Data Clustering

REPORT

BY ZAIN UL HAQ

1 Problem Statement

The dataset provided for the clustering task consists of only approximately 30% labelled data while the unlabelled data may also contains data belonging to classes present other than in labelled data.

2 Description of dataset

Dataset provided consists of 10443 readings that represent data from independent experiments where each column represents one measurement (feature) of the experiment. The first column (ref_group) defines the experimental condition.

A subset of these experiments belongs to the so-called reference groups which are apparent by the group column:

reference group A: group == group_A

reference group B: group == group_B

reference group C: group == group_C

reference group D: group == group_D

reference group E: group == group_E

These groups represent so-called control experiments. For the majority of the experiments (group == unknown), and its not known which group they belong to. They may belong to one of the reference groups or to a group that is not already defined.

3 Approach

1. Analyse dataset for its original shape and structure
2. Pre-process data for making it consumable by learning algorithms
3. Perform an unsupervised clustering on the data to visualize the natural clusters present in the original dataset.
4. Visualize separate clustering in unlabelled data to get gist of novel cluster if present and compare with clusters of labelled data.

5. Perform a semi supervised learning based on exiting labelled data and perform a probabilistic label propagation for the unlabelled data
6. Cluster data based on newly propagated labels to unlabelled data

4 Methodology

The adopted methodology involves a combination of unsupervised clustering and semi-supervised classification. The implementation consists of following major steps:

1. Perform an unsupervised clustering on the data to visualize the natural clusters present in the original dataset.
2. Perform analysis to get optimal number of clusters naturally present in the dataset.
3. Visualize separate clustering in unlabelled data to get gist of novel cluster if present and compare with clusters of labelled data.
4. Perform a semi supervised learning based on exiting labelled data and perform a probabilistic label propagation to the unlabelled data using a certain confidence threshold level.
5. Cluster all the newly labelled data.

4.1 Dataset Analysis

Provided data is visualized first for understanding the underlying structure of collected features in each independent experiment. On initial inspection of the data, many empty columns were found along with other missing sections as can be seen in the Figure 1. Along with that main row contain missing values as well. In figure-1, white gaps show empty columns and rows with no data points at all. In order to make data useful for learning, data processing is required to get rid of empty columns as well as to identify and eliminate redundant feature columns.

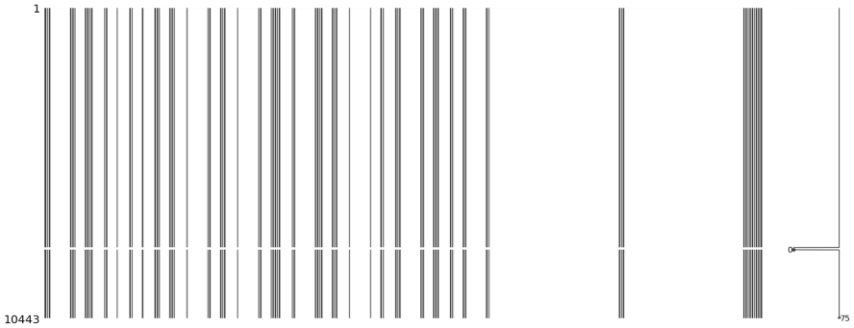

Figure 1 Visualization of first 75 feature columns of dataset

4.2 Data Pre-processing

Following steps are performed in pre-processing of the dataset.

1. Locating and eliminating empty rows and columns
2. Replacing missing values in feature with mean of the feature values
3. Identify highly correlated data and remove redundant data
4. Converting categorical data features to numerical values
5. Separation of labelled and unlabelled data.

Out of 1490 feature columns in dataset 265 columns were empty columns so they were identified and removed at first. Later on missing values in the rows were identified and replaced with the mean value of the feature columns to make the data consistent for clustering and classification algorithms. Correlation among the features were calculated and all the features having more than 95% correlation threshold are dropped out of dataset to get rid of redundant data. After performing all of the above steps 284 feature columns are left in the dataset where as one column describes the target class of each example.

4.3 Visualization of natural trend in dataset

After pre-processing of entire dataset, it is plotted on 2D planes to exactly visualize the natural distribution of the datapoint present in the dataset. PCA technique is used to reduce the dimension of the dataset from 284 to 2 principal component that

capture the major variation in the dataset within only 2 dimensions. Figure-2 shows the natural variation in the data captured by PCA.

