Abhishek Deo
CS 4641

Unsupervised Learning Assignment Three

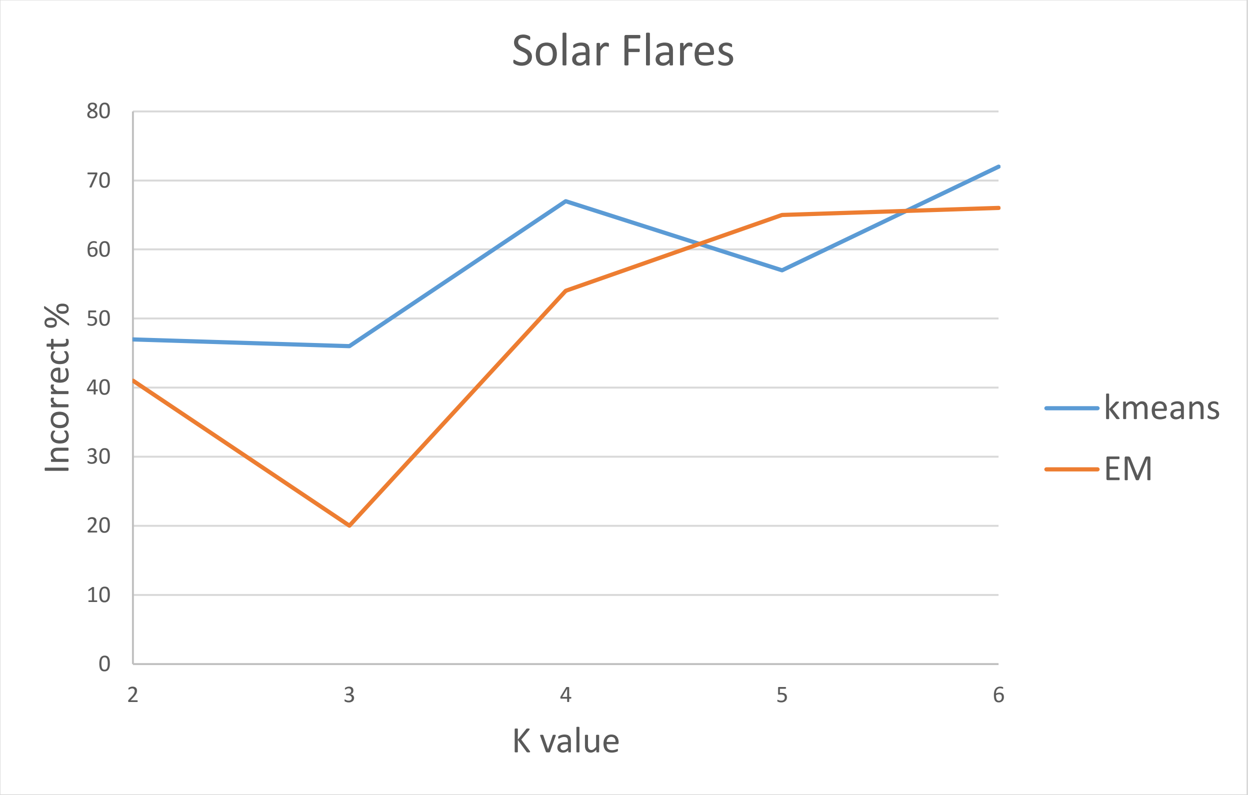
Datasets:

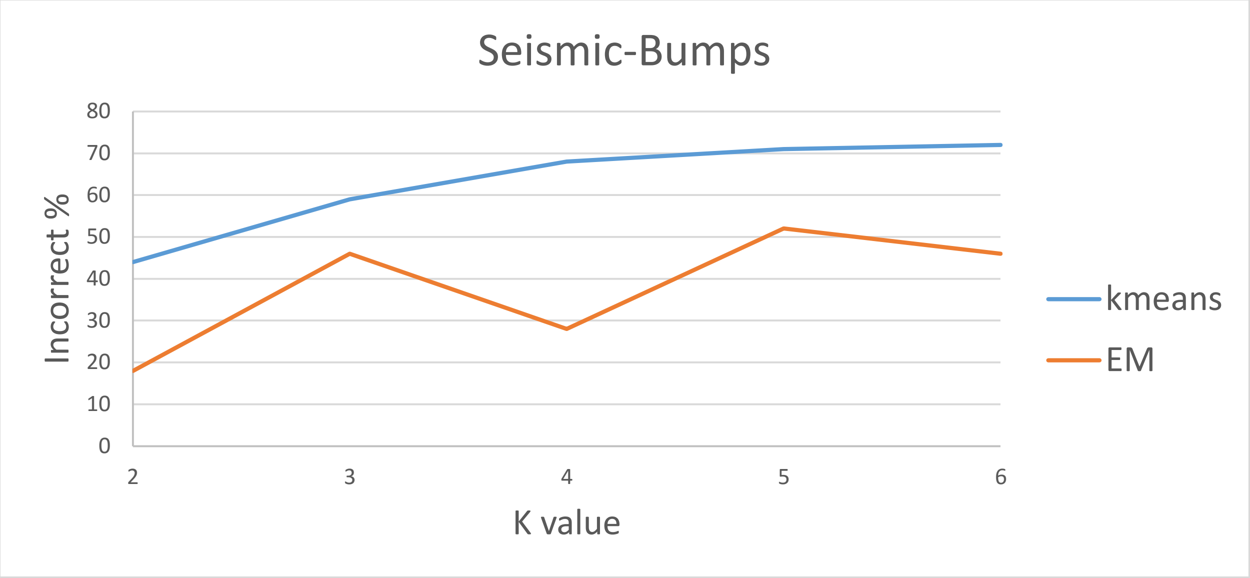
The two datasets I used were the same datasets I used for assignment one and assignment two. The first dataset is the seismic-bumps dataset from the UCI repository. This dataset classifies high-energy seismic bumps forecasting in a coal mine. Based on 19 attributes, an individual recording is classified in one of three classes. The second dataset is the solar flares dataset from the UCI repository. This dataset uses 10 attributes and three classification attributes which are based on the other 10. Each attribute counts the number of solar flares within a 24 hour period and classifies them as such.

