Exploring Gaussian Mixture Model and K-Means on Breast Cancer DataSet

Wenbo Han
wenbo.han@uwaterloo.ca
University of Waterloo
Waterloo, ON, Canada

Abstract

The project aims to explore the Wisconsin Breast Cancer (Diagnosis) Data Set in two clustering algorithms, namely Gaussian Mixture Model (GMM) and K-Means. With the advent of machine learning technologies, the researches show that the computer-aided applications in such advanced algorithms perform much better than traditional manual operations no matter in quality and efficiency. In addition, these intelligent algorithms are widely applied to the real-world industry to solve the complicated problems that human can not resolve manually. In this project, two algorithms are implemented respectively. The performance of these two models is compared as well. The experimental results show that K-Means outperforms the GMM model in running time but the GMM model has a more accurate prediction.

Introduction

Breast Cancer starts from the cells of breast, these malicious cells grow out of control and form a cancerous tumor which can colonize other cells, destroy nearby tissue and eventually metastasize to other parts of the body. Breast cancer usually occurs in female patients but some male patients suffer this disease as well. It is considered as one of the most common invasive cancers due to its high incidence and serious fatality rate. There are more than one million new suffers increased every year around all over the world and approximate 60% of them are died due to this disease. In Canada, it is the most common cancer among Canadian women and it causes the second most death from cancer in Canadian women. According to the annual report in 2017 from the Canadian Cancer Society, breast cancer takes around 25% of new cancer incidence cases and about 13% mortality cases among all kinds of cancers.

Although the above statistics illustrate the astounding results, the reassuring news is that the early detection of breast cancer can provide a high possibility of its cure. Starting from the early 2000s, some detection techniques such as mammography were applied into the treatment of breast cancer which gained significant achievements. However, the cultivation of an experienced mammography technologist takes a huge amount of cost in both time and resources. In addition, the accuracy usually floats at the range from 78% to 83% which is not perfectly accurate. With the advent of data science in the medical aspect, many advanced machine learning technologies are applied to the medical informatics to help to prevent the development of such serious disease. Many researches have already demonstrated that the computer-aided program has the same or better performance compared with the traditional detection technology(Palaniappan and Awang 2008; Thomas Martini Jrgensen and Jemec 2008). If a comprehensive method to detect the disease is developed and applied to make the early diagnosis, the survival rate of breast cancer will be enhanced tremendously.

This project aims to construct some machine learning models to predict the diagnosis of breast cancer with some given features. Two clustering algorithms, Gaussian Mixture Model (GMM) and K-means, are chosen as principle technologies which are implemented by Expectation-maximization algorithm. The Breast Cancer Wisconsin (Diagnostic) Data Set is separated as both training set and validation set to explore the data inside. In this experiment, accuracy, converge time and the number of iterations are the required measurements to evaluate the performance of each model. With the help of machine learning algorithms, the diagnosis with higher accuracy can be obtained and the relationship among features in original dataset behind the scene can also be figured out. More importantly, an enormous amount of resources will be saved if this technology is applied to the real-world industry. From this experiment, we have distilled three principal findings in the process of implementing these two algorithms:

- Both K-means and GMM algorithms are sensitive to the initial parameters. It does not only have a great influence on the accuracy of prediction result, but affecting the running time as well.
- Some features in original dataset are highly correlated to each other which provide the redundant information. In order to improve the quality of training process and reduce the cost of computation, the dataset can be optimized by removing such features.

- Although GMM has more complicated structure than K-means, it is more sophisticated to deal with the data with complex relationship.

Comparing with the traditional detection technologies, the advanced computer-aided models do not only improve the accuracy of predictions, but also save a huge amount of resources for real-world industry. The work makes the following technical contributions:

- Implementing machine learning model prototypes in GMM and K-means respectively.
- Comparing the performance of both models on Wisconsin Breast Cancer Data Set.
- Reducing the size of data set by evicting the features representing the redundant information.

