International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 3 Issue: 12

ISSN: 2454-4248
375 – 382

A Detailed Dominant Data Mining Approach for Predictive Modeling of Social Networking Data using WEKA

Ravi Arora
Computer Science & Engineering
LIET, Alwar, Rajasthan,
Indiaraviarora250785@gmail.com

Kuldeep Kumar Somvanshi
Computer Science & Engineering
LIET, Alwar, Rajasthan, India
sw.ee.dhanawat@gmail.com

Abstract:- Social network has gained popularity manifold in the last decade. Accessing social network sites such as Twitter, Facebook LinkedIn and Google+ through the internet and the web 2.0 technologies has become more affordable. People are becoming more interested in and relying on social network for information, news and opinion of other users on diverse subject matters. In this Paper, we present the first comprehensive review of social and computer science literature on trust in social networks. We first review the existing definitions of trust and define social trust in the context of social networks. Web-based social networks have become popular as a medium for disseminating information and connecting like-minded people. The public accessibility of such networks with the ability to share opinions, thoughts, information, and experience offers great promise to enterprises and governments. As the popularity increases and they became widely used as one of the important sources of news, people become more cautious about determining the trustworthiness of the information which is disseminating through social media for various reasons. For this reason, knowing the factors that influence the trust in social media content became very important. In this research paper, we use a survey as a mechanism to study trust in social networks. First, we prepared a questionnaire which focuses on measuring the ways in which social network users determine whether content is true or not and then we analyzed the response of individuals who participated in the survey and discuss the results in a focus group session. Then, the responses, we get from the survey and the focus group was used as a dataset for modeling trust, which incorporates factors that alter trust determination. The dataset preprocessing a total of 56 records were used for building the models. This Paper applies the Decision Tree, Bayesian Classifiers and Neural Network predictive data mining techniques in significant social media factors for predicting trust. To accomplish this goal: The WEKA data mining tool is used to evaluate the J48, Naïve Bayes and Multilayer Perception algorithms with different experiments were made by performing adjustments of the attributes and using various numbers of attributes in order to come up with a purposeful output.

Keywords *Social network, social trust, predictive analysis, weka, unsupervised learning.*

*****

1. INTRODUCTION

Social network is a term used to describe web-based services that allow individuals to create a public/semi-public profile within a domain such that they can communicatively connect with other users within the network [1]. Social network has improved on the concept and technology of Web 2.0, by enabling the formation and exchange of User-Generated Content [2]. Simply put, social network is a graph consisting of nodes and links used to represent social relations on social network sites [3]. The nodes include entities and the relationships between them forms the links (as shown in Fig. 1).

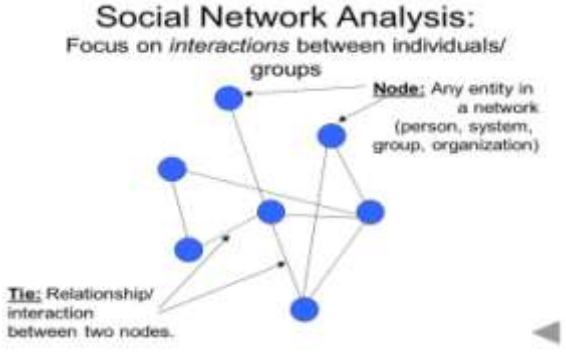


Fig.1 Social Network showing nodes and links

In social networks people keep in touch with their friends by posting some kind of content in their walls and sharing news, clips and any kinds of activities they have inclination to and preserve their involvement on the social media. Forming new relationship in these sites doesn't have any limitation of both place and time, which makes it quite easy and attractive. This days the number of people who use social media as a source of news is increasing rapidly even though they have still to a certain extent a doubt about truthfulness of the contents which are propagated across the social network in a daily basis.

A social network is a heterogeneous and multi relational dataset represented by a graph. Vertexes represent the objects (entities), edges represent the links (relationships or interaction), and both objects and links may have attributes3. Social networks are usually very large. Social network can be used to represents many real-world phenomena (not necessarily social), such as electrical power grids, Phone calls, spread of computer virus. Network construction from general, real-world data presents several unexpected challenges owing to the data domains themselves, e.g., information extraction and preprocessing, and to the data structures used for knowledge representation and storage

Since social networks are organized around the people who use them, trusting the content which is propagated in them is solely dependent on the determination ability of the users. If the users don't trust the information then he/she will not propagate it.

