CHAPTER 5

PHASE 3: METRICS FOR CLUSTER QUALITY

5.1 Introduction

5.1.1 Data mining

Data mining is also called as knowledge discovery in databases (KDD), which can be defined as the extraction of implicit, previously unknown and potentially useful knowledge patterns from huge amount of data. Data mining specifically refers to extract knowledge from large amounts of data through its various branches like classification, clustering etc. Data mining techniques are used in different applications to analyse and predict the data for decision support system (Jeyabalaraja and Edwin Prabakaran (2012)).

5.1.2 Clustering

Cluster is a collection of data objects which are similar to one another within the same cluster and dissimilar to the objects in other cluster. Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters is called as Cluster analysis. Ensemble analysis improves classification accuracy and the general quality of cluster solution (Strehl and Ghosh (2012)). Clustering is a vital process to distinguish the objects according to either their shapes or characters. Clustering is used as unsupervised learning process (Raj and Singh (2010))and its goal is to discover a new set of categories which is one of the most useful tasks in the data mining process for discovering groups and identifying
interesting distributions and patterns in the underlying data (Jeyabalaraia and EdwinPrabakaran (2012)).

Thus, the main concern in the clustering process is to cluster "sensible" groups, which allow us to discover similarities and differences as well as to derive useful inferences about them. Clustering is also called as unsupervised classification.

5.1.3 Role of Metrics in Data Mining

Data mining has been the potential technology for supporting, enhancing and understanding data. The absence of the metric leads to redundancy while producing clusters. Irrespective of the research domain the efficiency of each process performance can be measured by means of metrics such that productive relationship among features can be modeled. Intelligent computing techniques such as data mining can be applied in the study of software quality by analyzing software metrics Bingbing et al (2006). Not much research is carried out in using the metrics in data mining. If technological metrics are used extensively, the knowledge discovered from the data mining tools would be of better quality. This proposed research work considers these technologies and methods that have to be taken into account for effective clustering.

5.1.4 Role of Metrics in cluster quality

The best method in predicting software quality is dependent on practical dataset. Clustering analysis technique has advantages in software quality prediction since it can be used in the case having little prior knowledge (Bingbing et al (2006)). The role of software clustering is to identify concrete entities for which a mapping decision is “easy enough” to be made automatically. The goal of clustering
was to identify related components in the software system (Mark Shtern and Vassilios Tzerpos (2012)). Effective usage of the following specification on clustering techniques leads to quality clusters.

5.2 Techniques and Issues

Irrespective of the algorithm used the following technological and issues have to be more stressed on to obtain and fine tune the results effective during clustering (Jeyabalaraja and Edwin Prabakaran (2012)).

Technology is an application of science (the combination of the scientific method and material) to meet an objective or solve a problem. For the clustering the most dominant four techniques and four issues are focused.

i. Cluster – size, validity

ii. Complexity

iii. Coupling

iv. Cohesion

In parallel to four issues

a. Cost(time involved in cluster formation, repositioning, hardware, familiarity of the mining expert and overheads)

b. Factors(quality of mining expert, experience and staff attrition ratio)

c. Quality(by considering noise and rate at which faults are found)

d. Maintainability

5.2.1 Cluster – size, validity

The cluster size may either increase or decrease according to the size of the data and the working pattern of the algorithm. The growth of the cluster should grow
gradually. In data mining, as the database contains historical data, the size of the cluster is considered to be an issue for the researchers. A well-defined mathematical model/framework is the need of the hour.

The size of each cluster after fixing the number of clusters can be obtained by using either Gompertz function or Binomial distribution. (Gupta and Kapoor (2010))

*Gompertz function*

\[ y(t) = ae^{be^{ct}} \]  

(1)

This method is applicable only when the time factor is known and it is cumbersome to estimate the size of the clusters. For computation purpose, we use logarithm

\[ \log y(t) = \log a + be^{ct} \log e \]  

(2)

*Binomial distribution*

The notations of Binomial distribution are stated hereunder.

- \( p \) – Probability of selecting a cluster
- \( q \) - Probability of not selecting a cluster
- \( n \) - Number of clusters
- \( N \) - the number of members in the graph.

