THE UNIVERSITY OF
WARWICK

Original citation:
Gibson, Jon, Tekiner, Firat, Halfpenny , Peter, Nazroo, James Y., Fagan, Colette, Procter , Robert N. and Lin, Yuwei (2007) CF URGENT: NCeSS Project : Data mining for social scientists. In: 3rd International Conference on e-Social Science, Ann Arbor, Michigan, USA, 7-9 Oct 2007. Published in: Proceedings of the Third International Conference on e-Social Science pp. 1-8.
Permanent WRAP url:
http://wrap.warwick.ac.uk/52923

warwick**publications**wrap

highlight your research

http://wrap.warwick.ac.uk/

NCeSS Project: Data Mining for Social Scientists

Jon Gibson[1], Firat Tekiner[2], Peter Halfpenny[3], James Nazroo[4], Colette Fagan[4], Rob Procter[3] and Yuwei Lin[3].

[1]Research Computing Services, University of Manchester, U.K.

[2]National Centre for Text Mining, University of Manchester, U.K.

[3]National Centre for e-Social Science, University of Manchester, U.K.

[4]School of Social Sciences, University of Manchester, U.K.

jon.gibson@manchester.ac.uk

Abstract. We will discuss the work being undertaken on the NCeSS data mining project, a one year project at the University of Manchester which began at the start of 2007, to develop data mining tools of value to the social science community. Our primary goal is to produce a suite of data mining codes, supported by a web interface, to allow social scientists to mine their datasets in a straightforward way and hence, gain new insights into their data. In order to fully define the requirements, we are looking at a range of typical datasets to find out what forms they take and the applications and algorithms that will be required. In this paper, we will describe a number of these datasets and will discuss how easily data mining techniques can be used to extract information from the data that would either not be possible or would be too time consuming by more standard methods.

Data Mining

A useful definition is that "data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner" (Hand et al, 2001). In recent years there has been a huge increase in economic, marketing and financial databases, with examples including government surveys and statistics, supermarket sales information and minute by minute stock prices. The amount of data generated and collected is overwhelming for the typical user. Data Mining is the process of analyzing databases to discover and extract knowledge in a more automated way than with classical techniques.

Data mining tasks are usually divided into two major categories, predictive tasks and descriptive tasks (Tan et al, 2006). The former is used to predict the value of a particular attribute based on the values of the other attributes, whereas the latter is concerned with deriving patterns (correlations, clusters, trajectories and anomalies) that summarise the underlying relationships in the data. The most common predictive modelling tasks are classification (used for discrete values) and regression (used for continuous target values). On the other hand, association analysis is used to discover patterns that describe associated features in the data, cluster analysis seeks to group attributes that are closely related to each other and anomaly detection is concerned with identifying attributes whose characteristics are significantly different from the others.

An important aspect of social science research involves the analysis of survey data and the case studies that we look at here all fall into this category. Its main purpose here is descriptive. By helping explain and summarise the underlying relationships in the data, it will hopefully suggest further areas for investigation and provide new insights into the data. The data mining algorithms we use in these cases are of the clustering (or cluster analysis) type. Clustering has already been used in many application domains, including biology, medicine, anthropology, marketing and economics. Clustering applications include plant and animal classification, disease classification, image processing, pattern recognition and document retrieval.

Clustering

Clustering assigns the data elements into *clusters*, which are groups of data whose elements have similar characteristics (Dunham, 2003). Conversely, elements from different clusters are dissimilar relative to those within a cluster. Clustering algorithms can be either hierarchical or partitional. With hierarchical clustering, a hierarchy of clusters is created. At the lowest level, each element is a cluster in itself and at the highest, all the elements comprise a single large cluster. With partitional clustering, on the other hand, the algorithm creates only one set of clusters and the number of clusters generated is an input parameter to the routine. There are many different algorithms within each of these categories, each having its own properties and features.

Most of the clustering algorithms in our study were run using the Weka package (Witten and Frank, 2005), although cluto was also used. Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato. Weka is free software available under the GNU General Public License. Weka supports all the standard data mining tasks, including a number of clustering algorithms, via both a GUI and a command-line interface and allows some preprocessing of the data and visualisation of the data and results. Weka provides seven different clustering algorithms: Cobweb, DBScan, Expectation-Maximisation (EM), Farthest First, Optics, K-means and X-means.

K-means

This is a simple partitional algorithm, in which the user specifies at the start the desired number of clusters, the parameter *k* of the name. Initially, *k* points are randomly selected as cluster centres and instances are assigned to the cluster whose centre they are nearest to according to the standard Euclidian distance function. The centroid, or mean, of all the instances in each cluster is then calculated, the *means* part of the algorithm (alternatively, modes can be used with categorical data). These centroids become the new centre values for their respective clusters and the whole process repeats. It will continue to iterate in this way until the contents of each cluster, and hence also the cluster centres, does not change between consecutive iterations.

