Euclidean Distance Clustering in Spectroscopy, with applications for polymorphs, formulated products, and other areas.

Michael Boruta

Advanced Chemistry Development Inc.
(ACD/Labs), Toronto, Canada

ABSTRACT

Finding which samples are similar to each other is a task often encountered in many areas of chemistry, from looking at polymorphs and salt forms in preformulation to determining similar groupings of data in competitive analysis or formulated products. There are many approaches to clustering data; this poster will look at the Euclidean Distance algorithm, some of the limitations in this approach and the data review steps required when clustering data.

ALGORITHMS

Euclidean Distance

$$HQI^2 = \frac{\sum\left(\left(U_i/\overline{U}\right)-\left(R_i/\overline{R}\right)\right)^2}{(npts)}$$

1st Derivative Euclidean Distance

$$HQI^2 = \frac{\sum\left(\left((U_i - U_{i+1})/\overline{U}\right)-\left((R_i - R_{i+1})/\overline{R}\right)\right)^2}{(npts)}$$

Where

$$\overline{U} = \sqrt{\sum U_i^2} \quad \text{and} \quad \overline{R} = \sqrt{\sum R_i^2}$$

INTRODUCTION

Definition: "Cluster analysis or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics."[1]

Using ED clustering (Euclidean Distance) or 1st derivative ED is similar to spectral searching – we are using the same algorithm to generate a sorted list of HQI values (Hit Quality Index). Rather than using the reported match value or the position of a hit in the result, the difference or gap in the HQI between successive hits can be used as an indicator of the quality of a match.

Clustering data will always give you some type of result – good or bad – reviewing those results is a requirement. Reviewing tools include overlaying spectra for comparison, moving spectra between groups, displaying other analytical data or meta data, and marking spectra as being reviewed.

Advanced Chemistry Development, Inc.
Tel: (416) 368-3435
Fax: (416) 368-5596
Toll Free: 1-800-304-3988
Email: info@acdlabs.com
www.acdlabs.com

PROCESS

The process involves 2 steps. The first step analyzes the gaps to create an initial set of clusters, while the second step tests each member of cluster against all the other clusters.

To create the clusters in the first pass, we start by comparing each spectrum in the dataset to all other spectra in the dataset. From this we create a sorted list of HQI values and calculate the gap and % gap for each comparison.

The HQI values generated by the ED algorithm are not by themselves the best way to evaluate the quality of a cluster. This is due to several reasons, including the fact that the HQI's generated by the Euclidean Distance algorithm do not decrease linearly. Meaning a gap of 5 units near 90 has a different meaning than a gap of 5 units near 60. Because of this none linearity we also included a fudge factor – the % gap is modified by its position in the sorted list, giving greater weight to gaps farther down the list than if we had just used the % gap itself.

The sorted lists are then scanned to find the largest discriminator and all spectra above this point become members of the first group. The process is repeated until there are no spectra left.

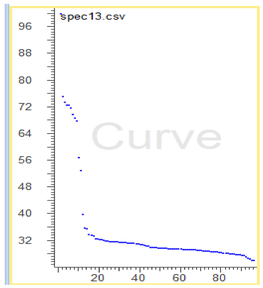


Figure 1A. Graph

	CS	CT	CU	CV	CW	CX
1	spec#	Spec13	delta	delta%	Discrim	OrigIndex
2	13	100	24.77	33.48	33.4833	1
3	85	75.2325	1.85	2.51	5.0126	2
4	77	73.3786	0.76	1.02	3.07	3
5	21	72.6217	0.11	0.15	0.5958	4
6	37	72.5115	0.84	1.13	5.6631	5
7	29	71.6737	1.81	2.45	14.6788	6
8	61	69.8641	1.05	1.42	9.9295	7
9	69	68.8148	1.03	1.39	11.1221	8
10	5	67.7864	10.85	14.67	132.0111	9
11	93	56.9367	3.9	5.27	52.6749	10
12	53	53.0403	13.1	17.7	194.7476	11
13	8	39.9445	4.17	5.64	67.7179	12
14	45	35.7703	0.15	0.21	2.6893	13
15	40	35.6173	1.73	2.34	32.7789	14
16	65	33.8854	0.27	0.37	5.4789	15
17	90	33.6152	0.22	0.3	4.7609	16
18	7	33.3951	0.8	1.09	18.4472	17
19	72	32.5924	0.11	0.15	2.6624	18
20	42	32.483	0.04	0.06	1.138	19
21	18	32.4387	0.04	0.06	1.1623	20
22	47	32.3957	0.3	0.4	8.4929	21
23	95	32.0966	0.1	0.14	3.0178	22
24	10	31.9951	0.08	0.11	2.5324	23
25	82	31.9137	0.11	0.15	3.4917	24

Figure 1B. Table

Figure 1 A and B. Graph of spectrum 13's HQI values versus all members of the dataset, along with a table of the calculated values for the gap, %gap and the discriminator.

Once the first pass is complete, a second analysis step occurs whereby each spectrum is compared to the average spectrum of each of the other clusters. (The spectrum is removed from it's own average cluster prior to its comparison.)

Figure 1 shows the results for a single spectrum during the clustering of a 96-well plate of XRPD data. Figure 1A shows a graph of the HQI values for spectrum number 13 compared to all other spectra. Figure 1B shows the values used to make the comparison for the first pass. In this case the largest discriminator occurs at position 11 – therefore all spectra from position 11 and above become members of the first group.

