Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

Comparative Analysis of PCA and Recursive Feature Elimination Technique for Feature Extraction in Community Mining using E-Commerce

Spoorthi C B.E., (M.Tech)[1]
Dept. of Computer Science & Engineering
Adichunchanagiri Institute of Technology
Chikkamagaluru, Karnataka, India

Dr. Pushpa Ravikumar B.E., M.Tech., Ph.D[2]
Dept. of Computer Science & Engineering
Adichunchanagiri Institute of Technology
Chikkamagaluru, Karnataka, India

***Abstract*—Advancement of innovation has brought about a critical development in every single field beginning from business to analyze. At present the examination is being completed in each division to decide the regular clients, break down their conduct in wording their buy and different components. Key player is the person who is found to make the regular buy which serves to expand the income of retailers. In this proposed work a proficient model is created by applying the Data mining techniques like preprocessing, Feature extraction, Network build for shopping dataset and Community mining. Preprocessing is completed utilizing the calculation regex and mean vector. The preprocessed information must be decreased as far as its measurement for which the element choice is connected. The calculations PCA with precision 90%, recursive component disposal with precision 72% and Karl Pearson Correlation is likewise utilized to compare the accuracy came for preprocessed data is 90.89% and accuracy came for feature extraction is 92.25%.**

Keywords*— Data Mining, Customer behaviour, E-commerce, Feature extraction, Community Mining

I. INTRODUCTION

Social network is concerned much about substantial quantities of clients acting together with some relationship. Community mining is one of the vital headings in informal organization investigation. The informal communities are various, heterogeneous and dynamic in nature. Which speaks to a specific relationship dependent on some regularly shared properties, and every sort of relationship may participate in an alternate job in a specific task.

Data mining techniques is used for shopping dataset to identify the loyal customer and to give a more benefit, promotion and to manage a community network in a superior way.

The enormous data from various fields like guidance field well being area, web business and significantly more systems. So separating the huge volume of data and building system to perceive the eagerness among the different identities of the overall public in tremendous data is fundamental in these days. Mining gatherings or network in a framework is basic for examining and fundamental administration in different structures is basic [1].

Nowadays the central issue in network mining is streamlining. None of the strategies decisively recognize the basic hub in the framework is major. The most necessities for such an enormous data is its examination. The condition is made such a way, that the data is more with us anyway information is less concerning data. So the information extraction is a trying errand for getting increasingly proficient information.

Not by try, mining or perceiving groups in immense frameworks have saved applications, since hubs in a practically identical assembling everything considered have essential properties or relationship as to center points interconnecting arranged social occasions [2]. For instance, individuals in a get-together of customers may have basic interests a get-together of stubbornly related proteins as frequently as conceivable team up for an offered sensible insistence, and tweets under a near point constantly spread the for all intents and purposes indistinguishable estimation.

At the point when the relationship among the clients is perceived in an online business [2], it is definitely not hard to foresee the energy among the customers. The main aim of this project is to identify the key player in a community. The community is build for a preprocessed data. By using an algorthim called dependency the community is build and mining techniques are used to predict the key player in a network.

II. LITERATURE SURVEY

A. Behaviour analysis of customer

Varun E [1] has proposed information mining framework is important to consider purchasing conduct of the clients in Ecommerce stores. By thought of the profitable client the examination began on complete anticipating the acquiring conduct properties of client. The information mining structure is exceptionally useful for association to begin the relationship of client with different things

B. Frame of Mining community

The Pushpa Ravikumar [2] has proposed network exchange is basic for mining adventure supportability and accomplishment. The composition exhibits a sensible association between system sponsorship and supportability. Quantifiable measurements factors have been seemed to impact organize affirmation.

III. METHODOLOGY

Data mining method is choosed for community build and for mining for the purpose of finding the loyal customer in a shopping dataset which is explained using flow chart in the figure 1.

A. Data Collection

The process of collecting a dataset from a shopping website. The data is cleaned because the so much of noise is present. By using the two algorthim it is cleaned and compared their accuracy to know which the best algorthim for cleaning process.

B. Data Preprocessing

In a community build the very important step is data preprocessing. In this dataset the noisy like regular expression, delimiter, comma, and unknown string are present in the datset. To remove this the algorthim called regex is used for cleaning and for missing value the mean weighted average vector is used to fill the values.

