Parallel Computing on Depression Analysis

Yanan Mao
School of Computer Science
Carleton University
Ottawa, Canada K1S 5B6
yananmao@cmail.carleton.ca

December 11, 2021

Abstract

Parallel computing is an amazing answer for multitasking cores and calculating in a short time, while transfer learning is a great way of output the relatively high accuracy for small datasets. In this project, depression disease analysis is based on Convolution Neural Network and Parallel K-means.

1 Introduction

For various reasons, public data on mental illness are hard to come by. Inspired by the rapid transferability of human beings in multiple similar tasks, transfer learning was introduced into this project, hoping to take this opportunity to make contributions to the faster classification of diseases. In addition, parallel computing is another focus of the project, which aims to identify the essential factors in categorizing diseases more clearly. In general, the classification ability of this project is not ideal due to the impact of the database. However, the project can continually do more profound research in the short future. Aim to achieve the initial proposal of this project, improve the accuracy of classification, and finally perform better.

With the development of society, the parallel processor has been introduced into the current scientific life. For example, mobile phones, personal notebooks, and other small electrical appliances can also do parallel processing. As a result, it reduced the size of the electrical devices and improved the running speed and functions simultaneously. Therefore, parallel computing is suitable for dealing with enormous data and fitting into the speedup requirement. Furthermore, the advantage of parallel computing is to use a collection of processors to attack multitasking on several cores simultaneously[9]. Especially in the bioinformatics area, to approach an increasing number of people worldwide are still suffering from it. More and more teens cannot help themselves escape from this lifelong problem. This project aims to create a highly efficient parallel adaptive clustering algorithm to study depression-related information. Moreclassify the mental illness more preciously. Due to the limitation of the dataset, at the end of the project, the related datasets are chosen the heart disease, breast cancer, lung cancer, depression disease dataset with standardized health information. In the meantime, to explore this network system, the spam email dataset has been chosen to check if the network could work better in other situations.

Section 2,I will go over the state of art of analyze depression disease, and point out how the project is lead to apply both parallel K-means method and transfer learning network.

Section 3 describes the problems. Section 4 is detailed a proposed solution in parallel. Section 5 presents the experimental results of this project. Section 6 is the conclusion and future plans to improve this algorithm.

2 Literature Review

Depression is not the blues or sadness or simply down. It is also a lasting overwhelming negative. Depression will also cause feelings of sadness and a loss of interest in activities we once enjoyed. It can lead to various emotional and physical problems and decrease functioning at work and home.

There are kinds of research on approaches to digging down the depression disease. They cover several ways: the species of the gut microbial, the specific diet, and related social factors, etc.[5][8][6]. Moreover, Li et al. was carried out for the epidemic period in China to discover that people's living environment and social environment have increased the probability of depression [6]. Different mental pressure can also put people at risk of contracting depression for different groups of people.

In order to deal with this level of complex data and do analysis work, in this experiment, parallel K-means is the first method we focus on, as it could speed up the sampling process[3]. Also, as a standard clustering method, K-means could recognize the patterns from different objects features. Instead of effect by human labeling bias, cluster analysis will find the structure in data and exploratory the related features. Ideally, we could have high similarities among one group and low similarities between each group[4]. Dimension reduction plays a crucial role in improving the accuracy of clustering algorithms[3]. Joonas et, al.,[3] also mentioned that the dimension reduction methods are split into feature selection and feature extraction. And transfer learning could do both of the work.

Rather than the dataset, which we cannot quickly get, because of the doctor-patient protocols of hospitals, individual privacy, and kinds of social opinions. The majority of research papers on depression I have found are based on statistics and small-scale surveys and experiments. There are more or fewer limitations in various aspects, such as sample selection. Moreover, multiple situations can promote depression in a person with a healthy mental state from the actual situation. Such disease is not only related to age or gender but is also connected with all the information about the individual. Due to sample limitations, this experiment still cannot solve and explore the true causes of depression.