Let \( p(x) \) be the probability of selecting \( x^{th} \) cluster, \( x=0,1,2,\ldots,n \).

\( N*p(x) \) is the size of the \( x^{th} \) cluster.

The probability mass function of Binomial distribution is

\[ p(x) = n_c p^x q^{n-x}, x = 0,1,2,\ldots,n \]  

(3)
To find the Size of each cluster, consider p=0.6, q=0.4, (p+q=1), n=10 and N= 2000.

Table 5a: Size of clusters

<table>
<thead>
<tr>
<th>X</th>
<th>$n_x$</th>
<th>$p^n q^{100-n}$</th>
<th>size</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>$1.048 \times 10^{-4}$</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>$1.573 \times 10^{-4}$</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>45</td>
<td>$2.359 \times 10^{-4}$</td>
<td>21</td>
</tr>
<tr>
<td>3</td>
<td>120</td>
<td>$3.539 \times 10^{-4}$</td>
<td>85</td>
</tr>
<tr>
<td>4</td>
<td>210</td>
<td>$5.308 \times 10^{-4}$</td>
<td>223</td>
</tr>
<tr>
<td>5</td>
<td>252</td>
<td>$7.963 \times 10^{-4}$</td>
<td>401</td>
</tr>
<tr>
<td>6</td>
<td>210</td>
<td>$1.194 \times 10^{-3}$</td>
<td>502</td>
</tr>
<tr>
<td>7</td>
<td>120</td>
<td>$1.792 \times 10^{-3}$</td>
<td>430</td>
</tr>
<tr>
<td>8</td>
<td>45</td>
<td>$2.687 \times 10^{-3}$</td>
<td>242</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>$4.031 \times 10^{-3}$</td>
<td>81</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>$6.047 \times 10^{-3}$</td>
<td>12</td>
</tr>
</tbody>
</table>

It is noted that if one may assume the number of clusters n and total number of members N in different levels, the researcher may get other sizes of clusters.

Cluster validity.

Evaluating and assessing the results of a clustering algorithm is known as cluster validity.

It is mainly classified into three types (Halkidi et al (2002)).

- External validity criteria,
- Internal validity criteria
• Relative validity criteria.

External validity criteria measure how much the clustering results match the prior knowledge about the data. Internal validity criteria measure how well the clustering results match the future knowledge about the data. The above two validity criteria depend on statistical testing which needs high computational demands. Cluster validity helps to identify whether the particular cluster is of good quality.

5.2.2 Complexity

Computational complexity problem is understood to be a task solved by a computer, which is equivalent to stating that the problem may be solved by mathematical techniques. Computational complexity in clusters can be solved by landmark based dimensionality reduction algorithm, which results in less cost and time. To achieve cluster quality, reduce the strength of computational complexity in clusters. In the absence of computational complexity metric, clustering the data will be too complex.

Algorithmic complexity is concerned about how fast or slow a particular algorithm performs.

Big Omega (or) "Big-O" is one of the notations to express an algorithm runtime complexity. The time function T(n) can be related with Big-O. For example, the following statement

\[ T(n) = O(n^2) \]

says that an algorithm has a quadratic time complexity.

Further the big omega can be extended to

a) Constant time O(1)

b) Linear Time O(n)
c) Logarithmic Time $O(\log n)$

d) Quadratic Time $O(n^2)$

5.2.3 Coupling

A coupled cluster consists of elements that are similar to one another and distinct from members in other clusters (Zvika Marx et al (2002)). Coupling measures the number of collaborations that a cluster has with any other clusters. Higher coupling decreases the reusability of a quality cluster. Higher coupling complicates modifications and testing. Coupling should be kept as low as possible. If this coupling metric is not done, it is not possible to categorize the members which lead to absence of quality.

Coupling can metrically measured as "low" (also "loose" and "weak") or "high" (also "tight" and "strong") and few types of coupling are listed below.

