Data Mining - Blatt 05

Manuel, Marius

December 6, 2017

1 Abgabe Übung 5

Das Programm k_mean.py nimmt zwei Eingabeparameter. Der erste Parameter ist das maxK. Der zweite Eingabeparameter ist der Pfad zu der Datei. Bis zum maxK wird der k_mean Algorithmus für die gegebene Datei ausgerechnet. Danach wird mit den Funktionen von sklearn und einer Implementation von `http://www.caner.io/purity-in-python.html` den adjusted RAND index, the normalized mutual information, und the purity of clusters und berrechnet. Das beste k wird dann für jede dieser Funktionen bestimmt. Dabei war das Ergebnis. Den Datensatz S2 konnte nicht verwendet werden, da er keine waren Cluster Informationen beinhaltet.

Datensatz	RAND	normalized mutual information	purity of clusters
jain.txt	2	6	1
compound.txt	3	3	1

Bei zufälliger Initialisierung werden am Anfang k verschiedene Punkte gewählt. Somit kann der k_mean Algorithmus unterschiedlich konvergieren. Die draus vorhergesagten Cluster der einzelnen Punkte sind somit von den Startwerten abhängig. Vor allem bei Punkten, welche zwischen zwei Clustern liegen führt das zu einer unterschiedlichen Einteilung. Somit sind die Werte für adjusted RAND index, the normalized mutual information, und the purity of clusters von den Anfangswerten abhängig.