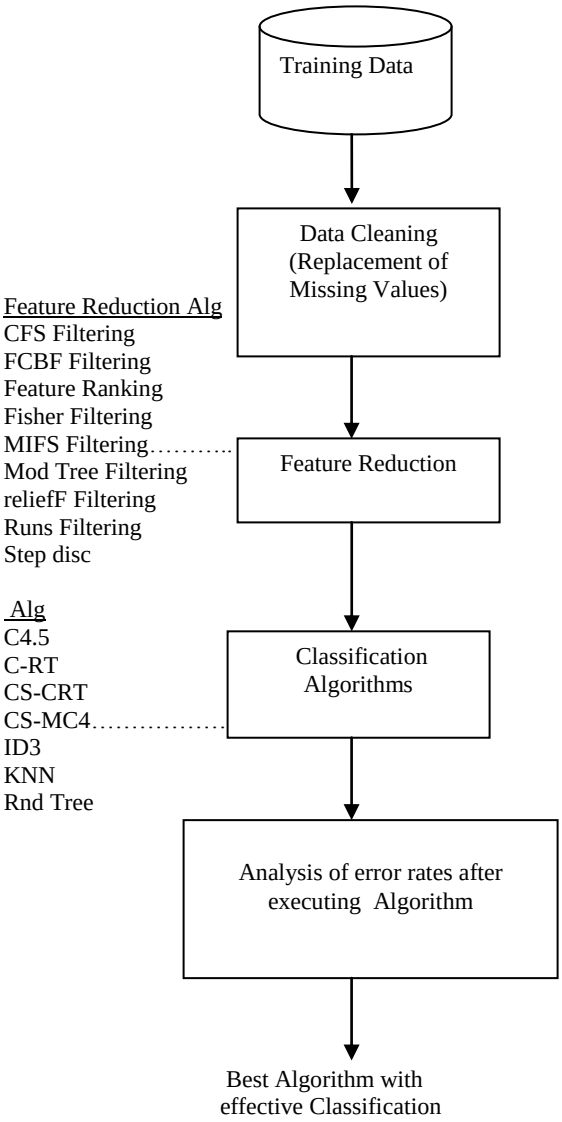

Fig 1: Architecture of the Proposed System Model

3.1 Data Cleaning

The wealth of data, coupled with the want for powerful data analysis tools, has been described as data affluent but information mearge situation. In World Wide Web, where data flow in and out like streams, huge volumes of data can be accumulated beyond databases and data warehouses. The original dataset contained some missing values for various attributes. To proceed with the work, those missing values were replaced as if the people answered don't know.

3.2 Feature Reduction: After replacing the missing values, some preprocessing of the data is to be carried out to proceed further. Feature Reduction is one of the preprocessing techniques. In this phase the important features required to implement the Classification Algorithm are identified. By Feature Reduction, the model complexity is reduced and it is easier to interpret. Moreover, the attenuation of the variables to collect is an advantage during the deployment of the model. In some cases, the variable selection enables to improve the model accuracy. Manual selection by an expert domain is certainly the best approach. But because the number of candidate descriptors is often large, it is not always possible in practice. [4]. So, we must select automatically the best variables. We can also use the automatic process as a preliminary approach in order to filter out the really irrelevant attributes. The various feature selection Algorithms that were tried includes:

3.2.1. Feature Ranking:

This Algorithm ranks the attributes based on their relevance. A cutting rule enables to select a subset of these attributes. It is a supervised Algorithm; we must define the discrete target attribute. This approach does not take into consideration the redundancy of the input attributes. [3]

3.2.2. reliefF Filtering:

This is a supervised Algorithm which will not consider the redundancy of the input attributes. At least two attributes must be available and the target attribute must be discrete. [3]

3.2.3. Fast Correlation based Filtering (FCBF):

It is a supervised feature selection Algorithm based upon a filtering approach i.e., processes the selection independently from the learning Algorithm. This Algorithm, unlike the ranking approaches, takes into consideration the redundancy of the input attributes. [3]

3.2.4. Fisher Filtering:

It is a supervised feature selection Algorithms based upon a filtering approach i.e., processes the selection independently from the learning Algorithm. This component ranks the inputs attributes according to their relevance. It is a supervised Algorithm; we must define the discrete target attribute. This approach does not take into consideration the redundancy of the input attributes. [3]

