Fast and Unsupervised Classification of Radio Frequency Data Sets Utilizing Machine Learning Algorithms

Phil Romero

Los Alamos National Laboratory
Los Alamos, New Mexico 87545
Email: prr@lanl.gov

Kalpak Dighe

Los Alamos National Laboratory
Los Alamos, New Mexico 87545
Email: kdighe@lanl.gov

***Abstract*—Collection of Radio Frequency data can overwhelm even the largest data storage capacities very quickly due to high sampling frequencies. There are many sources of possible error in maintaining an accurate record of the captured signals. These issues can be solved, in large part, through an automatic classification of data sets gathered that eliminates the possibility of human error and assures that the proper type of signals were captured in a timely fashion. In this paper, we will describe the process used to produce a classification system. The goal is to identify and use measures produced from the raw signal information and/or the spectrograms for input into an algorithm that produces clusters based on similarity that will classify the data into subsets with the least amount of computational complexity. K-means clustering and principal component analysis are utilized in a two step process to perform the classification of the data sets. Minimal amounts of measures have been found to produce satisfactory results in separating the raw signal data into dissimilar signal types based on a 32768 sample size. This minimizes computational complexity while still producing output used in the second stage of the process to classify the data sets. A method of classification was found that produces minimal false positive errors while selecting the proper number of clusters without resorting to more computationally complex methods thereby decreasing the time spent classifying.**

***Keywords*—Digital Signal Processing; Machine Learning; Radio Frequency.**

I. INTRODUCTION

Collection of Radio Frequency data can overwhelm even the largest data storage capacities very quickly due to high sampling frequencies. The sampling frequencies can range up to two or even five billion samples per second with many channels collecting the data simultaneously. Data rates can exceed 200 GB per second[1] and it is prohibitively expensive to store large samples in real time. Adding to the problem is the time required to verify the desired signals were recorded in the data collection and properly annotating the data for convenient retrieval at subsequent times. Also noteworthy is the problem created by both expected[2] and unexpected sources of radio frequency signals that can diminish the value of the data collected[3]. Human error can also lead to incorrect annotation of data whose consequences can be difficult to mitigate. These issues can be solved, in large part, through an automatic classification of data sets gathered that eliminates the possibility of human error, assures that the proper type of signals were captured in a timely fashion and eliminates the need for storage of uninteresting data sets. A methodology for signal discovery is proposed in Section II and is compared with a currently used alternative. Results are presented for an optimum sample size and input parameters in Section III. Results of the first clustering process are presented in Section IV, this process considers each data set independently. The results of the second clustering process, where data sets are compared, is discussed in section V. Finally, a conclusion is presented in Section VI.

II. METHODOLOGY

The radio frequency spectrum ranges from around 3kHz to 300GHz and is, in part, utilized to carry communication signals. These communications signals vary widely including AM radio broadcast signals, television broadcast signals, FM radio broadcast signals, Cell phone signals, GPS, and wireless computer networks. All of these signal sources can produce produce a significant amount of background noise in the RF spectrum. Typically, the background noise must be considered when capturing signals in the RF spectrum and the proper adjustments must be made to ensure they do not interfere with signals of interest, examples of which are shown in Figure 1 and in Figure 2.

A frequency analysis can be performed on the discrete-time signal by converting the time-domain sequence to an equivalent frequency-domain representation. This can be accomplished with the Fourier Transform of the discrete-time signal. Further processing can produce a spectrogram which shows the power level at given frequencies for the timespan in question as shown in Figure 3. The goal is to use measures/features produced from the raw signal information and/or the spectrograms for input into an algorithm that produces clusters based on similarity that will classify the data into subsets with the least amount of computational complexity. This would eliminate a time consuming process that must be undertaken by an expert in Digital Signal Processing that is prone to error. In order to discover signals within the data sets, spectrograms[4] would need to be produced for each segment of data, the known signals would need to be removed through the application of digital filters[5] without eliminating any part of the signal we may be interested in and the images produced would then have to be examined. Given the large numbers of data segments to examine and the possibility of a digital filter eliminating a signal of interest, this method vastly improves throughput in identifying signals within the data. Several measures were computed from both the raw signal and the spectrogram to