I. Pathological coupling

II. Global coupling

III. External coupling

IV. Control coupling

V. Data-structured coupling

VI. Data coupling

VII. Message coupling

VIII. Subclass Coupling

IX. Temporal coupling
The various measurements involved to estimate coupling are given below with notations.

- For data and control flow coupling
  
  \( d_{in} \) - Total number of input data parameters
  
  \( c_{in} \) - Total number of input control parameters
  
  \( d_{op} \) - Total number of output data parameters
  
  \( c_{op} \) - Total number of output control parameters

- For global coupling
  
  \( g_d \) - Total number of global variables used as data
  
  \( g_c \) - Total number of global variables used as control

- For environmental coupling
  
  \( w \) - Total number of modules
  
  \( r \) - Total number of modules calling the module under consideration

The mathematical expression for coupling is stated as

\[
C = 1 - \frac{1}{d_{in} + 2 \times c_{in} + d_{op} + 2 \times c_{op} + g_d + 2 \times g_c - (w + r)}
\]  

(4)

By considering the fixed values for the notations of the expression (4) as

\[C_{in} = 12, \quad d_{op} = 15, \quad C_{op} = 18, \quad g_d = 17, \quad g_c = 16, \quad w = 45 \quad \text{and} \quad r = 37\]

assuming different values for \( d_{in} \), the values of coupling are computed and presented in table 5a and figure5a.
Table 5b: Coupling

<table>
<thead>
<tr>
<th>$d_{in}$</th>
<th>Coupling (c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6667</td>
</tr>
<tr>
<td>3</td>
<td>0.8</td>
</tr>
<tr>
<td>5</td>
<td>0.8572</td>
</tr>
<tr>
<td>7</td>
<td>0.8889</td>
</tr>
<tr>
<td>10</td>
<td>0.9167</td>
</tr>
<tr>
<td>15</td>
<td>0.9411</td>
</tr>
<tr>
<td>20</td>
<td>0.9545</td>
</tr>
<tr>
<td>30</td>
<td>0.9688</td>
</tr>
<tr>
<td>35</td>
<td>0.973</td>
</tr>
<tr>
<td>40</td>
<td>0.9762</td>
</tr>
</tbody>
</table>

Table 5b and figure 5a reveal that the values of coupling increase for increasing the total number of input data parameter. It is observed that the least value 0.6667 indicates as low coupling and the maximum value 0.9762 reflects as high coupling.
On fixing the parameters showed and variables, the values of coupling showed an upward trend when varying the input data parameters. By increasing the values of parameters, the values of coupling must increase for different values of any other parameter or variables, after fixing the remaining parameters and variables, other values of coupling may get.

Figure 5a: Coupling
5.2.4 Cohesion

Cohesion is a qualitative measure meaning that the source code text to be measured is examined using a rubric to determine a cohesion classification. Cohesion may also be defined as the sum of the weights of all links within a cluster. The cohesion of a class is the degree to which its set of properties is part of the problem or design domain. The best cohesive cluster is formed when all elements of a cluster belong to the same category (Mesfin Silesi and Bjorn Gamback (2009)). The qualities of the clusters created by the algorithms are measured in terms of cluster cohesion.

The types of cohesion, in order of the worst to the best type, are as follows:

A. Coincidental cohesion
B. Logical cohesion
C. Temporal cohesion
D. Procedural cohesion
E. Communicational cohesion
F. Sequential cohesion
G. Functional cohesion

In the absence of cohesion metric, there is no chance for grouping the same category of members, which finally leads to less quality.

➢ Cost

The expenditure involved in collection and cleaning data, cluster formation, repositioning, hardware, familiarity of the mining expert, overheads etc., is generally referred as cost. Without cost one cannot imagine to produce quality cluster. The
selection of quality cluster with less cost is essential. The time required for the above stated factors are taken into account.

The costs and times are symbolically stated as follows:

\[ c_c - \text{cost for collection and cleaning data} \]
\[ c_f - \text{cost for formulation of clusters} \]
\[ c_h - \text{cost for hardware} \]
\[ c_w - \text{cost for mining expert} \]
\[ c_m - \text{miscellaneous costs} \]
\[ t_1 - \text{time duration of collection and cleaning the data} \]
\[ t_2 - \text{time duration of formulation of clusters} \]

and

\[ t_3 - \text{time duration of machine processing (hardware)} \]

Thus, the total cost (TC) is defined as

\[
TC = c_c t_1 + c_f t_2 + (c_h + c_w) t_3 + c_m
\]  
(5)

The cost required for the performance of the algorithm can be obtained by using the cost equation (5).

