Pattern recognition algorithms for biology

Kernel Learning

Kernel learning methods [122, 397, 72, 76, 75, 74, 77, 78, 73], and in particular Vapnik's Support Vector Machine [51, 103, 425, 71] (as implemented in our toolbox [70]), currently represent the most vibrant and promising area of research in machine learning, due to a combination of state-of-the art performance, computational efficiency and mathematical elegance. Kernel methods aim to construct very simple linear statistical models in a "feature space" indirectly specified by a kernel function, which essentially measures the similarity between vectors in this feature space. The mathematical tractability and computational efficiency of kernel methods are a result of the underlying linearity of the model. Fortunately a simple linear model constructed in feature space corresponds to a flexible, non-linear model of the original data, allowing state-of-the art performance on a wide range of real world problems (e.g. [194, 53, 52]).

Kernel learning methods are particularly interesting in bioinformatics research as it is possible to construct kernel functions that operate on sequence data, for instance as DNA, (e.g. [442, 253]). In this case the induced feature space consists of the frequencies of all possible substrings of length k, however using a kernel approach it is not necessary to enumerate every substring explicitly (which would be computationally infeasible). Kernel machines have achieved impressive results in a range of bioinformatics applications, including analysis of gene expression microarray data [57, 197, 218], detecting remote protein homologies [252], analysis of promoter regions [338], recognizing translation initiation sites [462], protein fold classification [295] and protein localization [242].

Research Team: Dr. Gavin Cawley, Dr. Nicola Talbot, Dr. Michael Lincoln, K. Saadi

Identifying Gene Regulatory Systems

In this project, in collaboration with the John Innes Centre, we aim to identify gene regulatory systems in plants, using *Arabidopsis thaliana* as a model organism. This will be achieved by associating conserved regions, or "motifs" in the upstream non-coding regions of the DNA sequence with observed gene expression, as measured during microarray experiments. While we do not intend to use kernel methods exclusively, they are likely to be one of the most promising approaches, both for classification of microarray experiments (e.g. [57]) an for

associating elements within the promoter region with measured gene expression (c.f. [338]).

Research Team: Dr. Gavin Cawley, Dr. Nicola Talbot, K. Saadi. Funding: BBSRC.

Kernel Survival Analysis applied to microbial pathogens

Kernel learning methods to produce improved statistical procedures for modelling data sets describing the growth domain for a range of microbial pathogens [79], including *Clostridium botulinum*. This is an important area in food safety research as a methodology is required that is statistically sound and that quantifies the uncertainty in model predictions for inclusion in a wider analysis of risk (e.g. via Bayesian belief networks [33]). We have started by developing kernel survival analysis models, extending well established linear statistical models (e.g.[108]) to accommodate non-linear interactions between processing and environmental conditions in a principled manner. This work will be extended via the use of transparent kernel functions for visualizing of the model [217], and the development of a hierarchical Bayesian framework to assess the uncertainty of model predictions. This would also aid the collection of further experimental data to concentrate on areas where the uncertainty of the model is greatest.

Research Team: Dr. Gavin Cawley, Dr. Nicola Talbot, K. Saadi. Collaborators Dr Michael Peck (Food Research Institute, adjacent campus). Funding: BBSRC.

Data access in natural science

Web-based resources for oceanograhy

The world carbon dioxide concentration is sampled in many ways, one is from ships as they move around the globe. The Carbon VAriability Studies by Ships Of Opportunity (CAVASSOO) project is providing reliable estimates of the uptake of CO_2 by the North Atlantic, and how this varies from season to season and year to year. These assist in constraining estimates of European and North American terrestrial (vegetation) sinks, using atmospheric inverse modeling techniques. The project is making the datasets selectively available to research groups, and providing access to overview information and other resources for a general audience. A website has been developed which is home to interfaces which allow the user to specify a geographical area to retrieve relevant information. Different interfaces allow the user to specify longitude and latitude manually, to select an area on an interactive map, or to compose SQL queries

Future work involves designing a self-populating digital library using information retrieval and computational linguistic techniques. The user will be able to supply a small number of relevant documents for the system to analyse. From this, the main keywords and phrases relevant to the topic will be extracted using information retrieval techniques and computational linguistic techniques. These key features will be parsed through to a web crawler which will scan the web for documents matching the features extracted from the sample documents. These will be used to populate the digital library database, using a mixture of links, downloaded papers and other resources, together with metadata and summary information. Users will then be able to search for particular papers or other resources focusing in on particular sub-topics, Figure 7.6.