 ISBN: 978-1-61208-423-7

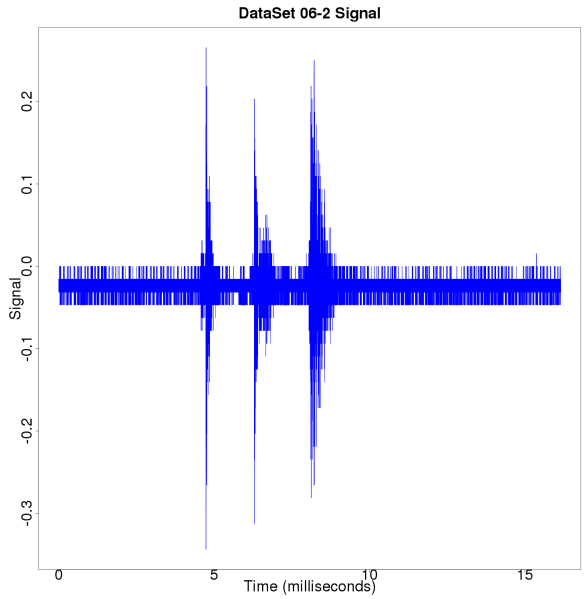

Figure 1. This figure shows a time-discrete signal waveform collected on a regular time interval in the RF spectrum.

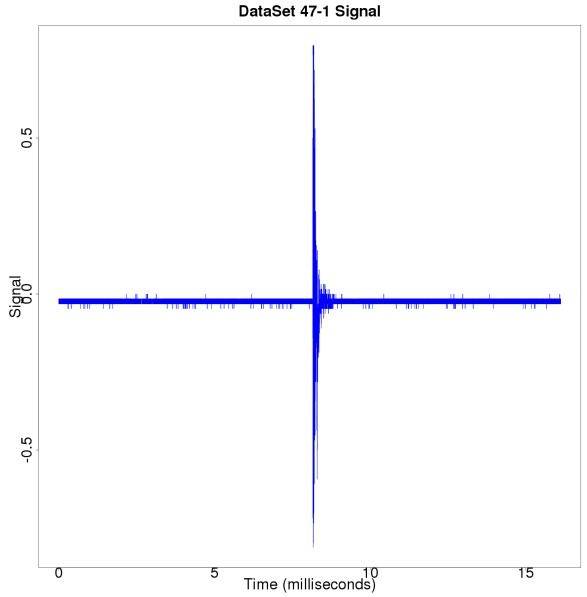

Figure 2. This figure shows another example of a time-discrete signal waveform collected on a regular time interval in the RF spectrum.

be input into the clustering algorithm. Among these features were:

1) The maximum number of frequencies identified above a given power threshold at every time processed from the spectrogram.
2) The maximum number of continuous frequencies above a given power threshold at every time processed from the spectrogram.
3) The mean power produced over the entire timespan of each spectrogram.
4) The standard deviation of power produced over the entire timespan of each spectrogram.
5) The minimum power produced at every time processed from each spectrogram.
6) The maximum power produced at every time processed from each spectrogram.
7) The median power produced over the entire timespan of each spectrogram.
8) The mode of power produced over the entire timespan of each spectrogram.
9) The mean value of the covariance of power produced over the entire timespan of each spectrogram.
10) The mean of the unprocessed signal data of a given timespan.
11) The standard deviation of the unprocessed signal data of a given timespan.
12) The minimum of the unprocessed signal data of a given timespan.
13) The maximum of the unprocessed signal data of a given timespan.
14) The absolute value of the minimum of the unprocessed signal data of a given timespan.
15) The median value of the unprocesssed signal data of a given timespan.
16) The mode of the unprocessed signal data of a given timespan.
17) The absolute value of the mean value of the unprocesed signal data of a given timespan.

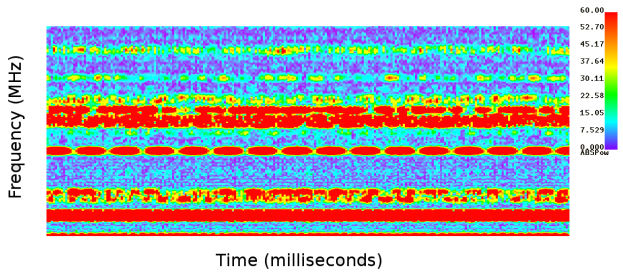

Figure 3. This figure shows an example spectrogram in the radio frequency range. The red and yellow lines are background noise caused by communication signals.

