CHAPTER 5
RESULTS AND DISCUSSIONS

5.1 INTRODUCTION

Software engineering is a complex engineering activity, which involves the interactions between people, processes and tools to develop a complete product. In this work, the Replacement Access, Library and ID Card (RALIC) dataset is used to collect the stakeholder’s requirements. RALIC is a combination of development and customization of an off-the-shelf system. The primary reason for selecting RALIC dataset is that it satisfies the following criteria:

*Large Scale* – It contains a complex stakeholder base with more than 60 stakeholder groups and 30,000 users. These stakeholders consists of various and conflicting requirements.

*Well-Documented* – RALIC is accurately documented and an external project support team documents the meeting minutes, post implementation report, project highlight reports in order to increase their objectiveness.

*Available Stakeholders* – Most of the stakeholders are available for interviews.

The dataset consists of:

- 1714 recommendations from 61 stakeholders (OpenR).
- 839 recommendations from 50 stakeholders (ClosedR).
- 439 ratings from 76 stakeholders on 10 project objectives (RateP-Obj).
- 1514 ratings from the same 76 stakeholders on 48 requirements (RateP-SReq).
- 262 ratings from 79 stakeholders on 10 project objectives (RankP-Obj).
- 469 ratings from the same 79 stakeholders on 51 requirements (rankP-Req).
- 1109 ratings from the same 79 stakeholders on 132 specific requirements (rankP-SReq).
- 276 ratings from 77 stakeholders on 10 project objectives (PointP-Obj).
- 670 ratings from the same 77 stakeholders on 45 requirements (PointP-Req).
1219 ratings from the same 77 stakeholders on 83 specific requirements (PointP-SReq).

- 410 raw textural description of requirements provided by stakeholders (Raw-requirements).

- Stakeholders and their roles (Stakeholders-and-roles).

The data preprocessing is one of the crucial steps in software engineering processes, which deals with preparation and transformation of initial datasets. Data gathering approaches are often loosely controlled, impossible data combinations (e.g., Gender: female), resulting in out-of-range values (e.g., Income: -100), missing values, etc. Analyzing data that has not been carefully separated for such problems can produce misleading results. So, it is necessary to represent and preprocess the data before running the analysis. Data preprocessing includes the following steps:

- Cleaning
- Normalization
- Transformation

5.2 FEATURE SELECTION

Feature selection has been widely used in the area of software engineering for increasing the accuracy and robustness of software generation modules. It reduces the dimensionality of datasets, minimizes the complexity and time required to reach an estimation using a particular technique. In this work, a hybrid approach is proposed for selecting the features. It takes the advantage of filter and wrapper models by developing their different evaluation criteria in various stages. The main intention of proposing a hybrid approach is to handle the large stakeholder’s requirements. Feature selection involves finding an optimum number and subset of features, which provides the most accurate prediction. A hybrid approach yields better results at the cost of high computational power and low generalization of the feature subset.

Initially, the feature selection process is performed on original datasets by using a proposed hybrid approach. Figure 5.1 depicts the comparative analysis among the various feature selection algorithms for original datasets. In this graph, the red colored bar describes the original dataset. Here, the hybrid approach is compared with the FCBC and
ReliefF algorithms. FCBC(2014) is a feature selection algorithm, which is used to remove the redundant features and keep the feature that most relevant to the class. It provides an efficient way to handle feature redundancy in feature selection. ReliefF(2010) is a supervised feature weighting algorithm of the filter model.

Figure 5.1 Comparative analysis between various feature selection algorithms

It assigns weight to a specific features based on the difference between feature values of the nearest neighbor pairs. When compared with the existing feature selection algorithms, the proposed approach provides the reduced set of selected features.