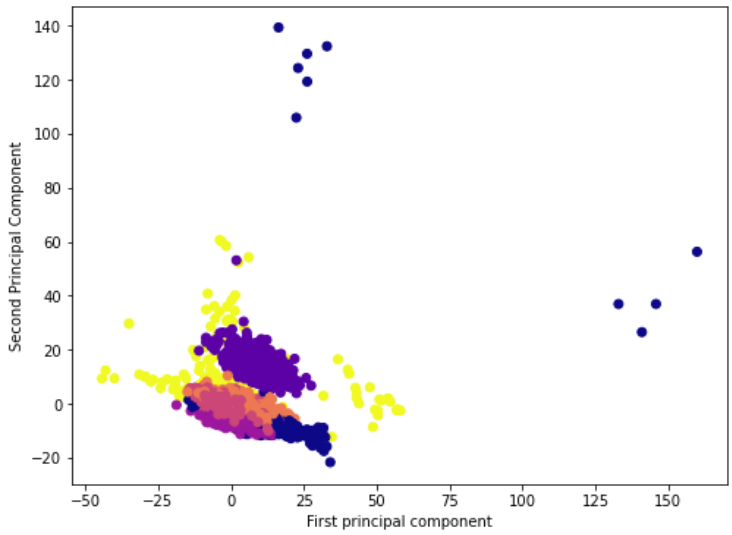

Figure 2 PCA Analysis of entire dataset after pre-processing step

As can be seen in Figure 2, “plasma” colour scheme from matplotlib library shows naturally captured variation in the data before clustering. A few outliers have also been detected but they are not removed for later learning with regards to discovering novel classes from the unlabelled examples in dataset.

4.4 Unsupervised clustering

In order to identify natural grouping in the dataset, unsupervised clustering is utilized. Since the dataset is supposed to contain some undiscovered novel groups as well so exact number of clusters is not known in advance. In order to get optimal number of clusters, Elbow method is employed. As can be seen in Figure 3, a prominent bent is seen at 6 which gives an indication of optimal number of clusters present naturally in the dataset.

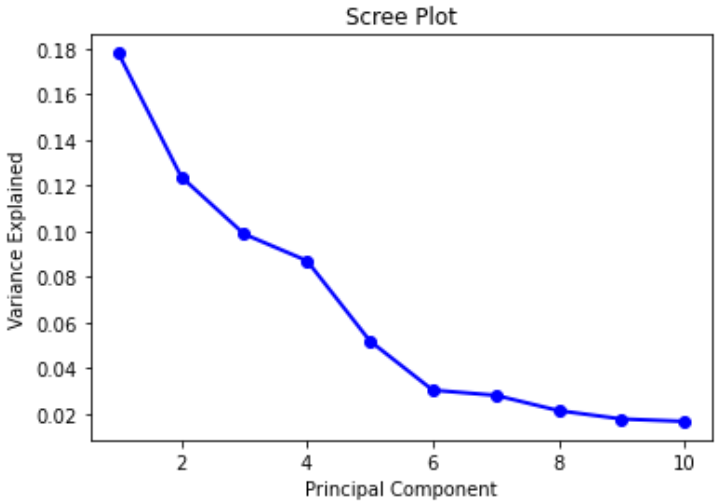

Figure 3 Scree Plot for finding optimal numbers of clusters

In order to visualize this natural grouping, Hierarchical clustering algorithm was used with agglomerative approach that start with many small clusters and merge them together to create bigger clusters using the 6 clusters to begin with.

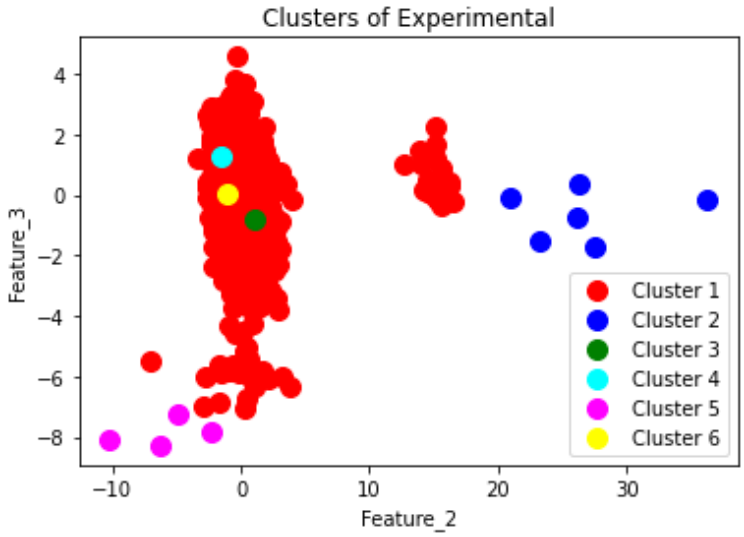

Figure 4 Agglomerative Clustering with 6 clusters.

