Data Mining for Online Retail Dataset: RFM model-based Customer Clustering and Classification

WENYU ZHANG
EE Student
Engineering & Computer Science, Syracuse University
SIXUAN CHEN
EE Student
Engineering & Computer Science, Syracuse University

Abstract: With increasing competition in fields of business and commercial, efficient and precise marketing has been one of the most crucial features ensuing enterprises' profit. Comparing with rising of data mining techniques, marketing methods with data mining show their power and exploit a brand new area. Methods like sentiment analysis, customer segmentation and commercial behavior prediction have been utilized widely, especially on online market for its special ease for data mining. In this project, we present a customer-clustering commercial intelligence based on RFM model with three metrics: Recency, Frequency and Monetary. Based on the most satisfying result from clustering, we also build classification models to classify belonging clusters with his or her transaction history.

I. Introduction

With repaid increase of online retail sales, the way of shopping has been significantly changed. International and influential online shopping websites changed our lives in many ways, including transferring of the focus of marketing strategies to online retail areas.

Comparing with traditional shopping methods, online shopping is equipped with several features which make it more suitable for analysis of marketing strategies: useful records of customers' shopping history and transaction details could be stored properly and tracked accurately; customers' personal information like shipping address, contact information and favorite payment methods are associated with each customer account. Such features enable online shopping to be a approach more suitable for analysis and applications of commercial intelligence. If both online retailers and customers can treat these valuable online shopping data and records properly, it's not difficult to build a win-win relationships between customers and retailers: online retailers can utilize business intelligence to build a customer-centric shopping environment in order to keep profits, meanwhile, customers can enjoy the personalized service provided by retailers with better online shopping experience.

In order to achieve commercial intelligence, data mining methods like clustering, classification and association rules have been widely adopted in online shopping. For instance, for many online retailers in the United Kingdom and internationally alike, especially the leading companies including Amazon, Walmart, Tesco, Sainsbury's, Argos, Marks and Spencer, John Lewis, and EasyJet, data mining has now become a common practice and an integral part of the business processes in creating customer-centric business intelligence and supporting customer-centric marketing.[1][2]

In this project, in order to achieve customer-centric marketing, we will apply customer clustering based on RFM model. Then, with the results of clustering, we can perform classification to predict a certain customer's group based on his or her transaction history.

This paper is organized as follow: section II briefly introduces the data set we utilize in this project. Section III presents the preprocessing steps and main idea of RFM model. Section IV provides customers clustering based on RFM model and compares results of different methods. Section V describes classification methods we use in this project and evaluates the classification models. Finally, section VI concludes this paper.

II. Online Retail Data set

In this project, we choose *Online Retail Data Set* from UCI Machine Learning Repository. This is a transnational data set which contains all the transactions occurring between 12/01/2010 and 1/20/2011 for a UK-based and registered non-store online retail.The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.[3]
The *Online Retail Data Set* has 65499 objects(different from the information showing in the website) in total and 8 attributes shown in bottom of this page. After a brief glance of this data set, several features need to be mentioned here: firstly, the data set is presented in detail, which corresponds that each instance in this data set merely stands for one certain item of a transaction of a customer. Therefore, if we need to apply high-level data mining methodology like customer segmentation, multi-aggregation should be performed before applying algorithms or models. Then, returned transactions are provided meanwhile which will be presented as negative *UnitPrice* and *StockCode* beginning with “c”.

III. Preprocessing

Data prepropressing is a broad area and consists of a number of different strategies and techniques that are interrelated in complex ways. Roughly speaking, preprocessing methods fall into two categories:selecting data objects and attributes for the analysis or creating/changing the attributes. In both cases the goal is to improve the data mining analysis with respect to time, cost, and quality.
As mentioned in previous section, this dataset is in detailed and not suitable for high-level potential information mining, preprocessing takes a lot of time and attention in this project. Generally, preprocessing of this project consists of following steps: dealing with missing values, attribute reduction, attribute transformation, aggregation and dealing with outliers. However, preprocessing is always purposeful. In this project, the purpose of preprocessing is to build a RFM model suitable for customer clustering. Therefore, before dive into preprocessing steps, we try to provide a clear idea for RFM model first.