REVIEW

The results of the clustering can be displayed as a graph of each member of a cluster compared to the average spectrum for the cluster, Figure 2. The graph view allows one to visualize how tight each cluster is and easily see the members that are farthest away from the center – the possible outliers. Along with the graph view, the spectrum and any meta data can be displayed to assist in the review of the clusters.

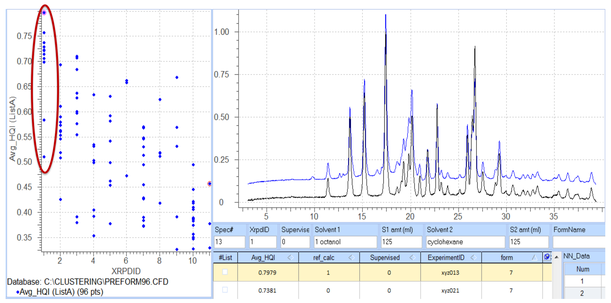


Figure 2. Results of clustering a 96-well plate of XRPD data. In the graph on the left we have the distance of each member of a cluster from the average spectrum for that cluster along the Y axis, and the individual clusters along the X axis. Each vertical column represents a group or cluster of spectra. The red oval highlights cluster number 1. The blue spectrum is the average spectrum for group 1. The black spectrum is the highlighted spectrum.

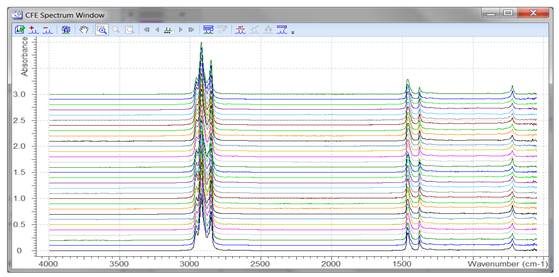


Figure 3. 34 IR spectra of competitive oils

Figure 3 shows an example of IR spectra for several oils. Visually the series of spectra look very similar and would be difficult to classify manually.

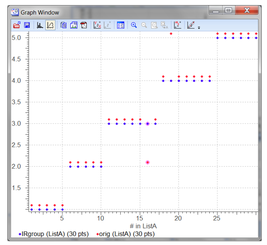
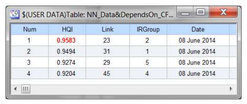


Figure 4 shows the spectra from figure 3 clustered into groups. The red dots indicate the clusters found by PCA and the blue dots show the results of the ED clustering. This is not meant to show that one method is better than another, however it does show that while the 2 methods are in agreement for most spectra, there are two points that are not in agreement and should undergo further review.

Figure 4. Graph of PCA results compared to ED results



Num	HQI	Link	IRGroup	Date
1	0.9583	23	2	08 June 2014
2	0.9494	31	1	08 June 2014
3	0.9274	29	5	08 June 2014
4	0.9204	45	4	08 June 2014

Figure 5.

Figure 5 shows a nearest neighbor table for one point of disagreement in figure 4. Because clustering results can sometimes be ambiguous, this implementation of clustering allows a user to move members from one cluster to another, or join or split clusters. The nearest neighbor table show the next best match for any member of a cluster. In this case the next best match for the highlighted blue dot (a member of ED group 3) is group 2 – the same as suggested by the PCA analysis. The HQI listed in the table is highlighted in red, meaning that it is closer to an individual spectrum in group 2 than it is to the center of group 3. Clicking on the "link" field in the table would allow a user to overlay that spectrum for comparison.

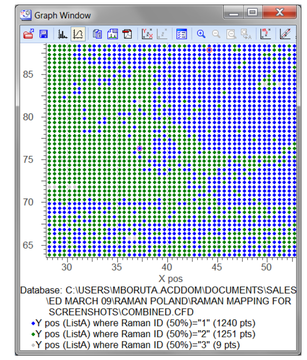


Figure 6. Raman image analysis

Figure 6 shows a graph of pixels from a Raman image. The results of the clustering gave 3 groups. In this view the X and Y axis represent the position of pixels from the Raman image. The graph points, or "pixels", are colored by their group number.

DISCUSSION AND LIMITATIONS

Generally we have been using a 1st Derivative Euclidean Distance rather than a straight Euclidean Distance to generate our initial list of HQI values. This was done primarily to minimize issues with baselines commonly due to scattering in IR spectra and amorphous content in XRPD spectra.

The clustering process is essentially unsupervised, however there are 2 parameters which can be modified and can have some effect on the size of the clusters. During the initial pass of the analysis, the software is instructed in how much of the data will be considered and the minimum size of a cluster – in most cases we have been 25% as the maximum size and 1 as the minimum size. Both of these numbers are only used during the first pass of the analysis. The second pass, the comparison of each member with average of each cluster, does not follow these constraints.

The maximum size constraint (expressed as a percentage of the entire dataset) is used primarily to prevent the software selecting some bad spectra near the very bottom of the sorted results as the largest gap.

Size of the dataset and time are two limitations to this approach. As the number of spectra in the dataset increases the time can go up dramatically. Currently we are limited to approximately 13,000 spectra – however this can take upwards of 2-3 days to generate results.

CONCLUSION

Any clustering process will generate results. It is important to have available some reviewing tools to compare clusters along with the ability to allow users to override or "supervise" the results. Once valid clusters have been identified, it is an easy process to compare new samples to existing clusters and obtain an initial quick analysis as to the class of materials it most closely resembles.

REFERENCES

1. Wikipedia - http://en.wikipedia.org/wiki/Cluster_analysis