University of Glasgow | School of Computing Science

Applications of Clustering Algorithms in the Analysis of Mass Spectrometry Data

Daniel Ramsay (2031365)

April 16, 2017

ABSTRACT

Mass spectrometry (MS) is an experimental technique in chemistry that is used to assist in the identification of chemical compounds by means of fragmentation. However due to the amount of noise in the data, current methods of analysis are very manual and since the process can produce vast amounts of data, this makes the detection of key features very time-consuming. In this paper we present the results of applying various clustering techniques from the machine learning sphere to this domain. The results that were gathered are promising, in both the quantity and quality of features that were detected, indicating strong potential for automation in a key stage of the MS data analysis pipeline.

1. INTRODUCTION

In mass spectrometry an unidentified chemical compound, is broken down into smaller charged fragments whose individual *m/z* (mass to charge ratio) and *intensity* (relative abundance) can then be recorded. These pairs of m/z and intensity values are referred to as MS peaks which can then be collectively plotted to produce a *mass spectrum*, see Figure 1. By analysing patterns of fragmentation in the mass spectrum it is hoped that key characteristics of the underlying chemical can be identified.

However, while current mass spectrometry methods are known to be extremely accurate in calculating the m/z values of chemical fragments, it can often be very difficult to determine whether or not two peaks with similar m/z values represent different underlying chemical features. It is currently an open question on how best to approach this, that is, to group MS peaks in a way which maximises the number of features identified without overfitting the data.

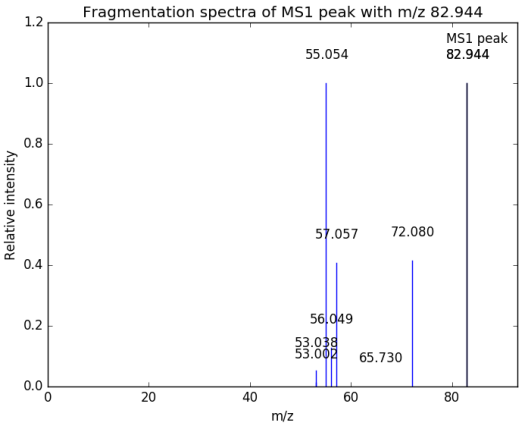
In this paper we apply algorithms based on k-means and Gaussian mixture models to detect clusters of peaks and then apply information criterion methods to determine the optimal number of clusters. We also consider the effectiveness of other clustering methods such as DBSCAN and OPTICS which take a density-based approach to identify the number of features directly. Once the final set of clusters are identified, we then assume that peaks belonging to the same cluster represent the same underlying chemical feature, thus reducing a larger set of real-valued peaks to a significantly smaller "vocabulary" of chemical fragments. By improving the process by which we formulate such a dictionary of features, we intend to supplement the work of [10] which describes a topic detection algorithm MS2LDA that maps the features identified here to chemical structures known as *Mass2Motifs*, which characterise the chemical behavior of the underlying compounds being fragmented.

2. BACKGROUND

Mass spectrometry

MS is a common analytical tool used in chemistry to derive information about a chemical compound's constituent substructures. There are many different implementations of MS, but they all revolve around the central premise of ionising chemical compounds and then sorting those compounds according to their mass to charge ratio.