C. Feature Extraction

After preprocessing the feature extraction technique is used from a data mining techniques. The preprocessed data has to reduce dimension from a feature selection method. The two algorthim is used for comparison.

D. Building a Community

In this area system can be viewed as the most fundamental unsupervised learning issue alongside as each other issue of this sort it makes sense of how to find a structure in a gathering of unlabelled information showed up.

E. Community Mining

Once the community is built the next step is to find the loyal customer from a network. From each network the centrality measures is calculated to find the key player in a whole community network.

```
Algorthim: Recursive Feature Elimination
1. Train the model on the training set
2. Calculate model performance
3. Calculate variable importance that is ranking
4. for Each subset size $S_{ip,}$ i = 1……..S do
5.    Keep the $S_i$ most important variable
6.    [optional] pre-process the data
7.      train the model on the training set using predictors
8.    Calculate model performance
9.    Recalculate the ranking for each predictor
10. end
11. Calculate the performance profile over the $S_i$
12. Determine the appropriate number of predictors
13. Use the model corresponding to the original $S_i$
```

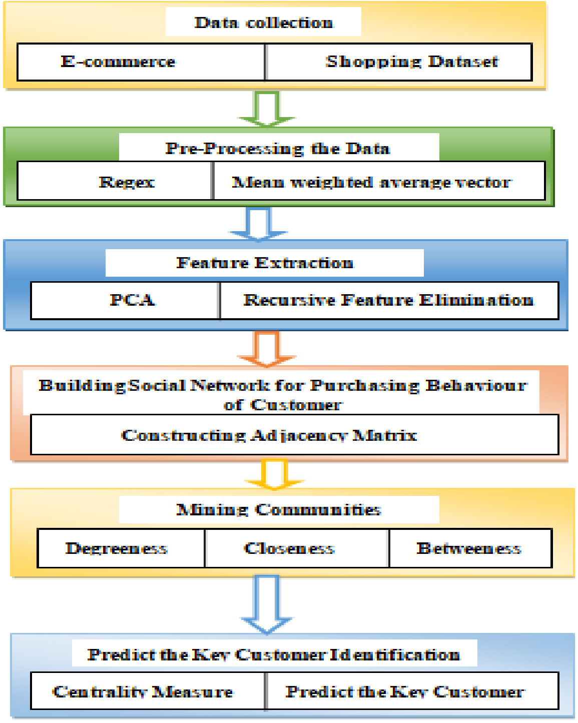

Data collection
E-commerce
Shopping Dataset
Pre-Processing the Data
Regex
Mean weighted average vector
Feature Extraction
PCA
Recursive Feature Elimination
Building Social Network for Purchasing Behaviour of Customer
Constructing Adjacency Matrix
Mining Communities
Degreeness
Closeness
Betweeness
Predict the Key Customer Identification
Centrality Measure
Predict the Key Customer

Figure 1: Community Mining for Analyzing Customer Behaviour

Table 1 depicts the contiguousness lattice of a five client which is higly associated with one another.

Table 1: Adjacency Matrix of five customer

	6	29	33	34	35
6	1.0000	-0.0030	0.0032	0.0028	0.0313
29	0.0145	1.0000	-0.6650	-0.5782	0.6024

IV. RESULT AND DISCUSSION

The snapshot gives a thought of a proposed work. These are the depictions taken when the framework is effectively executing the code.

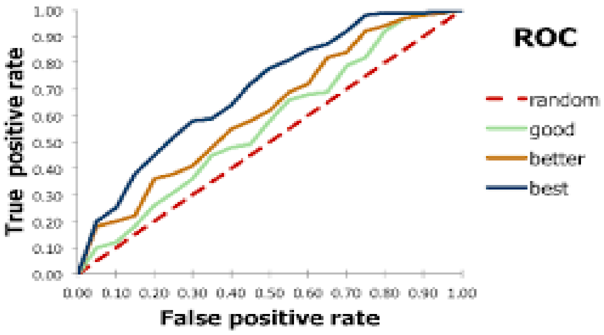


Figure 2: Raw data for preprocessing

Figure 2 describes the before preprocessed shopping raw dataset which is taken from a known shopping websites. The data consists of noise and it is present in the format excel.

