CSE-6324 Advanced Topics in Software Engineering

Project Inception

Clustering Simulink Projects

Team – 07

Shiva Katukoori

Sai Venkata Bhanu Shasank Bonthala

Priyanka Nallabirudu

Instructor

Christoph Csallner

Abstract

Simulink project clustering is an unsupervised classification of Simulink projects into groups(clusters) based on the Simulink metadata. The projects with similar metadata are grouped into one cluster. Projects with dissimilar metadata are usually grouped into different clusters. The main goal of this project is to cluster similar Simulink projects based on the metadata present. In the current project, using multiple clustering methods and approaches we come up with iterating and optimizing clustering results. We are going to use K-Means, Spectral clustering, and DBSCAN. Our data consists of attributes regarding the Simulink projects like the author, summary, category, content, etc. Project clustering involves data preprocessing and data clustering using the above-mentioned clustering algorithms. In this project, we will be using only one metric to evaluate the clustering results in the form of Human judgment. In the process of optimizing the code further, we would be using various tunable parameters to derive the best accuracy.

Introduction

Clustering has almost become a part of our life, which is also a leading topic in the research of various fields such as Statistics, Pattern Recognition, and Machine Learning. When we consider Data Mining, we can see clustering deals with very huge datasets with a bulk amount of data that has many different kinds of attributes. This leads to the unique computational requirements on similar clustering techniques. Many clustering algorithms have been invented that meet these requirements and can be utilized in real-life scenarios. We can differentiate the Clustering methods into two basic types: ***hierarchical*** *and* ***flat*** *clustering. These two types of clustering have a lot of subtypes and many different kinds of algorithms to the cluster. The goal of the* ***Flat clustering*** *algorithm is to create such clusters that are coherent internally and can be distinguished easily. The main goal here can be phrased as follows, The data in the same cluster must be very similar and data in the different clusters should be as dissimilar as possible. When coming to* ***Hierarchical clustering****, the output cluster hierarchy should be a tree of clusters. A cluster can be a child, a parent, or a sibling. The only drawback of Hierarchical clustering when compared to Flat is that it is computationally intensive when finding relevant hierarchies.*

Competition

For this project, we have decided to work using python scikit learn library because it has a very active community and also supports more clustering methods than the competition. Here our competition for clustering includes KEEL(Knowledge extraction based on Evolutionary Learning), mlpy, and Weka[1].

KEEL supports only one clustering algorithm, that is ClusterKMeans, so we cannot perform this project using this tool since we need to investigate using different clustering algorithms to find the best one for dividing the given Simulink projects into different clusters[2].

*The "**mlpy** is a Python module for **Machine Learning** built on top of NumPy/SciPy and the GNU Scientific Libraries."[3]. It only supports K-means, Hierarchical clustering algorithms and memory saving hierarchical clustering,[4] which helps us to investigate different clustering techniques on the Simulink metadata but here hierarchical and memory saving hierarchical clustering may yield similar results which may hinder our chances of finding the best clusters.*

Weka is a major competitor to our choice of project implementation for the clustering of Simulink projects based on metadata. It also has many clustering techniques at disposal like skit learn library. Weka is implemented using java and is used for experimenting on small datasets whereas using scikit learn we can do clustering on medium to large datasets.

Risks

1)Attaining knowledge on the clustering techniques that we are going to use and tuning the algorithms to get optimal results. - Shiva
2)Gain knowledge related to Simulink to be able to distinguish bad clusters from good clusters. - Shiva
3)The need to depend on the project guide/mentor for reviewing the models, to get the actual accuracy of clusters obtained vs manually created clusters. - Shiva
4)Providing the correct time for the deliverables. - Bhanu
5)Identifying the appropriate roles and assigning them to the respective teammate. - Bhanu

Features

1)When Simulink users need to refer to an existing project linked to a certain topic and utilize it as a knowledge base, this project will prove to be very helpful.
2)This is also useful when someone develops a project using Simulink and is trying to figure out if the project can be useful in a different domain.
3)By iteration 1, we will implement the K-means clustering algorithm, identify the clusters and get feedback from the project mentor to see the accuracy. This plan is Based on the risk that we need to know more about the hyperparameter optimization and get a better understanding of Simulink projects.
4)By iteration 2, we will implement the Spectral clustering algorithm, spot the clusters and get feedback from the project mentor to analyze the accuracy. Based on the risk that we need to know about how to implement the spectral clustering to the data.
5)By iteration 3, we will implement the DBSCAN clustering technique, recognize the clusters and get feedback from the mentor to obtain the accuracy of the model. Then we will analyze the clusters we got using all the three algorithms and come up with the best clustering technique for the Simulink data.

Customers/Users

The customers of our project will be organizations which use Simulink for their engineering designs.
The users of our project will be Engineers who use Simulink to implement designs and simulate their systems.

References

1. *http://www.butleranalytics.com/10-free-data-mining-clustering-tools/*
2. *http://www.keel.es/*
3. *http://mlpy.sourceforge.net/*
4. *http://mlpy.sourceforge.net/docs/3.5/cluster.html*
5. *https://vtechworks.lib.vt.edu/bitstream/handle/10919/52341/ClusteringReport.pdf#:~:text=Abstract%20Document%20clustering%20is%20an%20unsupervised%20classi%0Ccation%20of,a%20structure%20in%20a%20collection%20of%20unlabeled%20data*

GitHub

https://github.com/Shiva-K-7/ASE-07