Autumn 2006
Vol 9, No 1
ISSN 1465-4091
£12.50

The Specialist Group on
Artificial Intelligence

EXPERT UPDATE

SGAI

Special Issue on the 2nd UK KDD Workshop

BCS
THE BRITISH COMPUTER SOCIETY

EXPERT UPDATE
(ISSN 1465-4091)

SGAI OFFICERS AND COMMITTEE MEMBERS

CHAIRMAN
Prof. Max Bramer, University of Portsmouth
Technical Programme Chair AI-2006,
max.bramer@port.ac.uk

MEMBERSHIP OFFICER
Dr. Miltos Petridis, University of Greenwich
UK CBR Organiser, AI-2006

Richard Ellis, Stratum Management Ltd
Applications Programme Chair AI-2006
NCAF Liaison

Dr Tony Allen, Nottingham Trent University
Deputy Applications Programme Chair AI-2006

Dr Alun Preece, University of Aberdeen
Deputy Conference Chair (Electronic Services) AI-2006

Alice Kerly, University of Birmingham
Research Student Liaison AI-2006

TREASURER
Rosemary Gilligan,
University of Hertfordshire
Research Student Liaison AI-2006
r.e.gilligan@herts.ac.uk

Dr Andrew Tuson, City University
Conference Chair AI-2006
Schools Liaison

Dr Frans Coenen, University of Liverpool
Deputy Technical Programme Chair AI-2005,
Deputy Conference Chair (Local Arrangements) AI-2006

Dr Nirmalie Wiratunga, Robert Gordon University
Poster Session Organiser AI-2006
Expert Update Editor

Prof. Adrian Hopgood, Nottingham Trent University
Tutorial/Workshop Organiser, AI-2006

Allan Smith, Department of Trade and Industry
Representing DTI

HONARARY MEMBERS

Maurice Ashill, Prof Ann Macintosh, Martin Merry, Prof Donald Michie, Dr Stuart Moralee

What is Expert Update?
Expert Update (www.comp.rgu.ac.uk/staff/nw/expertUpdate.htm) is the bulletin/magazine of the SGAI, the British Computer Society's Specialist Group on AI (BCS-SGAI: www.bcs-sgai.org). The purpose of Expert Update is to foster the aims and objectives of the group by publishing news, conference reports, book reviews, conference announcements, calls for papers and articles on subjects of interest to the members. Expert Update is generally published 3 times per year by BCS-SGAI. The group's official postal address is: SGAI, The BCS, Davidson Building, 5 Southampton Street, London, WC2E 7HA.

How do I subscribe?
It is free to all SGAI members. Please visit www.bcs-sgai.org/sgai/sgai.htm for details on joining the group.

How do I subscribe?
It is free to all SGAI members. Please visit www.bcs.sgai.org/sgai/sgai.htm for details on joining the group.

How do I contribute?
Submissions are welcome and must be made in electronic format sent to the editor or sub-editor.

Who owns Copyright?
All non-reprinted material in Expert Update is the intellectual and literary property of the author(s). Original articles are unrefereed (unless otherwise stated), and are not copyrighted by the BCS or SGAI. Permission to reprint an article must be obtained from the author(s). All opinions are those of the authors and do not necessarily reflect the position of the SGAI or the BCS.

EDITORIAL

Welcome to Expert Update's autumn 2006 issue.

BCS-SGAI members will be pleased to hear that selected back copies of Expert Update is now downloadable from the new SGAI member's page area accessible from the BCS-SGAI home page (www.bcs-sgai.org). Here you will also have access to the ECCAI journal, AI Communications and other AI related information. This area is accessible with the password "peterhouse". We would appreciate any feedback on this new facility and any further items for inclusion.

In this issue of Expert Update we have included papers presented at the 2nd UK Knowledge Discovery and Data mining Symposium (UKKDD'06) earlier on this year. UKKDD'06 was organised by Dr George Smith from the University of East Anglia, Dr Frans Coenen from the University of Liverpool (also BCS-SGAI) and Dr Alex Freitas from the University of Kent. BCS-SGAI once again sponsored this year's event and is pleased to be able to sponsor the event again in 2007.

The seven papers in this issue cover both theoretical and application issues in KDD. Technology topics range from unsupervised algorithms to fuzzy reasoning to Bayesian learning applied to a variety of interesting application problems such as credit card fraud detection to fraudulent telephone usage detection to genetic structure discovery. Finally the paper by Alex Freitas argues the case for evaluation measures that go beyond learner accuracy.

Nirmalie Wiratunga
Editor Expert Update
sgai-newsletter@bcs.org.uk

Max Bramer
Richard Forsyth
John Nealon
Frans Coenen
Editors Emeriti

Clustering in Metric Spaces for the KDD Practitioner

V. J. Rayward-Smith
School of Computing Sciences, University of East Anglia, Norwich

Abstract: Approaches for clustering records in real-world databases are discussed. Particular attention is paid to defining "similarity" when the fields are correlated and to the problem then posed by databases where fields are of differing types.

1 Introduction

Clustering is one of the most widely used techniques in Knowledge Discovery in Databases (KDD) but it is arguably one of the most difficult to accomplish well. In *non-hierarchical* clustering, the database is partitioned into separate sets of similar records; in *hierarchical* clustering, there are multiple levels of decomposition resulting in a tree structure with the database at the root and, at each level, a set of records being partitioned into further subsets. This paper only addresses non-hierarchical clustering. In *partitional*, non-hierarchical clustering, the clusters form a partition of the database, D, in the sense that each record belongs to exactly one cluster. This *strict* definition of a partition is relaxed in *fuzzy clustering* where each record is assigned a fractional degree of membership to each cluster. The focus here is on strict partitioning since most KDD work has been in this area.

Databases contain data with different characteristics and it is usual to classify data into one of two types. Real-valued data contains real numbers and commonly arises from measurements. However, in many cases, the data is not real-valued but is *categorical*; values are drawn from a domain comprising a finite set of possible values. Categorical data can either be *nominal* or *ordinal*. It is called nominal iff there is no assumed ordering between the elements of the domain. Thus EYE-COLOUR with domain $\{brown, blue, green\}$ is an example of nominal data whilst DEGREE-CLASS, perhaps with domain $\{pass, 3rd, 2(ii), 2(i), 1st\}$, is ordinal because there is a clear ordering of the elements of the domain. When clustering, it is advisable to replace ordinal data by a real-valued encoding that reflects the relative distances between successive values. This should be done by a domain expert. For example, $\{pass, 3rd, 2(ii), 2(i), 1st\}$ might be encoded as $\{37, 45, 55, 65, 78\}$ by an academic with experience of marking ranges. The ordinal data can then be processed as if it were real-valued and this considerably simplifies the clustering process.

2 Metric Spaces

One of the problems in clustering is to determine some notion of "similarity". This can be defined in terms of a distance metric, $d : D \times D \to R^+$, satisfying the three properties:
1. for all $x, y \in D$, $d(x, y) = 0$ iff $x = y$,
2. (symmetry) for all $x, y \in D$, $d(x, y) = d(y, x)$,
3. (triangle inequality) for all $x, y, z \in D$, $d(x, y) \leq d(x, z) + d(z, y)$.

A *pseudometric* is sometimes used where $d(x, y) = 0$ may not imply $x = y$ but all the remaining conditions of a metric are still satisfied. Some researchers have even used measures that do not satisfy the symmetric property and/or the triangle inequality. Even defining a suitable metric is not an easy task; for real-valued data, there is an infinite number of possible metrics between which to choose.

Consider a database of n records and assume each field k, $1 \leq k \leq m$, is real-valued. The ith record corresponds to a row vector, x_i, of reals where $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})$. There are many ways of defining a metric between x_i and x_j, the most obvious being the *Euclidean* metric, $\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\ldots+(x_{im}-x_{jm})^2}$.

A scaled Euclidean distance uses real values, $\lambda_i \geq 0$, $1 \leq i \leq m$, to scale the successive fields so that the distance function becomes $\sqrt{\lambda_1(x_{i1}-x_{j1})^2+\ldots+\lambda_m(x_{im}-x_{jm})^2}$. Note that the Euclidean metric is a special cases of the *Minkowski* or L metric defined by $\sqrt[\lambda]{|x_{i1}-x_{j1}|^\lambda+|x_{i2}-x_{j2}|^\lambda+\ldots+|x_{im}-x_{jm}|^\lambda}$, where $\lambda > 0$. The Euclidean corresponds to the case where $\lambda = 2$ and the Manhattan to $\lambda = 1$.

Naïve use of the Euclidean metric is dangerous. The first consideration is that of scale. Let F_i denote the ith field with $n = |D|$ real values $x_{1i}, x_{2i}, \ldots, x_{ni}$. Then we define $max(F_i) = \max\{x_{ji} \mid 1 \leq j \leq n\}$ and $min(F_i) = \min\{x_{ji} \mid 1 \leq j \leq n\}$. If the ranges of different fields, $|max(F_i)-min(F_i)|$, vary then the use of Euclidean distance may not give an expected result. Scaling each field should be considered, using perhaps

$$x \mapsto x' = \frac{x}{max(F)-min(F)}$$

so that the range is 1. However, using such a scaling in the presence of outliers can cause problems – beware of the use of 999 to denote a missing value!

Missing values are generally a problem and can be handled using various strategies. Fields or records containing missing values may be removed or techniques can be used to try to complete missing values. Alternatively, missing values can be suitably flagged and the clustering algorithms adapted to handle missing data.

Another concern that is more subtle is that of correlation. If care is not taken, correlated fields can cause some concepts to have an overly large influence on the clustering. The covariance of two features measures their tendency to vary together. Let μ_i be the mean of the feature i values, and μ_j be the mean of the feature j values. Then the *covariance* of feature i and feature j is defined by

$$S_{ij} = \frac{\sum_{k=1}^{n}(x_{ki}-\mu_i)(x_{kj}-\mu_j)}{n}.$$

The *variance* of the sequence of real values $x^i = (x_{1i}, x_{2i}, \ldots, x_{ni})$ associated with field i is defined as

$$S_{ii} = \frac{\sum_{j=1}^{n}(x_{ji}-\mu_i)^2}{n}.$$

The square root of the variance is known as the *standard deviation* and is denoted by σ_i. For large n, S_{ij} and S_{ii} are not changed much if the divisor is $n-1$ instead of n and the use of both values is to be found in the literature. The correlation between two real sequences, x^i and x^j corresponding to fields i and j is then defined by

$$corr(x^i, x^j) = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}} = \frac{S_{ij}}{\sigma_i\sigma_j}.$$

The Mahalanobis metric, denoted by $d_{\mathcal{M}}$, can be used in clustering to overcome the problem of correlated fields. This measure is given by: $d_{\mathcal{M}}(x,y) = \sqrt{(x-y)\mathcal{S}^{-1}(x-y)^T}$, where $\mathcal{S} = [S_{ij}]$ is the matrix of covariance values.

In Principal Component Analysis (PCA), linear combinations of the fields are constructed that are mutually uncorrelated. These linear combinations are simply computed in $O(m^3)$ time using the eigenvectors of the original covariance matrix. The linear combinations are known as factors or components and the principal components are those corresponding to the largest eigenvalues and thus are those explaining most of the variance in the original data. Say a component, C_j, is of the form $C_j = a_{j1}Field_1 + a_{j2}Field_2 + \ldots + a_{jm}Field_m$ and has eigenvalue, λ_j. Let a record in the database be denoted by $x = (x_1, x_2, \ldots, x_m)$. Then we can construct a new set of data $x' = (x'_1, x'_2, \ldots, x'_m)$ where $x'_j = a_{j1}x_1 + a_{j2}x_2 + \ldots + a_{jm}x_m$. This new data can be safely clustered using the Euclidean metric. Moreover, if each field, C_j, is scaled by dividing by $\sqrt{\lambda_j}$, the records x' and y' will appear in the same cluster iff x and y appear together in the clustering produced by the Mahalanobis metric on the original data.

When the data is of nominal or mixed type, alternative approaches to defining a metric are required. One technique is to replace a field with nominal data taking p distinct values by p fields of binary data. Thus, for example, the field EYE-COLOUR can be replaced by three binary fields, IS-BLUE, IS-BROWN, and IS-GREEN. The binary data can then be treated as numeric data. Such an approach can lead to an exponential growth in the number of fields with consequent implications on the efficiency of clustering algorithms.

Let d_i denote a metric defined on the ith field. If the field has real values, the metric on two values, u and v, is normally $|u - v|$. However, the metric on two nominal values, u and v, is given by the 0/1 metric, i.e. it is defined to be 0 if $u = v$ and 1, otherwise. Then the Euclidean metric (and likewise many others) can be generalised for mixed data by defining

$$d(x_i, x_j) = \sqrt{d_1^2(x_{i1}, x_{j1}) + d_2^2(x_{i2}, x_{j2}) + \ldots + d_m^2(x_{im}, x_{jm})}.$$

When using this approach, scaling should be considered so that for all $0 \leq k \leq m$, $0 \leq d_k(x_{ik}, x_{jk}) \leq 1$. Note that using this metric is equivalent to using the Euclidean metric where each field of nominal data is replaced by the necessary number of Boolean-valued fields. However, it is not so easy to generalise the Mahalanobis metric since it is not clear how to define the covariance between mixed data types. Thus, if the Mahalanobis metric is to be used, replacing nominal data by Booleans remains the recommended approach.

Let C be a set of records over a metric space, (M, d). Then the *centroid* of C is defined to be the $x \in M$ that minimises $\sum\{d(x, r) \mid r \in C\}$. The *medoid* is defined similarly but there is an added requirement that the medoid must itself be an element of C. If $M = R$ and $d(x, y) = |x - y|$ then the centroid is simply the mean of the values in C. If M is a finite domain of categorical values and d is the 0/1 metric then the centroid is the mode of C, i.e. the most commonly occurring value. Thus, for a set of records in a mixed database, if the extended Euclidean measure is used, the centroid will have components that are either the mean value of a real-valued field or the modal value of a categorical-valued field.

3 Clustering in a Metric Space

Even when the distance metric is determined, there remains the difficulty of determining a measure to define a good cluster. What exactly do we wish to optimise? Intuitively, a good clustering will have records that are close to one another allocated to the same cluster and those that are far from one another allocated to different clusters. There may also be a requirement that clusters are well separated and the total number of clusters to be used is also pertinent. Given any clustering $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$, a function, $f : \mathcal{C} \to R^+$ is required that determines the worth or fitness of the clustering. There are many alternative measures of cluster fitness but most researchers simply use the within-cluster, sum-of-squares fitness measure

$$\Sigma_{i=1}^k \Sigma_{x,y \in C_i} d^2(x, y).$$

Alternative measures might use sum of distances (possibly squared) from the cluster centroid or from its medoid.

Whilst intra-cluster measures must be minimised, inter-cluster measures must be maximised. The distance between two clusters, C_i and C_j, can be measured in many different ways including

1. $d_1(C_i, C_j) = d(\mu_i, \mu_j)$, where μ_i, μ_j denote the medoids of C_i, C_j, respectively,
2. $d_2(C_i, C_j) = d(c_i, c_j)$, where c_i, c_j denote the centroids of C_i, C_j, respectively,
3. $d_3(C_i, C_j) = \min\{d(x, y) \mid x \in C_i, y \in C_j\}$,
4. $d_4(C_i, C_j) = \Sigma\{d(x, y) \mid x \in C_i, y \in C_j\}$.

Once $d(C_i, C_j)$ is determined, we might define an inter-cluster measure using

$$\sum\{d(C_i, C_j) \mid 1 \leq i < j \leq k\}$$

or

$$\sum\{d^2(C_i, C_j) \mid 1 \leq i < j \leq k\}$$

perhaps dividing each of the summed terms by $|C_i| \times |C_j|$ and the total by $k(k-1)/2$.

In general, clustering is a multi-objective optimisation problem. However, in some cases, minimising one intra-cluster measure will automatically maximise an inter-cluster measure. For example, if our inter-cluster measure is simply

$$\sum\{d^2(x, y) \mid x \text{ and } y \text{ are in different clusters}\}$$

then minimising the sum-of-squares intra-cluster measure will necessarily maximise this inter-cluster measure since their sum is a constant.

3.1 Clustering Algorithms

Heuristic algorithms are widely used for clustering, e.g. the well known k-means algorithm and its variants, and Partitioning Around Medoids (PAM) [7] and its variant CLARANS [12]. These algorithms greedily search for a local optimum. Metaheuristics offer the opportunity to find better quality solutions but generally at a computational cost. A survey on the use of metaheuristics for clustering in KDD is given in [13]. When undertaking comparative studies of clustering algorithms, it is essential to determine what objective is being optimised and to compare like with like. In practice, databases tend to be large and, even if sampling is used, time constraints often mean that a fast algorithm is essential.

Most algorithms can be generalised to work with a variety of metrics although this may impact on their efficiency as well as on their effectiveness. As an example, we will consider a version of the familiar k-means algorithm and assume that we have mixed mode data and that the (extended) Euclidean metric is used. The resulting algorithm aims to minimise the sum of squares intra-cluster measure and can be called the *k-centroid algorithm.* A cluster is represented by its centroid and the number of clusters, k, is set in advance. The algorithm can be simply described as follows:

1. Choose k distinct records as centroids. Let each centroid represent a cluster.
2. For each record in the database, assign the record to the cluster represented by the closest centroid.
3. Update the values of the centroids.
4. If during the execution of steps 2 and 3 the centroids have changed their values then repeat from step 2; otherwise halt.

To implement this algorithm efficiently, associated with each cluster will be the following data:-

1. the value of the centroid,
2. the number of records in the cluster,
3. for each nominal field, F, and, for each value, u, of that field, the number of records in the cluster with F-value equal to u.

The average values of real-valued fields are easily updated from the old mean, the number of records and the corresponding value of a newly added record. The update of the centroid's value for any nominal field requires more data to be available. Note that the centroids, although initialised as records, may not correspond to any records at the termination of the algorithm. The k-medoids algorithm insists that cluster centres are themselves records and, although there are significant overheads in calculating medoids, the algorithm is less prone to the impact of outliers.

The direct application of metaheuristics to cluster databases is really only effective on small data sets. This is because the representation is too large and the search space becomes enormous. Added to this, the neighbourhood function or the crossover/ fitness function is sometimes expensive to compute. A more promising approach to clustering large databases using a metaheuristic is to hybridise a metaheuristic with a heuristic clustering algorithm, very often k-means, see e.g. [6, 11]. Metaheuristics might be used to initialise the centroids or a single iteration of the k-means algorithm might be used as an operation within the metaheuristic.

