CS 4641/7641 Assignment 3: Unsupervised Learning

Sitong Wu

School of Electrical and Computer Engineering Department

Georgia Institute of Technology

Atlanta, GA, USA

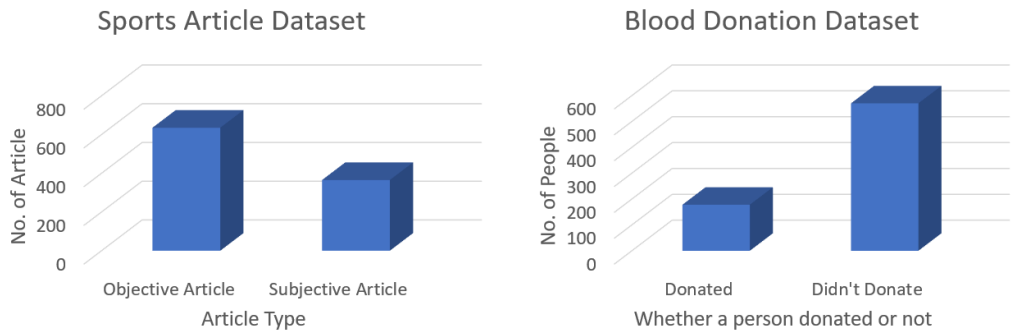
swu321@gatech.edu

Abstract – This is a report about three difference unsupervised learning algorithms. They are k-means. clustering, GMMs, and PCA.

I. INTRODUCTION

The first dataset is about whether a sports article is objective or subjective. People read articles everywhere and everyday online. The objective content would give readers general and comprehensive information that help readers understand what happened sufficiently. On the opposite side, the subjective article which might come with biases or preferences from the author. Or the author might partially report the truth with hiding the others to prove the personal idea for any purposes. In the subjective article case, those articles might annoy critical thinking readers or may blind folks who do not have back ground of the stories. Thus, a technology that can pre-filter those subjective articles away and only leave the objective ones may be necessary for a more efficient and more convenient reading experience. The dataset is only from sport article perspective. It contains 1000 samples (1000 articles). Each sample has 59 features, such as word count and quotes count. By investigate different features of those sports articles, it might be possible to find a way to label any sports article as objective or subjective. The dataset is illustrated in figure 1.a.

The second dataset is about whether a person did blood donation in March 2007. Blood donation tightly relates to blood storage as well as lifesaving. It shows how fast a city or a community could respond to any emergencies. Analyzing this data also helps investigate how enthusiastic the people are to participate into the philanthropy events. This blood donation data has 749 samples with four attributes include blood donation frequency and the total amount of blood they had donated. Like the first data set, this is also a double-class dataset. This dataset is shown in figure 1.b below.



a b

Figure. 1. a. Objective/subjective sports article dataset; b. blood donation dataset

Both datasets are idea for machine learning, because they do not have a super large number of samples in one class and have very trivial number of samples in the other class, as shown in figure 1. About the sports article dataset, it is kind of biased. The reason is that it is subjectively judged by people whether an article is objective or subjective. Contrarily, the second data set is about the absolute truth.

In the unsupervised learning parts, the process is running on the whole datasets. In the neural network parts, sports article dataset is used, and it needs to be divided so that cross validation can be utilized. The 80% data is for training which would be divided further, which are 80% as training data and 20% as verification data. The machine learning technique is trained based on the training data and is improved by utilizing the verification data. The left 20% of the whole dataset is the testing data which is used for evaluating the behavior of a machine learning algorithm.

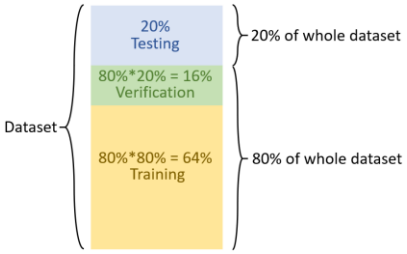


Figure. 2. Training data, verification data, and testing data

For the full dataset of sports article, the 1000 data is divided into 639 training data, 161 verification data, and 200 testing data. For the full dataset of blood donation, the 748 data is divided into 479 training data, 119 verification data, and 150 testing data.

II. Performance of Three Unsupervised Learning Algorithms on Two Datasets

The cluster algorithms are applied to both datasets for best understanding of both unsupervised learning algorithms. The data in both datasets are grouped together in a way that the data within the same group have the similar instances. Besides, principle Components Analysis is used for reducing the dimension of dataset. The definition and measurement vary per algorithm. Several tests are done to find the best cluster numbers for each algorithm.

A. K-mean Clustering

The most significant factor of choosing the best cluster number of K-mean is the within cluster squared error. As the number of clusters increases the within cluster squared error decreases, which is shown by the orange line in figure 3 and K is the cluster number. This makes sense because as the cluster number increases, the samples in each cluster decreases. Therefore, the sample in each cluster would be more close to the center of the belonging cluster. From the plot, it is clear to see that this decreasing is larger at the beginning and then become more linear as the cluster number become large enough. For the sports article dataset, this number is around 6. Another factor that might help to find the proper cluster number is the number of significant clusters. A significant cluster is the one that has at least 10% of the whole data. This is under consideration because clusters that identify a negligible number of instances are not as interesting and might not help explain the whole data generation. As the cluster number increases, the significant cluster number never exceeds 4, as figure 4 shows. Following the Occam's Razor principle of choosing the simplest solution, 6 seems like the best choice from the experiment which also match the result with considering the within cluster sum of squared error.