Choosing k:

The range of which I choose the value of k were between 2 to 6. For the seismic dataset, since it is classified into two classes based on the values of the attributes k=2 would be the best choice. For the solar flares dataset, since it is classified into three classes based on the value of the attributes k=3 would be the best choice. For variability I choose the range from 2 to 6 because anything less than two would not make sense and I chose 6 as the upper threshold after various experiments to show significant decline. For both datasets because of the variability in features the quality of the results significantly decline at and after k=3.

K-means and EM



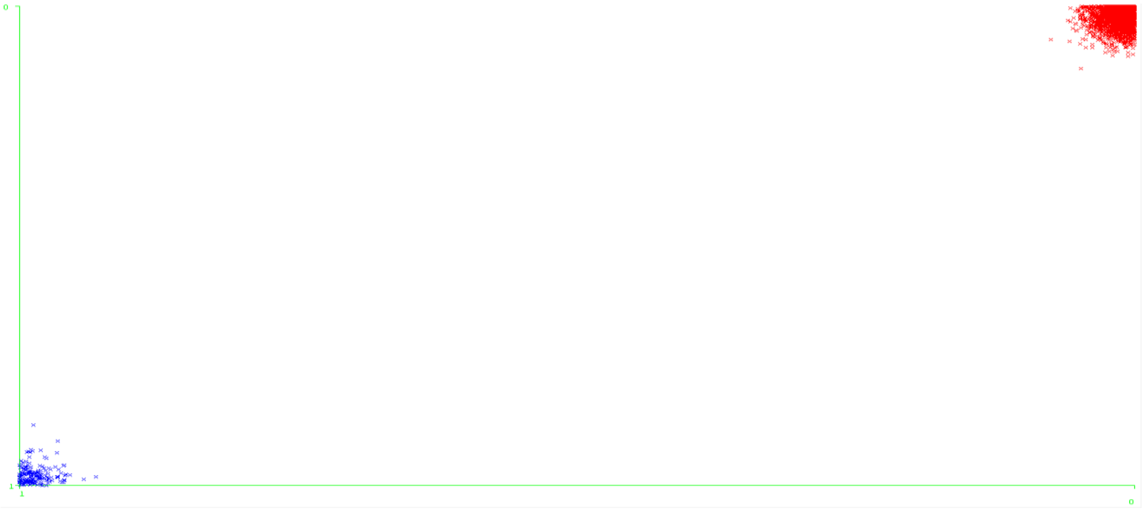


The Solar Flares dataset is classified into three separate classes. Based on the results, the percent error is lowest on k=3 which equivalent to the same number of classes. There is not much of an incorrect percentage at k=3 for simple k-means but the percent error rises for EM after k=3 in a sort of logarithmic fashion. One thing to note however at k=5 the percent error actually drops in comparison to k=4 and k=6 for k-means. After numerous trails runs as to see why, the best conclusion is because of how the data is divided. From the repository, it mentions that the data is divided into two sections where the second section has much more error correction applied to it. I ran the test multiple times around k= {4, 5, 6} and I got slightly different results each time. At 5 clusters it matches the biggest feature available where the available choices are greater than equal to 5. The reason the percent error increases at almost a linear rate for EM is because it isn't biased to spherical clusters. K-means is a variation of EM and makes use of the L2 Norm whereas EM soft assigns points to clusters and gives it a probability. EM also allows for a little grey area when assigning points. This results reflect the nature of the dataset where k-means allows for different variations in clustering.