Research Team: Marie-Claire Jenkins, Dan Smith. Collaborators Dr Nathalie Lefvre (Environmental Sciences).

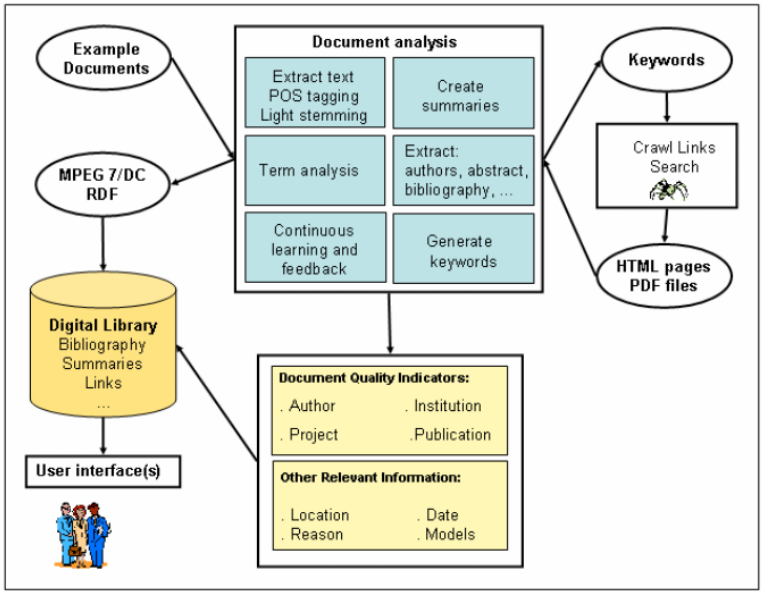

Figure 7.6: Web based resources for logging world carbon dioxide levels.

TLM in the life sciences

(This sub-section repeats a part of the Report on Mathematical modelling.) The absorption of food from the gut is usually modelled as a multi-compartment system. Such models can, in limited circumstances, predict the changes in blood concentration of for example, glucose, that arise when a bolus of food is swallowed. However, these models are too simplistic to account for changes in the way food passes through the gut. The group has been working on models that account for the way food is mixed within a bolus (food charge). A moving boundary diffusion model has been developed to describe the sloughing action which occurs at the periphery of a food charge during the human digestion process. The results are consistent with observations of stomach emptying undertaken using magnetic resonance imaging. This opens the way to extending the studies to more detailed problems.

In the field of chemical kinetics, an early success was the simulation of periodic precipitations [132]. A general treatment of first order- and the extension to higher order- kinetics was based on what was called the multi-compartment technique [392] and this was described in [3] Follow-up investigations after the death of Adnan Saleh led to the realisation that the non-linearities could be handled within one time interval, rather than spread over two (see [130] pp 195-196) and this led to considerable improvements in accuracy and stability. Nevertheless, the scheme could only be applied to well-stirred mixtures. Recent (as yet unpublished work) has led to the development of an alternative TLM model for chemical kinetics which can be implemented in one, two or three dimensions, Figure 7.10. The scheme has been outlined in the thesis of Owen Morris, a 2001/2 MSc student at UEA. There is some evidence of stability problems which could limit the spatial/temporal discretisations that could be used and these are being investigated.

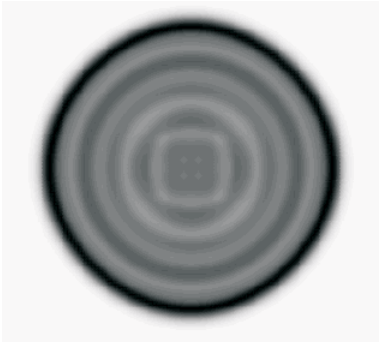

Figure 7.10: A sequence of diffusion waves based on a logistic equation with local fluctuations.

