**A close up of a sign

Description automatically generated**

**CS-619 Final Project File**

**Citizen Complaint Data Mining Application**

**Megha Godwal**

**U01695574**

**05/15/2020**

**Citizen Complaint Data Mining Application**

**Abstract**

Data mining is sorting the collected data to identify patterns which provide useful information regarding the subject. Data mining involves six common tasks which are anomaly detection, association rule learning, clustering, classification, regression and summarization. Data mining can be used to study citizen complaints in the year 2016. This application focuses on the citizen complaints in Detroit. The results of this application would be helpful in detecting areas with more complaints from the citizen. To detect the type of complaints, clustering can be used to cluster different types of complaints provided in the data set. Different data mining parameters like classification, clustering, visualization, prediction can be applied on this data to generated various conclusions.

**Introduction**

In past years, it was difficult to study the complaint types in different area as computer systems were not widely used in this area. There were no data sets available which could be studied. But due to the computerised complaint data sets it is now easy to perform data mining. Using data mining on complaint data set various prediction can be made, for example, number of complaints launched in a particular time period, type of complaints by the citizens, etc. Due to use of data mining in citizen complaint data, the difficulties faced by the citizens can be predicted faster and respective work can be done to keep the citizen safe and happy.

**Dataset Description**

The main idea of the project is to develop a data mining application which is based on citizen complaint analysis. The data used for this application is a citizen complaint data of Detroit for the year 2016. This data is provided by the Detroit Police Department on the open data portal for the city of Detroit. The results of this application would be helpful in detecting areas with more complaints from the citizen.

The data consists of 2539 records with 13 attributes. The attributes contain the BPC, CCR, Report date, Entry, Age, citizen race, citizen sex, complaint closed date, Unit, allegation, Finding, officer race, officer sex. BPC stands for Board of police commission. Report date is the date when the citizen lodged the complaint. Entry is type way how the citizen lodged the complaint. Unit is the department of police where the complaint was forwarded. Allegation is description made by the citizen and finding is the actual finding of the situation by the police. The officer sex and office race are also provided in the data set.

|  |  |
| --- | --- |
| Name of the Dataset | DPD\_\_Citizen\_Complaints\_\_2016 |
| URL | https://data.detroitmi.gov/datasets/dpd-citizen-complaints |

**Problem Statement**

Governments need to consider citizen’s satisfaction an important factor and provides means for to citizen to report complaints. Data mining of the citizen’s complaints data provide the police department and the government to evaluate the citizen’s needs. This project will analyse the citizen complaints data collected by the Detroit Police Department using clustering and classification techniques it is possible to find primary causes for complaints.

**Data Mining Approach**

To detect the type of complaints, clustering can be used to cluster different types of complaints provided in the data set. Different data mining parameters like classification, clustering, visualization, prediction can be applied on this data to generate various conclusions.

**Pre-processing of Data**

Data pre-processing contains of several techniques that are useful to pre-process the data before data mining. Data pre-processing also includes techniques to handle the missing data values from the data set. There are different ways to handle missing data values. Data pre-processing also includes techniques such as sampling the data, dimensionality reduction, feature creation, feature subset selection, Normalization and Standardization. Some techniques need only categorical data and some need only nominal data. Discretization can be used to discretise the data so that the numeric values are converted in to the nominal values.

In this project, we will be using discretization to discretise the age data as there are 73 distinct values of age for better classification and clustering. Before Discretization:

A screenshot of a cell phone

Description automatically generated

After Discretization:

A screenshot of a cell phone

Description automatically generated

After Discretization the numeric attribute age is converted into nominal attribute. The age is divided in to range of ages for every data point.

**Results**

**Clustering**

Clustering is finding similar type of objects and group them together in a cluster. Clustering can be used in pattern recognition, data mining and information retrieval. There are several algorithms developed for clustering analysis. Weka also has many different types of algorithms such as EM clustering, hierarchical clustering, simple K-means clustering, etc. For this project, K-means clustering has been used to cluster the data and analyse it.

**K-Means clustering algorithm**

K-means clustering algorithm is a simple unsupervised algorithm for the clustering. The data has a fixed priority while clustering. Centroid points are selected among the data. Cluster is formed by selecting the points which are close to the centroid point. Euclidian distance is computed for each point to the centroid.