Related Work

There are some prior works on applying K-means and Gaussian Mixture Model techniques respectively to make a clustering for various breast cancer datasets. In addition, some researches also involve another machine learning algorithms such as Bayesian Network, Decision Tree and Support Vector Machine which explore the same data set in this project (i.e. Breast Cancer Wisconsin (Diagnostic) Data Set). All these prediction models indicate that they have the potential to classify Breast Cancer Wisconsin (Diagnostic) Data Set.

K-Means is considered as one of the most popular clustering approaches which are frequently applied to the real-world dataset. The paper proposed by Dubey et. al. in 2016 (Ashutosh Kumar Dubey and Jain 2016) analyzed the breast cancer Wisconsin dataset in K-Means approach with different computation measurement. Based on the prediction accuracy of several measurements, distance (Euclidean and Manhattan) has the best performance overall other parameter settings. Similarly, both works implemented by Joshi et. al. in 2014 and Radha et. al. in 2014 applied K-Means on the same UCI dataset to analyze the clustering problem of breast cancer. (Jahanvi Joshi and Patel 2014; R.Radha and P.Rajendiran 2014). In Joshi's work, the K-Means model predicted 83% cases correctly and authors concluded that this algorithm is more helpful than other algorithms analyzed in the experiment (e.g. EM Method and Hierarchical Cluster Method) to the early diagnosis of breast cancer. In Radha's work, they did similar work to Dubey et. al., they seek the best classification average by using various distance measure, scoring method and initial value. In their experiment, the agglomerative clustering obtained a 100% correct classification. In addition, they suggest that the average accuracy can be improved if some referential decisions are combined.

In addition, the Gaussian Mixture Model is also widely deployed in machine learning models to solve the problems in the real industry. In the work of Prabakaran et. al. (Indira Prabakaran and Guvakova 2019), they applied a GMM-based classifier to exploit the multi-class classification from the huge amount of breast cancer data. In addition to provide a consistent classification result, it also demonstrated additional layers of information. After the test from several datasets, it obtained an average 85% prediction accuracy. Beyond that, the general flexibility is also a highlight of this model which can be extended to other clinical applications. Similarly, the work proposed by Liu et. al. (Jialu Liu and He 2010) applied both GMM and K-Means on the Breast Cancer Wisconsin Data Set. GMM obtained a 94.7% which is nearly 10% higher than the accuracy of K-Means. In addition, they also introduced a novel concept locally consistent regularizer which is based on the GMM and implemented in EM algorithm. The performance of this innovative improved model reached to 95.5%.

From the comparison of above prior works, the prediction accuracy of K-Means is not competitive as GMM, but it is more straightforward than GMM. Based on Radha's suggestion, the prediction accuracy can be raised if there are some other referential works. The work which demonstrates the integration of K-Means and other machine learning algorithms to predict some real-world problem is also studied. The work proposed by Kuo et. al. (R. J. Kuo 2014) integrated the K-Means algorithm with the artificial immune network to analyze Wisconsin Breast Cancer Data Set. The accuracy is 97.2% significant improved comparing with the sole K-Means algorithm. In addition, the work also illustrated that this algorithm consumed the least computation resource. Beyond that, Zheng et. al. combined the K-Means algorithm with support vector machine together (Bichen Zheng 2014). The innovative method obtained high accuracy in 97.3%. The running time in significant reduced without any accuracy loss by extracting six features from the original 32 features.

Beyond that, there are also some prior works analyzing the Wisconsin Breast Cancer (Diagnosis) Data Set in different algorithms. In Lucas Borges's work (Borges 2015) , he proposed to applied Bayesian Network and J48 algorithms to explore the imbalanced data set. Both models obtained good performance in 97.8% and 96.5% respectively. In addition, Polat et. al. (Kemal Polat 2007) also used the least square support vector machine classifier algorithm to predict Wisconsin Breast Cancer Data Set. After the k-fold cross-validation, the performance of the model reached 98.53% accuracy. Both works represent that the structure of Wisconsin Breast Cancer Data Set is well-defined for the clustering problem.

From the study of prior works, the Gaussian Mixture Model has a better predict performance than K-Means in terms of accuracy. On the other hand, the K-Means algorithm has a simpler interpretation and more straightforward implementation of its work. In addition, the dataset chosen in this project is also well-defined for the clustering problem.