The process of identifying the timespan to process the data with was determined by processing with several different numbers of sample sizes including 128, 256, 512, 1024, 2048, 4096, 8192, 16384, 32768 and 65536. The sample sizes were restricted to powers of two due to considerations of applying a discrete fourier transform to calculate the spectrograms for each timespan. This is essentially an optimization problem where the sample size needs to be able to resolve complex signal information into repeatable patterns while minimizing computational complexity. The k-means clustering algorithm[6] works by dividing a large sets of points into any number of ”neighborhoods” requested by the user. In our case, the points are all the above measurements for computed for every chosen timespan for a given sample size within a data set. It is important to note that three separate data set file lengths were utilized, consequently, the result has to be independent of the size of the data set processed. Formally, the k-means algorithm is used to solve the following problem:

Given: a set of observation (x_1, x_2, .. x_n) where each

 ISBN: 978-1-61208-423-7

observation is a y-dimension vector.

Task: Partition the n observations into k sets ("neighborhoods") to minimize the within-cluster sum of squares.

The output of the k-means algorithm is an assignment of each observation into one of the k clusters, the sum of squares within each cluster, the distance between clusters, along with other statistical measures. A technique called Principal Components Analysis (PCA) is also used to simplify a complex multivariate data set to expose the underlying sources of variation in the data. A full description of Principal Components Analysis is extremely complicated and better left to more authoritative resources[9]. A challenge with the k-means algorithm arises from the fact that it can produce different data clusters in subsequent runs because it may have found a local minima rather than the global minima. Another problem with the algorithm is that a user must select the k (or number of data clusters) prior to starting the algorithm. There are several options to guard against picking the wrong value for k including the elbow method[8], using the X-means algorithm[10], using the Gmeans algorithm[11], and a proposed manipulation of the k-means output parameters are also investigated in an attempt to minimize computational complexity.

It was unknown if the features/measures needed to have equal or unequal weightings before the methodology was implemented, however, k-means allows for changing the weightings should the need arise. Agglomerative heirarchical clustering methods were eliminated from consideration due to the added complexity deemed unecessary. There are certainly other clustering approaches that could also have been considered such as k-medoids[12] or DBSCAN[13] that are considered more robust than k-means, however, we have found in extensive use of k-means that we have never encountered any instability issues. Therefore, k-means was selected over other methods for it's flexibility and simplicity as well as it's low computational complexity.

The data was collected in an RF laboratory environment with a commercial, programmable broadband signal generator. Repeatability of the experiment is not an issue as the signal generating codes are archived.

We investigated the output of the k-means/PCA algorithms after they are applied to each of the 96 sample data sets in order to classify each of the data sets by the patterns found within them. This allows for the possibility that a given data set has more than one type of signal within the data set. It also means that this is a two stage process whereby the initial k-means/PCA process serves as input into another k-means/PCA process to classify each data set from the combinations of data found by the first process. It should also be noted that combinations of patterns found in the data set are important to find thereby rendering the two step process as necessary.

III. Results Optimizing the Sample Size and Input Parameters

Data was processed from all 96 data sets in sample sizes of 128, then doubling in size until 65,536 was reached. This produced 10 different complete sets of data that were analyzed for suitability. The smaller sample sizes produced larger amounts of clusters and longer processing times than the larger sample sizes. The combinations of larger number of clusters, when combined in the second clustering processes, would produce a more complex classification set, consequently the small sample sizes were eliminated from consideration. The larger sample size of 65,536 was thought to produce too few samples from the clustering process with smaller data sets, thereby diminishing the value of the clustering precision. An optimal sample size of 32,768 was decided upon as the proper balance between precision, output complexity, and computational complexity. The input parameters were compared to determine whether it was necessary to perform the more computationally complex calculations necessary to produce spectrograms. The R statistical packages provides output showing the importance of variables in producing the principal components analysis, from these results it was clear that it was not necessary to perform the more computationally complex work required to produce the spectrograms since input produced from processing only the raw signal data could be produced with less work (in less time) without any significant loss in clustering precision. A subset of the variables calculated from only the raw signal data were further reduced due to two factors. The first factor was that in order to produce a clustering output, all input variables must have a non-zero variance for all data sets, this eliminated many of the variables from consideration into the final optimal method. The next factor that eliminated variables for input into the clustering was the significance upon the clustering, again as determined by the principal components analysis. This process left only the following three variables that need to be calculated on the raw signal for the clustering process:

1) The mean of the unprocessed signal data of a given timespan.
2) The standard deviation of the unprocessed signal data of a given timespan.
3) The absolute value of the mean value of the unprocessed signal data of a given timespan.