SSE =

The sum of squared error is found for every point. This error is the distance to the nearest cluster. A good clustering has low sum of squared error (SSE).

In the citizen complaint dataset, K-means clustering can be applied to cluster different types of complaints and the findings. This information is useful to find which area or department has more number of complaints in a particular period of time. There are 7 distinct findings in the dataset. Hence, 7 clusters can be formed and studied the finding can be compared with allegations and can be found which allegations are made for a particular finding.

Class attribute: Finding.

Classes to Clusters:

0 1 2 3 4 5 6 <-- assigned to cluster

180 38 83 118 105 159 14 | Not Sustained

106 50 54 35 86 23 61 | Admin. Closure

144 67 52 52 97 58 64 | No Charge

127 4 57 35 75 38 1 | Exonerated

74 2 20 24 38 31 1 | Sustained

101 6 43 35 60 32 4 | Unfounded

24 18 6 12 5 7 10 | Void

Cluster 0 <-- Exonerated

Cluster 1 <-- Void

Cluster 2 <-- Unfounded

Cluster 3 <-- Sustained

Cluster 4 <-- No Charge

Cluster 5 <-- Not Sustained

Cluster 6 <-- Admin. Closure

A picture containing screenshot

Description automatically generated

Fig. Finding vs allegation clustering

Another clustering can be done with respect to allegations. Types of allegations can be clustered and compared with the reporting date. The output is helpful in finding what allegations were made by the citizens in a particular period of a year. It can be also compared with the age and found which age group people make which type of allegations.

Class attribute: Allegation

Classes to Clusters:

0 1 2 3 4 5 6 <-- assigned to cluster

231 73 180 100 158 139 35 | Procedure

172 45 71 74 91 135 30 | Demeanor

44 18 31 31 15 23 10 | Force

14 4 8 1 9 6 4 | Entry

40 22 20 22 11 31 5 | Property

46 39 55 10 100 48 14 | Service

23 2 16 20 21 10 1 | Arrest

28 9 20 16 25 13 7 | Search

34 12 10 17 22 12 3 | Harassment

Cluster 0 <-- Procedure

Cluster 1 <-- Property

Cluster 2 <-- Force

Cluster 3 <-- Arrest

Cluster 4 <-- Service

Cluster 5 <-- Demeanor

Cluster 6 <-- Search

A screenshot of a video game

Description automatically generated

Fig. Allegations vs Age

It can be found from the plot that people of age group 0-54 make more allegations that the people of age group 54 and above.

**Classification**

Classification in data mining is a problem of identifying to which category a new data value belongs. Classification can be used for the pattern detection. There are several algorithms used for classification. Algorithms like Naïve Bayes algorithm and decision tree algorithms like C4.5 are widely used in data mining.

For this project, decision tree algorithm C4.5 is used to classify the data. In weka, Java implementation of C.4.5 is known as J48 algorithm. It gives a confusion matrix and the accuracy of the classification.

For this data set, complaints from the citizens can be classified according to their age. Citizen can also be classified according to their sex and race. There are different types of entries mentioned in the data set that citizens prefer to lodge the complaint. Some of them are walk-in some are through telephone and some are via email. So, types of entries can be classified according to the age of the citizens. This might be helpful in providing more options for the citizens which would make easier for them lodge the complaint.

For example, if most of the citizens use telephone as mean to lodge complaint, then more telephone line can be made open for the citizens. Below are few classifications applied on the data set of citizen complaints in Detroit.

Classification according to age:

=== Confusion Matrix ===

a b c <-- classified as

339 112 0 | a = Female

237 174 0 | b = Male

0 0 0 | c = Unknown

Correctly Classified Instances 513 59.5128 %

The decision tree consists of 16 leaves and the size of the tree is 31.

Number of female: 339+112 = 451

Number of male: 237+174 = 411

These are correctly classified males and females.

The tree is shown below:

A close up of a map

Description automatically generated

Fig. Decision tree classification of age and sex

From the above classification we can conclude that there are more females above age 33 who have lodged complaints.