Figure 5.2 illustrates the comparative analysis among feature selection algorithms for original and preprocessed datasets. In this graph, the red colored bar describes the original dataset and the blue colored bar describes the preprocessed dataset. The selected features of original and preprocessed datasets for FCBC, ReliefF and Hybrid approach are described. In preprocessing, the irrelevant noises are removed and the data cleaning is performed on the dataset. After performing the preprocessing, a hybrid approach provides a precise reduced set of selected features compared than the other algorithms.
Table 5.1 describes the feature selection value and fuzzy classification accuracy for FCBC, ReliefF and hybrid approach algorithms. From this, it is observed that the proposed hybrid approach provides high accuracy level (90%), when compared to other algorithms.

Table 5.1 Comparison between existing and proposed algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Selected Features (%)</th>
<th>Fuzzy Classification Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCBC</td>
<td>10.5</td>
<td>85%</td>
</tr>
<tr>
<td>ReliefF</td>
<td>17%</td>
<td>75%</td>
</tr>
<tr>
<td>Hybrid Approach</td>
<td>9%</td>
<td>90%</td>
</tr>
</tbody>
</table>

From Figure 5.3, it is observed that the hybrid approach achieves better results, when compared with the FCBC and ReliefF algorithms. In this graph, the red colored bar defines the preprocessed dataset. Accuracy is defined by the overall correctness of the model that is calculated as the sum of the correct classification divided by the total number of classifications. The accuracy of the proposed method is increased by 96%, which is calculated by using:
\[
\text{Accuracy} = \frac{(TN+TP)}{(TN+TP+FN+FP)} = \frac{\text{Number of true correct assessment}}{\text{Number of all assessment}} \tag{5.1}
\]

Figure 5.3 Comparative analysis between feature selection algorithms based on fuzzy classification accuracy

Where, TP represents True Positive rate, TN represents True Negative rate, FP represents the False Positive rate and FN represents the False Negative rate.

5.3 CLASSIFICATION

After performing the feature selection process, a Fuzzy Rule Based Classification (FRBC) algorithm is applied to extract the required features. Classification is the process of automatically grouping a given set of data into separate clusters. In this analysis, the proposed FRBC classification algorithm is compared with the other existing algorithms such as, Naïve Bayesian (NB) and C4.5. The Naïve Bayesian (NB) (2014) is a descriptive and predictive type of classification algorithm, which is used to predict the class membership for a target tuples. NB requires a small amount of training data to estimate the parameters that is the merit of this classifier.

The C4.5 (2011) is a type of decision tree classification algorithm, which is the most powerful and popular method for classification. It is an inductive supervised learning system that employs decision trees to represent a quality model.

Figure 5.4 depicts the comparative analysis between the proposed and existing classification algorithms, where the red colored bar labels the original dataset and the blue colored bar labels the preprocessed dataset. The proposed FRBC algorithm provides the
high classification accuracy, when compared with the NB and C4.5 classification algorithms. Table 5.2 describes the fuzzy conditions for critical, high and medium fuzzy consequents.

Table 5.2 Conditions for Fuzzy Rule Based Classification (FRBC)

<table>
<thead>
<tr>
<th>Fuzzy Consequent</th>
<th>Fuzzy Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical</td>
<td>At least three of the clusters being evaluated have high value.</td>
</tr>
<tr>
<td>High</td>
<td>At least two of the clusters being evaluated have high value.</td>
</tr>
<tr>
<td>Medium</td>
<td>At least one of the cluster being evaluated have high value.</td>
</tr>
</tbody>
</table>

Figure 5.4 Comparative analysis between various classification algorithms

Figure 5.5 represents the comparison graph for various classification algorithms based on the accuracy. From this, it is proved that the proposed FRBC provides the highest classification accuracy rate, when compared with the other methods.
Figure 5.5 Classification accuracy for original and preprocessed datasets

Precision and recall are the basic measures, which are used mainly used to evaluate the searching strategies. Normally, precision is the proportion of retrieved material that is literally relevant to the query requirement and recall is the proportion of relevant material that is actually retrieved. Precision is a measure of the accuracy that provides a predicted class, which is defined by,

\[
\text{Precision} = \frac{TP}{(TP+FP)} \tag{5.2}
\]

