Social Networks of Buying – Likely Patterns

M. Mohamed Sathik,
Reader in Computer Science,
Sathakathullah Appa College,
Tirunelveli. Tamilnadu. India.
mmdsadiq@gmail.com

A. Abdul Rasheed
Assistant Professor in Computer Applications,
Valliammai Engineering College
Kattankulathur. Chennai. Tamilnadu. India.
profaar@gmail.com

Abstract **Social network is the group of individuals who have common interest. Data mining has the greatest attention of research over the past decade. The Machine Learning field evolved from the broad field of Artificial Intelligence, in which the machine (computers) can be enabled to think or act as intelligent as humans. Classification is the machine learning problem in which the given dataset can be classified upon known classes. Social network has great impact in product marketing. In this paper we consider the classification as a social network problem. To prove this, we have taken a dataset and classify it using ID3 implementation. The dataset contains multivariate and multi – classes. The results obtained by classification are considered as social networks.**

Keywords: Data mining, Social Networks, Classification.

I. INTRODUCTION

Social network is the grouping of individuals into specific groups, having common interest. The grouping can also be done over the relationship that exists among the individual or even on organizations. Social networking involves grouping specific individuals or organizations together. This grouping is possible, on the basis of common or similar interest. In today's scenario, Internet is playing a key role in changing the trend from physical social networks into online social networks. For example, students studying in schools or colleges, and people working in a workplace provide opportunities for grouping the individuals as to treat as social networks.

In today's world, websites are commonly used as a platform to meet like – minded people. These websites are known as "Social Sites" or "Social Networking Sites". Depending on the commonality of interest among the users of the Internet or website, the members share a common interest like hobbies, religion etc., Today's online social networking sites provides opportunities for dating kind of service also. Some of the popular online social networking sites like MySpace, YouTube, FriendWise, FriendFinder, FaceBook and Orkut, even hi5 are extending their service in multiple services so as to promote social networking.

Data mining is the process of analyzing data form different perspectives and summarizing it into useful information. It is a process used to extract some useful hidden information from a very large database or dataset. Data mining has the greatest attention in research over the past decade. There are spectrums of techniques that can be applied to find the useful and hidden information from the database. The techniques include association rule mining, classification, clustering, neural networks, to name a few. Machine learning, computational learning and similar terms are often used in the context of data mining, to denote the application of generic model – fitting or classification algorithms for predictive data mining. Machine learning is divided into two broad areas: one as supervised learning and another one is unsupervised learning. In a supervised learning environment, the number of class labels is known. The given dataset is to be grouped over the known class labels. Hence the name "Supervised". Data mining technique – classification – falls into the category of supervised learning. When the number of class labels is unknown, and we need to classify the dataset into number of classes, then we can go for the other machine learning method called Clustering. When we do clustering, the objects in the same cluster (intra cluster) are similar, and the objects in different clusters (intra cluster) are dissimilar.

In this problem, we would like to classify the dataset into number of known classes. The group of data that falls into common classes is considered as social network. That is, the tuples that shares the same interest belongs to the same class. For this purpose, we used a variation of ID3 algorithm and its implementation in Java.

II. RELATED WORK

Social networks have greatest research interest in computer science in the recent past. Social network sites are emerging to provide support to identify people among interests in similarity. Leading social network sites like Orkut, FaceBook, MySpace and all are providing a common platform to meet such a like – minded people over the Internet through liberalization of World Wide Web. Social networks are an application of Web 2.0, as it facilitates users to host their feel. It also provides an opportunity to respond for any query or any user – related information instantly. Data mining facilitates different techniques like Association Mining, Classification, Clustering and Neural Networks to identify similarity in search to offer user – elicitation. Concept of social networks are applied in spectrum of areas includes marketing to introduce promotional options and also to identify future as well as customer prediction. Coffman (2004) demonstrated how pattern classification can be used by applying social networks for aerospace application, as a case study. Christian

978-1-4244-4711-4/09/$25.00 ©2009 IEEE

Bird (2006) discusses the general ideology of how to make use of social networks in mining e – mails. Blaz Fortuna (2007) tried to identify similarity among newsgroup messages. Social networks methods are used by Che Fu Yeh (2007) to identify e – mail intention finding mechanism. Rolf Holzer (2007) used social networks methods to identify e – mail alias detection.

