Chapter – 6

Results and Discussion
CHAPTER - 6
RESULTS AND DISCUSSION

The different methods were proposed for the clustering of categorical data and each method has been implemented and validated for their efficiency. The methods has been evaluated using various data sets and the details of data sets being used has been listed below.

Table 6.1: Details of data set being used

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of Data Points (N)</th>
<th>Attributes (d)</th>
<th>Attribute Values (AA)</th>
<th>Classes (K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>zoo</td>
<td>101</td>
<td>16</td>
<td>36</td>
<td>7</td>
</tr>
<tr>
<td>20 news group</td>
<td>1000</td>
<td>6084</td>
<td>12,168</td>
<td>2</td>
</tr>
<tr>
<td>Mushroom</td>
<td>8124</td>
<td>22</td>
<td>117</td>
<td>2</td>
</tr>
<tr>
<td>KddCup99</td>
<td>1,00,000</td>
<td>42</td>
<td>139</td>
<td>20</td>
</tr>
</tbody>
</table>

The Table 6.1 shows the details of data sets being used to evaluate the performance of the proposed method.

6.1. RELATIONAL GRAPH BASED CATEGORICAL DATA CLUSTERING

To improve the performance of categorical data clustering the relational graph based method has been developed and the method has been implemented and the method has been evaluated using the above mentioned data sets. The proposed method produced efficient results with various data sets. The algorithm with the following data sets were evaluated.
Figure 6.1: Snapshot of result produced by relational graph method

The Figure 6.1 shows the result of clustering produced by the proposed method and it shows that the proposed method has produces quality clustering.

Graph 6.1: The time complexity of the proposed system

The graph 6.1 shows the time taken by the proposed method and each data set has different number of data points with different number of attributes. Even the kddcup99 has 1,00,000 data points it is still has very few i.e. only 42 attributes so that
it takes less time than 20 news group with 1000 records with 6800 attributes. The proposed method takes less time than other algorithms.

Graph 6.2: The data point overlap of different algorithm

From the graph 6.2, it is very clear that the number of overlaps produced by the proposed method is very less than other algorithms at the first iteration and will be removed at the next iteration.

Graph 6.3: Comparison of clustering accuracy
The Graph 6.3 shows the comparison result of clustering accuracy produced by different methods and it shows that the proposed relational graph based method has produced more accuracy than other methods.

6.2. PROBABILISTIC APPROACH FOR CATEGORICAL DATA CLUSTERING USING CMM

The proposed Class Match Measure based multi-attribute clustering algorithm generated efficient results with various data set and various number of attributed data points. The proposed method have been evaluated with different number of attributes and data points. The iterative nature of the proposed system reduced the false indexing and overlapping of data points and the final clustering generated is more efficient than the clustering produced with earlier methods. The proposed method produced efficient results with various data sets. The algorithm have been evaluated with the following data sets.

![Probabilistic Categorical Data Clustering Using CMM](image)

**Figure 6.2: Snapshot of results produced by probabilistic approach**
The Figure 6.2 shows the snapshot of result produced by the proposed probabilistic approach for categorical data clustering and the method has produced efficient clustering than other methods.

**Graph 6.4: The time complexity of the proposed system**

The graph 6.4 shows the time taken by the proposed method and each data set has different number of data points with different number of attributes. Even the kddcup99 has 1,00,000 data points it is still has very few i.e. only 42 attributes so that it takes less time than 20 news group with 1000 records with 6800 attributes. The proposed method takes less time than other algorithms.

**Graph 6.5: The data point overlap of different algorithm**
From the graph 6.5, it is very clear that the number of overlaps produced by the proposed method is very less than other algorithms at the first iteration and will be removed at the next iteration.

Graph 6.6: Comparison of clustering accuracy

The Graph 6.6 shows the comparison of clustering accuracy produced by different methods and it shows clearly that the proposed method has produced more accuracy than other methods.

6.3. SEMANTIC ONTOLOGY BASED CATEGORICAL DATA CLUSTERING

The proposed semantic weight based multi-attribute clustering algorithm generated efficient results with various data set and various number of attributed data points. The proposed method have been evaluated with different number of attributes and data points. The iterative nature of the proposed system reduced the false indexing and overlapping of data points and the final clustering generated is more efficient than the clustering produced with earlier methods.

The proposed method produced efficient results with various data sets. The algorithm with the following data sets have been evaluated.
Figure 6.3: Snapshot of semantic categorical data clustering

The Figure 6.3 shows the snapshot of result produced by the semantic ontology based categorical data clustering method and shows that it has produced higher clustering accuracy than other methods.

Graph 6.7: The time complexity of the proposed system
The graph 6.7 shows the time taken by the proposed method and each data set has different number of data points with different number of attributes. Even the kddcup99 has 1,00,000 data points it is still has very few i.e. only 42 attributes so that it takes less time than 20 news group with 1000 records with 6800 attributes. The proposed method takes less time than other algorithms.

Graph 6.8: The data point overlap of different algorithm

From the graph 6.8, it is very clear that the number of overlaps produced by the proposed method is very less than other algorithms at the first iteration and will be removed at the next iteration.

Graph 6.9: Comparison of clustering accuracy
The Graph 6.9 shows the comparison of clustering accuracy produced by different methods and it shows clearly that the proposed method has produced more accuracy than other methods.

6.4. COMPARATIVE STUDY

The methods have produced efficient results and the efficiency of the methods has been computed by measuring various parameters.

**Entropy**

Entropy is the measure which shows the accuracy of clustering and which is measured based on the true and false classifications. The entropy of any method can be computed as follows:

\[
\text{Entropy} = \frac{TP + TN}{FP + FN}
\]

TP - True Positive

FP - False Positive

FN - False Negative

TN - True Negative

<table>
<thead>
<tr>
<th>Method</th>
<th>Clustering Accuracy %</th>
<th>False Classification Ratio %</th>
<th>Time Complexity in seconds</th>
<th>Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relational Graph</td>
<td>95.6</td>
<td>4.4</td>
<td>28</td>
<td>0.62</td>
</tr>
<tr>
<td>CMM Based</td>
<td>97.7</td>
<td>2.3</td>
<td>21</td>
<td>0.75</td>
</tr>
<tr>
<td>Semantic Weight Based</td>
<td>99.4</td>
<td>0.6</td>
<td>14</td>
<td>0.93</td>
</tr>
</tbody>
</table>

The Graph 6.2 shows the comparison of various measures of clustering accuracy and it shows clearly that the proposed method has produced more efficient results than other methods.