Recall is a measure of the ability of a prediction model to select instances of a certain class from a data set, which is defined by,

\[
\text{Recall} = \frac{TP}{(TP+FN)} \tag{5.3}
\]

Table 5.3 Precision and Recall Values for NB, C4.5 and Fuzzy Classification

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>81%</td>
<td>82%</td>
</tr>
<tr>
<td>C4.5</td>
<td>83%</td>
<td>87%</td>
</tr>
<tr>
<td>Fuzzy Classification</td>
<td>90%</td>
<td>95%</td>
</tr>
</tbody>
</table>
Table 5.3 describes the precision and recall values for NB, C4.5 and Fuzzy classification algorithms. From this table, it is analyzed that the proposed fuzzy classification approach provides the highest precision (90%) and recall (95%) values. Figure 5.6 shows the precision and recall values for the original and preprocessed datasets. Here, the proposed FRBC method is compared with the existing NB and C4.5 classification algorithms. In this graph, it is observed that the FRBC provides the highest precision and recall values, where the red colored bar represents the precision and the blue colored bar represents the recall.

5.4 NETWORK FORMATION

In this analysis, the Bayesian network formation is applied to assign the stakeholders as nodes. It is a Directed Acyclic Graph (DAG), where the nodes represent the random variables and the directed arcs defines a causal influences or functional influences. Figure 5.7 shows the Bayesian network formation for the number of stakeholders. In this graph, the green colored nodes represent the stakeholders and link denotes the connectivity between the nodes. This network is a directed graph whose nodes represents the uncertain variables and edges defines the causal or influential links between the variables. It enables reasoning under uncertainty and combines the advantages of an intuitive visual representation with a sound mathematical basis in Bayesian probability.
5.5 CLUSTERING

Clustering is a division of data into groups of similar objects and each group is called a cluster. It is an unsupervised classification mechanism, where a set of patterns are classified into groups. The collected requirements from the stakeholders are decomposed based on the common characteristics by using a Stake Requirement Clustering Algorithm (SRCA). It encapsulates the requirements based on the similarity and association relations. Moreover, the similar characteristic clusters are grouped together based on the Jaccard similarity. Figure 5.8 illustrates the computation time of various clustering algorithms, where the proposed SRCA method is compared with the other existing clustering algorithms, such as, K-Means and Make Density Based Clustering (MDBC).

K-Means (2010) clustering algorithm follows a simple and easy way to classify the given data set through a certain number of clusters. It partitions a dataset by minimizing a sum of squares cost function. MDBC (2013) is a type of clustering algorithm that is constructed based on the density properties of the database. It is most extensively used.
density based algorithm, which uses the concept of density reachability and density connectivity.

**Figure 5.8** Computation time for various clustering algorithms

**Figure 5.9** Accuracy level of various clustering algorithms
MAE is an error rate that is used to measure how far the estimates are from actual values, which could be applied to any two pairs of numbers, where the first one is actual value and the other is an estimate prediction.

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |X_i - Y_i|
\]  

(5.4)

Where, \(X_i\) represents the predicted value for data point \(i\) and \(Y_i\) represents the actual value for data point \(i\).

From this analysis, it is observed that the SRCA clustering algorithm provides the reduced computation time. Figure 5.9 and 5.11 depicts the comparative analysis among the clustering algorithms based on the level of accuracy and MAE rate. The proposed SRCA algorithm provides the high accuracy (75%) and low error rate (50%), when compared with the other algorithms. Figure 5.10 represents the MAE rate for the proposed SRCA algorithm based on the number of clusters. From this analysis, it is proved that the proposed SRCA method performs well and provides the best performance results, when compared with the existing algorithms.
Figure 5.11 Comparative analysis between various clustering algorithms based on MAE

The computation time, accuracy and MAE values of K-Means, MDBC and SRCA are shown in Table 5.4. From this table, it is evaluated that the proposed SRCA provides high accuracy, minimized computation time and MAE.