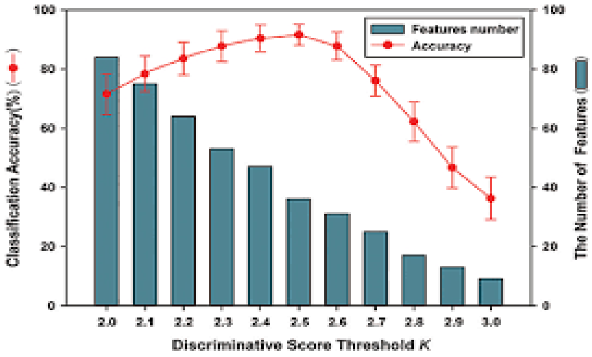


Figure 3: Feature extraction analysis graph

Figure 3 describes the preprocessed data graph which is filled with missing values. By applying the algorthim of feature extraction it gives the important attributes for community build.

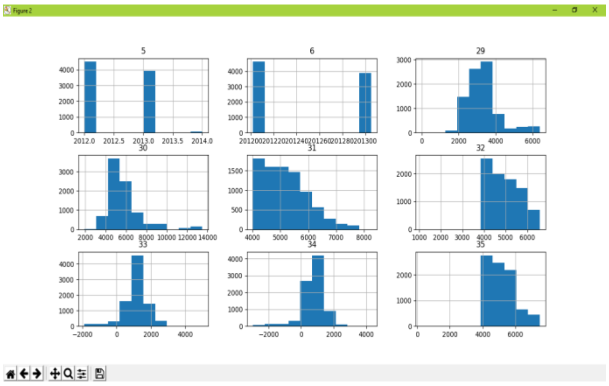


Figure 4: Histogram with the selected feature for community build

Figure 4 gives the result of a selected attributes with a rank using a histogram. The correlation matrix is obtained from a feature extraction.

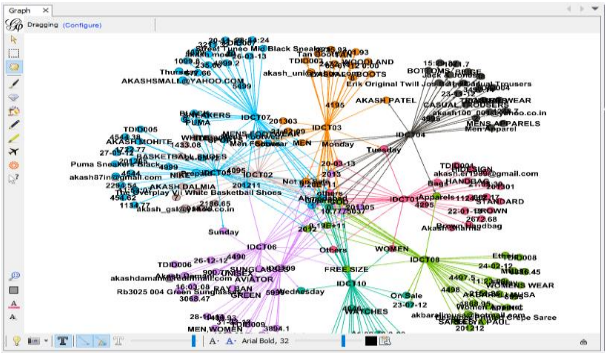


Figure 5: Community building for each network

Figure 5 present the size of a node and edges with colors. It is specifying the each community in a small network. The algorthim modularity is used for building a network.

Figure 6 provides the closeness centrality with a value of a degree, closeness, and betweeness to calculate and to average these three values to find the community.

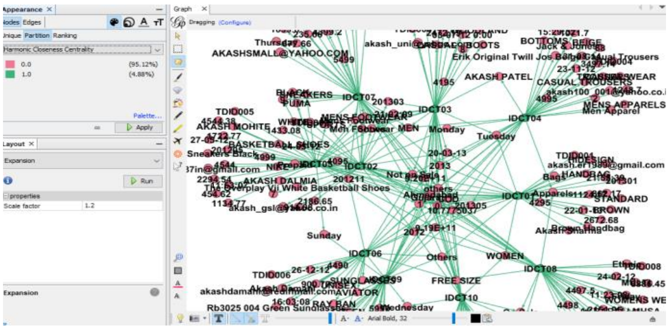
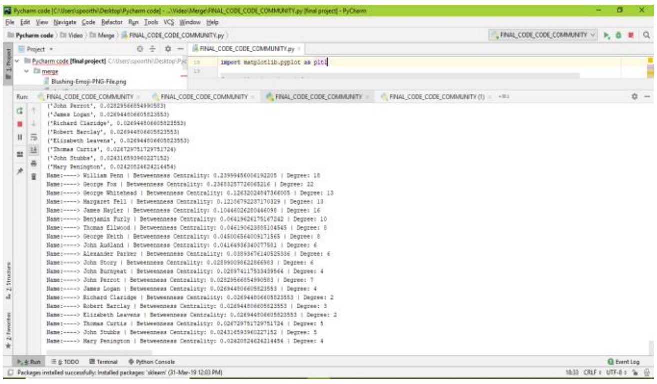


Figure 6: Centrality measure for community

Figure 7 describing the key customer from a calculation of a centrality measure using a community mining techniques. The community is mined with a key player who get more benefits from a owner of a shop.


