Identifiying Key Players in online Shopping Datasets using Centrality Measures

Kavitha H.M[1]
Student,(M.Tech)
Dept of CS&E,
Adichunchangiri Institute of Technology,Chikkmagaluru

Dr. Pushpa Ravikumar[2]
B.E, M.Tech, Ph.D, LMISTE
Dept of CS&E,
Adichunchanagiri Institute of Technology, Chikkmagaluru

Varun E[3]
B.E,M.Tech.
Dept of CS&E,
Adichunchanagiri Institute of Technology, Chikkmagaluru

***Abstract* —E commerce is growing rapidly in such a way that in future everything is in fingertips of each person. No need to go outside for purchasing in these days but still the retailers have huge competition among them for marketing products and they are now interested and targeting towards the most important customers as in order to improve their business. So identifying important customers in large dataset using different tools is a challenging task. And knowing how the customers are related each other in various aspects is challenging. These challenges are addressed in this paper. These may help to increase the revenue of the retailers. Rapidly identifies the important customers in community structure.**

Keywords:-Boruta,online shopping, keycustomer, centrality measures

1. INTRODUCTION

E-Commerce is the purchasing and selling of goods and services over the internet. The technology is being developed rapidly from last two decades, especially internet and world wide development of information technology in digitalization and also being developed worldwide. After the improvement of internet technology the firms can improve the images of their product and services in their websites. The more and depth information and improved services attracts more people to purchase the products through online. Thus the traditional mode of purchasing is replaced by online shopping. Therefore the internet shopping and its impact on consumer behaviour help to increase the revenue of the retailers. There are many online shopping website in which each has competitions among online shopping sites to increase revenues. The major online shopping sites are Amazon, ebay, flipkart,Wal-Mart Online, Macy's etc.,

The fast improvement of information technology, a large volume of data is collected on many fields. The data collected may be wide variety and valuable and there are different techniques for mining the individual data. When an individual data is compared with the others data though which identifying the similar minds in a large dataset. The huge data from various fields like education field, health sector, ecommerce and many more systems. So analyzing the vast volume of data and building community to identify the interest among the different minds of the people in vast data is important in these days. Mining communities or groups in a network is valuable in analysing and decision making in many systems is important.Nowdays the main problem in community mining is optimization. None of the techniques accurately identify the central nodes in the network. The most need for such a large data is its analysis. The situation is created such a way that the data is more with us, but information is less with respect to data. So the information extraction is a challenging task for obtaining efficient information.

Community is a set of entities that splits or shares the same characteristics or connects to each other via certain relationships. Social network structure is built through nodes, represents objects from various cluster that are connected from various types of relationship. Identifying community characteristics and placing objects in different communities is a major objective of ‘community mining’ and can have different applications in many fields. Many of individual units are interconnected in a network with different conditions. Analysing the network with different interest will result in making a decision. Mining the community helps in making decision in a fields like ecommerce.

Nowadays technology is in finger tips of each people. Everyone can do their work from anywhere because technology. But the analyzing the various activities of each people through their data is much important in providing better service for each people. Here by considering individual customer as a nodes or vertices, the relationship among the customers is considered as the edges or links. For analysing those data by building a community among different customers using large set of data. Community detection and Identifying the key node in a community is important properties in network. It is carried out by considering different parameters for evaluating. Community structure consist of objects that are clustered into set of nodes that have similarities among two of nodes are more correlated if both belongs to same community otherwise two nodes are less connected. Main aim of analysing the network is to detect the community. After building a community grouping the nodes into a cluster is an important task. This is sometimes referred to be community detection. Community detection has a broad application because as the community will possess the same characteristics or properties among them. This will

help in building intelligent services to the society. These tools can be helpful in the fields of marketing, mining on social networks, trend prediction in the ecommerce fields. Once the relationship among the customers is identified in an ecommerce it is easy to predict the interest among the customers. So that it also helps in increase in the business. An optimization algorithm in which novel dynamical system employed, that detects the central node through parameter using quality function, and helps to identify the optimal communities and central nodes.

Community structure helps to identify the suspicious events that may happen in the network of telecommunication.Terrosits more often started hacking on the social networks so prediction of the terrorist groups can through community structure. Communities are referred in terms of partition of set of vertices that each of the nodes or vertices are put into one and only community just as shown in fig 1.

