Project Plan and Progress Report

Team Members

- Yang Ruizhi, A0105733M
- Chen Shaozhuang, A0134531R
- Li Yuanda, A0078501J

Introduction to Our Project

Many data mining applications deal with high dimensional data. For example, when dealing large amount of text and image data, it is common to represent the data in a high dimensional vector space. Some classic data mining algorithms including k-means clustering and k-nearest-neighbor (kNN) become computationally expensive when dealing with such high dimensional data. Therefore, it is desirable to reduce the dimensionality of the data with low distortion. Fortunately, according to Johnson-Lindenstrauss lemma, such dimensionality reduction can be done by applying random projection on the high dimensional data.

In this MiniProject, we give a introduction on random projection and its sparse version. Then we will give some experimental results of random projection by applying it on high-dimensional text and image datasets, and verify that the distances between data points can be nearly preserved. In addition, as J-L lemma gives a worst case bound for the choice of reduced dimension, it is also interesting to experimentally study how the dimension after reduction affect the performance of random projection on the two datasets.

After empirically studying random projection and J-L lemma, we will also apply some classic data mining algorithms on both the original high-dimensional data and low-dimensional data after dimensionality reduction, and compare the results and time complexity of these algorithms on these data. Since random projection has the nice property of nearly preserving interpoint distance of the data, we will be particularly interested in experimenting with algorithms which utilize interpoint distances. In particular, k-means clustering and kNN will be studied in this report. In addition, we will also explore a sublinear k-means clustering algorithm, and compare its performance with the regular k-means clustering algorithm.

We will use two datasets on text and image respectively. The text dataset we use is the 20 newsgroups dataset, which comprises around 18000 newsgroups posts on 20 topics. The image dataset we use consist of image taken by surveillance cameras. Both of these two datasets are high dimensional, and Euclidean distance is important to cluster the data into different set of topics or scenes.

Current Progress

- Implemented RP and sparse RP.
- Implemented the algorithm to determine the dimension bounded by J-L lemma.

- Surveyed on vector space of text data.
- Found suitable text and image dataset to experiment with.

Plan

- By 3 Nov: apply RP to both text and image datasets, and see how the distances between data points can be preserved, and how different choices of reduced dimension can make a difference.
- By 5 Nov: apply k-means clustering and kNN to both datasets and get some basic results.
- By 9 Nov: finish interim report, with basic results and conclusions.
- By 12 Nov: perform more experiments, include sublinear k-means clustering.
- By 16 Nov: finish presentation and final report.