III. MATERIALS AND METHODS

Social network is the individuals of like – minded people. Classification is the supervised learning method in which the given dataset will be classified according to the known classes. We have taken Car Evaluation dataset from the UCI Machine Learning Repository[1]. It is a multivariate dataset with 6 attributes and 1728 instances. As it is reported, there are no missing values found in the dataset. This data set is applicable for classification problem. There are four class variables in this dataset. We have taken the rules to be tested by using the dataset.

There are numerous classification algorithms available, like ID3, C4.5 and CART. We used Shih Data Miner tool[7] that uses Quinlan's C4.5 classification algorithm. To test the applicability of rules for classification we applied this tool. We consider various rules to get different classification. These rules (classification) forms to a particular group. That is, the classes under which the group of tuples falls in will be considered as a social network. Hence, the results obtained by using different rules are different social networks.

We used the classification method using Entropy selection method. Hence, the tool compute gain at every node, and then it splits according to the information gain of the attribute among the dataset taken for study.

The following are the different rules that we treated as patterns. The results obtained by classification are considered as individual "Social Networks".

Rule 1: buying=vhigh, maint=low, doors=4, persons=4, lugboot=med, safety=high, class=acc.
Rule 2: buying=vhigh, maint=low, doors=4, persons=4, lugboot=med, safety=high, class=unacc.
Rule 3: buying=vhigh, maint=low, doors=4, persons=4, lugboot=med, safety=high, class=good.
Rule 4: buying=vhigh, maint=low, doors=4, persons=4, lugboot=med, safety=high, class=vgood.
Rule 5: buying=high, maint=med, doors=4, persons=5+, lugboot=big, safety=high, class=acc.
Rule 6: buying=high, maint=med, doors=4, persons=5+, lugboot=big, safety=high, class=unacc.
Rule 7: buying=high, maint=med, doors=4, persons=5+, lugboot=big, safety=high, class=unacc.
Rule 8: buying=high, maint=med, doors=4, persons=5+, lugboot=big, safety=high, class=good.
Rule 9: buying=high, maint=med, doors=4, persons=5+, lugboot=big, safety=high, class=vgood.

We have taken the following minimum characteristics of attributes as patterns from the dataset: a vehicle which is most liked among the public, it should have low or medium maintenance, it can have minimum of four doors, medium or big lug boot. We considered the attribute value as high against safety. All the four Different classifiers are taken along with

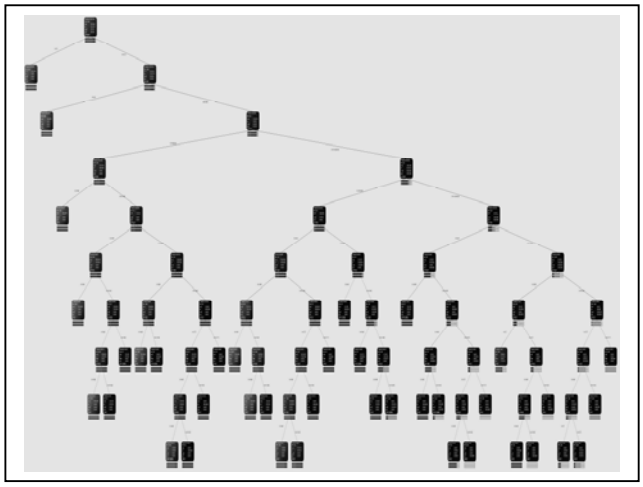

Figure 1. The Classification tree generated for the dataset

the above attribute combinations. The following Fig. 1 is the overall tree generated by the tool for the dataset we have taken for study.

During the initial setting, we selected the Entropy selection method. Hence, the tool computes the information gain at every attribute. The attribute which has maximum gain will become as a root, and it generates subsequent subtree according to the information gain obtained for the particular attribute. The node which has maximum information gain will become as a root again and then its own subsequent subtree is generated. In this way the final tree is obtained by computing the information gain for all the attributes at every level, and then the subsequent root node is identified by finding the maximum information gain for the attributes by comparison.

