ANL/NSE-21/67

Detection of Anomalies in Gamma Background Radiation Data with K-Means and Self-Organizing Map Clustering Algorithms

Consortium on Nuclear Security Technologies (CONNECT) Q1 Report

Nuclear Science and Engineering Division

ANL/NSE-21/67

Detection of Anomalies in Gamma Background Radiation Data with K-Means and Self-Organizing Map Clustering Algorithms

Consortium on Nuclear Security Technologies (CONNECT) Q1 Report

prepared by
Allen Herrera[1,2] and Alexander Heifetz[1]

[1]Nuclear Science Engineering Division, Argonne National Laboratory
[2]Department of Electrical and Computer Engineering, University of Texas at San Antonio

December 1, 2020

Table of Contents

List of Figures

List of Tables

Abstract

Environmental screening of gamma radiation consists of detecting weak nuisance and anomaly signal in the presence of strong and highly varying background. In a typical scenario, a mobile detector-spectrometer continuously measures gamma radiation spectra in short, e.g., one-second, signal acquisition intervals. The measurement data is a 2D matrix, where one dimension is gamma ray energy, and the other dimension is the number of measurements or total time. In principle, gamma radiation sources can be detected and identified from the measured data by their unique spectral lines. Detecting sources from data measured in a search scenario is difficult due to the highly varying background because of naturally occurring radioactive material (NORM), and low signal-to-noise ratio (S/N) of spectral signal measured during one-second acquisition intervals. The objective of this work is to explore *unsupervised* machine learning (ML) algorithms for detection and identification of weak nuisances and anomalies events in the presence of highly fluctuating background. The challenge is that spectral lines of isotopes are difficult to observe in one-second measurements. Averaging over the entire measurement campaign data set reveals spectral lines of most common background isotopes. Spectral lines of orphan sources, which might appear only in a few measurements during the campaign, will be washed out if averaging is performed over the entire measurement data set. The approach we have explored consists of extracting one-second measurements containing weak spectral features through data clustering. Averaging one-second spectra in a cluster should reveal the presence of anomaly sources. We created two ML models using K-means clustering and Neural Network Self-organizing Map (SOM). Performance of these ML models was benchmarked using search data. One data set contained ^{137}Cs source, and another dataset contained ^{131}I source.

1. Introduction

Environmental screening of gamma radiation consists of detecting weak nuisance and anomaly signal in the presence of strong and highly varying background. In a typical scenario, a mobile detector-spectrometer continuously measures gamma radiation spectra in short, e.g., one-second, signal acquisition intervals [1-3]. The measurement data is a 2D matrix, where one dimension is gamma ray energy, and the other dimension is the number of measurements or total time. In principle, gamma radiation sources can be detected and identified from the measured data by their unique spectral lines. Detecting sources from data measured in a search scenario is difficult due to the highly varying background because of naturally occurring radioactive material (NORM), and low signal-to-noise ratio (S/N) of spectral signal measured during one-second acquisition intervals. The objective of this work is to explore *unsupervised* machine learning (ML) algorithms for detection and identification of weak nuisances and anomalies events in the presence of highly fluctuating background.

As an example, Figure 1 shows images of gamma counts obtained with NaI detectors placed on a mobile platform in a drive through portions of the city of Chicago. Gamma counts per second (CPS), which are integrated over the energy spectrum, are displayed on the city map with pseudo color. Brighter counts indicate larger number of total counts. As seen in the figure, there is significant fluctuation of gamma counts due to NORM in an urban setting.