Methodology

The Breast Cancer Wisconsin (Diagnostic) Data Set is the major resource used in this experiment which consists of 569 individual samples and each sample is the digitization of a breast mass FNA (fine needle aspirate) image. There are 30 distinct features for individual samples. These features are computed based on ten real-valued features of the individual cell nucleus in each image listed below:

- **Radius**: the distance from center to points on the perimeter of cell nucleus
- **Texture**: gray-scale values in the image
- **Perimeter**: the length of the outline of the cell nucleus
- **Area**: the amount of space taken by the cell nucleus
- **Smoothness**: local variation in radius lengths
- **Compactness**: $\frac{Perimeter^2}{Area-1.0}$
- **Concavity**: severity of concave portions of the contour
- **Concave points**: number of concave portions of the contour
- **Symmetry**: the extent of symmetry of cell nucleus
- **Fractal dimension**: coastline approximation -1

Each sample has a unique ID number to distinguish itself from other samples, and the diagnostic result "Diagnosis" is recorded as either malignant (represented as M) or benign (represented as B). In addition, other feature values, namely, the mean value, standard error and "worst" or largest mean value are derived from the real-valued features listed above. This dataset is a perfect representation to reflect cases at the period it created since one of its authors using it to report his clinical cases periodically. After several revisions, the dataset is considered as a noise-free dataset without any missing numbers. However, there are also two limitations existing in this dataset: a) The dataset contains approximately 63% of observations indicating the absence of cancer cells and 37% observations showing the cancerous cell. But the typical medical analysis distribution usually has more cancerous cases than the malignant ones. b) Some highly correlated features provide redundant information which raises the computation cost. Overall, this dataset provides a fair estimation of the prediction of breast cancer. The dataset is available to download at `https://www.kaggle.com/uciml/breast-cancer-wisconsin-data`

In this experiment, two clustering machine learning algorithms Gaussian Mixture Model (GMM) and K-means are going to be used to solve this problem respectively. Gaussian Mixture Model (GMM) can be considered as a linear combination of various Gaussian components. Each cluster in GMM follows the Gaussian distribution and consists of the grouped samples with the same parameters. The cluster is implemented by Expectation-maximization (EM) algorithm to estimate parameters in each cluster. In each cluster, the probability density function is defined as this:

$$p(x|\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}\Sigma^{1/2}} exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$$

where μ is the mean value and Σ is the covariance matrix. In this model, each Gaussian distribution is associated with some weights π_j which represents the probability of one point falls into this distribution. Then the overall probability of a single data point falling into the mixture model is expressed as:

$$p(x) = \sum_{j=1}^{K} \pi_j \cdot p(x|\mu_j, \Sigma_j)$$

where $1 \leq$ j $\leq$ K, $\sum_{j=1}^{K} \pi_j = 1$ and $0 \leq \pi_j \leq 1$.

For the given data set, the aim of this EM algorithm is to optimize the maximum likelihood function with regard to the parameter $\theta(\mu, \Sigma, \pi)$ which is

$$\theta^* = \arg\max_{\theta} p(X|\theta) = \arg\max_{\theta} \prod_{i=1}^{N} p(x_i|\theta)$$

In this algorithm, an indicator variable $z_{n,k}$ is introduced to represent if the point x_n is in cluster k, where $z_{n,k}$ has either value 0 or 1. If the point x_n is in cluster k, then the value will be 1. Otherwise, the value of $z_{n,k}$ is 0. Note that $\sum_{k=1}^{K} z_{n,k} = 1$. The variable $z_{n,k}$ is also known as the latent variable, and is helpful in simplifying the maximization of the likelihood. In this case, $p(z_{n,k} = 1) = \pi_k$. Suppose that every $z_{n,k}$ is identically and independently distributed, the we can have:

$$p(z) = p(z_{n,1}) \cdot p(z_{n,2}) \cdot ... \cdot p(z_{n,k}) = \prod_{k=1}^{K} \pi_k^{z_{n,k}}$$