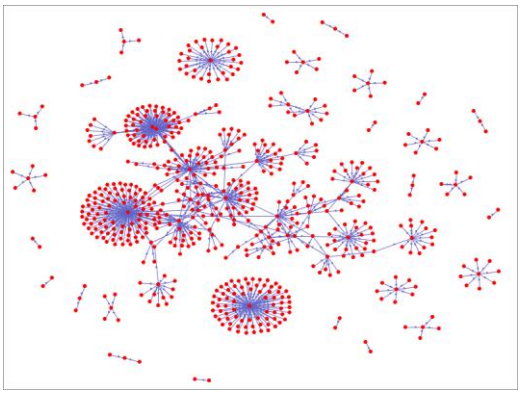


Fig 1: community structure

2.MOTIVATION

A hierarchical data structure might show many levels of grouping nodes, with tiny clusters enclosed inside giant clusters those successively giant clusters then on. Impact of social interaction of the folks contrariwise impacts of the social interaction of the folks impact the community structure. The nodes during a higher level ought to be necessary which will be called central nodes.

Today analyzing knowledge and predicting the long run is a lot of necessary for the business. Thinking one step quite customers in business square measure a necessary. Technique to search out the clump and central nodes don't offer the strict optimized result. The standard improvement strategies square measure used support d assumption that community square measure the cluster of nodes kind of like one another. The previous techniques limits like resolution limit and misidentification is comparatively are avoided. Modularity is validity indicator for the community structure of the network in community detection.

3.RELATED WORK

"Relevant feature choice"[1] could be a comparatively new sub-field within the domain of feature choice. Matter of all-relevant feature choice is initial outlined, so key algorithms help delineated. Finally the Boruta rule explains in an exceedingly bigger detail and applied each to a group of real-world knowledge sets and artificial. The algorithm is both sensitive and selective. The stage of wrongly discovered important variables is low on average less than one wrongly most important t variable is identified for each data set. The affectivity of the algorithm is 100% for data for which classification is not difficult, but may be lesser for data sets for which grouping is difficult.it is possible to increase the sensitivity of the algorithm at the cost of growth computational effort without affecting the wrong discovered level. It is achieved by random increasing the number of trees in the rfe that convey the importance estimate in " Boruta" algorithm. Different relevant feature selection algorithms are capable of discerning between relevant and non-relevant variables. The "Boruta algorithm", helped to understand the wide of ranges of data that are currently available in our day to day life.

Algorithm demonstrates especially well models for which a better models that are available by "random forest classification algorithm". By using random forest algorithm sensitivity can be improved with has more count of decision trees. Wrongly identifying of the attributes is less in case of this therefore considered significant attributes selection is best fit for generation of hefty knowledge. Only one iterative set of the rfe(random forest algorithm) that much more time for computation. Random forest algorithm in the better case it takes the more than the 1 cpu week to finish the refute random forest is mathematical calculations forces its implementation in studio or R language, it is very useful, that is not terribly economical for issues. Especially, whereas the random forest is trivially laterally its implementation is strictly consecutive. The appliance of the rule is finite just for analysis of really massive datasets delineated with over 10 or maybe a whole lot thousands variables and thousands of objects thought-about.

The centrality [2] measures are changed to apply the electrical parameters of the power. The paper presents the three different measures to measure centrality. Finding the critical nodes in a system through standard test using simulation. New research area is that "complex network" in power system. Based on electrical parameters different centrality measures are used. To identify the important nodes in a system centrality can be applied. Power flow and network impedance are new definition proposed in a system. Through examples all these types are explained. Different simulations provide that the descriptions used provided by the paper as popular for study of power grid in network complex theory. Using different simulations that are suggested can also be used to find the critical nodes in an network.

"Machine learning(ML) techiques [3]", proposes the different classifying methods in an large attributes provided in large dataset. Grouping the irrelevant the attributes is great tasking it as many scopes. Initially we cannot take any decision regarding the irrelevant data it may lead to the less accuracy. Different algorithm is also available to find the significant attributes in a system using already generated information system. One of the methods is rfe which provides the importance generated by original dataset. By comparing in an iterative manner the importance of the original attributes based on the predictor value. Randomised copies are generated.Analysizing the data using synthetic and also in biogical dataset is done but the not in commercial dataset. For the business analytics this algorithm can also be applied.

