Machine Learning for Bioinformatics: Uncertainty Management with Fuzzy Rough Sets

Sarah Vluymans[1,2,3], Chris Cornelis[1,3] and Yvan Saeys[2,4]

[1] Dept. of Applied Mathematics, Computer Science and Statistics,
Ghent University, Belgium
[2] Data Mining and Modeling for Biomedicine,
VIB Inflammation Research Center, Zwijnaarde, Belgium
[3] Dept. of Computer Science and Artificial Intelligence University of Granada, Spain
[4] Dept. of Internal Medicine, Ghent University, Belgium
{sarah.vluymans, chris.cornelis, yvan.saeys}@ugent.be

1 Introduction

Machine learning is the research domain concerned with *learning from examples*. A general knowledge model is constructed based on specific examples. A machine learning method can discover previously unknown patterns or construct a model to make future predictions. For example, in bioinformatics, machine learning methods are used to analyze biological processes. The application of a computer algorithm allows for an automatic processing and interpretation of possibly large and complex data.

2 Fuzzy rough set theory

In any real-world problem, a level of uncertainty may be expected. Key concepts (like similarity) can be inherently vague, collected data may not be fully reliable due to the presence of noise and the description of observations can be incomplete. One way to deal with uncertainty is to use machine learning methods based on fuzzy rough set theory [4], a hybridization of fuzzy sets and rough sets. The former have been introduced in [8] and are used to represent vagueness by allowing partial membership of elements to sets. The latter approximate incompletely described concepts [5]. By merging the two, fuzzy rough set theory forms a useful mathematical tool to model both vagueness and incompleteness in data. In our recent survey [6], we reviewed the various machine learning paradigms in which fuzzy rough set theory has been employed.

3 Case study: multi-instance learning

Multi-instance learning (MIL, [3]) is a setting in supervised learning dealing with a data format more complex than the one used in traditional learning. In the latter, also referred to as single-instance learning, observations can be modeled by a single vector gathering the values of the observation for all descriptive features. When dealing with a classification problem, a class label is associated

with this feature vector. In MIL, two distinct levels are present in the data. Data samples are groups of individual instances, called bags, and can be modeled as collections of feature vectors. A class value is assigned to the entire bag, not its constituent elements.

The multi-instance data representation is required in applications where an observation (bag) has multiple alternative representations or consists of several parts. These separate components or alternate descriptions correspond to the instances. A variety of application domains include image recognition and text categorization [1]. Due to its flexible data representation, MIL also forms a natural learning framework for several bioinformatics applications. In the original proposal of MIL [3], drug activity prediction was indicated as requiring a multi-instance solution. In this task, the aim is to predict the ability of a molecule to bind to a target, in which case it is a suitable drug molecule. However, biochemistry teaches us that a given molecule can have different conformations (shapes) and not all of them will be able to bind to the target. Without any prior knowledge, it is therefore important to collect all alternative views of the molecule as instances in a bag in order to correctly predict whether the molecule has the binding ability at all. Another example of a bioinformatics problem handled with MIL is protein identification (e.g. [2]).

In [7], we considered the problem of imbalanced multi-instance classification, where the class distribution is skewed, a situation commonly occurring in bioinformatics problems. Few existing multi-instance classifiers take into account the possibility of class imbalance, which causes them to fail in this situation. We presented a classifier framework relying on the use of fuzzy rough set theory. Our methods interpret both classes and bags as fuzzy sets, capturing the inherent data ambiguity present in MIL. The classification step is based on fuzzy rough set measures. The flexibility of fuzzy rough set theory is reflected in the excellent performance of our methods, that are shown to outperform the state-of-the-art in an extensive experimental study.

References

1. Amores, J.: Multiple instance classification: Review, taxonomy and comparative study. Artif. Intell. 201, 81–105 (2013)
2. Minhas, F., Ben-Hur, A.: Multiple instance learning of Calmodulin binding sites. Bioinformatics 28(18), i416–i422 (2012)
3. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell. 89(1), 31–71 (1997)
4. Dubois, D., Prade, H.: Rough fuzzy sets and fuzzy rough sets. Int. J. Gen. Sys. 17(2-3), 191-209 (1990)
5. Pawlak, Z.: Rough sets. Int. J. Comput. Inf. Sci. 11(5), 341–356 (1982)
6. Vluymans, S., D'eer, L., Saeys, Y., Cornelis, C.: Applications of Fuzzy Rough Set Theory in Machine Learning: a Survey. Fund. Inform. 142(1–4), 53–86 (2015)
7. Vluymans, S., Sánchez Tarragó, D., Saeys, Y., Cornelis, C., Herrera, F.: Fuzzy Rough Classifiers for Class Imbalanced Multi-Instance Data. Pattern Recogn. 53, 36–45 (2016)
8. Zadeh, L.: Fuzzy sets. Inform. Control 8(3), 338–353 (1965)