Since there is only one of $z_{n,k}$. it is assigned a value of 1, with all other elements being assigned a value of 0. Therefore, the above constraint $\sum_{k=1}^{K} z_{n,k} = 1$ still holds. Every component in Gaussian Mixture Model is normally distributed, so we can have $p(x|z_{n,k} = 1) = \mathcal{N}(x|\mu_k, \Sigma_k)$. Therefore, the likelihood function based on the normal distribution can be expressed as:

$$p(x|z) = \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_{n,k}}$$

Based on above equations, the probability density function of a single data point x in Gaussian Mixture Model (GMM) can be derived by considering the probability density function of x in each cluster and the weight of each cluster in the

Gaussian Mixture Model (GMM):

$$\begin{aligned} p(x) &= \sum_{k=1}^{K} p(z) \cdot p(x|z) \\ &= \sum_{k=1}^{K} (\prod_{k=1}^{K} \pi_k^{z_{n,k}} \cdot \mathcal{N}(x|\mu_k, \Sigma_k)^{z_{n,k}}) \\ &= \sum_{k=1}^{K} \pi_k^{z_{n,k}} \cdot \mathcal{N}(x|\mu_k, \Sigma_k) \end{aligned}$$

In Bayesian statistics, the posterior probability $p(z|x)$ can be computed since the probability p(x) and the conditional probability $p(x|z)$ are already known:

$$\begin{aligned} \gamma(z_{n,k}) &= p(z_{n,k} = 1|x) \\ &= \frac{p(z_{n,k} = 1) \cdot p(x|z_{n,k} = 1)}{p(x, z_{n,k} = 1)} \\ &= \frac{p(z_{n,k} = 1) \cdot p(x|z_{n,k} = 1)}{\sum_{j=1}^{K} p(z_j = 1) \cdot p(x|z_j = 1)} \\ &= \frac{\pi_k \cdot \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \cdot \mathcal{N}(x|\mu_j, \Sigma_j)} \end{aligned}$$

The maximum likelihood function of these three parameters π, μ, Σ are able to be obtained by calculating the derivative of probability density function of GMM respectively. The parameters and latent variables are then updated iteratively by EM steps until the stopping criteria is reached. For each iteration of the EM algorithm:
In E-step, estimating the probability of a single point that each Gaussian distribution generated it by giving the data and the current value of parameters, the expected value of $z_{n,k}$ introduced above can be computed as:

$$\gamma(z_{n,k}) = \frac{\pi_k \cdot \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \cdot \mathcal{N}(x|\mu_j, \Sigma_j)}$$

In M step, modifying the parameters θ according to the expected value of $z_{n,k}$ to maximize the likelihood of the given data, we using the following equations to obtain the parameters for every subsequent iteration:

$$\pi_k = \frac{1}{N} \sum_n \gamma(z_{n,k})$$

$$\mu_k = \frac{1}{\sum_n \gamma(z_{n,k})} \sum_n \gamma(z_{n,k}) x_n$$

$$\Sigma_k = \frac{1}{\sum_n \gamma(z_{n,k})} \sum_n \gamma(z_{n,k})(x_n - \mu_k)(x_n - \mu_k)^T$$

After each iteration, the stopping criteria is examined by computing the change of log-likelihood value.

$$\log p(x) = \sum_i \log[\sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(x_i|\mu_k, \Sigma_k)]$$

In this case, the criteria is simply to check whether the updated changes is less than the threshold ϵ (the default value is 10^{-3}). If the criteria is met, then the training process is stopped. Otherwise, the parameter θ is kept to update by repeating E-step and M-step until the model is converged.

On the other hand, K-means clustering aims to divide all observations in the dataset into K (K = 2 in this case) clusters. Each observation in the same cluster has the minimum distance value to its centroid. The algorithm starts with initializing the value of K centroid. After the initialization, the algorithm will be iterated in the following steps until the stopping criteria is met:

1. Each data point is assigned to its nearest centroid based on Euclidean distance:

$$\theta^* = \arg\min_{c_i \in C} dist(c_i, x)^2$$

where x is the data point, c_i is one of the central point in centroid set C.

2. Then update the centroids by computing all data points assigned to the same cluster.

$$c_i = \frac{1}{|S_i|} \sum_{x_i \in S_i} x_i$$

If no points change clusters then the training process of the model is completed.