3.2.5. Stepwise discriminant:

Step disc is always associated to discriminant .We implement the FORWARD and the BACKWARD strategies in TANAGRA. In the FORWARD approach, at each step, we determine the variable that really contributes to the discrimination between the groups. We add this variable if its contribution is significant. The process stops when there is no attribute to add in the model. In the BACKWARD approach, we begin with the complete model with all descriptors. We search which is the less relevant variable. We remove this variable if

the removing does not significantly damage the discrimination between groups. The process stops when there is no variable to remove. [3]

3.2.6. Correlation based Feature Selection (CFS):

It is a supervised feature selection Algorithm based upon a filtering approach .i.e. processes the selection independently from the learning Algorithm. This Algorithm unlike the ranking approaches, takes into consideration the redundancy of the input attributes. [3]

3.2.7. MIFS Feature Filtering:

It is a supervised feature selection Algorithm based upon a filtering approach. .i.e. processes the selection independently from the learning Algorithm. This Algorithm unlike the ranking approaches, takes into consideration the redundancy of the input attributes. [3]

3.2.8. Multivalued Oblivious Decision Tree Feature Selection (MOD Tree):

It is a supervised feature selection Algorithm based upon a filtering approach. .i.e. processes the selection independently from the learning Algorithm. This Algorithm unlike the ranking approaches, takes into consideration the redundancy of the input attributes. [3]

3.2.9 .Runs Filtering:

It is a supervised feature selection Algorithm based upon a filtering approach. .i.e. processes the selection independently from the learning Algorithm. This component ranks the input attributes according to their relevance. [3]

3.3 Classification Algorithms

The goal of Classification is to build a set of models that can correctly foresee the class of the different objects. Classification is a two-step process, 1. Build model using training data. Every object of the data must be pre-classified i.e. its class label must be known. 2. The model generated in the preceding step is tested by assigning class labels to data objects in a test dataset. The test data may be different from the training data. Every element of the test data is also reclassified in advance. The accuracy of the model is determined by comparing true class labels in the testing set with those assigned by the model. The input to these methods is a set of objects (i.e., training data), the classes which these objects belong to (i.e., dependent variables), and a set of variables describing different characteristics of the objects (i.e., independent variables). [4]. The key advantage of supervised learning methods over unsupervised methods (for example, clustering) is that by having an explicit knowledge of the classes the different objects belong to these Algorithms can perform an effective feature selection if that leads to better prediction accuracy. The following are brief outline of some Classification Algorithms that had been used in data mining and machine learning area and used as base Algorithms in this researcht.

3.3.1. k-Nearest Neighbour (KNN) Algorithm:

KNN classifier is an instance-based learning Algorithm which is based on a distance function for pairs of observations, such as the Euclidean distance or Cosine. In this paradigm, k nearest neighbors of a training data is computed first. Then the similarities of one sample from testing data to the k nearest neighbors are aggregated according to the class of the neighbors, and the testing sample is assigned to the most similar class.

3.3.2. ID3 (Iterative Dichotomiser 3) Algorithm:

It is an Algorithm used to generate a decision tree invented by Ross Quinlan. ID3 is precursor to the C4.5 Algorithm. The work flow of the Algorithm is shown in Figure 2.

```
A = The Attribute that best classifies examples.
Decision Tree attribute for Root = A.
For each possible value, vi, of A,
Add a new tree branch below Root, corresponding to test
 A = vi.
Let Examples(vi) be the subset of examples that have value
vi for A
If Examples(vi) is empty
Then below this new branch add a leaf node with
label = most common target value in the examples
Else below this new branch add the sub tree
ID3(Examples(vi), Target Attribute, Attributes – {A})
End
Return Root
```

Fig 2: ID3 Algorithm

3.3.3. C4.5 Algorithm:

It is also called as statiscal classifier. The pseudo code of the general Algorithm is as follows:
Check for base cases. For each attribute a, Find the normalized information gain from splitting on a. Let a_best be the attribute with the highest normalized information gain .Create a decision node that splits on a_best. Recurse on the sub lists obtained by splitting on a_best, and add those nodes as children of node

3.3.4. RndTree (Random Forest):

Each tree is constructed using the following Algorithm:
Let the number of training cases be *N*, and the number of variables in the classifier be *M*.

We are told the number *m* of input variables to be used to determine the decision at a node of the tree; *m* should be much less than *M*.