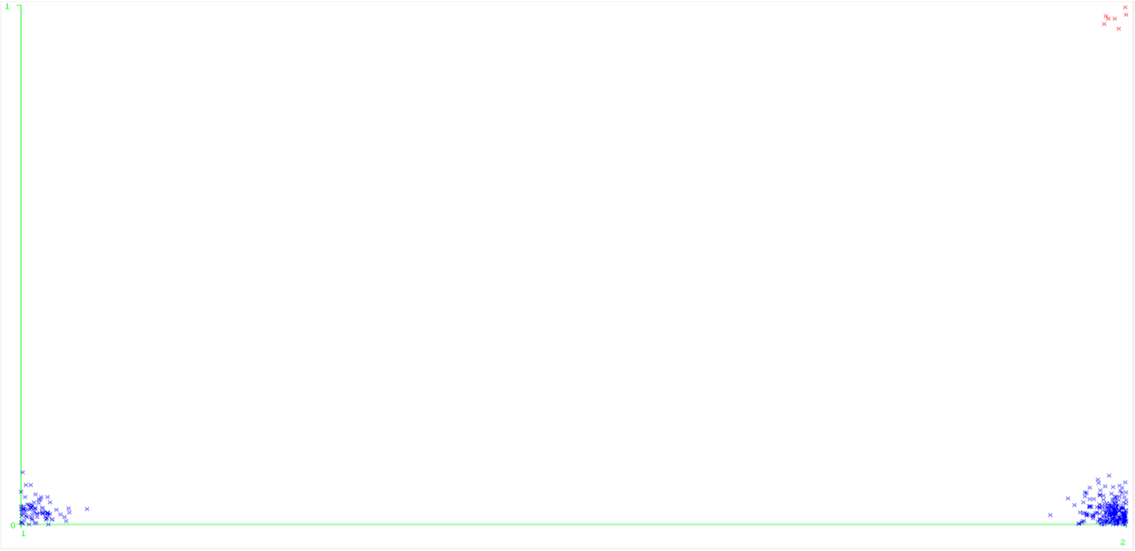
The seismic-bumps dataset is classified into two separate classes. Based on the results, the lowest percent error is on k=2 which is also equivalent to the same number of classes. For the k-means lines, the percent error rises in very logarithmic fashion and the variance decreases through every consecutive k-step. For EM however, at k=4 the percent error decreases as well as at k=6 compared to the previous step. To explain this, it should be noted that in the dataset only a portion (170 instances) are positive examples representing class 1 which means an instance seismic activity will occur. As a result, one class has significant majority in terms of instances which means that classifying into two classes would work better with some dimensionality reduction.

The time taken to run the algorithms were virtually zero with 5 iterations for solar flares and 7 iterations for seismic-bumps to find a solution.

Clustering Illustration

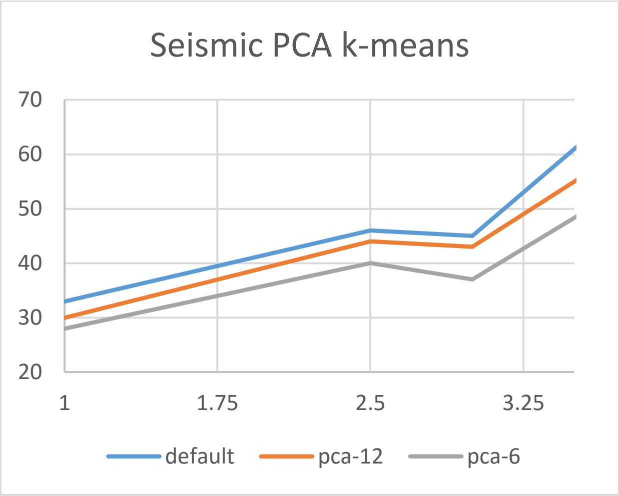


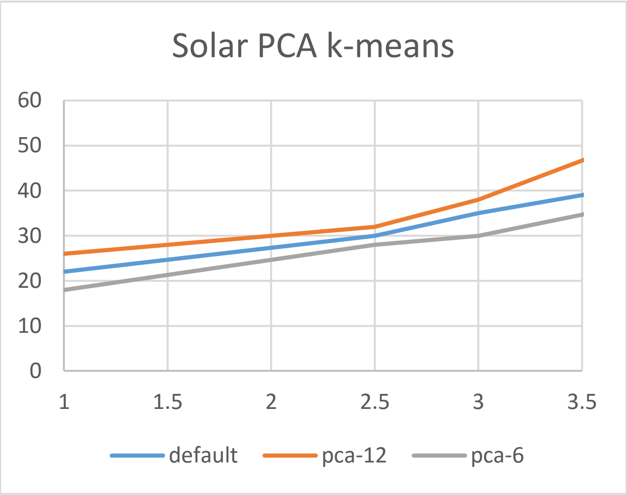
The clustering shows how biased the seismic dataset is. (From the description of the dataset on the UCI website). There are clearly many more points for the negative classifier as opposed to the (170 instances) for the positive classifier.