A. RFM Model

The RFM model was proposed by Hughes (2000). This model is popular in customer value analysis and has been widely used in measuring customer lifetime value and in

TABLE I: Attributes in Online Retail Data Set

Attribute name	Data type	Description
InvoiceNo	Nominal	Invoice number, a 6-digit integral number uniquely assigned to each transaction.
StockCode	Nominal	Product code, a 5-digit integral number uniquely assigned to each distinct product.
Description	Nominal	Product (item) name.
Quantity	Numeric	The quantities of each product (item) per transaction.
InvoiceDate	Numeric	Invice Date and time, the day and time when each transaction was generated.
UnitPrice	Numeric	Unit price, Product price per unit in sterling.
CustomerID	Nominal	Customer number, a 5-digit integral number uniquely assigned to each customer.
Country	Nominal	Country name, the name of the country where each customer resides.

customer segmentation and behavior analysis. The RFM model also has been used in several cases, especially in choosing clustering indexes. The RFM segmentation model is a model that differentiates important customers according to three variables: customers? consumption interval, frequency and amount of money: R represents "Recency", defined as the interval between most recent transaction and the present, F represents "Frequency", defined as how many times has a customer ever conducted commercial behaviors; M represents "Monetary", defined as total amount of consumption of a certain customer.

The RFM method is very effective at customer segmentation. Each customer is positioned in a three-dimensional space, corresponding to a coordinate of R, F and M. With these RFMs sorted in descending order, the groups of customers are classified proportionately.[4]

B. Preprocessing Procesure

After introducing RFM model, now we can represent our preprocessing steps. The procedure below corresponds to the processed oder in our program:

1) Missing Value: In *Online Retail Data Set*, missing values only exist in attribute *CustomerID*. By our understanding, we have no way to replace these missing values. Therefore, we merely choose to remove them.
2) Attribute Reduction: Keep in mind that our desired model has only three features: Recency, Frequency and Monetary. We can remove attributes *StockCode*, *Description* and *Country*, which are irrelevant to these 3 features.
3) Attribute Transformation: we utilize attribute transforming in two directions: amount of certain items in a transaction and recency. In raw data set, one instance only provides information about unit price and item quantity(items are always bought with quantity more than one), but no information about total amount of a certain item. We choose to create a new attribute showing this information by multiple *Quantity* with *UnitPrice* in each raw. For the recency, in raw data set, we only have attribute *InvoiceDate* showing time information which is not desirable to RFM. In order to obtain time interval data from this attribute, we use package *lubridate*. We set the date 2011-01-21 as the current date and use function "difftime" from this package to calculate time intervals as recency(all items in the same transaction share the same recency).
4) Aggregation: as we mentioned above, aggregation is the major task for preprocessing. Actually, we apply aggregation 3 times in this project: firstly we aggregate items into transaction by summing all amount values, now each row represents one transaction. Then, we create a new attribute "Frequency" with initial value equals to 1. We aggregate transactions and obtain customer data set by summing transaction amount values and Frequency. Finally, for recency feature, we aggregate the recency attribute obtained in attribute transformation step by choosing the minimal recency value and assign this recency information to the customer data set. After the final aggregation, we will notice that some monetary values are negative for that we remove the instances without *CustomerID*. We choose to set a lower threshold as 0 and replace all negative values with 0.
5) Outliers: it's obvious that some clustering methods are very sensitive to outliers. However, due to the characteristics of the raw data set, outliers can't be simply removed or replaced by mean or mode, for that these values have their meanings, especially in order. In this project, we write a new function to replace outliers with third quantile plus IQR(generally all outliers are high values). In this way, we

can keep the outliers' order in dataset and get rid of affect by outliers.

After performing the steps above, we can obtain a RFM model with only 1204 instances and 4 attributes: *CustomerID*, *Amount(Monetary)*, *Frequency* and *Recency*, which can be shown in following table. Now, this preprocessed data set is ready for clustering which will be discussed in next section.

TABLE II: Attributes in RFM model Data Set

Attribute name	Data type	Description
CustomerID	Nominal	Customer number
Amount	Numeric	Total value of consumption
Frequency	Numeric	Frequency of consumption
Recency	Numeric	Recency of consumption

IV. RFM MODEL-BASED CLUSTERING

Clustering is the process of grouping a set of physical or abstract items into classes of similar items where the groups are either meaningful or useful, or both. In this section, we will use both k-means and hierarchical algorithms for customer clustering. Then, comparisons of there different methods and parameters will be presented as well as a reasonable explanation to resultant customer clusters.