IV. Results of the First Clustering/PCA Process

The R program that was written to produce the clustering/pca analysis for the first step in the process also creates several plots. A small sample of the plots produced are shown starting with the cluster plot for the first selected data set in Figure 4. The unique signal that exists in cluster number eight can be found on the first data line and is shown in Figure 5. It should be noted that a principal components analysis plot would look exactly like the cluster plot without the ovals and with the green numbers shown as symbols or labels indicating the data set identifiers.

Another set of plots for a selected data set is shown in Figure 6 and in Figure 7. This time there are two signals that are separated from all other data points in the file, the line shows that there are two members in cluster number 12. The third set of plots for yet another selected data set is shown in Figure 8 and in Figure 9. This time there are three signals that are separated from all other data points in the file, the ellipse shows that there are three members in cluster number 12. More sets of plots can be shown that show similar patterns with most of the identified clusters being very near each other on the plot and a few small clusters very isolated from the rest. However, there is another pattern shown in some plots where there are no clear outliers amongst the clusters, this occurs when no signal has been found in the data set (and when only

 ISBN: 978-1-61208-423-7

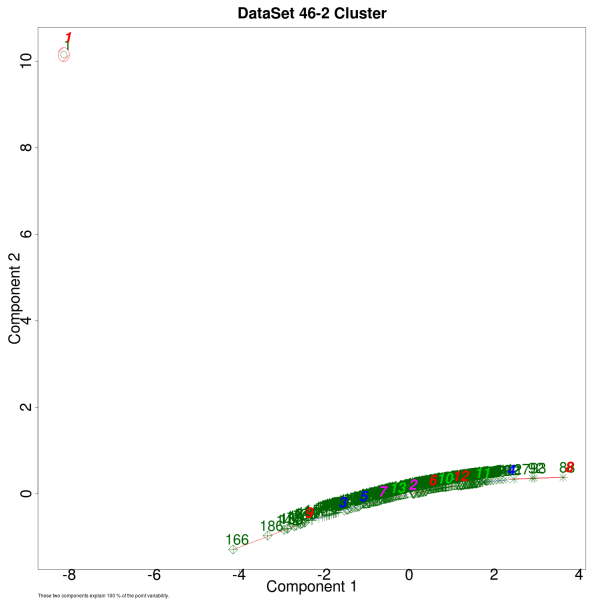

Figure 4. This figure shows a cluster plot shown with the axes being the first two principal component axes. The data point lines are the thinner font dark green labels and the clusters are identified with the thicker font.

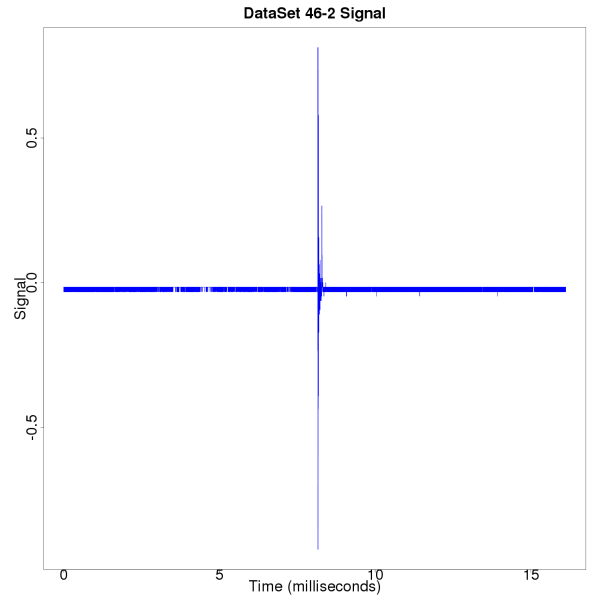

Figure 5. This figure shows the unique signal identified in the cluster plot in Figure 4.

