Lab Assignment 5

CIS660/EEC 525

Sunnie Chung

Clustering with NASA Webserver Log Data

Choose any two Clustering Algorithms covered in class and apply to Nasa Webserver Log data Set

Plan your experiment with:

1. Determine Data preprocessing methods and Distance metric to apply for each of your Clustering algorithm.

2. For each clustering algorithm,
   2-1. Compare the accuracy of the classifier with at least two different sets of input parameters, if applicable
   Or

   2-2. Experiment for Feature Selection with PCA tools or Your Own Experiment (See Below for an example) – Extra Credit

3. Compare the accuracy of two Clustering algorithms

4. Discuss about your results:
   - Why your inducted model is different for the same training data as you change the parameter values.
   - Why a certain parameter setting shows with better accuracy than the others that you tried.

Data:

You may choose one from the followings.

Download from in Lab5 Section or its web sites:

- Data from Nasa Webserver Log File
- Data from NIJ Challenging
- For EEC 525 Students, you can choose your data source from any sensor data set or machine generated signals

Phases:

1. Determine Data preprocessing methods to apply for each of your Clustering

2. Apply two different versions of a Clustering Methods of Your Choice. Design your Data Analytic Experiment.

   2-1. Experiment To Find the Best Parameter Setting for your Clustering Methods.

   - Measure
   - Different Thresholds
   - The Number of Clusters

   OR For Extra Credit

   2-2 Experiment for Your Own Experiment as follow:

   Simple Experiment to choose the best K, the number of the clusters

   2-2-1. Pick the best parameter setting from Phase 2.

   2-2-2. Apply Your Clustering algorithm with the best parameters set to each different number of the clusters to see if there is any significant difference in the result for each iteration.

3. Validate your result with your Test Set to compare the Accuracy of your models for each Clustering method or with different Parameter settings.

4. Discuss about your results:
   - Why your inducted model is different for the same training data as you change the parameter values.
   - Why a certain parameter setting shows better accuracy than the others that you tried.

Available Platforms:

You can use any data analytic systems/tools of your choice. Some of those systems/tools are in the followings:

- R
  https://www.r-project.org/
  http://www.rdatamining.com/

- Python has the most recent Machine Learning Library and data analytic Algorithms

- SQL Server Analysis Services (SSAS) Data Tools: You can use R in 2016 Data Tool
  https://msdn.microsoft.com/en-us/library/mt604845.aspx
  or Stand Alone R Server
  https://msdn.microsoft.com/en-us/library/mt674874.aspx
  https://msdn.microsoft.com/en-us/library/mt671127.aspx

- Other useful data mining tool sites

http://www.cs.waikato.ac.nz/~ml/weka/

http://www.kdnuggets.com/software/classification-decision-tree.html

http://www.salford-systems.com/downloadspm

Submission: Submit your report in Doc File including:

1. Screen Captures of your Installation Procedure and related Source info (Which software, Link to the Site, Which Clustering Algorithm, etc).
2. Show all the Data Preprocessing Steps
3. All your models with each the different parameter settings and the result in Accuracy
4. Report on Discussion and Analysis on:
    - Why your inducted model is different for the same training data as you change the parameter values.
    - Why a certain parameter setting or a clustering method shows with better accuracy than the others that you tried.
5. Report on Discussion and Analysis on Your Results