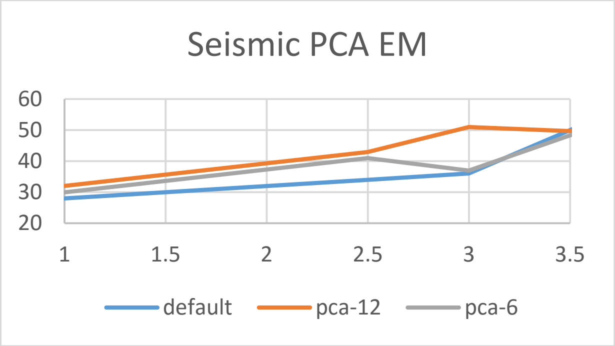


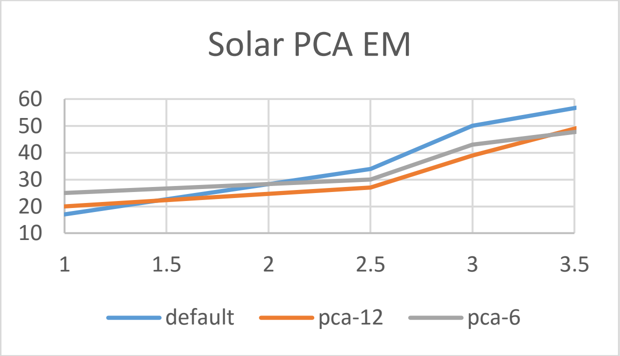
This clustering shows the result of solar dataset. The clusters are separated in three classes and shows that one clearly outweighs the other in terms of frequency. However compared to the Seismic dataset there is still a big difference in distribution.

(PCA): (The graphs are below for all the DR algorithms!)









Seismic Results

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
5.1	4.1	2.3	2	1.9	1.7	1.74	1.7	1.2	.9	.72	.6	.4	.2	.11

Solar Results

1	**2**	**3**	**4**	**5**	**61**	**7**	**8**	**9**	**10**
2.13	2	1.6	1.41	.94	.77	.53	.44	.39	.12

The graphs above illustrate the percent error for the seismic and solar datasets. The data was split into a combination of 12 and 6 attributes respectively as well as using all the attributes which is denoted as default. This way, it provides a comparison and provides clarity by reducing the dimensionality and remove things based on variance. The tables represent the eigenvalues results based on the combination of attributes. It seems that comparing the percent error as the Eigen values drop, the solar dataset does worse than the seismic dataset albeit not as much. The default PCA percent error is better than when the number of elements is at 12.

There are 13 elements that represent the solar dataset if including the classes as attributes. In the case of pca-12 in the solar dataset, as the Eigen value lowers the accuracy lowers as the number of dimensions lower which is contrary to the case in the seismic dataset. This may be a result of the combination of attributes which all have very similar formats and dependencies. Another interesting occurrence is that the default run on the EM algorithm for the Seismic dataset actually did the best for the most part compared to k-means. This is possible because the increase in accuracy in k-means as the dimensionality went down is due to the fact that the attributes are reorganized but was grouped using hard clustering. As a result, the classes for the seismic dataset seem to be more spread-out. This behavior can also be noted with the solar dataset but is performed at a lower level of quality.

Both the Seismic dataset and solar dataset return the default run to have the lowest percent error. Reducing the dimensions causes the variance to increase and the information lost is important and not categorically considered noise (represented by the decrease in accuracy in pca-12 and pca-6). EM behavior is determined by spherical clustering and strictly removes noise while trying to find maximum likelihood data. The noise would be considered missing values and simply ignores this by assuming the existence of additional unobserved data points. The best Eigen value for the seismic dataset at 5.1 shows that a specific combination of attributes provided a lot of information. Most of the Eigen values were around 0.7-1.7 but do not convey actual viable information. The results got better which can also be shown in the increase of variance as the Eigen values decrease. The same can be said for the Eigen values for the solar dataset but there is smaller variance between the Eigen values as they decrease. All-in-all, seismic activity performed the best under k-means after the pca modification by improving accuracy as dimensionality decreased. Solar activity also performed fairly well under k-means but due to the nature of the elements, pca-12 did worst then the default run. If you compare the starting error percentages of both the seismic and the solar datasets, solar datasets starts off with better accuracy and also ends with better accuracy. But the discrepancy lies in both the organization of the attributes of the solar dataset. Running PCA and allowing for enough attributes to cover the variance in the seismic dataset also helps in starting at a lower accuracy. But it is more consistent between the run types compared to the solar dataset.