4 Conceptual Clustering

Even if a good clustering can be found according to agreed fitness measures, there will remain issues concerning the presentation of the clusters to the database owner and with their exploitation within an organisation. Simply reporting the centroids or medoids together with the maximum (and/or average) distance of records in each cluster from the centroid or medoid will seldom be adequate, especially since the metric used to measure this distance might itself be quite complex. For management to be excited by, and to develop strategies for using, the clustering, simple descriptions of each cluster will be required. *Conceptual clustering*, see e.g. [3] refers to techniques that are focussed on finding cluster descriptions. Since giving good descriptions is of such importance, it seems sensible to make this the primary target of a search. Instead of looking for k cluster centres and using metrics, the search should be for k simple tests, so that a large percentage of the database satisfies, or nearly satisfies, just one (or occasionally more) of these tests. The clusters defined by a test are then the records that satisfy that test. If a significant percentage of the database falls into these three clusters and each of these clusters are dense then this will be important information to the Company that can be easily understood and exploited. In a recent paper, Mishra et al. [10] use the term *conjunctive* clustering to refer to conceptual clustering where each test is simply a conjunction of simple attribute tests. To solve this problem, they formulate the problem in graph theoretic terms and develop and analyse an effective heuristic.

Metaheuristics have been used most successfully in finding rules in databases, either in the form of high quality individual rules (or *nuggets*), see e.g. [4, 5], or as sets of rules for classification, see e.g. [9]. Techniques used in finding such rules can be modified to seek tests defining clusters. In the search for a set of tests, any correlation of fields need not be an issue although multiple equivalent descriptions of a given cluster will complicate the search. A fitness measure will need to measure the density of the clusters using the sum-of-squares measure based on some metric, perhaps the Mahalanobis metric. The percentage of the database described will also need to be part of the fitness measure together with rewards for simplicity of description and penalties for overlapping clusters. Fuzzy rules with associated partial set membership can also be exploited to find good conceptual clusters. In conceptual clustering, as indeed with clustering generally, there is plenty of scope for using metaheuristics not just in the search for clusters but also in the search for new fields (perhaps constructed using simple arithmetic expressions over existing fields). de Jong et al. [1] used a GA for this task, which they refer to as *concept learning.* More recently, genetic programming has been the favoured approach, see e.g. [2, 8].

5 Conclusions

We have discussed the all-important role of "similarity" in clustering and discussed the design of suitable metrics to apply to a database. The choice of metric is crucial and, once determined, there are a wide variety of clustering algorithms available. Clustering is essentially a multi-objective optimisation problem; we seek a clustering that minimises intra-cluster variation and maximises inter-cluster variation. When comparing algorithms, it is important to understand what measures they are seeking to optimise. Even though clustering is a favourite KDD activity, it is beset by difficulties and is often done poorly. Presentation of the clustering to the database owner is also often a problem and, if simple descriptions of the clusters are of primary concern, then conceptual clustering is to be preferred.

References

[1] K. A. deJong, W. M. Spears, and D. F. Gordon. Using genetic algorithms for concept learning. *Machine Learning*, 13:161–188, 1993.

[2] F. E. B. Otero, M. M. S. Silva, A. A. Freitas, and J. C. Nievola. Genetic programming for attribute construction in data mining. In *Proceedings 6th European Conference on Genetic Programming, Lecture Notes in Computer Science, 2610*, pages 384–393, 2003.

[3] D. Fisher. Data mining tasks and methods: Clustering: conceptual clustering. In W. Klösgen and J . M. Żytkow, editors, *Handbook of Data Mining and Knowledge Discovery*, pages 388–396. Oxford University Press, Inc., 2002.

[4] A. A. Freitas. A genetic algorithm for generalized rule induction. In R. Roy, T. Furuhashi, and P. K. Chawdhry, editors, *Advances in Soft Computing – Engineering Design and Manufacturing (Proc WSC3 3rd On-line World Conf. hosted on the internet)*, pages 340–353, Berlin, 1999. Springer-Verlag.

[5] B. de la Iglesia, M. S. Philpott, A. J. Bagnall, and V. J. Rayward-Smith. Data mining rules using multi-objective evolutionary algorithms. In R. Sarker et al. editor, *Proceedings of the 2003 Congress on Evolutionary Computation*, pages 1552–1559. IEEE Press, Piscataway, NJ, 2003.

[6] P. M. Kanade and L. O. Hall. Fuzzy ants clustering with centroids. In *Proceedings of the International Conference on Fuzzy Systems.* IEEE, 2004.

[7] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley, New York, 1990.

[8] K. Krawiec. Genetic Programming-based construction of features for machine learning and knowledge discovery tasks. *Genetic Programming and Evolvable Machines*, 3(4):329–343, 2002.

[9] R. E. Marmelstein and G. B. Lamont. Pattern classification using a hybrid genetic program decision tree approach. In J. R. Koza et al. editor, *Genetic Programming 1998: Proceedings of the Third Annual Conference*, pages 223–231. Morgan Kaufmann, 1998.

[10] N. Mishra, D. Ron, and R. Swaminathan. A new conceptual clustering framework. *Machine Learning*, 56(1-3):115–151, 2004.

[11] N. Monmarche. On data clustering with artificial ants. In A. A. Freitas, editor, *Data Mining with Evolutionary Algorithms: Research Directions*, pages 23–26, Orlando, Florida, 1999. AAAI Press.

[12] R. T. Ng and J. Han. Efficient and effective clustering methods for spatial data mining. In *VLDB '94: Proceedings of the 20th International Conference on Very Large Data Bases*, pages 144–155. Morgan Kaufmann, 1994.

[13] V. J. Rayward-Smith. Metaheuristics for clustering in KDD. In *Proceedings of the 2005 IEEE Congress on Evolutionary Computation (CEC2005).* IEEE Press, 2005.

Multiobjective approaches to Unsupervised Classification

Julia Handl and Joshua Knowles
Manchester Interdisciplinary Biocentre, University of Manchester, UK

Abstract: Three problems in exploratory data analysis, clustering, feature selection and semi-supervision, are considered in this paper. We discuss how multiobjective optimization provides a flexible means to overcome some of the fundamental difficulties that arise in each of these problems, and refer to recent experimental work that has demonstrated its practical performance benefits compared to traditional single-objective approaches.

1 Exploratory data analysis

Existing methods for exploratory data analysis differ along a number of different dimensions, one of the most fundamental of which is the distinction between unsupervised and supervised learning techniques. Supervised learning refers to learning in the presence of training examples—in classification, a set of data samples for which the correct classification is known. If a sufficient amount of such data is available, a classifier can be trained in order to learn and correctly predict the class memberships of these data items in the hope that the trained classifier subsequently generalizes to the classification of new unlabelled data. Supervised methods can be very powerful for the classification of complex data, but may suffer from problems related to overtraining, resulting in poor generalization capabilities.

In contrast to supervised learning, unsupervised learning can be applied in the absence of any prior knowledge about the number of classes, or any correct training examples. It relies on the assumption that the main class structure of the data is reflected by the actual distribution of the data, that is, that clusters of homogeneous data items can be identified and that this grouping will lead to a meaningful classification. Evidently, unsupervised approaches are prone to fail if no distinct cluster structure is present in the data, and they are, in this sense, less powerful than supervised methods. However, their positive aspects include the facts that, in contrast to supervised approaches, they can be used for exploratory data analysis in scenarios where little prior information is given and that overtraining is not an issue.

In our recent work [5, 6, 7, 8, 9], we have developed multiobjective approaches to three of the most fundamental problems met in exploratory data analysis, namely clustering, feature selection and semi-supervision. In all three applications we were able to demonstrate distinct advantages of the multiobjective approaches, and we will summarize this work in this paper.

2 Clustering

Most respected texts on data clustering define the problem informally, e.g. Arabie *et al.* [1] define clustering as "Those methods concerned in some way with the identification of homogeneous groups of objects". The use of informal definitions such as these reflects one of the prevailing and fundamental problems in cluster analysis: the difficulty of providing a single formal (but sufficiently broad) definition of clustering and of the concept of a cluster. This is because the concept of a cluster is a generalization (to arbitrary dimensions) of what humans perceive, in two or three dimensions, as densely connected 'patches' or 'clouds' within data space, a human intuition which is inherently difficult to capture by means of individual objective criteria.

Consequently, different formulations of the clustering problem vary in the optimization criterion used, which capture different aspects of the properties of a good clusters such as the compactness of clusters, their local connectedness or the spatial separation between them. Importantly, though, most existing clustering methods attempt, explicitly or otherwise, to optimize just *one* such criterion, and it is this confinement to a particular cluster property that explains the fundamental discrepancies observable between the solutions produced by different algorithms on the same data, and will cause a clustering method to fail in a context where the criterion employed is inappropriate.

2.1 Motivation of a multiobjective approach

In practice, the problem of choosing an appropriate clustering objective (viz. algorithm) can be alleviated through the application and comparison of multiple clustering methods [10], or through the *a posteriori* combination of different clustering results by means of ensemble methods [14, 15]. However, a more principled approach may be the consideration of clustering as a multiobjective optimization problem, as suggested in [4].

The set of Pareto-optimal solutions to a multiobjective clustering problem (where a partitioning is optimized with respect to a set of objective criteria $\{P_1, P_2, \ldots, P_m\}$) always comprises the optimal solutions to the single-objective clustering problems (where a partitioning is optimized with respect to a single objective $P \in \{P_1, P_2, \ldots, P_m\}$). For *ideal* single- and multiobjective clustering algorithms (i.e. algorithms that always identify all globally optimal solutions, and the entire Pareto-optimal set, respectively), we therefore trivially know that the multiobjective algorithm will *always* find a solution *as good or better* (equal in terms of the clustering objective optimized and equal or *possibly* better in terms of external knowledge) than those of the single-objective algorithms. In situations where the best solution corresponds to a trade-off between the different objectives *only the multiobjective clustering algorithm* will be able to find it. An example of a data set, for which this is relevant is shown in Figure 1. Here, different possible clustering solutions are plotted in two-objective space and it can be seen that the correct solution represents a trade-off between the two objectives.

2.2 Implementation and results

In [5, 7], we have described a multiobjective evolutionary algorithm (MOEA) for clustering, MOCK (Multiobjective clustering with automatic k-determination). MOCK optimizes two, conceptually different clustering criteria and explores solutions across a range of different numbers of clusters. It has been compared to various single-objective clustering methods, ensemble techniques and validation techniques, and results indicate a clear advantage to the multiobjective approach.

3 Feature selection

Feature selection, or subset selection, is a process commonly used for dimensionality reduction in machine learning. It refers to the process of obtaining a lower-dimensional projection of a data set by selecting a subset of the original features and discarding the remaining ones. This has the advantage that all features in the reduced feature space correspond to a single feature in the original high-dimensional feature space, and are, therefore, directly interpretable. Generally, dimensionality reduction can be crucial in learning tasks for a number of reasons. First, for large feature sets, the processing of all available features may be computationally infeasible. Second, many of the available features may be redundant, noise-dominated or irrelevant to the classification task at hand. Consequently, the inclusion of all features will be detrimental and the subset most relevant for the learning task at hand needs to be identified. Third, high-dimensionality is also a problem if the number of variables is much larger than the number of data points available. In such a scenario, dimensionality reduction

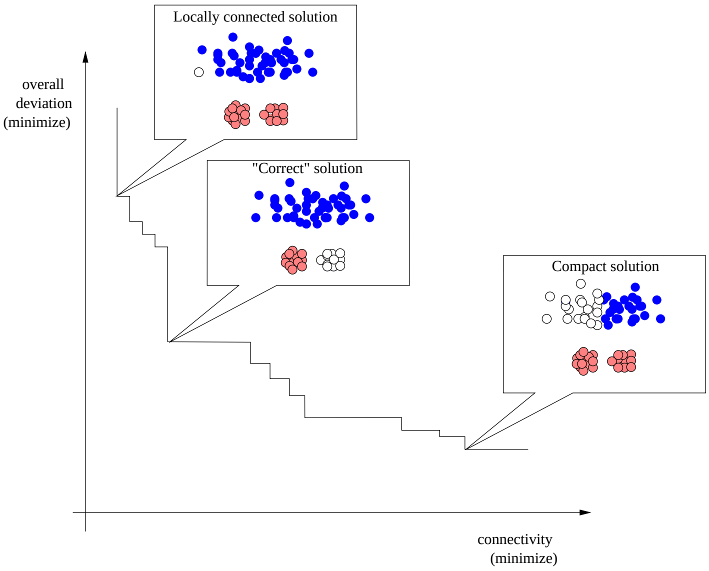


Figure 1: The correct clustering solution often corresponds to a trade-off between two or more clustering objectives. A Pareto front (depicted as a line) and three different Pareto optimal clustering solutions are shown for a simple three-cluster data set, plotted in two-objective space. The solution to the top left is generated by an algorithm like single link agglomerative clustering, which minimizes local connectedness of clusters only. The solution to the bottom right is generated by an algorithm like k-means, which minimizes cluster compactness only. The correct solution is situated some way between these two solutions and so it would not usually be discovered by either method. A multiobjective approach considering the trade-offs between the two objectives should be able to access this solution much more readily. For sake of clarity, the approximation set in this example only contains solutions for $k = 3$. More generally, the number of clusters can also be kept dynamic — in this case an approximation set is obtained in which the number of clusters varies along the Pareto front.

is crucial in order to overcome the curse of dimensionality [2] and allow for meaningful data analysis. The problem has been well-studied in the supervised scenario but only little research to date has dealt with the unsupervised problem (for a recent overview of research efforts in both areas, see [12]). Yet, several of the challenges faced in the unsupervised problem are very different to those encountered in a supervised scenario: in particular the assessment of the quality of an individual feature or a feature subset becomes even more intricate in unsupervised classification.

3.1 Motivation of a multiobjective approach

Methods for unsupervised feature selection rely on the use of an unsupervised measure to assess the quality of a given feature subspace. This can either be done through the combination of a clustering algorithm and an internal cluster validation technique (in a so-called ‘wrapper approach’), or using a measure such as entropy that can directly assess the degree of structure present in a given feature subspace (a so-called ‘filter approach’).

Unfortunately, all available unsupervised measures are based on some form of distance computation in feature space and this is problematic for their use in feature selection, as it automatically induces a bias of these measures with respect to the dimensionality of the feature space. The existence of this bias is related to the fact that, when moving to high dimensions, the histogram of distances between items in data space changes: the mean of

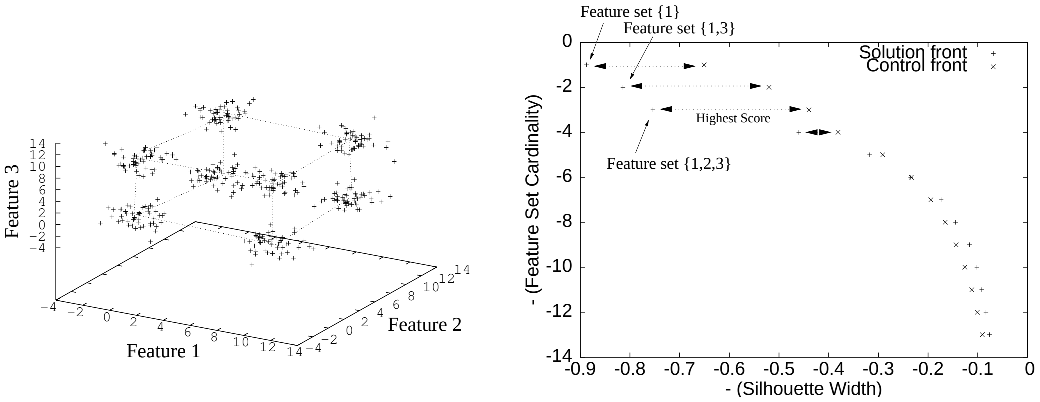


Figure 2: (Left) Plot of Square3d, a 13-feature data set, containing eight clusters arranged in a cube pattern in the first three dimensions, and Gaussian noise in the remaining 10 dimensions. (Right) Pareto front obtained on this data and a random control distribution. The distance between the solution and the control point obtained for a given feature cardinality can serve as an indicator of quality.

the histogram tends to increase and the variance of the histogram tends to decrease. In other words, the distances between all pairs of points tend to become highly similar and (dependent on the specific form of the unsupervised measure used) this causes a bias to low or high dimensions. If this natural dimensionality-bias is not accounted for, an unsupervised feature selection method will always favour extreme feature spaces (i.e. the lowest- or highest-dimensional feature spaces available).

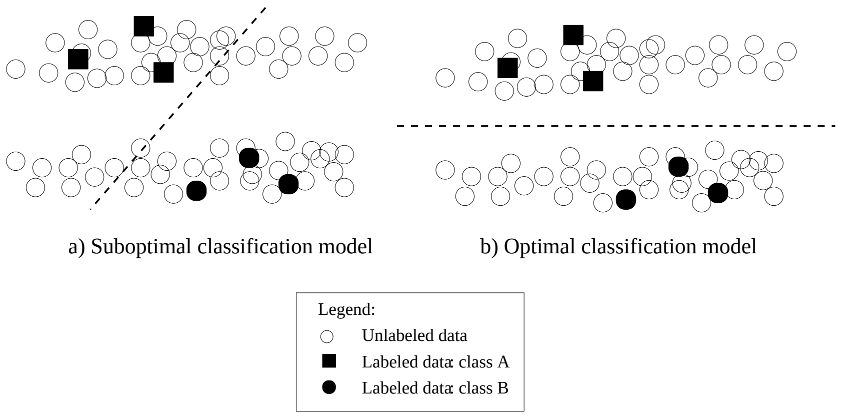
In the literature, three different approaches have been proposed to tackle the issue of bias. The first approach is a simple ad-hoc normalization of the evaluation function by means of an appropriate scaling factor (usually expected to be a function of the feature cardinality) [3, 11]. This type of normalization may reduce the bias or overcompensate for it, but will not usually remove it cleanly. An alternative approach is the cross-projection technique proposed by Dy and Brodley [3], which attempts to reduce the cardinality-specific bias by comparing pairs of clustering solutions in both of the associated subspaces. This relation can be used for pairwise comparisons between features sets but it is not transitive, which makes its use in global optimization techniques problematic. Finally, two recent papers [11, 13] have suggested dealing with the bias by considering feature cardinality as a separate objective and applying a Pareto multiobjective optimization algorithm.

3.2 Implementation and results

In [8], we investigated a multiobjective formulation of unsupervised feature selection, and found it to be an effective method of dealing with the inherent cardinality-bias of unsupervised measures for the evaluation of feature subspaces. A good performance of the resulting algorithm was demonstrated in a comparison to a range of alternative approaches.

4 Semi-supervision

In certain classification scenarios it can be advantageous to combine the advantages of both unsupervised and supervised classification techniques, that is, to exploit both previous knowledge of class labels and the underlying data distribution: semi-supervised approaches aim to do this. Through the combined use of labelled and unlabelled data it becomes possible to give a degree of external guidance to the classification algorithm, while still permitting intrinsic structure in the data to be taken into account. This is considered particularly useful when dealing with data sets consisting of a large number of unlabelled data items and relatively few labelled ones, and, more generally, in the case of very limited prior knowledge. For example, in cases where the classes within a particular data set are only partially known, additional ones may be identified by taking the data distribution into account (see Figure 3). Also, due to the combination of two fundamentally different sources of information, semi-supervised approaches would be expected to be more robust than both unsupervised and supervised approaches, and may be less sensitive towards both annotation errors and the occlusion of structures in the data due to noise. Data sets with the above properties are frequently encountered in application domains where the categorization of individual data items is accompanied by high computational, analytical or experimental costs, such as in bioinformatics or in text-mining.



a) Suboptimal classification model

b) Optimal classification model

Figure 3: Illustration of the fundamental idea behind semi-supervision. The unlabelled data points can help to avoid suboptimal solutions and identify the classification model that is optimal with respect to the given data.

