2009-03

Clustering and Outlier Analysis For Data Mining (COADM)

Seng, Choo Chwee

Clustering and Outlier Analysis For Data Mining (COADM)

Choo Chwee Seng[1]
DSO National Laboratories, Singapore
Ng Ee Chong
DSO National Laboratories, Singapore
Chua Ching Lian
DSO National Laboratories, Singapore

INTRODUCTION

The Clustering and Outlier Analysis for Data Mining (COADM) tool is one of the three key components delivered under the Systematic Data Farming (SDF) project [1]. SDF was sponsored by the Singapore Armed Forces (SAF) Centre for Military Experimentation (SCME) and was completed in 2005.

OBJECTIVE

The objective of COADM is to provide an additional dimension to data analysis, especially when there is a large amount of output generated through data farming. It aims to complement statistical analysis by grouping the data into "good" and "bad" clusters, and identifying the associated parameters so as to provide insights on how to get into "good" clusters and avoid the "bad" ones. COADM also identifies the outliers in each cluster, and in doing so try to discover "surprises".

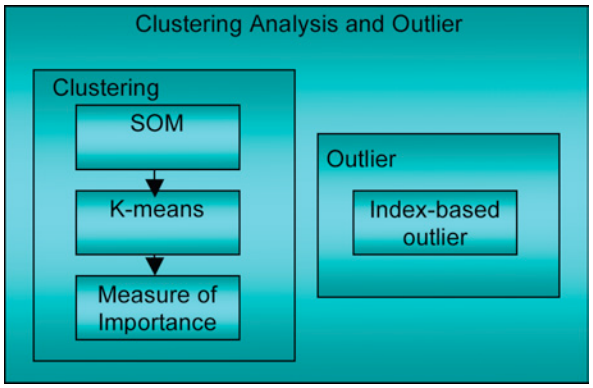

Figure 1: Key Features of COADM

KEY FEATURES

Figure 1 shows the key features of COADM and the underlying techniques and algorithms used.

The Clustering Analysis was based on K-Means methodology coupled with Self-Organising Maps (SOM) to help organise the data into clusters. The incorporation of K-means was to help improve the clustering and segregation capability of the SOM [2].

Based on the Clusters identified, a search was carried out within to identify the points that are "most different" from the rest of the data points within the same cluster, i.e. the outliers. This was achieved by comparing the Euclidean Distance of each data point with its k-nearest neighbour in each cluster and finding the one with the largest Euclidean Distance [3].

COADM was developed from several open source software packages and DSO contributions were in synthesizing the various algorithms/packages to form a package (coded in JAVA) capable of extracting information from numerical data sets. The SOM program used in this package was derived from the SOM toolbox in Matlab [3]. This toolbox is capable of visualizing complex data set, courtesy of Matlab's great visualization tools; moreover it keeps track of much information which greatly facilitates the data mining process. The outlier algorithm was coded and modified slightly for integration with other packages. There is also a WEKA package provided as an extra data visualizations tool for a more detail examination of the clustering results.

DEMONSTRATION

Scenario

An Urban Scenario was used to demonstrate the key features of COADM (see Figure 2).

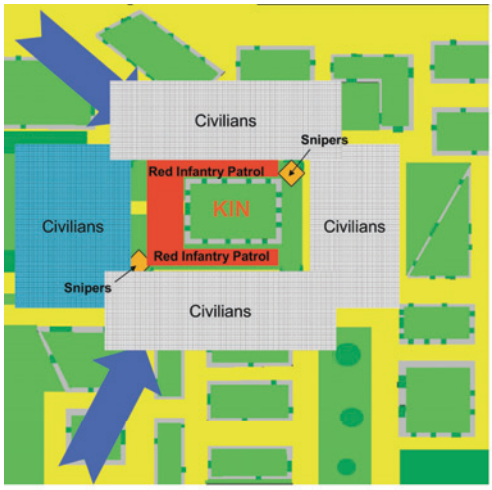

Figure 2: Urban Scenario Setup in MANA

An Urban Area of Operations (AO) 2km by 2km in size was set up in MANA. The scenario was set in this Urban AO where 2 platoons of Blue Infantry soldiers (21 soldiers

[1] For more information contact: Choo Chwee Seng, cchweese@dso.org.sg

per platoon), each platoon was supported by 3 MG-mounted soft-skin vehicles, attempted to take over a Key Installation (KIN) held by a platoon of Red Infantry soldiers (21 soldiers). The Red Infantry defence was assisted by two teams of Red snipers (4 snipers in total). The Blue agents' task was made more difficult by the crowd of hostile Civilians congregating near to the KIN and randomly attacking the Blue agents when they were encountered.

