Exploring Classification & Clustering Techniques for Predictive Analytics

Manisha, Pawan Lathwal

Department of Computer Science, Department of Computer Science, B.M.I.E.T, Sonipat (Haryana)

ABSTRACT: Diseases have been increased tremendously. Eight person from ten have diabetes, so the data in the hospital is increasing day by day. A huge amount of data has to be handled and for that we are using classification and clustering techniques which classifies the data and forms the clusters as well as we developed a tool for Apriori which predicts the strong rules. The tool is developed using .NET framework. For Clustering and Classification WEKA tool is used for extracting the results.

Keywords: Classification, Clustering, Predictive Analytics.

I. INTRODUCTION

Data Mining is the extraction of the knowledge from large amount of data. Data Mining is a detailed process of analysing large amounts of data and picking out the relevant information. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically. There are many types of Data Mining like Clustering, Classification, and Association Rules Mining.

Classification in the data mining is to classify the data into similar and dissimilar classes. Clustering is the technique which is used for finding the clusters. Apriori is a seminal algorithm for finding frequent item sets using candidate generation. It is characterized as a level-wise complete search algorithm using anti-monotonicity of item sets, “if an item set is not frequent, any of its superset is never frequent”.

II. CLASSIFICATION

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks. Here we use the diabetic data of the hospital in CSV format. Classifications are discrete and do not imply order. Continuous, floating-point values would indicate a numerical, rather than a categorical, target. A predictive model with a numerical target uses a regression algorithm, not a classification algorithm.

The simplest type of classification problem is binary classification. In binary classification, the target attribute has only two possible values: for example, high credit rating or low credit rating. Multiclass targets have more than two values: for example, low, medium, high, or unknown credit rating.

ZeroR predicts the mean (for numeric class) or mode (for nominal class). Here ZeroR predicts class value and to build a model WEKA took 0 seconds.
a) **ZeroR:** It is the simplest classification method which relies on the target and ignores all predictors.

b) This classifier simply predicts the majority category (Class).

c) **TP Rate:** Rate of true positives (instances correctly classified as a given class).

d) **FP Rate:** Rate of false positives (instances falsely classified as a given class).

The Confusion Matrix for classification.

### III. CLUSTERING

Clustering technique finds groups of items that are similar. It divides a data set so that records with similar content are in the same group, and groups are as different as possible from each other. Cluster data using the k means algorithm. If the Manhattan distance is used, then centroids are computed as the component-wise median rather than mean.

```
This model took 0.01 second to build. Some implementations of K-means only allow numerical values for attributes. In that case, it may be necessary to convert the data set into the standard spreadsheet format and convert categorical attributes to binary. It may also be necessary to normalize the values of attributes that are measured on substantially different scales (e.g., "age" and "gender").
```

### IV. APRIORI TOOL

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database.

- **F** = (Frequent Itemsets);
- for (k=1; F_k≠Ø; k++) do begin
  - C_{k+1}=Apriori_gen(F_k); // New Candidates
  - for all transactions t of Database do begin
    - c.count++;
  end
  - F_{k+1}={C ∈ C_{k+1} | c.count>= minimum support}
end

Answer U_kF_k:

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>6</td>
</tr>
</tbody>
</table>

The tool was developed on visual studio. We are using Visual Studio .NET tool, and the coding is done in C Sharp (C#). For Front End WPF (Windows Presentation Foundation) is used. WPF is a powerful framework for building Windows applications. WPF employs XAML, an XML-based language, to define and link various interface elements.

The main window has options to add and delete transactions as well as we can set minimum support and minimum confidence on this page. If process transaction is clicked after selecting transaction, the result window will appear.

The result window will calculate strong rules by frequent items. It will find all frequent itemsets and get frequent items. Items whose occurrence in database is greater than or equal to the min.support threshold. Get frequent itemsets: Generate candidates from frequent items. Then prune the results to find the frequent itemsets and will generate strong association rules from frequent itemsets, rules which satisfy the min.support and min.confidence threshold.
We took min sup 50% and min confi 75% after taking transactions as abc, bcd, bde, ace and from this strong rules are c→b, e→d.

V. CONCLUSION

The paper shows the clustering & classification rule techniques for predictive analysis and evaluates the performance of association rule mining algorithms in the context of database partitioning. The focuses on Apriori and sampling algorithms for frequent item sets mining when the data is partitioned into a number of given segments. Apriori algorithm scans the database multiple number of times for counting the support for the item sets. The strong rules can be predicted by using this tool. We shows the clustering & classification of data on WEKA and it classified the data perfectly & shows the confusion matrix.

VI. REFERENCES

[1]. Association Rule Learning – Wikipedia, the free encyclopedia.


[8]. Huang, A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining

[9]. Shraddha, A Literature Review on Data Mining and its Techniques, Volume : 5 | Issue : 6 | June 2015 | ISSN - 2249-555X

[10]. Antonie, Application of Data Mining Techniques for Medical Image Classification, SIGKDD conference, 2001