4.1 Motivation of a multiobjective approach

To date, no thorough analysis of the advantages and disadvantages of the different existing methods of semi-supervised clustering is available, but a number of observations can be readily made. In particular, when integrating unsupervised and supervised information by means of a distance or an objective function, it is not usually clear what the best weighting between these components will be. It is possible that the weighting chosen may have a significant effect on the final outcome, and may cause an algorithm to be sensitive to small annotation errors.

Tackling the semi-supervised clustering problem within the framework of multiobjective optimization may provide a more flexible framework for the integration of both unsupervised and supervised components into the clustering process. Specifically, the use of Pareto optimization provides the means to avoid the need for a fixed weighting between unsupervised and supervised objectives. Consequently, we would expect a multiobjective approach to semi-supervised clustering to perform more consistently across different data sets, and to show a higher robustness towards annotation errors.

4.2 Implementation and results

In [6, 9] we have described multiobjective approaches to semi-supervised clustering and semi-supervised feature selection. In both applications, we were able to observe clear performance advantages of the multiobjective approaches. In particular, the algorithms proposed were observed to outperform purely unsupervised and supervised methods, as well as semi-supervised methods based on the linear or non-linear combination of objectives.

5 Conclusion

In this paper, we have summarized our recent work on the use of multiobjective optimization in the context of exploratory data analysis. We have shown that the advantages of a multiobjective formulation arise as a consequence of different aspects of a problem, such as the difficulty of choosing a single objective (in clustering), the need to counter-balance a bias intrinsic to the primary objective (in feature selection) and the availability of multiple data sources (in semi-supervision).

Acknowledgements

JH acknowledges support of a doctoral scholarship from the German Academic Exchange Service (DAAD) and the Gottlieb Daimler- and Karl Benz-Foundation, Germany. JK is supported by a David Phillips Fellowship from the Biotechnology and Biological Sciences Research Council (BBSRC), UK.

References

[1] P. Arabie, L. J. Hubert, and G. De Soete. *Clustering and Classification.* World Scientific, New Jersey, NJ, 1996.

[2] R. Bellman. *Adaptive Control Processes: A Guided Tour.* Princeton University Press, Princeton, NJ, 1961.

[3] J. G. Dy and C. E. Brodley. Feature selection for unsupervised learning. *Journal of Machine Learning Research*, 5(5):845–889, 2004.

[4] A. Ferligoj and V. Batagelj. Direct multicriterion clustering. *Journal of Classification*, 9:43–61, 1992.

[5] J. Handl and J. Knowles. An evolutionary approach to multiobjective clustering. *IEEE Transactions on Evolutionary Computation*, 2006. (In press).

[6] J. Handl and J. Knowles. Feature subset selection in unsupervised learning via multiobjective optimization. *International Journal on Computational Intelligence Research*, 2006. (In press).

[7] J. Handl and J. Knowles. Multiobjectve clustering and cluster validation. In Y. Jin, editor, *Multi-Objective Machine Learning*, chapter 2. Springer-Verlag, Heidelberg, Germany, 2006.

[8] J. Handl and J. Knowles. On semi-supervised clustering via multiobjective optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference*, 2006. (In press).

[9] J. Handl and J. Knowles. Semi-supervised feature selection via multiobjective optimization. In *Proceedings of the International Joint Conference on Neural Networks*, 2006. (In press).

[10] J. Handl, J. Knowles, and D. B. Kell. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–3212, 2005.

[11] Y. Kim, W. N. Street, and F. Menczer. Evolutionary model selection in unsupervised learning. *Intelligent Data Analysis*, 6(6):531–556, 2002.

[12] H. Liu and L. Yu. Towards integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17(3):1–12, 2005.

[13] M. Morita, R. Sabourin, F. Bortolozzi, and C. Y. Suen. Unsupervised feature selection using multiobjective genetic algorithms for handwritten word recognition. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, pages 666–671. IEEE Press, New York, NY, 2003.

[14] A. Strehl and J. Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *Journal on Machine Learning Research*, 3:583–617, 2002.

[15] A. Topchy, A. K. Jain, and W. Punch. Clustering ensembles: models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1866–1881, 2005.

The Importance of being Fuzzy

Trevor Martin
Artificial Intelligence Group, Department of Engineering Mathematics
University of Bristol, Bristol BS8 1TR, UK
email: trevor.martin@bris.ac.uk

Abstract: Probability is the mathematician's tool of choice in modelling uncertainty, and is the only rational option when dealing with well-defined problems. However, human language is far more subtle and expressive than a formal model, and deals with many concepts that are defined by common usage rather than by necessary and sufficient conditions. Such vague concepts are widespread in hierarchical classification structures - for example, subject categories. We consider that fuzzy sets are generally a better approach in modelling hierarchies than artificially precise yes/no definitions.

1 Introduction

1.1 How should we deal with uncertainty ?

From a mathematical perspective, it is very difficult to argue against the use of probability as a tool for handling uncertainty. Indeed, it is relatively easy to show that if one is prepared to bet as an indication of one's level of certainty about an event then it is not rational to base one's behaviour on anything other than the laws of probability. However, it is worth remembering the words of Einstein:

> As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality ("Geometry and Experience" 1921)

The assumptions underlying probability include

- the availability of precise definitions for events and
- procedures to determine whether or not a given event has occurred.

Much human knowledge and communication is based on natural language, and one of the strengths of natural language is its capacity to efficiently convey a large amount of information relatively compactly. This relies on a shared understanding of terms, without necessarily sharing precisely the same definition of terms. For instance, an Englishman announcing that he is "travelling to Europe" would be understood to mean that he is going somewhere on the opposite side of the channel to England; an American saying the same thing would probably include the UK as a possible destination. The word "Europe" denotes a collection of countries, but its precise definition is elusive - is it a set of countries marked as Europe on a particular map, the members of the European union (now? in 1970 ? in 1975 ? in 2010 ?), the countries eligible to enter European Championship football, countries eligible to enter the Eurovision song contest .... ?

Each "source" has its own definition of the term, but we are able to understand and use the concept in communication without the inconsistency causing a problem. It is at this level that we argue for the use of fuzzy sets -to model human understanding of concepts using nested and graded sets. Mathematically we can represent a fuzzy set by a membership function on a universe of objects U

$$\mu : U \to [0, 1]$$

noting that the interval [0, 1] is a special case, and that use of a lattice of membership values is more general. Lattice based approaches allow us to work with partially ordered symbolic memberships and thus handle the intrinsic uncertainty in an elegant manner. What is needed for most purposes is an interpretation of the membership which can be used by others; a partial ordering may be more useful than precisely calculated membership values.

Without humans in the loop - in a formal mathematical system, for instance, where everything is precisely defined - there is no need to handle fuzzy uncertainty. In any application involving implicit or explicit use of "commonly understood" definitions, we argue that there is a need either to model the fuzzy uncertainty or to hide the problem by adopting artificially precise definitions. To take a simple example, using online ontologies we would interpret the notion of "*a few minutes drizzle*" as

> "2-10 minutes of uniform precipitation composed exclusively of fine drops with diameters of less than 0.02 inch (0.5 mm) very close together".

Even allowing for the fact that this does not define *uniform* or *very close*, we claim that such precision is not compatible with normal human usage. Cases of over-precise definitions abound, and giving necessary and sufficient conditions can cause confusion because humans understand the concept without necessarily sharing a common agreement in every single case.

We treat fuzzy sets using the mass assignment theory of Baldwin [2, 3], which gives a convenient bridge between fuzzy and probability - see also the pignistic probability function of Smets [18]. A mass assignment M on a set B is a distribution over the power set of B

$$M = \{B_i, m(B_i)\}$$

where

$$m(B_i) \geq 0 \text{ for all } B_i \subseteq B \text{ and } \sum_{B_i \subseteq B} m(B_i) = 1.$$

This is related to a fuzzy subset of B,

$$F = \{b_i/\mu(b_i)\}$$

by $\mu(b_i) = \sum_{b_i \in B_k} m(B_k)$. The corresponding least prejudiced distribution is calculated by

$$LPD(b_i) = \sum_{b_i \in B_k} \frac{m(B_k)}{|B_k|}$$

The mass assignment - fuzzy set transformation relies on a voting (or possible worlds) model. For example, suppose that 10 people are asked which dice values they would accept as being *large.* If two will only accept {6} as large, five will accept {5, 6} and the remaining three will accept {4, 5, 6} then we obtain a mass assignment

$$\{6\} : 0.2, \{5, 6\} : 0.5, \{4, 5, 6\} : 0.3$$

from which we can extract a fuzzy set by treating the memberships as the proportion of voters who accept a particular value as satisfying "large"

- all accept 6 so its membership is 1
- 80% accept 5, so its membership is 0.8
- 30% accept 4, so its membership is 0.3

Thus the fuzzy set *large* is {6/1, 5/0.8, 4/0.3}. The reverse process is possible, i.e. we can go from a fuzzy set to mass assignment.

1.2 Fuzzy Control

The first generation of fuzzy applications - fuzzy control - was successful partly because it allowed knowledge expressed in a limited form of natural language to be converted into control rules fairly easily. It provided a huge boost to the fuzzy community by showing clear benefits in consumer goods such as video cameras and washing machines, and in industrial processes such as cement kilns and trains. The life-cycle of the 1990s "fuzzy boom" ran through a fairly standard pattern - initial research leading to demonstrator applications, followed by wider interest and the development of fuzzy software toolkits, which then led to high profile commercial applications.

1.3 The next generation of fuzzy applications

There are a number of other areas in which fuzzy has been applied but made a smaller impact - for instance, fuzzy databases (including fuzzy conceptual models and fuzzy object-oriented databases), decision support systems and artificial intelligence (such as image understanding, learning, robotics). Generally in these areas, fuzzy has not gone beyond the first phase (initial research) mentioned above; however, the recent explosive growth in data and the need to impose structure is an opportunity to exploit the human understandable nature of fuzzy sets again.

2 Digital Obesity and the need for Fuzzy Categories

It is commonly agreed that we are facing a potentially disruptive increase in the amount of information we can store and access. Music, images, e-mails, and texts are hoarded on mobiles, cameras, laptops and PDAs, and there is a large volume of paper-based information as well as the explosive growth in web-accessible data. A report by Toshiba[1] used the term "Digital Obesity" to summarise this problem.

According to recent estimates [10], the amount of new information stored on paper, film, magnetic, and optical media increased by approximately 30% per annum in the period 1999-2002. At that time, the "surface" web was estimated to contain about 170 terabytes of data. Since then, it appears that the volume of information has continued to increase at a similar rate. The problem of information location is fundamental to the successful use of the world's information resources, and much research effort (including initiatives such as the semantic web) is intended to address this issue by enabling the use of meta-data to classify and describe information, services, etc. Depending on the degree of human interpretation and judgment required, we can define a spectrum from "understanding-free" approaches involving simple syntactic operations (e.g. keyword matching) to "understanding-rich" approaches, needing considerable human expertise.

Librarians have provided standardised solutions for many years, by analysing and categorising books within carefully designed taxonomies. Clearly this is a very understanding-rich approach, as it requires a human to read or view the material in order to judge exactly where it fits within the taxonomy and how it should be indexed for later retrieval. A user needs knowledge of the subject and the categorisation scheme in order to find relevant sources within a reasonable time.

The opposite end of the spectrum arises from the advent of electronic storage and machine-searchable content. This has led to new techniques of information retrieval based on syntactic analysis of content. Most approaches treat documents as bags of words, and queries as smaller sets of words. A document's relevance to a query can then be computed from the overlap. Probably the most common method of computing relevance is TF-IDF, which automatically

[1] news.bbc.co.uk/2/hi/technology/4079417.stm

determines the most discriminating query terms (see [16] for an introduction and [1] for a more complete discussion of classical information retrieval methods). Many intranet search engines and document retrieval systems use the vector representation with the TF-IDF technique or some variation of it. An additional 'syntactic feature' exploited by Google is the link structure of the web which can be used as a guide to page quality [4] giving rise to the PageRank measure. If we treat the web as a graph with pages as vertices and hyperlinks as edges, then the PageRank corresponds to the principal eigenvector of the normalised adjacency matrix.

Even with the outstanding success of Google, the volume of data on the web means that a typical query can retrieve millions of pages. If the desired page is not found within the first 20-30, it may not be seen at all as users do not have the time or patience to sift through large numbers of potential matches. A further problem with syntactic approaches is the possibility of misleading information - whether by concealing text to increase a page's rating in keyword searches, or by manipulating the link structure to achieve a higher ranking (so-called googlebombing).

2.1 Combining Categorisation and Syntactic Features

The ability to create hierarchical classification structures is a core part of many approaches to indexing and organising information, including the semantic web which envisages meta-data describing the content of web pages and ontologies to define "meanings" of the meta-data vocabulary and to organise the web pages hierarchically. Of course, the word "semantic" is somewhat misleading. It is not equivalent to "meaning" in the human sense of the word - instead, it refers to a mapping where we can draw links between the terms in the language and the terms in an abstract mathematical structure, in such a way that the provable properties of the mathematical structure also apply to the language.

Uschold [19] asked "Where are the Semantics in the Semantic Web?", pointing out that the machine processable semantics envisioned by Berners Lee is at the extreme of a spectrum that also ranges through formal specifications for human use, dictionary definitions and "shared understanding" where human consensus defines the meaning of terms. An example is the use of XML tags, where the tag names are used to indicate meaning without having formal definition within a schema or DTD. Most research is focused at the formal end of this spectrum - because that is where the elegant theorems are to be proved - but most available meta-data is at the informal end. Commercial interests dictate that there is advantage in using XML, as it enables electronic commerce; there is no comparable driving force for creating mark up in a form suitable for the semantic web. We can build systems from the top down (define the knowledge representation and populate it) or from the bottom up (take existing marked-up data and add intelligent processing).

Sheth et al. [17] also discusses the spectrum of semantics, ranging from implicit semantics - arising from syntactic features of data such as keyword occurrence - to formal semantics - the model theoretic approach. In between these extremes he places "powerful" or soft semantics, including probabilistic, possibilistic and fuzzy approaches. He argues strongly that the semantic web needs to move away from its almost exclusive focus on formal semantics and embrace soft semantics in the effort to build practical tools implementing intelligent web applications.

Underlying the semantic web are two key assumptions - that first order logic is an adequate framework for expressing the required knowledge, and that computationally tractable logical inference is adequate for processing that knowledge. The aim of the semantic web is to make the content both human and machine understandable. Our position is that this goal will be better served by making use of a key advantage of fuzzy sets - that they are easily interpreted by humans.

3 Fusion of Information

We treat information fusion as the integration of material retrieved from two or more sources into a single consistent answer. There are two key aspects of this problem - how can we tell when two sources refer to the same entity, and how can we compare the classifications of that entity? There is no universal hierarchy of knowledge - and hence different information sources will almost inevitably differ in the hierarchical classification scheme they adopt. To take a very simple example, the set of tracks or albums classified in one online music store as

music > rock > classic rock > 70's classics

may correspond to another's

music > rock&pop oldies.

3.1 Instance matching (the Record Linkage Problem)

This question - when are two entities in an information system the same? - is the basis of the "record linkage" problem identified by [13] and formalised by [8]. It remains a problem for information systems [7] as well as the semantic web [14]. Most information systems (explicitly or implicitly) require a unique identifier for every individual. "Instance-matching" is the process of determining that objects from different sources are the same - for example, to deduce with a reasonable degree of certainty that an author known in one database as "Lewis Carroll" represents the same individual as the author known in a second database as "C L Dodgson". The SOFT (Structured Object Fusion Toolkit) method [11] for instance matching attempts to establish an approximate relation

$$h : A \rightarrow \tilde{P}(B),$$

(where $\tilde{P}(B)$ is the set of all fuzzy subsets of B) by comparing their attributes. For example, if sets A and B refer to films, attributes could be *title*, *director*, *directedBy*, *year*, *releaseDate* etc. This method has been used in diverse areas such as identifying when different news reports are concerned with the same underlying story, merging information about films from different sources and combining classified directories.

3.2 Multiple Taxonomies

Large AI projects in knowledge representation e.g. [9] have shown that it is impossible to create a single unified ontology. "Mediator" systems which aim to answer a query by combining responses from multiple sources form an important area of current research. Several tools have been proposed to aid in the automation of this problem, and were surveyed from various perspectives by Rahm and Bernstein [15]. They present a taxonomy covering many existing approaches based on the split between meta-data matching and content (instance) matching. The need for uncertainty has also been noted by others. For example, Chang et al. [6] developed an approach for the precise translation of Boolean queries across different information sources. In a subsequent paper [5], Chang and Garcia-Molina presented a real-world case study (combining book searches from four web sites), and found that it was only possible to make exact mappings in 30% of the rules ,while 70% required approximation.

Integration of semi-structured information from heterogeneous sources is an unsolved and important problem which is ideally suited to the use of fuzzy sets. We have used the SOFT instance matching method to find correspondences between hierarchies, when instances are classified according to different categorisations. We use the equivalence of instances from different sources to learn a soft mapping between categories in these hierarchies, allowing us to compare the hierarchical classification of instances as well as their attributes. Such correspondences may in turn be used to improve the identification of equivalent instances.

In general, we consider two sets of instances A and B with corresponding sets of labels LA and LB, each of which has a hierarchical structure, i.e. there is a partial order defined on the labels. Each label $l_i \in LA$ denotes a subset of A (and similarly for B), i.e. we have a denotation function

$$den : L_A \rightarrow A$$

such that $l_i > l_j \Leftrightarrow den(l_j) \subseteq den(l_i)$

For example, if A and B are sets of films then LA and LB could be genres such as *western*, *action*, *thriller*, *romance*, etc. Given a label $l_i \in LA$, we consider its denotation $den(l_i)$ under the mapping h and compare it to the denotation of $l_j \in LB$. In the ideal case if the two labels are equivalent,

$$h(den(l_i)) = den(l_j)$$

Given that h is approximate and that the correspondence between labels may not be exact, we use semantic unification [2] to compare the sets.

$$Pr(l_i \rightarrow l_j) = Pr(h(den(l_i)) = den(l_j))$$

This gives an interval-valued conditional probability which expresses the relation between a pair of labels; we then extract the most likely pair to give a crisp relation

$$g_c : L_A \rightarrow L_B.$$

Ideally, it should be possible to map such categories into a user's personal hierarchy, and this is the subject of ongoing research. The method has been applied to two film websites, "rotten tomatoes" and the internet movie data-base which are "user-maintained" datasets aiming to catalogue movie information. Each contains in the region of 100,000 film records. Since they are produced by two different movie web sites, there is inevitable "noise" existing in the film data; i.e. different tag sets, different genre names and missing elements.

The similarity between two genres is relatively hard to decide from text string matching. For example, "animation" is not similar to "children's", but the extension of the sets of films in these two categories shows considerable overlap, as do *Horror* and *Suspense* - further detail is given in [12].