ICA:

Seismic Kurtosis Table (bold values are negative):

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
.5	**.9**	.2	1.4	.6	1.2	1.1	1.8	**.8**	**.4**	**.7**	1.9	2.7	1.7	**.1**	**.3**	**.2**	.5	**.5**

Solar Kurtosis Table:

1	2	3	4	5	6	7	8	9	10	11	12	13
.7	**.2**	1.2	3.1	2.4	.8	**.1**	.6	1.1	.5	**.9**	1.3	.4

The tables above show the absolute value for the kurtosis for each of the sources and the bolded values were initially negative values. These values are scaled to zero so that it is easier to consider zero Gaussian as opposed to the standard three. ICA was run on the datasets and then performed on both clustering algorithms without removing sources. Sources closer to 0 were removed to reduce dimensionality and improve variance as well as accuracy. When removing values close to zero (zero being Gaussian) the clustering results improved slightly but not by much because the distribution of the results are not Gaussian. This is evident by the number of negative values which are close to zero. Keeping the sources in the datasets was better than removing them because the attributes are independent of each other in the seismic dataset. This is true for both EM and k-means. But, removing information from mutually exclusive sources would cause a further increase in accuracy percentage and spread out the clustering algorithm which is shown between k-intervals in the graph. The farther spread out the data points the less grey area meaning the appropriate class is most likely closer. This is applicable to both the seismic and solar dataset. The nature of the attributes are independent of another but they factor in to the overall classification. There is more error for seismic dataset due to having more attributes but is almost balanced equivalently to the solar dataset by having fewer classifications. It seems running ICA modified through EM is the best for the solar dataset and ICA modified through EM is the best for the seismic dataset as well. Spacing out due to removing mutually exclusive sources helped in this matter. Because the data points tended to be closer for the seismic dataset, ICA was not able to outperform PCA in k-means. However that was not the case for the solar dataset. Both PCA and ICA were relatively close through each iteration of k and seemed to almost converge after a certain amount of iterations.

Random Projection (Seismic):

For random projection, among all the datasets the most interesting result was the line produced by the seismic dataset run under k-means. The second most interesting result was the random projection line under EM for the seismic dataset. After k=5 the percent error started to significantly increase but for the first for k iterations random projection had the lowest percent error. This may be because randomly projecting in a lower cluster space would work better at least for the seismic dataset because of the number of attributes. While they are independent of each other, there are 19 present to classify an instance in one of two classes. These conditions make it easier for random projection to cluster a data point. In addition, all attributes provide a decent amount of unique information. As for random projection under EM, it seems at one point the accuracy improves between k=5 and k=6. However it started off very high in percent error and at k=2 clusters percent error increased tremendously. Equally however, it surprised me that it also went down by almost the same amount of difference at k=3. This result is inverted compared to the line in k-means where it started off low but eventually became high in percent error. This may be due to the fact that EM is very stringent in clustering data points. Another contributing factor would be that this dataset is heavily biased towards the positive classification which may influence the results on random projection.

(Solar):

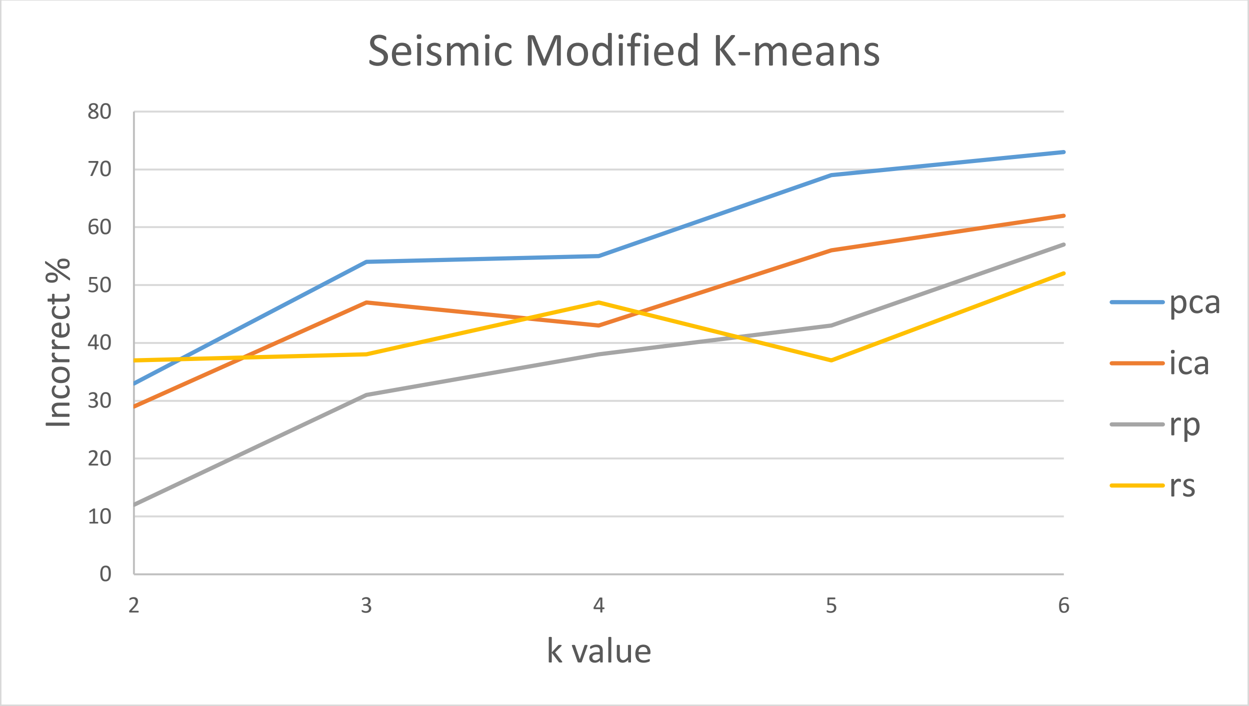
For the solar dataset on both k-means and EM random projection behaved as expected. In the case of EM it was surprising to see that random projection was not the highest percent error at k=2 but was soon overtook the other algorithms. Another interesting point is that the same situation occurs at k=6 where PCA overtakes random projection in percent error but it may be because of the number of attributes which affect PCA significantly. For the first aforementioned case of EM, the same situation occurred where random projection did not start off with the highest percent error in k-means but it quickly became apparent that it was due to that specific algorithm. (Random Subset).

