ILP, the Blind, and the Elephant: Euclidean Embedding of Co-Proven Queries

Hannes Schulz[1] and Kristian Kersting[2] and Andreas Karwath[1]

[1] Institut für Informatik, Albert-Ludwigs Universität
Georges-Köhler-Allee 79, 79110 Freiburg, Germany
{schulzha,karwath}@informatik.uni-freiburg.de
[2] Dept. of Knowledge Discovery, Fraunhofer IAIS
Schloss Birlinghoven, 53754 St Augustin, Germany
kristian.kersting@iais.fraunhofer.de

Abstract. Relational data is complex. This complexity makes one of the basic steps of ILP difficult: understanding the data and results. If the user cannot easily understand it, he draws incomplete conclusions. The situation is very much as in the parable of the blind men and the elephant that appears in many cultures. In this tale the blind work independently and with quite different pieces of information, thereby drawing very different conclusions about the nature of the beast. In contrast, visual representations make it easy to shift from one perspective to another while exploring and analyzing data. This paper describes a method for embedding interpretations and queries into a single, common Euclidean space based on their co-proven statistics. We demonstrate our method on real-world datasets showing that ILP results can indeed be captured at a glance.

1 Introduction

Once upon a time, there lived six blind men in a village. One day the villagers told them, "Hey, there is an elephant in the village today." They had no idea what an elephant is. They decided, "Even though we would not be able to see it, let us go and feel it anyway." All of them went where the elephant was and touched the elephant. Each man encountered a different aspect of the elephant and drew a different inference as to its essential nature. One walked into its side, concluding that an elephant is like a wall. Another, prodded by the tusk, declared that an elephant is like a spear. The chap hanging onto the tail was convinced that he had found a sort of rope. The essential nature of the elephant remained undiscovered.

The tale is that of "The Blind and the Elephant", which appears in many cultures. It illustrates the problem many ILP users face, to make sense of relational data and models, the elephants, before applying their algorithms or while interpreting the results. Due to the complexity of the data and the models, the user can only touch small parts of them, like specific queries. Hence, he often gets only a narrow and fragmented understanding of their meaning.

Presented at the 19th International Conference on Inductive Logic Programming (ILP'09).

In contrast, visual representations make it easy to shift from one perspective to another while exploring and analyzing data. How to visually explore relational data and queries jointly, however, has not received a lot of attention within the ILP community. This can be explained by the fact that relational data involves objects of several very different types without a natural measure of similarity.

Our paper addresses this problem by creating embeddings from statistical associations. Observing that interpretations can be seen as documents, atoms as words appearing in the documents, and queries as topics, we compute their empirical co-occurrence statistics and find Euclidean embeddings of these items representing their statistics[3].

2 Euclidean Embedding of Co-proven Queries

Computing joint embeddings of interpretations K and queries Q essentially requires three steps: (1) collecting embeddable queries, (2) embedding queries and interpretations into a single Euclidean space, and – as an optional postprocessing step – (3) labelling the representation by extracting local representatives.

Step 1 – Queries: Given a finite set K of observed interpretations, any ILP algorithm can be used to preselect embeddable queries Q for K. In this paper, we use *Molfea* [5] and *C-armr*[4] [2] to mine databases of molecules. Both systems are inspired by the Agrawal's *Apriori* algorithm [1]: they construct general queries and specialize them only if they are frequent. Only queries more frequent than some threshold are retained and further expanded, i.e., specialized. While *Molfea* constructs linear fragments only (atom, bond, atom, ...), *C-armr* constructs general queries and can take background knowledge into account as well. In addition to the queries, we also store the interpretations in which they were true. This will prove useful for the next step and can efficiently be represented in a binary matrix $\mathcal{C} \in \{0,1\}^{|Q|\times|K|}$, where $|Q|$ is the number of queries and $|K|$ is the number of observed interpretations.

Step 2 – Embedding: We wish to model the statistical dependence between K and Q through mappings $\phi : K \mapsto \mathbb{R}^d$ and $\psi : Q \mapsto \mathbb{R}^d$ for a given dimensionality d. These mappings should reflect the dependence between K and Q such that the co-occurrence C_{qk} of some $k \in K$ and $q \in Q$ determines the distance between $\phi(k)$ and $\psi(q)$.