4 Summary

We claim that systems which represent and reason with real world knowledge and combine multiple semi-structured (or structured) information sources should avoid artificially precise definitions and model human-understandable categorisations by using fuzzy sets. Such systems are widely applicable in areas such as database integration, data warehousing and e-commerce and web searching.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* Addison Wesley, Harlow, UK, 1999.

[2] J. F. Baldwin. The management of fuzzy and probabilistic uncertainties for knowledge based systems. In S. A. Shapiro, editor, *Encyclopedia of AI*, pages 528–537. John Wiley, 1992.

[3] J. F. Baldwin, T.P. Martin, and B.W. Pilsworth. *FRIL - Fuzzy and Evidential Reasoning in AI.* Research Studies Press (John Wiley), 1995.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc International world wide web conference, Brisbane, Australia*, pages 107–118. Elsevier Science, 1998.

[5] C.C. Chang and H. Garcia-Molina. Approximate query mapping: Accounting for translation closeness. *VLDB Journal*, 10:155 – 181, 2001.

[6] C.C. Chang, H. Garcia-Molina, and A. Paepcke. Boolean query mapping across heterogeneous information sources. *IEEE Transactions On Knowledge And Data Engineering*, 8:515–521, 1996.

[7] M.G. Elfeky, V.S. Verykios, and A.K. Elmagarmid. TAILOR: A record linkage tool box. In *Proc International Conference on Data Engineering*, pages 17–28, San Jose, CA, 2002. IEEE Computer Society.

[8] I.P. Fellegi and A.B. Sunter. A theory for record linkage. *J. American Stat Assoc.*, 64:1183–1210, 1969.

[9] D.B. Lenat. CYC: A large-scale investment in knowledge infrastructure. *Comm. ACM*, 38:32, 1995.

[10] P. Lyman and H.R. Varian. How much information? www.sims.berkeley.edu/how-much-info-2003, 2003.

[11] T. P. Martin and B. Azvine. Soft integration of information with semantic gaps. In E. Sanchez, editor, *Fuzzy Logic and the Semantic Web.* Elsevier, 2005.

[12] T.P. Martin and Y. Shen. Soft mapping between hierarchical classifications. In *Proc IPMU-2006*, Paris, 2006.

[13] H.B. Newcombe, J.M. Kennedy, S.J. Axford, and A.P. James. Automatic linkage of vital records. *Science*, 130:954–959, 1959.

[14] J. Novak, P. Raghavan, and A.Tomkins. Anti-aliasing on the web. In *Proc WWW04*, pages 30–39, New York, 2004.

[15] E. Rahm and P.A. Bernstein. A survey of approaches to automatic schema matching. *The VLDB Journal*, 10:334–350, 2001.

[16] G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval.* McGraw Hill, New York, 1983.

[17] A. Sheth, C. Ramakrishanan, and C. Thomas. Semantics for the semantic web: The implicit, the formal and the powerful. *Journal on Semantic Web and Information Systems*, 1:1–18, 2005.

[18] P. Smets. Constructing the pignistic probability function in a context of uncertainty. In M. Henrion, et al., editor, *Uncertainty in Artificial Intelligence 5*, pages 29–39. 1990.

[19] M. Uschold. Where are the semantics in the semantic web? *AI Magazine*, 24:25–36, 2003.

Fraud Detection in Consumer Credit

Niall M. Adams[1], David J. Hand[1,2], Giovanni Montana[1], David J. Weston[2]
[1]Department of Mathematics, Imperial College London
[2]Institute for Mathematical Sciences, Imperial College London

Abstract: Fraud detection poses major computational and statistical challenges to plastic-card credit providers. These challenges include the difficulty of both performance assessment and setting control parameters. We describe these challenges and illustrate them in the context of detecting fraudulent behaviour in a relatively simple and small artificial problem.

Keywords: Fraud detection, anomaly detection, time series, detection performance

1 Introduction

Plastic card fraud is a very serious problem. It is estimated that such losses in the UK in 2004 amounted to £505 million. This number showed a year on year increase over the decade from 1995 (apart from an apparent tiny drop from £425m to £420m from 2002 to 2003) [1]. This financial loss is absorbed by lenders, merchants, and legitimate customers. While it is clear that fraud is a critical problem, banks face a number of serious challenges in dealing with it. One problem is the need for a rapid authorisation response. This requirement dictates that currently only simple, rule-based transactions filters can be used for fraud detection. A second problem is the need to avoid false positives. For customer relationship purposes it is crucial to minimise the number of incorrect "authorisation-denied" responses. A third problem is simply the urgency of detecting fraudulent card usage as rapidly as possible. Another difficulty is that, compared to the volume of non-fraudulent transactions, fraud is a very rare.

There are a number of distinct types of fraudulent behaviour, as well as emergent modes of attack. The latter point illustrates that there is an "arms-race" between fraudsters and lenders - with fraudsters continually adapting to the barriers put in place by lenders. The Chip and PIN system which was launched in the UK on 14th February 2006 will prevent certain types of fraud, but this merely means that fraudsters will switch to other modes. A topical example of an emergent attack is "sleeper" fraud, where a fraudster sets up a line of credit with the intention of building up a good credit history to increase lending limits. Later, the seemingly good customer rapidly empties the account completely. Obviously, this is a particularly difficult type of fraud to detect.

Tackling fraud presents a variety of opportunities for the application of data mining and knowledge discovery tools. Phua et al. [4] provide a recent review. Classes of methods which have been applied include supervised classification [2], novelty detection [7], and outlier detection techniques (reviewed by Hodge and Austin [12]). There is also scope for other types of tool, for example, pattern detection and discovery [10] and change-point detection [6]. It may not be that any of these approaches will immediately lead to improved methods at the coalface, but the research process should ultimately yield powerful and deployable insights.

A team at Imperial College is working on a long term project, *Statistical and machine learning tools for plastic card and other personal fraud detection*, funded by the EPSRC. This project is in collaboration with several banks, who are supplying us with data. (Regrettably, of course, these data are highly confidential.) The systems in these banks which make the complete authorisation process rapid, automatic, and relatively fraud resistant are incredibly complex, and the authorisation of a single transaction is a process that can accumulate a large amount of highly-structured data. For example, one of the banks has provided card authorisation data in which each transaction has up to 76 fields, including free form text fields. These data described 4 months of transactions, a total of 175 million transactions, with a fraud rate of 0.03%.

In this paper, we illustrate some of the difficulties of fraud detection with a simple problem relating to detecting fraud in ATM (Automated Teller Machine) cash utilisation. We stress that this is a very small example, involving a single continuous variable and a relatively small sample. However, this example does illustrate some of the interesting characteristics of fraud problems. Note that the approaches here are all examples of *unsupervised* detection, in the sense that the true fraud indicator is never explicitly utilised when building the detectors, though it is available for performance assessment.

The aims of this paper are twofold. First we describe several approaches, each of which combines a method for modelling normal behaviour with a method for raising a fraud *alarm* - that is, for finding a departure from normal behaviour and asserting that fraud has occurred at a specific time. We will sometimes use the tactic [9] of looking for sequences of departures. In these cases, a single departure from normal behaviour is termed an *alert*, and certain sequences of alerts are required to raise a fraud alarm. The modelling strategies vary in technical sophistication from simple nearest neighbour approaches to dynamic linear models. Secondly, we develop a criterion for comparing the performance of different fraud detectors for situations such as the ATM problem and evaluate our methods using it. Criteria rather more advanced than those developed for standard two-class classification problems are required (though those are difficult enough - see [5]), because of the need for timely decisions. Our measure accounts for both accuracy and timeliness.

2 Data

The data comprised the daily cash output of 700 UK ATMs over a period of more than two years. A selection of these (non-fraud) time series is displayed in Figure 1. The figure illustrates some interesting features of the data. Firstly, the top left frame illustrates a problematic characteristic of such data - an occasional, very large extreme value. This behaviour is normal, possibly explained by other machines in the locale being unavailable, resulting in a single day of extreme heavy demand. Secondly, the plots in Figure 1 include occasional zeroes. For this data set, typical of data mining, the symbol zero is overloaded since it can represent no cash requested or failure to meet demand, either by being empty or inoperable. Although it is not apparent from all these plots, extensive modelling revealed strong weekly and monthly periodic structure.

For the purpose of this investigation, a small subset of the machines was randomly selected and distorted. The distortion was intended to replicate a type of staff (or internal) fraud, and is manifested as a small change in the behaviour of the machine after a certain time. Our objective then is to construct unsupervised, sequential fraud detection mechanisms that correctly and rapidly detect the consequent changes in behaviour.

3 Measuring Detection Performance

Fraud alarms in our present context have two aspects. Firstly, there is the recognition that a fraud has occurred (that a particular ATM has been subject to a fraudulent withdrawal), and secondly there is the timeliness of detecting that a fraud has occurred. There is limited value in a batch mode analysis which detects that a fraud occurred on a particular machine a year ago. Some proposed approaches are based on ROC curve analysis, including the AMOC curve used for detecting behaviour change [11]. We propose the following fraud detection

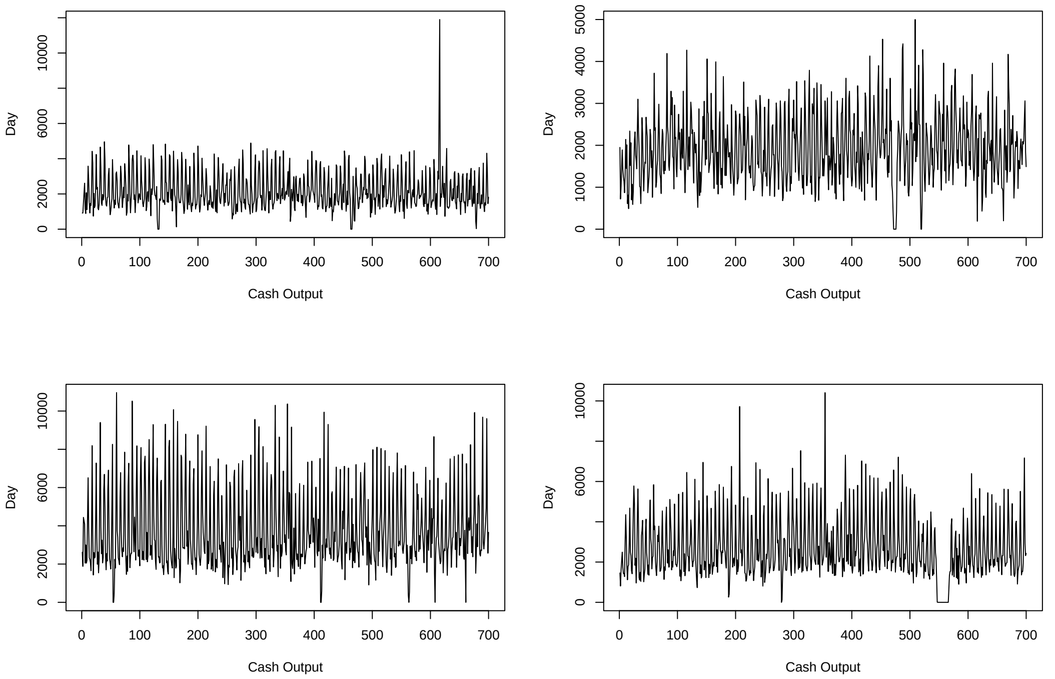


Figure 1: Examples of non-fraud ATM series

performance assessment score which combines both accuracy and timeliness into a single measure.

In considering flagging a fraud, we have to consider the usual decision matrix:

		Fraud Event	
		Yes	No
Fraud Alarm	Yes	True alarm (a)	False alarm (b)
	No	Missed Fraud (c)	True negative (d)

We also need to account for the time frauds and alarms occur. In particular, we consider t_1, the time fraud occurred, t_2, the time a fraud alarm occurred and t_3 ($> t_2$) the time it was deemed that the fraud occurred. This later case accounts for the need to accumulate evidence (in the form of alerts) before raising an alarm. Thus, if fraud occurs at time t_1, at time t_3 we have sufficient data to claim fraud happened at time t_2. We consider the elements of the decision matrix separately, as follows.

TRUE ALARMS (a)

The core of our performance measure is

$$M = \epsilon \left(1 - \left(\frac{(1 + |t_1 - t_2|)^{-1} + (1 + t_3 - t_2)^{-1}}{2} \right) \right)$$

Like most work in classifier performance assessment, this measure is a cost, so that small is good. This attempts to reward both t_2 and t_3 for being near t_1, and treats t_2 and t_3 equally. Here a score of zero means that fraud is detected immediately. The constant ϵ, with a value less than 1, is used to balance the cost of true alarms, and true negatives. The derivation of this measure will be given in detail in a forthcoming paper.

MISSED FRAUD AND FALSE ALARM (b and c)

These errors will usually have asymmetric costs. We measure the cost of these two errors as proportional to the number of frauds, n_F, and the number of non-frauds n_N. It will usually be the case that n_F will be much smaller n_N. If these counts were not available we might estimate them from quoted population prevalence values.

TRUE NEGATIVE (d)

To ensure that correct alarms (a) incur smaller cost that true negatives we set this cost to ϵ.

This finally yields:

		Fraud Event	
		Yes	No
Fraud Alarm	Yes	M	C/n_N
	No	C/n_F	ϵ

Sensible arguments to induce a reasonable measure, reflect structural constraints and treat the incorrect alarms asymmetrically, lead to $C = 2$ and $\epsilon = 0.05$. This provides the means to compute a score for all machines. We simply sum these scores to obtain *fs*, our measure of performance for a fraud detector.

4 Fraud Detectors

We consider a collection of fraud detectors derived from three modelling approaches. Some detectors accumulate alerts prior to triggering an alarm, and can be parameterised by both the magnitude of the alert threshold, O, and some function of the number of alerts, N, required before raising an alarm.

EXPLICIT MODEL (EM):

This method was derived from a thorough analysis of the ATM data. The method uses a model of cash output based on a 31-day moving window (although the size of the window is a parameter to be selected, this choice is well motivated by data analysis). The model considers the log cash output as an additive linear function of day of month and linear and quadratic terms in day of week, after suitable scaling. There is strong support for this particular seasonal structure in the data. Model parameters are estimated using least squares.

The five detection mechanisms derived from this model keep track of various model (and derived) parameters. This tracking involves fitting a 31-day moving average to daily differences in the parameters, thereby smoothing the differences from one window to the next. An alarm is generated if this (smoothed) difference exceeds an alert threshold, O. The value t_2 associated with an alarm is taken as 15 days prior to the alarm, reflecting the 31-day moving window. This alert threshold can be set on a set of unlabelled training data to induce a specific alert rate.

DYNAMIC LINEAR MODEL (DLM):

This detector assumes that all the ATMs are realisations of a common underlying stochastic mechanism described by a seasonal dynamic linear model [8]. The model is formulated to capture some of the structure described above. ATMs are modelled independently and the models are updated sequentially using the Kalman filter. The marginal distribution of one-step ahead forecast errors is available analytically under our modelling assumptions. This provides the means to use hypothesis testing as a means to detect outliers, and hence raise alerts. Early experiments found the theoretical distribution unsatisfactory, since it gave too many outliers, so we use the empirical distribution instead. The null hypothesis of normal behaviour is rejected at a specified fixed significance level, O, related to the alert threshold.

Rather than address the problem of multiplicity, this approaches adopts a heuristic rule to raise a fraud alarm. This rule says if N outliers starting at time t are separated by gap G, an alarm flag is raised at time t. We consider four choices of the parameters N and G, attempting to capture different modes of fraud. Sensible choices of N and G were obtained from a study of training data.

NEAREST NEIGHBOUR DETECTOR (NND):

Of all the methods we have explored, this nearest neighbour approach uses the least structural knowledge about the ATM time series. The method starts by holding back the first 90 days of data. Then, for a target machine on a specific day, we find the $K = 20$ nearest 7-day windows (in the sense of Euclidean distance) across all machines that end on the same day. This search is only conducted across the data observed up to the specific day. The 7-day period is motivated by structural knowledge of the data. For these K machines, we obtain the cash output for the specific day. If the actual cash output of the target machine is more than O standard deviations from the mean of the K outputs, raise an alert.

To set the alert threshold O, the process was repeated on unlabelled training data, using random, rather than nearest neighbours. This corresponds to the situation where the nearest neighbours are not informative. A large number of simulations provided approximations of the probability of an outlier for different values of alert threshold, O. We then select O to induce a low probability of observing an outlier.

5 Results

The actual fraud rate with these data, in terms of frauds per day, is 3%. These frauds were distributed across 74 of the 700 machines. For our experiments, we randomly split the data into equal size samples, one for setting control parameters, like the alert threshold, O, and the other for formal assessment. Note that the data used for tuning did not include the fraud indicator.

Detector performance, in terms of *fs*, is presented in the following table. For comparison the percentage of fraud machines identified (FI%) and the correct alarm rate (CA%) are included. Note that the DLM model has a number of detectors associated with different choices of alert threshold and count. Similarly, the EM model has a number of detectors for different model parameters.

Method	*fs*	FI%	CA%
NND	3.42	54.3	27.1
DLM-1	4.39	8.6	100.0
DLM-2	4.03	25.7	100.0
DLM-3	3.97	28.6	100.0
DLM-4	4.02	25.7	100.0
EM-1	3.97	22.9	11.7
EM-2	3.39	40.0	21.9
EM-3	3.81	22.9	12.3
EM-4	3.87	17.1	9.5
EM-5	3.58	25.7	14.6

The detector that performs best in the sense of our detection measure, *fs*, is EM-2. This particular technique keeps track of the day-of-month parameter of the explicit model. In contrast, the simple nearest neighbour detector identifies most frauds, although this must be moderated by the associated low correct alarm rate. NND also performs very well in terms of *fs*. It is striking that such a simple procedure can perform well, on both metrics. The primary difference between EM-2 and NND is that the former provides much more timely alarms.

We note that a more formal comparison of these performance measures should account for finite sample variability and also for other sources of uncertainty. Bootstrap methods [3] provide one approach for calibrating these estimates, and this is something on which we will report in the future. Certainly, it is difficult to draw very strong conclusions from the above results.

6 Conclusions

Our illustration demonstrates some of the difficulties of unsupervised fraud detection. It will usually be the case that the alert threshold, O, has to be determined such that the number of false alarms is minimised. Our experiments explored a number of ways of thinking about setting this threshold. This issue remains an open research problem. Unfortunately, the nature of the problem usually means that we cannot think about selecting O as an optimisation problem. It will often be the case that other control parameters are also required. The different EM approaches produce a wide range of results, suggesting that tuning this model appropriately might be critical. On the other hand, when tuned well, it performs well.

The importance of timely alarms cannot be overlooked. Typically, the longer fraud is allowed to proceed the more cost it will accrue. Our experiment illustrates the importance of selecting fraud detection methodology on the basis of a performance measure that rewards timely alarms. A more crude measure, like the correct alarm rate, may lead us to use the wrong method.