The biologically relevant results of the current study give another example to show that the selection of the important attributes can reveal important information. It opens, for example, a possibility of application of the random forest classifier as a filter for finding aptameric sequences in genes. On the other hand, the analysis of the synthetic data shows that the results of the analysis could be possibly misleading. In the system with small sample sizes and large number of attributes correlations between decisions attribute and random order of attributes may be present. That random correlation between attributes can lead to formation of wrong dependencies between attributes and decision, which are strong enough, to pass the statistical test of validity. In particular, correlations of the less important attributes with important attributes for small subsets of data are also possible. The machine learning classifier might not be able to discern such correlations from certain correlations with decision attribute. It means that for a considered data set the attributes which are non-informative by technique, might still be informative by chance. Therefore the importance of the attribute in the machine learning classifier may be used as a hint for existence of a relationship between variables of the information system and decision attribute, not as a crucial proof.

4. METHODOLOGY

Initially data is collected from the shopping websites which consist of various attributes which some are irrelevant for mining. So the attributes of customers is removed using wrapper method. For data reduction attribute subset selection method is applied. Then the preprocessed data is applied to find or identify the most or more important customers in the graph .To predict the customer relationship and predict customer more active customer in an online shopping with fast and accurate using set of parameters in an network or community structure built for customers. The raw data is collected from the online shopping websites. But the processing of the data is done for the few set of data using the function Head in R.The statistical data gathered is preprocessed using set of function in an Rstudio. The preprocessed dataset is imported into a Studio for the further processing. The manipulation of the dataset can also be done in further steps as shown in the fig 2.

Nowadays there large data set is available as improvement in technology mainly in field of machine learning, to predict the future, to understand the mind of people. Many of the shopping sites started collecting the large dataset with hundreds or even thousands of attributes values. But sometimes all these wouldn't be required for the processing. It may be difficult to predict the attribute required for processing so thus algorithms are used to identify the relevant attributes for information extraction theses is done using many algorithms such as correlation, subset elimination, random forest etc. The raw dataset of the online shopping sites is considered for the comparison of two algorithms such as Boruta and also traditional method. In which dataset consist of more than 40 attributes which consist of both categorical and also numerical attributes

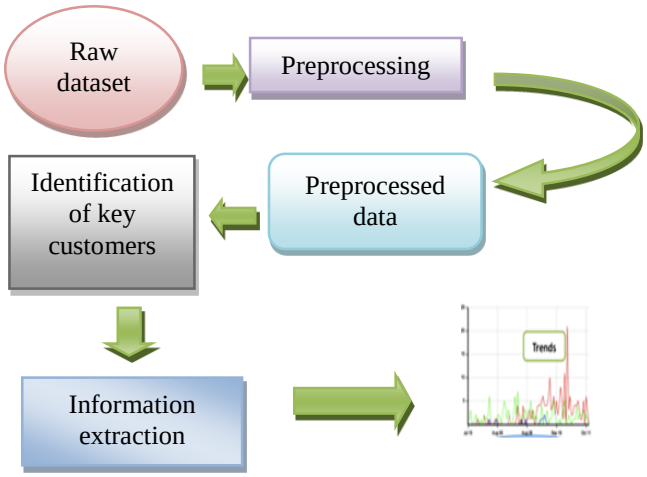


Fig 2: Architecture diagram of key customer identifying system.

4.1 Selecting important attributes

Boruta algorithm is a variable selection algorithm. It is used to select the attributes required Step wise working of Boruta algorithm:

1. The original dataset are forms shuffled copies of another attribute which is called "shadow attributes", which consist of the same objects as that of the original dataset but order is different and data is shuffled.

2. The shadow attributes that are created using the shuffled values the original dataset, shadow attribute dataset is made to find the maximum z score value. For each of the attributes maximum z score is calculated and in next step it compared with the maximum z score of the shadow attributes.

3. At each every iteration, it checks whether a real attribute has a higher value than the shadow features and statically removes features which are deemed highly unimportant.

4. Finally, the algorithm stops either when all variables get confirmed or rejected.

4.2 Identification of key customer

Initially find the correlation among the customer using set of attributes methods like Pearson is used for finding the correlation. The value is considered as weights of the edges in the network and community structure or graph is built for certain customers. For the network or the graph that is built is applied to find the key customers. Using the centrality measured in an R tool. Centrality measures consist of the betweeness, closeness, in degree, out degree, is applied to find total centrality of each node where a node represents the customers.