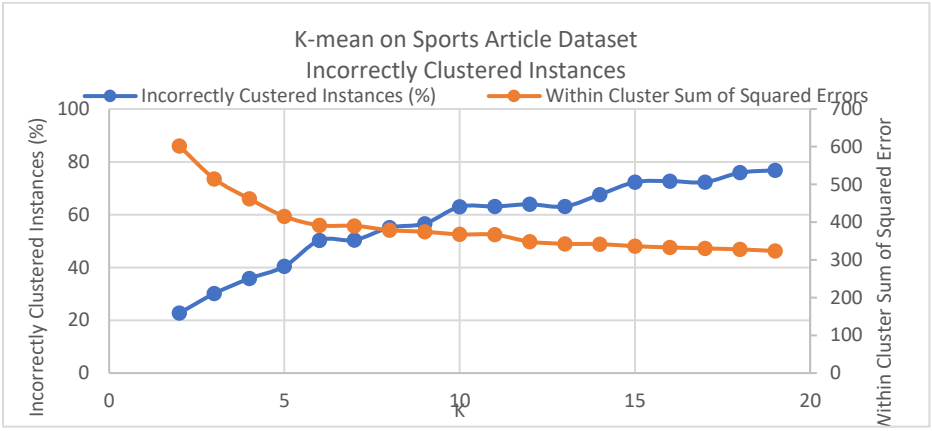


Figure. 3. K-mean on Sports Article Dataset about Incorrect Clustered Instances and Within Cluster Sum of Squared Error

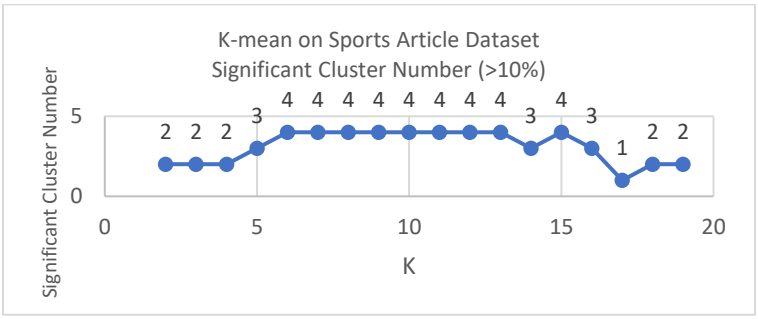


Figure. 4. K-mean on Sports Article Dataset about Significant Clusters

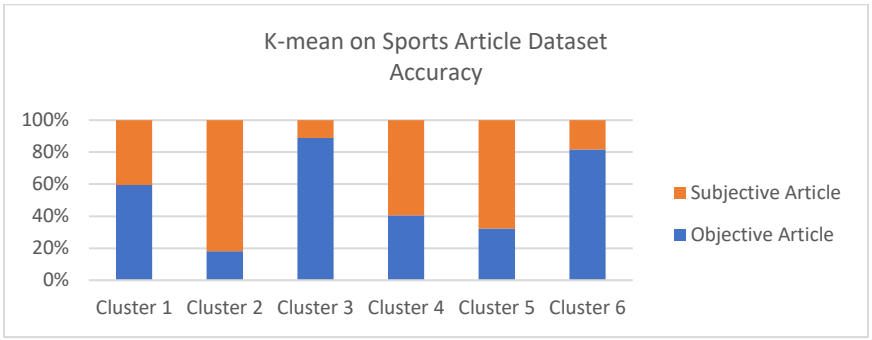


Figure. 5. K-mean on Sports Article Dataset Clusters with K = 6

Figure 5 is the distribution of each clusters within the six clusters. The charts do a kind of good job at discriminating between subjective article and objective articles. The only chart that might have a larger errors stand for cluster 1 and cluster 4, which have the distribution of each type close to 40% vs 60%.

Similar process is applied to the blood donation dataset. The results are in figure 6, figure 7, and figure 8. The best cluster number from thoses plot is 7 by looking at both figure 6 and figure 7. Similar case happens with the blood donation dataset, there are two clusters are not as good as the others, which are cluster 5 and cluster 7. Especially in cluster 7, the distribution is half-half; however, there are only 14 samples inside cluster 7. Considering and whole dataset population of blood donation is 748. And from figure 7, it is known that there are 6 significant clusters out of 7. So this bad cluster 7 is definitely the lest significant one that is no need to be worried.

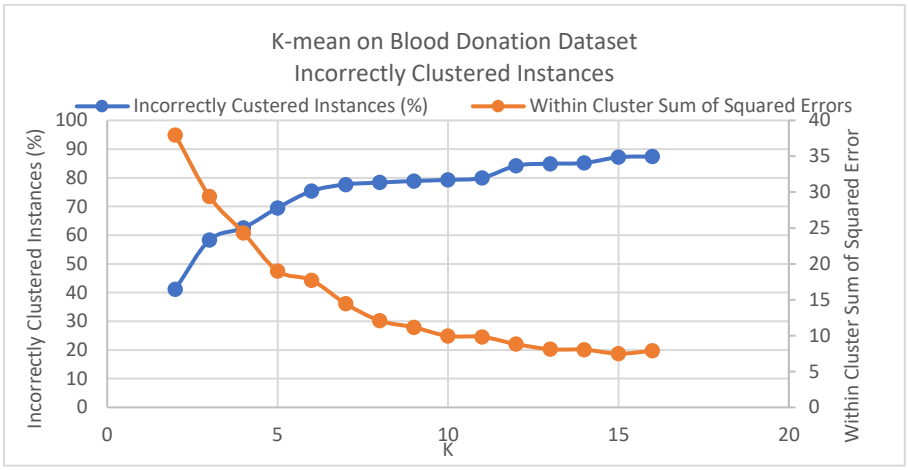


Figure. 6. K-mean on Blood Donation Dataset about Incorrect Clustered Instances and Within Cluster Sum of Squared Error