This is exactly what Globerson *et al.*'s *CODE* algorithm [4] does. *CODE* models the empirical joint distribution $\bar{p}(k,q) \propto \mathcal{C}_{qk}$ over K and Q as exponentials of Euclidean distances in a low-dimensional embedding space. The positions of the points are selected in a maximum-likelihood fashion. More precisely, we consider an asymmetric co-occurence model of interpretations and queries,

$$p_{QK}(k,q) \equiv \frac{1}{Z} \cdot \bar{p}(k) \cdot \exp(-\|\phi(k) - \psi(q)\|^2),$$

[3] Embedding algorithms often use Euclidean distances in some feature space as a measure of similarity. However, we deal with objects of different types having different representations, making this approach less appealing.

[4] in the CLASSIC'CL implementation [9]

where k is an interpretation and q a query. $Z = \sum_{k,q} \bar{p}(k) \exp(-\|\phi(k) - \psi(q)\|^2)$ is a normalization term to avoid the trivial solution. We multiply by the empirical marginal $\bar{p}(k)$ such that the number of queries which are true in an interpretation does not influence the embedding. Now, *CODE* initially assigns random positions to all queries and interpretations in ϕ and ψ. It then changes ϕ and ψ so that they maximize the likelihood $\sum_{k,q} \bar{p}(k,q) \log p_{QK}(k,q)$, see [4] for details.

For our problem at hand, however, the results are unsatisfactory: using p_{QK} only, *CODE* tends to map the interpretations to a circle. This can be explained by the fact that the marginal distribution of interpretations is almost uniform in our case. To generate more expressive embeddings, we use the interpretations and queries to generate a non-binary query-query co-occurence matrix $\mathcal{D} = \mathcal{C}\mathcal{C}^T$ and model $p_{QQ}(q,q') \propto \mathcal{D}_{qq'}$. This co-proven statistics of queries q and q' should be represented by distances in the embedding as

$$p_{QQ}\left(q,q'\right) \equiv \frac{1}{Z} \cdot \exp(-\|\psi(q) - \psi(q')\|^2) \; .$$

Again, we assign initial positions randomly but now adapt them so that they maximize the "log-likelihood" of the combined models

$$\sum\nolimits_{k,q} \bar{p}(k,q) \log p_{QK}(k,q) + |K|/|Q| \cdot \sum\nolimits_{q,q'} \bar{p}\left(q,q'\right) \log p_{QQ}\left(q,q'\right) \; .$$

Step 3 – Condensation: Literally thousands of queries and instances can be embedded into a single Euclidean space and can – as a whole – provide useful insights into the structure of the data. However, we would also like to get a grasp on what queries in certain regions focus on. To do so, we propose to single out queries q in an iterated fashion. Specifically, we assign to each query q the weight $\mathrm{w}(q) = F(q)/\operatorname{length}(q)$, where $F(q)$ is q's F_1 or F_2-measure, see e.g. [7], and $\operatorname{length}(q)$ is its description length. We now locally remove queries with a low weight in a two-step process. First, we build the k-nearest neighbour graph of the embedded queries. From the weight, we subtract the weight of its graph neighbours, thereby removing large-scale differences. Second, we increase weights of queries with lower weighted neighbours and decrease weights which have higher weighted neighbours. The last step is repeated until the number of queries q with a positive weight is not changing anymore. In other words, we prefer short queries with high F-measures on a local neighbourhood.

3 Showcases

We tested our approach on several real-world datasets for the two-dimensional case. To provide a qualitative assessment of our method, we apply it to datasets where some structures or models have already been discovered.

On **Mutgenesis** [8], the problem is to predict the mutagenicity of a set of compounds. In our experiments, we use the atom and bond structure information only (including the predefined predicate like *ball3s*, *ring_size_5s*, and others). The dataset consists of 230 compounds (138 positives, 92 negatives). The 2D Euclidean embedding is shown and discussed in Fig. 1.