The above cost expression is a multiple linear equation in terms of different time periods. By fixing the cost constraints and time periods, the total cost is estimated. Among the three specified time periods, fix two time periods and considering the remaining one varies.
The fixed cost are considered as

\[ C_c = 550, \ C_4 = 60, \ C_h = 170, \ C_w = 350 \text{ and } C_m = 70 \]

*Table 5c: Cost Analysis \([t_2 = 0.25, t_3 = 0.60\text{hrs}]\)*

<table>
<thead>
<tr>
<th>Time (hours) (t_1)</th>
<th>Total cost (Rupees)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>672</td>
</tr>
<tr>
<td>1.0</td>
<td>947</td>
</tr>
<tr>
<td>1.5</td>
<td>1222</td>
</tr>
<tr>
<td>2.0</td>
<td>1497</td>
</tr>
<tr>
<td>2.5</td>
<td>1772</td>
</tr>
<tr>
<td>3.0</td>
<td>2047</td>
</tr>
</tbody>
</table>
The cost for cluster formulation was obtained corresponding to the time of collection and cleaning the data. They provided upward linear trend. This trend helped to estimate the cost for a long run of time.

Figure5b: Cost Analysis \([t_2 = 0.25, t_1 = 0.60\text{hrs}]\)
Table 5d: Cost Analysis \( t_1 = 1.5, t_3 = 0.60 \text{ hrs} \)

<table>
<thead>
<tr>
<th>Time (hours) ( t_2 )</th>
<th>Total cost (Rupees)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>1421.6</td>
</tr>
<tr>
<td>0.20</td>
<td>1427.6</td>
</tr>
<tr>
<td>0.30</td>
<td>1433.6</td>
</tr>
<tr>
<td>0.40</td>
<td>1439.6</td>
</tr>
<tr>
<td>0.50</td>
<td>1445.6</td>
</tr>
<tr>
<td>0.60</td>
<td>1451.6</td>
</tr>
</tbody>
</table>
The cost for clustering was obtained for different time of cluster formulation. The cost was steeply increasing. There was a chance to predict the cost for larger amount of time.

*Figure 5c: Cost Analysis  \( t_1 = 1.5, t_3 = 0.60 \text{ hrs}\)
The tables 5c, 5d, 5e and figures 5b, 5c, 5d reveal that the total cost increases by increasing time irrespective of all the three cases.
The cost for the formulation of clusters corresponding to machine processing time showed slightly upward trend. The trend helped to provide future cost when increasing the time of machine processing.

*Figure 5d: Cost Analysis  \[t_1 = 1.5, t_2 = 0.25 \text{ hrs}\]*
There are various factors while processing with issues. One of the most dominant factors is the human factor (HF) which is not that much considered but, in this research work, it is considered as the vital issue and has been focused on.

Quality of mining expert should have the attitude towards Data Mining, knowledge about the clusters, awareness of datasets, interest on studying data mining problems and knowledge to interpret results, etc. Experience of mining expert will provide accurate and concrete results in lesser time to utilize future studies. Attrition ratio describes the rate at which employees leave a company. Staff attrition ratio must be very less and it will give quality clusters. If anyone of the human factors fails, the resulting cluster will not be of good quality. HF is a function of the sum of experience, psychological measures and attrition ratio.

Let

a- Number of years of experience
b- $\varphi(\alpha, \beta, \gamma, \delta)$-psychological measures
   $\alpha$-attribute
   $\beta$-knowledge
   $\gamma$-awareness
   $\delta$-Interest
c- Attrition ratio

$$HF = f(a, b, c) = f[a + \varphi(\alpha, \beta, \gamma, \delta) + c]$$ (6)
★ Quality

Quality means the standard of something measured against other things of a similar kind or the degree of excellence of something. Number of lines of code, load time, execution time, size of program (binary), Modularity and density of the clusters are of the key focus.

To improve the above said time the efficiency of the programming skill has to be improved to obtain better results.