1. Choose a training set for this tree by choosing *n* times with replacement from all *N* available training cases (i.e. take a bootstrap sample). Use the rest of the cases to estimate the error of the tree, by predicting their classes.
2. For each node of the tree, randomly choose *m* variables on which to base the decision at that node. Calculate the best split based on these *m* variables in the training set. Each tree is fully grown and not pruned as done in constructing a normal tree classifier.

For prediction a new sample is pushed down the tree. It is assigned the label of the training sample in the terminal node it ends up in. This procedure is iterated over all trees in the ensemble, and the average vote of all trees is reported as random forest prediction. The other Algorithms that were tried include

C-RT, CS-CRT, and CS-MC4. The experiments conducted and results obtained are described in the next section.

4. ANALYSIS AND RESULTS

This section shows the analysis after executing various Classification Algorithms as per the requirements and explores the results of the same. The whole experiment is carried out with the Data Mining tool TANAGRA. The analysis of Feature Reduction technique is described in section 4.1 and the analysis of execution of the Classification Algorithm is described in section 4.2.

4.1 Analysis of Feature Reduction

The features selected by feature reduction technique are chosen as input attributes with necessary class variable as the target attribute and various classification Algorithms were executed for all selected features one by one. The total number of attributes in the original dataset is 162. After performing feature reduction for the required subsets as shown in Table 1, important attributes were selected whose counts are shown in Table 3 & Table 4.

Table 3. Attributes selected after Feature Reduction for subset 1 & 2

Feature Selection Algorithm	Focus towards participation in Face book	Focus towards participation in twitter
CFS	8	8
FCBF	4	4
Feature Ranking	123	125
Fisher Filtering	120	123
MIFS Filtering	56	58
Mod Tree Filtering	9	10
ReliefF Filtering	36	38
Runs Filtering	14	16
Step Disc	50	55

It does not imply that higher the number of attributes selected higher the accuracy of the classification algorithm. Even if less number of attributes were used, the attributes selected should be highly relevant for the target attribute or class attribute. For subset 1, the features selected by Feature ranking gave good results. For subset 2, reliefF Filtering produced good results. For subset 3, with a set of nine questions, same feature reduction Algorithms were applied and relevant attributes were identified and the counts of attributes selected are shown in Table 4.

Different algorithms gave different attributes and the best is selected for every survey question separately and necessary graph is drawn for the same, a sample of which is shown in Figure 3.

Table 4. Attributes selected after Feature Reduction for subset 3

Feature Selection Algorithm	1	2	3	4	5
CFS	8	8	7	8	8
FCBF	5	5	5	4	5
Feature Ranking	126	125	121	124	125
Fisher Filtering	120	112	118	124	123
MIFS Filtering	60	55	56	58	59
Mod Tree Filtering	10	8	9	12	11
reliefF Filtering	38	37	43	41	35
Runs Filtering	29	14	19	36	17
Step Disc	60	45	47	48	45

Feature Selection Algorithm	6	7	8	9
CFS	9	8	8	8
FCBF	7	2	9	5
Feature Ranking	125	126	123	124
Fisher Filtering	124	123	120	123
MIFS Filtering	61	56	61	59
Mod Tree Filtering	11	8	11	11
reliefF Filtering	45	36	39	40
Runs Filtering	29	12	8	35
Step Disc	50	50	59	56

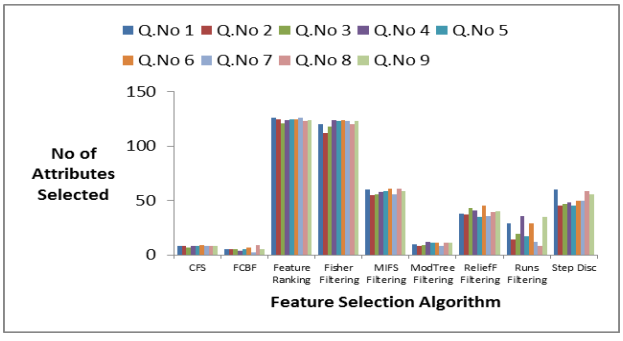

Fig 3: Attributes Selected after Feature Reduction for subset 1 & 2

4.2. Analysis of Classification Algorithm

In this section we present a comparative study of various data mining classification algorithms on the Dataset "Social side of the Internet". For subset 1, the features selected by Feature

ranking gave good results. For subset 2, reliefF Filtering produced good results.