Implementation

In this section, the implementation of K-Means and GMM are demonstrated respectively. From prior works, the GMM and K-Means are both sensitive to initial states. In order to see the influence on various initialization, the accuracy and running time are both measured as the major benchmark. Moreover, the original dataset is shrunk by evicting features which provide redundant information. The whole project is implemented in Python 3.7 and the numpy library is utilized for the complex computation.

K-Means

As the introduction in the previous section, points are clustered by calculating the Euclidean distance to find the nearest centroid. The values of centroids are then updated by computing the average values in the same cluster. The specific algorithm is implemented as Algorithm 1. In addition, two different initialization states are configured, namely random initialization and furthest-point initialization.

Random Initialization In this configuration, the K-Means model starts with two samples which are randomly selected from the original dataset. Then the model follows the workflow of Algorithm 1 to cluster the dataset.

Data: Wisconsin Breast Cancer (Diagnosis) Data Set
Result: The list of clustered label
Initialize two centroids, clustered label list and iteration number;
while *no points change clusters* ***or*** *maximum iteration not reached* **do**
  for *each point in dataset* **do**
    assign the point to its nearest centroid
  end
  update the label list;
  compute and update two centroids;
  increase the iteration number by 1;
end
return clustered label list;

Algorithm 1: K-Means Implementation

Furthest-point Initialization Instead of selecting both centroids randomly, this configuration only keeps one the random-selected sample and computes the other centroid which has the longest Euclidean distance from the first centroid.

Gaussian Mixture Model

The GMM model is implemented by the EM (Expectation-Maximization) algorithm. In each E step, computing the posterior probability $\gamma(z_{n,k})$ is computed to estimate the probability of a point falling into a single Gaussian distribution. Then the values of μ, π and σ are updated iteratively to maximize the log-likelihood value of this single point. The specific algorithm is implemented as Algorithm 2.

Random Initialization The intuition in this configuration is to randomly initialize the μ, π and Σ in GMM. In order to be consistent with the comparison results, the μ value in GMM adopts the same value of initial centroids in K-Means Random Initialization section. Each cluster in GMM has the equal value in initial weight (i.e. π = 0.5 since there are 2 clusters). Moreover, Σ is initialized with the covariance value of the original dataset.

K-Means based Initialization The most common way of initializing the parameters in GMM is to utilize the results (i.e. final clusters' centroids) of K-Means. From the previous study, this initialization method does not only guarantee the convergence of the model, but also improves the accuracy of prediction results. The other values such as π and Σ remain the same configurations as above random initialization.

In the next section, the accuracy and running time will be measured to compare the performance of all configurations in above models.

Simplified DataSet

From the statistical perspective, a pair of high-correlated features refer that these two features provide similar information for the same label. In this case, the correlation values of every two features are examined. If the correlation value is higher than the certain threshold value (default value is 0.9), then the feature with a higher expected value between these two features will be reserved. Based on above methods, there are 10 features in original dataset which provide redundant information: perimeter_mean, area_mean, concavity_mean, perimeter_se, area_se, radius_worst, texture_worst, perimeter_worst, area_worst and concave points_worst. In this case, a robust and compact dataset can be simplified by removing these features from the original dataset.

Data: Wisconsin Breast Cancer (Diagnosis) Data Set
Result: The posterior probability matrix
Initialize the centroids (i.e. μ), Σ, π, likelihood threshold, max iteration, posterior probability matrix, clustered label list;
while *prev_likelihood* $-$ *current_likelihood* $>$ *likelihood threshold* ***or*** *maximum iteration not reached* **do**
  update the likelihood in previous iteration;
  Perform E step;
  update posterior probability matrix;
  Perform M step;
  update the likelihood in current iteration;
  increase the iteration number by 1;
end
for *each entry in posterior probability matrix* **do**
  label the sample as the cluster with higher probability;
  save the label to clustered label list;
end
return clustered label list;

Algorithm 2: GMM Implementation

Evaluation

In this section, the accuracy and running time are considered as the performance benchmark tests in both models. Specifically, the performance of various initialization in different models is going to be compared. Also, the effect of the simplified dataset is also tested by comparing with the original dataset.