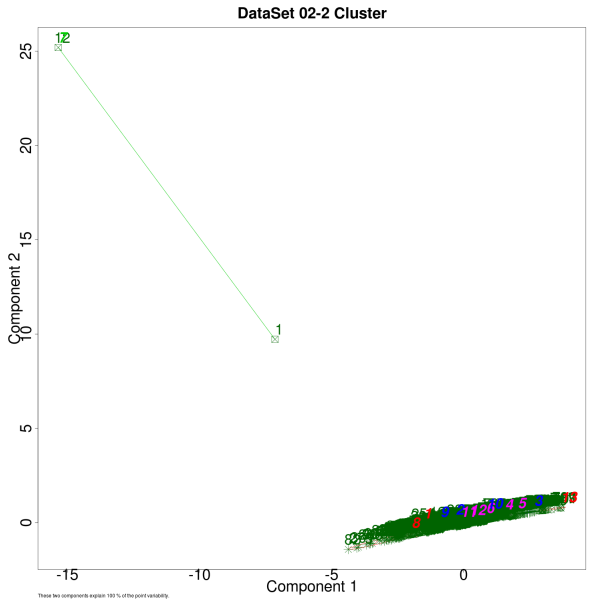

Figure 6. This figure shows a cluster plot shown with the axes being the first two principal component axes.

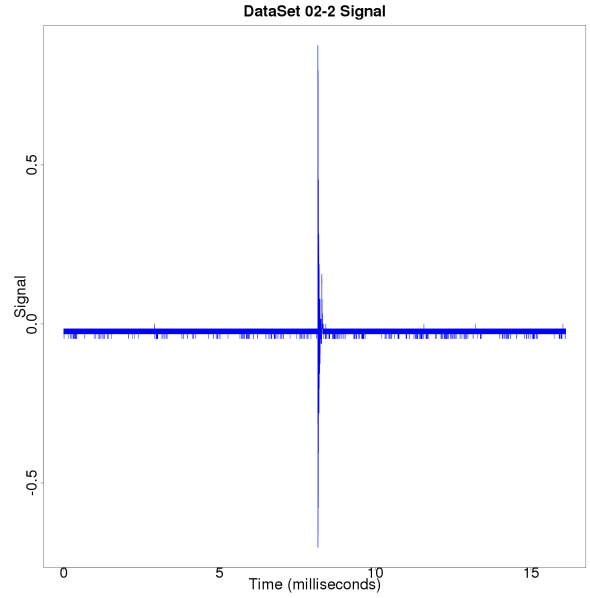

Figure 7. This figure shows the unique signal identified in the cluster plot in Figure 6.

one type of signal occurs in the data set) and the differences in the data are small througout the data set.

An additional comment should be made here noting that with very large data sets, larger than approximately 32GB, it may be better to switch to the k-medoids algorithm instead of the k-means algorithm since it is more robust to noise and outliers. This is because k-medoids minimizes a sum of pairwise dissimilarities rather than the k-means algorithm which minimizes a square of Euclidean distances[12].

V. Results of the Second Clustering/PCA Process

With promising results found from similarities of the principal component plots, we hypothesize that information contained in the clustering/PCA analysis might be enough to classify all the data sets into subsets. Data for each clustering of all 96 data sets were gathered to provide as much statistical data as possible. This was done in two differing techniques, the first was an attempt to characterize the data set by cluster size alone and yielded 22 columns of data for each of the data sets. The hypothesis here is that the distribution of cluster sizes

 ISBN: 978-1-61208-423-7

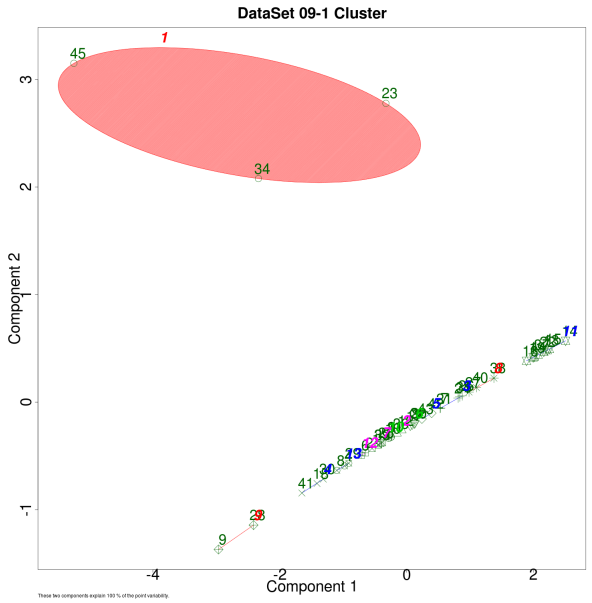

Figure 8. This figure shows a cluster plot shown with the axes being the first two principal component axes.

