Prediction of Molecular Substructures from Mass Spectrograms Using Constraint Based Clustering

Pieter-Jan Drouillon PIETER-JAN.DROUILLON@CS.KULEUVEN.BE
Dept. of Computer Science, Katholieke Universiteit, Leuven, Belgium

Hendrik Blockeel HENDRIK.BLOCKEEL@CS.KULEUVEN.BE
Dept. of Computer Science, Katholieke Universiteit Leuven, Belgium
Leiden Institute of Advanced Computer Science, Leiden, The Netherlands

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 the result of a first preliminary experiment.

1. Introduction

Learning predictive models is a very common data mining task. In most cases, the predictive models that are learned take as input a vector; sometimes the input is a set, graph, or other object with complex structure, in which case we speak of relational learning. Very few learning methods, however, have the ability to learn predictive functions where the *output* of the function has a complex structure. This is sometimes called structured output prediction, or prediction in structured output spaces. Many methods that do consider this setting, are still limited to predicting a structured output value where the structure is in fact known in advance, and only values for given elements of the structure need to be predicted (e.g., the labels of nodes in a tree or graph). Predicting the structure itself is an even harder problem. We use the term "structure prediction" to refer to this particular problem.

An example of the problem is predicting the structure of a molecule solely based on its mass spectrogram. 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, viewed as a predictive learning task, the input for a single example is a set of (x, y) couples with x the mass-to-charge ratio and y the frequency, and the output to predict is the structure of the original molecule.

In earlier work (Drouillon & Blockeel, 2007) we suggested two possible approaches towards solving this problem. In this abstract we describe in more detail one of these approaches and report on preliminary experiments with it.

2. Method

A very general approach to prediction, which may be useful for structure prediction, is *predictive clustering* (Blockeel, 1998). The basic idea behind predictive clustering is the following: given an input space X and output space Y, clusters are formed with high predictiveness of X and high predictability of Y; that is, given the projection of a new instance on X, we

Appearing in the 6^{th} *International Workshop on Mining and Learning with Graphs*, Helsinki, Finland, 2008.

can accurately predict the cluster it belongs to, and given the cluster an example belongs to, we can accurately predict Y. Put differently, while clustering generally tries to maximize inter-cluster-distances and minimize intra-cluster-distances, in predictive clustering, we try to maximize inter-cluster-distances in the X space (clusters are well-separated in the X space) and minimize intra-cluster-distances in the Y space (clusters are highly coherent in the Y space).

In the context of molecular substructure prediction, X is the space of mass spectrograms, Y the space of molecular structures (represented as graphs). We use the following iterative 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 constraints that the clustering algorithm can use;
4. Repeat steps 1, 2 and 3 until no more new frequent substructures are found, and thus no more new constraints will be formed.

In the first step clusters are formed using a hierarchical clustering algorithm. This algorithm uses the distances between each pair of mass spectrogram as input. In (Ramon & Bruynooghe, 2001) the matching distance is proposed. This metric measures the distance between two sets of points. Since a mass spectrogram is a set of (x, y) points, the distance between two mass spectrograms is computed using this metric.

In the next step each cluster is mined for frequent substructures. More specifically, we look for substructures that occur in many molecules of this cluster, and in few molecules of the other clusters.

Since we want the clusters to be predictive with respect to substructures, we will try to improve the clusters as follows: if a substructure occurs very often in one cluster and rarely in other clusters, we will push the clustering process towards finding a clustering where all molecules containing this substructure are in the same cluster. We do this by imposing must-link and cannot-link constraints on the clustering process.

In the following subsections we elaborate on how the found substructures are used to generate the constraints.

2.1. Constraint representation

Frequent substructures mined from molecules from the same cluster C should ideally be found only in that particular cluster. If a molecule m in a different cluster contains the same molecular substructure, we can formulate a 'must-link' constraint. This constraint states that this molecule m should be clustered with the molecules from cluster C also containing this particular substructure. As a consequence, must-link constraints between m and each molecule of C containing the substructure can be added.

In a similar vein, if a substructure is very frequent in cluster C and not frequent in other clusters, a molecule m' in cluster C that does not contain this substructure should not be clustered with the other molecules from C containing the substructure. A cannot-link constraint between this molecule m' and cluster C can be added. This means that the molecule m' cannot be clustered with any other molecule in C that does contain this frequent substructure.

2.2. Constraint generation

Constraints as described in section 2.1 can only be generated if the following condition holds:

$$\forall C_k, k \neq j : \text{freq}_{C_k}(SS_{ij}) \ll \text{freq}_{C_j}(SS_{ij}) \quad (1)$$

This condition states that a frequent substructure SS_{ij} from cluster C_j should be omnipresent in cluster C_j and not frequent at all in all other clusters C_k. If this condition holds for both absolute and relative frequency, then substructure SS_{ij} is a valuable candidate to generate the must-link and the cannot-link constraints.

2.3. 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.

First the molecule is assigned to the cluster where its mass spectrogram fits best. Next 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.

3. Experiments

3.1. Data set

A data set of 5031 molecules was compiled from (SDBS-Web, 2007). For each molecule, the name, molecular formula, weight, mass spectrogram and the

structure are stored in a database. Molecular structures are stored in the SMILES format (Anderson et al., 1987), which uses ASCII strings to represent unambiguously the molecule's structure.