Accuracy

The feature values (except "diagnosis" and "ID" attribute) of Wisconsin Breast Cancer (Diagnosis) Data Set is applied as the training data in this project. After clustering all the points in the dataset, the predicted results are compared with the corresponding true label ("diagnosis" attribute) of the original dataset.

In Figure 1, the accuracy of both models in different initialized states is measured. In this figure, both random ini-

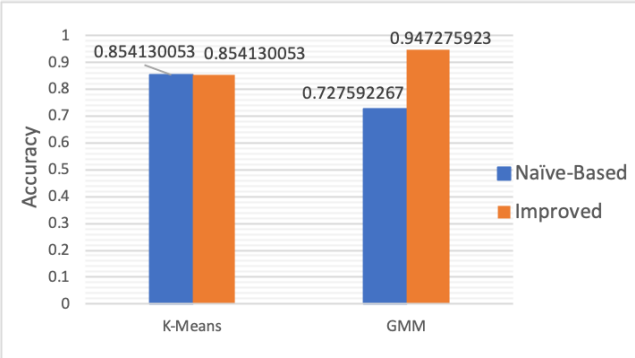

Figure 1: Accuracy of various models in original dataset

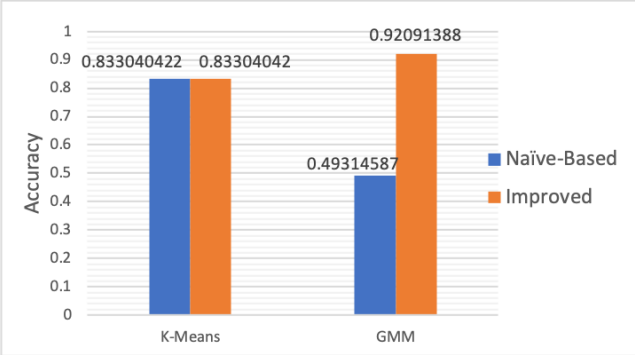

Figure 2: Accuracy of various models in simplified dataset

tialization states in two models are called the naive-based method and the furthest-point initialization and K-Means based initialization are named as the improved method. Both random initialization methods apply the same initial centroids values. The improved GMM method deploys the final centroids from K-Means as the initial μ value. In original dataset, K-Means has the same accuracy 85.4% in both initialization states. However, the accuracy of GMM in improved initialization 94.2% outperforms the one in naive-based initialization 72.8%.

Figure 2 demonstrates the accuracy of both models which are trained in two initialized states with the simplified dataset. Compared with the results in Figure 1, the accuracy is only dropped 2%, but the size of data shrinks to 66.7% of its original size. The more obvious benefit will be stated by comparing the running time in the following part. As a result, the simplified dataset is qualified as a robust training set.

Both figures above illustrate that the GMM is more sensitive to the initial states since the accuracy in K-Means remains the same but the one in GMM differs from 23% to 46%. In addition, it also can be concluded that the GMM in a stable state (improved version) is approximate 10% more accurate than K-Means model.