The features selected by feature reduction technique are chosen as input attributes with necessary class variable as the target attribute and various classification Algorithms were executed for all selected features one by one**.** For subset 1 & 2, relevant attributes identified by feature reduction are executed by various Classification Algorithm and different error rates were identified and mentioned in the Table 5.

Table 5 . Error rates after executing Classification Algorithms for subset 1& 2

Classification Algorithm	Error rates	
	Face book	Twitter
C4.5	0.0860	0.1042
C-RT	0.1798	0.1976
CS-CRT	0.1798	0.1976
CS-MC4	0.1099	0.1107
ID3	0.1650	0.2084
KNN	0.1871	0.2097
Rnd Tree	**0.0004**	**0.0004**

From Table 5, it is clear that the error rate generated by Rnd Tree Algorithm is very less compared to all other Algorithms. The misclassifications identified were very less. A Graph drawn for the error rates after executing the Algorithm for the attributes selected by Feature reduction is shown in Figure 4.

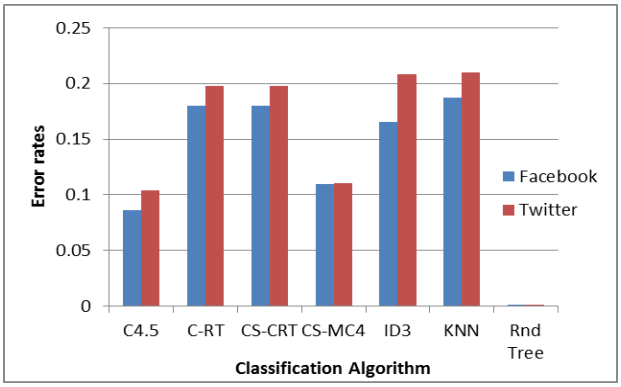

Fig 4: Comparison of error rates for subset 1 & 2

A confusion Matrix is obtained. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. One benefit of a confusion matrix is that it is easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another). A sample Confusion Matrix for RnDTree Classification Algorithm is shown in Figure 5. In the Figure 5, n, d, s, N and r are various identifiers and the descriptions are shown in the Table 7.

	n	d	s	N	r	Sum
n	1187	0	0	0	0	1187
d	0	782	0	0	0	782
s	0	0	269	0	0	269
N	1	0	0	59	0	60
r	0	0	0	0	5	5
Sum	1188	782	269	59	5	2303

Fig 5: A Sample Confusion Matrix for RnD Tree Algorithm for subset 2

Table 7. Description of Confusion Matrix.

Identifier	Description
n	Not a Twitter User
d	Don't Know
s	Twitter User
N	Non user of Internet
r	Refused to answer

Similarly for subset 3 also different Algorithms were tried and the corresponding error rates for different survey questions are shown in the Table 6.

Table 6. Error rates after executing Algorithms for subset 3

Algorithm	**1**	**2**	**3**	**4**	**5**
C4.5	0.1481	0.1676	0.1481	0.1233	0.1937
C-RT	0.3183	0.3122	0.2723	0.2280	0.3930
CS-CRT	0.3183	0.3122	0.2723	0.2280	0.3930
CS-MC4	0.2132	0.2141	0.2102	0.1776	0.2501
ID3	0.3330	0.3231	0.2827	0.2397	0.4108
KNN	0.3092	0.3261	0.2875	0.2280	0.3474
Rnd Tree	**0.0035**	**0.0026**	**0.0026**	**0.0026**	**0.0100**

Algorithm	**6**	**7**	**8**	**9**
C4.5	0.1424	0.1575	0.1198	0.1432
C-RT	0.2966	0.3183	0.2579	0.2145
CS-CRT	0.2966	0.3183	0.2579	0.2145
CS-MC4	0.1785	0.2162	0.1811	0.1584
ID3	0.3040	0.3300	0.2779	0.2953
KNN	0.2631	0.3265	0.2575	0.2144
Rnd Tree	**0.0056**	**0.0048**	**0.0035**	**0.0038**

From Figure 5, we can infer that n, d, s and r have no misclassifications whereas N has one misclassification where it has been identified as n. After analysis of the results it is clear that the Classification Algorithm RndTree gave lesser error rates when compared to other Classification Algorithms for this dataset and declared as best Algorithm with efficient as for as the dataset " Social Side of the Internet" is concerned.