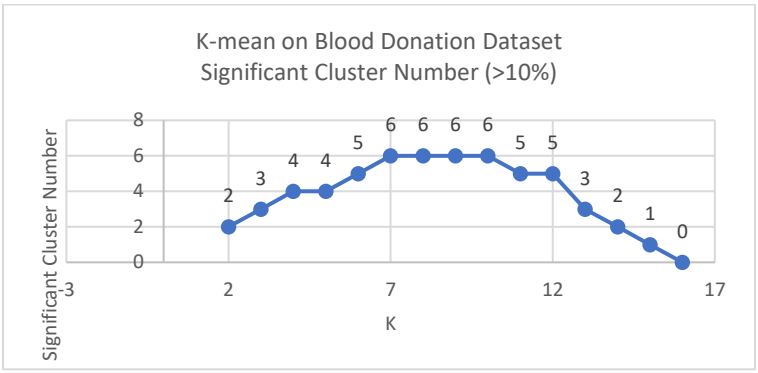


Figure. 7. K-mean on Blood Donation Dataset about Significant Clusters

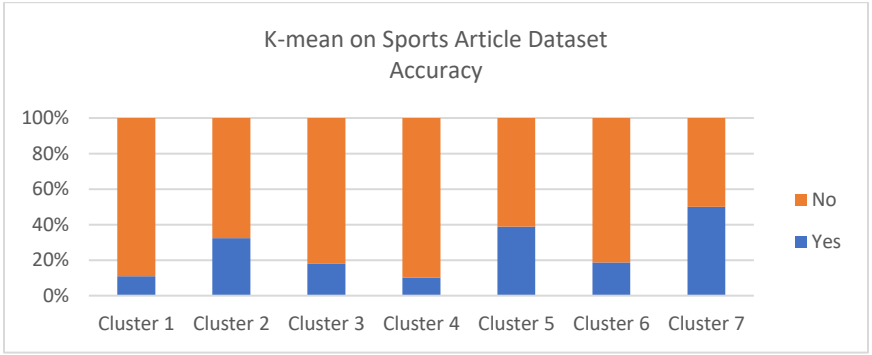


Figure. 8. K-mean on Blood Donation Dataset Clusters with K = 7

Weka was used for this part of experiment. The simple K-mean algorithm was applied here with the default Euclidian distance.

There is another factor which is plotted in figure 3 and figure 6, which is the incorrect cluster instances percentage. This factor is not the main factor to find the cluster number. The way Weka generate this variable is only consider two classes in the clusters generated as the classification cluster. Both datasets used in this experiment only have two classes. For sports article case, they are objective and subjective; for blood donation case, they are YES and NO. The data falls inside those two clusters would be considered as been classified. All the samples inside the classification clusters which are wrongly classified and all the samples that are out side

the classification clusters would be summed up as the incorrect cluster instances. This is only the calculation from Weka and unsupervised learning should not consider the label. So this factor is not very useful.

B. GMM

This experiment was run with Scikit-learn. Bayesian Information Criterion is used to evaluate the behavior of Gaussian Mixture Models. BIC stands for Bayesian information criterion or Schwarz criterion. It utilizes the likelihood function. Generally, the lower the BIC value, the better the clustering. When donging the BIC analysis for GMM, only the cluster number with the first local minimal BIC number is chosen as the proper cluster number. Though it is not shown in figure 9, the value will eventually decrease as the cluster number increases, which sometimes might go negative. The general decrement tendency will continue until the cluster number is equal to the sample population number. It is noticed that the case cluster number equal to the total data number is not practical and useful at all. It is also not practical to have a super larger cluster number to achieve a small BIC value; for instance, 500 clusters for a 1000-sample dataset with a very low BIC. Therefore, the first local minimal BIC is chosen to locate the cluster number. Figure 10 shows the plot of log-likely hood which also can help to determine the proper cluster number. Figure 9 and figure 10 both agree that 3 is a good value in this case.

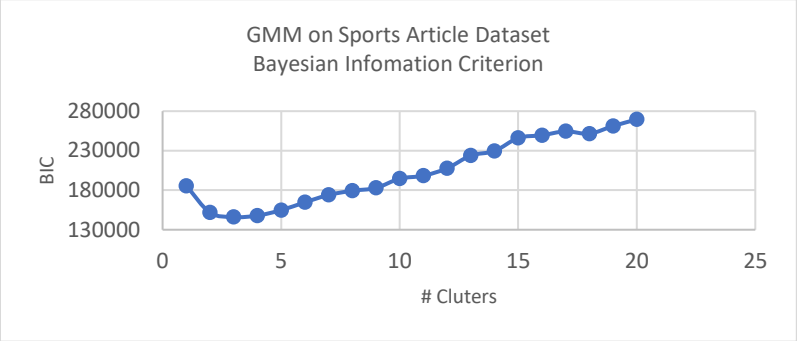


Figure. 9. GMM on Sports Article Dataset BIC

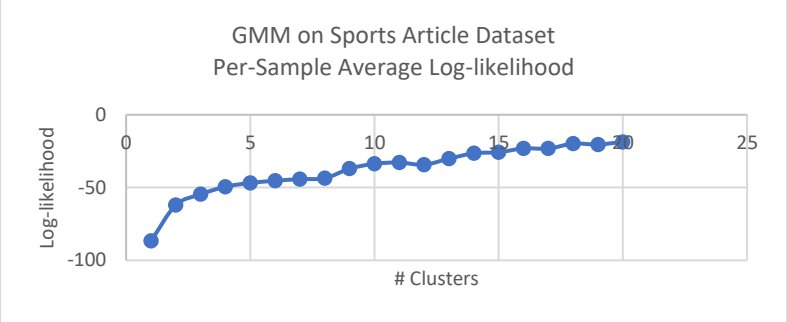


Figure. 10. GMM on Sports Article Dataset Per-sample Average Log-likelihood

Go through the same process for the blood donation dataset, the results are shown in figure 11 and figure 12. The local minimal BIC and the maximum log-likelihood is found when the cluster number is 7.