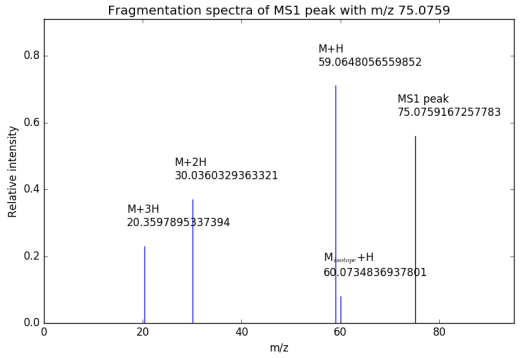
Figure 1: An example mass spectrum



The charge of an ion in MS can either be positive, which is achieved with the loss of electrons or gain of protons, or it can be negative which is usually achieved with the loss of protons. Furthermore when the charge is positive, this in general tends to be in the range +1, +2 or +3, charge is always a whole number and is unlikely to be any higher. When the charge is negative, this is also likely to be in the range in -1, -2 or -3 and similarly is unlikely to be any higher. The mappings of m/z values to intensity are commonly referred to as peaks and if we were to review a peak for a chemical compound, the peak would not correspond to the chemical directly but rather its corresponding ion. So in the case of an element M being ionised with an additional proton (which we denote as a hydrogen ion H) then the peak would refer to the ion $[M + H]^+$ with m/z value being the mass of M plus the mass of H. However in the instance where M is ionised by two hydrogen ions then the peak would refer to the ion $[M+2H]^{2+}$ and would have a m/z value equivalent to $(M + 2H)/2$, see Figure 2. This means that two fragments with different charges that represent the same underlying chemical feature are given very different m/z values whilst also adding to the issue of noise in the data. Furthermore

we have the added complication of isotopes [1], an example being an element such as carbon which predominately exists in the form carbon-12 but also exists less frequently in the form carbon-13. Such isotopic behavior results in what is known as *mass shift* which can mean two peaks which would otherwise represent the same underlying chemical are similarly represented with very different m/z values again raising the risk of misclassifcation and noise in the data set.

Figure 2: Example mass spectrum showing an element M ionised with one, two and three H ions along with a commonly occurring ionised isotope of M



Although current approaches used to perform mass spectrometry are known to be extremely accurate, they still exhibit some degree of random uncertainty in their measurements such that the same molecular fragment exhibited in two separate MS experiments' mass spectra cannot be expected to return the exact same peak in separate instances. Another complication is that two molecular fragments that are isobaric, that is chemicals that have the same molecular weight, are very hard to distinguish. Similarly isomers, that is compounds with the same chemical formula but different structural formulae, suffer the same problem and are difficult to distinguish by m/z value alone.

The raw experimental data used by the algorithms investigated in this paper were produced using a tandem mass spectrometry approach which consists of two stages: MS1 and MS2. The MS1 phase involves ionising a chemical mixture's constituent chemical compounds from which we can derive their individual m/z and intensity values which we refer to as MS1 peaks. In the MS2 phase the different molecules are separated by liquid chromatography and at regular time intervals the ionised compounds are broken up into their constituent ionised fragments, however not all the molecules in the mixture will be able to be fragmented in this way, due to shortcomings of current technology. For these fragments we again are able to derive m/z values and intensity values which map to their respective MS2 peaks. It is the case that each of the fragments' MS2 peaks belong to a single MS1 peak and each MS1 peak will correspond to a constituent chemical compound of the mixture.

The conventional approaches taken to identify a chemical compound from its MS data were reliant on a trained expert comparing the resulting mass spectrum of the unknown compound to reference spectra of known chemicals. This process is incredibly time-consuming and is made difficult due to spectra exhibiting large amounts of loss and noise. Tools currently exist to automate this process, [17] [9], comparing the experimental spectrums to spectra held in public databases, such as [15]. However relatively poor coverage of reference spectra means this approach of directly mapping is limited unless the corresponding reference data exists in the database.

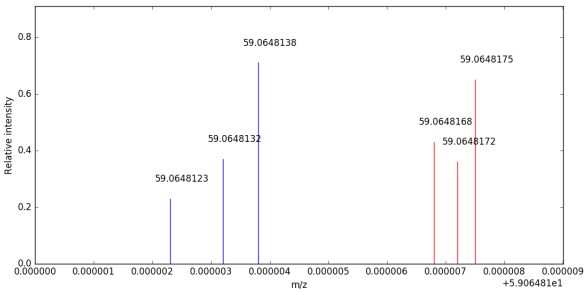
[1] Isotopes are atoms of the same element which have different numbers of neutrons but the same number of protons in the nucleus.