The term "social media" refers to the wide range of Internet-based and mobile services that allow users to

 ISSN: 2454-4248
375 – 382

participate in online exchanges. Just as the Internet has affected how people interact socially. Through the use of social media, people can exchange photos and videos, share news stories, post their thoughts on blogs, and participate in online discussions. Social media also allow individuals, companies, organizations, governments, and parliamentarians to interact with large numbers of people. Online social networks facilitate connections between people based on shared interests, values, membership in particular groups (i.e., friends, professional colleagues), etc. They make it easier for people to find and communicate with individuals who are in their networks using the Web as the interface.

The main objective of this dissertation is to assess the different ways of trust determination factors and to find the most important factors which can be used to model trust in social media content.

1.1 Classification Of Social Media

Facebook: Facebook is an online social networking service. Its name stems from the colloquial name for the book given to students at the start of the academic year by some American university administrations to help students get to know one another.

Google +: Google+ is a social networking and identity service owned and operated by Google Inc. It is the second-largest social networking site in the world, having surpassed Twitter in January 2013.

Twitter: Twitter is an online social networking and micro blogging service that enables users to send and read "tweets", which are text messages limited to 140 characters. Registered users can read and post tweets but unregistered users can only read them. Users access Twitter through the website interface, SMS, or mobile device app.

Youtube: YouTube is a video-sharing website. All YouTube users can upload videos up to 15 minutes each in duration. Users who have a good track record of complying with the site's Community Guidelines may be offered the ability to upload videos up to 12 hours in length, which requires verifying the account, normally through a mobile phone.

Blog: A blog (a contraction of the words web log) is a discussion or informational site published on the World Wide Web and consisting of discrete entries ("posts") typically displayed in reverse chronological order (the most recent post appears first).

LinkedIn: LinkedIn is a social networking website for people in professional occupations. One purpose of the site is to allow registered users to maintain a list of contact details of people with whom they have some level of relationship, called Connections. Users can invite anyone (whether a site user or not) to become a connection. Users can post their own photos and view photos of others to aid in identification.

The comparison of Social Media based on Users is shown in Fig.2 using collected data in Table 1.

TABLE 1Comparison of Social Media

Social Media	Users (Million)
Twitter	900
Facebook	700
Google +	401
Youtube	350
Blog	200
LinkedIn	175

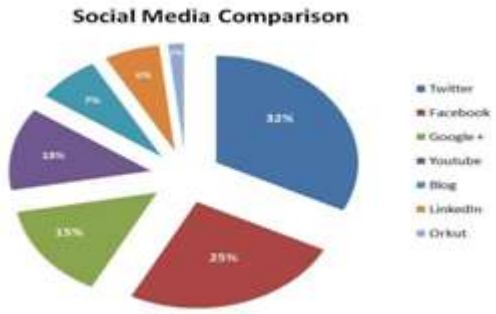


Fig.2 Comparison of Social Media based on Users

The organization of this papers is as follows ; In Section we have discussed the related work to our work , In Section 3 we have collected the data on the basis of questionnaires prepared and this section we have also done the analysis using different data mining approach on WEKA tool. The final section 4 concludes our paper.

2. RELATED WORK

This factor can be used for modeling of trust in this study, because in social network sites this factor have a huge influence on trusting a content which is shared by people who have already get a credibility because of their previous posts quality. In social network sites the most important factors for building trust are reputation and influence. When we say reputation in social media it means the way you are perceived by others solely based on your posts.

A trust indicates a positive belief in another person, or content in this particular case. Ordinary users are more

 ISSN: 2454-4248
375 – 382

likely to trust people who share information which is solely based on actual facts, like by attaching the links related to the contents they share, which will most likely increase the credibility of the information they share. Even though it is quite new area of research there are some useful researches which are done in the last few years.

Such as "Propagation Models for Trust and Distrust in Social Networks" by **Cai-Nicolas Ziegler and Georg Lausen [4],** proposes a model for both trust and distrust in social networks.And also the researches made by likes of "Models and Methods in Social Network Analysis" by **Carrington P. J., Scott J., and Wasserman S.(2005) [5]** and **"A Flexible Trust Model for Distributed Service Infrastructures" by Liu Y., Yau S., Peng D., and Yin Y. (2008)[6]** were really helpful in introducing some of the already existing trust metrics.

