CSE633 Project: Transfer Learning on Unsupervised Imbalanced Dataset

Kang Li

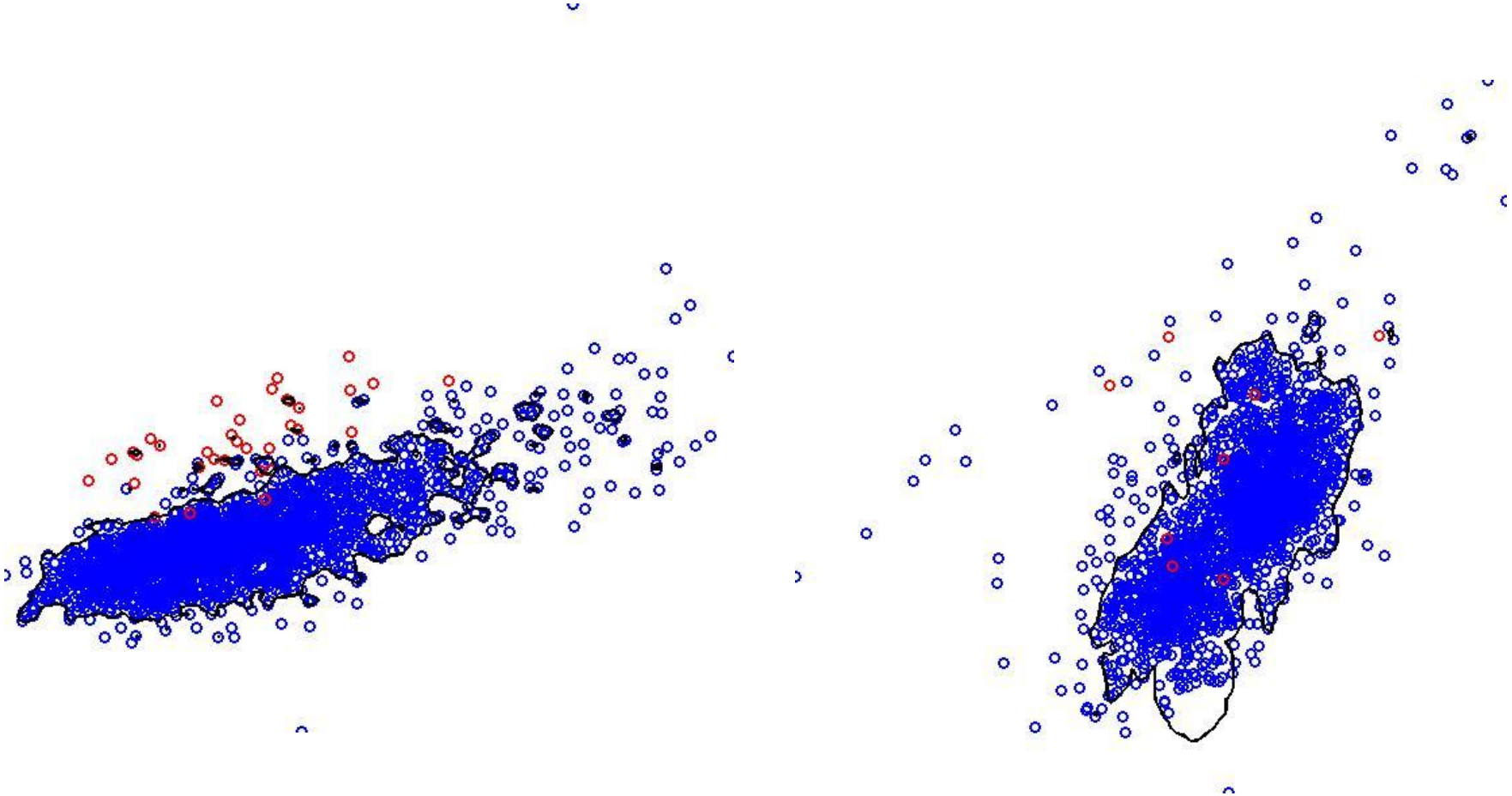
Person #: 36485128

12/6/2011

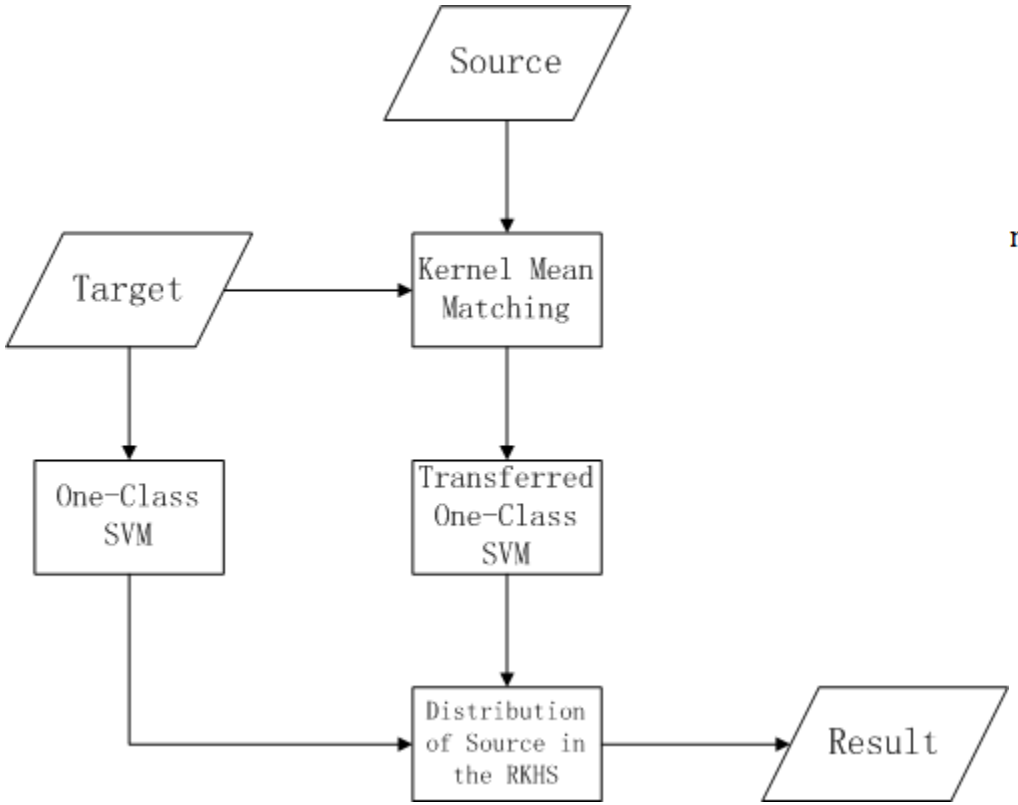
Introduction

- What are imbalanced datasets?
  - Number of instances in one class is significantly less than in others
- Why learning on such data is meaningful?
  - There are many such dataset:
    - Electro-cardio-graph (ECG)
    - Network Intrusion
    - PPI
  - There are many related areas:
    - Outlier Filtering;
    - Anomaly Detection;
    - Novelty Detection;

How to handle the Imbalance?



Methodology



$$\min \sum_{s=1}^{S} || \frac{1}{m_s} \sum_{i=1}^{m_s} T_s(x_i)\varphi(x_i)_+ - \frac{1}{m_t} \sum_{i=1}^{m_t} \varphi(x_i)_+ ||^2$$

$$= \min \sum_{s=1}^{S} (\frac{1}{m_s^2} T_s^T K_{s+} T_s - \frac{2}{m_s m_T} K_{T+} T + \text{const.})$$

$$\min R^2 + \frac{1}{S\gamma} \sum_{i=1}^{S} T_i \zeta_i$$

$$\text{s.t.} \; ||T_i \varphi_i - O||^2 \leq R^2 + T_i \zeta_i, \zeta_i \geq 0 \; \text{for} \; i \in [1, S]$$

Dataset for Experiments

- **MIT-BIH Arrhythmia Database:**
  - **Contain 48 half-hour excerpts of two-channel ECG signals, obtained from 47 subjects between 1975 and 1979.**
  - **Randomly select 10 records and only use channel #1 signals.**

- **Media-Mill Challenge Datasets:**
  - **ACM-Multimedia 2006**
  - **General video indexing data;**
  - **Contains five dataset for five challenge topics, dataset #1 is extracted feature space for images in each category.**

Experiments Set Up

- In each data, randomly select 1000 majority instances and keep all minority instances;
- Compare performance to both supervised and unsupervised methods:
  - Unsupervised:
    - k-NN: kernel distance, select threshold by observe the histogram;
    - One-Class-SVM;
  - Supervised: first 500 for training and remained for testing:
    - Over-sampling SVM;
    - Under-sampling SVM;