To further observe the natural clustering in the dataset k-mean clustering was employed with variable number of clusters to see where the cluster centroids converge. In Figure 5, it can be clearly seen that cluster centroid does not shift significantly after k=6 clusters which indicates that dataset can clustered optimally in 6 groups.

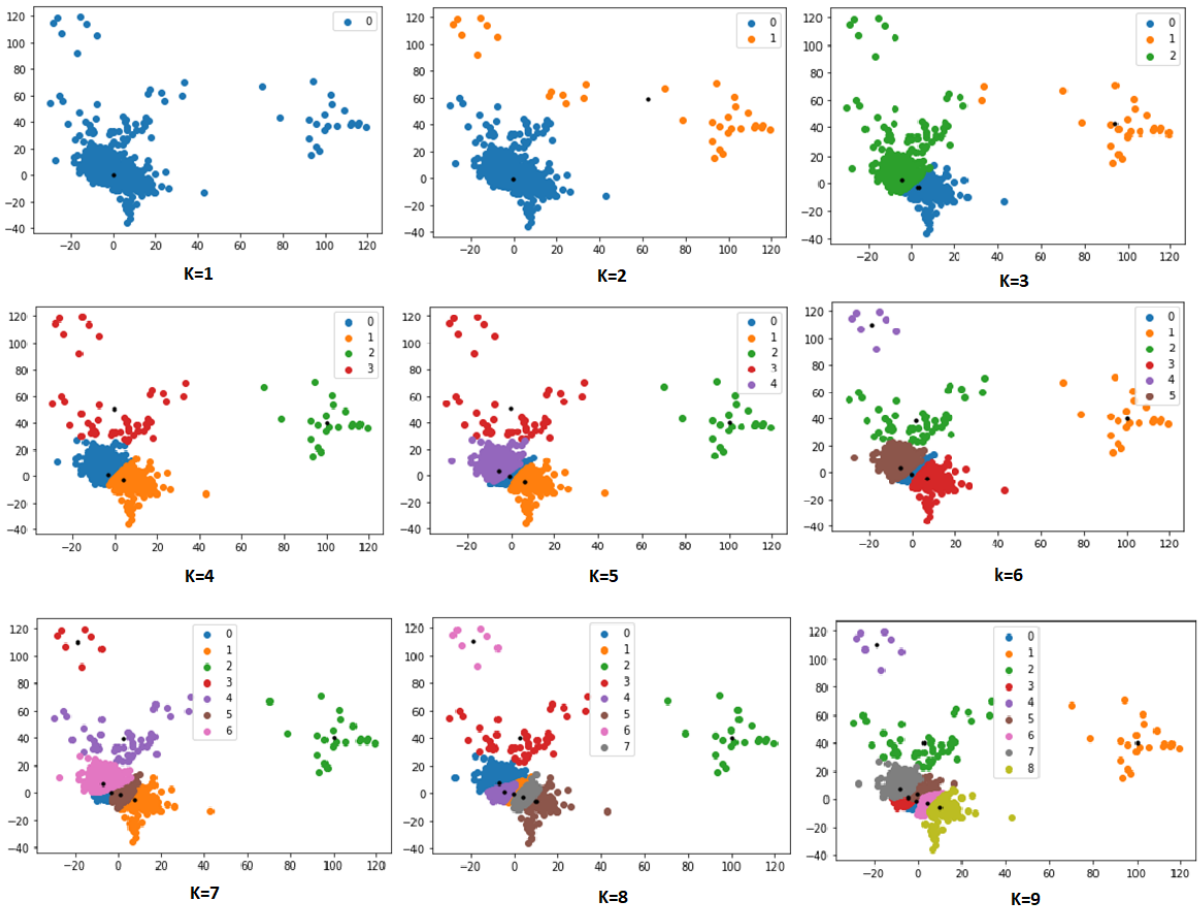

Figure 5 K-mean clustering with different value of K

4.5 Semi-Supervised Learning

Since almost 30% of the given dataset is labelled hence a semi-supervised learning approach becomes a natural choice. The fact that some of the unlabelled data also belong to existing classes hence label propagation using some probabilistic approach can be utilized to assign the unlabelled examples partially to a class and train a classifier accordingly in iterative manner to improve the accuracy.

