Machine Learning: Theoretical Foundations and Research Topics

Thomas Gärtner

17th December 2021

9 Application: Diterpene Structure Elucidation

NMR spectroscopy is one of the most important techniques in analytical chemistry. It is used as a tool in the search for new pharmaceutical products to help in determining the structure-activity relationships of biologically active compounds. Once these have been determined, it is clear which variations of the compound do not lose the biological activity. In NMR experiments the sample is placed in an external magnetic field and the nuclei are excited by a pulse over a range of radio frequencies. The signal emitted by the nuclei as they return to equilibrium with their surrounding is analysed to obtain an NMR spectrum of radio frequencies.

The problem of diterpene structure elucidation from ^{13}C nuclear magnetic resonance (NMR) spectra was introduced to the machine learning community by Džeroski et al. (1998). There, different algorithms were compared on a dataset of 1503 diterpene ^{13}C NMR spectra. Diterpenes are compounds made up from 4 isoprene units and are thus terpenes – the general term used for oligomers of isoprene. Terpenes are the major component of essential oils found in many plants. Often these oils are biologically active or exhibit some medical properties most of which are due to the presence of terpenes.

The uses of logic as a representation language in learning algorithms are manifold. The simplest type of logic that can be used to represent instances is propositional logic. Lloyd (2003) suggested the use of a typed and polymorphic higher-order logic for learning. There, instances are represented by basic terms in a typed higher-order logic. The alphabet of the logic consists of a set of type constructors with a given arity, a set of parameters, a set of constants with given signature, and a set of variables. The types of the logic are built up from a set of type constructors and the set of parameters using the symbol $\to$ (for function types) and $\times$ (for product types). They are defined inductively. Every parameter is a type. Given a type constructor T of arity n and types $\alpha_1, \ldots, \alpha_n$, the expression $T\ \alpha_1 \ldots \alpha_n$ is a type. Given types α, β, the expression $\alpha \to \beta$ is a type. Given types $\alpha_1, \ldots, \alpha_n$, the expression $\alpha_1 \times \cdots \times \alpha_n$ is a type.

In the dataset considered by Džeroski et al. (1998), each spectrum is described by the frequency and multiplicity of all peaks. Depending on the number of protons connected to the carbon atom, the multiplicity of a peak is either a singlet (no proton), a doublet (one proton), a triplet (two protons), or a quartet (three protons). The formal specification of the data follows an extension of the Haskell language (Jones and Hughes, 1998) and is as follows:

```
type Spectrum = Frequency -> Multiplicity
type Frequency = Real with modifier gaussian 0.6
data Multiplicity = s | d | t | q | 0 with default 0
```

In addition to the multiplicities s(ingulet), d(oublet), t(riplet), and q(uartet) we introduced also the multiplicity 0 and declared it as the default data constructor of the type `Multiplicity`. The abstraction `Spectrum` then maps every frequency (every real number) that is not emitted by the molecule to 0 and every emitted frequency to the multiplicity of the corresponding carbon atom.

The dataset consists of 1503 spectra of diterpenes, classified into 23 different classes according to their skeleton structure as follows (number of examples per class in brackets): Trachyloban (9), Kauran (353), Beyeran (72), Atisiran (33), Ericacan (2), Gibban (13), Pimaran (155), 6,7-seco-Kauran (9), Erythoxilan (9), Spongian(10), Cassan (12), Labdan (448), Clerodan (356), Portulan (5), 5,10-seco-Clerodan (4), 8,9-seco-Labdan (6), and seven classes with only one example each.

The accuracies reported in literature range up to 86.5%, achieved by RIBL (Emde and Wettschereck, 1996). Other results were reported for FOIL (Quinlan, 1990), TILDE (Blockeel and De Raedt, 1998), and ICL (De Raedt and Van Laer, 1995).

References

H. Blockeel and L. De Raedt. Top-down induction of first order logical decision trees. *Artificial Intelligence*, 101(1-2):285–297, June 1998.

L. De Raedt and W. Van Laer. Inductive constraint logic. In K.P. Jantke, T. Shinohara, and T. Zeugmann, editors, *Proceedings of the 6th International Workshop on Algorithmic Learning Theory*, volume 997 of *LNAI*, pages 80–94. Springer Verlag, October 18–20 1995. ISBN 3-540-60454-5.

S. Džeroski, S. Schulze-Kremer, K.R. Heidtke, K. Siems, D. Wettschereck, and H. Blockeel. Diterpene structure elucidation from ^{13}C NMR spectra with inductive logic programming. *Applied Artificial Intelligence*, 12(5):363–383, July-August 1998. Special Issue on First-Order Knowledge Discovery in Databases.

W. Emde and D. Wettschereck. Relational instance-based learning. In *Proceedings of the 13th International Conference on Machine Learning*, pages 122–130. Morgan Kaufmann, 1996.

S. Peyton Jones and J. Hughes, editors. *Haskell98: A Non-Strict Purely Functional Language*. 1998. URL `http://haskell.org/`.

J.W. Lloyd. *Logic for Learning*. Springer-Verlag, 2003.

J. R. Quinlan. Learning logical definitions from relations. *Machine Learning*, 5:239–266, 1990.