Azra Shamin et. al [7], proposed a framework for bio data analysis data mining technique on bio data as well as their proprietary data, Bio database is often distributed in nature. In this system take input from the user, preprocess the query and load it into local bio database. System will search the knowledge from Database and send it back to the user, if the data related to user query exists.

Rumi Ghosh et. al [8], proposed a framework to aggregate multiple heterogeneous documents to extract data from them. Therefore, in this paper, they propose a novel topic modeling framework, Probabilistic Source LDA which is planned to handle heterogeneous sources. Probabilistic Source LDA can compute latent topics for each source maintain topic-topic correspondence between sources and yet retain the distinct identity of each individual source. Therefore, it helps to mine and organize correlated information from many different sources.

Prakash R.Andhale et. Al [9], In this paper author represent the characteristics of HACE theorem which provides the description of heterogeneous data and proposes a model for processing of heterogeneous data from the view of data mining. This information extraction model involves the information extraction, data analysis and provides the security and privacy mechanism to the data.

Amir Ahmad et.al [10], performs k-mean clustering algorithm for both numerical and categorical data. In this paper, they present a modified approach of cluster center to overcome the numeric data only restriction of k-mean algorithm and provide a better categorization of clusters. The performance of the k-mean algorithm has been studied on real world data sets.

Dr. Goutam Chakra borty et.al [11], in this paper they represent a way at how to organize and analysis of textual data for getting useful data from huge collection of documents which improve the performance and business operations.

According to definition.net [12] trust means reliance on the integrity, strength, ability, surety, etc., of a person or thing; confidence.

"*Trust is both and emotional and logical act. Emotionally, it is where you expose yourvulnerabilities to people, but believing they will not take advantage of your openness. Logically, it is where you have assessed the probabilities of gain and loss, calculating expected utility based on hard performance data, and concluded that the person in question will behave in a predictable manner. In practice, trust is a bit of both. I trust you because I have experienced your trustworthiness and because I have faith in human nature*." **[13]**

Ming- Syan Chen et.al [14] , Data mining is the process of automatically collecting large volumes of data with the objective of finding hidden patterns and analyzing the relationships between numerous types of data to develop predictive models.

Data mining is an interdisciplinary field, the confluence of a set of disciplines, including database systems, statistics, machine learning, visualization, and information science. Moreover, depending on the data mining approach used, techniques from other disciplines may be applied, such as neural networks, fuzzy and/or rough set theory, knowledge representation, inductive logic programming, or high performance computing.

This research uses classification techniques [15, 16] for predicting trust. The three types of classification techniques that were used to construct prediction models are Decision Tree(j48), Neural Network(Multilayer perception) and Bayesian(Naïve Bayes) Classifiers.

Moreover, the three algorithms that were used to construct the models and the output matrices of the algorithms that were used to measure the performance of the algorithms and comparison are explained thoroughly.

As Han & Kamber [17] have stated Classification has two distinct processes, namely learning and classification. Throughout the learning process, a classifier will be built portraying a set of beforehand determined classes that will later portrayed in the form of classification rules.

V. S. Jagadheeswaran et.al describes [15], A decision tree is a data mining technique that generates a graphical illustration and analysis of the model it generates [15]. The model that is generated by decision tree could be either predictive or descriptive model.

Decision trees are widely used for classification purpose; they can be used also for different kinds of regression

 ISSN: 2454-4248
375 – 382

analysis.

Al Jarullah et.al [18] explains J48 is an implementation of the well known C4.5 algorithm for producing either pruned or unpruned C4.5 tree. The C4.5 algorithm was built based on the concept of information obtaining or entropy reduction to select the most efficient split.

In general, It assumes that individual attributes of the data can be used to make a decision by splitting the original data into minor subsets.

The J48 decision tree algorithm is the one that is used in this research to classify the social media content as trusted or non-trusted.

The main reason J48 decision tree was chosen to serve as a model for classification is that it produces simpler rules and remove insignificant parameters before it begins a process of tree induction. Usually, J48 decision trees happen to had a relatively higher accuracy than other classification algorithms, In addition, J48 also provides extremely fast and pretty powerful way of fast and powerful way to show structures for a data.

According to Quinlan [19] , Neural network make use of a multilayered approach which estimates sophisticated mathematical functions to process a specific data.

Neural networks are well known for their learning efficiency. They perform much better in comparison with the other classifier algorithms when the majority of variables are weakly relevant. One disadvantage of neural networks is that they took longer time to learn.