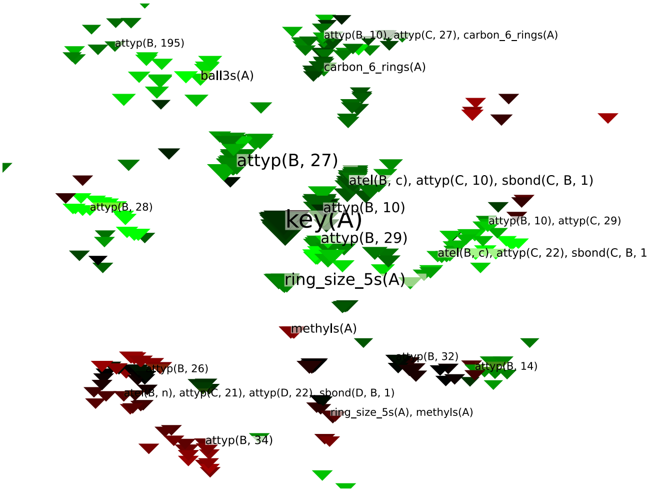

Fig. 1. Mutagenesis embedding. We show all frequent queries (triangles) distinct w.r.t. the interpretations. Black/colored queries have low/high precision, small/large queries have low/high recall. Red/green queries indicate negative/positive class. The queries with textual descriptions were automatically selected, the trivial `key` attribute was omitted in all but the central queries. The embedding reflects rules we could induce employing Srinivasan's ALEPH on the same dataset such as `active(A) :- attyp(A,B,29), ring_size_5s(A)` or `active(A) :- ball3s(A)`.

The **Estrogen** database was extracted from the EPA's DSSTox NCTRER Database[5]. The original dataset was published by Fang *et al.* [3], and is specially designed to evaluate QSAR approaches. The NCTRER database provides activity classifications for a total of 232 chemical compounds, which have been tested regarding their binding activities for the estrogen receptor. The database contains a diverse set of natural, synthetic, and environmental estrogens, and is considered to cover most known estrogenic classes spanning a wide range of biological activity [3]. Here, "activity" is an empirically measured value between 0 and 100, which we averaged and used as a query's color. The 2D Euclidean embedding is shown and discussed in Fig. 2.

The DTP **AIDS** Antiviral Screening Database originating from the NCI's development therapeutics program NCI/NIH[6] consists of SMILES representations of 41,768 chemical compounds [6]. Each data entry is classified as either active, moderately active, or inactive. A total of 417 compounds are classified as active, 1,069 as moderately active, and 40,282 as inactive. We have converted this dataset into SDF format using the OpenBabel toolkit and randomly sampled 400 active and 400 moderate/inactive compounds. The 2D Euclidean embedding is shown and discussed in Fig. 3.

[5] http://www.epa.gov/ncct/dsstox/sdf_nctrer.html

[6] http://dtp.nci.nih.gov/

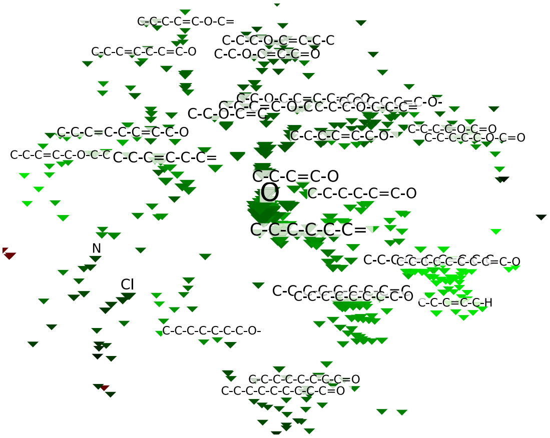

Fig. 2. Estrogen embedding. The coding is as in Fig. 1; only black/colored queries indicate now low/high activity. In their original publication Fang *et al.* have identified that a phenolic ring connected by one to three atoms to another benzene ring is one of the key features that have to be present regarding the likelihood of a compound being a ER ligand. A phenolic ring is a 6-carbon benzene ring with an attached hydroxyl (OH) group. In the embedding, it can be seen that this is reflected in features like `C-C-C=C-O`, which indicates that there is a path of one carbon atom to (a part of) a ring structure (`C-C=C`) connected to an oxygen.

4 Concluding Remarks

In our opinion, to unveil its full power, ILP must incorporate visual analysis methods. With the work presented here, we have made a step in this direction. We have presented the first method for embedding interpretations and queries into the same Euclidean space based on their co-occurrence statistics. As our experiments demonstrate, the spatial relationships in the resulting embedding are intuitive and can indeed reveal useful and important insights at a glance. Aside from their value for visual analysis, embeddings are also an important tool in unsupervised learning and as a preprocessing step for supervised learning algorithms. In future research, we will explore this direction.

