Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 163 (2019) 62–69

Procedia Computer Science

www.elsevier.com/locate/procedia

16th International Learning & Technology Conference 2019

Analysis of Customer Complaints Data using Data Mining Techniques

Amani Ghazzawi, Basma Alharbi

Taif University, Taif ,Saudi Arabia
University of Jeaddah, Jeddah, Saudi Arabia

Abstract

The Metropolitan Transportation Authority (MTA) is a public transportation service provider for the NewYork region. It is considered as the largest transportation network in North America, serving a population of 15.3 million people in the 5,000-square-mile area fanning out from New York City through Long Island, southeastern New York State, and Connecticut [1]. The public transportation services offered includes: subways, buses, and railroads which provide 2.73 billion trips each year to New Yorkers [1]. MTA receives and manages customers' complaints through Customer Relationship Management System (CRM) for all agencies [2]. In this work, we apply data mining techniques on a real domain problem using a data set of MTA customer feedback. We analyze customer complaints data for useful information that help companies to improve the quality of services and identify factors that lead to low levels of customer satisfaction. We then classify agencies based on the different areas of received customer complaints. The results of our work can be used to aid decision makers in MTA to focus on areas which needs more improvement.

Keywords: Data mining; Public services; Public transportation; Customer complaints; Customer relationship management.

* Corresponding author. Tel.: +966127272020.
E-mail address: Amghazzawi@tu.edu.sa, Bmalharbi@uj.edu.sa

10.1016/j.procs.2019.12.087

1. Introduction

1.1. Background

Customer behavior and complaints is an important issue that needs to be addressed and resolved in both public and private sectors of service providers. Customer complaints can be used as an accurate measure of how successful a service is, especially with the transformation of information technology into concepts and ideas. The competitive advantage of some companies is not the service offered but rather the attention to customer complaints and resolution. In fact, there is a risk to the company when the client is silent and not complaining. When the client is faced with a problem, usually he/she has two options to go with: either grumble from the company and disconnect from them permanently or complain. Therefore, it is very important to rely on a mechanism that helps the Customer Relationship Management (CRM) division in each company to analyze and handle complaints. In this paper, we analyze customer complaints dataset from a public service provider, the Metropolitan Transportation Authority (MTA) public transportation service provider, which provides services such as subway, bus and rail.

1.2. Problem *Definition*

The transportation network (MTA) is one of the largest public transportation companies in North America. The CRM receives numerous client complaints from various agencies. Each record i in the dataset represents a single complaint filed for an agency, a_i , and described by a set of features, x_i where

$$\boldsymbol{x_i} = \{x_{i,j}\}_{j=1}^{J}, \tag{1}$$

and J is the total number of features in the dataset. The dataset, which is a collection of all the records of complaints is denoted by X, where $X \in R^{I \times J}$:

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,J} \\ \vdots & \ddots & \vdots \\ x_{I,1} & \cdots & x_{I,J} \end{bmatrix} \tag{2}$$

where I is the total number of records, i.e., complaints, in the dataset, and J is the total number of features in the dataset. Further, each record x_i belongs to an agency a_i.

Due to the large amount of complaints and the existence of multiple agencies, it is difficult to manually analyse and study these complaints individually. In this work, we:

- Study and analyse the different relations/correlations between various features in the dataset.
- Identify the main causes of complaints in each agency in order to assist decision makers in identifying areas of improvement.
- Predict the agency a_i that a record belongs to, using classification models, where the objective is to obtain an accurate mapping of the function:

$$a_i = f(\boldsymbol{x_i}) \tag{3}$$

Table 1. Summary of related work 3

Reference	Service Sector	Data Mining Technique
Birim et al [3]	Airline	Clustering of customer complaints
Chugani et al [4]	Banks	Clustering of customer complaints
Xu et al [5]	Mobile	Clustering of customer complaints
Hsiao et al [6]	Restaurants	Classification of customer complaints given restaurants context attributes
Yang et al [7]	Telecommunication	Clustering of customer complaints

The contribution of this work is in the application of data mining and machine learning tools in the public service sector, with an overall objective to improve customer service.

2. RELATED WORK

Analysis of customer complaints using various data mining techniques is an active area of research. Recent studies explored and analyzed a wide range of service provides including as an example airlines and banking services. The importance of these studies is mainly due to the potential benefits that would arise from understanding and identifying the factors governing these complaints and would eventually help in proposing suitable solutions to improve these services. Table I provides a summary of the top most recent work in the area of customer complaint analysis. Related work are compared with regards to the service sector they study, and the problem they are trying to tackle, and more details on each work is provided next.