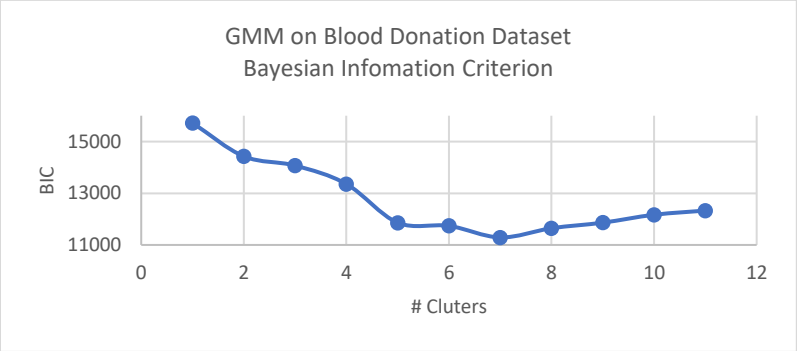


Figure. 11. GMM on Blood Donation Dataset BIC

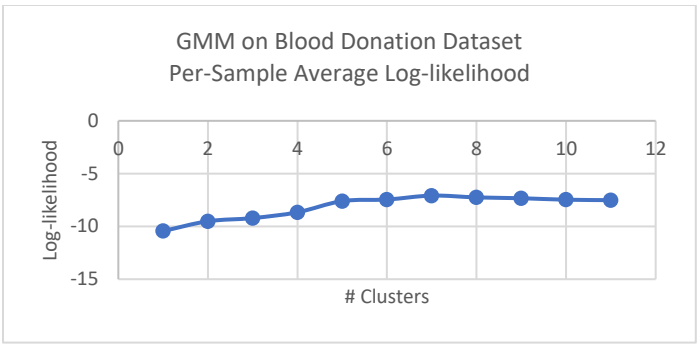


Figure. 12. GMM on Blood Donation Dataset Per-sample Average Log-likelihood

C. *PCA*

Principle Components Analysis allows for a marginal decrease in the number of dimensions for the dataset. It orthogonal project data onto lower-dimensional space that maximizes variance of projected data and minimized mean squared distance between data point and projections.

For the sports article dataset, the original 59 features are transformed into 30 features with variance coverage 95%. Figure 13 shows all the 30 features. Each line in the figure 13 is a feature. Notice that features are ranked, and they have different weight. The heavier the feature, the more significant it is. Another thing about those new features is that each new feature is a linear combination of several old features. For example, the most important new feature, which has the heaviest weight 0.4941, is formed by 0.184 of total words count, 0.178 of CD, 0.178 FW, and several other old features. The new features which have weight less than 0.1 will not be considered due to the reason that they are not as important as the rest. Another reason is that 30 features is still a large number.

Ranked attributes:
0.4941 1 0.184totalWordsCount+0.178CD+0.178FW+0.178semanticobjscore+0.176baseform...
0.4401 2 0.445Quotes+0.43 RP+0.276past+0.275VB+0.231CC...
0.4035 3 0.326colon+0.326CC+0.29 PDT-0.255RP-0.249POS...
0.3747 4 0.502txtcomplexity-0.363exclamationmarks-0.287ellipsis-0.275questionmarks+0.197past...
0.3501 5 0.538ellipsis+0.534exclamationmarks+0.218VB+0.216past-0.193present3rd...
0.3282 6 0.495semicolon+0.323VBZ-0.294NN+0.262WP+0.26 RB...
0.3082 7 0.683JJS+0.334colon-0.259RB-0.227RBR-0.209compsupadjadv...
0.2896 8 -0.656sentencelast-0.601sentence1st+0.171pronouns1st+0.141TOs+0.129colon...
0.2716 9 -0.538TOs-0.411WP+0.373ellipsis+0.245JJS-0.202sentencelast...
0.2542 10 0.625sentence1st-0.484sentencelast-0.293JJS-0.214TOs+0.185PDT...
0.2375 11 -0.373TOs+0.366WP+0.349sentencelast-0.342sentence1st+0.272NN...
0.2212 12 -0.497RBR-0.336WP-0.289JJS+0.235sentencelast+0.219CC...
0.206 13 -0.385RBR-0.357semicolon+0.296WP+0.245TOs+0.229sentencelast...
0.1919 14 -0.453WP+0.337TOs+0.254RBR-0.251JJS+0.229semicolon...
0.1785 15 -0.621NNS+0.334pronouns1st+0.249RBR+0.23 RB+0.203JJS...
0.1655 16 -0.552NN+0.393colon+0.368RBR+0.189pronouns1st+0.181NNS...
0.1535 17 -0.335NNS+0.288pronouns3rd+0.272DT+0.268ellipsis+0.252TOs...
0.1422 18 0.379NN-0.377DT+0.301semicolon-0.255VBZ+0.235RB...
0.1315 19 -0.316DT+0.302EX+0.293txtcomplexity-0.249NN-0.241questionmarks...
0.121 20 0.592JJR-0.407RB-0.218PDT+0.187exclamationmarks+0.175WP$...
0.1113 21 0.332txtcomplexity+0.327colon-0.303CC-0.292RBS-0.287pronouns1st...
0.1023 22 0.465PDT-0.319RB-0.275colon-0.247WP$-0.218EX...
0.0937 23 0.375EX+0.329DT+0.324exclamationmarks-0.254questionmarks-0.224WDT...
0.0852 24 0.58 pronouns2nd-0.378RBS+0.241PDT-0.218pronouns1st-0.19ellipsis...
0.0772 25 0.394DT+0.36 RBS+0.255pronouns2nd-0.226pronouns1st+0.22 colon...
0.0697 26 0.358JJ-0.356RBS-0.315RB+0.279txtcomplexity+0.279questionmarks...
0.0629 27 -0.329exclamationmarks-0.255RBS-0.252questionmarks-0.251WP$+0.239imperative...
0.0566 28 -0.366JJR+0.339colon-0.321questionmarks+0.311JJ-0.266WDT...
0.0506 29 0.387WP$-0.346JJ-0.311WDT-0.272VBZ-0.258EX...
0.045 30 -0.517WDT-0.436LS+0.333pronouns2nd+0.269VBZ-0.232exclamationmarks...