According to Bhargavi et.al [20], a Naïve Bayes classifier works under the assumption of that the presence of a specific feature of a class have no association to the presence of any other constituent.

The Naïve Bayes algorithm makes use of Bayes' Theorem, which is a formula that determines a probability by estimating the frequency of values and mixture of values in the previously collected data. It determines the probability of an event happening provided that the probability of another event that has already happened.

The Bayes' theorem is stated as follows

$$P(H/X) = P(X/H)\, P(H) / P(X)$$

The Naive Bayes algorithm provides a way to mix the prior probability and conditional probabilities within a single formula that can be used to determine the probability of each of the classifications in turn.

3. DATASETS AND RESULTANALYSIS

The source of data for this research is my own data set, which is obtained by using a questionnaire and focus group to collect information. The questionnaire was chosen to collect information because it makes it is easier to distribute to as many people as you want, however, it is quite difficult to get a detailed analysis by using just the data which is collected by questionnaire . As a result, we decided to use the focus group method to supplement the information we get from the questionnaire by discussing with people who have information technology educational back ground and pretty good technical know-how of the research area. Before starting to write the questions which were used in the questionnaire we made extensive research by reading articles related to the topic of our project, in particular about " trust ".

The whole questionnaire can be seen in the appendix section. After the data was gathered, the diagrams were created and analyzed with the help of Google form.

Following questionnaires are prepared using Google form.

Questionnaire

1, Please specify your Gender

- Male
- Female

2, Are you using any Social Networking? (Example - Face book, Tweeter, Google+, LinkedIn)

- Yes
- No

3, Is the number of people who commented or like a link which is shared in social media important for you when it comes to trusting the information.

- Unimportant
- Less important
- Neither
- Important
- Very Important

4, Is Knowing the person who shared the information (it could be personally) important for you?

- Unimportant
- Less important
- Neither /
- Important
- Very Important

 ISSN: 2454-4248
375 – 382

5, Age

6, Do you think engaging actively in social media will make a person more trustworthy?

- Yes
- No

7***, Do you use more than one social media networks.***

- Yes
- No

8, In your opinion, how important it is for a person to increase his trustworthiness by being actively engaged in more than one social media networks .

- Unimportant
- Less important
- Neither /
- Important
- Very Important

9, Is the number followers or friends the person sharing the information have influences your assessment of the credibility of the content.

- Yes
- No

10, Does the trustworthiness of a person depends on the quality of the previous posts, comments and links he/she shares.

- Yes
- No

11, On average, how many people should share a content before you start trusting the information.

- 1-5
- 6-10
- 11- 15
- 16 – 20
- More than 20

12, Do you think the information which is shared in social media is higher quality (trust worthy) than the traditional media outlets such as television, radio and newspapers?

- Yes
- No

13, Which social media platform is your favorite?

- Twitter
- Face book
- Google+
- LinkedIn

14, Have you ever blocked or "unfriended "people from your friends list because of the untrustworthiness of the information they share?

- Yes
- No

15, Which of the following is your most important news source? /

- TV
- News paper
- Tweeter
- Face book
- Websites
- Other

16, How much trust do you have in social media as a source of news? In a scale of 0 to 5

(5 if you fully trust them and 0 if you don't trust them at all).

- 0
- 1
- 2
- 3
- 4
- 5

17, How long have you been using social sites? (Example- 3 years)

18, What is your field of Study?

19, Do you forward/share any content that you do not fully trust?

- Yes
- No

20, Which of the following do you need to trust to a social media content? (you can select multiple) Please also order these criteria from the most important to the least.

- The source is known and well reputed by you
- High number times the content is liked, shared and forwarded
- Verified by conventional media
- Verified by friends and colleagues
- Common sense or your intuition

21, Do you have any other criteria that you need to

 ISSN: 2454-4248
375 – 382

trust to a social media content?

22, Which of the following make you NOT trust to social media content? (You can select multiple) Please also order these criteria from the most important to the least.

- Denial by the government or a governmental organization
- Denial by a trusted nongovernmental organization
- Denial by the subject of the content
- Number of denying social media content
- Inconsistent social media content
- Inconsistent conventional media content
- Bad reputation of the source
- Common sense/your intuition

23, Do you have any other criteria that makes you NOT trust to a social media content?

