Journal of Cheminformatics

POSTER PRESENTATION Open Access

Probabilistic classifier: generated using randomised sub-sampling of the feature space

Jonathan D Tyzack*, Hamse Y Mussa, Robert C Glen

From 7th German Conference on Chemoinformatics: 25 CIC-Workshop
Goslar, Germany. 6-8 November 2011

Nowadays supervised classification, based on the concept of pattern recognition, is an integral part of virtual screening. The central idea of supervised classification in chemoinformatics is to design a classifying algorithm that accurately assigns a new molecule to one of a set of predefined classes.

Naturally, probabilistic classifiers can be far more useful than hard point classifiers in making a decision on problems [1], such as virtual screening, where there is an associated risk in classifying an instance to one class or the other.

For their conceptual simplicity and computational efficiency probabilistic classification methods based on the Naive Bayes concept are widely employed in chemoinformatics. The simplicity of the Naive Bayes is due to the assumption that the descriptors representing the molecule one desires to classify are statistically independent. Unfortunately it is well documented that when the molecular descriptors are binary-valued - which is often the case in chemoinformatics - and thus take values of 0 or 1 the Naive Bayesian classifier can only act as a linear classifier in the descriptor space.

Techniques such as the Parzen-Window approach can address the above shortcomings but suffer from being computationally expensive as they require one to retain all the training dataset in core memory [2,3].

In an attempt to address the above mentioned drawbacks, a new probabilistic classifier is proposed which uses randomized sub-sampling of the descriptor space. The proposed algorithm generates better class membership predictions than its Naive Bayesian counterpart on classifying molecules that are non-linearly separable in descriptor space.

We present a realistic test of the new method by classifying large chemical datasets generated from the ChEMBL database [4].

Published: 1 May 2012

References

1. Duda RO, Hart PE: **Pattern Classification and Scene Analysis.** John Wiley & Sons, Ltd : New York, NY; 1973.
2. Parzen E: **The Annals of Mathematical Statistics.** 1962, **33**:1065-1076.
3. Harper G, Bradshaw J, Gittins JC, Green DVS, Leach AR: **.** *J Chem Inf Comput Sci* 2001, **41**:1295-1300.
4. **ChEMBL.** *J Comput-Aided Mol Des* 2009, **4**:195-198.

doi:10.1186/1758-2946-4-S1-P40
Cite this article as: Tyzack *et al.*: Probabilistic classifier: generated using randomised sub-sampling of the feature space. *Journal of Cheminformatics* 2012 **4**(Suppl 1):P40.

* Correspondence: jdt42@cam.ac.uk
Unilever Centre for Molecular Sciences Informatics, Department of Chemistry, University of Cambridge, Cambridge, CB2 1EW, UK