Unsupervised Clustering of Requirements

B. Saranya, R. Subha, Dr. S. Palaniswami

1Department of CSE, Sri Krishna College of Technology, Coimbatore
2Assistant Professor, Department of CSE, Sri Krishna College of Technology, Coimbatore
3Principal, Government College of Engineering, Bodinayakanur, India

ABSTRACT

Requirement engineering emphasizes the use of systematic and repeatable techniques that ensure the completeness, consistency and relevance of system requirements. But when requirements are collected from the users it involves many challenges like data inconsistency, different views of users, and different users will have different priorities and preferences for the same requirements. In this paper, a methodology is proposed to overcome the above challenges. It uses mean substitution method to handle the missing data and the requirements are prioritized using Analytical Hierarchy process. Ward's algorithm is used to determine the optimal number of clusters. K-means algorithm is used to cluster the requirements which will be useful for future projects.

Keywords: Requirement Engineering, Prioritizing Requirements, Requirement Clustering, Unsupervised Clustering

I. INTRODUCTION

Requirements engineering provides the appropriate mechanism for understanding the customer needs. The requirements engineering process can be broken down into requirement elicitation, requirement analysis, requirement validation. The processes may vary depends upon the users since every user will have different needs and preferences [1]. Requirement analysis plays an important role in the success of the project; it helps to know about the inconsistent and conflicting requirements. Requirements should be documented, measurable and testable and defined in a way that is suitable for system design. Clustering is the process of partitioning a set of data into meaningful subclasses called clusters. It is a machine learning technique for analyzing data and dividing it into groups of similar data. The data are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the inter class similarity. Outliers can be identified using clustering. Clustering comes under unsupervised classification that is it does not require a known class label to form clusters. The basic types of clustering methods are hierarchical and partitional clustering. Hierarchical clustering algorithms repeat the cycle of either merging smaller clusters into larger ones or dividing larger clusters to smaller ones. Partitional clustering algorithms generate various partitions and then evaluate them by some criterion.

Unsupervised clustering of requirements proposes a methodology to cluster the requirements based on the priority of the requirements. Before clustering, the requirements should undergo preprocessing. Missing data in the requirements are handled using mean substitution method and the requirements should be prioritized. When requirements are clustered based on the priority it will can guide the target for each development sprint.
The paper is structured as follows: After the introduction Sect. 2 provides a review of related work, Sect. 3 describes the methodologies used; Sect. 4 provides a short discussion on the outcome of the project and future work that concludes the paper.

II. RELATED WORK

Requirements are identified with the assistance of stakeholder using interviews, group sessions or using online tools. Machine learning algorithms can be used to validate the requirements and to determine the requirement clusters. Incomplete data in the requirements were handled using EM-algorithm. Ward’s approach was used to find the optimal number of clusters. By using k-means algorithm clusters were identified. Requirements were verified by building the number use cases that represents system functionality [2].

Requirements prioritization is a process to find which subset of requirements should be implemented in the early stages, in order to satisfy the customers. Incomplete data are handled by identifying relevant stakeholders for the project under consideration and assign weights. Utility theory is used to compute ranks of given requirements. Finally, the output should be distributed to the stakeholders, to obtain the degree of satisfaction [3]. Preference goals can be distinguished from mandatory ones and alternative ways are used to fulfill mandatory goals. A state-of-the-art planner is used to identify alternative solutions to mandatory goals that satisfy the preferred goals and priorities among them [4]. When requirements are properly collected from the users it can be processed efficiently. Hence wikis can be used to collaborate between the stakeholders. Wikis provide a flexible platform for asynchronous collaboration of stakeholders in requirement engineering. They are used as a base technology to create software project requirements with a variety of stakeholders [5]. One of the main problems in processing the requirements is handling of missing data. It can be handled using various techniques such as ignoring or discarding missing data, parameter estimation or use of different imputation methods [6].

Duan et al. [7] describes about the application of hierarchical clustering and probabilistic techniques to provide support for automated software requirement extraction. Hsia [8] proposed a methodology to decompose the requirements into certain number of useful, usable and semi-independent partitions. When requirements are partitioned in such a way then it would facilitate incremental development.

Ward’s algorithm is an agglomerative hierarchical clustering algorithm. It is used as an exploratory statistical technique to determine the number of clusters. In the first stage each of the n objects to be clustered is considered as a single cluster. The nearby clusters are merged. This procedure is repeated until the desirable number of is achieved [9].Oyelade, O. J [10] proposed a system for analyzing students’ results based on cluster analysis. K-means clustering algorithm is the efficient tool to monitor the progression of students’ performance in higher institution. As the system results are based on cluster analysis, it enhances the decision making by academic planners to monitor the candidates’ performance semester by semester and to improve the future academic results.