Evaluation Methods

- **The selection optimal source for a target is achieved by cross-validation.**
  - **Assumption: if a source is good for other targets, then it should be good for the aimed target.**
- **Evaluation:**

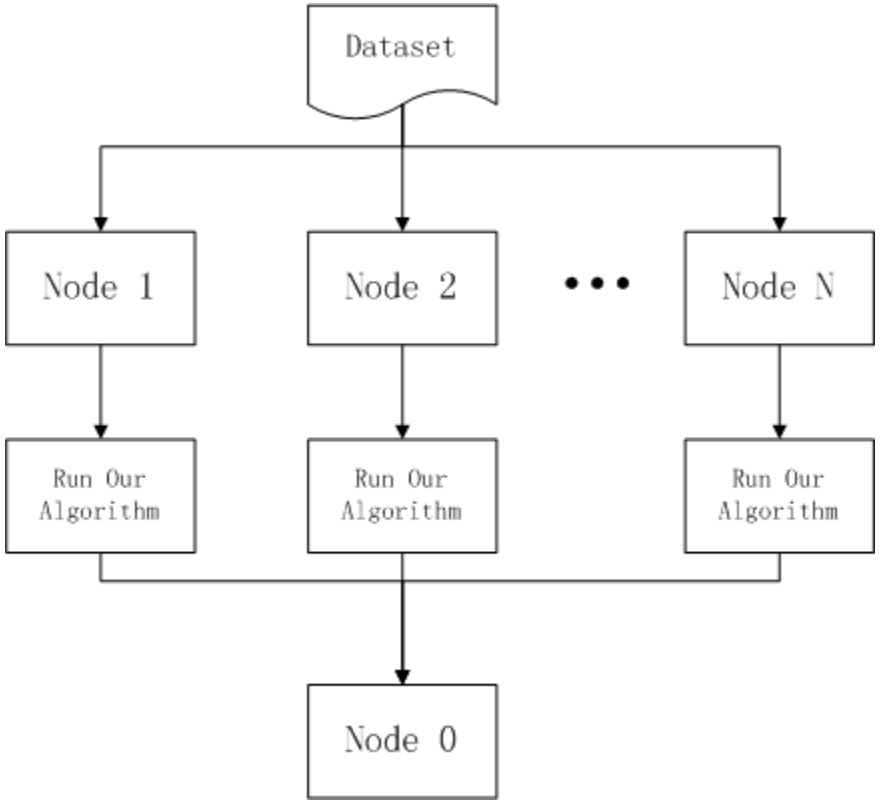
	True	False
Positive	TP	FP
Negative	FN	TN

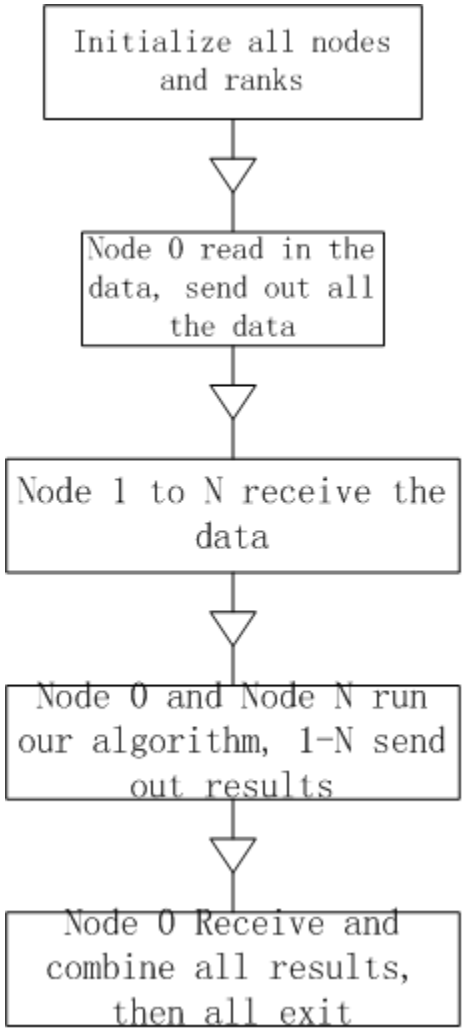
$$G-Mean=\sqrt{\frac{TP}{TP+FN}\times\frac{TN}{TN+FP}}$$

How to select a good source?