The performance of the nearest neighbour method is very impressive. Techniques such as this that do not use parametric models for normal behaviour are very attractive, for two reasons. First, they have the potential for dealing with gross changes in behaviour, while parametric models cannot, without re-modelling. Second, such techniques do not have the same modelling overhead; they can be deployed largely automatically. This is particularly important for large scale transaction fraud applications described by multivariate, mixed type observations. Some of our current research is concerned with extending the nearest neighbour detector to select the window width automatically.

We have illustrated some of the difficulties of the fraud detection problem with a very small example. These difficulties, and others, are enormously amplified by large and more complex fraud data sets. Our on-going research attempts to address such problem with a variety of tools including classification, outlier detection, change point detection and pattern detection and discovery. The work reported here is part of a much more elaborate project.

Acknowledgements

The work of David Weston was supported by grant number EP/C532589/1 from the UK Engineering and Physical Sciences Research Council. The work of Christopher Whitrow was supported by the Institute of Mathematical Sciences, Imperial College London. We are indebted to our commercial collaborators for providing data and useful discussions about the fraud detection problem.

References

[1] Association for payment clearing services: Cardwatch. http://www.cardwatch.org.uk/default.asp? sectionid=5&pageid=82, 2005.

[2] Webb A.R. *Statistical Pattern Recognition.* Wiley, 2002. Second edition.

[3] Efron B. and Tibshirani R.J. *An Introduction to the Bootstrap.* Chapman and Hall, 1993.

[4] Phua C., Lee V., Smith K., , and Gayler R. A comprehensive survey of data mining based fraud detection research. *Artificial Intelligence Review.* Submitted 2005.

[5] Hand D.J. *Construction and assessment of classification rules.* Wiley, 1997.

[6] Basseville M. and Nikiforov I.V. *Detection of Abrupt Changes.* Prentice Hall, 1993.

[7] Markou M. and Singh S. Novelty detection: a review-part 1: statistical approaches. *Signal Processing*, 83:2481–2497, 2003.

[8] West M. and Harrison P.J. *Bayesian Forecasting and Dynamic Models.* Springer-Verlag, 1997. Second edition.

[9] Adams N.M., D.J. Weston, Bolton R.J., and P.R. Cohen. Anomaly-based intrusion detection in categorical data streams using a goodness-of-fit statistic. *Machine Learning.* Submitted 2006.

[10] Bolton R.J. and Hand D.J. Statistical fraud detection: a review. *Statistical Science*, 17:235–255, 2002.

[11] Fawcett T. and Provost F. Activity monitoring: noticing interesting changes in behavior. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pages 32–41. ACM Press, 1999.

[12] Hodge V.J. and Austin J. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22:85–126, 2004.

Germplasm collections: Gaining new knowledge from old datasets

Robert Davey[1,2], George Savva[1,2], Runchun Jing[3], Martin Lott[4], TH Noel Ellis[1], Michael Ambrose[1], Vincent Moulton[4], Andrew Flavell[3], Ian Roberts[2] and Jo Dicks[1]

Abstract: Over several decades, germplasm collections have been developed across the world to capture the genetic diversity of crop plants vital to food and agriculture. Recently, the genetic characterisation of many of these collections has begun, using a variety of genetic marker technologies. Here, we describe some first attempts at uncovering the genetic structure of a single collection characterised by high-throughput marker techniques. This research both hints at the knowledge that may be gained by analysing such datasets and identifies areas of research that should be targeted for the future.

1 Introduction

The production of crops, a large part of the worldwide food supply, relies on intensive agricultural practices that can lead to genetic uniformity. Such uniformity creates risks regarding maintaining protection against pests, disease and environmental change. Plant breeders wish to develop new crop varieties that can overcome old adversities and deal with new ones as they arise. In order to achieve this, the breeder must have access to a wealth of genetic diversity in their crop of interest. The development of germplasm collections, which capture this diversity, has been ongoing for decades. Collections have been developed though national projects and through international collaborations. For example, the work of the Consultative Group on International Agricultural Research (CGIAR; http://www.cgiar.org/) has led to the systematic collection of specimens of landraces, old cultivars, wild species, advanced cultivars and breeders lines, for cassava and sweet potato to rice and maize. The eleven CGIAR international genebanks currently maintain over 600,000 crop, forage and agroforestry samples in the public domain, providing massive datasets to be analysed in the coming years.

The historical driver for the development of germplasm collections was to preserve and document crop genetic diversity, thus ensuring future food security, whilst also evaluating and distributing the germplasm. Recent advances in genetic marker technology are leading to the growing genetic characterisation of germplasm collections, allowing for informed exploitation of the germplasm in future breeding studies. Germplasm collections may hold important *alleles* (or versions of a gene) for agronomic traits such as disease resistance, yield and tolerance to a broad range of environmental conditions. With the molecular characterisation of germplasm collections comes the ability to carry out detailed analyses on their genetic structure. For example, we may wish to search for associations between traits and alleles or perhaps traits and haplotypes (allelic combinations of adjacent genes). We may wish to understand the evolutionary history of the species, in particular the balance between the different processes of genetic marker evolution (vertical evolution) and introgression (introduction of a gene or haplotype from one variety to another via hybridisation - horizontal evolution). We may wish to know how closely related two members of a collection, or accessions, are to one another. We may wish to examine all relationships within the collection and use this to develop a *core collection* that maximises the diversity for a small, fixed number of accessions.

[1] John Innes Centre, Norwich Research Park, Colney, Norwich, NR4 7UH, UK.

[2] Institute of Food Research, Norwich Research Park, Colney, Norwich, NR4 7UA, UK.

[3] University of Dundee at SCRI, Invergowrie, Dundee, DD2 5DA, UK.

[4] University of East Anglia, Norwich, NR4 7TJ, UK.

In this article, we will introduce ongoing research that attempts to answer some of these questions. We will begin by describing the molecular characterisation of a germplasm collection for pea by a recently developed high-throughput marker technique. Such techniques bring with them new challenges for determining marker scores from the resultant raw datasets. We will then discuss how we can use the marker scores to assess the genetic difference between accessions and estimate the structure of an entire germplasm collection. Finally, we will touch upon ongoing research into the estimation of efficient core collections.

2 Marker prediction from high-throughput datasets

With the desire to analyse the genetic structure of a germplasm collection, there comes an interesting debate as to which type of molecular marker is most appropriate for the task. The last decade has seen the rapid development of marker technologies in the plant domain, from RFLPs, SSRs and AFLPs through to SNPs (single nucleotide polymorphisms), SSCPs (single stranded conformation polymorphisms) and RBIPs (retrotransposon-based insertion polymorphisms), with some marker types targeting genic regions of the genome and others deriving from alternative genomic features.

RBIPs [1] are based on a genomic element known as a retrotransposon. Such an element can be thought of as a mobile piece of DNA that inserts itself within a genome and subsequently jumps to a new genomic location, whilst leaving a copy of itself behind. Thus retrotransposons accumulate within the genome, leading to an observed growth in plant genome size. As these elements can only be gained, and not lost, through their normal mode of evolution, they can help us to understand the direction of evolution by the order of their accumulation. However, introgression can lead to the appearance of an element being lost or gained without a deletion or a jump taking place. Each retrotransposon type possesses a number of locations within a genome at which it can be inserted. Therefore each plant accession can be characterised by a particular pattern of presence and absence of the retrotransposon at each of these locations. Formally, for a particular retrotransposon, there is a fixed order i (along the genome, if this information is known, otherwise a conceptual order for clarity only) and number N (i.e. $i = 1, \ldots, N$) of locations at which the retrotransposon may be present or absent. Thus, for each plant j, each value $m_{i,j}$ denotes the presence or absence of a copy of the retrotransposon at position i in plant j. We say that $m_{i,j} = 0$ when the retrotransposon is absent and $m_{i,j} = 1$ when the retrotransposon is present.

A high-throughput experimental technique for the assaying of a single RBIP marker in a large number of plant accessions (e.g. several thousand) has recently been developed [2]. This technique is known as the tagged microarray marker (TAM) approach. TAM microarrays have recently been used to characterise the John Innes *Pisum* Collection using the PDR1 retrotransposon (see http://www.jic.bbsrc.ac.uk/germplas/pisum/). This characterisation has taken the form of 76 experiments (one for each insertion site) over 3,029 *Pisum* accessions together with 171 positive and negative controls. The experiment measures the relative levels of red and green fluorescently labelled probes for each plant. The probes are specific to each genomic location such that, at a particular insertion site, the red probe is designed to be indicative of the absence of the retrotransposon at that particular location and the green probe is designed to be indicative of its presence. Thus for each experiment i, we are given a measure of intensities of green and red for each plant j, $g_{i,j}$ and $r_{i,j}$ respectively. The first problem presented to us is to use these values of $g_{i,j}$ and $r_{i,j}$ to predict the corresponding values of $m_{i,j}$.

The recent widespread use of gene expression microarrays in biological research has taught us many lessons about analysing such datasets. We know that, prior to comparison of our red and green intensities, we need to *normalise* the raw data. In particular, we have chosen to use the *vsn* routine [9] within the BioConductor suite [8] of the R statistical package (http://www.r-project.org/). This algorithm both calibrates the red and green values (i.e. brings them onto a common scale so that the intensities can be directly compared) and stabilises their variance (i.e. transforms the values so that their variance is no longer a function of intensity but is more or less constant across the intensity scale). Figure 1 shows two distributions of the ratios of red and green intensity levels after they have been analysed with *vsn*, one for each of two markers. The left distribution is bimodal, with the left peak representing accessions where the retrotransposon is present (i.e. the green intensity level is significantly higher than the red intensity level) and the right peak representing accessions where the retrotransposon is absent (i.e. the red intensity level is significantly higher than the green intensity level). The right distribution is more difficult to analyse, with four significant peaks. In this example, it is most likely (comparing the distribution to that of other markers) that the leftmost peak represents low intensity values that cannot be analysed with any certainty. The second left peak represents the accessions with the retrotransposon present and the rightmost peak accessions with the retrotransposon absent. The remaining peak represents "yellow" spots where the red and green intensity levels are comparable. At first sight, one would presume that these were plants that were *heterozygous* for the retrotransposon insertion (i.e. on one copy of the relevant chromosome the retrotransposon was present and on the other it was absent). However, it is known that the plants within the *Pisum* collection are *homozygous* for these retrotransposons (i.e. both copies are present or both are absent). What appears to be happening is that some insertions reside in repeated sequences, so a plant containing an "occupied" signal from a locus might nevertheless produce an "unoccupied" signal from another copy of the repeat elsewhere in the genome. In some cases such problems can be solved but further research must be done to resolve this issue.

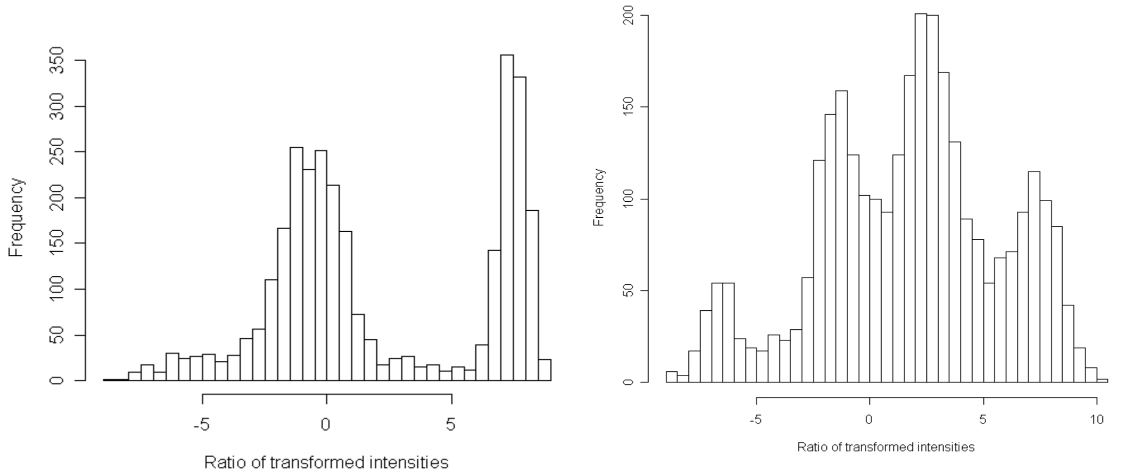


Figure 1: Two distributions of transformed intensity ratios for a single RBIP marker, each assayed by a TAM microarray in 3,029 varieties of *Pisum*

Once we have determined the meaning of these peaks, we can use mathematical techniques to predict the status of each marker within each plant accession. We have recently fitted Gaussian (normal) mixtures to the distribution of transformed intensity ratios to predict marker presence or absence. Furthermore, we are currently automating the analysis of TAM microarrays within our MPP (microarray-to-phylogeny pipeline) software (http://cbr.jic.ac.uk/dicks/software), which was originally developed for the analysis of Comparative Genomic Hybridisation (CGH) microarrays. By combining the results of each array analysis, we produce a table of 76 x 3,029 elements, with each element being a 1 or a 0. We can now use this table to find out more about the relatedness of our accessions and the overall structure of the germplasm collection.

3 Measures of distance

When comparing the marker scores of two or more accessions, we need to have some measure of comparison. Usually, we will use a measure of the distance between two sets of marker scores, where we would like this distance to be strongly correlated to the real evolutionary time separating them. There are many distance measures in the biological domain that are used to compare sets of binary characters, such as our RBIP marker scores. We will now discuss briefly two distance measures: the Jaccard distance and the retrotransposon distance.

Jaccard distance

A distance measure widely used with genetic markers is the Jaccard distance which, for our data, can be calculated as follows between two accessions x and y:

$$d_J(x,y) = a/(a+b+c)$$

where

$$\begin{aligned} a &= \text{no. of markers } i \text{ where } m_{i,x} = 1 \text{ and } m_{i,y} = 1 \\ \text{b} &= \text{no. of markers } i \text{ where } m_{i,x} = 0 \text{ and } m_{i,y} = 1 \\ \text{c} &= \text{no. of markers } i \text{ where } m_{i,x} = 1 \text{ and } m_{i,y} = 0 \end{aligned}$$

The Jaccard distance is very easy to calculate, even for large datasets. Furthermore it is widely used and understood and may be used flexibly for many types of marker. However, because it is widely applicable it may not maximise the information contained within a particular type of dataset. For this reason, we are looking to develop a custom distance measure for RBIP datasets.

Retrotransposon distance

We have recently begun looking at ways of modelling the retrotransposon insertion process. If we suppose that retrotransposons arise according to a simple birth process and, furthermore, that a proportion of insertion sites ρ are invariant (i.e. always empty) and that rates of insertion per site vary across the genome according to a Gamma distribution with shape parameter α then the maximum likelihood estimates (Savva, manuscript in preparation) of the retrotransposon distance between an accession x and the *reference* accession 0 (where all sites are empty) and between two accessions x and y are as follows:

$$d_R(x,0) = \left(\frac{1-\rho}{\frac{N_x}{N}-\rho}\right)^{\frac{1}{\alpha}} - 1$$

$$d_R(x,y) = \begin{cases} 2\left(\frac{1-\rho}{\frac{N_{xy}}{N}-\rho}\right)^{\frac{1}{\alpha}} - \left(\frac{1-\rho}{\frac{N_x}{N}-\rho}\right)^{\frac{1}{\alpha}} - \left(\frac{1-\rho}{\frac{N_y}{N}-\rho}\right)^{\frac{1}{\alpha}} & \text{if } \frac{N_x N_y}{N^2} < \frac{N_{xy}}{N} \\ \left(\frac{1-\rho}{\frac{N_x}{N}-\rho}\right)^{\frac{1}{\alpha}} + \left(\frac{1-\rho}{\frac{N_y}{N}-\rho}\right)^{\frac{1}{\alpha}} - 2 & \text{otherwise} \end{cases}$$

where N is the number of insertion sites, N_x is the number of empty sites in x, N_y is the number of empty sites in y and N_{xy} is the number of sites empty in both x and y.

At present, this distance is a very simple model of the retrotransposon insertion process and we have yet to compare its performance to that of the Jaccard distance. In the future, we intend to validate and extend the model. For example, we would like to be able to analyse more than one retrotransposon type simultaneously. Furthermore, we need to take into account that most crop plant germplasm collections contain strongly conserved, fragmented haplotypes, which have been distributed across the species by introgression (i.e. horizontal evolution). Thus, two apparently highly diverged plants might be almost identical for a large fraction of a particular chromosome(s). The new model should take into account the introgression process such that we will be able to formally estimate the relative contributions of insertion and introgression to the evolution of a group of accessions, while analysing more than one retrotransposon type simultaneously.

4 Deducing network-like structures

Once we have established methods of calculating distances between pairs of accessions, we can use these values to analyse the pattern of genetic diversity within the collection as a whole. Traditionally, many types of biological dataset have been viewed as tree structures, after the "tree of life" thought to connect all living organisms. However, it has become apparent in recent years that trees will not always describe adequately a biological dataset and that network-like evolutionary events may play an important role in shaping such datasets. Several algorithmic methods have been developed to estimate some type of network from a matrix of distances. One such method is the NeighborNet [3], which is implemented in the SplitsTree4 software [4]. NeighborNet essentially extends the widely used neighbor-joining algorithm [7], one of the most popular methods of tree construction in the biological domain, to one capable of deducing a planar phylogenetic network. NeighborNet allows the researcher to visualise areas of the graph that are inconsistent with a tree-like structure, via "box-like" features. For a germplasm collection, such features may represent introgression events, which do not follow a treelike evolutionary mode.

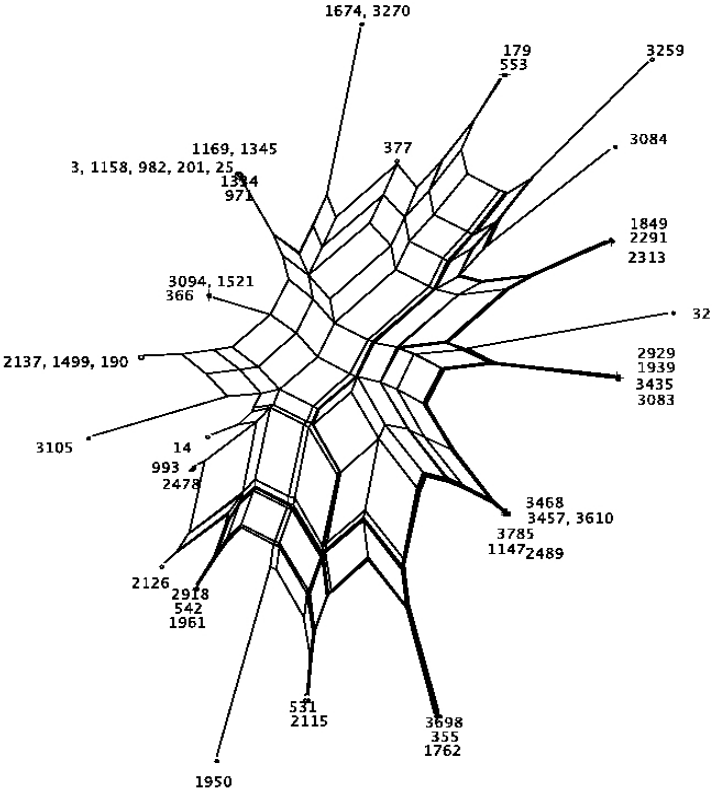


Figure 2: A NeighborNet of 50 *Pisum* accessions scored over a subset of 7 RBIP markers

Figure 2 above shows a NeighborNet of 50 *Pisum* accessions, assayed over 7 RBIP markers and with evolutionary distances estimated using the Jaccard distance. It will be interesting to see whether or not the apparently significant network-like structure seen within this figure remains when all 76 markers have been analysed.