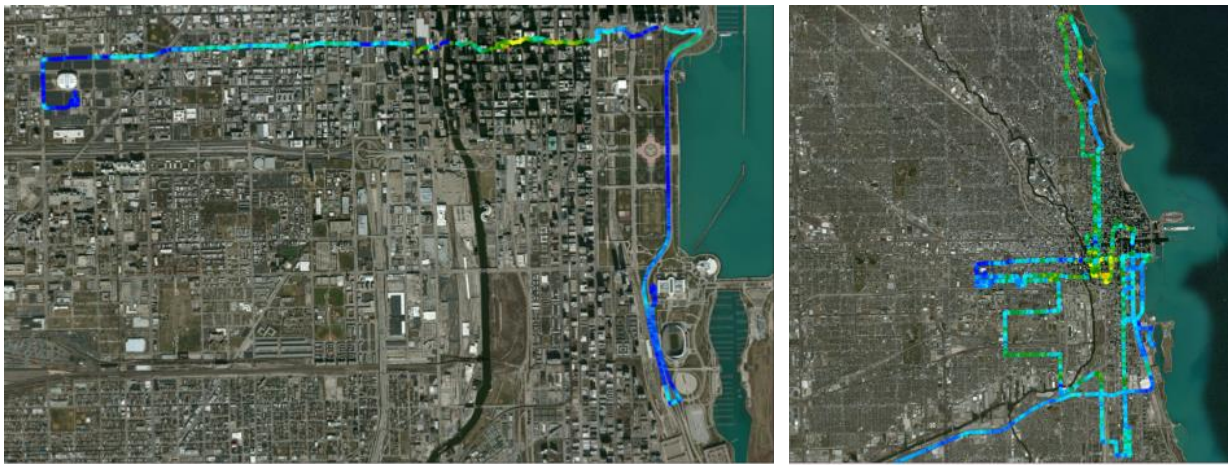

Figure 1 – Gamma counts, measured while driving with NaI detector through sections of the city of Chicago, displayed with pseudo color. Brighter colors indicate larger number of total counts.

We investigate detection of gamma emitting sources in the presence of complex background using unsupervised machine learning. Spectral lines of isotopes are difficult to observe in one-second measurements. Averaging over the entire measurement campaign data set reveals spectral lines of most common background isotopes. Spectral lines of orphan sources, which might appear only in a few measurements during the campaign, will be washed out if averaging is performed over the entire measurement data set. The approach we have explored consists of extracting one-second measurements containing weak spectral features through data clustering. Averaging one-

second spectra in a cluster should reveal the presence of anomaly sources. We created two ML models using K-means clustering and Neural Network Self-organizing Map (SOM). Performance of these ML models was benchmarked using search data. One data set contained ^{137}Cs source, and another dataset contained ^{131}I source.

2. Gamma Source Detection and Identification with Unsupervised Learning Clustering Algorithms

One-second spectra acquired with a moving platform show weak spectral signatures of isotopic sources. Averaging over multiple measurements will increase the S/N of isotopic spectral lines, which would allow for unambiguous detection and identification. However, the key decision consists of choosing the segment of measurement data for averaging. For example, if measurement is performed over all data taken during several-hour environmental screening campaign, spectra of orphan sources is washed out. Signal from an orphan source are most likely to be found in a small subset of total measurements. On the other hand, all measurements contain signals due to isotopes found in NORM. As an illustration, in Figure 2 we plot the spectrum averaged over search time in the data set of 4265 one-second gamma spectrum measurements performed with a moving NaI detector. The measurement set containing 96 one-second spectra of ^{137}Cs isotope, which is not part of the natural background. In the plot of Figure 2 of time-averaged number of gamma counts $<N_\gamma>_T$ as a function of energy E, the peaks are due to NORM. The peak at 662 keV corresponding to ^{137}Cs isotope is not visible.

