Tina Memo No. 2011-012

Internal, Initial draft of Leverhulme proposal.

Quantitative Use of Pattern Recognition in the Analysis of Complex Data Distributions.

N.A. Thacker.

Last updated
12 /4 / 2012

Imaging Science and Biomedical Engineering Division,
Medical School, University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT.

Quantitative Use of Pattern Recognition in the Analysis of Complex Data Distributions

Abstract
We aim to improve upon conventional approaches to Pattern Recognition (PR), facilitating the quantitative interpretation of data suitable for making scientific measurements – an enabling technology currently lacking. We seek to analyse sampled distributions, in the form of histograms, commonly used for expressing complex data within a range of disciplines (e.g. planetary science, spectroscopy, and image analysis). Our goal is to continue development of quantitative methods designed to summarise the composition of distributions in terms of predefined categories. We will demonstrate the method's utility in a range of applications, including the analysis of planetary surfaces and mass-spectra.

Proposal Summary
Pattern Recognition (PR) is a relatively modern approach to data interpretation that has developed alongside modern computing. It is predominantly the process of assigning pre-learned categories to individual instances of data, often for quantitative tasks (i.e. counting), and can be viewed as an estimation process. Many modern scientific data analysis problems could be amenable to these procedures making PR a significant enabling technique (e.g. genomics). However, it does not seem to be widely appreciated that PR differs from more conventional statistical estimators in important ways. In particular, conventional statistical methods, such as model fitting (regression), include theories for providing parameter error covariances and goodness-of-fit measures. Covariances are considered essential for the scientific use of data, and goodness-of-fits are important for confirming the success of analyses [1]. PR methods do not provide these outputs. Instead, they are typically assessed using receiver operating characteristic curves (ROC) which quantify identification and misidentification rates empirically against known reference data. This allows alternative methods (such as Support Vector Machines or Random Forests) to be treated as black-boxes (i.e. understanding of tuning parameters is unnecessary) and ranked when applied to standard datasets [2]. However, there are problems in using such empirical information in the context of a scientific study. In particular it is of no value in the general task of comparing predictions from theory with observation, where covariances and goodness-of-fits are far more appropriate.

We might reasonably ask why PR methods cannot provide parameter covariance and goodness-of-fit estimates? If these were available, they would have as much utility as other estimation procedures but with the potential to be applied to a wider range of applications. The answer may be that users of PR systems wish to apply their techniques to arbitrary datasets. To do so they generally assume that factors such as measurement noise are negligible, leading to simplified black-box methodologies. However, goodness-of-fit and parameter error covariances can only be estimated *with knowledge of such factors*. The estimation of errors in conventional statistics is achieved through links to quantitative probabilities (generally via Likelihood), and methods including the minimum variance bound and error propagation [3]. Most PR methods purport to be related to probability theory and should therefore be amenable to a similar ``white-box'' approach. However, those using PR are aware that the underlying probability theory is based on poorly approximate distributions. This is often a consequence of inherent variability, i.e. generating distributions are not fixed. Consequently, attempting to make theoretical predictions of performance is likely to be highly inaccurate unless additional steps are taken [4]. This important and necessary work has thus far not been undertaken by practitioners in the PR field, limiting applicability to serious scientific analyses.

It is the conviction of the PI on this project that work is needed which will make PR fully quantitative. To this end, we will apply knowledge of statistical methods and imaging science (as demonstrated by the PI [10-18]), pattern recognition and machine learning, and

application specific expertise in planetary science and spectroscopy, in order to create a widely applicable enabling PR technique.

Our preliminary work [5][6] has demonstrated that for *well-defined problems* it is possible to construct supervised PR systems capable of producing covariance estimates and goodness-of-fit measures. We have developed methods for constructing low parameter models of data distributions from training data, which can be fitted to incoming data in order to quantify the composition in terms of pre-learned categories. The method's validity has been tested using Monte-Carlo, corroborating all aspects including quantity estimates, covariances and goodness-of-fit measures. However, experiments on more realistic data (for making surface area measurements from Martian terrains) exposed limitations, believed to be caused by complex residual correlations in data distributions. The proposed project aims to address these difficult issues and so extend the applicability of our methods to real scientific applications.