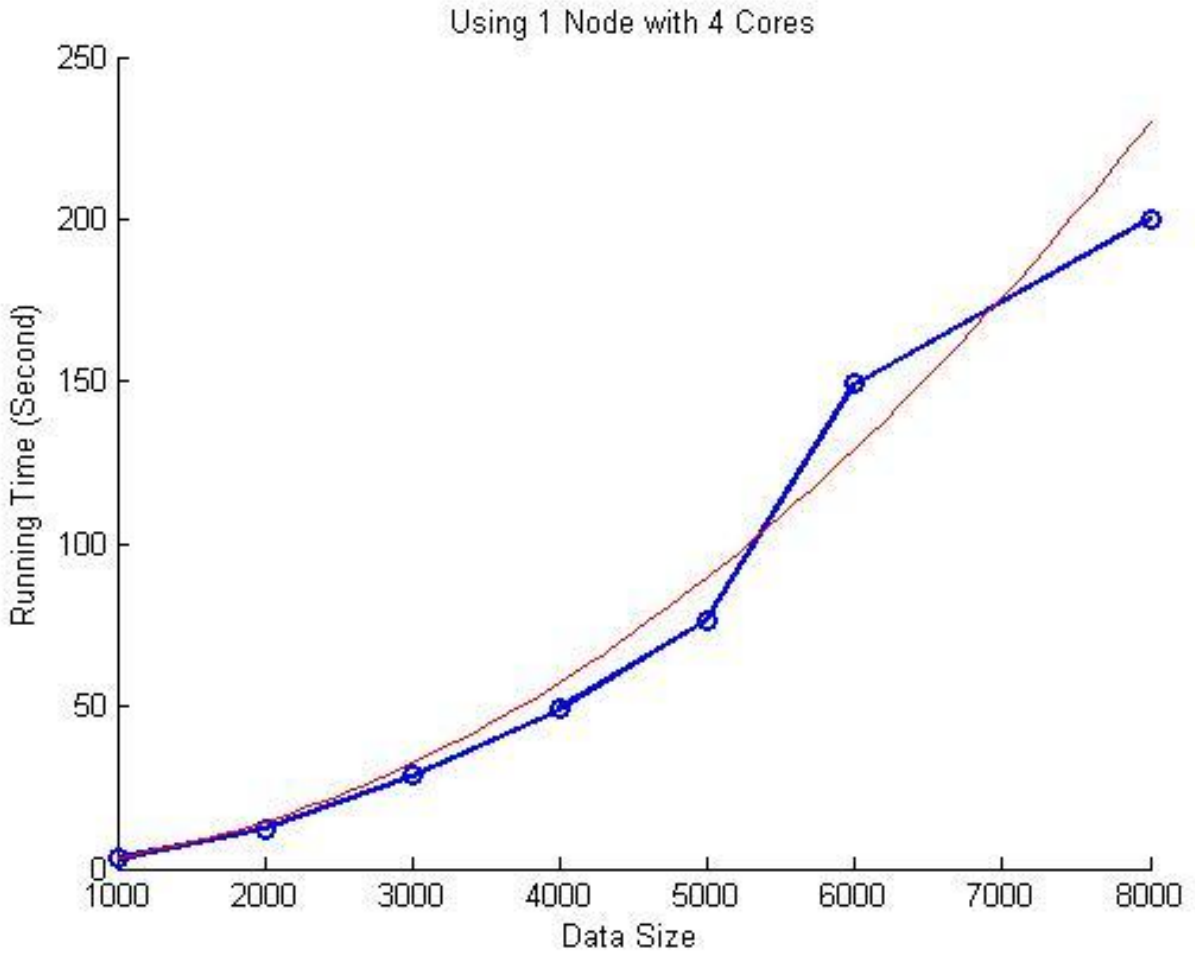
- **Cross Validation:**
  - Suppose we have 10 dataset, iteratively select one as target, then in the rest do pair-wise source-target exchange and find the source with highest accumulated performance as the 'good source' for the target;
  - In total we need to solve 10*9*2=180 Quadratic Programming Functions!
  - Let's parallelize it!