Methodology

Labelled and unlabelled data is separated first. A SVM classifier is trained on labelled training data first. For this purpose, label data is divided into train and test data in 70-

30 ratio. The trained classifier on labelled data is then used to predict probability for each unlabelled example belonging to a certain group.

From the visual inspection of unsupervised clustering of data, as shown in previous section, 6 classes have been selected for the classifier whereas 5 classes already exist in the labelled data and 6th class added for classifying examples to novel group if any exits. Trained model predicts probabilities of an example to belong to each class out of 6. Maximum probability of a class is then compared with a threshold criterion that gives a notion of confidence level in the predicted label of the class. If the maximum probability value obtained for an example is above the confidence threshold, then that example is added to the training examples and model is retrained on new data to further enhance the accuracy of the model. Figure 6 shows confidence levels obtained for training examples.

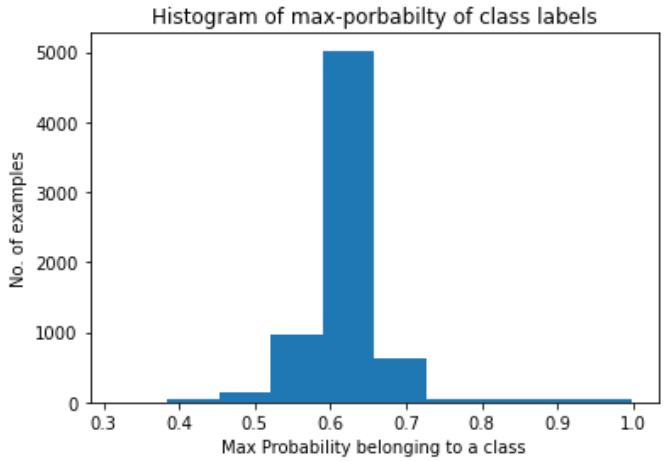

Figure 6 Confidence level with model trained only with labelled data

0.60-0.65 seems a good threshold value as it can be observed in Figure 6. All the examples from the unlabelled dataset are iteratively included in training set as per the confidence level to retrain the model.

Optimization and Evaluation strategy

Semi-Supervised learning algorithm mentioned above gives probability of belonging to each class. A threshold value can be set to compare this confidence level to consider the newly assigned label to the data and then include the new example in training of the model. This threshold value can be optimized to enhance the accuracy of the model on test data. Different threshold values have been tried to include

unlabelled examples to training set and evaluate model against test set. Trained model has been evaluated for its accuracy. An accuracy vs confidence level graph is shown in Figure 7 that shows optimization of the learned model.

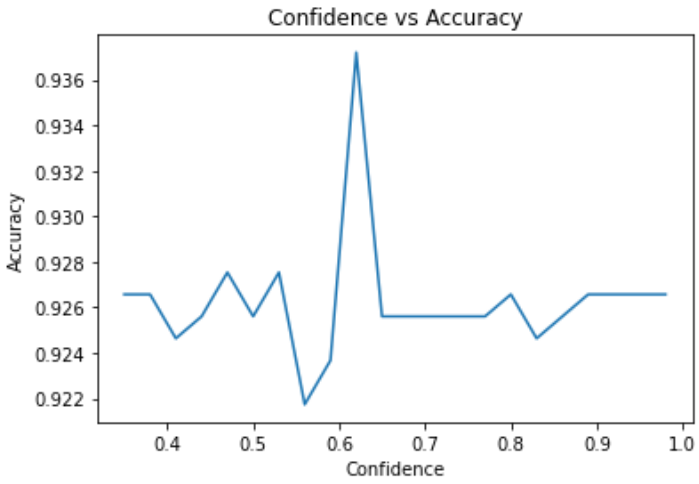

Figure 7 Accuracy of classifier with changing confidence threshold levels

It can be observed in figure 7 that after 0.65 no curve almost becomes flat showing that no new information is learned by model and hence accuracy is high around 0.65 confidence interval.

