Chemistry Central Journal

Oral presentation

A new approach to kernel based data analysis algorithms

HY Mussa* and RC Glen

Address: Unilever Centre for Molecular Informatics, Department of Chemistry, University of Cambridge Lensfield Road, Cambridge CB2 1EW, UK
* Corresponding author

from 4th German Conference on Chemoinformatics
Goslar, Germany. 9–11 November 2008

Published: 5 June 2009

Chemistry Central Journal 2009, **3**(Suppl 1):O6 doi:10.1186/1752-153X-3-S1-O6

This abstract is available from: http://www.journal.chemistrycentral.com/content/3/S1/O6

Kernel based methods (KBMs) [1,2] are arguably the best data analysis technique currently available [3,4]. Unlike Neural Networks in which, besides a global minimum, several local minima exist, a Kernel based fitting/classifying problem is a convex optimization problem with a single minimum. However, finding this minimum (and in doing so yielding optimal parameters of a given observational model) in practice requires the manipulation, such as inversion, of large matrices. This has been challenging even when the number of data points is just over a few thousands [5][6].

The well established direct methods for updating, or inverting huge matrices fail due to the expense of a large increase in core-memory storage and CPU-time, even for moderately-sized systems. The root of the problem is that direct methods have $O(N^2)$ core memory storage requirements and the CPU-time scales as $O(N^3)$, where N is the dimension of the matrix (the number of data points, here). Despite the advances in computer power, "conventional" computers can only solve relatively small problems ($N \approx 10^4$ to 10^5).

Another outstanding drawback of the KBMs is how to choose the appropriate kernel function for a given data set [4].

In this paper we would like to propose a computationally efficient training scheme for KBMs for obtaining the global minimum. We also present a systematic approach to selecting the appropriate kernel functions. Some preliminary results on chemical data sets will be illustrated.

References

1. Nadaraya EA: *Theory Prob Appl* 1964, **10:**186.
2. Watson GS: *Sankhya Ser A* 1964, **26:**359.
3. Vapnik V: *The Nature of Statistical Learning Theory Springer-Verlag, New York*; 1995.
4. Shawe-Taylor J, Cristianini N: *Kernel Methods for Pattern Analysis Cambridge University Press*; 2004.
5. Chua KS: *Pattern Recognition Letters* 2003, **24:**75.
6. Mangasarian OL, Musicant DR: *J Mach Learn Res* 2001, **1:**161.