The scientific goals we will address have been defined jointly with the project CI's as unsolved analysis problems and are: the analysis of satellite images of the Martian surface; the automated identification of lunar craters and the filtering of Citizen Science (Moon Zoo) data [7], and the analysis of laboratory spectroscopic data. Each of these applications can be viewed as PR problems:

The frequency distribution of locally repeating patterns found within Martian images can be learned for user-defined categories of terrain (e.g. different types of dunes, polygonal fracture patterns etc.). Using these histograms, we plan to demonstrate the abilities of our PR methods by quantifying regional distributions and orientations of dunes [8], which will provide insights into wind patterns and grain availability at selected locations. As a proof-of-concept we also plan to perform similar analyses for a range of indicative textures to illustrate the flexibility and generic applicability of our techniques within Martian surface analysis.
a

The techniques developed for Martian terrains will be adapted for large-scale quantitative crater counting (using NAC images from Apollo sites). They will be applied to estimate regional crater Size Frequency Distributions for the dating of lunar geological units [9]. Fully automated methods will be trained using expert crater definitions then applied to regions of interest. Semi-automated methods will be used to filter Moon Zoo crater data to separate false-positives from correctly identified citizen scientists' crater candidates. This latter work forms an important part of the Moon Zoo data reduction pipeline, and has the support of the Moon Zoo team. Such techniques are important for the reduction of large databases into science-oriented outputs.

For organic (MALDI) spectroscopy data, samples of spectra for pure molecular compounds will be obtained and used to construct individual models of spectra variability. Combination of pure-spectra samples will then be used to generate synthetic data for evaluation of the theory via a comparison of the prediction of known quantities and associated measurement uncertainties.

The performance of the quantitation methods will be assessed via a comparison with ground truth samples, constructed specifically for the purpose.

All results will be published in appropriate literature as exemplars of the suitability of the new theory in the quantitative analysis of data. Software will be made available as Open Source.

ADDITIONAL:

Mass spectrometry imaging (MSI) is now possible with a wide range of mass spectrometer types, differing mainly in the means of sample ionisation. These include SIMS, where ionisation is initiated by an ion beam, MALDI that uses UV laser pulses and DESI based on electrospray ionisation. Applications range from semiconductor device characterisation, imaging of geological samples to biological tissue imaging. MSI records a histogram-like mass spectrum at each image pixel. The spectrum recorded at each pixel is the sum of the mass spectra of the molecular components of that sample region, weighted according to their ionisation and detection efficiencies, together with instrumental noise. The spectra are approximately additive, although there are known interaction effects that can lead to signal suppression or enhancement. MALDI-MSI data will be used as representative of the MSI data type. MALDI spectroscopic data, will be analysed, starting with mass spectra of pure compounds and the spectra of simple mixtures to construct models of spectra variability and interaction. Combination of pure-spectra samples will then be used to generate synthetic data for evaluation of the theory via a comparison of the prediction of known quantities and associated measurement uncertainties. The performance of the quantitation methods will be assessed via a comparison with ground truth samples, constructed specifically for the purpose.

References

[1] Ref Numerical Recipes

[2] R. Caruana, A. Niculescu-Mizil An Empirical Comparison of Supervised Learning Algorithms, Proceedings of the 23d International conference on Machine Learning, Pittsburgh, PA, 2006.

[3] Alan Stuart, Keith Ord, Steven Arnold, Kendall's Advanced Theory of Statistics: Classical Inference & the Linear Model, Sixth Edition, vol 2A ch 17 pp 1-45 ISBN 0-340-66230-1, 1999

[4] Alexandru Niculescu-mizil and Rich Caruana, Obtaining Calibrated Probabilities from Boosting, Proc. 21st Conference on Uncertainty in Artificial Intelligence (UAI ’05, 2005

[?] K. Chatfield, V. Lempitsky, A. Vedaldi, A. Zisserman, The devil is in the details: an evaluation of recent feature encoding methods, Proc. BMVC 2013, 2013

[5] Tar P.D., Thacker N.A., Jones M.A., Gilmour J.D., A Quantitative Approach to the Analysis of Planetary Terrains, Proc. Remote Sensing and Photogrammetry Society Conference 2012, 2012

[6] Tar P.D., Thacker N.A., Linear Poisson Models: A Solution to the Histogram Composition Problem, SUBMITTED TO JCMSE, UNDER REVIEW (details of background work available at www.tina-vision.net/docs/memos.php)

