Institution: University of Leicester

Unit of Assessment: 10 Mathematical Sciences

Title of case study: Data maps with applications to medical diagnostics and monitoring

1. Summary of the impact

Advanced technologies for data visualisation and data mining, developed in the Unit in collaboration with national and international teams, are widely applied for development of medical services. In particular, a system for canine lymphoma diagnosis and monitoring developed with [Text removed for publication] has now been successfully tested using clinical data from several veterinary clinics. The risk maps produced by our technology provide early diagnosis of lymphoma several weeks before the clinical symptoms develop. [Text removed for publication] has estimated the treatment test, named [Text removed for publication], developed with the Unit to add [Text removed for publication] to the value of their business. Institute Curie (Paris), applies this data mapping technique and the software that has been developed jointly with Leicester in clinical projects.

2. Underpinning research

The problems related to *large data set analysis and visualisation, model reduction and the struggle with complexity of data sets* are important for many areas of human activity. The identification of hidden geometry and topology in noisy data sets is a challenging task. Many branches of data analysis aim to solve such problems under some additional assumptions that simplify the problem. However, the verification of these assumptions may be more complicated than the solutions of the problems. A universal technology for uncovering the hidden structure is very desirable. An answer to this challenge cannot be simple because it must potentially cover the majority of situations.

In the 1990s Levesley and Light produced theoretical results concerning the approximation power of neural networks. This work led to Levesley's involvement in a simple neural network model for the prediction of acute rejection of kidney transplants together with pathologists from the University of Leicester [**3**.7]. The Unit recognised the potential for impact of research in this area, leading to the development of a team under the leadership of Gorban with more specific expertise in the theory and practical application of neural networks.

In summary, we have developed a universal technology for revealing and visualising the hidden structure in data. For this purpose, we have used ideas both old and new:

- The oldest of them is the idea of self-consistency introduced by H. Steinhaus in 1957 (k-means) and then recognized as a very general and productive idea that can be used for construction of many principal objects like principal manifolds and principal graphs (Husty at al, 1984). This idea is an intrinsic part of the self-organizing maps (SOM) and many data approximation approaches also.
- The application of quadratic elastic energy functionals is a basic idea in spline approximation and is used by us for construction of principal manifolds, in the elastic maps technology [**3**.6].
- Gorban and Zinovyev (Curie) developed the topological grammars approach for data analysis [**3**.3] based on the idea of graph grammars.
- We use the pluriharmonic embeddings of graphs into data space as the ideal approximators [**3**.5] and developed optimization methods to minimize the deviation of data approximants from the pluriharmonic graphs.

- The idea of robust growth makes the whole approach more efficient. For the organization of robust grows, we use truncated energy functionals. In the splitting algorithms of optimization they also produce systems of linear equations, and make the construction of the approximators much more stable in presence of noise and outliers.

Most of the ideas are implemented in user-friendly software and can be applied to many real-life problems.

For the development of applied systems we combine our original technology with more classical approaches like decision trees, advanced kNN method and Bayesian networks. For example, for the canine lymphoma diagnosis we have tested more than 2,000,000 versions of combinations of known and our novel data mining approaches, and the best solutions have been implemented in JAVA (web-accessible) software. It is shown that for the differential diagnosis of clinically vulnerable patients, the sensitivity (proportion of correct prediction of positive results) of the system is 83.5%, and specificity (proportion of correct prediction of negative results) is 77%. For caninelymphoma screening purposes, the best data mining solution we found has sensitivity 81.4% and specificity >99%.

On base of case-study, which has been done, the best solution for each problem has been selected. The results obtained from case-study are extremely favourable compared to many current human cancer screening tests that rely upon single biomarkers. These include the current CA-125 screen for human ovarian cancer (sensitivity approximately 50% and specificity 98% [**3**.1]) and the male PSA test (sensitivity approximately 65% and specificity 35% [**3**.2]).