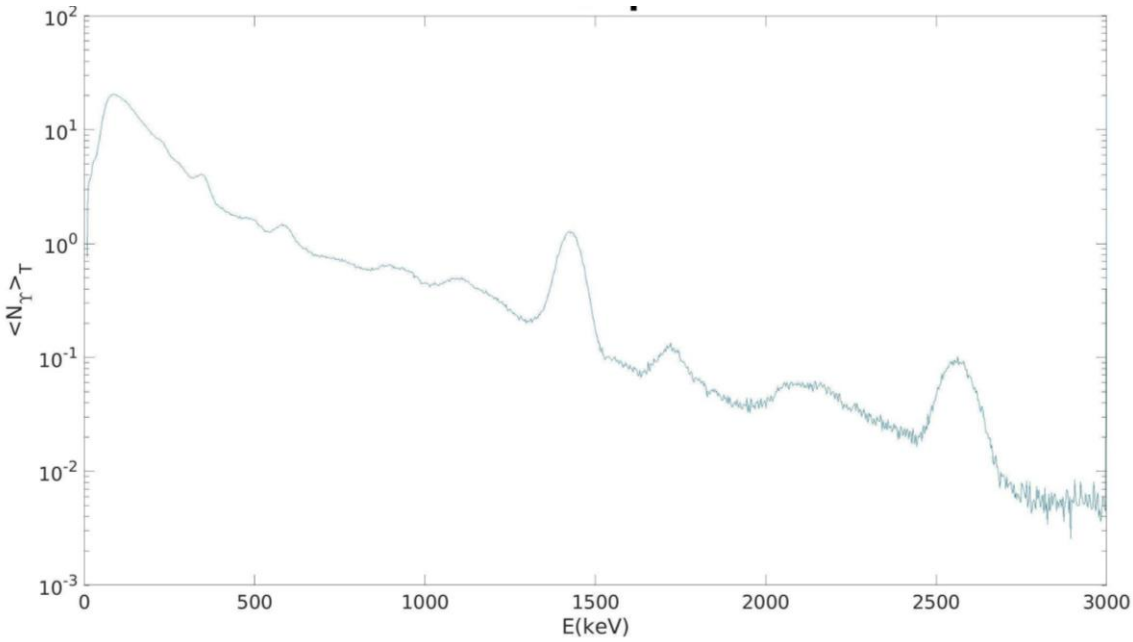

Figure 2 – Gamma spectrum in the energy range 0 – 3000keV averaged over 4265 total measurements. The line of ^{137}Cs isotope at 662keV is washed out.

In our approach, we select a subset of total measurements for averaging using two Unsupervised Learning clustering analysis techniques called Neural Network SOMs [4] and K-means clustering [5]. Clustering is one of the most common exploratory data analysis techniques used to get an intuition about the structure of the data. The method can assist in identifying subgroups in the data such that data points in the same subgroup or cluster are very similar while

data points in different clusters are very different. Clustering analysis can be done on the basis of features, where we try to find clusters of samples based on its feature

Clustering algorithms were used to detect orphan ^{137}Cs and ^{131}I isotopes. The first dataset contained 4265 one-second spectra from a NaI scintillation detector, including 96 one-second spectra of ^{137}Cs source. The second dataset contained 5827 one-second spectra from a NaI scintillation detector, including 89 one-second spectra of ^{131}I source. Both datasets contained 1024 channels ranging from 0 to 3000keV. The models were created using MATLAB Deep Learning Toolbox software. We performed a normalization procedure on both datasets so that the largest value in each spectra was scaled to unity. This normalization procedure ensures that clustering would not be sensitive to fluctuation in total counts. Once the datasets were clustered, we then determined its precision, recall, and F_1 score to evaluate the model with the following equations:

$$Precision = \frac{tp}{tp + fp} \quad (1)$$

$$Recall = \frac{tp}{tp + fn} \quad (2)$$

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \quad (3)$$

where t_p is true positives, f_p is false positives, and f_n is false negatives. For K-means clustering, due to the varying centroids, we took an average of the precision and recall between ten trials, which we then used to create the average F_1 score.

3. Gamma Source Detection and Identification with K-Means Clustering Algorithms

K-means algorithm is an interactive algorithm that tries to partition or separate the dataset into sections or K clusters. Each data point will belong to only one cluster and the algorithm tries to make the intra-cluster data point as similar as possible while also keeping the clusters as different as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid is at the minimum. Therefore, for each model we will need to specify the number of clusters K, initialize centroids by shuffling the dataset, then randomly selecting the centroids for each K, iterate until there is no change in the centroid, and compute the sum of the squared distance between data points and all centroids.

3.1. K-means clustering of dataset with ^{137}Cs source

Applying the K-means clustering algorithm to the dataset containing ^{137}Cs source, we used K = 11 for a total of 11 clusters. Each iteration of K-means was also able to visualize ^{137}Cs in one of the clusters with the number of predictions ranging from 82 - 115 one-second spectra, where the original dataset had 96 one-second spectra of ^{137}Cs source. Averaged gamma spectrum of K-means cluster with 84 one-second spectra, is plotted in Figure 4. Averaging of one-second spectra in the anomaly cluster reveals ^{137}Cs 662keV line.