Latent Dirichlet allocation

A new approach MS2LDA [10], applies a technique known as *latent Dirichlet allocation* (LDA) [7], from text mining which models individual documents from a corpus as a finite mixture of a set of underlying topics. The process assumes that a set of words which make up the a document are chosen from a set of topics, where a topic is defined as a set of words which share a common theme.

The key utility of this process is that given a vocabulary of words and a set of known topics, then given a new document we should be able to derive the set of topics which generated that document. MS2LDA extends this analogy to mass spectra, with the set of MS2 peaks of a molecule representing a document and the MS2 peaks representing words. The topics in this case would be commonly occurring chemical structures which we refer to as Mass2Motifs. A Mass2Motif topic would represent a common chemical structure such as a functional group, with the "vocabulary" specific to that topic represented by the different ways that particular chemical structure might fragment in different instances.

Figure 3: A mass spectrum with two sets of peaks derived from two distinct fragments, here the clustering may be easy to spot but difficulties can arise as the number of peaks increases



However this approach is hindered by the presence of noise in the MS data. Using the LDA approach requires a fixed vocabulary as one of the parameters, but due to the nature of the experimental procedure it is rare for two instances of the same underlying chemical fragment to be represented by the exact same m/z value, with the difference being in the parts per million, see Fig 3. Hence we are not able to use the real-valued MS2 peaks as a vocabulary directly since the dictionary size would increase linearly with the number of peaks in the data.

It is therefore the objective of this paper to investigate the potential of clustering algorithms in grouping together sets of MS2 peaks so that they can be mapped to a fixed vocabulary of underlying chemical fragments which can then in turn be used by topic modelling algorithms, such as MS2LDA, to identify Mass2Motifs of unknown chemical compounds.

3. CLUSTERING METHODS

There is a variety of clustering algorithms available and the key focus of this investigation has been identifying which approaches are most effective in extracting features from our MS data. In this section we outline some of the technical characteristics of various clustering methods.

Iterative clustering algorithms

K-means is one of the oldest and most commonly used clustering algorithms, [13], [12]. The algorithm itself is iterative and simple to implement and requires a predetermined number of clusters k to be searched for in a set of data points. However it is known to be NP-Hard [2], but there are a number of heuristic algorithms that exist which are able to reduce computational time considerably such as *K-means++* described in [3].

Another approach is to use expectation maximisation in the context of a Gaussian mixture model (GMM) [8]. This can be considered a more generalised algorithm, as opposed to k-means which assumes all distributions are spherical. The GMM approach is iterative with each iteration having two stages, first the expectation-step where each object is assigned to a centroid and then the maximisation-step where a point is assigned to the cluster with the highest likelihood.

Choosing a value of k for k-means or GMM which best represents a data set can be subjective. The number of clusters is dependent on several factors such as shape and distribution of data points. A factor which makes finding an optimal k difficult is that performance functions based on the distance between all points and their respective allocated centroids will always favour increasing the number of clusters since the distance between a data point and its assigned cluster centroid will always decrease when the number of clusters increases. Therefore in order to find the optimal k we must find the point of equilibrium which is able to achieve the least distance cost between points and centroids using the least number of clusters.

One approach is to use an information criterion, with one such example being the Akaike information criterion (AIC) [1], which is a statistical approach for calculating the quality of a set of clusters to the data. The process works by assigning a penalty to each new cluster, for this we need a general objective function that has the parameters of distortion, which can be defined as how much a point deviates from its prototype cluster, and a measure of model complexity, which in this instance would be the number of clusters. Another IC approach is that of the Bayesian information criterion (BIC) as described in [18], which makes use of a likelihood function, with the number of clusters chosen being based on which number of clusters returns the lowest BIC.

Another algorithm x-means [16] combines both the k-means algorithm and BIC into a single algorithm whilst taking a slightly different approach. Rather than simply varying k over a wide range of values which can be computationally expensive, x-means performs an initial k-means sweep for clusters and then identifies which candidate clusters would be best suited for refinement into subclusters using BIC, it then performs this operation recursively until no more viable divisions can be identified.