3.1 Results of Survey

In this section we will explain the results we get from the questionnaire and the focus group. This questionnaire was sent out via Google form and distributed to participants by Face book, email and other social media network; as a result a response from 56 participants was acquired.

The majority, 61.1 % of the participants was Male and 38.9 % of the participants were Female, The youngest age 15 and the oldest 49.

Result 1

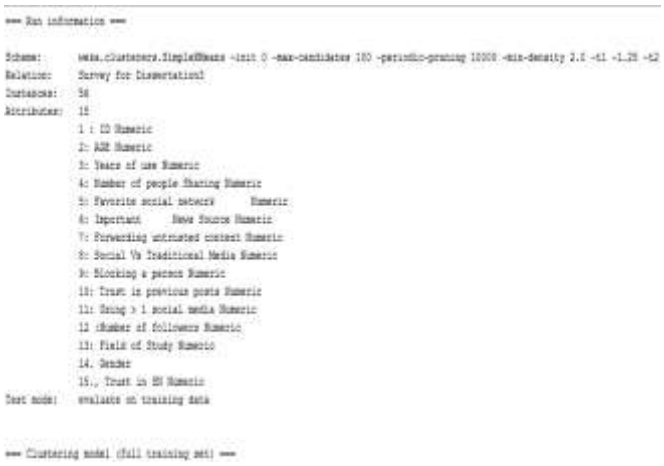
This experiment was performed for K=2, with default values of seed and distance function. Every one of the final chosen 14 attributes and 56 records were used in this experiment. For the purpose of clustering the records according to their values this model was trained by

Using the default values of the K-Means algorithm. The table below shows the outcome of the experiment and cluster distribution of the data set.

TABLE 2

The values of the parameters used for the first experiment

Cluster Result				
Distance Function	**Seed Value**	**K**	**Cluster Distribution**	
Euclidean Distance	10	2	C0	C1
			31(55%)	25(45%)



According to the above *Table 2*, we can clearly observe that the first experiment was performed with default values of the algorithm (Euclidean distance, K = 2 and Seed Value= 10).

Moreover, theoutput of the experiment exhibits us that within cluster sum of squared error is a little bit high, which leads to the fact that instances within the same cluster have a tendency to not have similarity. In order to improve this result the next experiment was done with a seed value of 100.

Result 2

The second experiment was carried out with a default K value, a default distance function (Euclidean Distance) and seed value of 50.

TABLE 3

The values of the parameters used for the second experiment

Cluster Result				
Distance Function	**Seed Value**	**K**	**Cluster Distribution**	
Euclidean Distance	50	2	C0	C1
			36(64%)	20(36%)

As in the first experiment, the result is showing us the togetherness of the clusters, "1" means all of them in that cluster share the exact same value of one, and a "0" means all of them in that cluster has a value of zero for that particular attribute. The other numbers are mostly the average value within in the clusters. Individual clusters exhibits a type of behavior in our participants, based on which we can start to draw some conclusions. In addition, we can observe each cluster visually in the same manner as

it's explained in the first experiment.

This experiment gives a much improved result in comparison with the first experimentation, the value of within clustered sum of squared error is minimized to 124.86 and also the number of iteration that the K-Means algorithm used to converge was also lowered from 5 to 4. Moreover, the number of trust claims 64 %(36) was also higher than the distrust claims 36% (20) in this experiment.

The result of this experiment looks quite satisfactory, however performing other experiments by changing the type of distance function and seed values seems quite important in case we find much better clustering model.

4. CONCLUSION

This research work presents data mining techniques can be used efficiently to model and predict trust. The outcome of this dissertation can be used to help people to make more consistent prediction of trust to social media content.

The data set used in this dissertation was gathered from my own survey, which was prepared solely for the purpose of collecting data that can be used in this study. After the data was collected, it was preprocessed and prepared in a way suitable for the data mining tasks. Then the was carried out in two sub phases, first the cluster modeling which then followed by classification modeling.

One of the main objectives of this dissertation was to conduct an experiment for observing how a person can decide on the trustworthiness of the information available in social media and to determine the significant factors that affect the trust to social media content. Some of the key findings from the dissertation are listed below:

- The effect of engaging actively in social media on the overall trust is much weaker than originally predicted.
- Previous posts quality in social media is hugely influential when it comes to trust towards future posts of a particular user.
- The traditional media outlets are still more trusted than social media sites like Face book and twitter. Websites were found to be clear favorite as the most important news source by more than half of participants of the survey.
- Women tend to trust Social network sites as most important news source than men.. In addition, participants who have been members of social networks for more than five years tend to prefer social media outlets as their most important news source in comparison with those who have been members for less than five years.
- Even though the overwhelming majority of the participants have less trust in social media outlets in relation to traditional media outlets, they are still using social media outlets as their important source of news. Websites were found to be clear favorite as the most important news source by more than half of participants of the survey.