Irrespective of the language used for clustering Cyclomatic complexity, Halstead complexities are the statistics metrics which focus on the quality of codes.

Cyclomatic approach

The cyclomatic approach counts the number of linearly independent paths through the source code or level of confidence in the program.

\[ e = \text{Number of edges of the graph} \]
\[ n = \text{Number of nodes of the graph} \]
\[ p = \text{Number of connected components} \]

Cyclomatic Complexity \( M = e - n + 2p \)

According to the number of decision making syntaxes, measurements can be done.

Halstead approach

Halstead reflects the implementation or expression of algorithms in different languages, but it is always independent of their execution on a specific platform. The metrics are computed statistically from the code so that their goal is to identify measurable properties and the relations among them.
The difficulty measure is related to the difficulty of the program to write or understand.

Let

\[
O_1 = \text{Number of distinct operators}
\]

\[
O_2 = \text{Number of distinct operands}
\]

\[
N_1 = \text{Total number of operators}
\]

\[
N_2 = \text{Total number of operands}
\]

Program vocabulary: \( O = O_1 + O_2 \)

Program length = \( N_1 + N_2 \)

Program length = \( O_1 \log_2 O_1 + O_2 \log_2 O_2 \)

Volume = \( N \log_2 O \)

Difficulty = \( \frac{O_1}{2} \times \frac{N_2}{O_2} \)

Effort = Difficulty \times Volume

Thus, Halstead complexity would provide the tentative results, for the particular source code can be calculated by the above.

★ Maintainability

Maintainability metric can be used to maximize the cluster's long life, cluster efficiency, cluster reliability, minimize noise and detect faults. The produced cluster need not be perfect and it may fail during its operation due to time and cost. The ease with which repair and enhancement may be made to the Cluster. If the produced clusters are not maintained properly, it leads to quandary. The proposed approach will take this issue into consideration during the prototype development.
The mathematical concepts like measures of central tendencies, dispersion measures, Laplace transforms, probability distributions, failure rates and reliability measures are used to achieve maintainability. The mathematical formulation of maintainability is not fixed model but it changes according to the concern flexibility of the research work which is used.

5.3 Results and discussions

In the beginning of this chapter, some technical terms and few metrics are described. The sub divisions and components of defined metrics are presented along with their importance.

The mathematical expressions for probability distribution function, coupling and total cost are given. The probability mass function of Binomial distribution is used and computed size of clusters against the number of selected clusters. By changing the number of clusters and number of members in the graph, we may get different sets of cluster sizes.

There are several parameters involved in the expression for coupling. Except the parameter din, all other parameters are assumed as fixed and coupling values are estimated against din.

The values of coupling increase when total number of data input parameter increases. The range of coupling (0.67-1.00) is justified. The cost constraints constitute the total cost. The different times t1, t2 and t3 are considered as variables and other cost parameters are kept constant, then the total cost is estimated. In the three cases, total cost increases when the time increases. The size of clusters, coupling and total cost in one way or other reflect the cluster quality.
The remaining metrics such as complexity, cohesion, factors, quality and maintainability are described theoretically. This is the open problem to analyse numerically and assure cluster quality through these metrics.

5.4 Chapter summary

This chapter describes various types of metrics in order to estimate cluster quality. Data mining, cluster quality, role of metrics in graph clustering are explained. Various metrics are listed under two heads such as techniques and issues.

The detailed descriptions of these metrics are presented along with their divisions and components. The suitable illustrations with numerical results are given. In particular, size of cluster, coupling and total cost are numerically and graphically exhibited. Based on the above stated metrics, the required cluster quality has been measured.

5.5 Conclusion

This chapter discusses various metrics to measure the quality of cluster. The metrics are explained theoretically and few mathematical expressions are given. The illustrations are provided and the cluster quality is assessed. The cluster size and its validity are presented. By assuming fixed values for the parameters except one independent parameter which takes different values. By using standard probability distribution, the cluster sizes are computed. The coupling values against the total number of input data parameters are computed and the low and high coupling are decided.

The most important concept of any research is cost analysis. The multiple linear cost equation is used and total costs are computed for different time periods