



A Local Search Method to K-clustering

Supervisor: **Noureddine Bouhmala**

Group members:

Xiong Wen, Wang Wenjuan, Huang Jiaquan



Introduction

- Background
- Problem description
- Requirements
- Design of clustering
- Test and comparison
- Further work

Background

- Data mining
- Clustering
- Partional clustering
 - K-means clustering
 - Advantages and disadvantages
- RLS method
 - Previous work

- most common measure is Sum of Squared Error(SSE)

$$SSE = \sum_{i=1}^K \sum_{x \in C_i} dist^2(m_i, x)$$

k is the number of clusters; x is a data point in cluster C_i and m_i is the representative point for cluster C_i .



Problem Description

■ Objectives

Find best clusters with the lowest SSE by using the new method

Test whether the RLS method to K-Clustering is better than K-means method



Why use the RLS?

- The RLS method is supposed to be better in clustering
- some improvements based on K-means method
- But further testing is still necessary before we reach the final conclusion



Requirements

- **Functional Requirements**
- **Non-functional Requirements**



Functional Requirements

Data resource: points to be clustered from supervisor

Input: decide the file to test and number of clusters desired

Randomly grouping: divide all the points into k clusters randomly

Exchange points: possible to exchange the points in two random clusters randomly.



Functional Requirements

Calculate SSE: calculate the SSE in order to find the best clusters.

K-Means algorithm: compare the quality between RLS method and k-means algorithm

Output: present the clusters that are best clustered in order.



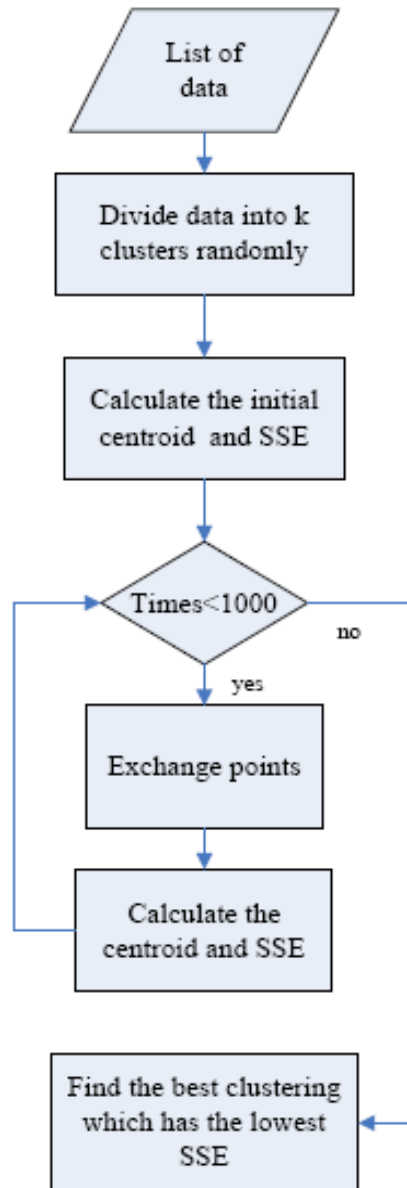
Non-functional Requirements

- **Time consuming:** try to short run-time
- **System requirement:** able to be implemented under Windows and Linux
- **User friendly:** easy for user to debug and input data



Design of Clustering

The new algorithm can be presented as a flow chart below :



Flow Chart

Test and Comparison

■ One test:

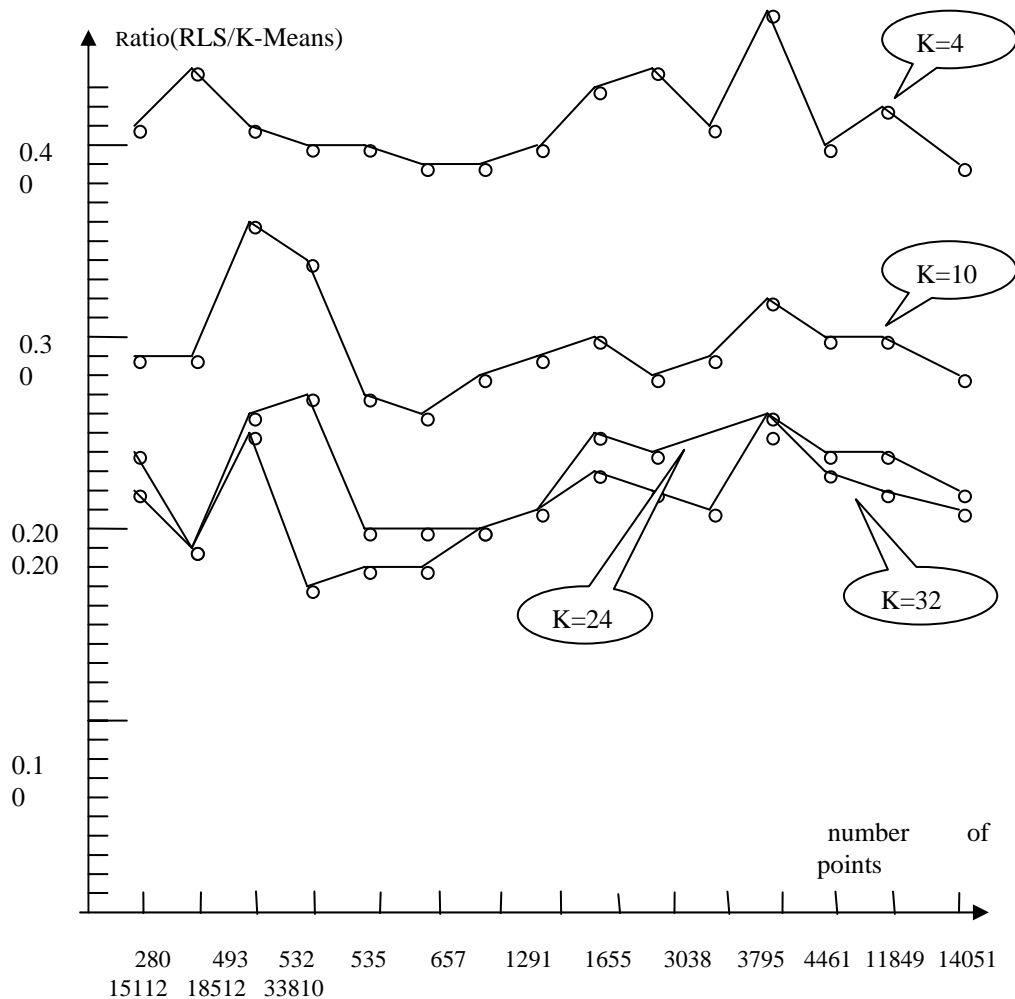
test1. Assign all the points to 4 clusters

Number of points	280	493	532	535	657
RLS method	2.560e+6	2.875e+8	3.488e+9	2.062e+6	6.862e+8
K-Means	6.233e+6	6.504e+8	8.473e+9	5.168e+6	1.724e+9
Ratio of results	0.41	0.44	0.41	0.40	0.40

Number of points	1291	1655	3038	3795	4461
RLS method	1.446e+9	1.758e+9	5.920e+9	2.175e+9	9.793e+9
K-Means	3.724e+9	4.502e+8	1.475e+10	5.022e+9	2.233e+10
Ratio of results	0.39	0.39	0.40	0.43	0.44

Number of points	11849	14051	15112	18512	33810
RLS method	3.996e+11	8.194e+10	7.474e+11	1.193e+11	1.611e+15
K-Means	9.74e+11	1.75e+11	1.86e+12	2.82e+11	4.164e+15
Ratio of results	0.41	0.47	0.40	0.42	0.39

Table 2: comparison of the lowest SSE of the two algorithms when the points are divided into 4 clusters, Ratio=RLS method/K-Means



The ratio of RLS and K-Means (note: the ratio goes down when increase the number of clusters, but the ratio becomes stable after 20 clusters)

Further Work

- Multilevel local search method

a new approach combining a new local search method and the multilevel paradigm is introduced for solving the k-clustering problem.

Agenda

Project progress ↻		Time plan ↻
Analysis project and divide tasks ↻		12.sep.2006-28.sep.2006 ↻
Learning Python and Java ↻		12.sep.2006-3.oct.2006 ↻
Decide the algorithm ↻		28.sep.2006-3.oct.2006 ↻
Coding: ↻	Open file to read data and store them into coordinate type ↻	3.oct.2006-19.oct.2006 ↻
	Complete the algorithm ↻	20.oct.2006-30.oct.2006 ↻
	Test and improve the algorithm ↻	31.oct.2006-10.nov.2006 ↻
Complete the report ↻		11.nov.2006-25.nov.2006 ↻

Reference

- U.Fayyad and R.Uthurusamy. Data mining and knowledge discovery in databases. Communication of the ACM, 39(11): 24-26, 1996
- Margaret H.Dunham. Data Mining introductory and advanced topics. 4: 17-18.
- Magnus Lie Hetland, Beginning Python: From novice to professional. Apress,2005
- J.MacQueen. Some methods for classification and analysis of multivariate observation. In Proceedings of the 5th Berkeley Symposium on Math. Statist. Problems, pages 281-297, 1967
- George Chang, Marcus J.Healey, James A.M. McHugh, Jason T.L.Wang, Mining eh World Wide Web: An Information Search Approach. 78



QUESTIONS?



Thanks for your patience