Blue Force has 3 Courses of Actions:

d. OCA 1. The Blue agents advanced from the northwest and southwest direction of the map towards the objective, attempting to take out the Red from both sides.

e. OCA 2. The Blue agents were concentrated in the southeast area of the map and advance as a force towards the Red, attempting to punch through the Red defence from a single direction.

f. OCA 3. The Blue agents were spread out on the northern portion of the map and attempted to flush out the read through a swarming approach.

Red Force has 2 Courses of Actions:

g. ECA 1 - All Red agents resided within the building's compound and defended their base from there.

h. ECA 2. A section minus of 6 Red agents lay hidden in an adjacent building as backup to the other two sections in the defended locality. They were called in when the Red agents came in contact with Blue Forces.

Design of Experiment

A hybrid design was formed using the Excel-based Latin Hypercube (LHC) Generator by crossing the 30-factor LHC with the 2-factor Full Factorial design for the OCA and ECA factors. The resultant hybrid design had 6000 design points and sent for data farming (with 100 replications each).

Analysis of Results

The large dataset of MOEs obtained from the data-farming output was analyzed using COADM and some interesting insights were derived. Figure 3 shows some of the selected component plots of the SOM clusters generated by the COADM. Similar distribution of colours on the component plots implies correlation. Hence correlation between the factors and the MOEs can be discovered. Factors found to be correlated to MOEs are also the main factors contributing to the MOEs.

Both the OCA and ECA factors were observed to be uncorrelated with the MOEs. The distribution patterns of the OCA and ECA factors (shown on Figure 3) were observed to be rather independent from the distribution patterns of the MOEs. Hence, varying the OCA and ECA would not contribute to significant changes to the MOEs.

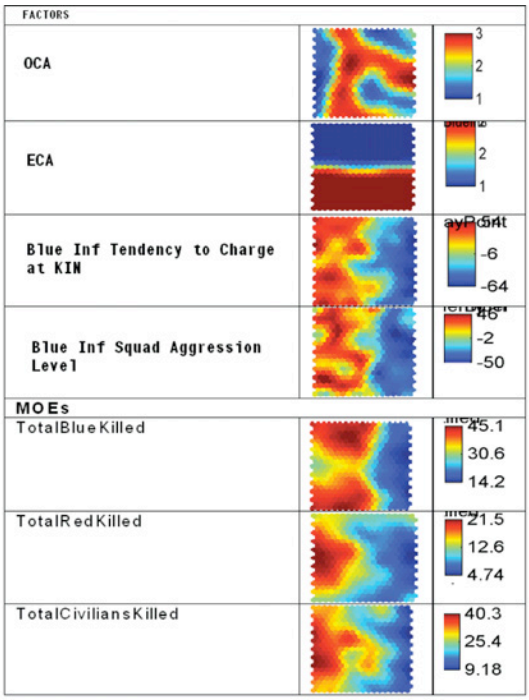

Figure 3: Component Plots of SOM clusters for selected Factors and MOEs

The MOEs were observed to be somewhat correlated. This suggested that achieving high Red attrition would likely coincide with high Blue and Civilian attrition levels. The Red and Civilian casualties were more closely correlated with each other compared with that of the Blue casualties. Therefore, it would suggest that larger number of civilian casualties was unavoidable in this scenario, if the Blue agents or Red agents attempted to maximize the casualties on either sides.

However, there were exceptions. A region that contained outcomes that corresponded to moderate Blue attrition but very high Red attrition was shown in Figure 4. This would be the region of most interest to Blue as the parameter values defined in this region allowed Blue to achieve its mission of killing as many Red as possible, while incurring moderate losses.

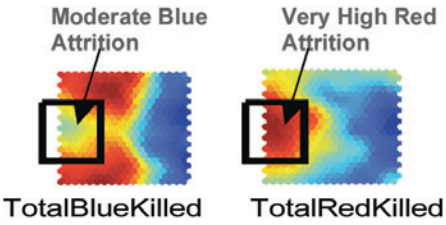

Figure 4: Region of Outcomes corresponding to Moderate Blue Attrition but Very High Red Attrition

Of the 32 farming parameters, it was observed that "Blue Infantry Tendency to Charge at KIN" and "Blue Infantry Squad Aggression Level" correlate most closely with the MOEs, and were hence most influential on the MOE outcomes.

It was interesting to revisit the region spotted under Figure 4, where Blue suffered moderate attrition but Red

suffered high attrition. As shown in Figure 5, in this region, the parameter values for "Blue Infantry Tendency to Charge at KIN" and "Blue Infantry Aggression Level" should define the Blue's behavior that would inflict high Red attrition while sustaining moderate Blue attrition.

