Multiple-Instance Learning Based Heuristics for Mining Chemical Compound Structure

Cholwich Nattee,[†] Sukree Sinthupinyo,[†] Masayuki Numao[†] and Takashi Okada[††]

Inductive Logic Programming (ILP) is a combination of inductive learning and first-order logic aiming to learn first-order hypotheses from training examples. ILP has a serious bottleneck in an intractably enormous hypothesis search space. This makes existing approaches perform poorly on large-scale real-world datasets. In this research, we propose a technique to make the system handle an enormous search space efficiently by deriving qualitative information into search heuristics. Currently, heuristic functions used in ILP systems are based only on quantitative information, e.g. number of examples covered and length of candidates. We focus on a kind of data consisting of several parts. The approach aims to find hypotheses describing each class by using both individual and relational features of parts. The data can be found in denoting chemical compound structure for Structure-Activity Relationship studies (SAR). We apply the proposed method to extract rules describing chemical activity from their structures. The experiments are conducted on a real-world dataset. The results are compared to existing ILP methods using ten-fold cross validation.

1. Introduction

Inductive Logic Programming (ILP)[3] aims to learning first-order rules from examples and background knowledge. ILP combines inductive learning and first-order logic to overcome limitations of inductive learning which is based on propositional logic or attribute-value language. First-order logic representation provides capability to handle data which consist of complicated relations. Such as, data that is scattered over many tables with relations among tables. Though, propositionalization allows attribute-value learning system to handle this kind of data, it causes the number of features to become larger and difficult to be managed. Another advantage of ILP is comprehension. Learning results are given in form of first-order rules which are understandable by human. Nevertheless, a bottleneck of ILP is an intractably enormous search space caused from flexibility of first-order logic.

To reduce the search space size, two techniques are mainly used: language bias and informed search. Language bias aims to define description of learning results to limit possibility in candidate generation. Informed search uses heuristic function to cut unnecessary parts from searching process. In this research, we focus on using heuristic function to limit search space and lead to appropriate rules. Heuristic functions used in the existing ILP systems are based only on quantitative information, such as, the number of examples covered by the considered candidate or length of the candidate. This causes the existing approaches sometimes perform worse than attribute-value learners. To overcome the shortcoming, qualitative information is required, such as the quality of the covered examples should be considered.

We therefore propose an improved heuristic function based on Multiple-Instance Learning (MIL)[1]. MIL is an extended two-class propositional learning approach for data that cannot be labeled individually, albeit several instances of data are gathered and labeled as a bag. Each positive bag may consist of both positive and negative instances. Nevertheless, MIL aims to obtain models that predict instances not bags, thereby rendering itself similar to supervised learning where there are noises in positive examples. Algorithms from MIL solve the ambiguity by using similarity or distance among instances within feature space. Using distance, target concept is an area where several instances from various positive bags are located together and that area is far from instances from negative bags. We derive this basic idea

† The Institute of Scientific and Industrial Research, Osaka University
†† School of Science and Technology, Kwansei Gakuin University

of MIL to evaluate quality of first-order objects consisting of multiple parts. Each object is considered as a bag containing several parts. We evaluate each part using MIL based measure using similarity or distance among parts. Therefore, the part whose features are common compared to parts from various positive objects is evaluated as high value. We evaluate all parts and incorporate obtained values as weights into heuristic function. Then, hypothesis candidate covering high-valued parts is evaluated as high value and selected first.

The paper is organized as follows. In the next section, we present details of proposed method that improves heuristic function used in ILP to efficiently learn rules from objects consisting of multiple parts. We focus on classifying chemical compound according to their structures. The experiments conducted on real-world datasets are then presented in Section 3. Finally, we conclude the paper and consider future directions in Section 4.

2. Using ILP in Structure-Activity Relationship Studies

The studies of Structure-Activity Relationship aim to find structures in chemical compounds describing their characteristics or activities. Knowledge discovered will be useful for developing new drugs. In recent years, advance in High Throughput Screening (HTS) technology has produced vast amount of SAR data. Once the rules which predict the activities of existing SAR data are found, it significantly helps the screening process. Since each compound consists of multiple parts, we then gain benefits from the improved heuristics for a large-scale data.

The proposed approach incorporates existing top-down ILP system (FOIL)[4] and applies MIL based measure to find common features among parts of positive compounds. The measure is then used as the weight attached to each part of the example and the common parts among positive examples are attached with high-valued weights. With these weights and heuristic function based on example coverage, the system generates more precise and higher coverage hypotheses from training examples. Next, we explain first-order representation used in the paper. After that, the improved heuristic function is then be explained.