For future research we will investigate different kinds of statistical methods to find more accurate measurement mechanism of trust and will make simulation experiments based on the findings. In this dissertation we have done a survey of 56 people of age between 15 and 50, mainly consisting of university students, so our next step is to make a survey for a larger audience which consists of people from various demographic groups.

5. REFERENCES

[1] Chen, Z. S., Kalashnikov, D. V. and Mehrotra, S. Exploiting context analysis for combining multiple entity resolution systems. In Proceedings of the 2009 ACM International Conference on Management of Data (SIGMOD'09), 2009.

[2] Kaplan, A.M. and Haenlein, M.: Users of the world unite! The challenges and opportunities of social media. Science direct, 53, 59-68, 2010.

[3] Borgatti, S P.: "2-Mode concepts in social network analysis." Encyclopedia of Complexity and System Science, 8279-8291, 2009.

[4] Cai-Nicolas Ziegler and Georg Lausen (2005) "Propagation Models for Trust and Distrusting Social Networks

[5] Carrington P. J., Scott J., and Wasserman S., 2005. "Models nd Methods in Social Network Analysis." Cambridge University Press, New York, 2005.

[6] Liu Y., Yau S., Peng D., and Yin Y., 2008. "A Flexible Trust Model for Distributed Service Infrastructures." In Proceedings of the 2008 11th IEEE Symposium on Object Oriented RealTime Distributed Computing, Orlando, USA, 108-115.

[7] Azra Shamim, Vimala Balakrishnan, Madiha Kazmi, and Zunaira Sattar, "Intelligent Data Mining in Autonomous Heterogeneous Distributed and Dynamic Data Sources", 2nd International Conference on Innovations in Engineering and Technology (ICCET'2014) Sept. 19-20, 2014.

[8] Rumi Ghosh, Sitaram Asur, "Mining Information from Heterogeneous Sources: A Topic Modeling Approach" ACM 978-1-4503-2321,2013.

[9] Prakash R.Andhale1 , S.M.Rokade2, "A Decisive Mining for Heterogeneous Data", International Journal of Advanced Research in Computer and Communication

Engineering Vol. 4, Issue 12,pp. 43-437, December 2015.

[10] Amir Ahmad, Lipika De, "A k-mean clustering algorithm for mixed numeric and categorical data" Data & Knowledge Engineering Elsevier,pp. 503-527,2007.

[11] Dr. Goutam Chakra Borty,Murali Krishna Pagolu, Analysis of Unstrucured Data: Application of Text Analytics and Sentiment Mining,2014.

[12] http://www.definitions.net/definition/trust

[13] http://changingminds.org/explanations/trust/what_is_trust.htm

[14] Ming-Syan Chen, Jiawei Han, and Philip S.Yu, "Data Mining – An Overview from Database Perspective",Knowledge and Data Engineering, IEEE Transactions on ,Volume 8 , No.6 , pp 866-883,Dec 1996.

[15] Nicholas J Belkin and W Bruce Croft,"Retrieval techniques", Annual Review of Information Science and Technology,Volume 22,pp 109-45,Information Today,1987.

[16] Romero, Cristobal, Sebastián Ventura, and Paul De Bra, "Knowledge discovery with genetic programming for providing feedback to courseware authors." Volume 14,Issue 5, pp 425-464, 2004.

[17] Jiawei Han and Micheline Kamber "Data Mining: Concepts and Techniques, 2nd ed." The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers, March 2006.

[18] Al Jarullah, A.A., "Decision tree discovery for the diagnosis of type II diabetes," Innovations in Information Technology (IIT), 2011 International Conference on , vol., no., pp.303,307, 25-27 April 2011

[19] Quinlan J (1993) Programs for Machine Learning Morgan Kaufmann Sait

[20] Bhargavi, P, & Jyothi, S. (2009). Applying Naive Bayes data mining technique for classification of agricultural land soils. International journal of computer science and network security, 9(8), 117-122.