Birim et al [3] suggested a model for analyzing customer complaints. The research in [3] focused on studying the impact of customer complaints on business performance in the airline sector. The study linked between the fees and quality of service and important events such as economic contraction. These variables were analyzed to predict customer complaints that affect the purchase of tickets in the future. Chugani et al [4] focused on customer complaints analysis in several banks in more than one region. The study suggested a model for identifying and solving problems using data mining. Both cluster analysis and predictive modeling have been applied to find places where complaints are frequent as well as to find what caused them. This kind of analysis and studies help to process problems, win customer loyalty, and thus to increase profits.

In another study, Xu et al [5], used K-means cluster analysis algorithm to cluster customer complaints from mobile service providers. The study referred to the formulation of explanatory notes for the complaint orders and then designing of these orders. The study conducted statistical analysis of the complaints group as well, as this provides great support to the customer complaints group and their clustering.

According to Hsiao et al [6], an experimental study was conducted for a group of restaurants in Taiwan. The study aims to analyze customer complaints to process it, as well as to forecast and improve service quality. The decision tree methodology was integrated with Six Sigma analysis tools in this study and the results indicate a decrease in the level of customer complaints.

Yang et al [7] presented a model for classifying customer complaints at a telecommunication company. Previous researches focused on decision support systems in handling complaints, while this study suggested using decision support systems based on (ER) evidential reasoning. The proposed model provides high performance in improving customer complaints handling.

Compared to related work, the research presented in this paper studies and analyzes customer complaint data from a public sector, i.e., public transportation. The overall goal of our work is to identify the main complains associated with different agencies in order to help decision makers to: 1) understand the underlying causes of these complaints, 2) propose suitable solutions, and thus 3) improve the overall quality of the public services.

3. METHODOLOGY

3.1. Data Preprocessing

The MTA Customer Feedback dataset is publicly available online* and can be obtained from data.gov, which is a federal open government data site. The customer feedback dataset was collected from 2014 to 2015 from Customer Relationship Management System (CRM) of all agency. It contains information about areas of customer services, and details of how that service was rated and recorded. In this work, the original dataset was preprocessed by removing records with missing values, eliminating features that are not important in our problem to reduce the feature space of the original dataset.

3.2. Data Analysis

The preprocessed dataset has three features, one label and 10K records. The features are: Subject Matter, Subject Detail, and Issue Detail. The label, Agency, has four distinct values: Subways, NYC Buses, Long Island Rail Road, and Metro-North Railroad in which the customer may reports complaints against. The unique values of the label are equally distributed in the dataset. That is, there exist almost the same number of records per each agency in the dataset, making this dataset a balanced one. Table 2 shows a sample from the dataset.

TABLE 2 . Sample from the MTA dataset

Agency (Label)	Subject Matter	Subject Detail	Issue Detail
Long Island Rail Road	Employees	Ticket Clerk / Station Agent	Rude / Inappropriate Language
Long Island Rail Road	Employees	Ticket Clerk / Station Agent	Rude / Inappropriate Language
Long Island Rail Road	Employees	Ticket Clerk / Station Agent	Rude / Inappropriate Language
Long Island Rail Road	Employees	Ticket Clerk / Station Agent	Rude / Inappropriate Language
Long Island Rail Road	Employees	Ticket Clerk / Station Agent	Rude / Inappropriate Language
Long Island Rail Road	Employees	Ticket Clerk / Station Agent	Rude / Inappropriate Language

To better understand the relationship between the different attributes in the dataset, the correlation coefficient matrix is computed as shown in Table III. The correlation coefficient is a statistical technique that measures the degree of association between a pair of attributes. This statistical measure produces a value between -1 and +1. A positive value for the correlation implies a positive association. In this case, values close to +1 indicates that the pair of attributes are positively correlated. Thus, it can be observed from the table that the agency and subject matter attributes are positively correlated. A negative value for the correlation, on the other hand, implies a negative or inverse association. In this case, no negative correlations exist between the main attributes in the dataset.

TABLE 3 . Correlation Matrix

Attributes	Agency	Subject Matter	Subject Detail	Issue Detail
Agency	1	0.883737	0.502644	0.172114
Subject Matter	0.883737	1	0.56234	0.124806
Subject Detail	0.502644	0.56234	1	0.056948
Issue Detail	0.172114	0.124806	0.056948	1

* https://catalog.data.gov/dataset

The correlation matrix in Table III indicates that there is a strong correlation between the agencies and the subject matter. This is further demonstrated in Figure 1 where the scatterplot illustrates the correlation between the agency attribute and the subject matter. The x-axis represents the four different agencies in the dataset and the y-axis represents the subject matter, where there is a total of 9 distinct subject matters.