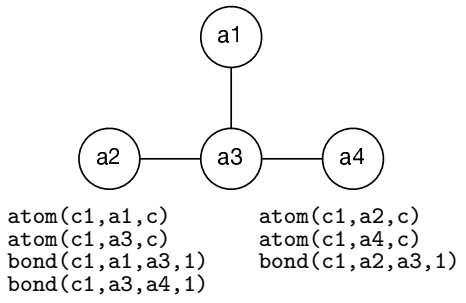

Fig. 1 Examples of first-order representation of chemical compound

2.1 First-Order Representation

To apply ILP for SAR studies, training examples are required to be denoted in form of the first-order logic. Each chemical compound is considered a first-order object. We denote them based on their structure using two predicates: `atom(Compound, Atom, Element)` and `bond(Compound, Atom1, Atom2, Type)`. Features related to atom and bond are put as parameters of predicate. Predicate `atom` denotes an `Atom` of `Element` in a `Compound`. Predicate `bond` denotes a bond of `Type` consisting of two atoms (`Atom1` and `Atom2`). Figure 1 shows an example of first-order representation. More features can be used to represent atoms and bonds in real-world dataset. In this research, we consider an atom as a part of compound, and a bond is a relation between two parts. In other words, a compound is a group of atoms relating to each other. An atom as a part is used in the improved heuristic function explained in the following section.

2.2 Improved Heuristics

The original heuristic function used in FOIL is based on information theory. Partially developing hypothesis is evaluated by using the number of positive and negative tuples covered.

$$I(T_i) = -\log_2 \frac{|T_i^+|}{|T_i^+| + |T_i^-|} \qquad (1)$$

Thus, FOIL selects the literal that covers many positive tuples but few negative tuples. To make heuristics select better literals, we derive Diverse Density (DD)[2], a measure for MIL data based on probabilistic model. DD is defined to be high in the area that instances from various positive bags locate together and that area is far from negative instance. From Equa-

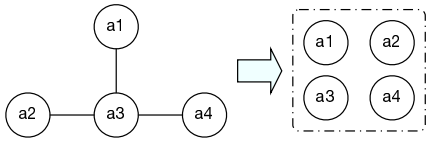

Fig. 2 A bag representation of compound

tion 1, T_i^+ and T_i^- denote the set of positive and negative tuples respectively. We consider each compound as a bag and each part of compound as an instance in the bag (Figure 2). DD of all parts is then computed and used as a weight attached to the parts. Therefore, a part with common features among parts from positive compound is given a high-valued weight. The weights are incorporated to the heuristic function by altering $|T_i^+|$ to be the sum of weight. If the heuristic is high, it means that the candidate covers many common parts among positive compounds. The heuristic function is modified as follows.

$$DD_s(T) = \sum_{T_i \in T} DD(T_i) \quad (2)$$

$$I(T_i) = -\log_2 \frac{DD_s(T_i^+)}{DD_s(T_i^+) + |T_i^-|} \quad (3)$$

Nevertheless, we still use the number of negative tuples $|T_i^-|$ in the same way as the original heuristics, since we know that all parts of negative examples show the same strength. Therefore, it is similar to weighing all negative parts with value 1. From the improve heuristics, we implement a prototype system called FOILMP.

3. Experiments and Discussions

3.1 Datasets

We aim to discover rules describing the activities of dopamine antagonist. Dopamine antagonist dataset contains 1,366 compounds separated into four classes; D1, D2, D3 and D4. They are obtained from MDDR database of MDL Inc. After preprocessing, three kinds of predicates are used to denote a compound as shown in Figure 3.

3.2 Comparing to existing ILP approaches

We conduct ten-fold cross validation to predict D1, D2, D3, and D4 activities and compare the experimental results with Aleph[5]. Aleph is an ILP system based on inverse entailment. It has adopted several search strategies, such as randomized search which helps improve the performance of the system. In this experiment, we set Aleph to use GSAT where the best results can be generated. Table 1 shows the prediction accuracy computed for both positive and negative examples, and then, for only the positive examples. The table also shows the results of significance test using one-tailed paired t-test. The experimental results show that FOILMP predicts more accurately than Aleph in both accuracy computation methods. The significance tests also show the confidence level in the difference between accuracy. Figure 4 shows the details of the rules obtained by FOILMP. We also found that FOILMP generates rule with higher coverage than Aleph.

Table 1 Ten-fold cross-validation test on dopamine antagonist data; Superscripts denote confidence levels for the difference in accuracy using a one-tailed paired t-test: * is 95.0%, ** is 99.0%; no superscripts denote confidence levels below 95%.

	FOILMP		Aleph	
	Acc(%) (all)	Acc(%) (pos)	Acc(%) (all)	Acc(%) (pos)
D1	97.0	85.5	96.0*	78.6**
D2	88.1	79.1	86.4*	70.5*
D3	93.4	78.4	93.1	75.1*
D4	88.4	85.1	87.6*	83.2*

```
d1(A) :- atom(A,B,C,D,E,F), E>=3.7, F=3.3,
         bond(A,L,B,H,M,N), bond(A,G,H,I,J,K),
         K=1.5, bond(A,O,B,P,Q,R),
         not_equal(H,P).
```

The rule shows a compound contains an atom `B` with distance to nearest oxygen is larger than 3.7$\mathring{A}$, and distance to nearest nitrogen is 3.3$\mathring{A}$. From `B`, there are two bonds to `H` and `P`. There is another bond from `H` to `I` of length 1.5$\mathring{A}$.