Figure. 13. PCA on Sports Article Dataset features of 95% Variance Coverage

VC = 0.95	
Attributes	StdDev
Attribute 1	5.37
Attribute 2	1.753
Attribute 3	1.446
Attribute 4	1.28
Attribute 5	1.186
Attribute 6	1.117
Attribute 7	1.066
Attribute 8	1.031
Attribute 9	1.011

Figure. 14. PCA on Sports Article Dataset SteDev of features

Another factor was considered here is the standard deviation of each features. The biggest standard deviation is 5.37. Based on this biggest standard deviation value, any new features with a standard deviation less than 1 will not be considered here. Therefore, only nine features left. They are shown in figure 14 with its corresponding standard deviation value.

There is an interesting factor that are worthy to be discussed here, which is the variance coverage. The variance of the dataset reflects how many information contains in the features. When there are 100% variance be covered, the feature number is the original feature number which is 59 in the sports article dataset case. The variance coverage drops to 0.95 will generate the data in new spaces that only contain 30 features in total. As the variance coverage decreases further, the number of features required to illustrate the information would also decrease further until reaches 1. Figure 15 describes such relationship. As the variance drop from 1 to 0.9, feature number drops by 36; as the variance drops from 0.9 to 0.8, feature number drops by 9; as the variance drops from 0.8 to 0.7, feature number drops by 6; as the variance drops from 0.7 to 0.6, feature number drops by 4; as the variance drops further, the drop of feature number will drop less and less until reach 0. Another aspect to explain this phenomenal is by the elbow theory graph. Choose feature number 9 as the point that between large variance change and small variance change is reasonable.

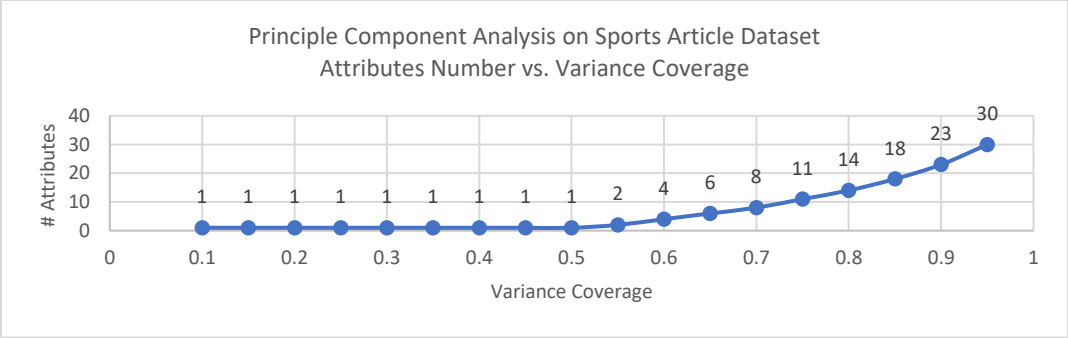


Figure. 15. PCA on Sports Article Dataset Attribute number vs. Variance Coverage

Similar results are got for the blood donation dataset. Notice that different from sports article dataset, blood donation dataset originally only has 5 features. After PCA of 95% variance coverage, there are only 3 features. The standard deviation of these 3 features are also kind of close, so all of them are kept.

Ranked attributes:

0.3647389769955831	1	-0.611Frequency (times)-0.611Monetary (c.c. blood)-0.494Time (months)+0.093Recency (months)
0.08939850714038855	2	0.915Recency (months)+0.384Time (months)-0.085Monetary (c.c. blood)-0.085Frequency (times)
-0.000000000000000222	3	0.78 Time (months)-0.391Recency (months)-0.345Monetary (c.c. blood)-0.345Frequency (times)

Figure. 16 PCA on Blood Donation Dataset features of 95% Variance Coverage

VC = 0.95	
Attributes	StdDev
Attribute 1	1.594
Attribute 2	1.049
Attribute 3	0.598

Figure. 17 PCA on Blood Donation Dataset SteDev of features

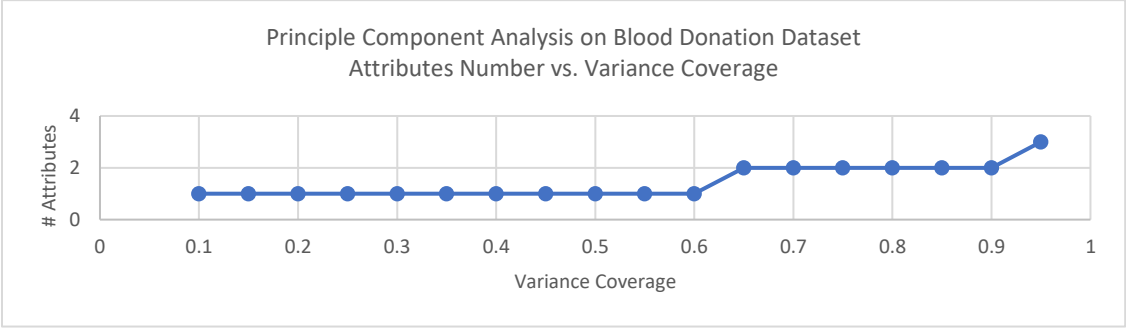


Figure. 18. PCA on Blood Donation Dataset Attribute number vs. Variance Coverage

III. Performance of K-means and GMM via PCA on Two Datasets

This part of experiment is to run PCA on the dataset first. Based on the result of previous PCA experiment, generate the best new dataset with new dimension. Then, run the cluster algorithms on the new datasets.