5 Estimating a core collection

Having calculated both a distance matrix and a neighbor-joining tree or NeighborNet (whichever is most appropriate) for a set of accessions within a germplasm collection, we would then like to use this information further to find an efficient *core collection*. A core collection may be thought of as a subset of the overall collection that describes most of its diversity (whether genetic, geographical or phenotypic) for a fraction of its size. Typically, a core collection comprises 10% of the number of accessions seen in the whole collection. Therefore, for a fixed number c, the required size of our core collection, we need to find the network that displays the maximum amount of diversity over c accessions.

We have recently begun to develop new approaches for the estimation of core collections based on genetic diversity. The starting point of this research is the greedy algorithm proposed by Steel [6]. This algorithm proceeds as follows, either from a distance matrix or a neighbor-joining tree, where c is the number of accessions within the core collection and G is the set of all accessions:

```
Choose the pair of accessions most diverged from one another within set G
Add both accessions to the current accession set S
While ($|S| < C$) do
        (Choose the accession from G that is most diverged from S
        Add this accession to S)
```

This algorithm is simple to implement and has been shown to give a guaranteed solution to the problem of finding an optimal genetically-based core collection from a distance matrix or a neighbor-joining tree, with no constraints on collection members or their properties. However, for a computational solution to be of real practical benefit to germplasm collection managers, other factors need to be considered. For example, it would be useful to be able to place constraints on datasets, as different managers will have different priorities for selecting core collections such as requiring allelic variation at a particular site or only including accessions with particular characteristics. Furthermore, many traditionally created core collections attempt to maximise variation not only genetically but also geographically and phenotypically and this needs to be taken into account in algorithmic approaches. We also need to develop techniques to account for missing data. For example, in our current marker analysis of the JIC Pisum dataset, roughly 10% of the dataset is uninterpretable, an unfortunate but common downside to high-throughput techniques. If we were to use basic techniques for dealing with missing data, we would ignore any marker where could not determine a score for one or more accessions. In some datasets this could mean ignoring a large proportion of the dataset. Clearly, more research in this area is required to maximise the information gained from germplasm collections with missing marker scores (or indeed any other type of missing data).

6 Discussion

Germplasm collections are essential resources for maintaining and documenting crop diversity and for developing efficient and targeted plant breeding studies. They contain useful alleles and allelic combinations that may help to combat crop disease and to overcome environmental pressures. High-throughput marker technologies present us with large, complex datasets that describe these collections in much greater genetic detail than has been available before now. Such information will enable us to understand the structure of crop plant species and therefore help us to develop strategies for the development of new varieties. Here, we have presented recent research in the analysis of such datasets, describing how marker scores may be predicted, evolutionary distances be estimated, and collection structures and core collections be determined. These approaches are essentially first efforts at understanding these datasets. In addition to the approaches touched upon here, other techniques may also be of considerable value. For example, a pilot study on the use of data mining algorithms (C4.5 and simulated annealing) for rule-based classification of trait-allele associations, in particular the association of marker scores with disease status, has been promising [5]. For all our methods, we need to evaluate formally their efficiency and utility, possibly through simulation. Ultimately, we aim to develop more sophisticated methodologies that will allow us and others to exploit these datasets to their full potential.

Acknowledgements

RD was funded by a BBSRC Committee Studentship. GS was funded by a John Innes Foundation PhD Studentship. RJ was funded by a grant from the European Union. THNE, MA, AF, IN and JD acknowledge support from the Biotechnology and Biological Sciences Research Council and the EU Framework 6 project Grain Legumes (FOOD-CT-2004-506223).

References

[1] Flavell A.J., Knox M.R., Pearce S.R., and Ellis T.H.N. Retrotransposon-based insertion polymorphisms (RBIP) for high throughput marker analysis. *Plant J.*, 16:643–650, 1998.

[2] Flavell A.J., Bolshakov V.N., Booth A., Jing R., Russel J., Ellis T.H.N., and Isaac P. A microarray-based high throughput molecular marker genotyping method - the tagged microarray marker (TAM) approach. *Nucleic Acids Res.*, 31:e115, 2003.

[3] Bryant D. and Moulton V. Neighbornet: an agglomerative method for the construction of planar phylogenetic networks. *Molecular Biology and Evolution*, 21:255–265, 2002.

[4] Huson D.H. and Bryant D. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2):254–267, 2006.

[5] Davenport G., Ellis T.H.N., Ambrose M, and Dicks J. Using bioinformatics to analyse germplasm collections. *Euphytica*, 137(1):39–54, 2004.

[6] Steel M. Phylogenetic diversity and the greedy algorithm. *Systematic Biology*, 54(4):527–529, 2005.

[7] Saitou N. and Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406–425, 1987.

[8] Gentleman R.C., Carey V.J., and Bates D.M. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004.

[9] Huber W., von Heydebreck A., Sultmann H., Poustka A., and Vingron M. Variance stabilisation applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18(1):S96–S104, 2002.

Bayesian Profiling to Identify Fraudulent Telephone Usage

Mark Girolami
Department of Computing Science, University of Glasgow
Glasgow G12 8QQ, Scotland, UK

Abstract: This paper presents a Bayesian solution to the problem of identifying anomalous, and therefore potentially fraudulent, usage of a telephone service by individual accounts. Creating and maintaining account specific profiles, which are represented by discrete multivariate probability distributions, provides a theoretically consistent and highly practical means of scoring and ranking individual service usage in terms of potential malfeasance. The computational overhead of the proposed method is such that millions of account profiles can be stored and maintained and tens of millions of transactions can be scored on a daily basis using standard computing capabilities. A commercial prototype system based on this proposed technology[2] has been developed by Memex Technologies and evaluated in an operational environment by the telecom operator NTL[3].

1 Introduction

Each year in the telecommunications sector, fraudulent transactions account for a substantial loss of annual revenue for telecom providers. The detection of such fraudulent activity is an arduous task and presents a significant challenge to researchers and practitioners alike. This is due to the nature of the telecommunications domain where a high volume of transactional call data is produced. However, only a small percentage of call transactions are actually fraudulent and to detect these in real time compounds the problem. Various solutions have been proposed [4], for example in [1] a system comprising of rule-based and artificial neural network components is developed, whilst in [3] and [2] signature based methods are proposed. In this paper we further develop the signature based methods by specifically adopting a Bayesian [5] approach to defining customer specific signatures of service usage. We very briefly demonstrate the ability of this Bayesian signature based scoring method to identify fraudulent service usage in a high volume transaction environment. The following section presents the statistical basis to fraud detection adopted in this paper.

2 Testing for Deviations from Normal Behavior

Given a customers telephone service usage profile and a series of recent telephone calls attributed to the customer account then a test is required to assess whether the logged transaction data provides evidence which is sufficiently high to accept that the calling activity genuinely originated from the account. Having logged a series of telephone calls, which supposedly have been made by a customer then the null hypothesis $\mathcal{H}_0$ is that the call has genuinely originated from the customers account and is consistent with all previous patterns of calling behavior from the account. The alternate hypothesis $\mathcal{H}_1$ regarding the telephone calls is that they were not made from the owner of the account but in fact originate from another account which we will denote as $\mathcal{F}$[4].

[2] Patent application number 0520789.9 filed by Memex Technologies.

[3] Memex (`www.memex.co.uk`) and NTL (`www.ntl.co.uk`) were industrial partners in a project funded under the DTI Management of Information (LINK) Programme and by EPSRC grant GR/R55184.

[4] We employ $\mathcal{F}$ to denote possibly fraudulent usage of the account which is discordant with the previous service usage.

From classical hypothesis testing there will be a rate α of TYPE I errors made for any test procedure. In this case the TYPE I errors (rejection of $\mathcal{H}_0$ when it is true) correspond to calls from an account which are genuine being labeled as *fraudulent* or having not originated from the customer. Given that the number of telephone calls being tested in a 24 hour period is of the order of tens of millions then the TYPE I error rate α has to be very carefully controlled and kept low to ensure that the number of calls exceeding the threshold are kept to a manageable level for operators who may be required to process the accounts which raise *alarms.* On the other hand the TYPE II error rate β (acceptance of $\mathcal{H}_0$ when it is false) indicates the number of deviant, and possibly fraudulent, telephone calls which are classified as normal by the test. This also needs to be kept to a very small level to ensure that the test is particularly sensitive to deviations from normal patterns of usage which may be highly indicative of fraudulent behavior.

The practical reality of such an anomaly detection system is that the false rejection rate (TYPE I error rate) will have to be controlled. Consider a number of independent and identically distributed random vectors $C_1, C_2, \cdots, C_N$ denoting the representation of N logged telephone calls and these have a probability distribution $P(C = \mathbf{c}|\mathbf{a}_m)$ under the null hypothesis, that is they are generated from a customer account $\mathbf{a}_m$. Further, there is a probability distribution, $P(C = \mathbf{c}|\mathcal{F})$, defining the distribution of telephone calls under the alternate hypothesis i.e. that another alternate signature, possibly a fraudulent one, is responsible for the generation of the phone calls. Then Neyman-Pearson state that the most powerful test (that which maximizes the power of the test $1-\beta$) for a fixed significance level α is obtained by using the likelihood ratio as the test statistic.

$$\lambda = \frac{\prod_{n=1}^{N} P(C = \mathbf{c}_n|\mathbf{a}_m)}{\prod_{n=1}^{N} P(C = \mathbf{c}_n|\mathcal{F})} \tag{1}$$

The null hypothesis will be rejected when the value of the test statistic λ is smaller than λ_{crit} such that $Prob(\lambda < \lambda_{crit} : \mathcal{H}_0 \text{ is true}) = \alpha$. Given the theoretically proven statistical optimality of the test we shall now adopt this as the core of the proposed method for detecting abnormal usage of a telephone service.

There are two areas which now require to be developed, that is the definition of the probability distributions under both $\mathcal{H}_0$ and $\mathcal{H}_1$, and the definition of λ_{crit} to set the level of significance of the tests, i.e. the false rejection rate. The section which follows elaborates on the definition of the required probability distributions.

3 Bayesian Multinomial-Dirichlet Account Profiles

Consider a population of customers, denoted by the set $\mathcal{A}$, each of whom have an account with the telecom provider. The m^{th} customer makes a series of N_m telephone calls during a given period $\mathcal{T}, \cdots, \mathcal{T}+\epsilon$, defined by $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_{N_m}]$, where each $\mathbf{c}_n$ defines the counts of the number of times that each of the events which defines a telephone call has occurred. The account for the m^{th} customer will be characterized by a consistent pattern of service usage over a given period of time $1, \cdots, \mathcal{T}-\delta$ from account initiation (time point 1) until δ time points prior to the set of telephone calls initiated during period $\mathcal{T}, \cdots, \mathcal{T}+\epsilon$. This will be reflected in a set of sufficient statistics, $\mathbf{a}_m$, describing the number of times, for this account, that a particular event related to the initiation and completion of a particular telephone service has occurred. This set of sufficient statistics, $\mathbf{a}_m$, will consist of, for example, the number of times a call is initiated in the morning between 6.00 am and midday, or the number of times that a call lasted longer than 15 minutes given that the call was international. For the purposes of this paper we define four independent sets of events which define a simple account model as `Day` of `Week` ($\mathcal{W}$), `Call Start Time` ($\mathcal{S}$), `Call Destination` ($\mathcal{D}$), `Call Duration` ($\mathcal{L}$) each denoted as $\mathrm{w} \in \mathcal{W}$, $\mathrm{s} \in \mathcal{S}$, $\mathrm{d} \in \mathcal{D}$, $\mathrm{l} \in \mathcal{L}$. If the number of possible values for each event is defined as $|\mathcal{W}|$, $|\mathcal{S}|$, $|\mathcal{D}|$, and $|\mathcal{L}|$ (e.g. seven days in the week to make telephone calls then $|\mathcal{W}| = 7$) in which case $\mathbf{a}_m = [a_{m,w=1,\cdots,|\mathcal{W}|}, a_{m,s=1,\cdots,|\mathcal{S}|}, a_{m,d=1,\cdots,|\mathcal{D}|}, a_{m,l=1,\cdots,|\mathcal{L}|}]^{\mathrm{T}} \in \mathcal{N}^{|\mathcal{W}|+|\mathcal{S}|+|\mathcal{D}|+|\mathcal{L}|}$. The definition for telephone calls, during the period $\mathcal{T}, \cdots, \mathcal{T}+\epsilon$, follows in a similar manner such that $\mathbf{c}_n = [c_{n,w=1,\cdots,|\mathcal{W}|}, c_{n,s=1,\cdots,|\mathcal{S}|}, c_{n,d=1,\cdots,|\mathcal{D}|}, c_{n,l=1,\cdots,|\mathcal{L}|}]^{\mathrm{T}} \in \mathcal{N}^{|\mathcal{W}|+|\mathcal{S}|+|\mathcal{D}|+|\mathcal{L}|}$.

We can consider further conditional events such as `Call Duration GIVEN Call Destination` and `Call Destination GIVEN Start Time`. However the detailed exposition of such a model shall be deferred to a further publication and, for now, we make the assumption that given the customer account all call related events are independent of each other i.e. $w \perp s \perp d \perp l | m$.

The series of telephone calls, $\mathbf{C}$, made during time period $\mathcal{T}, \cdots, \mathcal{T}+\epsilon$, are made with probability $P(\mathbf{C}|\mathbf{a}_m)$, in other words this series of calls were likely to have been made by the m^{th} customer account with probability $P(\mathbf{C}|\mathbf{a}_m)$. For the simplest case where it is assumed that each call is independent of all previous calls made from the customer account then $P(\mathbf{c}_1, \cdots, \mathbf{c}_N|\mathbf{a}_m) = \prod_{n=1}^{N} P(\mathbf{c}_n|\mathbf{a}_m)$.

Now each $P(\mathbf{c}_n|\mathbf{a}_m)$ will be defined by the distribution over the available features which characterize the call which in this case will be the day that the call was made (w), the time that the call was initiated (s), the destination of the call (d), and the duration of the call (l). Assuming conditional independence then $P(\mathbf{c}_n|\mathbf{a}_m) = P(l_n|\mathbf{a}_m)P(d_n|\mathbf{a}_m)P(s_n|\mathbf{a}_m)P(w_n|\mathbf{a}_m)$. What is now required is a representation of each account in terms of a set of parameters which define each of the conditional probability distributions employed in each $P(\mathbf{c}_n|\mathbf{a}_m)$. Assuming independence of the features each account, m, is then defined by the following set of multinomial parameters as $(\boldsymbol{\theta}_m^w, \boldsymbol{\theta}_m^s, \boldsymbol{\theta}_m^d, \boldsymbol{\theta}_m^l) \in [0,1]^{|\mathcal{W}|+|\mathcal{S}|+|\mathcal{D}|+|\mathcal{L}|}$ where the strictly positive parameters define the multinomial distributions such that $\sum_{i=1}^{|\mathcal{W}|} \theta_{m,i}^w = 1$, $\sum_{i=1}^{|\mathcal{S}|} \theta_{m,i}^s = 1$, $\sum_{i=1}^{|\mathcal{D}|} \theta_{m,i}^d = 1$, $\sum_{i=1}^{|\mathcal{L}|} \theta_{m,i}^l = 1$.

3.1 Bayesian Predictive Customer Account Profiles

The distribution over each of the required Multinomial parameters will be defined with a Dirichlet prior probability distribution which is the conjugate of the Multinomial likelihood [5] such that, for example, the `Start Time` parameters have a Dirichlet prior distribution defined as

$$P(\boldsymbol{\theta}_m^s|\boldsymbol{\alpha}_m^s) = \frac{\Gamma(\alpha_m^s)}{\prod_{i=1}^{|\mathcal{S}|}\Gamma(\alpha_{m,i}^s)} \prod_{i=1}^{|\mathcal{S}|} (\theta_{m,i}^s)^{\alpha_{m,i}^s - 1} \tag{2}$$

where $\alpha_m^s = \sum_{i=1}^{|\mathcal{S}|} \alpha_{m,i}^s$, each $\alpha_{m,i}^s \geq 0$ and Γ denotes the Gamma function. The corresponding multinomial likelihood for the start time is

$$P(\mathbf{a}_m|\boldsymbol{\theta}_m^s) = P(\mathbf{a}_{m,s}|\boldsymbol{\theta}_m^s) = \frac{a_{m,s}!}{\prod_{i=1}^{|\mathcal{S}|} a_{m,s=i}!} \prod_{i=1}^{|\mathcal{S}|} (\theta_{m,i}^s)^{a_{m,s=i}} \tag{3}$$

where $a_{m,s} = \sum_{i=1}^{|\mathcal{S}|} a_{m,s=i}$. Then the marginal distribution for the account based on, for example `Start Time` alone, is

$$\begin{aligned} P(\mathbf{a}_m|\boldsymbol{\alpha}_m^s) &= \int_{\boldsymbol{\theta}_m^s} P(\mathbf{a}_m|\boldsymbol{\theta}_m^s) P(\boldsymbol{\theta}_m^s|\boldsymbol{\alpha}_m^s) d\boldsymbol{\theta}_m^s \\ &= \frac{a_{m,s}!\Gamma(\alpha_m^s)}{\prod_{i=1}^{|\mathcal{S}|} a_{m,s=i}!\Gamma(\alpha_{m,i}^s)} \int \prod_{i=1}^{|\mathcal{S}|} (\theta_{m,i}^s)^{(a_{m,s=i}+\alpha_{m,i}^s-1)} d\boldsymbol{\theta}_m^s \\ &= \frac{a_{m,s}!\Gamma(\alpha_m^s)}{\prod_{i=1}^{|\mathcal{S}|} a_{m,s=i}!\Gamma(\alpha_{m,i}^s)} \frac{\prod_{i=1}^{|\mathcal{S}|}\Gamma(a_{m,s=i}+\alpha_{m,i}^s)}{\Gamma(a_{m,s}+\alpha_m^s)} \end{aligned} \tag{4}$$

and this follows for the other terms $P(\mathbf{a}_m|\boldsymbol{\alpha}_m^l)$, $P(\mathbf{a}_m|\boldsymbol{\alpha}_m^d)$ and $P(\mathbf{a}_m|\boldsymbol{\alpha}_m^w)$. Now the specific account profile is dependent on the Dirichlet parameters $\alpha_{m,i}^s$, $\alpha_{m,i}^l$, $\alpha_{m,i}^d$, and $\alpha_{m,i}^w$ as the multinomial parameters have been integrated out due to the conjugacy of the Multinomial-Dirichlet distributions.