Density-based clustering (DBC) algorithms

DBC algorithms are another approach which can be used to directly specify the number of clusters. one such example being *density-based spatial clustering of applications with noise* (DBSCAN), [14]. This algorithm groups data points into clusters that are in close proximity to each other, it does this by designating points as being either core, reachable or outlier points. Whether or not a point is designated a core point is based on whether a predetermined minimum number of points lie within an epsilon of that point, points are considered reachable if they are directly in the neighbourhood of a core point or a point that is itself reachable, while outliers are not in range of either reachable or core points. The benefits of this approach is that clusters can be arbitrary shapes and noise from outliers can be ignored. However the choice of epsilon or minimum number of points can have a large impact on the final choice of clusters [6] such that domain specific knowledge is often required for configuration. Another DBC algorithm, "Ordering points to identify the clustering structure" (OPTICS) [11] takes a similar approach to DBSCAN but has the benefit that its clusters can vary in density allowing it to be more flexibility than DBSCAN in certain instances. However OPTICS can also be highly sensitive to its choice of initial parameters so requires careful configuration.

4. APPROACH

Methodology

The clustering algorithms used to extract features from our MS data included:

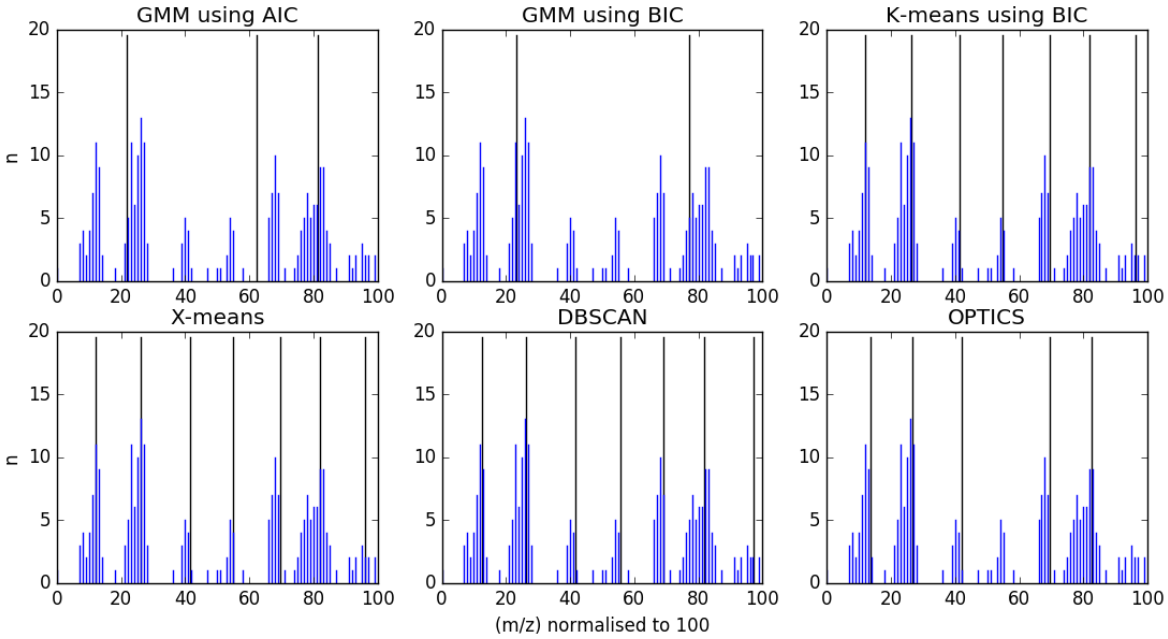
1. GMM using AIC
2. GMM using BIC
3. k-means using BIC
4. x-means
5. DBSCAN
6. OPTICS

Algorithms (1), (2) and (3) involved varying the number of clusters k and then choosing the number of clusters that returned the optimal IC value. Note that while (3) is not a probabilistic algorithm, it can still calculate the probablistic BIC indirectly by using the centroids as the mean values of the distributions.