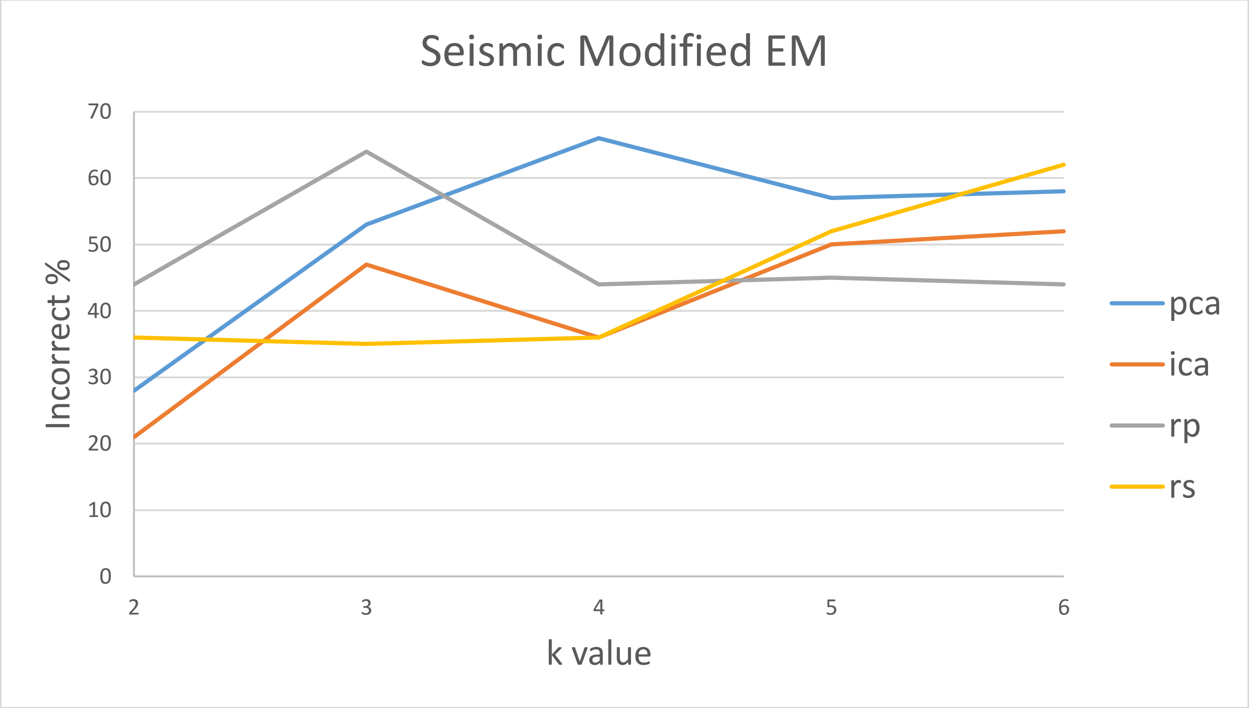
Random Subset (Seismic):