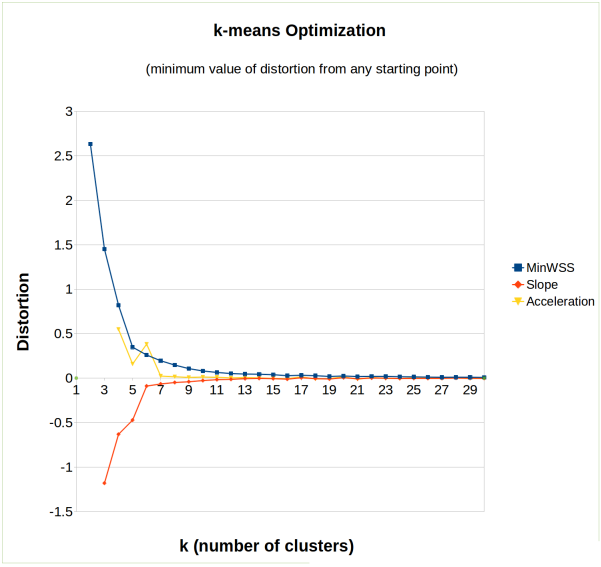

Figure 10. This figure shows a plot of the minimum value of distortion derived from any choice of starting points of the clustering.

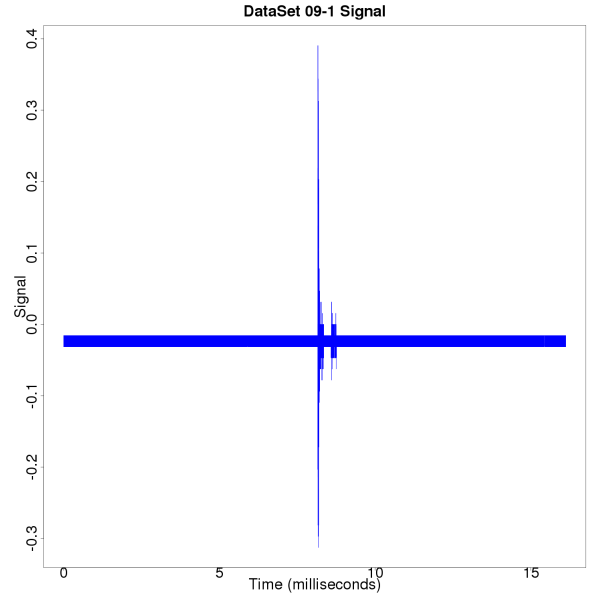

Figure 9. This figure shows the unique signal identified in the cluster plot in Figure 8.

might prove to be enough to classify the data sets. The second technique attempts to capture information from the clustering analysis about the geometry of the clusters by separating them into three parts, the cluster with the smallest number of members, the cluster with the next smallest number of members and by all the number remaining clusters combined. Centers of mass for all three input factors were then calculated for all three inputs and distances were calculated between them. The sizes of the members were normalized and added to this data producing 24 columns of data. The data was processed by cluster size alone to determine the total within sum of squares metric, also known as distortion, this yielded the plot in Figure 10. This plot also includes the first and second derivatives to clarify how the distortion curve moves with increasing numbers of clusters. This shows that there is no clear choice that stems from minimizing distortion. The value of distortion continues to decline as more clusters are added and it is clear that the limit will be reached when there are 96 clusters (the same number as the data points). This is due to the fact that there is no penalty for creating new clusters. A penalty for creating new clusters was added to the calculation by dividing the distortion value by the median number of members in the clusters, this is shown in Figure 14. This works well for geometrically processed data sets (but not for the cluster size processed data sets) and shows a clear minimum of seven as a choice for k.