III. METHODOLOGY

The next subsections describe the methodology which constitutes our proposal of unsupervised clustering of requirements.
Fig 1: Overview of Methodology.

Figure 1 shows the overview of methodology.

It involves the following steps:

1. A large set of requirements is identified with the help of stakeholders and the users of the system are asked to grade the importance of the requirements. Thus n*m requirement matrix is formed where n denotes different users and m denotes various requirements.

2. Requirement matrix will have some missing values; hence mean-substitution method is used to handle the incomplete data.

3. Analytical hierarchy process (AHP) is used to rank the requirements.

4. Ward’s algorithm is used to find the optimal number of clusters.

5. Finally in order to obtain clusters k-means algorithm is used.

Fig 2: Requirements Matrix

Figure 2 shows the n*m requirements matrix with missing values
1. Handling Missing Data

Missing data is a major problem, when requirements are collected from different classes of users. Different approaches can be followed in order to handle such situations. Mean substitution method is the traditional approach that is used to resolve the missing data problem.

In this method, missing values are imputed with the mean value of that variable on the basis of the existing values for that variable. It is a single imputation method; the theory is that in the absence of any information, the mean is the best single estimate of any participant’s score.

Figure 3 shows the complete requirements matrix in which the missing data are filled using mean substitution method.

2. Ranking of Requirements

When users are asked to grade the requirements, different users will give different grades to the same requirement. Hence in order to know the exact priority of the requirements, it is necessary to rank the requirements. The common approach for this purpose is to use Analytical hierarchy process. AHP is one of Multi Criteria decision making methods that were originally developed by Prof. Thomas L. Saaty Figure 4 shows the rank of the each requirement in the requirement matrix.

Fig. 3 Complete Requirement Matrix.

Fig. 4 Rank of the Requirements.
3. Determining the Optimal Number of Clusters

In order to obtain the most desirable user requirements, we cluster the requirements using k-means algorithm. But it does not provide any information about the value of k, and also the output of the k-means varies depends upon the value of k. So it is necessary to determine the optimal number of clusters using an efficient technique. Ward’s algorithm is the most efficient technique uses for this purpose. Ward’s algorithm is an agglomerative hierarchical procedure, where the clusters are merged until single cluster is formed. Ward’s method considers the distance between two clusters A and B by calculating the sum of squares when we merge them.

Merging cost is calculated whenever the clusters are merged. The merging cost D of the two clusters is given by:

\[ D(A,B) = \sum_{i \in A \setminus B} \| x_i - \bar{m}_A \|^2 - \sum_{i \in A \setminus B} \| x_i - \bar{m}_A \| \]
\[ - \sum_{j \in B} \| x_j - \bar{m}_B \| - \frac{n_A n_B}{n_A + n_B} \| m_A - m_B \| ^2 \]

The optimal number of clusters is inferred by considering the step for which the merging cost start rises significantly.

4. Clustering of Requirements

The next step is to cluster the requirements based upon their priorities. The common approach used for clustering is k-means algorithm [12]. It works as follows

1. It accepts the number of clusters k given by ward’s algorithm.
2. It then creates the first K initial clusters from the dataset.
3. The K-Means algorithm calculates the Arithmetic Mean of each cluster formed in the dataset. The Arithmetic Mean of a cluster is the mean of all the individual records in the cluster.
4. Next, K-Means assigns each record in the dataset to only one of the initial clusters. Each record is assigned to the nearest cluster (the cluster which it is most similar to) using a measure of distance or similarity like the Euclidean Distance Measure.
5. K-Means re-assigns each record in the dataset to the most similar cluster and re-calculates the arithmetic mean of all the clusters in the dataset.
6. The preceding steps are repeated until stable clusters are formed and the K-Means clustering procedure is completed.

IV. CONCLUSION

In this paper, we proposed a methodology that uses machine learning algorithms to process the requirements. Incomplete data in the requirements matrix are handled using the Mean substitution method. Requirements are prioritized using Analytical hierarchy process and optimal numbers of clusters are determined using Ward’s algorithm. Finally, Clusters are determined using k-means algorithm.
This approach also has some limitations. If variables are huge, then K means algorithm is computationally faster than hierarchical clustering and it produces tighter clusters than any other clustering algorithms. But for different initial partitions it will result in different final clusters. Mean substitution method does not work efficiently when the number of missing values is large. In future to improve the efficiency of the method EM algorithm can be used to handle the incomplete data.

REFERENCES