3.2 Population Specific Priors

The parameters of the Dirichlet can be written as a product of a normalised measure over, for example, $\mathcal{S}$ in the case of `Start Time` and a positive real value, i.e. $\alpha_m^s = \sum_{i=1}^{|\mathcal{S}|} \alpha_{m,i}^s = \mu_m^s \sum_{i=1}^{|\mathcal{S}|} m_i^s = \mu_m^s$. The values of the parameters of the Dirichlet prior probabilities have a direct effect on the predictive probability assigned to a series of calls given a particular account. For the case where the prior parameter values for all variables is set to the value of one, then it is implicitly being assumed that all parameter values are equally likely *a priori* as for $\alpha_i^s = 1 \;\; \forall \;\; i$ then $P(\boldsymbol{\theta}_m^s) = \Gamma(|\mathcal{S}|)$. In practical terms, given that the m^{th} account is new and has made no calls then we are assuming that all possible behaviors or modes of service usage are equally likely to emerge.

The form of prior probability just discussed is particularly naive in that given the existing population of customer accounts $\mathcal{A}$ it ignores all the information available regarding specific characteristics of service usage from the population or market segmented parts of the population. Therefore in the absence of account specific information i.e. a brand new account, our prior should be guided by the population average signature (or the part of the market segment the new customer is attributed to) in which case each $\alpha_{m,i}^s = \mu_m^s m_i^s = \mu_m^s P(s_i|\mathcal{A})$ where the $P(s_i|\mathcal{A})$ is the probability of a call being initiated at `Start Time` equals i (the i^{th} start time event e.g Morning given the whole population of accounts $\mathcal{A}$. The values of the coefficients μ_m^s will be account specific and can be identified via some form of grid search. Alternatively to obtain the scalar values for each account $\mu_m^w, \mu_m^s, \mu_m^d, \mu_m^l$ we can employ Empirical Bayes [5] (Type II maximum likelihood) such that $\hat{\mu}_m^s = \underset{\mu_m^s}{\mathsf{argmax}} \; \log P(\mathbf{a}_m|\boldsymbol{\alpha}_m^s)$. Now denoting $f'(\mu_m^s) \equiv \frac{\partial}{\partial \mu_m^s} \log P(\mathbf{a}_m|\boldsymbol{\alpha}_m^s)$ and $f''(\mu_m^s) = \frac{\partial^2}{\partial^2 \mu_m^s} \log P(\mathbf{a}_m|\boldsymbol{\alpha}_m^s)$ then

$$f'(\mu_m^s) = \Psi(\mu_m^s) - \Psi(a_{m,s} + \mu_m^s) + \sum_{i=1}^{|\mathcal{S}|} P(s_i|\mathcal{A}) \left\{ \Psi(a_{m,s=i} + \mu_m^s P(s_i|\mathcal{A})) - \Psi(\mu_m^s P(s_i|\mathcal{A})) \right\}$$

$$f''(\mu_m^s) = \Psi'(\mu_m^s) - \Psi'(a_{m,s} + \mu_m^s) + \sum_{i=1}^{|\mathcal{S}|} P(s_i|\mathcal{A})^2 \left\{ \Psi'(a_{m,s=i} + \mu_m^s P(s_i|\mathcal{A})) - \Psi'(\mu_m^s P(s_i|\mathcal{A})) \right\}$$

where Ψ denotes the digamma function and these expressions can be employed in a Newton iteration, $\mu_m^s \leftarrow \mu_m^s - \frac{f'(\mu_m^s)}{f''(\mu_m^s)}$, for each attribute of every account. We now have to consider how to assign the required probabilities to a new sequence of calls originating from the accounts.

3.3 Predictive Likelihood of a Series of Calls

The posterior probability over the parameters[5] follows as

$$P(\boldsymbol{\theta}_m^s|\mathbf{a}_m, \boldsymbol{\alpha}_m^s) = \frac{\Gamma\left(a_{m,s} + \alpha_m^s\right)}{\prod_{i=1}^{|\mathcal{S}|} \Gamma(a_{m,s=i} + \alpha_{m,i}^s)} \prod_{i=1}^{|\mathcal{S}|} (\theta_{m,i}^s)^{(a_{m,s=i} + \alpha_{m,i}^s - 1)} \tag{5}$$

[5] For brevity we only show the expressions for `Start Time`, the required expressions for the other variables follow trivially.

Now for N calls made during the new period $\mathcal{T}, \cdots, \mathcal{T} + \epsilon$ we require the following probability[6]

$$P(\mathbf{C}|\mathbf{a}_m) = \prod_{n=1}^{N} P(\mathbf{c}_n|\mathbf{a}_m) = P(\mathbf{l}|\mathbf{a}_m)P(\mathbf{d}|\mathbf{a}_m)P(\mathbf{s}|\mathbf{a}_m)P(\mathbf{w}|\mathbf{a}_m) \tag{6}$$

where each $P(\mathbf{w}|\mathbf{a}_m) = \prod_{n=1}^{N} P(w_n|\mathbf{a}_m)$, $P(\mathbf{s}|\mathbf{a}_m) = \prod_{n=1}^{N} P(s_n|\mathbf{a}_m)$, $P(\mathbf{d}|\mathbf{a}_m) = \prod_{n=1}^{N} P(d_n|\mathbf{a}_m)$, $P(\mathbf{l}|\mathbf{a}_m) = \prod_{n=1}^{N} P(l_n|\mathbf{a}_m)$.

Now defining $c_{s,i} = \sum_{n=1}^{N_m} c_{n,s=i}$ and $c_s = \sum_{i=1}^{|\mathcal{S}|} c_{s,i}$ then

$$\begin{aligned} P(\mathbf{s}|\mathbf{a}_m, \boldsymbol{\alpha}_m^s) &= \int_{\boldsymbol{\theta}_m^s} P(\mathbf{s}|\boldsymbol{\theta}_m^s) P(\boldsymbol{\theta}_m^s|\mathbf{a}_m, \boldsymbol{\alpha}_m^s) d\boldsymbol{\theta}_m^s \\ &= \frac{c_s!}{\prod_{i=1}^{|\mathcal{S}|} c_{s,i}!} \frac{\Gamma\left(a_{m,s} + \alpha_m^s\right)}{\prod_{i=1}^{|\mathcal{S}|} \Gamma(a_{m,s=i} + \alpha_{m,i}^s)} \int \prod_{i=1}^{|\mathcal{S}|} (\theta_{m,i}^s)^{(c_{s,i} + a_{m,s=i} + \alpha_{m,i}^s - 1)} d\boldsymbol{\theta}_m^s \\ &= \frac{c_s!}{\prod_{i=1}^{|\mathcal{S}|} c_{s,i}!} \frac{\Gamma\left(a_{m,s} + \alpha_m^s\right)}{\prod_{i=1}^{|\mathcal{S}|} \Gamma(a_{m,s=i} + \alpha_{m,i}^s)} \frac{\prod_{i=1}^{|\mathcal{S}|} \Gamma(c_{s,i} + a_{m,s=i} + \alpha_{m,i}^s)}{\Gamma(c_s + a_{m,s} + \alpha_m^s)} \end{aligned} \tag{7}$$

and this follows for the other terms required, i.e. $P(\mathbf{l}|\mathbf{a}_m), P(\mathbf{d}|\mathbf{a}_m)$, and $P(\mathbf{w}|\mathbf{a}_m)$, to compute the predictive likelihood of the series of calls originating from the specific account $P(\mathbf{C}|\mathbf{a}_m)$.

So we see that the signatures for each account comprise of the sufficient statistics (counts of each event) and the estimated values of the parameters of the Dirichlet priors which require little storage overhead. The scoring of the series of calls amounts simply to the iterated application of the Gamma function in each term of (7) defining $P(\mathbf{C}|\mathbf{a}_m)$.

3.4 Defining Account Specific Threshold Levels

It is clear that account specific thresholds are also required to capitalize on the individual descriptive statistics. For a given level of test significance α each account will require a corresponding λ_{crit}^m value such that $Prob(\lambda_n^m < \lambda_{crit}^m : \mathcal{H}_0 \text{ is true}) = \alpha$. This has important practical consequences in that the False Rejection rate, the number of calls which are actually genuine being rejected by the system as inconsistent with the current profile, will be controlled by this value.

To this end we employ a form of Parametric Bootstrap [6] by using the above predictive distributions to repeatedly simulate a series of calls from each account, compute their associated scores and then obtain the empirical distribution of the scores. These account specific empirical distributions can then be used to obtain the account specific threshold scores which will yield the required test significance levels (i.e. TYPE I error rates).

4 Experimental Evaluation and Conclusion

We briefly report results based on a study conducted over a four month period in a city whose population is approximately half a million. A small number of fraudulent accounts had been identified by the telecom provider and the first three months of logged calls were used to build the required signatures for city based customer accounts ($\sim$ 150,000) with the final week of the remaining period being used for test purposes (4,000,000 telephone calls processed during the test period). Prior telecom operator knowledge of recent fraudulent activity was encoded in obtaining $P(C = \mathbf{c}_n|\mathcal{F})$ the *Fraudulent Signature*. Receiver Operator characteristic (ROC) curves have been employed in assessing overall performance. Two forms of prior distribution

[6] Conditioning on each α is implicit.

were assessed, one which assumed that all behaviors were equally likely to emerge from a new account and the other employed a population specific prior as discussed. From the Figures below we observe that the overall Area Under the Curve (AUC) is consistently higher for signatures employing population specific priors (Figure 2) which in practical terms translates to fewer alarms being raised in capturing all fraudulent accounts. We also note that the use of a *Fraudulent* signature informed by prior knowledge improves the AUC in both cases (Solid Lines for performance when a *fraudulent signature* employed, Dashed Lines when no *fraudulent signature* is used). This proposed signature-based Bayesian profiler forms a component of an overall commercial fraud detection system which is capable of managing signatures and scoring telephone activity on a nationwide scale.

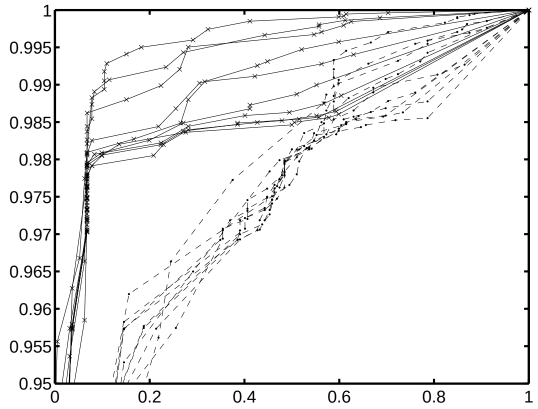


Figure 1: Naive Prior

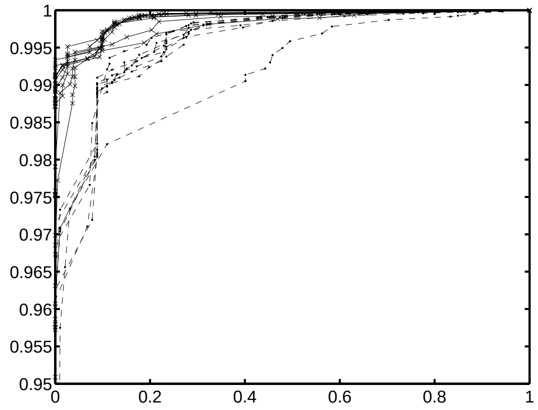


Figure 2: Population Prior

References

[1] Shawe-Taylor, J., Howker, K. and Burge, P. (1999) Detection of Fraud in Mobile Telecommunications. Information Security Technical Report 4(1):pp. 16-28.

[2] M. Cahill, D. Lambert, J. Pinheiro, and D. Sun., *Detecting Fraud in the Real World*, Handbook of Masive Data Sets, pp 911-929, 2002.

[3] C. Cortes, and D. Pregibon., *Signature Based methods for Data Streams*, Data Mining and Knowledge Discovery, 5, pp 167-182, 2001.

[4] T. Fawcett and F. Provost, *Adaptive Fraud Detection*, Data Mining and Knowledge Discovery, 1, pp 291-316, 1997.

[5] P. M. Lee, Bayesian Statistics: An Introduction, 3rd Edition, Hodder & Stoughton Educational, 2004.

[6] B. Efron and R. Tibshrani. An Introduction to the Bootstrap, Chapman and Hall, 1993.

Are We Really Discovering “Interesting” Knowledge From Data?

Alex A. Freitas
Computing Laboratory, University of Kent
Canterbury, CT2 7NF, UK
A.A.Freitas@kent.ac.uk , http://www.cs.kent.ac.uk/∼aaf

Abstract: This paper is a critical review of the literature on discovering comprehensible, interesting knowledge (or patterns) from data. The motivation for this review is that the majority of the literature focuses only on the problem of maximizing the accuracy of the discovered patterns, ignoring other important pattern-quality criteria that are user-oriented, such as comprehensibility and interestingness. The word “interesting” has been used with several different meanings in the data mining literature. In this paper interesting essentially means novel or surprising. Although comprehensibility and interestingness are considerably harder to measure in a formal way than accuracy, they seem very relevant criteria to be considered if we are serious about discovering knowledge that is not only accurate, but also useful for human decision making. The paper discusses both data-driven methods (based mainly on statistical properties of the patterns) and user-driven methods (which take into account the user’s background knowledge or believes) for discovering interesting knowledge. Data-driven methods are discussed in more detail because they are more common in the literature and are more controversial. The paper also suggests future research directions in the discovery of interesting knowledge.

1 Introduction

A well-known definition of knowledge discovery is as follows [7]:

> Knowledge Discovery in Databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

Although this definition is often quoted in the literature, in general it has not been taken very seriously by the research community. This claim is supported by the fact that the vast majority of data mining works focus on discovering knowledge that is accurate - e.g. trying to maximize predictive accuracy in the classification task. This can be considered as aiming to discover valid patterns, and perhaps aiming at discovering “potentially useful” patterns - to the extent that we believe that there is a high positive correlation between the accuracy of a pattern and its usefulness to the user. However, in practice the correlation between predictive accuracy and usefulness of discovered patterns is not so clear, and the focus on maximizing predictive accuracy does not seem to improve the chances of discovering novel, ultimately understandable patterns in the data. Actually, it is often the case that focusing on maximizing predictive accuracy only - ignoring other criteria to evaluate the quality of patterns - significantly harms the discovery of understandable, novel and useful knowledge. A few examples can illustrate this point, as follows.

Brin et al. [2] found, in a Census dataset, several rules which were very accurate but were also useless, because they represented obvious patterns in the data, such as “five year olds don’t work”, “unemployed residents don’t earn income from work” and “men don’t give birth”. Tsumoto [22] found 29,050 rules, out of which only 220 (less than 1% of them) were considered interesting or unexpected by the user. These two works are examples of the fact that high accuracy is not a sufficient condition for the usefulness or interestingness (novelty or surprisingness) of a pattern. In addition, high accuracy is not always a necessary condition for the usefulness or interesting of a pattern. For instance, Wong and Leung [23] found rules with 40-60% confidence that were considered, by senior medical doctors, novel and more accurate than the knowledge of some junior doctors.

The goal of this paper is to contribute to a better understanding of the limitations of the concepts and techniques used to discover comprehensible and interesting patterns. Hence, this paper can be considered a critical review of the literature on the discovery of comprehensible, interesting patterns. By "interesting" we mean novel or surprising. Note that we consider interestingness and comprehensibility to be different quality criteria, since patterns such as "men don't give birth" are comprehensible but not interesting at all. Hence, this paper focuses on two out of the four pattern-quality criteria mentioned in Fayyad et al [7]. Concerning the other two criteria, we interpret "valid" essentially as "accurate", a pattern-quality criterion that is not discussed here because it is already extensively discussed in the literature; and we follow Silberchatz and Tuzhilin [18] in using surprisingness or novelty as a proxy for usefulness, because usefulness is a concept whose formalization seems elusive. The remainder of this paper is organised as follows. Section 2 discusses the discovery of comprehensible (understandable) patterns. Section 3 discusses the discovery of interesting (novel or surprising) patterns. Section 4 presents the conclusions and future research directions.

2 On the Discovery of Comprehensible Patterns

In many application domains, in order for the user to trust the discovered patterns and make important decisions based on them, it is usually necessary that the user understand those patterns. For instance, in principle a medical doctor should not blindly trust the diagnosis output by a black box classification algorithm and recommend a surgery for the patient based just on that automatic diagnosis. The doctor should interpret the discovered patterns in the context of her/his previous knowledge about the application domain. Similarly, a user would probably hesitate in investing a large amount of money in a financial application based on some pattern automatically discovered by a black box prediction algorithm. In addition, in some applications the reason for a decision must be explained for legal reasons, which requires that the patterns on which the decision was based be understandable.

This does not mean that comprehensibility is always important. In principle the need for understandable patterns arises when the patterns will be used to support a decision to be made by a human user. In some applications the discovered patterns will be automatically used by a machine rather than support a human decision, and so they do not need to be understandable. A typical example is the pattern recognition task of automatically recognizing the post code in a letter and sending the letter to a pigeon hole containing letters for the appropriate destination.

In any case, in applications where a human user would like to make important, strategic decisions based on the discovered patterns, intuitively the comprehensibility of the discovered patterns improves the potential usefulness of those patterns - although of course just comprehensibility by itself is not guarantee that the patterns will be really useful to the user. Despite the importance of comprehensibility, there has been little progress towards techniques that improve the comprehensibility of discovered patterns. In general we can say that some knowledge representations lend themselves more naturally to comprehensible patterns than others. For instance, most researchers would agree that representations such as decision trees, IF-THEN rules or Bayesian networks tend to be more comprehensible than, say, neural networks or support vectors. However, as pointed out by Pazzani [14], there is no consensus on which of these representations is the most comprehensible in general, and there seems to be no cognitive psychology study comparing the comprehensibility of different representations from the point of view of human users. Pazzani also suggests some cognitive psychology-related criteria for evaluating pattern comprehensibility, such as the criterion that the pattern should be consistent with the user's prior knowledge, but there has been relatively little work in this area. In any case, note that although the criterion of consistency with prior knowledge tends to improve comprehensibility, intuitively it tends to hinder the discovery of novel or surprising patterns - see Section 3.

As for the usual measure of "comprehensibility" or "simplicity" often used in the literature, which consists of measuring the size (number of conditions or nodes) of a rule set, decision tree or Bayesian net, it should be noted that this is just a measure of syntactical simplicity, which is very different from semantic simplicity (which would need to involve the meaning of the attributes in the conditions or nodes of the rules, decision tree or Bayesian net). In any case, if a large number of patterns are discovered, one possibility to reduce the user's cognitive workload in interpreting the discovered patterns consists of selecting a subset of the most "interesting" (novel or surprising) patterns - using, for instance, some of the methods discussed in Section 3 - and show just those selected patterns to the user.

3 On the Discovery of Interesting (Novel or Surprising) Patterns

There are two basic approaches to discover novel or surprising (unexpected) patterns, namely the user-driven (or "subjective") approach and the data-driven (or "objective") approach. In essence, the user-driven approach is based on using the domain knowledge, beliefs or preferences of the user; whilst the data-driven approach is based on statistical properties of the patterns. Hence, the data-driven approach is more generic, independent of the application domain. This makes it easier to use this approach, avoiding difficult issues associated with the manual acquisition of the user's background knowledge and its transformation into a computational form suitable for a data mining algorithm. On the other hand, the user-driven approach tends to be more effective at discovering truly novel or surprising knowledge to the user, since it explicitly takes into account the user's background knowledge. This raises the question of to what extent the data-driven approach is effective in discovering interesting patterns to the user - an issue that will be discussed in Subsection 3.2.