As is common with clustering techniques, the cluster centres do not necessarily converge to a global minimum but a local one. It is therefore quite possible that starting from a different set of random cluster centres will lead to completely different clusters being found. For this reason, it is important to perform a number of different runs with varying initial conditions. The time complexity of the algorithm is $O(tkn)$, where t is the number of iterations and n is the number of instances. There are also a number of variations on this basic algorithm.

Weka's x-means is one of these but since it cannot handle nominal (categorical) attributes, it is unsuitable for our social science datasets.

Expectation-Maximisation (EM)

This method does not actually assign a given instance to a cluster but calculates the probability of it belonging to each one. For the purposes of analysing the results, the user is free to assume that each instance belongs to the cluster for which it has the greatest probability. Each cluster is assumed to be represented by a probability distribution that gives the probability that a given instance would have a certain set of attribute values if it was a member of that cluster. Each cluster will have a different distribution. They will not be equally likely and so their relative populations are also reflected in a probability distribution. Generally, each cluster is assumed to have a Gaussian, or normal, distribution (although others can be used) with different means and variances. Each cluster can therefore be characterised by its mean, its variance and the overall probability of an instance belonging to that cluster (i.e. the proportion of the population belonging to the cluster).

The EM algorithm itself is very similar to that used in k-means clustering. Initial guesses are made of each cluster's parameters. The cluster probabilities are then calculated for each instance, the *expectation* stage of the name. In the second *maximisation* stage, the distribution parameters are calculated so as to maximise the likelihood of the distributions given the data. A more detailed description of the algorithm can be found in Witten and Frank (2005). At the end of each iteration, an overall likelihood is calculated by multiplying the probabilities of the individual instances. This value is not actually a probability but its magnitude does reflect the quality of the clustering and will increase with every iteration of the EM algorithm. In practice, the logarithm of this value is actually calculated and the measure is known as the *log-likelihood*. The algorithm therefore should proceed until successive values of the log-likelihood are within a specified tolerance.

As with k-means, the EM algorithm is only guaranteed to converge to a local maximum, not the global one and so the procedure should be repeated a number of times with different initial guesses for the parameter values. In this case, however, the log-likelihood figure can be used to directly compare the final configurations obtained and so the user just has to choose the largest of the local maxima.

Some Other Clustering Algorithms

Weka's only hierarchical algorithm is called Cobweb. This is also an *incremental* algorithm. While the other routines we have looked at iterate over the whole dataset, this one adds instances into the cluster hierarchy one at a time. Updating could just mean finding the right place to put a new instance or could mean a radical restructuring of the part affected. Details of the algorithm can be found in Witten and Frank (2005).

DBScan (Density-Based Spatial Clustering of Applications with Noise) finds sets of clusters of a minimum specified size and density, where density is defined as a minimum number of points within a certain distance of each other. Due to these restrictions, the number of clusters is not known in advance and some points will not be assigned to clusters – these are treated as noise. The algorithm is explained in Dunham (2005) and has a time complexity of *O(n log n)*.

The Farthest First algorithm is an implementation of the "Farthest First Traversal Algorithm" by Hochbaum and Shmoys (1985). It finds fast, approximate clusters and may be useful as an initialiser for k-means.

Method

Weka uses a data format called ARFF (Attribute-Relation File Format). An ARFF file is an ASCII text file in a format that first describes all the attributes in a header section and then contains all the data itself. Weka can convert a CSV (comma-separated values) file, a format into which most packages can save their data, into the ARFF format but in many cases it is preferable to do this as a separate preprocessing stage using, for example, a perl script, in order to get the types of the attributes correct (numerical, nominal, etc) and to have complete control over the data file. Consideration also has to be given to how to deal with any missing values in the data before any clustering can be done.

It is important to realise that there is no one correct answer to a clustering problem and that many different answers may usefully be found. Similarly, the number of clusters required is often unclear and may require guidance from domain experts. Also, there may be no a priori knowledge concerning the attributes of the different clusters. In fact, interpreting the meaning of each cluster may be difficult and require a domain expert to do this. Different sets of clusters will also be found by clustering on different sets of variables in the dataset. If the user knows in advance the desired cluster characteristics, then a specific subset of the attributes may be appropriate. Otherwise, it may be sensible to use nearly all the attributes, only omitting those such as ID numbers that carry no information. The best clustering algorithm to choose for a given problem will similarly depend on the nature of the dataset and what is being looked for. It may be sensible to do a number of investigative runs.