A. K-mean via PCA

For sports article dataset, the best feature number is 9. After the new dataset is generated, run the same tests as previous K-mean experiment. From figure 19 and figure 20, 6 seems to be the best choice. Consider that after dimension reduced by PCA, less information might make it more difficult to do K-mean. From the within cluster sum of square error plot, the proper number is 6; however, when K equals 9, it has the maximum significant cluster number 6, which is the only highest one. Consider the lack of information which might cause the tolerance, the numbers in the range of 5 to 10 are all acceptable. Here, 6 is chosen. Even after dimension reduction by PCA, the cluster number after K-mean is the same as that of the sports article dataset only going through K-mean. The behavior of each cluster when K=6 is better. Almost every cluster shows a majority of one the article type, as shown in figure 21.

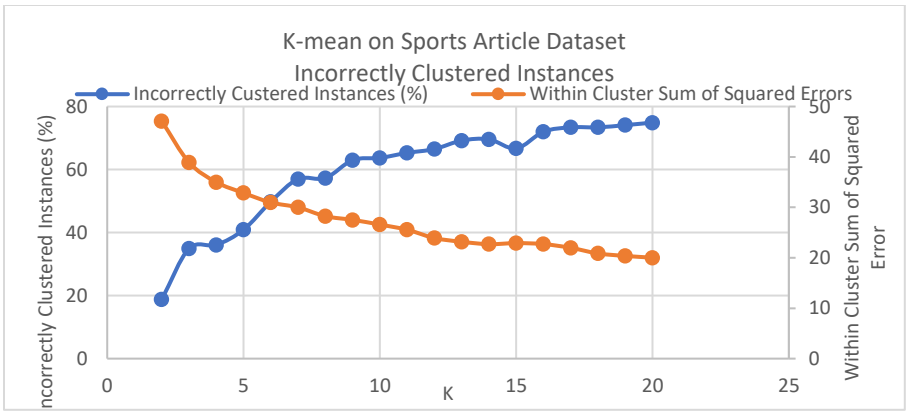


Figure. 19. K-mean via PCA on Sports Article Dataset about Incorrect Clustered Instances and Within Cluster Sum of Squared Error

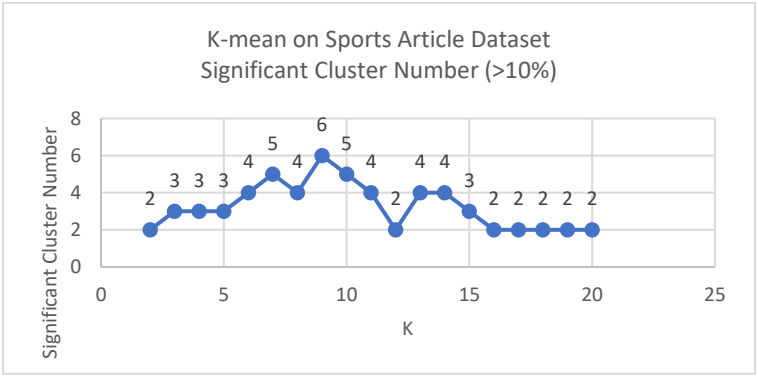


Figure. 20. K-mean via PCA on Sports Article Dataset about Significant Clusters

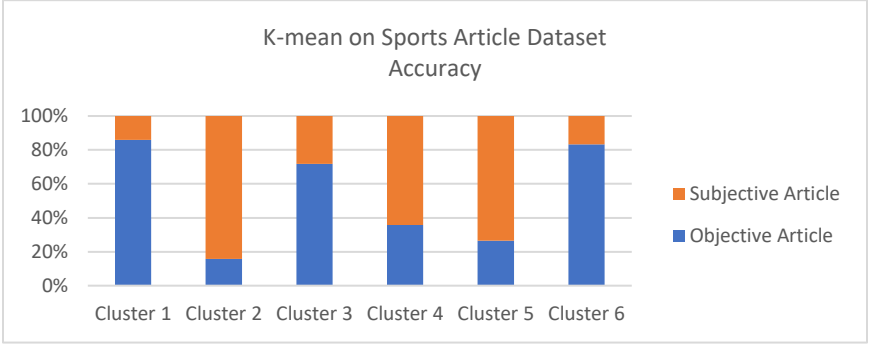


Figure. 21. K-mean via PCA on Sports Article Dataset Clusters with K = 6

The results for the blood donation dataset is shown in figure 22, figure 23, and figure 24. The proper cluster chosen is 7, which is also the same as the one that the dataset only goes through K-mean. Each cluster also behaves better than before. There is no such cluster that has 50% of each king of sample.

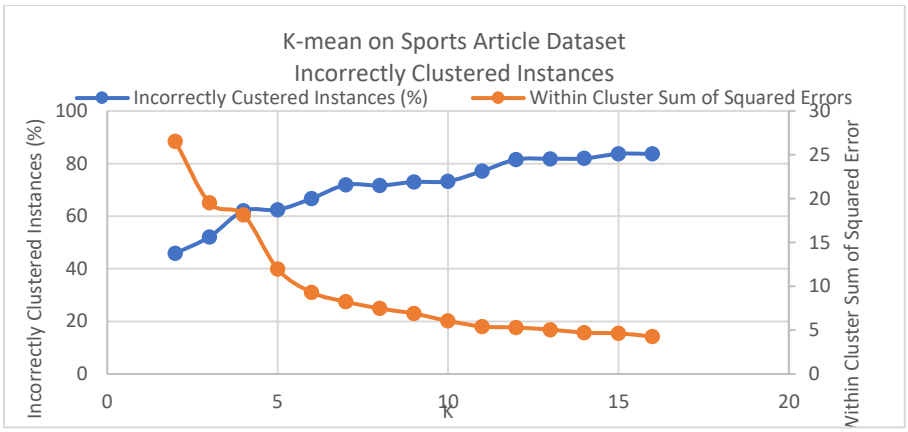


Figure. 22. K-mean via PCA on Blood Donation Dataset about Incorrect Clustered Instances and Within Cluster Sum of Squared Error