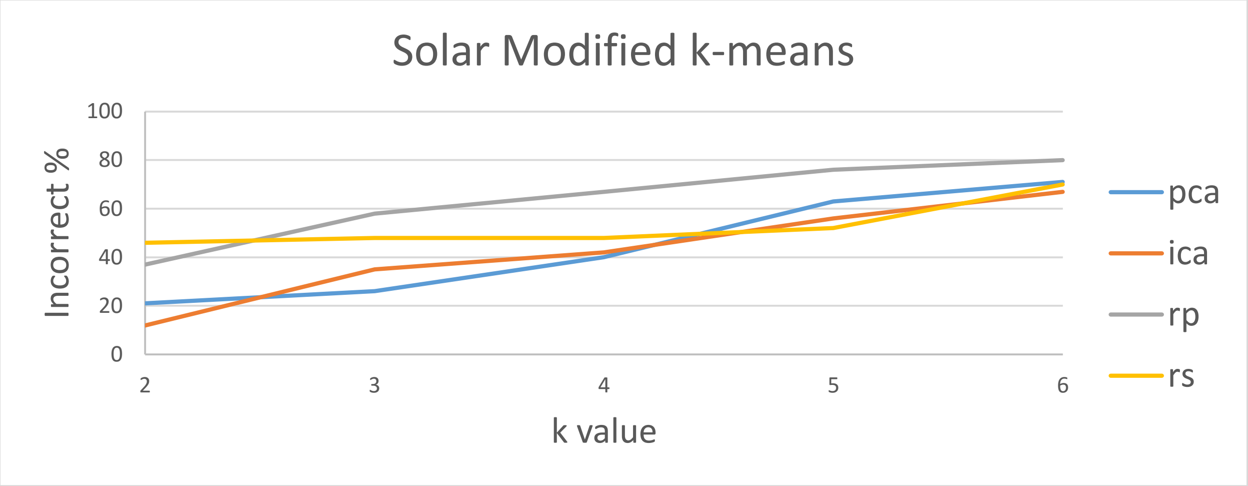
Random Subsets produced the most consistent result in comparison to each other between algorithms and datasets. For the seismic dataset, random subset under k-means seem to increase and decrease in percent error quite a lot. Since random subset removes random portions of data, it makes sense that most likely percent error would increase because it is possible that some of the information removed is relevant, it can balance out. Removing data which does not contribute to the overall classification can reduce complexity making clustering easier. This also is heavily dependent on the dataset and how bias it is as well on one class vs. another. Seismic dataset is very bias which means that most likely most of the information is relevant to some class which affects the clustering. However because only a small percentage of the data (170 instances out of around 2000+) where meant for the positive class, it was irrelevant on how many random subsets where removed. In the end, even after k=6 it had the best percent error in comparison to the rest of the algorithms. This was not the case for EM however because the probabilistic value of data points are dependent on the ones surrounding it which is the clustered into one centroid. Removing a decent amount of information even though it is heavily biased does not affect the result which is why only for the first few sets of k were the results consistent. Once k=5 or more the percent error increased significantly.

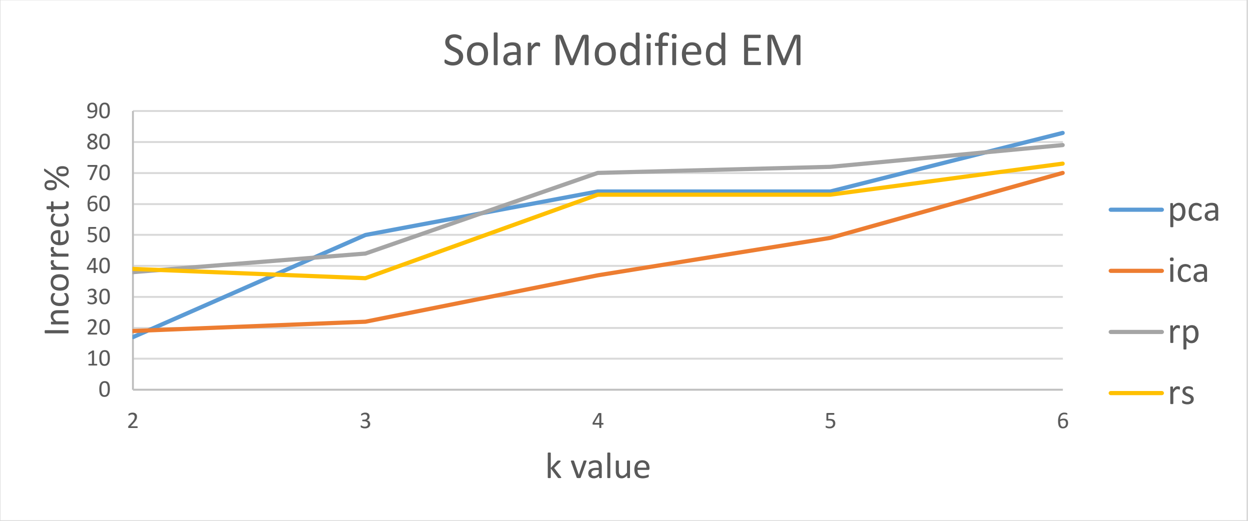
(Solar):

The results for random subset applied to the solar dataset were slightly different from the seismic dataset. It was every consistent but increasing in k-means but seemed to be the average among all the DR algorithms. There were no random jumps in difference between percent errors between k. For the solar dataset, this would make sense as results are classified into three different classes and most of the dataset is randomly distributed, it is more even then the seismic dataset. As such, the percent error slowly increased while k increased and eventually started to really increase after k=5. The solar dataset has significantly fewer instances in comparison to the seismic dataset which may explain why k=6 may be too much. For EM the biggest hiccup was the increase in percent error from k=2 to k=3. The number of classes for this dataset is three so that was not to be expected but it may be simply due to the nature of removing the random subsets. This is what makes it unpredictable but sometimes it may also produce quality results.