3. References to the research

Publications

(1) A.N. Gorban, A. Zinovyev, Principal manifolds and graphs in practice: from molecular biology to dynamical systems, *International Journal of Neural Systems* 20 (3) (2010), 219–232.

(2) A.N. Gorban, A. Y. Zinovyev, Principal Graphs and Manifolds, Chapter 2 in: *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques,* Emilio Soria Olivas et al. (eds), IGI Global, Hershey, PA, USA, 2009, pp. 28-59.

(3) A.N. Gorban, N.R. Sumner, and A.Y. Zinovyev, Topological grammars for data approximation, *Applied Mathematics Letters,* 20 (4) (2007), 382-386.

(4) A. Zinovyev, E. Mirkes, Data complexity measured by principal graphs, *Computers & Mathematics with Applications,* Volume 65, Number 10, 1471-1482.

(5) A.N. Gorban, B. Kegl, D. Wunsch, A. Zinovyev (Eds.), *Principal Manifolds for Data Visualisation and Dimension Reduction*, Lecture Notes in Computational Science and Engineering, Vol. 58, Springer, Berlin – Heidelberg – New York, 2008. (ISBN 978-3-540-73749-0).

(6) A. Gorban, A. Zinovyev, Elastic Principal Graphs and Manifolds and their Practical Applications, *Computing* 75 (2005), 359–379.

(7) Furness PN, Levesley J, Luo Z, Taub N, Kazi JI, Bates WD, Nicholson ML., A neural network approach to the biopsy diagnosis of early acute renal transplant rejection, *Histopathology*, Volume 35 (1999), 461-467.

Grant

Data Mining for Lymphoma Differential Diagnosis, A University of Leicester Innovation Partnership with [Text removed for publication], 2012. European Regional Development Fund.

4. Details of the impact

Joint work with Institute Curie (Paris, France) started in 2004. This is one of the top European cancer research and treatment centres. Together with the Bioinformatics Unite of Institute Curie, we have developed a software library which implements most of our methods. This software is now open for non-commercial use worldwide [**5**.2]. Institute Curie uses this software in various projects for visualization and analysis of microarrays for various types of cancer, for visualization of clinical and biochemical data [**5**.2].

Publication [**5**.3] demonstrates knowledge transfer impact as the IC-MSQUARE conference is dedicated to application of mathematics in other science and technology, and the author list of the paper has two member of the University (Gorban and Mirkes) and three colleagues from [Text removed for publication] (Alexandris, Slater and Tuli).

Use in Humans

Many institutions and clinics in various countries have reported successful use of these methods and software for clinical purposes [**5**.2]:

- The Ukrainian Medical Almanac [**5**.6] reported two new applications: (i) Prediction of treatment result of long bones fracture for diabetes patients, (ii) Pain management and quantitative estimation of pain.
- Dr. Arndt Benecke (joint affiliation at Institut de Génétique et de Biologie Moléculaire et Cellulaire, CNRS/INSERM/ULP, Collège de France and Institut des Hautes Etudes Scientifique, France) used the method of elastic maps for analysis of microarray data in cancer. This experience was reflected in the subsequent publication [**5**.8].

Use in Animals

The treatment of dogs is a vast and recession-resistant business: there are 80 million dogs in the United States alone, and even in recession most people keep spending on their pets. Research into the treatment of cancer in dogs also has relevance to the treatment of cancer in humans, particularly because it relates to spontaneous cancer which occurs in a domestic environment. “Lymphoma is one of the most common canine cancers, representing 5% of all malignancies. It has an annual incidence on 25 cases per 100,00 dogs” [**5**.7].

[Text removed for publication] has developed a lymphoma blood test, [Text removed for publication], [**5**.4, **5**.5] which gives vets an easier, less stressful, cheaper and quicker way of testing for lymphoma. This means that dogs are more likely to be tested for lymphoma when any suspicious symptoms show, and that results of the tests are available quickly – generally the same day. If lymphoma is caught early on it can be treated quickly. While researchers do not talk of a “cure” for lymphoma, early treatment can produce a healthier dog for longer, adding 12 months to two years to a dog’s average 12-year lifespan.
The blood test was developed from serum samples collected from several veterinary practices. The samples were fractionated and analysed by mass spectrometry. Two protein peaks, with the highest diagnostic power, were selected and further identified as acute phase proteins, C-Reactive Protein and Haptoglobin. Data mining methods were then applied to the collected data for the development of our online computer-assisted veterinary diagnostic tool.