Nevertheless, the neural network has excellent room for engaging in big data research and fine-tuning the performance of a model. We can also see that people use networks to analyze big data .animals image classification, create art, and classify spam emails. Although some researchers have tried to detect the user's mental state by crawling the user's web browsing records, and have made some progress. However, in general, mental illnesses are low-resource diseases. Therefore, transfer learning will also play its due role here.

It can be concluded from various works of literature on the Internet that transfer learning is a networking skill suitable for extracting characteristic information and classifying. Moreover, it is not new to study other fields through transfer learning. For example, the transfer learning network has been introduced to train the classification of plant diseases[10], and some use transfer learning to do language translation. Therefore, it is reasonable to train the network with abundant resource data to extract information and then use it for small data analysis. In addition, the case in the Crohns disease inspired the project to explore the possibility of parallel computing.

3 Problem Statement

As we can tell, the dataset of depression disease is hard to find and private. Therefore, the size of this kind of dataset is usually small. This project aims to build a combination transfer learning network and parallel K-means methods. The initial purpose is that the whole algorithm could deal with a small dataset. There are no related papers on this idea, so it is an innovation process. The questions I aim to answer are:

1. In what situations the algorithm will perform better.
2. How will the pretraining dataset affect the result we could get on classifying depression disease

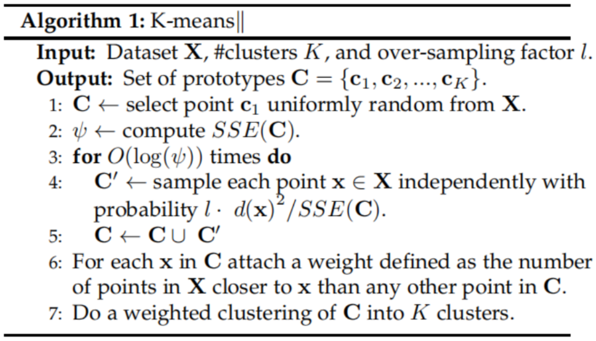
4 Proposed Solution

4.1 Parallel K-means (K-means||)

As an implementation of K-means, the parallel K-means method is not sensitive to the initialization selection of centroids. Ideally, the objects in the same cluster are more similar, and vice versa. Usually, the number of samples is way large than the number of clusters. The goal of the K-means clustering algorithm is to find suitable centroids by minimizing the sum-of-squares error(SSE), where SSE is defined as:

$$SSE(C) = \sum_{x \epsilon X} \min_{c \epsilon C} \|c - x\|^2 \tag{1}$$

The algorithm is inspired by Asarbaev A. [1]. and Kim A.[2].
Shown as Algorithm 1:Parallel K-means



```
Algorithm 1: K-means||
Input: Dataset $\mathbf{X}$, #clusters $K$, and over-sampling factor $l$.
Output: Set of prototypes $\mathbf{C} = \{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_K\}$.
$\mathbf{C} \leftarrow$ select point $\mathbf{c}_1$ uniformly random from $\mathbf{X}$.
$\psi \leftarrow$ compute $SSE(\mathbf{C})$.
for $O(\log(\psi))$ times do
    $\mathbf{C}' \leftarrow$ sample each point $\mathbf{x} \in \mathbf{X}$ independently with probability $l \cdot d(\mathbf{x})^2 / SSE(\mathbf{C})$.
    $\mathbf{C} \leftarrow \mathbf{C} \cup \mathbf{C}'$
For each $\mathbf{x}$ in $\mathbf{C}$ attach a weight defined as the number of points in $\mathbf{X}$ closer to $\mathbf{x}$ than any other point in $\mathbf{C}$.
Do a weighted clustering of $\mathbf{C}$ into $K$ clusters.
```

Assume we are given a dataset X, with n rows (n = the number of samples) and m columns (m = the number of features). Then, we initialize the random centroids C with m rows and k columns (k =the number of clusters).

