Prediction of Molecular Substructures from Mass Spectrograms

Pieter-Jan Drouillon, Hendrik Blockeel

Dept. of Computer Science, KULeuven, Belgium
{Pieter-Jan.Drouillon,Hendrik.Blockeel}@cs.kuleuven.be

Abstract

This paper describes work in progress. We propose a possible approach to predicting the structure of a molecule from its mass spectrogram. The main idea is to cluster molecules based on their mass spectrogram. On these clusters one can perform a frequent subgraph mining algorithm to find the most frequent substructures in these molecules. Substructures that are much more frequent in one cluster than in others are likely to have an important influence on the mass spectrogram. Once these substructures have been identified, they can be used in a second clustering step to improve the clustering, after which a new search for frequent substructures in the new clusters can be performed. This can be repeated until the process stabilizes, which should lead to clusters that are coherent with respect to mass spectrograms as well as those molecular substructures most related to them. We discuss two variants of the approach, both of which remain to be validated empirically.

1 Introduction

In the field of data mining, learning predictive models is a common task. Based on input values, a predictive model delivers a set of output values. These input values are a mixture of different types such as numerical, categorical or structured values. They are entered in the model which computes the outcome. The output is typically a single value or a vector of values.

Until now, the most common output values are numerical or categorical values. In the past, some attempts were made to predict structured values [7]. The focus of our research is to investigate what the needs and possible restrictions are to include structured values as output.

Among many applications of this type of prediction, the task of deriving the structure of a molecule solely based on its mass spectrogram is an example. This procedure is often done to identify an unknown compound. In mass spectroscopy, molecules of a compound are bombarded with electrons. Some break up to give a variety of charged fragments, characteristic of the original molecule. A mass spectrogram is basically a histogram of the mass-to-charge ratio of the different fragments versus the frequency. Thus the input for a single example is a set of (x, y) couples with x the mass-to-charge ratio and y the frequency, the output to predict is the structure of the original molecule.

In this abstract a possible way of using these mass spectra to predict the structure of a molecule is proposed. It is the outline of ongoing and future research.

2 Data set

A data set of 5031 molecules was compiled from [5]. For each molecule, the name, molecular formula, weight, mass spectogram and the structure are stored in a database.

A natural way of representing a molecule is a graph with the vertices representing the atoms and the edges the bond between two atoms. Such a graph can be encoded in a string of characters, for instance SMILES[4].

SMILES is an abbreviation for simplified molecular input line entry specification. This specification uses ASCII strings to represent unambiguously the molecule's structure.

3 Preliminary experiments

Preliminary experiments on clustering were conducted on the mass spectra of 50 molecules[3]. Conclusions from these experiments were that it is possible to obtain clusters of molecules with similar structures. For instance, some clusters could be labeled as molecules containing a chain of carbon atoms with at the end a cycle.

The validation of these clusters, however, was somewhat subjective. The labeling of the clusters was done by hand, thus influenced by the prior knowledge and experience of the validator. While this result showed that molecules with similar mass spectrograms are relatively similar to the human expert, it is not obvious how similar they are in terms of an "objective" similarity measure on molecular structures. A next step will therefore be to define such a similarity measure on molecular structures (using for instance standard measures for graphs) and see how coherent the resulting clusters are with respect to this metric.

It is not unlikely that similarity according to a human expert will turn out to differ significantly from similarity according to any standard similarity measure for graphs, since human experts may assign more value to certain kinds of substructures that they know are relevant. It may not be obvious to turn this expert knowledge into a similarity measure.

For this reason, it seems a good idea to use a kind of clustering process where the similarity measures is learned together with the clustering itself. The next section provides a tentative method for that.

4 Method

4.1 Iterative process

We propose the following iterative clustering process:

1. Cluster the molecules based on the mass spectra;
2. Mine all the clusters from the previous step separately for frequent substructures;
3. Use the frequent substructures to form structural constraints that the clustering algorithm can use, or to extend the data representation (for details, see further)
4. Repeat steps 2 and 3 until no more new frequent substructures are found, and thus no more new constraints will be formed.

Thus, the mass spectrograms are used to "bootstrap" the clustering process. Clusters are mined for frequent substructures; ideally we want to find substructures as large as possible that occur in (almost) all of the molecules in one cluster. When we have found a substructure that corresponds almost perfectly with one cluster, we can use it to improve the clustering process. We see two different ways in which this can be done: the substructure can be used to formulate constraints on the clustering, or it can be used to extend the data representation so that clusters are found that are more coherent with respect to the occurrence of these substructures. We next detail the mining for frequent substructures, the formulation of constraints and the extension of data.

4.2 Mining frequent substructures

As the structure of a molecule is representable as a graph, frequent molecular fragments can be detected using a frequent subgraph mining algorithm [6]. Similar to the frequent itemset mining algorithm *a priori* [1], possible frequent subgraphs of size $k+1$ are constructed from already found frequent subgraphs of size k.

An other approach is to use the SMILES strings directly. In [2], a method for mining frequent fragments of chemical components is proposed where the molecule's structure is represented by a string of characters. Constructing a variant to use SMILES representation is also an option.

4.3 Constraint representation

Frequent substructures mined from molecules in a cluster A should ideally be found only in that particular cluster, and should be omnipresent there. If a molecule M in a different cluster contains the same substructure, we can formulate a 'must-link' constraint, stating that M should in fact be clustered together with the molecules in A. We could add must-link constraints with each element of A, or, perhaps more economically, with the element from A that is most dissimilar to M with respect to the similarity measure according to which the clustering was formed.

In a similar vein, when a molecule M' in cluster A does not contain the (almost) omnipresent substructure from A, that is an argument to include a cannot-link constraint between this molecule and the other molecules in A, or at least with the element in A most similar to M'.

4.4 Extending the data with new features

A second option, besides the structural constraints mentioned above, is to add the relevant frequent substructures to the mass spectrogram as binary features. This way the substructures influence the clustering process in a "softer" way than when used for constraints.

4.5 Use of the clustering for prediction

Once clusters have been formed that are coherent with respect to mass spectra as well as (the relevant parts of the) molecular structure, we can use such clusters for predicting (part of) the structure of a molecule as follows: the molecule is assigned to the cluster where its mass spectrogram fits best; then the substructures that frequently occur in this cluster are predicted to be part of the molecular structure. This process works best if the clusters are indeed coherent with respect to the mass spectrograms as well as the molecular structure.

5 Conclusion

We have presented work in progress that aims at predicting the molecular structure of molecules from their mass spectrogram. The method we propose might be useful more generally for predicting structured outputs. It currently remains to be validated empirically.

Acknowledgements

Hendrik Blockeel is a postdoctoral fellow of the Fund for Scientific Research of Flanders (FWO). The authors thank the National Institute of Advanced Industrial Science and Technology[5] for providing the mass spectrograms of the molecules.

References

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, 1993.

[2] L. De Raedt and S. Kramer. The levelwise version space algorithm and its application to molecular fragment finding. In *IJCAI*, pages 853–862, 2001.

[3] P-J. Drouillon. Clustering of mass spectrograms. unpublished. 2007.

[4] The SMILES homepage. `http://www.daylight.com/smiles/`.

[5] SDBSWeb: national Institute of Advanced Industrial Science and Technology. `http://www.aist.go.jp/RIODB/SDBS/`.

[6] S. Nijssen and J.N. Kok. Frequent graph mining and its application to molecular databases. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC 2004)*, 2004.

[7] J. Ramon and L. De Raedt. Instance based function learning. *Lecture Notes in Computer Science*, 1634:268–279, 1999.