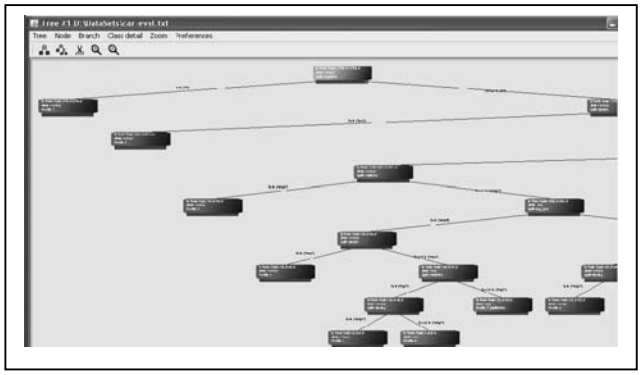

Figure 2. Node Generation by Computing Information Gain

Figure. 2 is the closer look of node generation by attribute's information gain comparison.

Finally, we are hereby producing the lift curve diagram that is generated by the tool, as shown in Figure 3.

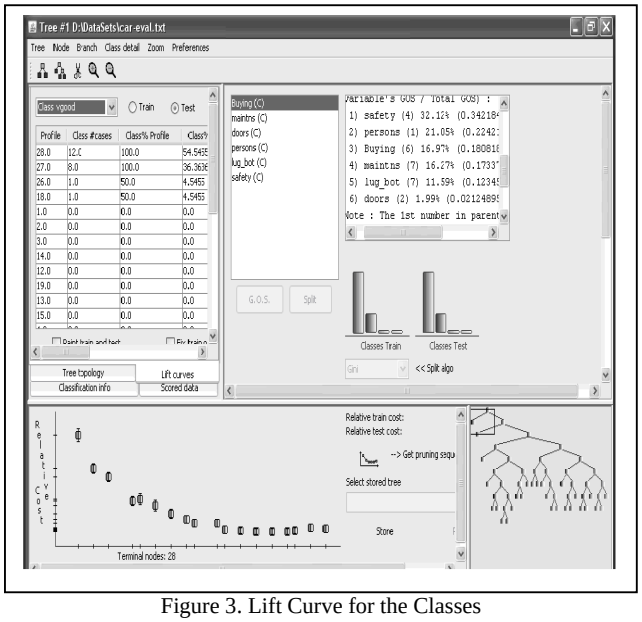

Figure 3. Lift Curve for the Classes

IV. CONCLUSION AND FUTURE WORK

Social network is the grouping of individuals having common interest. In this work, we would like to classify the car dataset that is published by UCI Machine Learning Dataset repository. In this dataset, we identified the dataset depends on customers' buying patterns. The customers who have similar interest are considered as social networks. We used Shih data miner tool for the purpose of classifying the dataset using Quinlan's C4.5 classification algorithm. In future, we would like to use other data mining technique for the same dataset to compare the result of what we obtained by the classification technique.

REFERENCES

[1] Asuncion, A. & Newman, D.J. (2007). UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.

[2] Blaz Fortuna, Eduarda Mendes Rodrigues, Natasa Milic Frayling, improving the classification of newsgroup messages through social network analysis, ACM, 2007. pp877-880.

[3] Che Fu Yeh, Ching Hao Mao, Hahn Ming Lee, Tsuhan Chen, Adaptive e- mail intention finding mechanism based on e-mail words social networks, ACM, 2007. pp 113-120.

[4] [4] Rolf Holzer, Bradley Malin, Latanya Sweeney, Email alias detection using social network analysis, ACM, 2007. pp 52-57.

[5] Christian Bird, Alex Gourley, Anand Swaminathan, Mining Email social networks, ACM, 2006, pp137-143.

[6] Dou Shen, Jian Tao Sun, Qiang Yang, Zheng Chen, Latent friend mining from blog data, Proc of Intl conf. on Data Mining, IEEE, 2006.

[7] Breiman, L., Friedman, J.H., Olshen R.A., & Stone, C.J. (1984). Classification and regression trees (pp. 112). New York: Chapman and Hall.

[8] Coffman, T.R. Marcus, S.E, "Pattern classification in social network analysis: a case study", IEEE Proceedings of Aerospace Conference, 2004, pp 3162 – 3175

[9] Jiawei Han, Micheline Kamber, "Data Mining Concepts and Techniques", 284-294.

[10] Gunnar Ratsch, "A brief introduction to machine learning", available online at http://www.tuebingen.mpg.de/~raetsch