SSE will calculate the distance between each data point and centroids. The distance values contribute to generating a D matrix in n rows and k columns. In each row, the k column containing the smallest number will be the cluster to which the point belongs.

The membership is supposed to save in the M matrix. The D and M matrices are changeable during the whole process until the centroids are fixed.

Parallel means we will run the code on multiple machines. For example, the master machine will work as an open-source, while worker machines will work locally.

The dataset X and centroids set C are stored in the master machine. The D and M matrix calculation belongs to workers and will pass the values to the master machine once finished.

4.2 Introduction of transfer learning

There are two domains in transfer learning, the source domain(S), and the target domain(T). Instead of learning from scratch like traditional machine learning, transfer learning will transfer the knowledge or weight matrix from the source domain to the target domain. The overall idea of transfer learning is to pre-train the network to fit into the source domain and fine-tune the trained network to predict the target domain. The limitation of transfer learning is that the source domain needs to have a deep-learning size of the dataset, meanwhile containing similar information as the target domain. For example, both of the domains are text information or image data. The definition of transfer learning is: Given a source domain D_s and target domain D_t and the learning task T_t to improve the learning of the conditional probability distribution $P(Y_t|X_t)$ in D_t with the information gained from $D_s$and T_s where $D_t \neq D_s$, or $T_t \neq T_s$ [7].

4.3 Convolution Neural Network(CNN)

In this project, CNN is needed to build up a network. Connect each hidden unit to a patch of previous layer neurons! Share matrix of parameters across units. But the inputs are different and local. The layer will take a fixed batch of information to produce the process.

ReLU is selected as the activation function. The returns the standard ReLU activation is: max(x, 0). Modifying default parameters allows you to use non-zero thresholds. ReLU is a linear function that will output the non-zero input directly, otherwise, the function will output zero. In this way, the related feature will increase their weights.

The dataset we have for this project is numerical. No other functions are needed in layers. It is tricky to set up the value of units, we need to adjust the values on ourselves.

The related fit function is to automatically split the dataset into training or testing data, call the loss function and reduce the loss by gradient descent method.

Evaluate method will Return the loss value metrics values for the model in test mode.

The output of this CNN shown as Figure 1.

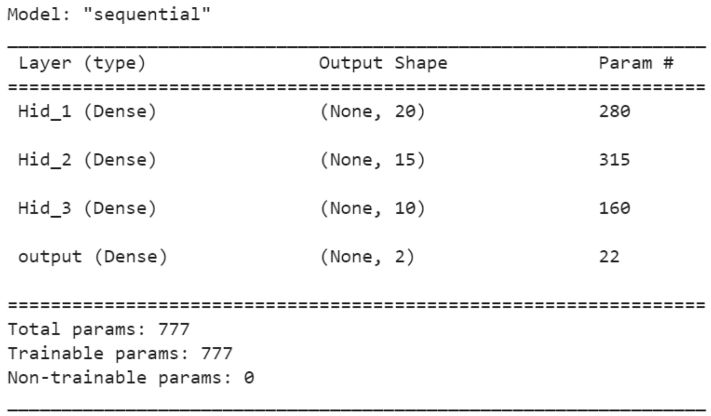
5 Experimental Evaluation

5.1 Datasets

All the datasets are found on the Kaggle website.

- Depression dataset (1432 * 23)

  The dataset is involved in the analysis of depression. The data consisted of a study about the living conditions of people who live in rural zones. The final features are shown as the following: Survey ID, Ville id, Sex, Age, Married, Number of children,



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 Hid_1 (Dense)               (None, 20)                280