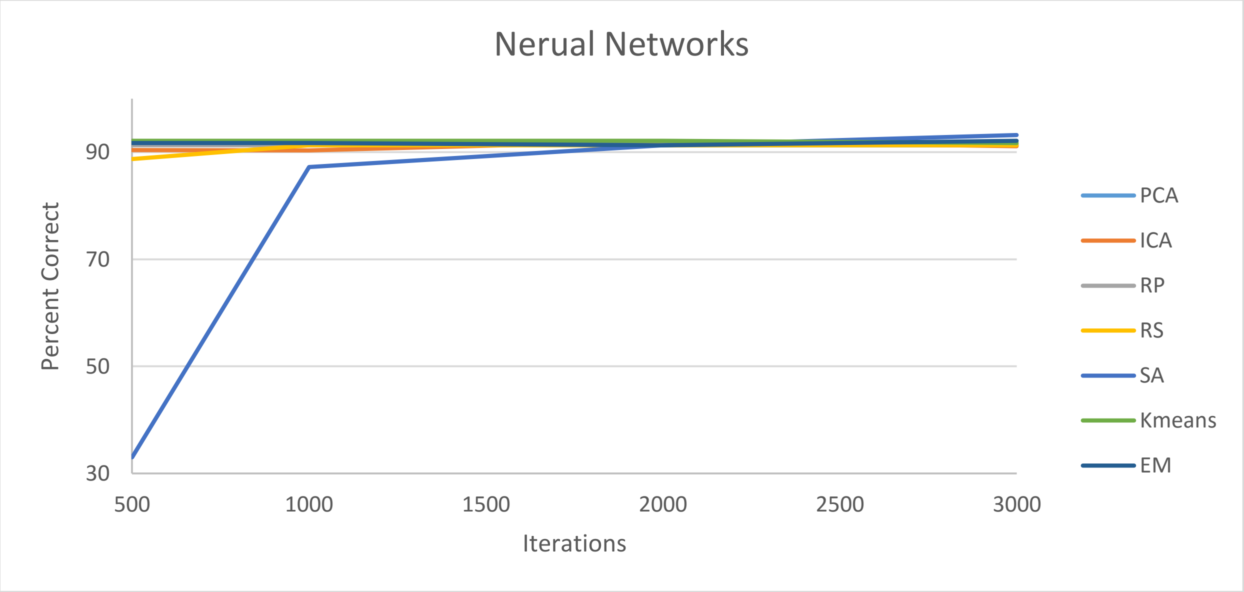








Neural Networks:



This graph is the set of neural networks for both steps 4 and 5 run on the seismic dataset. The k-means and EM line are the results of step 5 and the rest are the results of step 4. The results of the neural networks were very interesting considering that almost all of them with the exception of Simulated Annealing were very close to each other at each iteration step and percent accuracy all ranged from a low of 88.73% to a high of 92.133%. Even though simulated annealing did not start off well, it still overtakes many of the other algorithms and is apparent after a few more 1000 iterations steps it would become the best. (Although would most likely stay stagnant at one point and converge with the rest eventually.) To start off, comparing EM and k-means after treating the clusters as if they were features were very similar in both the training and testing sets for the seismic dataset. Although difficult to see, at 3000 iterations k-means was at 92.71% accuracy but EM was at 92.15% accuracy. However, for the remaining set of iterations k-means always beat EM. For the most part it acted how I expected because from the graph on the second page, the percent error slowly increased for k-means as the percent for EM hiccupped up and down but stayed mostly the same. Eventually EM would overtake k-means due to its own downfall. In terms of reducing dimensionality, ICA seemed to be the most average and consistent which correlates to its consistency during clustering. PCA performed better than k-means which was also true during clustering when comparing the difference between the percent errors but by a small margin. Most of the algorithms did not over-fit except k-means which slightly over-fit. After running numerous iterations it seemed that it would become constant almost. Random Projection using with added features was pretty much the same without because of the dataset nature itself. The same instance can be said for random subsets.

Conclusion:

Clustering overall returns very interesting unexpected results and like other types of algorithms, are heavily dependent on the features of the dataset. The seismic dataset faired decently across the board especially when it came to Random Subsets or Random Projections as dimensionality reduction methods. Considering the structure of the dataset which is composed of 19 attributes and 2 classifications in addition to the heavy bias towards one of the classes, this dataset favors these dimensionality reduction methods. The solar flare dataset performed better overall because of the number of instances in the dataset as well as a better classification distribution compared to the seismic dataset. Keeping this in mind, solar flare dataset did better on ICA and PCA under both k-means and EM.