The figure illustrates that there is a strong correlation between some agencies and subject matters. For example, the subways agency has a strong correlation with the travel disruption / trip problem subject matters, while the long island railroad agency is strongly correlated with the employees and busses subject matters. This simple analysis can be used to aid decision makers in identifying the main areas for improvement in each agency separately. Thus, utilizing the power of data analysis to improve public services.

Further analysis can also be done in order to provide decision makers with better and more precise identification of the issues of the complaint. As mentioned earlier, the dataset contains additional attributes such as the subject details and issue details which can be used to accurately identify the areas for improvement. Table IV lists the top three issues identified for each agency. Each issue is identified by the subject matter -> subject detail -> issue detail. For example, the results clearly indicate that for the long island railroad agency, most complaints filed for the Employee subject matter are about rude / inappropriate language. This precise identification of the issue behind the complaint help decision makers to make a better and well-informed decision on what measures to take in order to improve the quality of the service provided. This illustrates the importance of data analysis in improving public services.

TABLE 4 . Top three issues per Agency

Long Island Rail Road
Employees-> Train Conductor-> Rude / Inappropriate Language
Buses-> Bus / Vehicle - General-> Crowding
Buses-> Bus / Vehicle - General-> Add More / Not Enough
Metro-North Railroad
Employees-> Bus Operator / Driver-> Closed Door Before Customer Could Board
Employees-> Bus Operator / Driver-> By passed Requested Stop
Employees-> Bus Operator / Driver-> Abandoned Customer at Station/Stop
NYC Buses
Telephone / Website / Mobile Apps-> Website-> Information Not Available
MetroCard/Tickets/E-Zpass & Tolls-> Tickets-> Pricing / Payment / Billing Error
MetroCard/Tickets/E-Zpass & Tolls-> Tickets-> Damaged / Defective
Subways
Travel Disruption / Trip Problem-> Bus / Vehicle - General-> Late / Delay
Travel Disruption / Trip Problem-> Bus / Vehicle - General-> Failure To Make Scheduled Stop
Trains-> Horn / Whistle-> Noise

FIG .1. SCATTER PLOT OF AGENCY VS SUBJEC MATTER

3.3. Data Modeling

We adopt four classification models, one at a time, to learn a mapping from the input feature space to the output label. In our case, the input feature space is the details of the customer complaints and the output feature space is the agency name. The adopted models are: Naïve Bayes, K-Nearest Neighbors (K-NN), Random Trees, and ID3.

- Naïve Bayes is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. In simple terms, a Naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class (i.e. attribute) is unrelated to the presence (or absence) of any other feature, and computes the probability of occurrences based in this simplified assumption.
- K-Nearest Neighbors is based on learning by analogy, that is, by comparing a given test example with training examples that are similar to it.
- ID3 (Iterative Dichotomiser 3) is an algorithm used to generate a decision tree invented by Ross Quinlan. ID3 is the precursor to the C4.5 algorithm. Very simply, ID3 builds a decision tree from a fixed set of examples. The resulting tree is used to classify future samples.
- Random Trees learns decision trees using Quinlan's C4.5 algorithm but it selects a random subset of attributes before it is applied.

The overall architecture of the proposed model is shown in Figure 2, where it illustrates the architecture of the model when using the Naïve Bayes classifier. For the other adopted classifier, the same architecture is utilized, and each classifier will be used in place of Naiive Bayes.

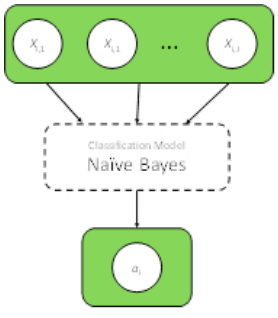

FIG.2. OVERALL ARCHITECTURE

TABLE 5. EXPERIMENTAL RESULTS

classification model	avg. accuracy + std	avg. recall + std	avg. precision + std
Naïve Bayes	91.54% +/- 9.67%	88.58% +/- 10.51%	86.56% +/- 14.76%
K-NN	87.56% +/- 10.63%	84.42% +/- 9.97%	81.50% +/- 14.32%
Random Trees	88.80% +/- 11.79%	86.92% +/- 13.19%	85.62% +/- 15.27%
ID3	95.62% +/- 7.35%	93.17% +/- 9.76%	91.04% +/- 13.05%

4. EXPERIMENTAL EVALUATION

4.1. Testing and Evaluation

In order to accurately assess the performance of each classification model, we use 10-fold cross validation. In 10-fold cross-validation, the dataset is split into 10 folds, where 9 folds are used for training the model and the remaining 1 fold is used to test the model. This process is repeated 10 times, iterating the training and testing folds. The average and standard deviation of the assessment measures are then computed and reported here for comparison.