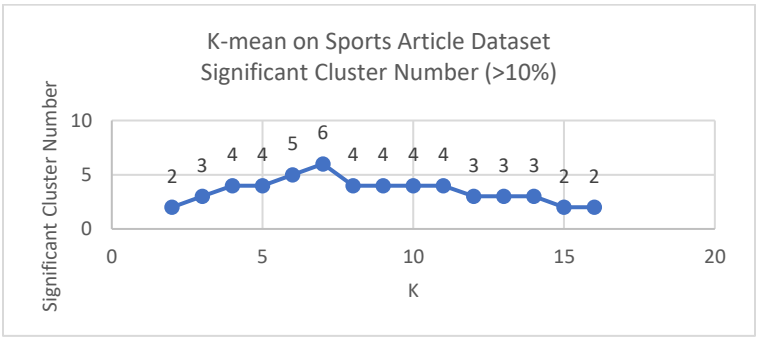


Figure. 23. K-mean via PCA on Blood Donation Dataset about Significant Clusters

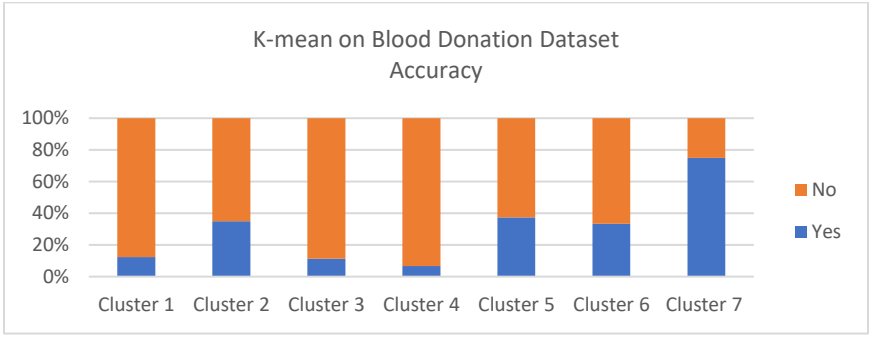


Figure. 24. K-mean via PCA on Blood Donation Dataset Clusters with K = 7

From the two experiments above, it seems like even go through PCA first the best cluster number of K-mean stays the same. This might because even get features reduced, similar samples would wtill be similar and got grouped together. Thoses samples which are not so similar before PCA may get more similar and be grouped together after K-mean but such samples are the minority.

B. GMM via PCA

Figure 25 and figure 26 are the results of the sports article dataset go through 9-feature PCA and GMM. 5 is picked as the idea K value. The analysis for getting the K value is the same as that for GMM. The graphs clearly show that the K value is around 5. Compare to the previous only GMM result K value, 3, 5 is acceptable. Figure 25 shows that the BIC value is almost the same when K equals to 3, 4, 5, or 6. The corresponding likelihood values in figure 26 are also very close. So a general conclusion can be that the cluster value is almost the same.

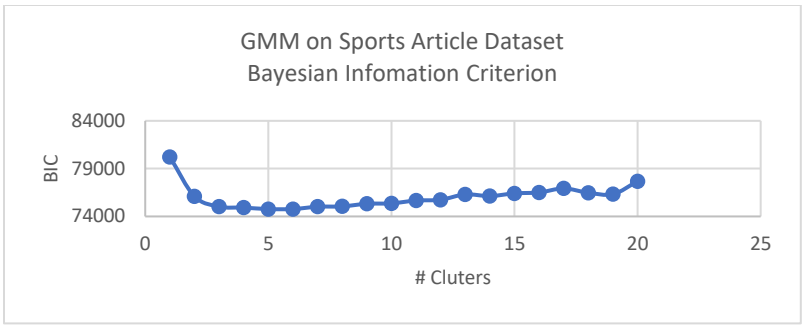


Figure. 25. GMM via PCA on Sports Article Dataset BIC

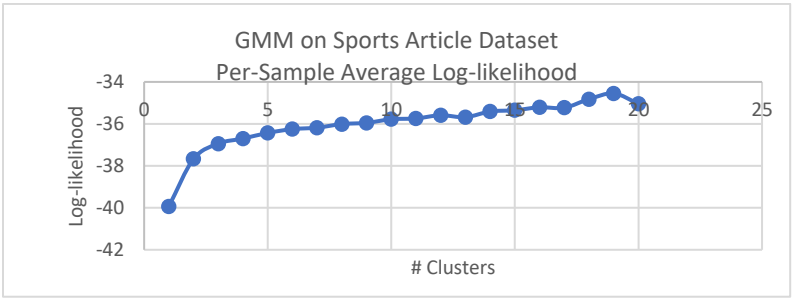


Figure. 26. GMM via PCA on Sports Article Dataset Per-sample Average Log-likelihood

The result for the blood donation is a mess. If only look at figure 27 which is the BIC plot, the first local minimal happens at K=4. Figure 28 is the likelihood plot which almost show a completely different increment tendency. This might because PCA reduces too much information when the features reduce from 5 to 3. Another reason might be that the original features are not good enough to represent the data so that the data generated from PCA would go even worse. Compare the datasets after PCA, the sports article has the feature with the largest standard deviation which is 5.37; however, the largest standard deviation that the blood donation dataset after PCA features have is 1.594 which is almost five times smaller. Which might explains that why the sports article dataset can have almost the same good and clear result with GMM and with GMM via PCA but the blood donation dataset would have a much worse result with GMM via PCA compare to that of only through GMM.