Fig. 4 Rules obtained by FOILMP using data for D1 activity.

3.3 Comparing to different parts

In the previous experiment, an atom is used as a part of compound. Its features are then used to compute DD for weighing. We can consider a compound composing of different kind of part, and features of that part can be used to compute DD for weighing. In this experiment, we consider two adjacent bonds as a part of compound. Thus, a new predicate `twobond(compound, twobond, bond1, bond2)` is generated and included into the dataset. Figure 5 shows a bag representa-

`atom(compound, atom, element, o-connect, o-min-dist, n-min-dist)` – describing an atom `atom` in `compound` with element `element`. It forms a bond with oxygen atom if `o-connect` is 1 and has distance `o-min-dist` and `n-min-dist` to the nearest oxygen and nitrogen atom respectively.

`bond(compound, atom1, atom2, bondtype, length)` – describing a bond `bond` in compound `compound`. This bond links atom `atom1` and atom `atom2` together with type `bondtype` and length `length`.

`link(compound, atom1, atom2, length)` – describing a relation `link` in compound `compound`. It links atom `atom1` and atom `atom2` with length `length`.

Fig. 3 Predicates used to describe dopamine antagonist compound.

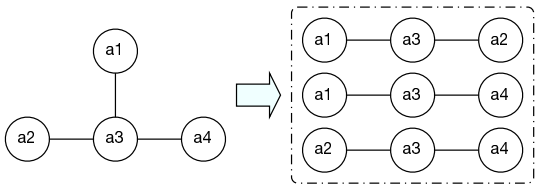

Fig. 5 A bag representation when considering two adjacent bonds as a part.

```
d1(A) :- twobond(A,B,C,D), bond(A,C,E,F,G,H),
         H<1.3, bond(A,D,E,I,J,K), K<1.5,
         K>=1.4, twobond(A,L,M,N),
         twobond(A,O,M,P),
         N\==P, D\==P, C\==P.
```

The rule shows a compound contains two adjacent bonds, `C` of length shorter that 1.3$\mathring{A}$, `D` of length between 1.4 and 1.5$\mathring{A}$ and three adjacent bonds `M`, `N`, and `P`.

Fig. 6 Rules obtained by FOILMP using `twobond` as a part.

tion which is different from one shown in Figure 2.

However, to compute DD, features of `twobond` are required. As a `twobond` composes of two bonds, features of bonds and atoms related to those bonds are used. In other words, we construct a new feature space for `twobond` based on features of bonds and atoms. This approach is useful when only features of atom or bond are unable to discriminate positive and negative compounds. For instance, if feature of atom is only an element type, they can be found in all compounds, such as carbon, oxygen or nitrogen. One way to solve this limitation is to append some special features like one used in the previous section. The other way is to consider a new part composing of simple parts as `twobond`. From this experiment, different rules are generated as shown in Figure 6.

4. Conclusion and Future works

We have presented an improved heuristic function for a data consisting of multiple parts. Diverse Density, a measure for MIL data, is applied to weigh parts, so that parts with common features among positive compounds have high-valued weights. The weights representing quality of examples enable ILP to cut off unnecessary searching paths from an enormous search space and produce more efficient rules.

For future works, scaling factor of features should be considered in DD computing to produce more suitable heuristics. We plan to evaluate the proposed approach on other domains.

Acknowledgments

This research was supported by the Active Mining Project (Grant-in-Aid for Scientific Research on Priority Areas, No.759).

References

1) Dietterich, T. G., Lathrop, R. H. and Lozano-Perez, T.: Solving the Multiple Instance Problem with Axis-Parallel Rectangles, *Artificial Intelligence*, Vol. 89, No. 1-2, pp. 31–71 (1997).
2) Maron, O. and Lozano-Pérez, T.: A Framework for Multiple-Instance Learning, *Advances in Neural Information Processing Systems* (Jordan, M. I., Kearns, M. J. and Solla, S. A.(eds.)), Vol. 10, The MIT Press (1998).
3) Muggleton, S. and Raedt, L. D.: Inductive Logic Programming: Theory and Methods, *Journal of Logic Programming*, Vol. 19,20, pp. 629–679 (1994).
4) Quinlan, J. R.: Learning Logical Definitions from Relations, *Machine Learning*, Vol.5, No.3, pp. 239–266 (1990).
5) Srinivasan, A.: The Aleph Manual (2001). http://web.comlab.ox.ac.uk/oucl/research/areas-/machlearn/Aleph/.