Three assessment measures are used to compare the performance of the classification models, which are: accuracy, recall, and precision, given by the equations below respectively:

$$accuracy = \ (TP + TN\ )/(TP + TN + FP + FN) \tag{4}$$

$$precision = \ TP/(TP + FP) \tag{5}$$

$$recall = \ TP/(TP + FN) \tag{6}$$

Where TP, FP, TN, and FN are the True Positives, False Positives, True Negatives, and False Negatives respectively. The three adopted measures: accuracy, precision and recall focus on the predictive capability of the model. The accuracy measures the overall predictive capability of the model, the precision focuses on the model's ability to predict True Positives correctly with respect to False Positives, and the recall focuses on the model's ability to predict True Positives with respect to False Negatives. Together, all these three measures, along with 10-fold cross validation, provide a comprehensive assessment of the models' performance.

The four classification models (Naïve Bayes, K-NN, Random Trees, and ID3) are applied to the preprocessed dataset to learn a mapping from the input space (set of features) to the output space (label). The average of the 10 runs as well as the standard deviation is reported per each model in Table VI.

It can be observed from Table VI that ID3 has the overall highest predictive capability. For ID3, the average accuracy, recall and precision are the highest, which indicates superior predictive capability in all these three metrics. In addition to that, the standard deviation of all three performance measures for ID3 is the lowest which indicates lower variability and thus higher stability in the results of all 10 runs of the 10-fold cross validation.

The Naïve Bayes classifier has the second best results, in terms of all three performance measures. This is followed by Random Trees classifier then K-NN. K-NN has the worst results which indicates that this type of classifiers, i.e., lazy learners where no training occurs, is not a suitable option for the complaint dataset we use in

this research.

Overall, the classification models achieved the required task with good performance. The results showed that the complaint records can be classified based on their agencies, which indicates high correlation between the feature space (issue details, subject detail, and subject matter) and the output space (agency name).

5. CONCLUSION

To conclude, it is well known that analysis of public services data is a critical task because it leads to better understanding of the processes, shortcoming, and ways of improvement. For MTA, analyzing customer complaints data is important to extract the main causes of customer dissatisfaction as well as to make appropriate improvements. The analysis and data mining tasks done in this paper illustrate how intelligent computing can be used to understand and improve public services.

ACKNOWLEDGMENT

This work was supported by funding from Taif University.

REFERENCES

[1] Web.mta.info. (n.d.). MTA - Transportation Network. [online] Available at: http://web.mta.info/mta/network.htm [Accessed 18 Apr. 2017].

[2] Anon, (n.d). [online] Available at: file:///C:/Users/HP/Downloads/MTA_CustomerFeedbackRIghtNowData_Overview%20(2).pdf [Accessed 20 Apr. 2017].

[3] Birim S, Anitsal MM, Anitsal İ. "A MODEL OF BUSINESS PERFORMANCE IN THE US AIRLINE INDUSTRY: HOW CUSTOMER COMPLAINTS PREDICT THE PERFORMANCE?". Business Studies Journal. 2016 Jul 1;8(2).

[4] Chugani S, Govinda K, Ramasubbareddy S. "Data Analysis of Consumer Complaints in Banking Industry using Hybrid Clustering". In 2018 Second International Conference on Computing Methodologies and Communication (ICCMC) 2018 Feb 15 (pp. 74-78). IEEE.

[5] Xu J, Zhao J, Zhao N, Xue C, Fan L, Qi Z, Wei Q. "The Research and Construction of Complaint Orders Classification Corpus in Mobile Customer Service". In CCF International Conference on Natural Language Processing and Chinese Computing 2018 Aug 26 (pp. 351-361). Springer, Cham.

[6] Hsiao YH, Chen LF, Choy YL, Su CT. "A novel framework for customer complaint management". The Service Industries Journal. 2016 Oct 25;36(13-14):675-98.

[7] Yang Y, Xu DL, Yang JB, Chen YW. "An evidential reasoning-based decision support system for handling customer complaints in mobile telecommunications". Knowledge-Based Systems. 2018 Sep 25.