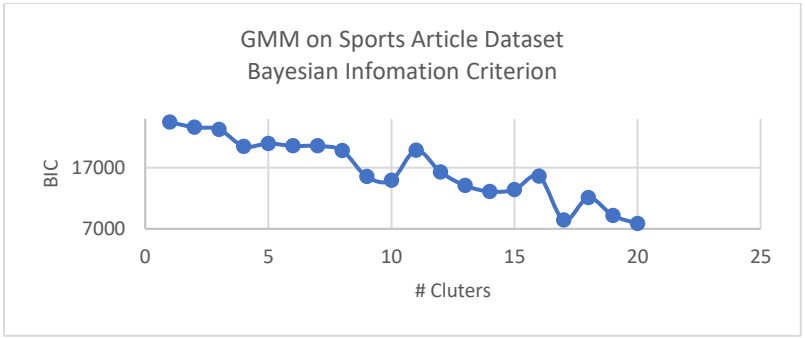


Figure. 27. GMM via PCA on Blood Donation Dataset BIC

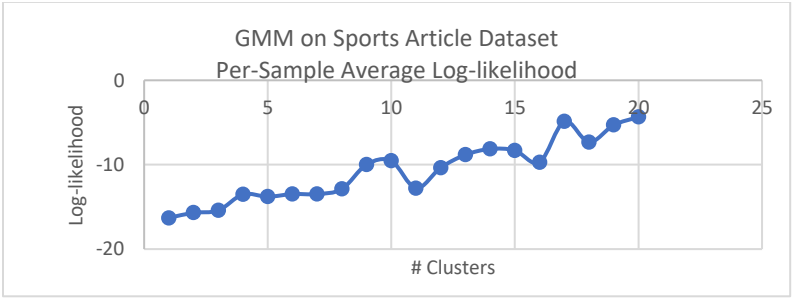


Figure. 28. GMM via PCA on Blood Donation Dataset Per-sample Average Log-likelihood

IV. Neural Network via PCA on one Dataset

NN SA				
# Layers	# Units	Training Accuracy	Verification Accuracy	Running Time (s)
25	15	0.83	0.87	0.22
25	25	0.87	0.82	0.13
50	25	0.87	0.82	0.15
50	50	0.83	0.79	0.17
100	50	0.84	0.83	0.19
100	100	0.82	0.86	0.19
200	100	0.83	0.85	0.14
200	200	0.83	0.82	0.2
200	500	0.83	0.83	1.16

Figure. 29. Neural Network on Blood Donation Dataset

NN SA via PCA				
# Layers	# Units	Training Accuracy	Verification Accuracy	Running Time (s)
25	15	0.78	0.83	0.03
25	25	0.81	0.76	0.11
50	25	0.83	0.78	0.09
50	50	0.76	0.75	0.02
100	50	0.87	0.77	0.09
100	100	0.85	0.81	0.07
200	100	0.86	0.79	0.11
200	200	0.84	0.79	0.15
200	500	0.82	0.72	0.23

Figure. 30. Neural Network on PCA Blood Donation Dataset

Figure 29 shows the result from Neural Network on the sports article dataset and figure 30 shows the result from Neural Network on the dimension reduced sports article dataset by PCA. It is clear to see that in all the combination of hidden layer number and unit number, the NN on the dimension reduced dataset always generate a worse verification accuracy but a faster funning time. The result is worse-off by PCA is easy to understand. The reason is losing information because of dimension reducing. The running time for NN on dimension reduced dataset contains two parts. One is the time requires for the PCA process to generate the new dataset; another part is the time for the NN itself. By comparing the results in figure 29 and figure 30, the advantage of dimensional reduction in time running is phenomenal.

V. Neural Network via K-means and Neural Network via GMM on one Dataset

NN SA via K-mean				
# Layers	# Units	Training Accuracy	Verification Accuracy	Running Time (s)
25	15	0.35	0.32	0.07
25	25	0.73	0.75	0.07
50	25	0.71	0.72	0.1
50	50	0.72	0.7	0.11
100	50	0.74	0.74	0.09
100	100	0.63	0.71	0.19
200	100	0.73	0.77	0.2
200	200	0.72	0.67	0.15
200	500	0.48	0.48	0.29

Figure. 31. Neural Network on K-mean Blood Donation Dataset

NN SA via GMM				
# Layers	# Units	Training Accuracy	Verification Accuracy	Running Time (s)
25	15	0.69	0.73	0.19
25	25	0.7	0.69	0.16
50	25	0.69	0.7	0.17
50	50	0.72	0.71	0.16
100	50	0.71	0.68	0.16
100	100	0.73	0.75	0.14
200	100	0.7	0.68	0.19
200	200	0.7	0.7	0.16
200	500	0.7	0.67	0.28

Figure. 32. Neural Network on GMM Blood Donation Dataset

Figure 31 contains the results of NN on dimension reduced sports article dataset by K-mean. Figure 32 contains the results of NN on dimension reduced sports article dataset by GMM. The running time calculations here are like that in the previous section for NN after PCA. By looking at the results from previous section, it is shown that K-mean and GMM get worse accuracies than NN and are much lower than PCA but still faster than NN in general. In one or two cases, the running times are longer than NN.

Notice that the experiments in this section and the previous section are done together. Although the dataset was shuffled each time the parameters of the NN changes, parallelly for the same NN size the dataset is the same.

VI. Conclusion

Unsupervised learning takes the dataset without labels and tries to draw inferences from the dataset. The three algorithms discussed in this paper are K-means, GMM, and PCA. K-means and GMM are cluster algorithms which means that they separate the dataset into many groups which contain the samples with similar properties. PCA is a dimension reduction algorithm, it transforms the original high dimensional dataset into a new space low dimensional dataset. The new dimension has very close relationship to the original high dimension. No matter through any of these three algorithms, when the dimension decreases, the new dataset will contain less information. The benefit of reduce the dimension is a simpler, easier, faster calculation for the following steps. When the data originally does not have many features, there is no need to do dimension reduction.

VII. References

[1] https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

[2] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[3] https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html

[4] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

[5] https://sourceforge.net/projects/weka/files/latest/download