The geometrical processing of the data sets produced better results that were better defined as evidenced by lesser distortion numbers output from the clustering. The x-means algorithm chose a value for k of 14 clusters whereas, the G-means algorithm chose a value for k of only 7 clusters. Verification of the clusters was done by noting if the data set had signals, what kind, and if there were high degrees of background noise in the data sets. A cluster plot shown in Figure 12 shows clusters 4 and 6 isolated on the left hand side, these clusters have no signals present in any of the data sets enclosed in these clusters. Cluster 1 has the most background noise of the clusters that have signals in them. This cluster also had three data sets in which we haven't found any signals, consequently they are thought to be misclassified by this method. All three of these data sets labeled 22-2, 23-2 and 24-2 in Figure 13 are shown as the points furthest from the center of cluster 1 and closest to cluster 4 by the cluster map plot in Figure 14. This result shows that the 3 misclassified data sets are closest to the boundary of the the data sets that have no signals in them raising the possibility of further improvement by adding another measure to the clustering to eliminate this error. However, due to the assymetric cost of missing a signal of interest over including false positives, we are confident that

 ISBN: 978-1-61208-423-7

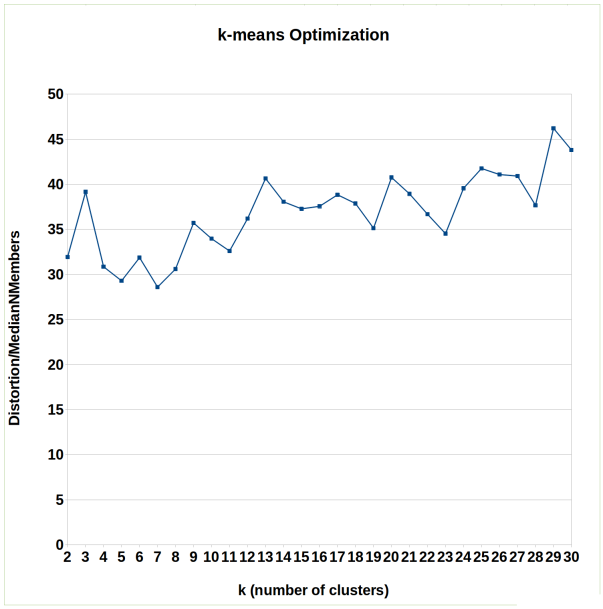

Figure 11. This figure shows a plot of the minimum value of distortion derived from any choice of starting points of the clustering, this time the distortion value is divided by the median number of members in each cluster.

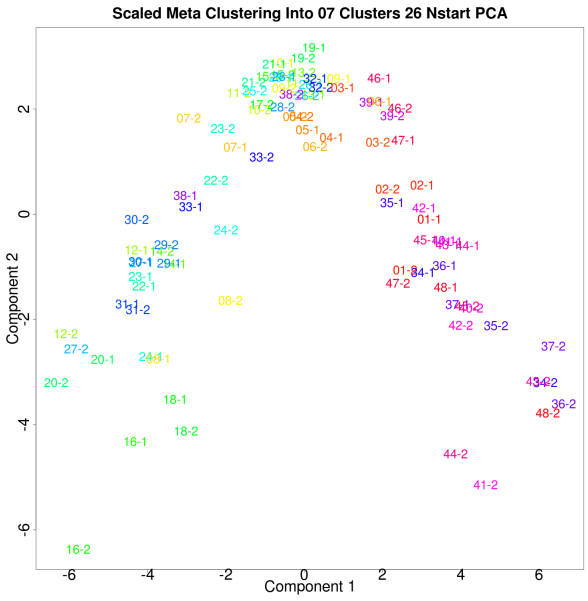

Figure 13. This figure shows a principal components analysis plot with the labels showing the data set identifiers.

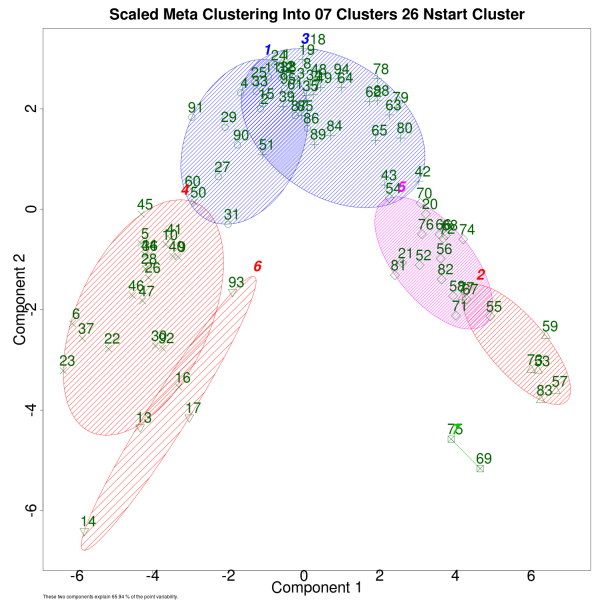

Figure 12. This figure shows a cluster plot of the the lowest distortion plot derived for seven clusters with geometrical processing of the data sets.