The chemical solutions from which the MS data used in this report was derived came from 19 different beer samples. In stage one of the tandem MS experiment, the beer was refined to separate each of the constituent chemical compounds in the solution, these chemical compounds were then ionised to produce the MS1 peaks and then each compound was fragmented to produce the associated MS2 peaks for each MS1 peak. It is the case that no two MS2 peaks relating to the same MS1 peak will represent the same underlying chemical fragment, however this constraint does not apply for two MS2 peaks originating from the same original mixture but different MS1 peaks.

The MS2 peaks from all 19 beers were then combined into a single data set consisting of hundreds of thousands of peaks. Instead of directly clustering all the peaks of the data set directly, an initial filtering was carried out to reduce the set of peaks into smaller neighbourhoods of peaks. This is because peaks need only be distinguished from other

Figure 4: A histogram with 100 bins illustrating the distribution of an example grouping of MS2 peaks over a short-range of m/z values. The **black** lines indicate each algorithms choice of centres.



peaks within an immediate vicinity of approximately 20ppm relative to their m/z values. For instance two peaks with m/z values of 217.1213449 and 217.1234009 could potentially represent the same underlying fragment but a peak with an m/z value of 218.12753703 will have no possibility of being related to the first two values.

The process for filtering peaks into local groupings to then be clustered was as follows, after sorting all the peaks according to their m/z values, the distance between every peak and its subsequent peak was calculated and if the difference was below a certain epsilon then they were grouped together, otherwise they would be assumed to be unrelated and a new grouping would be formed. This process involved iterating through the list of peaks until all peaks belonged to a grouping consisting of one or more peaks.

Note that groupings consisting of a single peak were classified as a single feature. Also groupings consisting of multiple peaks whose range fell below neighbourhood of radius epsilon were also assumed to represent a single feature. Those groupings whose range exceeded a radius of epsilon were assumed to represent two or more chemical features and the clustering algorithms were subsequently applied. It should also be noted that while the value of epsilon used for filtering the peaks was varied, the value of the neighbourhood radius used in the DBC algorithms was fixed as 10% of epsilon for each instance and the minimum number of neighbouring points for a peak to be considered a core point was fixed at 5 peaks for all instances.

Once the cluster centroids were identified they were then allocated formula labels by an elemental formula assigner [5], which used a knapsack algorithm [4]. This allowed for different algorithms' predicted centroids, which had similar m/z values, to be compared directly.

Table 1: The values indicate the number of different features detected by the various clustering algorithms, in the data set consisting of all 19 beers, as the relative epsilon was varied over 3 different values.

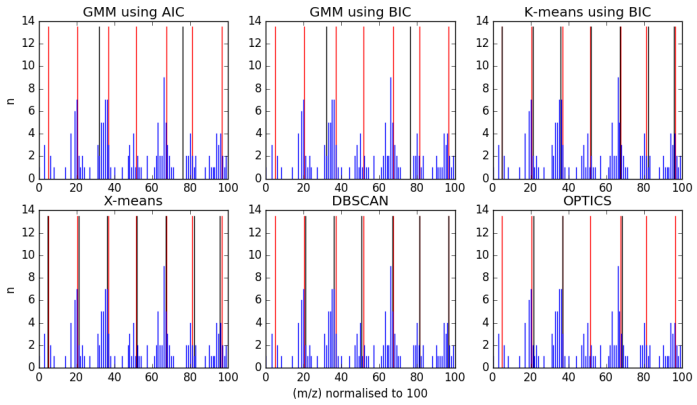
No. of features detected over varying epsilon			
Epsilon	20 ppm	10 ppm	5 ppm
GMM AIC	227	37	9
GMM BIC	183	31	9
k-means BIC	491	65	12
x-means	306	57	11
DBSCAN	337	54	12
OPTICS	249	45	0

Table 2: The percentages illustrate the proportion of overlapping features chosen by the same algorithm when applied to two different data sets, such that the first data set consisted of the peaks contained in the first 10 beer samples and the second data set consisted of the remaining 9 beer samples.