4.2.1 Degree

Degree centrality is used to measure of a customer's connected to the number of edges. [4] For this investigation, a customer's Degree centrality is defined in equation (1),

$$D_i = \sum_{i \neq j} a_{ij} \qquad (1)$$

4.2.2 Closeness

Closeness centrality is to find the customer's average distance to rest of the customers in the network.[4]. In this paper, closeness is defined in equation (2),

$$c_i = \frac{1}{\sum_{j \varepsilon T_J} d(i,j) + \sum_{\mu \exists T_i} |N|} \qquad (2)$$

4.2.3Betweenness

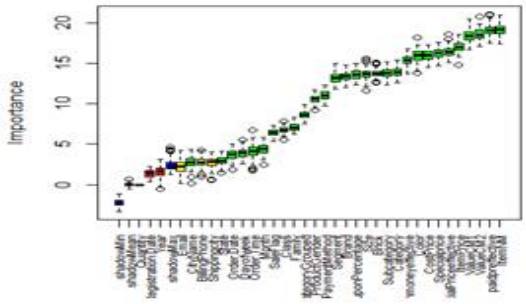
Betweenness is to find the shortest paths that pass through a one customer to another customer.[4], betweenness is defined in equation (3)

$$B_i = \frac{\sigma_{st}(i)}{\sigma_{st}} \qquad (3)$$

After calculating the centrality measures finding the total centrality. The customer with highest centrality measure is identified as the key customer. The key customer for each of the community is identified.

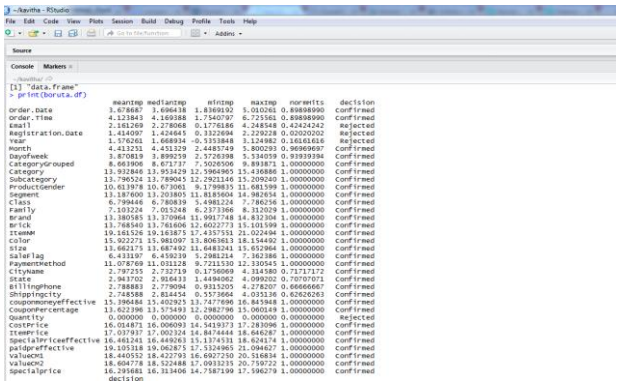
5. RESULTS AND ANALYSIS

Graph is plot based the importance of the attribute in an considered dataset.



Snapshot 3: Boxplot for attribute selection.

In an snapshot 3 boxplot is drawn based on the importance where the green color indicates the attributes which are selected ,red color indicates the attributes that are rejected, yellow attributes they are can be accepted or can be rejected.



Snapshot 4: List of attributes with calculated values.

In snapshot 4 each attributes mean important value, median important, minimum, maximum important values are displayed. Normal hits value of each attributes finally decision whether rejected or accepted.

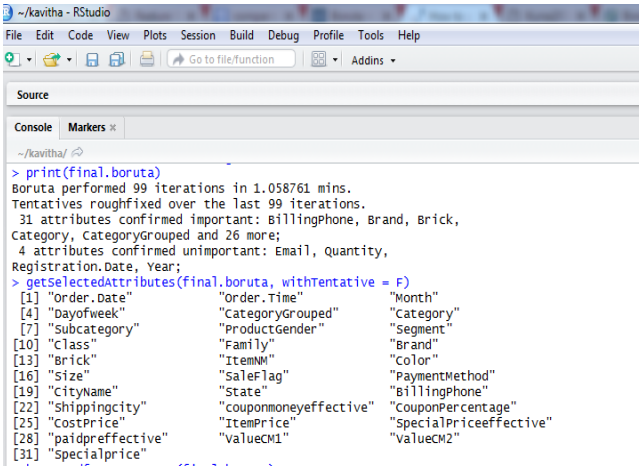
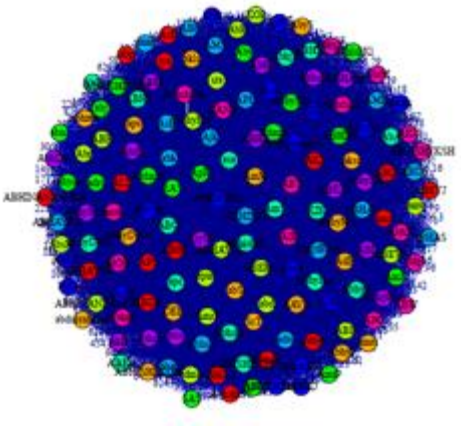