Pycharm code [C:\Users\spoorthi\Desktop\Pycharm code] - ...\Video\Merge\FINAL_CODE_CODE_COMMUNITY.py [final project] - PyCharm
File
Edit
View
Navigate
Code
Refactor
Run
Tools
VCS
Window
Help
Pycharm code
Video
Merge
FINAL_CODE_CODE_COMMUNITY.py
Project
Pycharm code [final project]
merge
Blushing-Emoji-PNG-File.png
Run:
FINAL_CODE_CODE_COMMUNITY
4: Run
6: TODO
Terminal
Python Console
Event Log

```
import matplotlib.pyplot as plt
```

```
('John Perrot', 0.02829566854990583)
('James Logan', 0.02694480660582 3553)
('Richard Claridge', 0.026944806605823553)
('Robert Barclay', 0.026944806605823553)
('Elizabeth Leavens', 0.026944806605823553)
('Thomas Curtis', 0.026729751729751724)
('John Stubbs', 0.024316593960227152)
('Mary Penington', 0.024208246242144 54)
Name:----> William Penn | Betweenness Centrality: 0.23999456006192205 | Degree: 18
Name:----> George Fox | Betweenness Centrality: 0.23683257726065216 | Degree: 22
Name:----> George Whitehead | Betweenness Centrality: 0.12632024047366005 | Degree: 13
Name:----> Margaret Fell | Betweenness Centrality: 0.12106792237170329 | Degree: 13
Name:----> James Nayler | Betweenness Centrality: 0.10446026280446098 | Degree: 16
Name:----> Benjamin Furly | Betweenness Centrality: 0.06419626175167242 | Degree: 10
Name:----> Thomas Ellwood | Betweenness Centrality: 0.046190623885104545 | Degree: 8
Name:----> George Keith | Betweenness Centrality: 0.04500656400917156 5 | Degree: 8
Name:----> John Audland | Betweenness Centrality: 0.04164936340077581 | Degree: 6
Name:----> Alexander Parker | Betweenness Centrality: 0.03893676140525336 | Degree: 6
Name:----> John Story | Betweenness Centrality: 0.028990098622866983 | Degree: 6
Name:----> John Burnyeat | Betweenness Centrality: 0.028974117533439564 | Degree: 4
Name:----> John Perrot | Betweenness Centrality: 0.02829566854990583 | Degree: 7
Name:----> James Logan | Betweenness Centrality: 0.026944806605823553 | Degree: 4
Name:----> Richard Claridge | Betweenness Centrality: 0.026944806605823553 | Degree: 2
Name:----> Robert Barclay | Betweenness Centrality: 0.026944806605823553 | Degree: 3
Name:----> Elizabeth Leavens | Betweenness Centrality: 0.026944806605823553 | Degree: 2
Name:----> Thomas Curtis | Betweenness Centrality: 0.026729751729751724 | Degree: 5
Name:----> John Stubbs | Betweenness Centrality: 0.024316593960227152 | Degree: 5
Name:----> Mary Penington | Betweenness Centrality: 0.024208246242144 54 | Degree: 4
```

Packages installed successfully: Installed packages: 'sklearn' (31-Mar-19 12:03 PM)

Figure 7: Calculating betweeness, closeness, degreeness for loyal customer identification