Compared to other types of data mining, where there are generally objective measures of success, such as a prediction being correct or not, clustering is intrinsically difficult to evaluate. The only realistic evaluation is whether the result of the clustering is useful to the domain experts, whether and to what extent it provides any insight into the data. Having said that, the algorithms can be compared with respect to robustness and accuracy for the dataset in question, robustness being a measure of the stability of the result and accuracy of the "quality" of the clusters.

Robustness can be assessed by setting aside a proportion of the dataset and then clustering on both parts and comparing. This could be repeated a number of times and there are a number of variations on this technique. Accuracy can be gauged for statistical algorithms such as EM by estimating the probability of the data given the clusters, the log-likelihood measure previously described. Techniques that do not normally work with probabilities can be converted into probability density based clusterers within Weka so that a log-likelihood value can be obtained.

Case Studies

Dataset #1: Establishment Survey on Working Time and Work-Life Balance (ESWT)

The establishment survey on working time and work-life balance was conducted to map working time policies and practices at the level of the establishment in the European Union, to survey the views of the management and employees at the establishment level on these policies and practices and to provide policy makers with a picture of the main issues and developments in the field. The survey focused on the following working time arrangements, which are likely to have an impact on work-life balance: part-time work; extended operating hours (night work, weekend work, shift-work); flexible working time arrangements (e.g.

flexi-time); overtime; child-care leave and other forms of long-term leave; and phased retirement and early retirement. (Bielenski et al, 2005)

The question that we wanted to explore was, given that some firms provide good 'work-life balance' working time arrangements and others have a hostile or harsh working-time regime, which variables predict the type of firm most likely to offer a good (or hostile) regime and why? The approach we took was to use a clustering algorithm to divide the data into two clusters in the expectation that those with good working time arrangements and those with bad working time arrangements would tend to cluster together. The dataset contains a number of filter questions which result in a significant number of system missing values. Variables that contain such system missing values were not used in the clustering to avoid skewing the result. This meant that we were clustering on 33 attributes. We had 21,031 instances (establishments) in the dataset. A number of different algorithms were run and compared for robustness and accuracy.

Given that we were looking for two clusters, it made sense to use a partitional rather than hierarchical algorithm. It also meant DBScan was unsuitable since the number of clusters it produces is not known in advance. The three potential algorithms therefore were k-means, EM and Farthest First. The Farthest First algorithm, which is fast, approximate and more suited to finding the initial conditions for k-means, tended to put most of the instances into just one of the clusters. Splitting the dataset showed that the results were also far from robust. The slower k-means algorithm still only took a couple of seconds to run on a 2GHz laptop. The results varied markedly with the initial conditions, as it found different local minima, although they did tend to divide the data into halves. Splitting the data, however, showed the answers to be fairly robust. The best algorithm for this dataset was found to be Expectation-Maximisation. This took about a minute to run on the same machine. The answer produced was actually independent of the initial conditions and so we can assume that we've found the global maximum. On splitting the data, it also looked quite robust. Using Weka to convert the k-means and Farthest First routines into density based clusterers, we were able to compare the three algorithms for accuracy using the log-likelihood values. Predictably, it showed that EM produced the best quality clusters, followed by k-means and then Farthest First.

In terms of answering our questions about the data, we did indeed find that the algorithm naturally divides the establishments into those whose working-time arrangements lead to 'good' and 'bad' work-life balances, as shown in Table I.

The 'good establishments' Cluster	The 'bad establishments' Cluster
Establishments in Denmark, Netherlands, Finland and Sweden are mainly in this cluster	Establishments in Greece, Spain, Italy and Portugal are mainly in this cluster
Public sector establishments are mostly in this cluster	Industry and private services tend to be in this cluster
Establishments with one or more domestic support services are nearly all in this cluster	Establishments without any domestic support services are nearly all in this cluster
Establishments with >=40% employee representation (or unknown) are nearly all in this cluster	Establishments with <40% employee representation are nearly all in this cluster