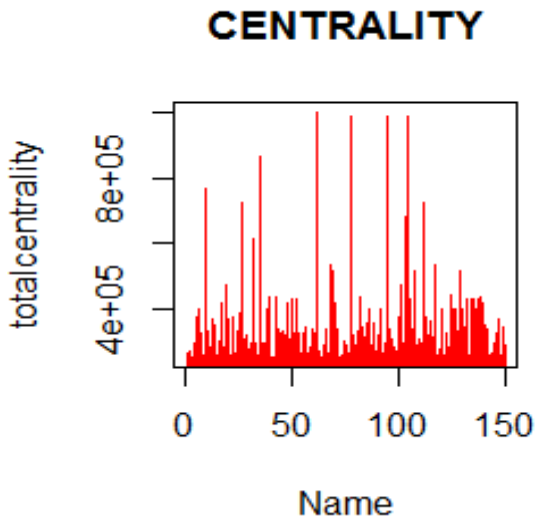
Fig 5: Snapshot of list attributes.

In snapshot 5 it is a list attributes which are finally selected. Considering the multivariate attributes in dataset. It also shows the number of iteration performed where it is 99 iteration in 1.058761 mins.31 attributes selected and 4 are rejected and list of rejected and selected attribute. Attributes listedbrick,segment,itemprice,costprice,state,cityname,month,size,family,brand,color,shipping city rest as shown in the snapshot 5.Boruta selects the important attributes less important that can be accepted or rejected.



Snapshot 6: Network for set of customers.

In a snapshot 6 shows the network structure for the customers such that correlation value are used as weights in a network or a graph. It is difficult to identify the key customer because of large customer so centrality measure is used to identify.



Snapshot 7: Graph indicating the node membership.

In snapshot 7, the graph is plot for customers' name or nodes versus centrality value. Highest centrality node is the key customer. Customers with higher node membership also have higher centrality measures.

6. CONCLUSION AND FUTURE ENHANCEMENT

The feature selection or the identifying important variable is required as if because of more number of attributes cannot be processed. So Boruta helps to identify the important variables are selected and identifies centrality based on centrality measures. In future different other approaches can also be applied to find the key customers and can be compared.

AUTHORS ACKNOWLEDGEMENT

I am thankful to my hod Dr.Puspha Ravikumar Professor & Head of the department of CS&E, A.I.T, chikmagaluru-577102 for her guidance and support. I also thank Mr.Varun E, Assistant Professor, Department of CS&E, Adichunchanagiri Institute of Technology, and Chikmagaluru-577102 for guidance,constant encouragement.

REFERENCES

[1] Witold R.Rudnicki,Mariusz 'n and Wieslaw paja,,"All relvant features selection method and applications",Springer-verlag berlin Heideberg 2015.
[2] A. B. M. Nasiruzzaman, H. R. Pota, and M. A. Mahmud , "Application of Centrality Measures of Complex Network Framework in Power Grid", 978-1-61284-972-0/11/$26.00 ©2011 IEEE.
[3] "Saikou Y. Diallo, Christopher J. Lynch, Ross Gore," Identifying key papers within a journal via network centrality measures, 1030 UniversityBoulevard, Suffolk, VA 23435, USA.Virginia Modeling Analysis and Simulation Center, Old Dominion University.
[4] Ba, S., Stallaert, J., Whinston, A. B. and Zhang, H., Choice of Transaction Channels: The Effects of Product Characteristics on Market Evolution, Journal of Management Information Systems Vol. 21, No. 4:173-198, 2005
[5] Linton C. Freeman. "Centrality in social networks: conceptual clarification. Social Networks", 1:215-239, 1979.
[6] Xi Zhaoa, Wei Dengc, Yong Shia,"Feature Selection with Attributes Clustering by MaximalInformation Coefficient", Information Technology and Quantitative Management , ITQM 2013.
[7] Miron B. Kursa, Aleksander Jankowski, Witold R. Rudnicki," Boruta – A System for Feature Selection", Fundamenta Informaticae 101 (2010) 271–285.
[8] S. P. Borgatti. The key player problem. Dynamic social network modeling and analysis: Workshop summary and papers. National Academy Press, 2003.
[9] Stephen P. Borgatti, and Martin G Everett. Network analysis of 2-mode data. Social Networks, 19(3): 243-269.
[10] F. Coulon. The use of social network analysis in innovation research: A literature review. Unpublished paper. Lund University, Lund, Sweden\ (2005).
[11] Linton C. Freeman. A set of measures of centrality based on betweenness. Sociometry (1977): 35-41.