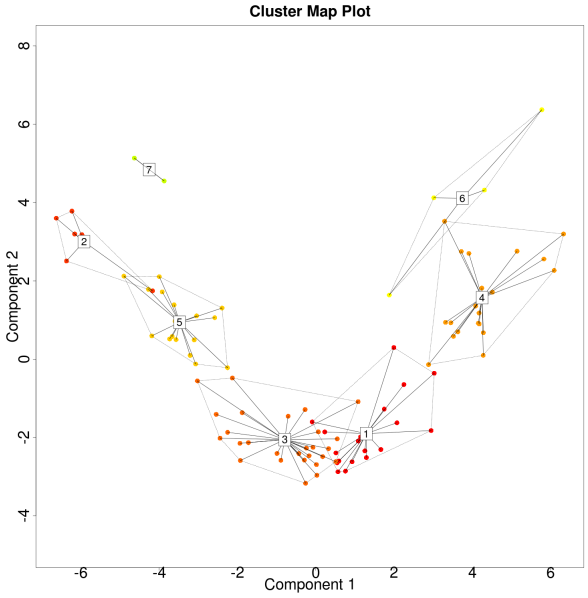

Figure 14. This figure shows a principal components analysis plot with the labels showing the data set identifiers.

method is exceptionally suited to our needs.

VI. Conclusion

A process has been described here that uses Machine Learning algorithms to classify data sets composed of RF signals. Only three measures/features have been found to produce satisfactory results in separating the raw signal data samples into classifications based on a sample size of 32,768 and the necessity of producing spectrograms was eliminated. This minimizes computational complexity while still producing output used in the second stage of the process to classify the data sets. A fast method of classification was found that produces minimal false positive errors while selecting the proper number of clusters without resorting to more computationally complex methods, thereby, decreasing the time spent classifying.

Acknowledgment

The authors would like to thank our colleague Daryl Grunau who was instrumental in making necessary resources available.

 ISBN: 978-1-61208-423-7

REFERENCES

[1] Teledyne Lecroy, "LabMaster 10-100Zi 100 GHz Oscilloscope" url=http://teledynelecroy.com/100ghz/

[2] R. W. Zeng and Y. Chen Li. "Design of Digital Multiple Frequency Notch Filter Based on Free Search Algorithm". Computer Engineering, vol. 40, number 12, 2014, pp. 209-213.

[3] R. G. Lyons. Understanding Digital Signal Processing, Third Edition. Prentice Hall. 2010, ISBN: 9780137027415.

[4] S. K. Mitra. Digital Signal Processing: A Computer-Based Approach, Fourth Edition. McGraw-Hill. 2010, ISBN: 9780077366766.

[5] J. G. Proakis and D. G. Manolakis. Digital Signal Processing: Principles, Algorithms, and Applications, Fourth Edition. Pearson Education Limited. 2013, ISBN: 9781292025735.

[6] A. Coates and A. Y. Ng. "Learning Feature Representations with K-means". Neural Networks: Tricks of the Trade, vol. 7700, 2012, pp. 561-580.

[7] T. W. Liao. "Clustering of time series data–a survey". Pattern Recognition, vol. 38, 2005, pp. 1857-1874.

[8] D. J. Ketchen, Jr and C. L. Shook. "The application of cluster analysis in Strategic Management Research: An analysis and critique" Strategic Management Journal, vol. 17, 1996, pp. 441-458.

[9] H. Abdi and L. J. Williams. "Principal Component Analysis" Wiley Interdisciplinary Reviews: Computational Statistics. 2010.

[10] D. Pelleg and A. Moore, "X-means: Extending K-means with Efficient Estimation of the Number of Clusters" url=http://pelleg.org/shared/hp/download/xmeans.pdf

[11] G. Hamerly and C. Elkan, "Learning the k in k-means" Proceedings of the 17th Annual Conference on Neural Information Processing Systems (NIPS), 2003, pp. 281-288

[12] S. M. Tagaram. "Comparison between K-Means and K-Medoids Clustering Algorithms" International Journal of Advanced Computing. April 2011.

[13] M. Ester, H. Kriegel, J. Sander and X. Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise". Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press., 1996, pp. 226231.

 ISBN: 978-1-61208-423-7