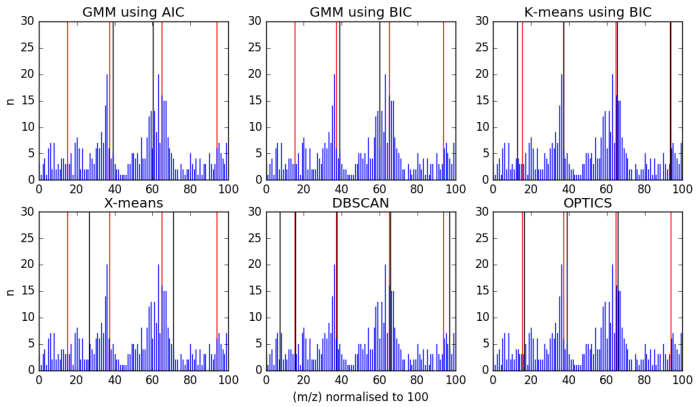
Proportion of overlapping features			
Epsilon	20 ppm	10 ppm	5 ppm
GMM AIC	36 %	47 %	69 %
GMM BIC	29 %	40 %	69 %
k-means BIC	78 %	83 %	85 %
x-means	49 %	53 %	85 %
DBSCAN	54 %	69 %	85 %
OPTICS	40 %	58 %	0 %

Figure 5: Each of the algorithms was applied to data sets of synthetic clusters. The true cluster centres used to generate the peaks are coloured in **red** while the predicted centres chosen by the algorithms are coloured in **black**.

(a) Synthetic set of 7 features



(b) Synthetic set of 4 features



(c) Synthetic set of 2 features

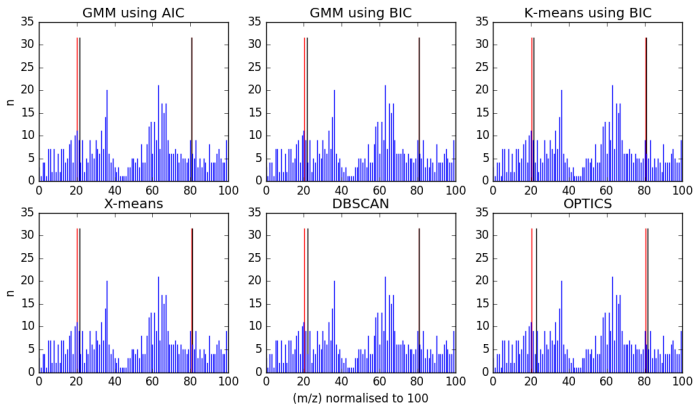


Table 3: Sets of synthetic clusters were generated using 7, 4 and 2 test nodes with known centres. Each algorithm was then applied and the number of predicted nodes was recorded.

Number of predicted features			
Test nodes	7	4	2
GMM AIC	3	2	2
GMM BIC	2	2	2
k-means BIC	7	4	2
x-means	7	2	2
DBSCAN	6	6	2
OPTICS	3	4	2

Table 4: Sets of synthetic clusters were generated using 7, 4 and 2 test nodes with known centres. Each algorithm was then applied and the average loss between each predicted node and its closest test node was recorded.

Avg. loss of predicted features (10^{-3})			
Test nodes	7	4	2
GMM AIC	0.9926	0.2669	0.08405
GMM BIC	0.9782	0.2669	0.08405
k-means BIC	0.1137	0.1137	0.09349
x-means	0.1137	0.814	0.09349
DBSCAN	0.1071	0.3396	0.09397
OPTICS	0.115	0.2529	0.0173

5. DISCUSSION

Analysis

The 6 clustering algorithms were compared against each other by 3 different approaches, these included:

1. comparing performance of the different algorithms against the same data set containing peaks from all 19 beers
2. comparing the overlap of features detected by the same algorithm on different data sets
3. comparing the average loss of the predicted centroids to the true mean value of various synthetically generated clusters

For deliberating on the effectiveness of the algorithms, we had the general preference for whichever approach detected the greatest number of features as well as the most consistent choice of features.