Research Team: Dr Donard de-Cogan, Dr. Pierre Chardaire

Data mining in Biology

(This sub-section repeats a part of the Report on Knowledge discovery.) Collaboration with biologists yields benefits for both sides: problems get solved and computing methods get refined. At present biologists are producing huge amounts of computer readable data. It is a rich source of data for analysis. Examples include, large gene bank depositories (see e.g. *http://www.sanger.ac.uk, http://www.embl-heidelberg.de, http://www.ebi.ac.uk*). Toolkits exist for analysing these databases but the exploitation of data mining techniques in this environment offers the prospect of less human intervention in the search process and better use of computational resource. This is going to be essential if the expected progress is to be made. Work in the group is currently focussed on the yeast with the goal of improving yeast identification techniques either using conventional properties of yeasts [141] or by exploiting properties of the DNA sequences. The work is being extended to determined yeast phylogeny.

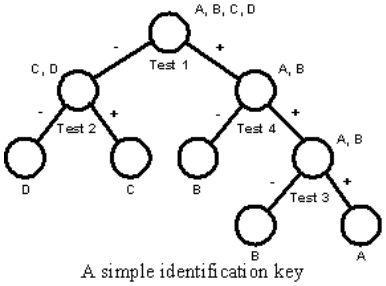

Figure 7.11: A simple identification key for yeasts.

Classification and Identification of Yeast Some yeasts may cause illness or food spoilage. Others may be entirely harmless. As such, the efficient classification and identification of yeast species is of importance. Our research concentrated on performing these two tasks, using either molecular or physiological characteristics to differentiate between species.

A database of the results of 96 physiological tests on 745 yeast species was donated by the CBS [427]. One avenue of our research was the investigation of the efficient selection of these tests, when attempting to identify an unknown yeast sample. There are three broad approaches to test selection:

1. Select a set of tests to perform, and perform all these tests in parallel. In this case, the aim is to find a test set that has maximal diagnostic power but minimal cost. We applied simulated annealing to find such sets and to determine how many of the tests are redundant [141].

2. Select one test at a time, and perform tests sequentially. Here an identification key must be constructed that minimizes the expected costs of identification. We have developed both a greedy algorithm and a simplified GRASP for the construction of such keys [444].

3. Perform tests in batches. The greedy algorithm for identification key construction was modified to permit keys where tests may be performed in batches. It was then shown that, when the aim is to minimize a weighted

sum of the material cost of identification and the time taken, the algorithm automatically produces keys where tests are performed in batches [444].

A second avenue of research led to the development of a hashing technique used in the search for unique DNA sequences in 702 26S rRNA genes. A unique DNA sequence was found for nearly every yeast species known. Surprisingly, it was found that most species can be identified with a sequence just eight base pairs long [444]. Research Team: Professor Vic Rayward-Smith, Dr. Bea de la Iglesia,

A.P. Reynolds, J.J. Wesselink, R.P.Davey, G.M.Sawa, G.Etherington, M.Burgess. Collaboration: Professor Dicks (JIC), Dr. Roberts (IFR). Funding: BBSRC (ref 83/BIO12037)

Data mining in Medicine

Medical data poses significant challenges to the data mining community for a number of reasons:

Diversity Historically, databases on specialist topics have been collected by enthusiastic individuals. Some hospitals and many general practitioners have also collected large databases of more general data. Some of these have used prescribed coding schemes (e.g. READ codes [370] SNOMED clinical terms [405] and HL7) but there is a large amount of medical data which does not use any such schemes. The result is a plethora of databases using different terms and designed for different groups of patients; they are at different levels of generality and targeted towards different medical specialities. The use of formal ontologies such as OIL + DAML (see e.g. http://www.w3.org/TR/daml+oil-reference) applied to specific areas of medicine will provide a way of overcoming these difficulties and will enable the mining of data from multiple sources. Some initial work has already been done on general medical ontologies [216] and we expect to build on the semantic definitions of SNOMED CT.

Time dependence and history In many applications of medical data mining, the interest focuses on the effectiveness (or otherwise) of certain treatments. The previous history of patients and, in particular, any treatments they have received must be considered [352]. Recording and representing this information poses a considerable challenge. Moreover, for this data to be used effectively, it may require a specialist in one field to understand specialist terminology used in another field. When Data mining, access to multiple and interlinked ontologies is going to be critical.