 Hid_2 (Dense)               (None, 15)                315

 Hid_3 (Dense)               (None, 10)                160

 output (Dense)              (None, 2)                 22

=================================================================
Total params: 777
Trainable params: 777
Non-trainable params: 0
_________________________________________________________________
```

Figure 1: Summary CNN

Education level, Total members (in the family), Gained asset, Durable asset, Save asset, Living expenses, Other expenses, Incoming salary, Incoming own farm, Incoming business, Incoming no business, Incoming agricultural farm expenses, Labor primary, Lasting investment, No lasting investment, Depression or not.

- Breast Cancer dataset (569 * 10)

  The features in this dataset are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe the characteristics of the cell nuclei present in the image. Ten real-valued features are computed for each cell nucleus: Radius (mean of distances from the center to points on the perimeter), Texture (standard deviation of gray-scale values), Perimeter, Area, Smoothness (local variation in radius lengths), Compactness ($perimeter^2/area - 1.0$), Concavity (severity of concave portions of the contour), Concave points (number of concave portions of the contour), Symmetry, Fractal dimension (coastline approximation - 1)

- Lung cancer dataset (284 * 14)

  The cancer prediction system's effectiveness helps people know their cancer risk at a low cost, and it also helps people make the appropriate decision based on their cancer risk status. The data is collected from the website online lung cancer prediction system. The attribute information is as followed: Gender: M(male), F(female); Age: Age of the patient; Smoking: YES=2, NO=1; Yellow fingers: YES=2, NO=1; Anxiety: YES=2, NO=1; Peer pressure: YES=2, NO=1; Chronic Disease: YES=2, NO=1; Fatigue: YES=2, NO=1; Allergy: YES=2, NO=1; Wheezing: YES=2, NO=1; Alcohol: YES=2, NO=1; Coughing: YES=2, NO=1; Shortness of Breath: YES=2, NO=1; Swallowing Difficulty: YES=2, NO=1; Chest pain: YES=2, NO=1; Lung Cancer: YES=1, NO=0.

- Heart attack dataset (76 * 14)]

  This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The "target" field refers to heart disease in the patient. It is integer-valued 0 = no disease and 1 = disease. The attribute information:Attribute Information; Age; Sex; Chest pain type (4 values); Resting blood pressure; Serum cholesterol in mg/dl; Fasting blood sugar >120 mg/dl; Resting electrocardiographic results (values 0,1,2); Maximum heart rate achieved; Exercise-induced angina; Aldpeak = ST depression induced by exercise relative to rest; The slope of the peak exercise ST segment; Number of major vessels (0-3) colored by fluoroscopy; Thal: 0 = normal, 1 = fixed defect; 2 = reversible defect

5.1.1 Results

As shown in Table 1, the accuracy for each dataset has great fluctuation. After changing all the parameters, the results are still not stable. Thus, only one set of results are showing here. The number of training epoch is 10, and we use cross-validation in this project. Cross-validation will automatically split the training dataset into 80% for training and 20% for validation.

Accuracy	Diabetes	Heart	Brease$_Cancer$	Depression
Round 1	54.08%	11.22%	35.71%	99.56%
Round 2	86.22%	75.51%	48.97%	59.18%
Round 3	82.65%	62.24%	53.06%	57.14%

Table 1: Table of Target Domain Accuracy.

The test accuracy might affect by the weight matrix of CNN, which will change each time we try to get the transfer learning output. Therefore, if we want to imply the algorithm, we need to generate a relatively fixed weight matrix for each dataset and relatively stable accuracy.

However, compared to the performance among all the datasets, the remaining depression dataset works the best. The largest size of training may cause the results. The depression dataset has been split into two parts in the pre-processing period. The first 100 samples are into the test set for all the source domains, and the rest is the training set. In this way, we still could get the idea that the training dataset's input data type and size will influence the test accuracy.