[7] Joy K. H., Crawford I. A., Grindrod P. M., Lintott C. J., Bamford S., and Smith A. (2011). “The Moon Zoo Citizen Science Project.” Astronomy and Geophysics. Vol. 52, pp. 2.10-2.12. doi:10.1111/j.1468-4004.2011.52210.x

[8] Hayward R.K. Et al. Mars Global Digital Dune Database and initial science results, Journal of Geophysical Research, Vol 112, E1100, 2007

[9] Stöffler D., Ryder G., Ivanov B. A., Artemieva N. A., Cintala M. J., Grieve R. A. F. 2006. Cratering history and lunar chronology (in New views of the Moon) Reviews in Mineralogy and Geochemistry (2006), Vol. 60. pp. 519-596.

[10] P.A. Bromiley, N.A. Thacker, P. Courtney, Non-Parametric Subtraction Using Grey Level Scatter-grams, Image and Vision Computing, 20, 609-617, 2002

[11] P.A. Bromiley, M.L.J. Scott, M. Porkic, A.J. Lacey, N.A. Thacker, Bayesian and Non-Bayesian Probabilistic Models for Magnetic Resonance Image Analysis, Image and Vision Computing, Special Edition: The use of probabilistic models in computer vision, 21, 851-84, 2003

[12] A. Nayak, E. Trucco, N.A. Thacker, When are simple LS estimators enough? An empirical study of LS, TLS and GTLS, IJVC, 68-2, 203-216, 2005

[13] N.A. Thacker, A. Clark, J. Barron, R. Beveridge, P. Courtney, W. Crum, V. Ramesh, C. Clarke, Performance Characterisation in Computer Vision: A Guide to Best Practices, CVIU, 109, 305-334, 2008

[14] T.F. Cootes, N.A. Thacker, C.J. Taylor, Automatic Model Selection by Modelling the Distribution of Residuals, Proc. ECCV 2002 (IV), LNCS 2353, 621-635,2002

[15] P.A. Bromiley, M. Porkic, N.A. Thacker, Computing Covariances for Mutual Information Co-registration, Proc. MIUA, 77-80, London, 2004

[16] N.A. Thacker, C. Leek, Retinal Sampling, Feature Detection and Saccades: A Statistical Perspective, Proc. BMVA, 990-999, 2007

[17] N.A. Thacker, Quantitative Verification of Projected Views Using a Power Law Model of Feature Detection, presented at Canadian Vission Conference, 2008

[18] N.A. Thacker, J.V. Manjon, A Statistical Interpretation of Non-Local Means, Proc. VIE 2008, Xi An, China, 250-255, 2008

[19] Hossein Ragheb, N.A. Thacker, P.A. Bromiley, D.Tautz, A.C. Schunke, Quantitative Shape Analysis with Weighted Covariance Estimates for Increased Statistical Efficiency, Frontiers in Zoology, 10(16), April, doi:10.1186/1742-9994-10-16, 2013

Why Leverhulme?

Our proposal is of a highly interdisciplinary nature, combining techniques from imaging science, statistics, pattern recognition, planetary science, chemistry and spectroscopy. In addition our work program does not deliver scientific outputs in a tightly focussed area. Our proposal therefore does not fit well with alternative disciplinary-focused research funding bodies.

Our quantitative measurement-oriented approach is motivated by our understanding of the limitations of current methods, yet is consistent with conventional methods found within the physical sciences (e.g. the use of statistical methods in particle physics). However, they run contrary to standard pattern recognition methodologies, marking a departure from traditions established within that field, and risking misunderstanding by current pattern recognition experts.

If successful, our methods will provide an enabling technique permitting a large range of previously ill-defined analysis problems to be tackled in a fully quantitative way. However, the immediate applications are largely scientific in nature and not applicable to industry (at this point in time), which also does not fit well with industry-focused funding sources.

There are immediate benefits of our work, in part facilitating the quantitative utilisation of existing data (i.e. Moon Zoo) that has been gathered with help from previous Leverhulme funds.

The quantitative approaches advocated are a natural extension of the PI's personal vision, who has successfully applied a quantitative ethos in a variety of projects, including biometric shape analysis, medical research and 3D vision.