TERAKI Candidate Quiz
Artificial Intelligence

Shorouk Awwad

May 11, 2020

Date Solved: May 10, 2020
Solved by: Shorouk G. Awwad

1 Objective

Using Clustering Algorithm to classify the given 22 .CSV datafiles into number of events (classes) according to the features measured by the **IMU**sensor: acc_x, acc_y, acc_z measured in m/s^2 and gyr_x, gyr_y, gyr_z measured in rad/s.

2 Data

2.1 Data assumbtion

First Assumbtions about The Data
The 22 . CSV files have been explored and gone through carefully and the following was noticed:

a) the uptime feature values seemed to be very close to ones in some of the oher files which made me assume those similarities might have effect on the training in two ways :

1. the classes with similar timestamps can belong to the same event (class) which makes it main feature to do the clustering. However, I couldn't assume the importance of the other features since there might be more than one device giving the reading at the same time .

2. the values of this feature was of very big numbers compared to the other features (different scale), which of course will affect the training.

b) The data wasn't labeled, which means there was no easy way to reduce or have a preference twords some features to another using easy tools like

RF .. etc.

2.2 Data Preprocessing

Preprocession Attempts

1 . combining the whole data set in one array of 3701 samples.

2. clearing the dtataset and imputing the NaN values that destroy the clustring model.

3. normaliziing the data set to get rid of the problems due to the huge scale the **Uptimes** has.

4. fitting and transforming the data with a PCA model to:

   a) reduce the redundancy effect if it does exist

   b) strengthen the effect of some important features

   c) reduce the dimesions of the dataset into two dimension to easen the scatter plotting of the data

3 Model Selection

3.1 Comparisons

Model Selection
Two followng models have been looked through to be considered for training, and here are a table of each model and its pros and cons.

Kmeans Pros:

* depends on the distance betweern samples

* simple to be implmented

* of complexity O(n)

Cons

* requires the nuber of classes pre-defined
* the center points are always randomized so this can yeild different results on different runs
* very sensetive to data scaling and outliars

Mean Shift

* Easy to implement
* Doesn't pre definition of number of casse
* depends on the density of the data to for centers of clusters
* has time complexity of O[n^2]

Gaussian Mix clustering

* Easy to implement
* depends on distance like KMeans
* of time complexity O[n^2]$doesn'tneedinitializationofnumberofclusters$
* doesn't depend on the distance to the center of clusters, but rather the distance between the samples and their distributions
* changes might happen in the pediction results on different runs

---

4 Execution

The three above mentioned algorihms wiith different approaches to do the best possible preprocessing and obtain the most desired results. However for Kmeans and Gaussian Mixture, the clustering process differed from run to run due to the ranoomised way the centroids are allocated with. So, all the previously mentioned preprocessing stages in the Data section has been done, and some other approaches have been added to different experiments to discover more the performance of the data.

Experiment 1

First approach was normalizing the whole data matrix at one along the Y access. and then pass the the resulting dataset to a PCA to reduce dimentionality and decrease the effect of redundany and dublicates. then applying the elbow method to get the optimal number of clusters for Kmeans. In this experiment, **fife distinct** clusters have been detected, however due to the imbalance in the dataset, two optimal clustrs have been determined by the elbow method.

Second experiment

the data wasn't normalized in any way and was left with its original scaled features. Again, by using the elbow method, the number of optimal cluters obtained in this case swang between 2 and three(Refer to the ipynb file),so three optial clusters.

Experiment 3

In the third experiment, the data has been normalized each feature on its own (independently) and the then the data has been passed to the PCA ...etc. The elbow method graph in this case was nearly a curve with no clear optimal point, however the number **11** was assumed to be the one.

Gaussian Mixture

this algorithm has been tried on the same dataset as for Kmeans; though it nearly showed the same results every time, it treated the outliars differently. For GaussianMixture, it considered the outliars of the clusters as separate clusters, which made it have more number of distinct clusters than Kmeans.

Experiment 4: mean shifting Clustring

after running the algorithm on the normalized and preprocessed **Not** PCA reduced, the number of clusters obtained was two clusters. please refer to Experiment4.ipynb.

5 Results and Conclusions

There was a conflict between the number of classes obtained in the different experiment depending on the preprocessing methods used at each.

Experiment no.	No. of classes	preprocessing
1	1 (2 clusters tho.)	data was normalized as a whole
2	5 classes	data wasn't normalized at all
3	5 classes	features were normalized independently
4	5 classes	data was normalized as a whole

from above, weconclude that the most likely number of classes is 5 distinct ones, though the data and the features are so close in values so when normalized as a whole one class is obtained.
please, refer to the results file.

6 Future work

* work more on implementing preprocessing algorithms
* try initializing the Kmeans algorithm before starting
* a thought to combine between both Shift Mean and Kmeans in the processes of finding the centroids, but this will cost complexity. (still thought!!)

7 References

* SKlearn documentation website
* https://spin.atomicobject.com/2015/05/26/mean-shift-clustering/
* https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68
* https://www.quora.com/What-is-the-difference-between-K-means-and-the-mixture-model-of-Gaussian