A. K-means Clustering

K-means is a well-known clustering algorithm. The basic idea of K-means is to discover k clusters, such that the records within each cluster are similar to each other and distinct from the records in other clusters. K-means is an iterative algorithm: an initial set of clusters is defined and the clusters are repeatedly updated until no further improvement is possible (or the number of iterations exceeds a specified threshold). The K-means algorithm is widely used to pre-process data or for clustering because of its simplicity and efficiency. K-means has been widely used to effectively identify the valuable customers and develop the related marketing strategies.[5] The first task for k-means algorithm is always to determine the number of cluster k. Based on our needs and normal knowledge, we expect that k falls into range [3,10]. Combining with measure Sum of the Squared Error(SSE), we can plot k-SSE line chart and the best k can be determined as the one with most rapid decrement in SSE.

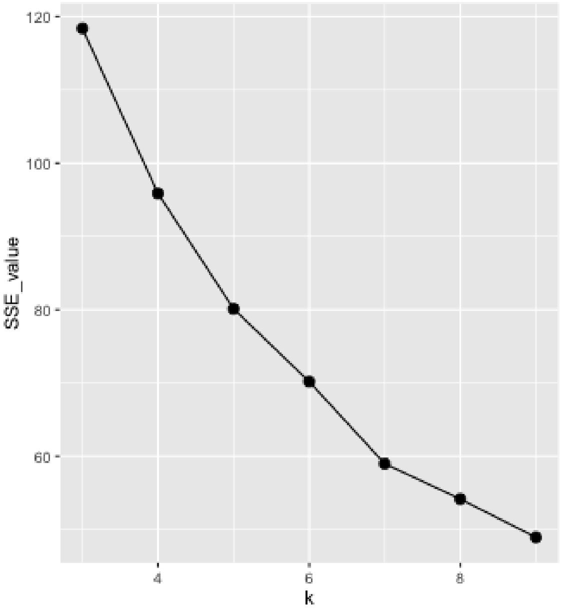

Fig. 1: k-SSE line chart

From the plot above, we can tell that when k changes from 3 to 4, we obtain the most rapid decrement in SSE, which corresponds to the slopes of lines. Therefore, we choose $k = 4$ for k-means.

Now, we can apply k-means algorithm by function *kmeans()*. Whereas, during the process of k-means clustering, we also have problem deciding whether to use normalization. In general, in order to avoid having a variable dominate the clustering result, we should apply normalization to our data set before clustering. But, the clusters derived from normalized variables are not that desirable comparing to those of non-normalized clusters. A comparison of both resultant clusters is shown below. We use min_max normalization method here and make 3 variables range in [0,100].

TABLE III: Clusters and their centroids derived from normalized variables

Cluster	Cluster Size	Amount	Frequency	Recency
1	475	16.8	0.9	83.1
2	328	28.1	12.1	16.2
3	187	41.9	44.4	74.7
4	214	67.5	84.11	20.0

TABLE IV: Clusters and their centroids derived from non-normalized variables

Cluster	Cluster Size	Amount	Frequency	Recency
1	415	121.9	1.2	32.2
2	411	347.1	1.5	30.0
3	180	689.5	2.3	19.9
4	198	1137.0	2.9	17.4

From the comparison above, we can tell that the clusters derived by non-normalized variables shares same correlations among all three features while *Recency* of normalized clusters does not have any constant correlation with other two attributes(or misorder to some extent). Therefore, we decide to use non-normalized clustering.

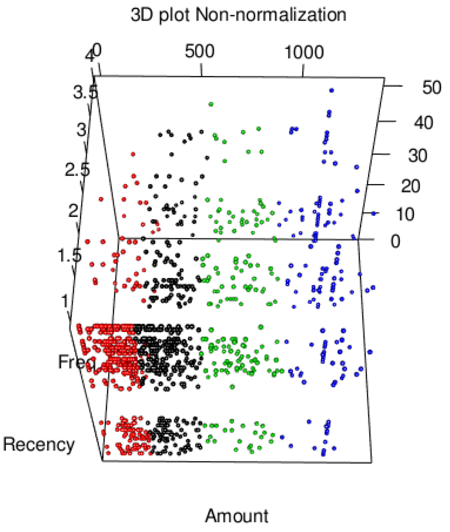

Fig. 2: K-means Clustering without Normalization

B. Agglomerative Hierarchical Clustering

Hierarchical clustering is a widely used method to obtaining hierarchical clusters, which are nested clusters that are organized as a tree. This project applies HAC algorithm, whose basic idea is: starting with individual points as clusters, successively merge the two closest clusters until only one cluster remains. A hierarchical clustering is often displayed graphically using a tree-like diagram called a dendrogram. Desired number of clusters can be derived by cutting the clustering tree at a proper height and we still $k = 4$ for hierarchical clustering.