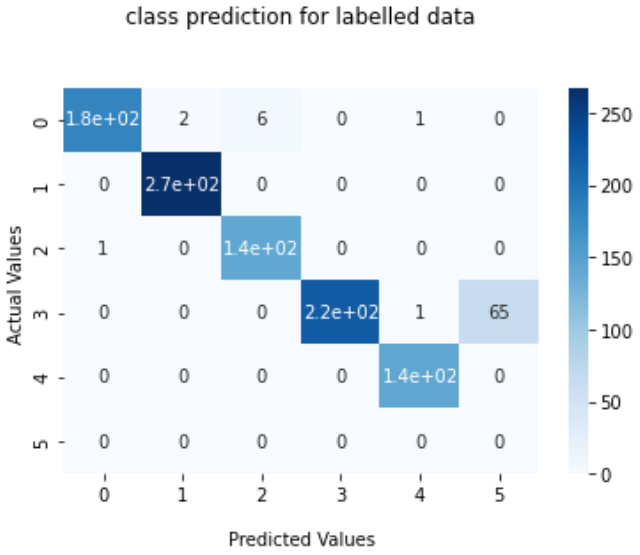

Figure 8 Confusion matrix for evaluation of trained model based on semi-supervised learning approach

Figure 8 has revealed a decent accuracy of the model on labelled test data. Another trend found in data is about group_X which is dented by class 5 in Figure 8, the confusion matrix clearly shows that group_X class is classified same as group_D indeed. To further validate the findings in results obtained about class “group-X”, a plot based on class naming of labelled data is shown in Figure 10 and observed relation between “group_D” and “group_X”. Since “group_D” and “group_X” show very little variation on PCA graph as shown in figure 10 hence model is not very well trained for distinguishing among group_D” and “group_X”.

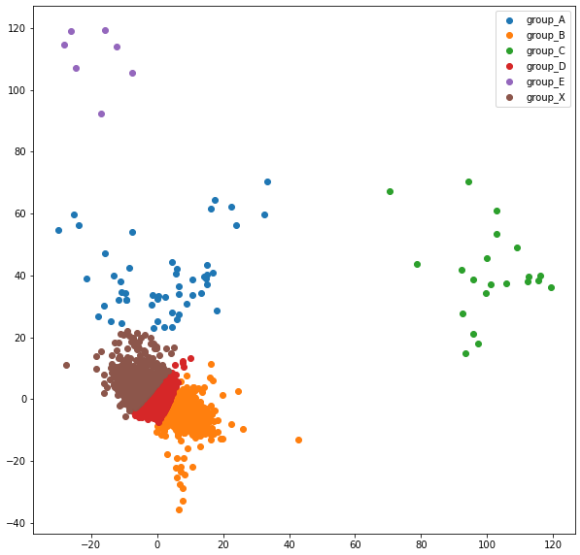

Figure 10 Clustering based on PCA predicted labelled data for labelled examples

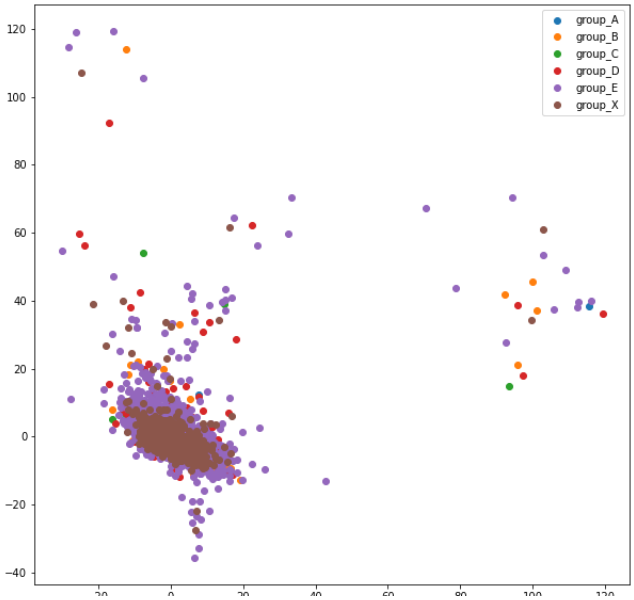

Figure 9 Clustering based on real labelled data

Figure 9 shows , how closely the experimental data is associated along first two principal components of PCA.

4.6 Classification of unlabelled data

After training a decent model, as mentioned in the previous section, all unlabelled data is clustered among 6 classes. A confidence level of 65% is used to label the examples to one of existing 5 classes. If max probability of a predicted class for an example in unlabelled data is more than 0.65 then example retains that class label else it is assigned to novel group i.e. 6^{th} group. After cluster.

