Editorial

Leif E. Peterson

Center for Biostatistics,
The Methodist Hospital Research Institute,
6565 Fannin, Suite MGJ6-031,
Houston, Texas 77030, USA
E-mail: lepeterson@tmhs.org

Paulo J.G. Lisboa*

School of Computing and Mathematical Sciences,
Liverpool John Moores University,
Byrom Street, Liverpool L3 3AF, UK
E-mail: P.J.Lisboa@ljmu.ac.uk
*Corresponding author

The papers in this special issue are extended versions of a special session at *ICMLA 2008* on 'Application of machine learning in constructing biopatterns and analysing bioprofiles.' The focus of the session was the assimilation of a wide range of information about an individual patient into a single predictive model to drive the delivery of healthcare. This theme is therefore central to implementing the agenda of personalised medicine and it is reflected in the selection of papers that follow.

Initial exploratory analysis of complex data sets often relies on the application of clustering methods. These coming in a wide range of different types, some agglomerative and hierarchical, others based on Voronoi partitions of the data space with discriminant surfaces, yet others are model-based. The paper on affinity propagation shows the effectiveness of combining pairwise similarities similar to aggregative methods with mixture model assumptions made about the cluster groups, with several examples relating to biomolecular sub-typing breast cancer. Ultimately, the consensus of several methodologies, together with principled approaches to determine cluster numbers, promises to open the door to significant advances in histological practice across many pathologies.

A second topic of focus in this special issue is the classification of high-dimensional data. This is a current problem studied with radically different approaches that include dimensionality reduction, blind signal deconvolution, wrapping with recursive feature elimination and filtering with statistical methods. An intuitive approach related to computational learning is to maximise the separation margin between classes. This is related to support vector machines (SVM) but can be implemented within the context of another intuitive method, nearest neighbour classifiers, as applied to mass spectrometry in a proteomic study by Liu et al.

An attractive alternative to single classifier methods is the combination of classifiers into ensembles or in some other way. For medical data, interpretation of discriminant functions can be helpful at the validation stage of decision support systems so this

motivates the use of decision templates, which are benchmarked against SVMs on single-voxel magnetic resonance spectra of brain tissue in the paper by Dimou and Zervakis, with promising results.

Feature reduction methods also include principal component analysis, radial basis methods such as K-means and neural adaptive learning with self-organising maps. The subject of feature extraction in high-dimensions is explored further in a fingerprint classification paper by Peterson et al, which evaluates image information retrieval with high-dimensional Gabor wavelets, dimensional reduction using non-linear self-organising maps, and classification of final image features with Hermite and Laguerre neural networks.

Having discussed exploratory modelling and classification, the last two papers turn to a topic that is less frequently studied yet of critical practical important in medical decision support – this is the derivation of prognostic predictions from time-to-event data. Longitudinal studies of clinical outcome are critical in the evaluation of medical treatment and important to the choice of therapy when this needs to be made shortly after diagnosis of initial surgical intervention. The evidence for these models comes from longitudinal studies, whose design is itself a matter of considerable interest. In the analysis of routinely acquired time-to-event survival data, however, it is frequently the case that important predictors are missing for some of the patients. The issue is then how to robustly incorporate missing data on covariates into a flexible model for right censored data. An example of how this can be done within a rigorous statistical framework is explained in the paper by Fernandes et al, again by reference to an application in breast cancer.

Finally, once data sub-typing is complete, high dimensional signals are appropriately represented, multiple modalities are assimilated and predictive models are prepared, taking into account the vagaries of routine medical records, it remains to validate the predictions in out of sample studies. It is generally known that there is a hierarchy of validation which begins with cross-validation often within a single retrospectively acquired dataset; this extends to temporal validation, that is to say predictions of a further cohort of patients whose data are acquired over a time period that is subsequent to the data used to fit and internally validate the initial predictive model; and last comes multi-centre validation, by which predictions are obtained from data in one hospital but for data from patients in another hospital with its own interpretation, measurement and recording methods for each covariate, which can introduce unexpected sources of variation and noise. The last paper in this group presents the results from a successful external validation of a prognostic model for ocular melanoma.

In summary, this special issue represents an order walk through recent advances in the important stages required to merge complex, multi-modal signals into a coherent bioprofile, which we call the Biopattern for the individual patient. Much work remains to be done before definitive methods are established for each stage in this process, and the biopattern is tracked over time to identify incipient trends, and these are applied in practice for preventive interventions as well as for making diagnostic and prognostic predictions about individual patients. Work such as that presented together in this special issue is in the critical path towards realising the ambitious goal of personalised medicine.