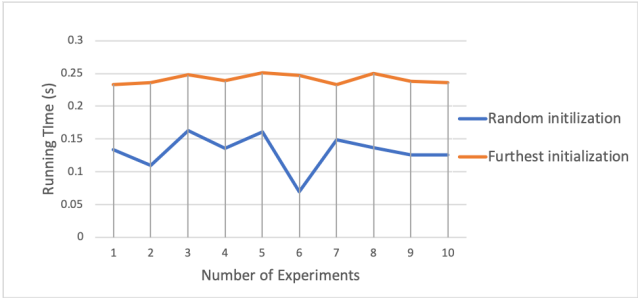

Figure 3: Running Time of K-Means in original dataset

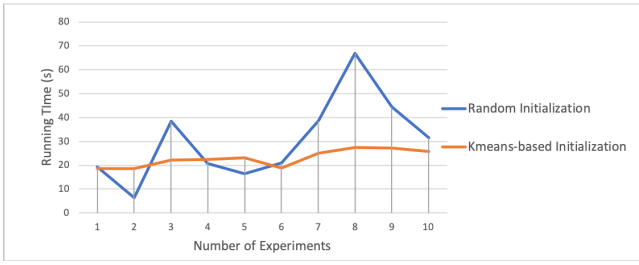

Figure 4: Running Time of GMM in original dataset

Running Time

Figure 3 and Figure 4 demonstrate the running time of both models in different initial configurations respectively. In Figure 3, the running time of furthest initialization is approximate 2 times of the one with random initialization. The longer running time in furthest initialization may be caused by the outliers in the dataset. On the other hand, two initial states in GMM have a significant difference in running time. The average running time of K-Means based initialization in 10 experiments takes only about 74% of the one with random initialization. From Figure 1 and Figure 4, it is obvious to see that the K-Means based initialization does not only outperform the random initialization in the degree of accuracy, but also have a better performance on running time.

In addition, the running time of both models on various versions dataset is also compared. In Figure 5, the advantage in the simplified dataset is not that obvious in K-Means model. It is mainly because the constraints (terms in calculating Euclidean distance) in the simplified version are less than ones in the original dataset. As a result, there can be more iterations than random initialization. However, the running time in each iteration reduces due to the fewer feature computations. Therefore, the overall running time in each experiment is not affected very much. On the other hand, it is known that the GMM model involves a lot of matrix computation which consumes substantial time. In this case, evicting some features from the original dataset is equivalent to reducing the dimension of matrices. Consequently, the running time of GMM in simplified version is signifi-

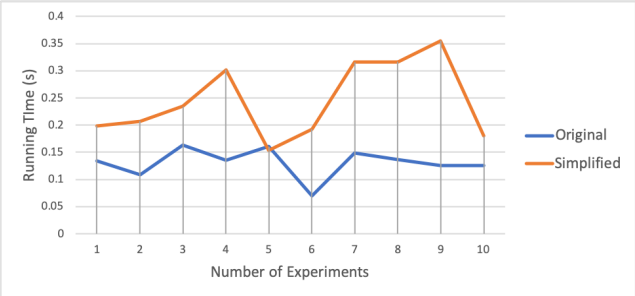

Figure 5: K-Means in original dataset vs. simplified dataset

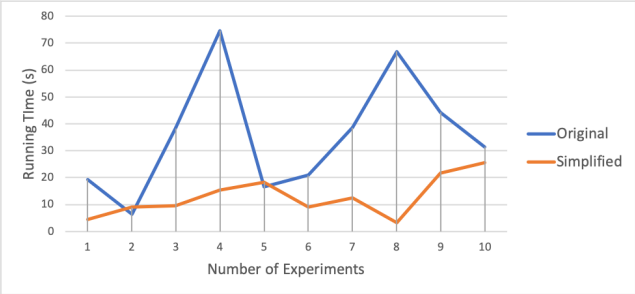

Figure 6: GMM in original dataset vs. simplified dataset

cantly lower than the one in the original dataset.

Discussion and Future Work

Both K-Means and GMM has decent performance on clustering the Wisconsin Breast Cancer (Diagnosis) Data Set which has 85% and 94% accuracy respectively. This result matches the prior work of Liu et. al. exactly (Jialu Liu and He 2010). Both models can separate the dataset into two distinct subsets based on their feature values. However, these two models apply totally different methods to achieve this similar goal. In K-Means, the data points are clustered based on the Euclidean distance. On the other hand, GMM labels data points into different clusters by estimating the parameters of individual Gaussian models. More importantly, K-Means is considered as a ”hard” clustering model which labels every data point into one of the clusters definitely. Nevertheless, instead of putting data points into some clusters directly, GMM provides a ”soft” clustering for the dataset. GMM returns the probability of every single point in different clusters. In other words, it tells the cluster that the data points are more likely located rather than assign them to one of the clusters directly. Thus, the ”soft” clustering comes with higher accuracy and it is better at dealing with some more complex datasets than the ”hard” clustering. The experimental results from the previous section show that the GMM has a more accurate prediction compared with K-Means but it consumes more time than K-Means. It is considered as a trade-off of these two models. In addition, the accuracy of GMM performance as we expected. In other words, the model trained with K-Means based initialization has a far better performance than the randomly selected one. Beyond that, the model running in simplified dataset consumes less time than the one in the original dataset. On the other hand, the accuracy of K-Means from random initialization and more logical initialization (furthest-point initialization) is measured as well. Theoretically, the accuracy of the logically selected one is expected to have a higher value than the randomly selected one. However, the practical results show that both methods have the same accuracy. It may be caused by the limited numbers of sample data. The original Wisconsin Breast Cancer (Diagnosis) Data Set only contains 569 individual samples which are a little bit fewer than the most training sets nowadays. Integrating multiple data sources into the current training set is considered as future work. Moreover, both models trained with simplified version dataset have a satisfying performance which shrinks the running time significantly (especially in GMM about 66% reduction) with only 2% accuracy loss. It turns out that all features in dataset provide unique information but may be very similar to each other. In this case, the method of evicting highly correlated features is an acceptable strategy if the dimension of the dataset is extremely large and the model is error-tolerated.

