BIOINFORMATICS EDITORIAL
Vol. 27 no. 24 2011, pages 3331–3332
doi:10.1093/bioinformatics/btr585

Editorial

The rise and fall of supervised machine learning techniques

Lars Juhl Jensen[1,*] and Alex Bateman[2]

[1]Department of Disease Systems Biology, The Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, DK-2200 Copenhagen N, Denmark and [2]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SA UK

Machine learning is of immense importance in bioinformatics and biomedical science more generally (Larrañaga *et al*., 2006; Tarca *et al*., 2007). In particular, supervised machine learning has been used to great effect in numerous bioinformatics prediction methods. Through many years of editing and reviewing manuscripts, we noticed that some supervised machine learning techniques seem to be gaining in popularity while others seemed, at least to our eyes, to be looking 'unfashionable'.

We were motivated to create a league table of machine learning techniques to learn what is hot and what is not in the machine learning field. In this editorial, we only include those that we considered major league and leave analysis of the minor league methods as an exercise for the interested reader. To create our league table, we created a list of supervised machine learning techniques commonly used in bioinformatics and their common synonyms, plural forms and abbreviations. We then searched this list against the PubMed titles and abstracts to identify the number of papers published per year for each machine learning technique. To match as many papers as possible, searches were case insensitive and allowed for variation in hyphenation.

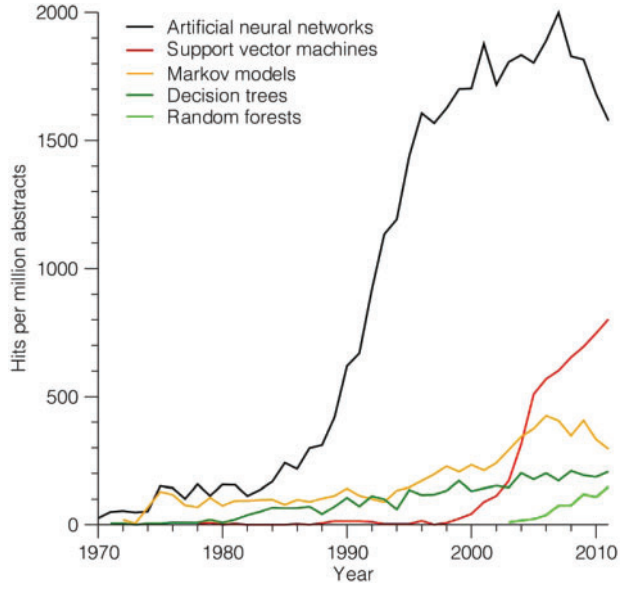

Fig. 1. The growth of supervised machine learning methods in PubMed.

To our surprise, the artificial neural network (ANN) is not only the dominant league leader in 2011 but has been in this position since at least the 1970s (see Fig. 1). However, in recent years the usage of support vector machines (SVMs) grew tremendously, and we predict that SVMs will challenge ANNs for the dominant position in the coming decade. Since 2007 the number of publications using ANNs has decreased by 21%, which we hypothesize may be directly attributed to researchers increasingly using SVMs in place of ANNs. SVMs caught up with and overtook Markov models in 2004 to gain second spot in our machine learning league.

As for the question of 'what is hot?', one can see that Random forests are a rapidly growing method with not a single mention of them before 2003 and now a total of 407 papers published to date.

We were hoping to find techniques that were not so hot and perhaps going out of fashion. The results show that none of the major league methods has gone out of fashion, but we do see moderate decreases in the use of both ANNs and Markov models in the literature.

We were also curious to find out if certain machine learning techniques were used in combination with each other. To investigate this, we looked at what machine learning methods are co-mentioned in articles (See Fig. 2). For all pairs of methods from the Supervised

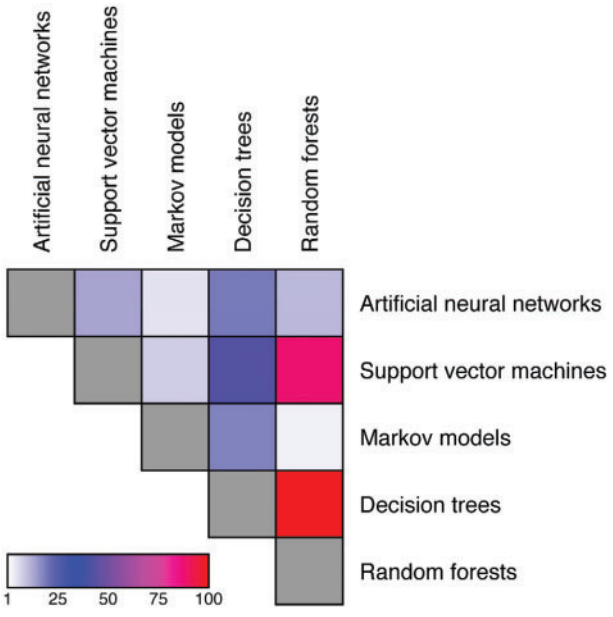

Fig. 2. Heatmap showing the co-occurrence of machine learning techniques within articles.

*To whom correspondence should be addressed

Machine Learning Top-5, we counted the number of abstracts that mention both methods and normalized the counts with the number of co-occurrences that would be expected by chance (based on the frequencies with which the methods are mentioned over the years). The strongest correlation (185 times higher than random expectation) is seen between decision trees and random forests, which is to be expected as random forests are ensembles of decision trees. Apart from this, the next strongest correlation (88 times higher than random expectation) is found between the two newest methods on the list, namely SVMs and random forests. We hypothesize that this is due to many researchers using these algorithms through machine learning frameworks such as Weka (Frank *et al*., 2004), which allows many different algorithms to easily be applied to the same dataset.

Applications of supervised machine learning methodology continue to grow in the biomedical literature. Despite new methods growing in usage, for example support vector machines and random forests, we see little evidence that any widely adopted methods are falling out of use.

Conflict of Interest: none declared.

REFERENCES

Frank,E. *et al.* (2004) Data mining in bioinformatics using Weka. *Bioinformatics*, **20**, 2479–2481.

Larrañaga,P. *et al.* (2006) Machine learning in Bioinformatics. *Brief Bioinform.*, **7**, 86–112.

Tarca,A.L. *et al.* (2007) Machine learning and its applications to biology. *PLoS Comput. Biol.*, **3**, e116.