3.1 User-Driven Methods for Discovering Interesting Patterns

A classic example of a user-driven method for discovering interesting patterns is the use of user-specified templates in the context of association rules [9]. In this case the user essentially specifies inclusive templates - indicating which items the user is interested in (among a large number of items available in the database) - and restrictive templates - indicating which items the user is not interested in. Then an association rule is considered interesting if it matches at least one inclusive template and it matches no restrictive template.

Another example of user-driven method is the use of user-defined general impressions [11, 17]. In this case the user specifies general impressions in the form of IF-THEN rules, such as "IF (salary = high) AND (education-level = high) THEN (credit = good)". Note that this is a general impression because its conditions are not precisely defined. By contrast, the data mining algorithm is supposed to produce rules with well-defined conditions, such as "$salary > £50K$". Once such rules are produced by the data mining algorithm, the system can match the rules with the general impressions, in order to find surprising rules. In particular, if a rule and a general impression have similar antecedents ("IF part") but different consequents ("THEN part"), the rule can be considered surprising, in the sense of contradicting a user's belief (general impression). For instance, the rule "IF ($salary > £50k$) AND (education-level $\geq$ BSc) AND (Mortgage = yes) THEN (credit = bad)" would be considered surprising with respect to the aforementioned general impression.

3.2 Data-Driven Methods for Discovering Interesting Patterns

There are more than 50 measures of rule quality that have been called rule "interestingness" measures in the literature. A review of these measures can be found in [8, 21]. One classical example of these data-driven rule interestingness measures is the one proposed by Piatetsky-Shapiro [16], defined as $Interest = |A \cap C| - (|A| \times |C|)/N$, where $|A \cap C|$ is the number of examples satisfying both the rule antecedent A and the rule consequent C, $|A|$ (alternatively, $|C|$) is the number of examples satisfying the rule antecedent A (rule consequent C), and N is the total number of examples. Hence, Interest is a measure of the deviation from statistical independence between A and C. Note that it measures the symmetric correlation between A and C, and not an asymmetric implication, i.e., Interest has the same value for the two"opposite" rules: IF A THEN C, IF C THEN A.

Until a few years ago, in general works proposing data-driven rule interestingness measures implicitly assumed that such measures were correlated with the user's real, subjective interest in the rules, and typically papers using those measures did not report any subjective evaluation of the rules by the user. More recently, some works have reported the results of experiments to assess to what extent the values of data-driven rule interestingness measures are correlated with the real, subjective interest of the user. The methodology used for this assessment can be summarized in three steps, namely: (a) rank the discovered rules according to each of a number of data-driven rule interestingness measures; (b) show (a subset of) the discovered rules to the user, who assigns an "interestingness score" to each rule based on her/his subjective interest in the rule; and (c) measure the linear correlation (or another measure of association) between the ranking of each data-driven rule interestingness measure and the real, subjective human interest on the rules. A couple of experiments following the basic idea of this methodology are as follows.

Ohsaki et al. [13] have done experiments with 39 data-driven rule interestingness measures, involving rules discovered from a hepatitis dataset. They report the results of two experiments. In the first one the highest correlation between a rule interestingness measure (out of the 39 measures) and the user's real interest was just 0.48, and only one measure had a correlation greater than or equal to 0.4 (on a scale from -1 to +1). In the second experiment the highest correlation was again 0.48, and only four rule interestingness measures had a correlation greater than or equal to 0.4. (It should be noted, though, that the paper also reports other indicators of performance of the rule interestingness measures, according to which those measures seem to obtain better results.)

Carvalho et al. [4] have done experiments with 11 data-driven rule interestingness measures, involving 8 datasets and one user for each dataset. Out of the 88 reported correlation values (involving 11 rule interestingness measures for each of 8 users), 31 correlation values were greater than or equal to 0.6. The correlation values associated with each measure varied considerably across the 8 datasets/users, so that no single rule interestingness measure performed consistently well across all datasets/users. In addition, more recent results reported in [3], in experiments involving 45 users (9 datasets and 5 users per data-set), suggest that, overall, the correlation between data-driven measures and real human interest is considerably lower than the correlation results obtained with 8 users in [4].

The aforementioned results support the intuitive argument that it is difficult to use a purely data-driven approach for discovering patterns that are truly novel or surprising to the user. There are some works that try to reduce this strong limitation of the data-driven approach, using not only statistical properties of the rules but also concepts or ideas that intuitively seem more likely to lead to the discovery of interesting patterns - although the extent to which these ideas capture real human interest seems somewhat controversial. Let us now briefly review some of these works.

One approach consists of automatically learning which combination of a number of data-driven rule interestingness measures is a good predictor of real human interest, as proposed by Abe et al. [1]. This work involves a kind of "meta-learning", constructing a meta-dataset where each meta-example corresponds to a classification rule discovered from a dataset, the 39 predictor meta-attributes are values of 39 data-driven rule interestingness measures for each of the meta-examples (rules) and the class meta-attribute is the user's real, subjective interest in each of the rules. (So, this is a hybrid data/user-driven approach.) The values of the class meta-attribute are manually specified by the user in the meta-training set and automatically predicted by the algorithm in the meta-test set. The authors applied five different classification algorithms to the meta-dataset, and report that the best predictive accuracy - measured by leave-one-out - was 81.6%. This seems a good result, but it should be noted that different classification algorithms selected different meta-attributes for the classification model.

Another approach consists of using a data-driven rule interestingness measure that is "more surprisingness-oriented" than the mere use of statistical properties, in particular discovering exception rules, as follows. Let $R1$ be a general rule of the form "IF $Cond1$ THEN $Class1$", and let $R2$ be an exception rule of the form "IF $Cond1$ AND $Cond2$ THEN $Class2$", where $Cond1$, $Cond2$ are conjunctions of conditions. Note that rule $R2$ is a specialization of, and predicts a different class from, rule $R1$. Hence, $R2$ is an exception of $R1$. In this kind of data-driven interestingness method, the exception rule $R2$ can be considered an interesting rule if both $R2$ and its generalized rule $R1$ have a high predictive accuracy. Rule interestingness measures based on these ideas are discussed, e.g. in [20, 19]. The rationale for this exception-based approach is that users tend to know the general data relationships in their application domain, but are less likely to know exceptions to those general relationships. Hence, exception rules tend to be more surprising or novel to users than general rules. A real-world example involves car accident data [19], where, in addition to the known general rule "IF (*used-seat-belt* = yes) THEN ($injury$ = no)", the system also discovered the surprising exception rule "IF (*used-seat-belt* = yes) AND ($passenger$ = child) THEN ($injury$ = yes)".

Another surprisingness-oriented data-driven method consists of discovering instances of Simpson's paradox in data, as follows. Let the event C be the apparent "cause" of an event E, the "effect". Simpson's paradox occurs if the event C increases the probability of the event E in a given population *Pop* and, at the same time, decreases the probability of event E in every subpopulation of *Pop* [15]. Let Z and $\neg Z$ denote two complementary values of a confounding variable, representing complementary properties describing two subpopulations of *Pop*. Then, mathematically, Simpson's paradox occurs if the following 3 inequalities hold for a given data set:

$$P(E|C) > P(E|\neg C), P(E|C,Z) < P(E|\neg C,Z), P(E|C,\neg Z) < P(E|\neg C,\neg Z),$$

where $P(X|Y)$ denotes the conditional probability of X given Y.

A classic example of Simpson's paradox occurred in a comparison of tuberculosis deaths in New York City and Richmond, Virginia, in 1910. Overall, the tuberculosis mortality rate of Richmond was higher than New York's one. However, the opposite was observed when the data was partitioned according to two racial categories: white and non-white. In both the white and non-white categories, Richmond had a lower mortality rate. In this example, the events C and $\neg C$ are Richmond and New York, the event E is tuberculosis death, and the events Z and $\neg Z$ are the categories white and non-white. A number of other occurrences of the paradox in real-world data are reported in [5], [6] and [10]. The two works by Fabris and Freitas also describe algorithms that systematically search for occurrences of Simpson's paradox in data.

Although Simpson's paradox is well-known by statisticians, it is usually surprising to data mining users, who typically have no formal statistical training. This makes the automatic detection of Simpson's paradox one of the few data-driven methods for discovering patterns that are likely to be considered surprising according to a user's subjective evaluation [12].

4 Conclusions and Future Research Directions

This paper presented a critical review of the current concepts and methods used for discovering comprehensible and interesting (novel or surprising) patterns in data. This is an important topic, because most works focus only on maximizing pattern accuracy (since accuracy is easier to measure), ignoring other aspects of pattern quality that, although harder to measure, are clearly related to the usefulness of the discovered patterns to the user.

We have discussed several methods for discovering interesting patterns, based on either a data-driven or a user-driven approach. The data-driven approach is normally easier to implement, but, since it does not take into account the user's domain knowledge, it has difficulty in discovering truly interesting knowledge to the user. In particular, recent results suggest that the effectiveness of a number of data-driven rule interestingness measures has been overrated in the literature. Three kinds of method that try to overcome some limitations of a data-driven approach based only on statistical properties of the data have been discussed, in particular: (a) a "meta-learning" method using a classification algorithm to learn which combination of data-driven rule interestingness measures best predicts the user's rule interest; and methods oriented towards the discovery of surprising patterns, namely: (b) the discovery of exception rules (which are less likely to be known by users than general rules); and (c) the discovery of instances of Simpson's paradox (which tend to be surprising to the user due to the nature of the "paradox"). However, even in the case of these methods there is not enough empirical evidence in the literature to show that they are effective in discovering patterns that are really interesting to the user, since most of the papers on these methods do not report the subjective evaluation of the discovered patterns by the user.

One research direction would be to try to significantly reduce the bottleneck of the user-driven approach, the manual acquisition of the user's background knowledge, by using text mining to automatically generate background knowledge about the application domain from the published literature. For instance, instead of asking the user to specify a comprehensive set of general impressions representing her/his background knowledge, in principle (at least in some application domains) a text mining algorithm could automatically extract general impressions from the literature. Presumably the user should still be in the loop to validate the general impressions discovered by the text mining algorithm, but intuitively it would be easier for the user to validate automatically-discovered general impressions than to specify a large number of general impressions herself/himself.

Another research direction would be to develop methods for discovering interesting patterns from the start of the KDD process - i.e. in the data preparation phase, rather than methods to be applied in the data mining phase or in the knowledge post-processing phase. For instance, current attribute selection methods in general are designed for maximizing the predictive accuracy of the data mining algorithm, and those methods normally show no concern for the interestingness (novelty or surprisingness) of the patterns to be discovered by the data mining algorithm.

References

[1] H. Abe, S. Tsumoto, M. Ohsaki, and T. Yamaguchi. Evaluating a rule evaluation support method with learning models based on objective rule evaluation indices. In *Proc. 5th Int. Conf. Hybrid Intelligent Systems (HIS-2005)*, pages 169–174. PUC-Rio, Rio de Janeiro, Brazil, 2005.

[2] S. Brin, R. Motwani, J.D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *Proc. KDD-97*. AAAI Press, 1997.

[3] D.R. Carvalho. *A decision tree / genetic algorithm to cope with the problem of small disjuncts (In Portuguese)*. Phd thesis, Federal University of Rio de Janeiro, Brazil, 2005.

[4] D.R. Carvalho, A.A. Freitas, and N.F. Ebecken. Evaluating the correlation between objective rule interestingness measures and real human interest. In *Proc. PKDD-2005, LNAI 3721*, pages 453–461. Springer, 2005.

[5] C.C. Fabris and A.A. Freitas. Discovering surprising patterns by detecting instances of Simpson's paradox. In *Research and Development in Intelligent Systems XVI*, pages 148–160. Springer, 1999.

[6] C.C. Fabris and A.A. Freitas. Discovering surprising instances of Simpson's paradox in hierarchical multi-dimensional data. *Int. J. on Data Warehousing & Mining*, 2(1):26–48, 2006.

[7] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: an overview. In *Advances in Knowledge Discovery and Data Mining*, pages 1–34. AAAI Press, 1996.

[8] R.J. Hilderman and H.J. Hamilton. *Knowledge Discovery and Measures of Interest*. Kluwer, 2001.

[9] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A.I. Verkamo. Finding interesting rules from large sets of discovered association rules. In *Proc. 3rd Int. Conf. on Information and Knowledge Management*, pages 401–407, 1994.

[10] R. Kohavi. Focusing the mining beacon: lessons and challenges from the world of e-commerce. Invited talk at PKDD-2005, January 2005. www.kohavi.com.

[11] B. Liu, W. Hsu, and S. Chen. Using general impressions to analyze discovered classification rules. In *Proc. KDD-97*, pages 31–36. AAAI Press, 1997.

[12] K. McGarry. A survey of interestingness measures for knowledge discovery. *Knowledge Engineering Review J.*, 20(1):39–61, 2005.

[13] M. Ohsaki, S. Kitaguchi, K. Okamoto, H. Yokoi, and T. Yamaguchi. Evaluation of rule interestingness measures with a clinical dataset on hepatitis. In *Proc. PKDD-2004*, pages 362–373. Springer, 2004.

[14] M.J. Pazzani. Knowledge discovery from data? *IEEE Intellig. Sys.*, pages 10–13, Mar/Apr 2000.

[15] J. Pearl. *Causality: models, reasoning and inference*. Cambridge Univ. Press, 2000.

[16] G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. In *Knowledge Discovery in Databases*, pages 229–248. AAAI/MIT Press, 1991.

[17] W. Romao, A.A. Freitas, and I.M.S. Gimenes. Discovering interesting knowledge from a science & technology database with a genetic algorithm. *Applied Soft Computing*, 4:121–137, 2004.

[18] S. Silberchatz and A. Tuzhilin. What makes patterns interesting in knowledge discovery systems. *IEEE Trans. Knowledge and Data Engineering*, 8(6), 1996.

[19] E. Suzuki. Discovering interesting exception rules with rule pair. In *Proc. Workshop on Advances in Inductive Rule Learning at PKDD-2004*, pages 163–178, 2004.

[20] E. Suzuki and Y. Kodratoff. Discovery of surprising exception rules based on intensity of implication. In *Proc. PKDD-98, LNAI 1510*, pages 10–18. Springer, 1998.

[21] P-N. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *Proc. ACM SIGKDD KDD*, pages 32–41. ACM Press, 2002.

[22] S. Tsumoto. Clinical knowledge discovery in hospital information systems: two case studies. In *Proc. PKDD-2000 LNAI 1910*, pages 652–656. Springer, 2000.

[23] M.L. Wong and K.S. Leung. *Data mining using grammar based genetic programming and applications*. Kluwer, 2000.

BCS
THE BRITISH COMPUTER SOCIETY
SGAI
SPECIALIST GROUP ON
ARTIFICIAL INTELLIGENCE

The Group has links with many outside bodies. It is a specialist group of the British Computer Society and a member of ECCAI, the European Co-ordinating Committee for Artificial Intelligence.

Since its inception the group has enjoyed a good working relationship with government departments involved in the AI field (beginning with the Alvey Programme in the 1980s). A succession of Department of Trade and Industry (DTI) representatives, have been co-opted as committee members. The Group acted as co-organiser of the annual DTI Manufacturing Intelligence awards and has included sessions presenting the results of the DTI Intelligent Systems Integration Programme (ISIP) in its annual conferences.

The group also has a good relationship with the Institution of Electrical Engineers (IEE), with which it has co-sponsored colloquia over many years, and with NCAF, the Natural Computing Applications Forum. We also host the annual UK-CBR (Case-Based Reasoning) workshops at our annual conferences. This year, the group will be co-hosting IJCAI-05 in Edinburgh.

Benefits of Membership

- Preferential rates for the Group's prestigious international conference on Artificial Intelligence, which has run annually since 1981.
- Discounted rates at other SGAI-sponsored events.
- Discounted rates for ECCAI organised events. The Group has been a member of the European Co-ordinating Committee for Artificial Intelligence since 1992.
- Discounts on international journals, and occasional special offers on books.
- Advance information on the SGAI Evening Lectures, which are held on a regular basis in central London.
- Free subscription to the *Expert Update* journal, containing reviews, technical articles, conference reports, comment from industry gurus and product news.
- The SGAI website at www.bcs-sgai.org and the AI-SGES list server to facilitate communication on all aspects of AI.
- A substantial proportion of the Group's membership is from industry. Providing a valuable forum where both academic and industrial AI communities can meet.

How to Join BCS-SGAI?

To join BCS-SGAI you do not need to be a member of the BCS. For further information please visit our website at www.bcs-sgai.org.

Subscription Rates

Annual subscription rates for Individual and Corporate Members are:

INDIVIDUALS

Standard Members (UK addresses)	£31.00
Standard Members (Overseas addresses)	£41.00
BCS Members (UK addresses)	£22.00
BCS Members (Overseas addresses)	£31.00
Students (UK addresses)	£11.00
Students (Overseas addresses)	£21.00
Proof of student status is required	
Retired (UK addresses)	£11.00
Retired (Overseas addresses)	£21.00

CORPORATE

UK addresses	£150.00
Overseas addresses	£190.00

Add £5 to all these rates if not paying by standing order.

UK KDD Symposium (UKKDD'07)

Wednesday 25 April 2007
University of Kent

Sponsored by the BCS-SGAI

Aims and Objectives

The philosophy behind the UK-KDD series of symposia is to establish and maintain a national forum to allow KDD practitioners to present and exchange their ideas within the KDD research community. The series seeks:

1. To present a collective view of the current "state of the art" of KDD research work, currently in progress within the UK, to commercial interests, academics, UK based researchers and post-graduate students.
2. To contribute to the overall quality of Computer Science research within the UK.
3. To emphasise and maintain the leading role of the UK within the global KDD Community.
4. To identify future directions and opportunities for the UK KDD community.

The inaugural UK-KDD symposium was held on Wednesday 6 April 2005 in Liverpool. It was a very popular event with over 50 delegates. The second symposium was held on Wednesday 26 April 2006 in Norwich, and attracted 110 delegates. It is intended that third symposium will build on this momentum, initiated by the first two UKKDD symposia, to maintain a common forum in the UK to facilitate the exchange of views and ideas within the UK KDD community.

Speakers

- **Beatriz de la Iglesia** (University of East Anglia)
- **Frans Coenen** (University of Liverpool)
- **Duncan Ross** (Advanced Analytics for Teradata)
- **James Cussens** (University of York)
- **Niall Rooney** (University of Ulster)
- **Andrew Secker** (University of Kent)
- **Jenny Harding** (Loughborough University)

Further Details

Contact	Email
Alex Freitas (Chair), University of Kent	`A.A.Freitas@kent.ac.uk`
Frans Coenen, University of Liverpool	`frans@csc.liv.ac.uk`
George Smith, University of East Anglia	`gds@ cmp.uea.ac.uk`