3.2. Experimental settings

In this preliminary experiment, a subset of 50 molecules was randomly selected from the database. The program was implemented in Java. The frequent subgraph mining algorithm is part of ParMol (Meinl et al., 2006), a Java library containing several graph mining algorithms and a parser to convert SMILES into graph representation. The graph mining algorithm used in this experiment is gSpan (Yan & Han, 2002) .

Each cluster was mined for substructures occuring in at least 50 % of the molecules in this cluster. Next the experiment was repeated with a support of at least 80 %. For each of these frequent substructures, the following condition was used to determine if this substructure was a good candidate to generate constraints:

$$\sum_{k \neq j} \text{freq}_{C_k}(SS_{ij}) < \text{freq}_{C_j}(SS_{ij}) \qquad (2)$$

If the absolute frequency of SS_{ij} in cluster C_j is higher than the sum of the absolute frequencies in all other clusters, then this substructure is used to formulate must-link and cannot-link constraints.

3.3. Results

After five (50% support) and seven (80% support) iterations no more new constraints were generated and the process converged to a solution. Both solutions contained seven clusters. Since most molecules of each cluster contained similar substructures, each cluster could be labeled, see table 1 . The resulting cluster solutions are similar in terms of labeling.This labeling was not possible when the molecules were clustered solely based on the mass spectrograms.

These clustering solution were then used to predict substructures of a test set of unseen molecules. An unseen molecule m is assigned to the cluster of m', the molecule whose mass spectrogram is closest to the mass spectrogram of m. The frequent substructures of the assigned cluster are then predicted as substructures of m.

In table 2, the number of found frequent substructures and the number of assigned molecules are listed for all clusters. For each frequent substructure SS_{ij} in cluster C_j, the molecules that are assigned to C_j and contain SS_{ij} are counted. These counts are then averaged over all assigned molecules (last column of the table).

Table 1. Cluster solution with constraints

Cluster	Molecules
0	1 small molecule
1	Small molecules with at least 1 double bond
2	Benzene ring
3	Long carbon chain
4	Medium alcohol
5	Small ring or aldehyde
6	Benzene ring with small chain

Table 2. Summary of prediction

Cluster	freq. sub-structures	assigned molecules	avg sub-structures
50 % support			
0	1	0	0
1	3	5	2
2	2	11	1.63
3	7	4	1.25
4	3	0	0
5	2	4	1.75
6	3	26	2.15
80 % support			
0	1	0	0
1	1	5	0.6
2	1	7	0.85
3	2	26	1.5
4	3	0	0
5	2	7	1.71
6	1	5	1

In the experiment with 50% support, the average number of frequent substructures of the assigned molecules are rather high with exception of cluster 3. In the second experiment (80% support), the average number of frequent substructures is still quite high for all clusters but the number of frequent substructures decreased.

4. Conclusion

We have presented work in progress that aims at predicting molecular substructures of molecules based on their mass spectrograms. We use a variant of predictive clustering that relies on frequent substructure discovery and constraint based clustering.

Future work includes (1) the use of alternative substructure miners, preferably systems that find structures that are frequent in one cluster and infrequent in other clusters, rather than just finding frequent struc-

tures (and filtering those in a second step); (2) more experiments to investigate how parameters (eg. the minimal frequency of the mined substructures in each cluster) influence the process, and (3) a more quantitative evaluation of the cluster solution, more specifically, measuring how accurately it can predict the molecular structure of a new instance from its mass spectrogram (what percentage of the molecule's structure is predicted on average, and how many predicted substructures are effectively part of the molecule).

Acknowledgements

Hendrik Blockeel is a postdoctoral fellow of the Research Foundation of Flanders (FWO). The authors thank the National Institute of Advanced Industrial Science and Technology (SDBS-Web, 2007) for providing the mass spectrograms of the molecules.

References

Anderson, E., Veith, G., & Weininger, D. (1987). *Smiles: A line notation and computerized interpreter for chemical structures* (Technical Report EPA/600/M-87/021). U.S. EPA, Environmental Research Laboratory-Duluth.

Blockeel, H. (1998). *Top-down induction of first order logical decision trees*. Doctoral dissertation, Department of Computer Science, Katholieke Universiteit Leuven. `http://www.cs.kuleuven.ac.be/~ml/PS/blockeel98:phd.ps.gz`.

Drouillon, P.-J., & Blockeel, H. (2007). Prediction of molecular substructures from mass spectograms. *Proceedings of the 5th International Workshop on Mining and Learning with Graphs* (pp. 171–174).

Meinl, T., Worlein, M., Urzova, O., Fischer, I., & Philippsen, M. (2006). The parmol package for frequent subgraph mining. *Proceedings of the Third International Workshop on Graph Based Tools*.

Ramon, J., & Bruynooghe, M. (2001). A polynomial time computable metric between point sets. *Acta Informatica*, *37*, 765–780. `http://www.cs.kuleuven.ac.be/cgi-bin-dtai/publ\_info.pl?id=34841`.

SDBS-Web (2007). National Institute of Advanced Industrial Science and Technology. `http://riodb01.ibase.aist.go.jp/sdbs/`.

Yan, X., & Han, J. (2002). gspan: Graph-based substructure pattern mining. *Proceedings of the 2002 IEEE International Conference on Data Mining* (pp. 721–724). Japan: IEEE Computer Society.