Table 5.4 Computation time, accuracy and MAE values for existing and proposed algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Computation Time (s)</th>
<th>Accuracy (%)</th>
<th>MAE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-Means</td>
<td>150</td>
<td>60</td>
<td>75</td>
</tr>
<tr>
<td>MDBC</td>
<td>125</td>
<td>65</td>
<td>65</td>
</tr>
<tr>
<td>SRCA</td>
<td>100</td>
<td>75</td>
<td>52</td>
</tr>
</tbody>
</table>

5.6 RISK EVALUATION

Risk is defined as the expectation and the probability of loss, which is an undesirable event. There are lots of risks involved while creating the highest quality software modules on the time and budget. It is important to identify the probability of occurrence and minimize the level of risk during software module generation. The severity of the risk is calculated by,
Risk minimization in the context of software engineering requires the understanding of software assurance and security engineering. Here, a Direct Acyclic Graph (DAG) method is proposed in this research in order to optimize the software requirements. When, the number of stakeholders is increased, the risk level is also increased.

Risk Severity = Probability of Occurrence × Potential Negative Impact \hspace{1cm} (5.4)

Figure 5.12 illustrates the risk minimization graph for various number of stakeholders, where the level of the risk is minimized by applying the DAG method. The software risk evaluation team analyzes each recognized risk in terms of its cost, performance, quality and schedule. Figure 5.13 represents the comparison graph between the existing and proposed methods based on the risk level. From this analysis, it is proved that the proposed DAG model efficiently reduces the risk, when compared with the existing algorithms.
Scalability is defined as the ability to handle increased workload by repeatedly applying a cost effective strategy for extending a system’s capacity. Scalability and reliability are the important attributes, which are mainly used to estimate the stakeholder’s requirements. It is the most significant and most measurable aspect of the software module generation. Reliability is an essential facet of software quality, which is dynamic and stochastic. It is defined as the probability of the failure free software operation for a specified period of time in a particular environment.

Figure 5.14 and 5.15 represents the scalability and reliability measure for various number of stakeholders. When the number of stakeholders is increased, the scalability and reliability measures are also increased. Reliability measure is a set of mathematical techniques, which is mainly used to estimate and predict the reliability behavior of software during module generation and development.
Figure 5.14 Scalability of proposed DAG

Figure 5.15 Reliability of proposed DAG

Figure 5.16 and 5.17 illustrates the comparative analysis between the preprocessed, classified and clustered datasets based on the time and accuracy. From this graph, it is observed that the clustered dataset requires less time (70,000 ms) and provides high accuracy (85%), when compared with the preprocessed and classified datasets.
Figure 5.16 Comparison between the datasets based on the time (ms)

Figure 5.17 Comparison between the datasets based on the accuracy level

The time and accuracy values of preprocessed, classified and clustered datasets are shown in Table 5.5. From this table, it is analyzed that the clustered dataset requires minimum time (78,000 ms) and provides high accuracy (82%).
Table 5.5 Time and Accuracy values for preprocessed, classified and clustered datasets

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Time (ms)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preprocessed</td>
<td>1,20,000</td>
<td>62%</td>
</tr>
<tr>
<td>Classified</td>
<td>82,000</td>
<td>71%</td>
</tr>
<tr>
<td>Clustered</td>
<td>78,000</td>
<td>82%</td>
</tr>
</tbody>
</table>

5.7 SUMMARY

The requirements engineering process provides the best opportunity to consider all of the various stakeholder’s interest in context with one another. Stakeholders are individuals, who have some level of influence over the requirements for that software product. Identifying and considering the needs of various stakeholders help to prevent software product requirements from being overlooked. In this analysis, the performance of the proposed work is evaluated in terms of accuracy, reliability, precision, recall, scalability and reliability. From the results, it can be concluded that the proposed hybrid approach, FRBC classification method, SRCA clustering technique achieves increased classification performance and yields results that are accurate in the cases of large number of stakeholders.

In future, the requirement prediction validation module can be developed by implementing the software testing analysis model. The accuracy of the proposed module prediction will be increased by using a module dependency analysis. Moreover, the prioritization technique will be developed for implementing the software modules.