Table 1 Class probability predictions for first 5 unlabelled examples

	C1Prob	C2Prob	C3Prob	C4Prob	C5Prob	C6Prob	lab	max
0	0.003531	0.000936	0.000663	0.610948	0.014493	0.369429	6	0.610948
1	0.003668	0.001092	0.000560	0.624936	0.018840	0.350904	6	0.624936
2	0.006752	0.001277	0.001436	0.605249	0.043742	0.341544	6	0.605249
3	0.006254	0.001123	0.001247	0.604561	0.028506	0.358309	6	0.604561
4	0.004113	0.000727	0.001090	0.634443	0.007816	0.351811	6	0.634443

Once label propagration based on trained model is complete new labelled data along with exiting labelled data is clustered again together using k-mean clustering. An accuracy of 0.61 is observed with whole dataset. Figure 11 shows clustering of the entire dataset based on newly propagated labels. Pink points in the figure 11 shows all the data points that showed low confidence level in belonging to any exiting class hence they are clustered together in as new novel group.

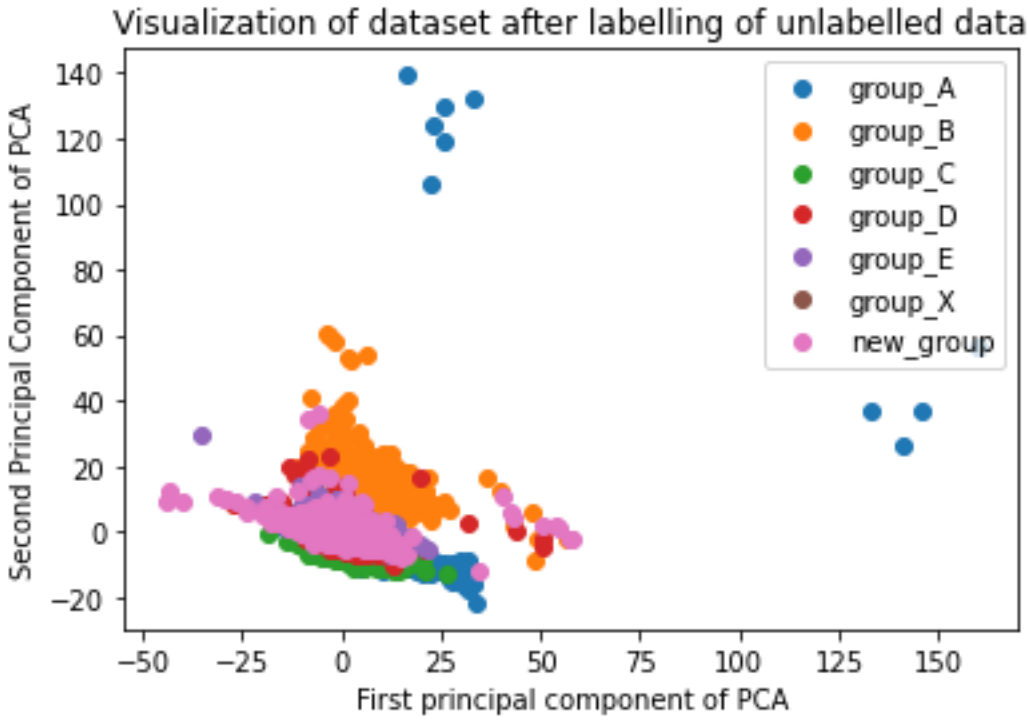

Figure 11 Clusters of whole dataset after labelling of unlabelled data

The approach presented in methodology section provided a classifier accuracy of 94 percent using semi-supervised learning and clustering accuracy of 62%. This low value of accuracy is due to presence of novel classes in unlabelled data which is not that well identified. Visual inspection of the predicted grouping has shown that labelling of unlabelled data through semi-supervised learning approach presented in section1.5 has done a decent job in achieving 94% classifier accuracy on test data and hence gives a good confidence in label propagation of unlabelled data. clustering of newly labelled data with confidence level of 62% has identified examples belonging to the existing classes with confidence level above 65% and rest are included as examples belonging to novel class.

References

[1]"A simple SVM based implementation of semi-supervised learning", *Medium*, 2022. [Online].Available:https://towardsdatascience.com/a-simple-svm-based-implementation-of-semi-supervised-learning-f44eafb0a970. [Accessed: 30- Mar- 2022].

[2]E. Bair, "Semi-supervised clustering methods", *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 5, no. 5, pp. 349-361, 2013. Available: 10.1002/wics.1270.

[3]E. Bair, "Semi-supervised clustering methods", *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 5, no. 5, pp. 349-361, 2013. Available: 10.1002/wics.1270.