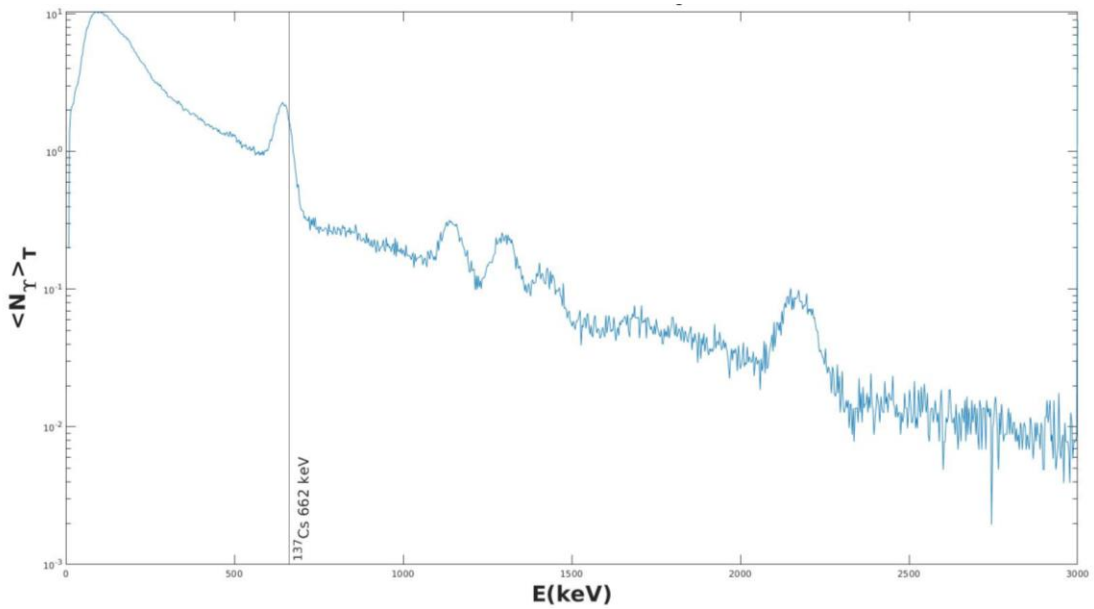

Figure 3 – Averaged gamma spectrum of K-means anomaly cluster with 84 one-second spectra. Averaging of one-second spectra in the anomaly cluster reveals ^{137}Cs line.

Average results for K-means clustering of the data set with ^{137}Cs source are shown in Table 1. The average F_1 score is 85.28%.

Table 1 – Precision, Recall, and F1 score for K-means clustering of data with ^{137}Cs source

Samples	96
# of trials	10
Predicted	82 - 115
Average Precision	82.57%
Average Recall	81.67%
Average F_1 score	**85.28%**

3.2. K-means clustering of dataset with ^{131}I source

Applying the K-means clustering algorithm to the dataset containing ^{131}I isotope, we used K = 3 for a total of 3 clusters. Each iteration of K-means was also able to visualize ^{131}I in one of the clusters with the number of predictions ranging from 91 - 92 one-second spectra, where the original dataset had 89 one-second spectra of ^{131}I. Averaged gamma spectrum of K-means cluster with 91 one-second spectra, is plotted in Figure 5. Averaging of one-second spectra in the anomaly cluster reveals ^{131}I isotope 364keV line.