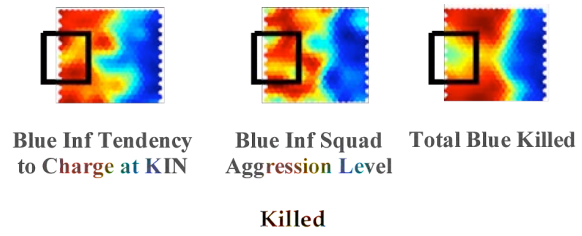

Figure 5: Comparison of Blue Inf Tendency to Charge at KIN, Blue Inf Squad Aggression Level, and Total Blue Killed

COADM tool revealed that the data points can be organized into 20 clusters (see Figure 6). The mean parameter values and MOEs for each cluster were obtained based on the data points within the cluster. By analyzing each cluster, it can identify the clusters that contained generally favorable outcomes for Blue and those that contained generally bad outcomes for Blue.

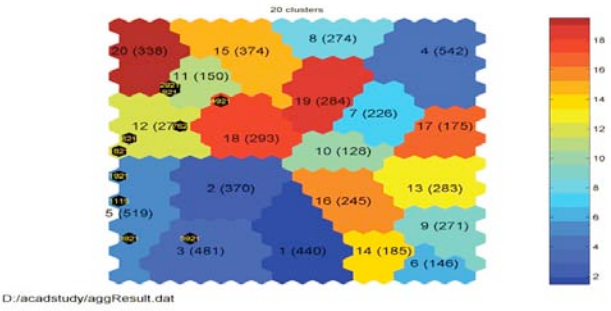

Figure 6: Clustering of Data Farming Output

COADM can also identify contributing factors and behavior that resulted in each of these clusters. Without going into each cluster in detail, with this analysis, Blue would know how to manipulate Blue factors and make decisions to avoid those bad clusters and shift towards the good clusters.

From the output generated by COADM, the outlier points were examined in greater detail and they were laid out in Table 1 in terms of the MOEs. The top outlier was case number 5921 (or Data Point 5921) amongst the 6000 cases in the Experimental Design. This case belonged to Cluster 3 and had 23.45 Red killed in total. COADM identified this case as an outlier because 23.45 red killed was 1.936 times more than Cluster 3's mean value of total Red killed. A value that is 1.5 times either side of the mean would normally be considered as an outlier.

In Cluster 3, Blue generally suffers high attrition and hence Blue should avoid parameter values that will cause them to fall into this cluster. This outlier Case 5921 is an interesting case because it is the best outcome in a bad cluster for the Blue, as Blue was able to inflict much higher Red attrition compared to other cases in Cluster 3.

Case 5921 described a Blue force that was very fast, highly aggressive and extremely stealthy. Although the Red force and Civilians were also generally aggressive, they were less so compared to the Blue force.

Hence, if factors uncontrollable by the Blue Force, such as Red Force tactics and behavior, resulted in the circumstances becoming unfavourable (e.g. falling into Cluster 3 outcomes), Blue force must attempt to exploit outlier case 5921 by moving swiftly and stealthily, and engaging more aggressively than the Red force inflict high Red casualties.

Case	Dist	Cluster	TotalBlueKilled	TotalRedKilled	TotalCiviliansKilled
5921	43.13	3	34.65 (+0.175)	23.45 (+1.936)	43.73 (+1.565)
4921	42.56	18	37.68 (+0.413)	22.88 (+1.838)	42.06 (+1.423)
1921	42.13	5	36.25 (+0.301)	23.63 (+1.966)	42.36 (+1.449)
921	41.93	11	37.89 (+0.430)	23.29 (+1.908)	40.92 (+1.327)
1115	41.31	5	40.47 (+0.633)	23.12 (+1.879)	46.93 (+1.835)
821	41.25	12	41.31 (+0.700)	21.83 (+1.657)	42.67 (+1.475)
2921	41.2	11	41.70 (+0.730)	20.24 (+1.385)	37.31 (+1.022)
1821	41.11	5	41.51 (+0.715)	20.69 (+1.462)	43.27 (+1.526)
3921	41.04	3	42.64 (+0.805)	20.34 (+1.402)	35.59 (+0.876)
762	40.99	12	29.84 (-0.205)	24.11 (+2.049)	45.98 (+1.755)

Table 1: MOEs in Outlier Cases

INSTALLATION

The installation requirements for COADM are as follows:

a. Java 1.4.2 and above.
b. Windows OS 2000/XP.
c. Memory recommended, 256MB Ram.
d. Disk storage space for files, 260MB

To request a copy of COADM, please contact Choo Chwee Seng at cchweese@dso.org.sg.

REFERENCES

[1] C. S. Choo, E. C. Ng, C. K. Ang, and C. L. Chua. Systematic Data Farming: An Application to a Military Scenario, Proceedings of Army Science Conference 2006, Florida, USA.

[2] J. Vesanto, E. Alhoniemi, K. Kiviluoto, and J. Parvianen. Self- Organizing Map for Data Mining in Matlab: The SOM Toolbox. www.cis.hut.fi/projects/somtoolbox.

[3] S. Ramaswamy, R. Rastogi, and K. Shim. 2000: Efficient algorithms for mining outliers from large datasets. In Proceedings of the ACM SIGMOD Conference, pages 427–438.