The choice of epsilon for deciding initial groupings is non-trivial, so results were generated for epsilon values of 5ppm, 10ppm and 20ppm in order to more comprehensively gauge the performance of the algorithms. It was the case that as epsilon became smaller, the groupings contained fewer peaks and more single-peak features were identified. Likewise as the value of epsilon became larger, clusters became more common and grew in complexity and fewer more general features were detected. The behavior of individual algorithms varied with the choice of epsilon such that while the number of features detected was broadly consistent for smaller epsilon, the number of features detected began to diverge as epsilon grew larger with GMM AIC, GMM BIC and OPTICS beginning to fall noticeably behind k-means BIC, x-means and DBSCAN in regard to number of features detected, see

Table 1, and in the consistency of feature intersection across different data sets, see Table 2.

Another area in which the algorithms' performance diverged was in the ability to detect the correct number of centroids in groupings with a high number underlying features. In general, the algorithms were quite consistent in predicting features when the number of underlying features was 2 or 3, and it is worth noting that groupings of this size made up the vast majority of filtered groupings. However the performance of the algorithms started to diverge for complex groupings of 4 or more underlying features with GMM AIC and GMM BIC struggling to detect more than 2 or 3 features in even the most complex of sets, while k-means BIC, x-means, DBSCAN and OPTICS facing fewer difficulties in detecting the same number of real features in the synthetic sets, see Table 3.

We also saw divergence in the performance between algorithms in respect to average loss, where we define loss as the distance of the predicted nodes to the nearest mean value used to generate a synthetic cluster. While the values for loss were rather consistent when there was 2 underlying nodes, when the number of underlying synthetic nodes increased, the more simplistic GMM AIC and GMM BIC algorithms began to lose any connection with the underlying nodes while the other 4 algorithms exhibited relatively conistent loss as the number of nodes increased, see Table 4.

None of algorithms displayed a universal advantage over the others, though k-means BIC, x-means and DBSCAN were broadly able to make similar choices on the number of features detected. However k-means BIC seems to have had a noticeable advantage for predicting the correct number of nodes whilst exhibiting the smallest overall loss between the predicted nodes and true underlying features. That being said, the average loss for GMM AIC and GMM BIC was the lowest among all the algorithms for groupings consisting of 2 underlying features implying that perhaps an approach of using different algorithms for groupings of different complexities may be best.

While further analysis could be carried out to deliberate on the performance of various algorithmic approaches, what is clear is that clustering algorithms do provide valuable insight into identifying features in mass spectrometry data. As can be seen from performance across both real MS data sets, see Figure 4, and synthetic data sets, see Figure 5, clustering algorithms can consistently reduce the noisy real-valued raw MS data to a smaller set of underlying features thereby considerably reducing the complexity of one stage of the MS data analysis pipeline.

Further work

There were a number of avenues for investigation to reduce noise in the MS data further that were not included in the scope of this report.

The main objective of this paper was to investigate the effectiveness of clustering techniques in reducing the number of features from a larger number of peaks to a smaller set of features which better represented the true underlying chemical fragments. Hence one opportunity to reduce noise in the data would have been to make corrections for potential isotopes present in the mass spectra. Transformations to account for mass shift could have been undertaken, such that peaks with high relative intensity could have had ranges of m/z values identified where isotopic representations of features could exist and if it was found that lower intensity peaks existed in those areas, those lower intensity peaks could have been associated with the more common high intensity isotopic representation.

However one should note that the possibilities for transformations would become extremely complex very quickly if we were to take into account every possible isotopic combination. Therefore one possible approach would be to limit our corrections to particularly common isotopes, such as carbon-13.

Similarly we also assumed that all the ionised fragments had a single charge. This is not always the case, with the specific rate of instance of fragments with non-singular charge varying between specific implementation of MS used and specific experimental instances. Hence in the same way we took into account mass shift for isotopes we could make corrections for the shift in the peaks' m/z values caused by the presence of multiple charges, thereby reducing the number of features to be annotated even further. However this approach would result in a large number of transformations needed to be considered for each outlier.