For HAC algorithm, we have three proximity measure choices: single-linkage, complete-linkage and average-linkage. Then, we will compare hierarchical clusterings with three measures with k-means resultant clustering.

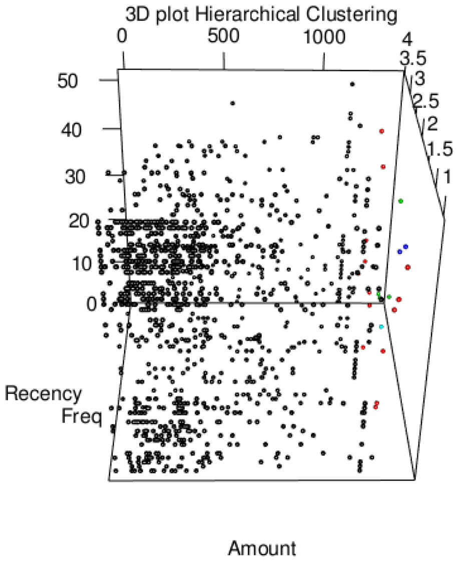

Fig. 3: Hierarchical Clustering with Single-linkage

Comparing with k-means methods, single-linkage creates a clustering who has one cluster in majority(over 95% of data size) and poor segmentation between clusters. Clearly, single-linkage is not suitable in this case.

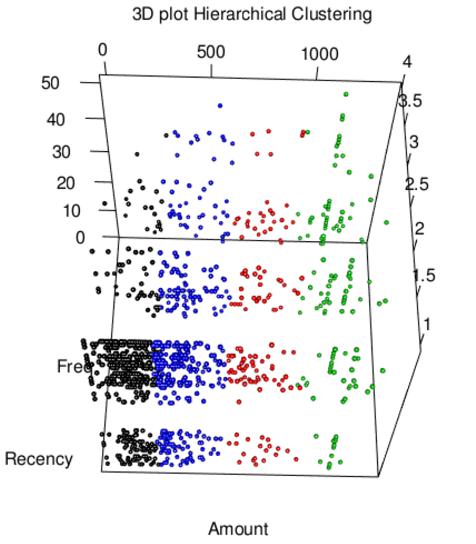

Fig. 4: Hierarchical Clustering with Average-linkage

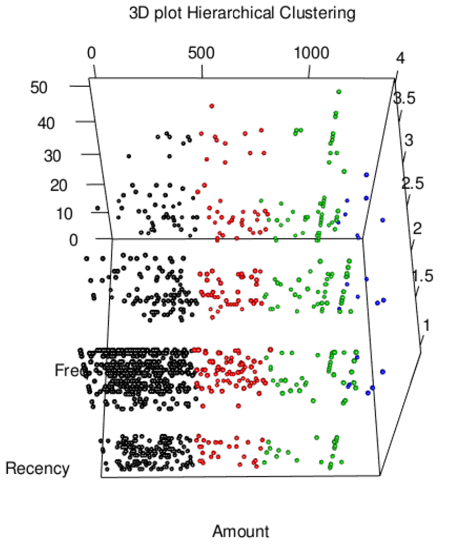

Fig. 5: Hierarchical Clustering with Complete-linkage

Hierarchical clustering using average-linkage produces clustering pretty similar to that of k-means and has good segmentation between clusters. For complete-linkage, even though it has desirable features and segmentation in 3 variables, it produces one cluster with pretty small size which does not fits the real-world application.

C. Understanding Clusters

Generating suitable and desirable clustering is merely one step in customer-centric commercial intelligence. Besides, interpreting and understanding each cluster is another crucial task, which enables us to propose accurate marketing strategies. We choose the clustering derived by k-means(similar to hierarchical clustering using average-linkage) to demonstrate.

Dive into Table IV and Figure 2, it's not difficult to figure out that each cluster actually is consisted by a group of customers sharing certain distinct and intrinsic features. We can use pie charts to compare size and consumption amount in order to better understanding the clustering.

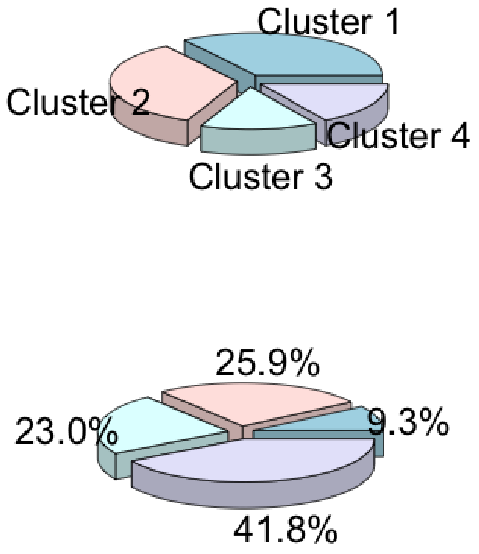

Fig. 6: Customer Clusters(upper) and Associated Consumption Amount(lower)

Cluster 1 consists of 415 customers, composed of 34.5% of the whole population. This cluster seems to be the least profitable group as it only contributes 9.3% of the whole consumption amount. Besides, this cluster has average frequency as 1.2 which shows that these customers are more likely to be one-time shoppers or visitors to this retailer.