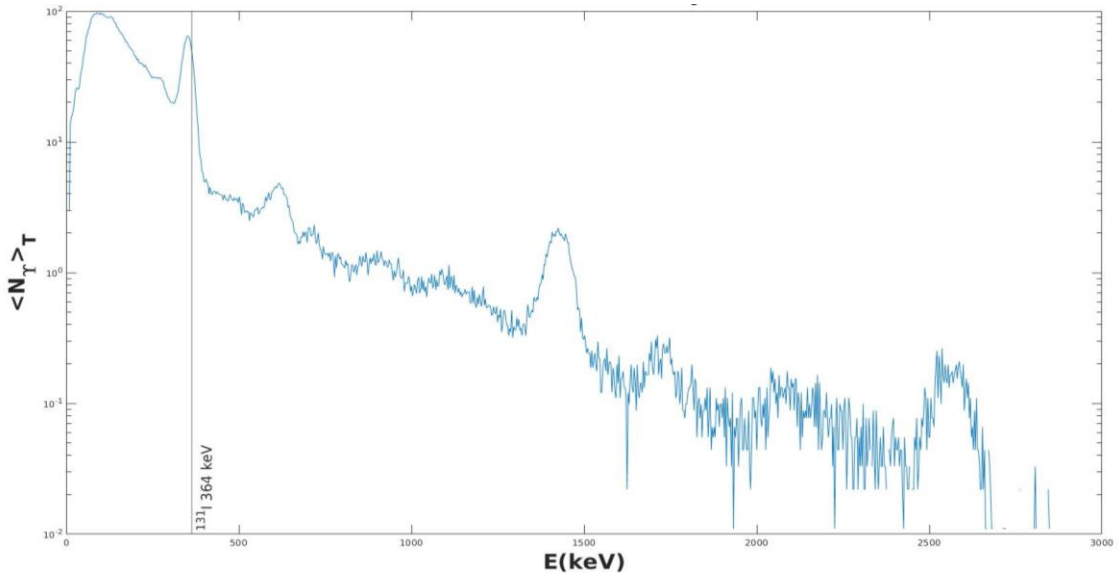

Figure 4 – Averaged spectrum on K-means anomaly cluster with 91 one-second spectra. Averaging of one-second spectra in the anomaly cluster reveals ^{131}I line.

Average results for K-means clustering of the data set containing ^{131}I source are shown in Table 2. The average F_1 score is 91.11%.

Table 2 – Precision, Recall and F_1 score for SOM for ^{131}I source detection.

Samples	89
# of trials	10
Predicted	91 - 92
Average Precision	91.11%
Average Recall	91.13%
Average F_1 score	**91.11%**

4. Gamma Source Detection and Identification with Neural Network Self-Organizing Map (SOM) Clustering Algorithm

A self-organizing map (SOM) is a type of artificial neural network (ANN) that uses unsupervised learning to produce a low dimensional, discretized representation of the input space of the training samples, called a map, and is therefore a method to do dimensionality reduction. With the use of competitive learning, as opposed to backpropagation like other ANNs, a SOM can use a neighborhood function to preserve the topological properties of the input space. The algorithm begins by first initializing each node's weight, then a vector is chosen at random from the set of training data where each node is examined to calculate which one's weights are most like the input vector. The winning node is known as the Best Matching Unit (BMU), which is a technique that calculates the distance from each weight to the sample vector, by running through all weight vectors. The weight with the shortest distance is the winner. The winning weight is rewarded with becoming more like the sample vector. The neighbors also become more like the sample vector. The closer a node is to the BMU, the more its weights get altered and the farther away the neighbor is from the BMU, the less it learns. This process then repeats for every input vector in the dataset for N epochs.

4.1. SOM clustering of dataset with ^{137}Cs source

For applying the SOM network to the dataset containing ^{137}Cs source, we used a map size of 3 and trained the network for 200 epochs which resulted in a total of 9 clusters. Plotting of all the partitions revealed that one of the clusters could visualize the peak at 662 keV for ^{137}Cs, as shown in Figure 6. The model had placed 101 one-second spectra into the same cluster while the original dataset had 96 one-second spectra of the ^{137}Cs source. Averaging of one-second spectra in the anomaly cluster reveals ^{137}Cs 662keV line.