Conclusion

The Wisconsin Breast Cancer (Diagnosis) Data set is explored by implementing K-Means and GMM clustering algorithms. Both models achieve satisfying results with 85.4 % and 94.7 % accuracy respectively. These two unsupervised learning models utilize the historical data to predict the diagnosis based on the features of digitized FNA image with high precision. The GMM model consumes more training time to converge due to its complex computation. However, this model has a better performance compared with K-Means when it deals with more complicated dataset due to its soft assignment. In addition, around 33.3% features are removed from the original dataset which reduces the converge time significantly (around 66.7 % reduction on the original time) without affecting the accuracy of prediction results. There could be some more future works to extend this project. As discussed in the previous section, this data set has only 569 individuals. Although this data set can be used to train some learning models effectively, integrating some more samples into the dataset can make the data set support more powerful analysis. Besides, both initial methods in K-Means contain randomly selected factors. The future work can focus on initializing centroids with some more logical methods, such as the K-Means++. Moreover, the Wisconsin Breast Cancer (Diagnosis) Data Set has already been proved which can be deployed to machine learning problems. As a result, there can be more advanced models involved to make a comparison of their performance.

References

Ashutosh Kumar Dubey, U. G., and Jain, S. 2016. Analysis of k-means clustering approach on the breast cancer wisconsin dataset. *Int J Comput Assist Radiol Surg*.

Bichen Zheng, Sang Won Yoon, S. S. L. 2014. Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms. *Expert Systems with Applications* 41(4):1476–1482.

Borges, L. 2015. Analysis of the wisconsin breast cancer dataset and machine learning for breast cancer detection. In *Workshop de Viso Computacional*.

Indira Prabakaran, Zhengdong Wu, C. L. B. T. S. S. G. K. P. J. Z., and Guvakova, M. A. 2019. Gaussian mixture models for probabilistic classification of breast cancer. *ACCR*.

Jahanvi Joshi, R. D., and Patel, J. 2014. Diagnosis of breast cancer using clustering data mining approach. In *International Journal of Computer Applications*, Volume 101 No.10.

Jialu Liu, D. C., and He, X. 2010. Gaussian mixture model with local consistency. In *AAAI'10 Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, 512 – 517. ACM.

Kemal Polat, S. 2007. Breast cancer diagnosis using least square support vector machine. *Digital Signal Processing* 17(4):694 – 701.

Palaniappan, S., and Awang, R. 2008. Intelligent heart disease prediction system using data mining techniques. In *2008 IEEE/ACS International Conference on Computer Systems and Applications*.

R. J. Kuo, S. S. Chen, W. C. C. C. Y. T. 2014. Integration of artificial immune network and k-means for cluster analysis. *Knowledge and Information Systems* 40(3):541–557.

R.Radha, and P.Rajendiran. 2014. Using k-means clustering technique to study of breast cancer. In *World Congress on Computing and Communication Technologies*. IEEE.

Thomas Martini Jrgensen, Andreas Tycho, M. M. P. B., and Jemec, G. B. 2008. Machinelearning classification of non-melanoma skin cancers from image features obtained by optical coherence tomography. *Skin Research & Technology* 14(3):364–369.