Contrast to cluster 1, cluster 4, composed of 198 customers, contributes to 41.8% to total sales and can be regarded as most profitable group. The average frequency of this group is about 2.9 which means that customers of this cluster shopped around 3 times averagely in around 50 days, and its average recency is 17 days, so those customers more likely in recent half month. These customers can be regarded as VIP customers of these retailers.

The other two clusters have their own characteristics, but are not in the extreme. These two clusters have similar amount contributions but differs a lot in sizes. Besides, *Frequency* and *Recency* averages also reveals huge differences. Generally speaking, cluster 3 can be grouped into "sub-VIP" but cluster 2 , at most, is a group of normal customers who have some needs on this retailer but are not loyal or keen to this retailer.

V. Classification based on Customer-Clustering

In this section, we build classification model based the clustering derived from k-means method previously. Every time we have a new customer, it's not applicable to re-calculate the clustering and determine which cluster this customer belongs to, which is too ineffective and time consuming. In stead, if we build a classification model based on a trustable training data set, what we need to do is merely predicting the cluster based on the model, which will be much more efficient. We choose to use tree induction and RFM model data set as training data set, defining *Cluster* as class attribute.

A. J48 Classifier

This project chooses the J48 decision tree classifier. For this classifier, in order to predict the testing data set, firstly it needs to create a decision tree based on the attribute values of the available training data. After that, whenever it encounters a set of instances (training data set), it identifies the attribute that discriminates the various instances most clearly. This feature that is able to tell us most about the data instances so the results can have the highest information gain. Now, among the possible values of this feature, if there is any value for which there is no ambiguity, that is, for which the data instances falling within its category have the same value for the target variable, then the instances can be terminated and assigned to the cluster. Basic parameters for the classifier are set as: the minimum number of instances per leaf of the tree is 5, the confidence threshold for pruning the tree is 0.1.

B. Classifier Evaluation

We use both Hold-out method and 10-Fold Cross Validation method to evaluate the classifier. For Hold-Out method, we split the training data set into two subsets, using one subset for training and the other one for testing. The splitting ratio is around 2:1. The first 802 instances in the RFM dataset are defined as training data set, while the rest instances from 803 to 1204 are defined as testing date set.

The classify accuracy for Hold-out method is 99.7506% and for 10-Fold CV is 99.7512%. It is obvious that both of the accuracies are pretty satisfying which means the classification model we build is effective in this online retail system.

VI. Conclusion

The sections above have demonstrated the basic procedure and results of this customer-centric commercial intelligence task for online retail data set. The distinct customer clustering can help the business better understand its customers in terms of their profitability, and accordingly, adopt appropriate marketing strategies for different consumers.

Further research is required to solve limitations underlying in this project. Firstly, the raw data for classification is well-processed which makes the resultant model less applicable for practice. Secondly, due to incompleteness of data set, problems like existing of negative values remaining in RFM model for clustering were dealt naively. Thus, more suitable and elaborated solutions are required for those problems.

REFERENCE

[1] Davenport, T.H. (2009) “Realizing the Potential of Retail Analytics: Plenty of Food for Those with the Appetite.” Working Knowledge Report, Babson Executive Education.

[2] Fuloria, S. (2011) “How Advanced Analytics Will Inform and Transform U.S. Retail. Cognizant Reports”, July, http://www.cognizant.com/InsightsWhitepapers/How-Advanced-Analytics-Will-Inform-and- Transform-US-Retail.pdf, accessed January 2012.

[3]http://archive.ics.uci.edu/ml/datasets/Online+Retail#

[4] Zhen, You. (2015) “A decision-making framework for precision marketing”, Expert Systems with Applications 42 (2015) 3357-3367.

[5] Wei, J. T., Lee, M. C., Chen, H. K., Wu, H. H. (2013). “Customer relationship management in the hairdressing industry: An application of data mining techniques.” Expert Systems with Applications, 40(18), 7513-7518.