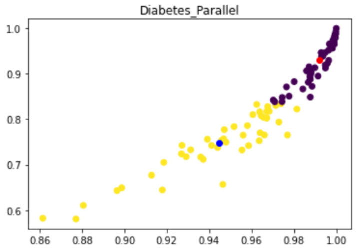


Figure 2: Parallel K-means on Linear Shape plot

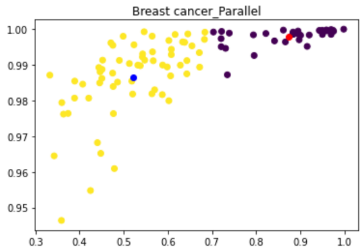


Figure 3: Parallel K-means on Scatter

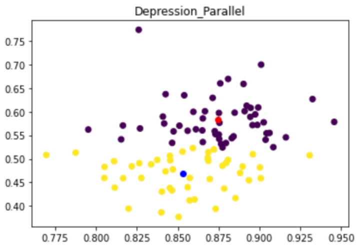


Figure 4: Parallel K-means on Ball Shape plot

Based on the figures, we could conclude that no matter how the inputs are distributed, parallel K-means always perfectly split the data into 2 clusters.

Compared to the stable high accuracy of VGG19, one of the famous image processing and classification network. The limitation of this project is that the datasets are still too small in size. Besides, the network only accepts the fixed number of inputs. Therefore, manually adjusting the datasets and randomly generating the additional columns is another problem in this project. Thus, the idea still needs to try image classification in the short future.

6 Conclusions

In this particular project, we could figure out that the selection of source domain dataset will affect the performance of the transfer learning network, and finally influence the clustering classification. Since the transfer learning network has a fixed number of input data. In the project, I create an additional supplement input data of the random repeat features or a random matrix with values bounced between 0 to 1. In this project, we clearly figure out the selection of datasets is important, which is necessary to find a dataset with the same pattern of information. Individually, transfer learning and parallel K-means methods perform better. This might be because the dataset is still limited in size. But a similar information pattern of the source domain would positively improve the test accuracy.

6.1 Summary of Contributions

This paper contributes in the following ways:

- Proposed a novel method of analysis data
- Explore the possible solution of clustering map-reduced methods.

6.2 Future research

- The images will be the next kind of dataset for this algorithm, we will introduce the transfer learning method into disease-related radiation image classification.
- Compare the performance with the VGG19 network, which is the most famous image classification neural network. To test if the combination of parallel computing and transfer learning will have better performance than the VGG19 network.

References

[1] Asarbaev A. Parallel-implementation-of-k-means, 2018.

[2] Kim A. Parallel k-means from scratch, 2020.

[3] Joonas Hämäläinen, Tommi Kärkkäinen, and Tuomo Rossi. Scalable initialization methods for large-scale clustering. *arXiv preprint arXiv:2007.11937*, 2020.

[4] Anil K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR).

[5] Dinan TG-Cryan JF. Anxiety Lach G, Schellekens H. Depression, and the microbiome: A role for gut peptides. *Neurotherapeutics*, 15:36–59, 2018.

[6] Yang Z. Qiu H.-Wang Y. Jian L. Ji J. Li K Li, J. Anxiety and depression among the general population in china at the peak of the covid-19 epidemic. *World psychiatry: official journal of the World Psychiatric Association (WPA)*, 19(2):249–250, 2020.

[7] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[8] Platzer M. Kohlhammer-Dohr A. Hamm C. Mörkl S. Bengesser S. A. Fellendorf F. T. Lahousen-Luxenberger T. Leitner-Afschar B. Schöggl H. Amberger-Otti D. Wurm W. Queissner R. Birner A. Falzberger V. S. Painold A. Fitz W. Wagner-Skacel J. Brunnmayr M. Rieger A. Dalkner N. Reininghaus, E. Z. Supplementary probiotic treatment and vitamin b7 in depression-a randomized controlled trial. nutrients. *Nutrients*, 12(11):3244, 2020.

[9] McLaughlin Benjamin R. S. and Kang Sung Ha. A new parallel adaptive clustering and its application to streaming data. pages 1–25, 2021.

[10] Abhinav Sagar and J Dheeba. On using transfer learning for plant disease detection. *bioRxiv*, 2020.