5. CONCLUSION

Data mining is a broad area that integrates techniques from several fields including machine learning, statistics, pattern recognition, artificial intelligence, and database systems, for the analysis of large volumes of data. Social network analysis applications have experienced tremendous advances within the last few years due in part to increasing trends towards users interacting with each other on the internet. There have been a large number of data mining Algorithms rooted in these fields to perform different data analysis tasks. In this paper, the comparison on the performance of various Data Mining Classification Algorithms were executed on the dataset "Social side of the Internet". To start with the entire dataset is categorized into 3 subsets. The entire attribute set includes 162 attributes which is very vast and hence feature reduction is performed to identify the highly relevant attribute for the target variable. The selected attributes were given as input to various Data Mining Classification Algorithm and the error rates were analysed and compared. From the results it is clear that in all the subsets considered for the research RndTree Algorithm produced less error rates when compared to all other Algorithms.

6. REFERENCES

[1] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. From Data Mining to Knowledge Discovery: An Overview. In Fayyad, U., Piatetsky-Shapiro, G., Amith, Smyth, P., and Uthurusamy, R. (eds.), Advances in Knowledge Discovery and Data Mining, MIT Press, 1-36, Cambridge, 1996

[2] Report on "Social side of the Internet" http://pewinternet.org/Reports/2011/The-Social-Side-of-the-Internet.aspx. This website provides a report with detailed information about Social side of the Internet.

[3] Tanagra Data Mining tutorials, http://data-mining-tutorials.blogspot.com/ This website provides detailed information on the basics of Data Mining Algorithms

[4] Dr. Varun Kumar, Luxmi Verma," Binary Classifiers for Health Care Databases: A Comparative Study of Data Mining Algorithms in the Diagnosis of Breast Cancer" in IJCST Vol. 1, Issue 2, December 2010

[5] Desouza, K.C. (2001) Artificial intelligence for healthcare management In Proceedings of the First International Conference on Management of Healthcare and Medical Technology Enschede, Netherlands: Institute for Healthcare Technology Management.

[6] D. E. Brown, V. Corruble, and C. L. Pittard. A comparison of decision tree classifiers with backpropagation neural networks for multimodal classification problems. Pattern Recognition, 26:953-961, 1993.

[7] J. Catlett. Megainduction: Machine Learning on Very large Databases. PHD Thesis, University of Sydney, 1991.

[8] M. James. Classification Algorithms. John Wiley, 1985.

[9] T. Cover and P. Hart. Nearest neighbor pattern classification. IEEE Trans. Information Theory, 13:21-27, 1967.

[10] Fayyad, Usama; Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996). "From Data Mining to Knowledge Discovery in Databases". Retrieved 2008-12-17.

[11] Fayyad, U. Data Mining and Knowledge Discovery: Making Sense Out of Data. IEEE Expert, v. 11, no. 5, pp. 20-25, October 1996. Exclusive Ore Inc. The Exclusive Ore Internet Site, http://www.xore.com, 1999.

[12] K. Cios, W. Pedrycz, and R. Swiniarski. Data Mining Methods for Knowledge Discovery. Boston: Kluwer Academic Publishers, 1998

[13] W. Ressom, Rency S. Varghese, Zhen Zhang, Jianhua Xuan, and Robert Clarke. 2008 Classification Algorithms for phenotype prediction in genomic and Proteomics Front BioScience.

7.AUTHORS PROFILE

Dr.R.Geetha Ramani is Professor & Head in Department of Computer Science and Engineering, Rajalakshmi Engineering College, India.. She has more than 15 years of teaching and research experience. Her areas of specialization include Data mining, Evolutionary Algorithms and Network Security. She has over 50 publications in International Conferences and Journals to her credit. She has also published a couple of books in the field of Data Mining and Evolutionary Algorithms. She has completed an External Agency Project in the field of Robotic Soccer and is currently working on projects in the field of Data Mining. She has served as a Member in the Board of Studies of Pondicherry Central University. She is presently a member in the Editorial Board of various reputed International Journals.

Mrs. P. Nancy completed her M.E. Computer Science & Engineering in Department of Science and Engineering at Bannari Amman Institute of Technology, Sathyamangalam, affiliated to Anna University, Chennai, India. She has more than 5 years of teaching experience. Presently she is pursuing her Ph.D in Computer Science and Engineering at Rajalakshmi Engineering College, affiliated to Anna University of Technology, Chennai. Her areas of interest include Data Mining, Data Structures, Computer Networks and Software Engineering. She has attended and presented few papers in National and International Conferences.