It should also be noted that while the clustering algorithms themselves were relatively computationally inexpensive to run, the knapsack algorithm which was used for annotating predicted centroids with formula labels proved to be a limiting factor in the size of data set that could be analysed, with the data used in this experiment, which consisted of several hundred thousand peaks, likely to have a running time of several hours on a conventional desktop computer. There are a number of ways this could have been approached such as caching formula ranges in a dictionary data structure or by restructuring the algorithms to compute in parallel.

6. CONCLUSION

In summary, we have strong evidence to suggest that clustering algorithms are effective in reducing noise and complexity in MS data by merging neighbouring peaks into more general and informative features.

Furthermore we have identified that k-means (with BIC), x-means and the DBC algorithms DBSCAN and OPTICS perform particularly well in clustering groupings of peaks which are likely to contain more than two nodes. In this investigation we were able to examine the effectiveness of the clustering algorithms across various parameters and while configuration of these values can have a substantial impact on the final choice of predicted nodes, noticeable reductions in complexity were achieved.

Hence we can strongly attest that clustering algorithms do provide impressive potential for automation in the detection of fragments in MS data to a level that is close to that of the true underlying features. It is therefore hoped, with further refinement, that the methods identified in this report could substantially improve the speed and quality of MS data analysis in the future.

Acknowledgements

I would like to express my sincere gratitude to my advisor Dr. Simon Rogers for his guidance and support during the course of this investigation.

References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory*, pages 267–281, 1973.

[2] D. Aloise, A. Deshpande, P. Hansen, and P. Popat. Np-hardness of euclidean sum-of-squares clustering. *Machine Learning*, 75:245–249, 2009.

[3] D. Arthur and S. Vassilvitskii. k-means++: the advantages of careful seeding. *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms.*, pages 1027–1035, 2007.

[4] S. Böcker, M. Letzel, Z. Lipták, and P. A. Sirius: Decomposing isotope patterns for metabolite identification. *Bioinformatics*, 25(2):218–224, 2009.

[5] S. Böcker and Z. Lipták. A fast and simple âĽăâĽăgorithm for the money changing problem. *Algorithmica*, 48(4):431–432, 2007.

[6] R. J. Campello, D. Moulavi, and J. Sander. Density-based clustering based on hierarchical density estimates. *Proceedings of the 17th Pacific-Asia Conference on Knowledge Discovery in Databases*, pages 226–231, 2013.

[7] M. J. D. Blei, A. Ng. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[8] Y. Guoshen. Solving inverse problems with piecewise linear estimators: From gaussian mixture models to structured sparsity. *IEEE Transactions on Image Processing*, 21(5):2481–2499, 2012.

[9] F. Hufsky, K. Scheubert, and B. S. Computational mass spectrometry for smallmolecule fragmentation. *Trends in Analytical Chemistry*, 53:41–48, 2014.

[10] J. Johan, J. Hooft, J. Wandy, M. P. Barrett, K. E. V. Burgess, and S. Rogers. Topic modeling for untargeted substructure exploration in metabolomics. *Proceedings of the National Academy of Sciences*, 113(48):13738–13743, 2016.

[11] H.-P. Kriegel, P. Kröger, J. Sander, and A. Zimek. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3), 2011.

[12] S. P. Lloyd. Least square quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137, 1957.

[13] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1:281–297, 1967.

[14] E. Martin, H.-P. Kriegel, J. Sander, and X. Xiaowei. 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*, pages 226–231, 1996.

[15] MassBank. High quality mass spectral database. `http://www.massbank.jp`, 2016. Accessed: 2016-12-01.

[16] D. Pelleg and A. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. *Proceedings of the 17th International Conf. on Machine Learning*, pages 727–734, 2000.

[17] L. Ridder. Automatic chemical structure annotation of an $lc - msn$ based metabolic profile from green tea. *Analytical Chemistry*, 85(12):6033–6044, 2013.

[18] G. E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):10724–10731, 1978.