Parallelization Implementation

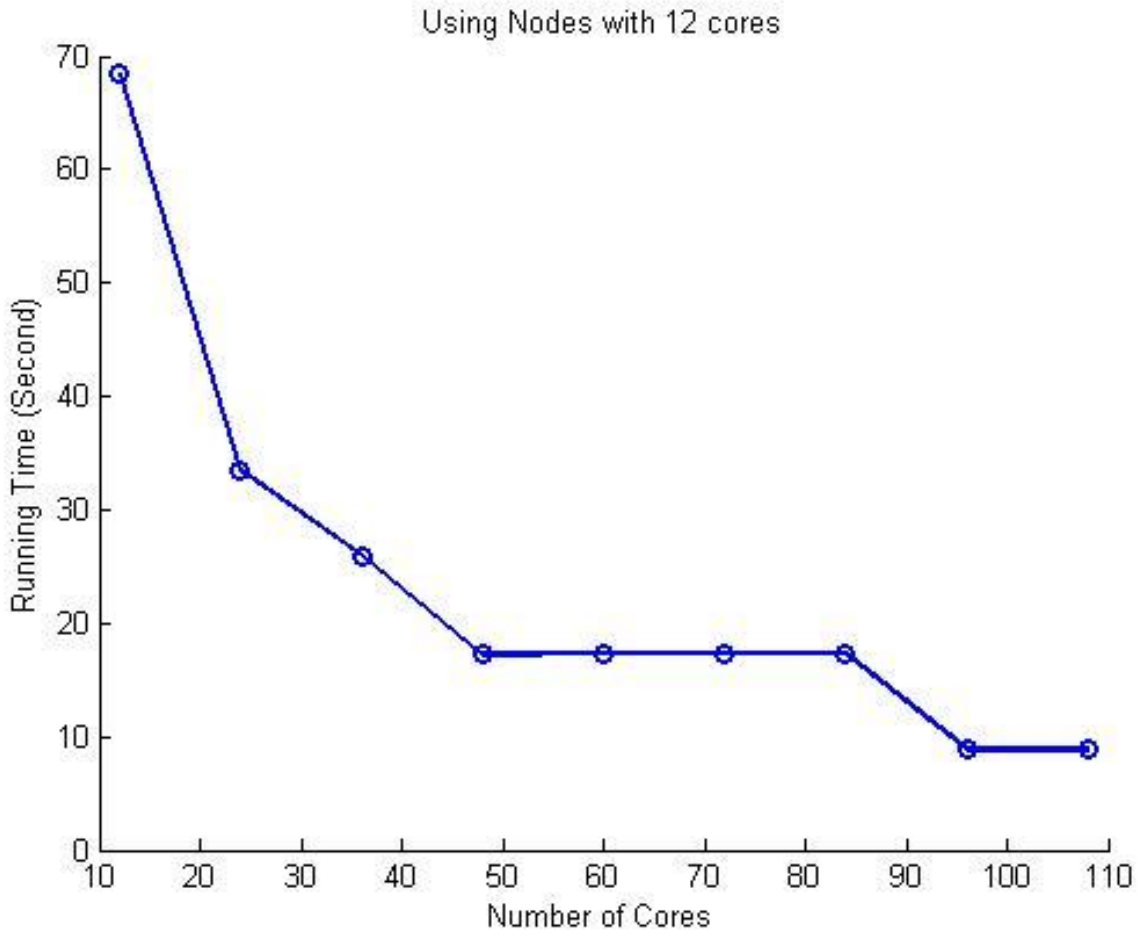




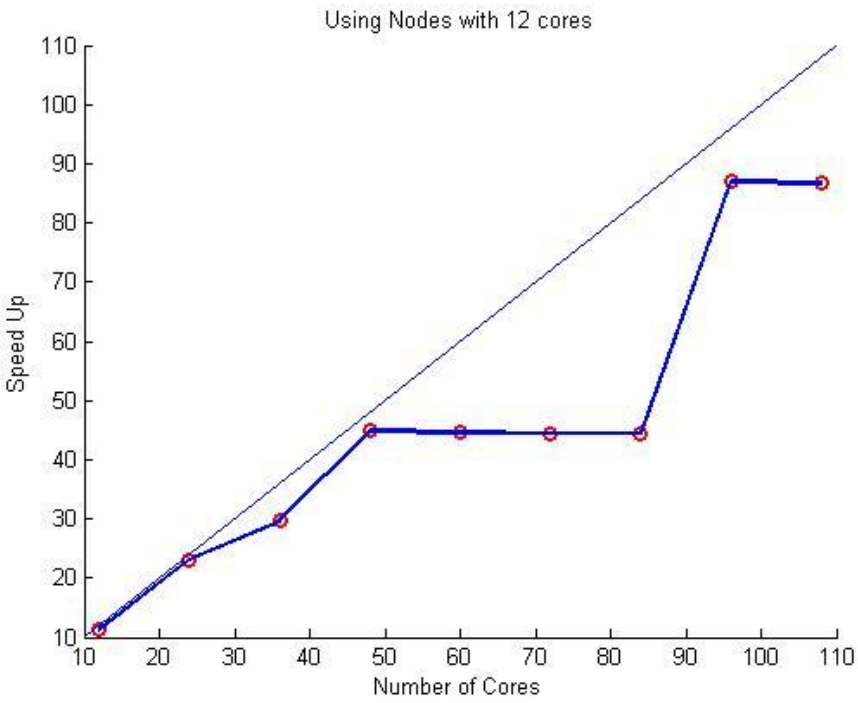
Experiment 1

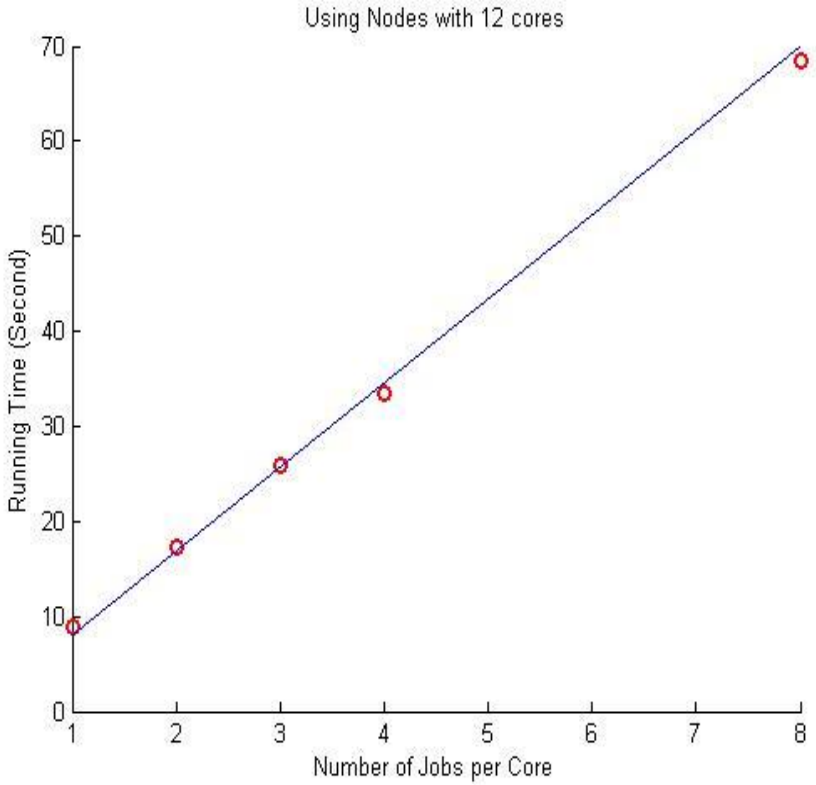


Experiment 2



Analysis


Using Nodes with 12 cores
Speed Up
Number of Cores
10
20
30
40
50
60
70
80
90
100
110


Using Nodes with 12 cores
Running Time (Second)
Number of Jobs per Core
0
10
20
30
40
50
60
70
1
2
3
4
5
6
7
8

Experiment 4

Data Size	Total Cores	Nodes	Cores/Node	Time (s)
8000	32	16	2	44.660853
8000	32	2	16	32.058259
8000	32	1	32	29.711242

1. Communication time between two nodes should be higher than between two cores in the same node;
2. Decreasing the number of nodes while increasing the number of cores per node enjoys significant benefits:
    1. 16*2->44.660853 s;
    2. 2*12->33.448988 s;

Conclusions and Future Works

This project successfully applies parallel programming into our proposed algorithm, and proves the powerful capability and advantages of parallel computing.

The future work of this project may include:

1. **Parallelize the Quadratic Programming Solver;**
2. **Adaptive evaluating parameters, and merge them into the new parallel computing process**

Performance of the Algorithm