After testing of more than 2,000,000 versions of the combinations of the known and original data mining approaches, the best solutions were found. It is tested on the clinical data of several veterinary clinics worldwide. The generated software is a tool for diagnostic, monitoring and screening. Initially, the diagnosis of lymphoma was formulated as a classification problem and then later refined as a lymphoma risk estimation. Three classical methods, decision trees, advanced kNN and probability density evaluation, were used in combinations with original approaches for classification and risk estimation and several pre-processing approaches were implemented to create the diagnostic system.

For the differential diagnosis the best solution gave a sensitivity and specificity of 83.5% and 77%, respectively (using three input features, CRP, Haptoglobin and standard clinical symptom). For the screening task, the decision tree method provided the best result, with sensitivity and specificity of 81.4% and >99%, respectively (using the same input features). Furthermore, the development and application of new techniques for the generation of risk maps allowed the visualisation of risk maps in a more user-friendly manner.

This is a potentially useful tool for explanatory data analysis and testing other theoretical input features in the final diagnosis. The risk maps provide early diagnosis of lymphoma return several weeks before the clinical symptoms are developed. In this monitoring lymphoma return the risk maps perform significantly better than most of the veterinary practitioners. The generated lymphoma software (JAVA) has the potential of being web-accessible.

In a letter to the Vice-Chancellor of the University of Leicester from [Text removed for publication] reports "The new treatment monitoring test has the potential to add a further [Text removed for publication] to our projected turnover. It has also bought forward the collaboration with the largest veterinary corporation in the UK who were specifically interested in the treatment monitoring application of our test. They are now planning to launch the new test developed with University of Leicester which will have an immediate impact on both our short and long term revenues" [**5**.1].

In short -- this system is significantly changing veterinary practice in the UK.

5. Sources to corroborate the impact

1. Factual statement by [Text removed for publication]

2. Factual statement from Director of U900 Institut Curie and references to the clinical projects.

3. E. M. Mirkes, I. Alexandrakis, K. Slater, R. Tuli, A. N. Gorban, Computational Diagnosis of Canine Lymphoma, Presented at the conference IC-MSQUARE 2013, Prague September 2013 (Short version is published in the Book of Abstracts IC-MSQUARE 2013), Accepted for publication in IC-MSQUARE 2013 Proceedings (IOP Conference series), extended version is invited to the Special Issue of Physics in Medicine and Biology. Preprint version is published in arXiv: arXiv:1305.4942 [q-bio.QM]

4. Canine lymphoma blood tests – results explained, [Text removed for publication], internal publication.

5. Guidance notes for [Text removed for publication], the canine lymphoma blood test system.

6. Ivchenko V.K., Galchenko V.Ya., Ivchenko A.V.: Part I: Prediction of treatment result of long bones fracture for diabetes patients by means of intellectual and statistical data analysis. Part I. Visual data mining for multidimensional data, Ukrainian Medical Almanac , 2013, Vol. 16, Iss. 2 (Supplement), pp. 4-7; Part II. Production of prognostic classification rules, Ukrainian Medical Almanac , 2013, Vol. 16, Iss. 2 (Supplement), pp. 8-11; Part III. Analysis of efficiency of produced prognostic classification rules, Ukrainian Medical Almanac , 2013, Vol. 16, Iss. 2 (Supplement), pp. 12-15.

7. [Text removed for publication]

8. Bécavin C, Benecke A. New dimensionality reduction methods for the representation of high dimensional 'omics' data. *Expert Rev Mol Diagn*. 11(1) (2011), 27-34.