Multi-media Medical data comes in a wide variety of forms. As well as databases, medical information is kept in free text format (sometimes poorly structured), as images (e.g. scans, x-rays) and as charts. To effectively datamine such a wide variety of data requires access to a range of skills. To be successful in medical data mining, skills in computational linguists and medical imaging are required.

Availability The medical community appears to have an ambivalent attitude to the collection and analysis of data. Hospitals can vary widely in their use of IT. In many hospitals, computing has been treated as a management overhead and doctors have had little or no incentive to become involved; relatively few consultants are fully exploiting IT [39, 40]. However, almost all general practitioners use computers in their consulting rooms.

Credibility There is a credibility issue in that the medical profession are now encouraged to practice 'evidence based medicine'. This has meant it relying heavily on the evidence gained from double blind randomised prospective clinical trials and their meta-analysis. The extraction of knowledge from large databases containing details of patients with widely varying backgrounds is relatively new and can be treated with suspicion. The medical community needs to be convinced that there are valid techniques to allow for patient disparity in a database. Of course such techniques do exists and have been widely used in commerce and, indeed, the School of Information Systems has been working for many years with Norwich Union to better understand their large customer database.

Patient confidentiality and ethics Anyone working with health data has to appreciate the confidential nature of the data. Often this can be partially overcome in data mining work by removing all references to named individuals but that sometimes results in lost linkages with other data sources. Results of data mining research can generate highly sensitive conclusions which can cause individual distress if not presented in a correct and supportive manner.

However, despite these problems, medical data mining has been making steady progress and has enormous potential [95, 240]. The goal of the research is to be able to trawl medical data recognising important trends and patterns which, in turn, will inform and develop medical opinion and practice.

Practical medical data mining projects have been undertaken in

1. Osteoporosis [Ray, Wang and Partridge, (ref 337)]

2. Diabetes [Richards and Rayward-Smith]

Further work is planned in both of these areas plus

1. Colonoscopy
2. Hip replacement
3. Breast cancer
4. GP demand and usage

Most projects involve close co-operation between researchers in the School and colleagues in medicine, often based in the University's School of Medicine, Health Policy and Practice or in the Norfolk and Norwich University Hospital. The School intends to become a major UK centre for medical data mining. Successful medical data mining will need to include techniques to extract knowledge from image data, natural language text, DNA and diagrams as well as from single or multiple databases, possibly web-based and supported by medical ontologies. Close co-operation with the medical vision researchers and researchers in natural languages within the School will be necessary.

Research Team: Professor V. J. Rayward-Smith, Dr. G. Janacek, Dr. A. J. Bagnall, Dr. B. De la Iglesia, Dr. G.D. Smith, Dr. G. Richards, Dr W. Wang, Sami Al-Harbi, Martin Burgess, Ellen Yi. Collaboration: Professor A Barrett, Consultant Oncologist, Norfolk and Norwich University Hospital and Professor in the School of Medicine, Health Policy and Practice, UEA, Professor Duncan Bell, Consultant Gastroenterologist at the Norfolk and Norwich University Hospital, Professor Peter Sonksen MD FRCP, Professor of Endocrinology, St. Thomas' Hospital, London. Funding: EPSRC

Environmental Sciences

General circulation models are considered to provide the best basis for estimating future climates that might result from anthropogenic modification of the atmospheric composition. However, output from these models cannot be widely or directly applied to many impact studies because of their relatively coarse spatial resolution. Statistical down-scaling methods seek to model the relationship between large scale atmospheric circulation on, say, a European scale and climatic variables, such as temperature and precipitation, on a regional scale. Work in SYS on the EU funded STARDEX project, in collaboration with the School of Environmental Sciences, aims to build on current modeling techniques used for

statistical downscaling [235, 447, 207, 463] such as neural network approaches, and investigate the use of new techniques such as Support Vector Machines.

Research Team: Dr. Gavin Cawley, Dr. Micheal Lincoln. Collaboration: Dr. Clare Goodness, CRU. Funding: EU Fifth Framework