Acknowledgments: The authors would like to thank Amir Globerson for providing the *CODE* code. The research was partially supported by the EU seventh framework program under contract no. Health-F5-2008-200787 (*OpenTox*), and by the Fraunhofer ATTRACT fellowship STREAM.

References

1. R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499. Morgan Kaufmann, San Francisco, CA, USA, 1994.

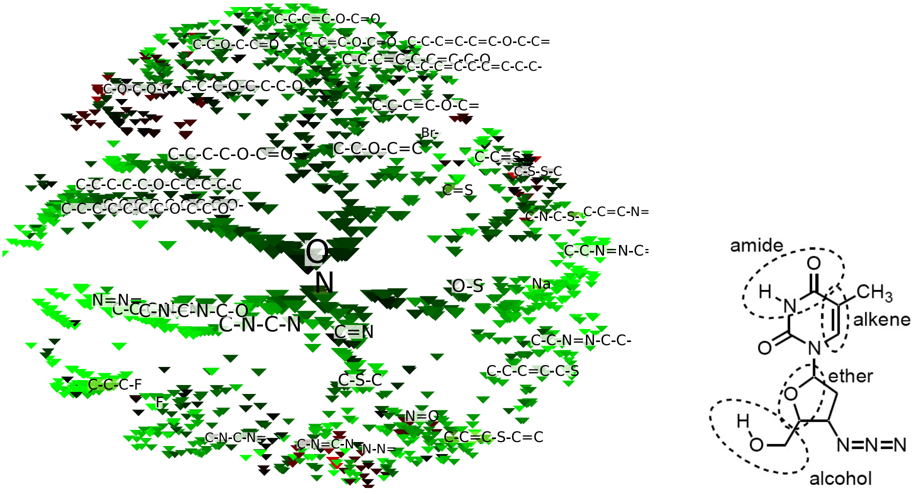

Fig. 3. AIDS embedding. The coding is as in Fig. 1. The embedding (left hand side) clearly indicates compounds that are derivatives of Azidothymidine (AZT, right hand side), a potent inhibitor of HIV-1 replication. AZT contains a number of functional groups like an amide group (`N-C=O`) or an ether group (`C-O-C`) as well as a nitrogen group (`N=N=N`), one of the prominent features of AZT.

2. L. De Raedt and J. Ramon. Condensed representations for inductive logic programming. In *Proceedings of 9th International Conference on the Principles of Knowledge Representation and Reasoning*, pages 438–446, 2004.
3. H. Fang, W. Tong, L.M. Shi, R. Blair, R. Perkins, W. Branham, B.S. Hass, Q. Xie, S.L. Dial, C.L. Moland, , and D.M. Sheehan. Structure-activity relationships for a large diverse set of natural, synthetic, and environmental estrogens. *Chem. Res. Tox*, 14:280–294, 2001.
4. A. Globerson, G. Chechik, F. Pereira, and N. Tishby. Euclidean Embedding of Co-occurrence Data. *The Journal of Machine Learning Research*, 8:2265–2295, 2007.
5. C. Helma, S. Kramer, and L. De Raedt. The molecular feature miner MolFea. In *Proceedings of the Beilstein-Institut Workshop*, 2002.
6. S. Kramer, L. De Raedt, and C. Helma. Molecular feature mining in HIV data. In Foster Provost and Ramakrishnan Srikant, editors, *Proc. KDD-01*, pages 136–143, New York, NY, USA, August 26-29 2001. ACM Press.
7. D.D. Lewis. Evaluating and optimizing autonomous text classification systems. In *Proceedings of the 18th Int. ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 246–254, 1995.
8. A. Srinivasan, S. H. Muggleton, R. D. King, and M. J. E. Sternberg. Theories for Mutagenicity: A Study of First-Order and Feature -based Induction. *Artificial Intelligence Journal*, 85:277–299, 1996.
9. C. Stolle, A. Karwath, and L. De Raedt. CLASSIC'CL: An Integrated ILP System. In A. G. Hoffmann, H. Motoda, and T. Scheffer, editors, *Discovery Science*, volume 3735 of *Lecture Notes in Computer Science*, pages 354–362. Springer, 2005.