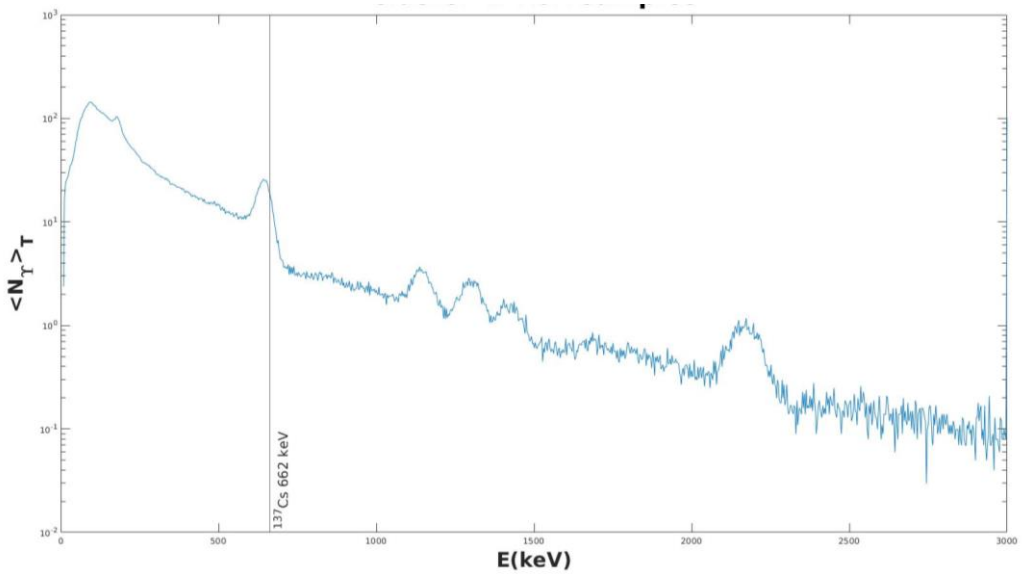

Figure 5 – Averaged gamma spectrum of SOM anomaly cluster with 101 one-second spectra. Averaging of one-second spectra in the anomaly cluster reveals ^{137}Cs line.

Table 3 shows the calculated precision, recall, and F_1 score of the SOM model for the dataset containing ^{137}Cs source. The F_1 score is 85.28%.

Table 3 – Precision, Recall and F_1 score for SOM for ^{137}Cs source detection.

Samples	96
Predicated	101
True Positives	84
False Negatives	12
False Positives	17
Precision	83.17%
Recall	87.50%
F_1 score	**85.28%**

4.2. SOM clustering of dataset with ^{131}I source

For applying the SOM network to the dataset containing ^{131}I isotope, we used a map size of 2, and trained the network for 200 epochs which resulted in a total of 4 clusters. Plotting of all the partitions revealed that one of the clusters could visualize the peak at 364 keV for ^{131}I as shown in Figure 7. The model had placed 91 one-second spectra into the same cluster, while the original dataset had 89 one-second spectra of the ^{131}I source. Averaging of one-second spectra in the anomaly cluster reveals the ^{131}I isotope 364keV line.

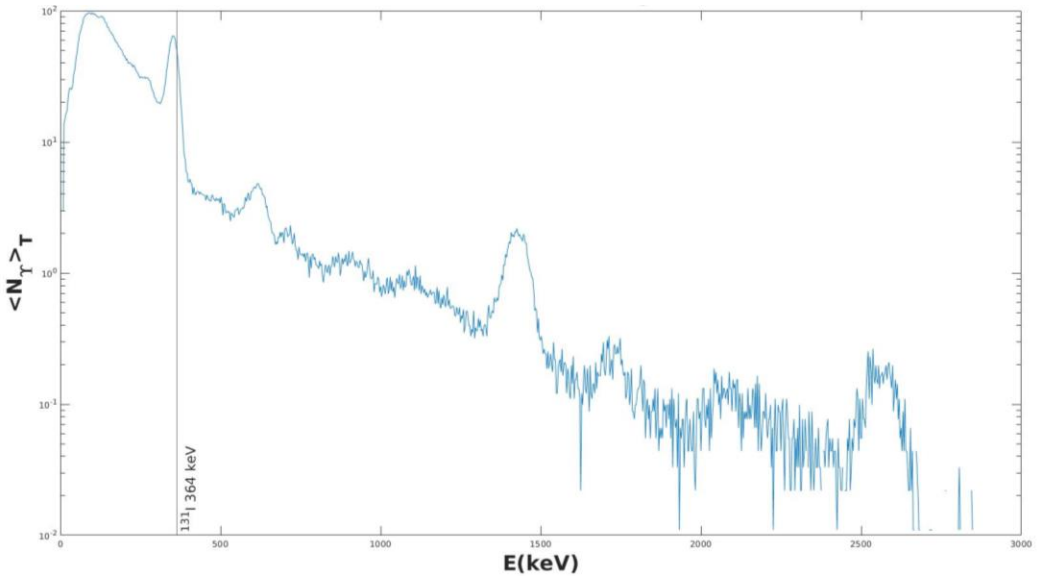

Figure 6 – Averaged gamma spectrum of SOM cluster with 91 one-second spectra. Averaging of one-second spectra in the anomaly cluster reveals the ^{131}I line.