Classification according to race:

=== Confusion Matrix ===

a b c d e f g h <-- classified as

692 0 6 3 0 0 0 0 | a = Black

20 6 1 0 0 0 0 0 | b = Other

51 0 10 1 0 0 0 0 | c = White

38 0 4 11 0 0 0 0 | d = Unknown

4 0 0 0 0 0 0 0 | e = Hispanic

6 0 0 0 0 0 0 0 | f = Arabic

8 0 0 0 0 0 0 0 | g = Biracial

0 0 0 1 0 0 0 0 | h = Asian

Correctly Classified Instances 719 83.4107 %

The tree is shown below:

A close up of a map

Description automatically generated

Fig. Decision tree classification of age and race

This classification tree shows classify the citizens according to their race. This classification might prove helpful to know which citizens face more number of problem compare to other citizens.

Classification according to the type of entry of complaint:

=== Confusion Matrix ===

a b c d e f g h i <-- classified as

147 205 0 0 0 0 0 0 2 | a = Walk In

74 330 0 0 0 0 2 0 0 | b = Telephone (or TDD)

3 21 0 0 0 0 0 0 0 | c = In-Custody

0 5 0 0 0 0 0 0 0 | d = Email

15 23 0 0 0 0 0 0 0 | e = Online

6 6 0 0 0 0 0 0 0 | f = Outside agency

0 7 0 0 0 0 4 0 0 | g = Unknown

4 1 0 0 0 0 0 0 0 | h = Telephone

4 3 0 0 0 0 0 0 0 | i = Letter

There are various types of entry types for complaints as shown in the confusion matrix above This classification as an accuracy of 55.80%.

The decision tree is as shown below:

A screenshot of a map

Description automatically generated

Fig. classification of type of entry of complaint and age

From the above classification tree we can predict that people of all ages prefer to lodge complaint through telephones. Little amount of citizens prefers to lodge complaints through walk ins and email. The prediction can be done that in next year also more number of citizens would use telephones to lodge complaints hence the number of telephone help lines should be increased so as to minimize the effort of the citizen. We can also say that, awareness about lodging the complaint online or via email should be spread among the people so that more citizens will lodge the complaints online or via email.

**Visualization**

Visualization is process of conversion of data into graphical or tabular format so that the characteristics of the data attributes are viewed and analysed properly. Visualization is a powerful tool of data exploration. It shows general trends in the data and help to find different and new patterns for analysis. Visualization also helps in detection of outliers. Visualization includes the histogram plots, pie charts, scatter plots, box plots, etc.

Weka also provides visualization tool to visualize the data attributes.

For this project visualization helps in finding some relationships among the attributes which are difficult to find by using classification and clustering.

A screenshot of a cell phone

Description automatically generated

Fig. Visualization of Allegations Vs Citizens race

This visualization of the allegations vs the ctznRace attribute shows that more allegations have been made by the citizens with Black race. Citizens with White race have made second largest allegations. The citizens with race Asian, Biracial and Arabic have the least number of allegations in the data.

A screenshot of a cell phone

Description automatically generated

Fig. Visualization of Allegations Vs Citizens sex

In the above visualization of allegations vs citizen sex, it can be observed that more Demeanor allegations have been made by males than females. Other allegations such as property and service allegations are mostly made by males.

A screenshot of a computer

Description automatically generated

Fig. Visualization of Allegations Vs Officer sex

Above visualization shows the allegation vs officer sex distribution. It shows that more male officers work on all type of allegations then the female officers.

**Conclusion**

Governments need to consider citizen’s satisfaction an important factor and provides means for to citizen to report complaints. Data mining of the citizen’s complaints data provide the police department and the government to evaluate the citizen’s needs.

This project analyses the citizen complaints data collected by the Detroit Police Department. Using clustering and classification techniques it is possible to find primary causes for complaints.

The results predict that telephone is most widely used mean to lodge complaint. The visualization results show that more number of Black as well as White race citizens lodge allegations as compared to other ethnicities. Furthermore, there is possibility of applying more data mining techniques like association rules which will help to achieve better results. If the attributes in the data are increased and some more meaning full data is gathered, then it will help in generating better results.