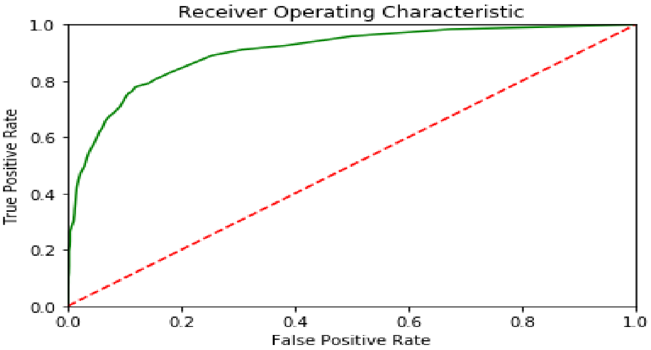


Figure 8: Pre-processed data for accuracy measure

Figure 8 provides the measure for pre-processed data to find for accuracy in a cleaning process. The confusion matrix is also built under ROC curve. The accuracy came for preprocessed data is 91.89%. The X-axis specifies the false positive rate and Y-axis specifies the true positive rate.

Figure 9 provides a brief description about the feature selection process. The accuracy for predicting the loyal customer in a E-commerce using a confusion matrix. The ROC curve gives the description about the X-axis gives the false positive and Y-axis specifies the true positive with accuracy 92.25%.

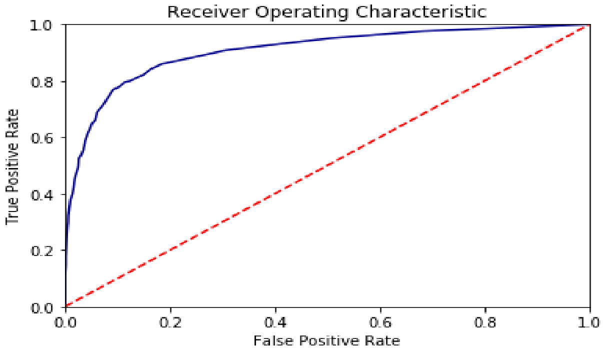


Figure 9: Feature extraction analysis graph

V. CONCLUSION

The client conduct can be broke down utilizing the network assembled utilizing the prepared dataset. The preprocessing strategy regex and mean weighted normal vector expels accentuation marks, supplant invalid an incentive with proper esteem individually. Highlight determination calculation PCA is found to give better precision of 90% when contrasted with recursive element end procedure which gives the exactness 69%. Precision and disarray grid of a pre-handled information is 91.89%. The exactness for highlight separated information is 92.25%. The centrality measure is utilized to recognize the key client in a shopping dataset which expands the income of retailers or organizations. The people group is assembled utilizing the dataset and every one of the key players in the network distinguished. The centrality measure can be additionally utilized in various areas to recognize the key player. In this proposed work the key player ID utilizing data mining is observed to be progressively proficient.

VI. ACKNOWLEDGMENT

I am highly thankful to my Hod **Dr.Pushpa RaviKumar** B.E., M.Tech., Ph.D., LMISTE, Professor & Head of the department of CS&E, Adichunchanagiri Institute of Technology, Chikkmagaluru-577102 for her consistent support & guidance all through the paper for her help regarding the usage of Comparative Analysis of PCA and Recursive Feature Elimination Technique for Feature Extraction in Community Mining using E-commerce & its aspects. Last but not the least in would like to thank all those who directly or indirectly assisted in the successful completion of the paper.

REFERENCES

[1] Varun E and Dr. Pushpa Ravikumar, “Attribute Selection for Telecommunication Churn Prediction”, International Journal of Engineering & Technology, Vol. 7, No 4.39, 2018, pp.506-509.

[2] Varun E and Dr. Pushpa Ravikumar, “Community Mining In Multi-Relational and Heterogeneous Telecom Network”, IEEE 6th International Conference on Advanced Computing (IACC-2016), DOI:10.1109/IACC.2016.15.

[3] Chun Fu Lin, Yu Hsin Hung, and Ray I Chang, “Mining Customer Behavior Knowledge to Develop Analytical Expert System for Beverage Marketing”, International Journal of Computer Trends and Technology (IJCTT) - volume4Issue4 –April 2013.

[4] Afolabi Ibukun.T, Olufunke Oladipupo, Rowland E. Worlu & Akinyemi I. O, “A Systematic Review of Consumer Behaviour Prediction Studies”, Covenant Journal of Business & Social Sciences (CJBSS), Vol. 7, No.1, June 2016.