Table 4 shows the precision, recall, and F_1 score of the SOM model for the data set containing ^{131}I source. The F_1 score is 91.23%.

Table 4 – Precision, Recall and F_1 score for SOM for ^{131}I detection.

Samples	89
Predicated	91
True Positives	82
False Negatives	7
False Positives	9
Precision	90.11%
Recall	92.36%
F_1 score	**91.23%**

5. Summary of Clustering Algorithm Benchmarks

The results have shown that both algorithms can successfully cluster both dataset sources into a single cluster with an accuracy of greater than 82%. Comparing both algorithms together, Neural Network SOMs outperform K-means clustering in both the F_1 scores metric for the datasets, and in the algorithm run time metric. Table 5 shows performance benchmarking results of the two algorithms for two different data sets.

Table 5 – Benchmarking of clustering algorithms performance

Algorithm	**^{137}Cs Data Set (F_1 score)**	**^{131}I Data Set (F_1 score)**	**Run time**
Neural Network SOM	85.28%	91.23%	~0.1 seconds
K-means clustering	82.11%	91.11%	~0.5 seconds

6. Conclusions

We have investigated several unsupervised machine learning algorithms for analysis of gamma spectrum measurements obtained in environmental wide are screening with a moving NaI detector-spectrometer. For weak anomaly and nuisance detection, we developed K-means clustering and neural network self-organizing maps (SOM) algorithms. The validation study consisted of two data sets of spectra measured in one-second intervals with a Sodium Iodide detector. The first dataset with over 4000 spectra consisted mostly of urban background measurements and approximately 90 measurements of ^{137}Cs. The second dataset consisted of over 5000 spectra mostly of urban background and approximately 90 measurements of ^{131}I. Using clustering analysis, we observed that a majority of spectra containing nuclear source signals clustered away from the background, producing point-like clusters when visualizing in search-time averaged number of gamma counts. One cluster in both datasets contained spectra with strong ^{137}Cs peaks and ^{131}I in over 85% of the clustered samples. The other clusters contained no visible peaks of the nuclear source in their spectra.

Development of a robust clustering technique for detection of sources will require further algorithm optimization sensitivity of cluster feature identification in the spectrum to the detector response function of NaI. This would provide a better understanding of the limits of the source detection capability of cluster based techniques. It is expected that cluster performance will depend on such factors as source spectrum and signal strength, background isotopic composition and variability with time. Such studies will also provide an indication if the detector spectral channels can be ranked in order of importance for cluster analysis. Channels with least importance could be excluded from data to reduce its size, and hence increase the speed of analysis.

References

1. Weinstein, M., Heifetz, A., & Klann, R. (2014). Detection of nuclear sources in search survey using dynamic quantum clustering of gamma-ray spectral data. The European Physical Journal Plus, 129(11), 239.
2. Alamaniotis, M., Heifetz, A., Raptis, A. C., & Tsoukalas, L. H. (2013). Fuzzy-logic radioisotope identifier for gamma spectroscopy in source search. IEEE Transactions on Nuclear Science, 60(4), 3014-3024.
3. Bai, E. W., Heifetz, A., Raptis, P., Dasgupta, S., & Mudumbai, R. (2015). Maximum likelihood localization of radioactive sources against a highly fluctuating background. IEEE Transactions on Nuclear Science, 62(6), 3274-3282.
4. Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. IEEE transactions on pattern analysis and machine intelligence, 24(7), 881-892.
5. Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. IEEE Transactions on neural networks, 11(3), 586-600.

Nuclear Science and Engineering (NSE) Division
Argonne National Laboratory
9700 South Cass Avenue, Bldg. 208
Argonne, IL 60439

www.anl.gov