The 'good establishments' Cluster	The 'bad establishments' Cluster
Establishments in the following sectors are mostly in this cluster: public administration; education; health and social work; other community, social and personal services.	Establishments in the construction and retail/repair sectors are mainly in this cluster. This is dependent on country though, e.g. retail/repair is mainly in the 'good' cluster in Finland, Sweden and the U.K.
Establishments with different workload variations are slightly more likely to be in this cluster but country is a more important factor	Establishments with no or don't know for the different workload variations are slightly more likely to be in this cluster but country is a more important factor
Establishments with >20% part-time proportion of employees (as well as don't know and no answer) are mostly in this cluster	Establishments with a 0-20% part-time proportion of employees are mostly in this cluster
Establishments where changing from full-time to part-time in both skilled and unskilled jobs is possible are mainly in this cluster, although those in Greece, Spain, Italy and Portugal are exceptions to this	Establishments where changing from full-time to part-time in both skilled and unskilled jobs is not possible are mainly in this cluster, although those in Finland and Sweden are exceptions to this
Establishments that work at night, Saturday or Sunday are mainly in this cluster, although those in Greece, Spain, Italy, Portugal and Cyprus are exceptions	Establishments that do not work at night, Saturday or Sunday are mainly in this cluster, although those in Netherlands, Finland and Sweden are exceptions
Establishments where the employees have changing working hours are mostly in this cluster, although those in Greece, Spain, Italy and Portugal are exceptions	Establishments where the employees do not have changing working hours are mostly in this cluster, although those in Denmark, Netherlands, Finland, Sweden and U.K. are exceptions
Establishments with employees in parental leave tend to be in this cluster, although those in Greece, Spain, Italy and Portugal are exceptions	Establishments with no employees in parental leave tend to be in this cluster, although those in Netherlands, Finland and Sweden are exceptions
Establishments with long term leave tend to be in this cluster, although those in Greece, Spain, Italy, Portugal and Hungary are exceptions	Establishments with no long term leave tend to be in this cluster (also the DK/NA's), although those in Finland, Sweden and U.K. are exceptions
	Establishments which did not agree to further contact are mainly in this cluster, along with those that did Greece, Spain, Italy, Austria, Portugal and Cyprus

Table I. The clusters of establishments providing a 'good' and 'bad' work-life balance

There were also a few variables that did not show significant difference between the clusters. These related to whether the establishment had flexible working time arrangements; the proportion of staff working overtime; whether it had training programmes for returning people; the possibility of early retirement; and whether the management considered work-life balance to be a task of the company. The results of this clustering have therefore gone some way towards addressing the question of which variables predict the type of firm most likely to offer a good (or hostile) regime. It is now necessary to get further input from domain experts

in order to assess the significance of, and perhaps explain, these results and to discuss how data mining could go on to further explore the patterns in the data.

Dataset #2: English Longitudinal Study of Ageing (ELSA) Wave 2

The English Longitudinal Study of Ageing (ELSA) is a study of people aged 50 and over and their younger partners, living in private households in England. The sample was drawn from households that had previously responded to the Health Survey for England (HSE) in 1998, 1999 or 2001. Every two years they hope to interview the same group of people to measure change in their health, economic and social circumstances. ELSA can complete the picture of what it means to grow older in the new century, and help us understand what accounts for the variety of patterns that are seen. The dataset we looked at relates to data gathered by the interviews from the second Wave of ELSA, which were carried out between June 2004 and July 2005.

The ELSA Wave 2 interview covered a wide range of topics. It was similar to the questionnaire used in Wave 1, although every module was reviewed to ensure that it would provide data that measured change over time. This was achieved by repeating some measures exactly (for example, to measure income and assets), by asking directly about change (for example, to capture perceived changes in memory and concentration) and by adapting questions to allow people to update or amend past responses (for example, about work, pensions and specific health conditions). The Wave 2 interview was also expanded to answer a variety of additional research questions. The new items included: quality of health care received; household spending on leisure, clothing and transfers; perceptions of deprivation relative to others; perceptions of ageing; levels of literacy; perceived effort and reward for care-giving; and voluntary activities. Core sample members who completed a main interview were also offered a nurse visit. This included tests of blood pressure, lung function, blood tests, anthropometric measures and physical performance measures.

A small part of the ELSA dataset consists of documented verbal responses to open questions about the positive and negative aspects of growing older. All of the different responses have been categorised by hand, whereby they are assigned a numerical reference code, which is a slow and painstaking manual process. Each person's response may contain up to seven positive and seven negative codes. We investigated the potential of using a clustering algorithm to automatically categorise the documented responses, hoping that this could do so as effectively as doing it by hand, and in considerably less time. Unfortunately, it was found that the data here was full of spelling and grammatical errors to the extent that cleaning it became impractical and so the data was unsuitable for data and text mining techniques. This has implications for a lot of free text social science data.

However, no such restrictions apply to mining the ELSA dataset itself. The dataset consisted of 8,688 instances and 1,410 attributes. Applying an Expectation-Maximisation algorithm to form two clusters from the data items that have no missing values seems to divide the interviewees largely according to age, given that there are a large number of variables that are, to a large extent, age dependent. For example, interviewees in the 'younger' cluster include most of those with a child in the household, most of those who had taken paid employment in the last week, most of those with good results on the health variables, most of those who had undertaken formal training or education in the last 12 months, and so on, with the opposite characteristics being exhibited by the members of the 'older' cluster. We are currently waiting for guidance from domain experts before we decide how to proceed with mining this dataset.

Future Work

The two case studies in this paper are both work in progress and we intend to further mine these datasets with guidance from the social scientists involved. We are also looking at other datasets, including the European Social Survey Data, which has a lot of apparently unrelated variables and so data mining might be useful as a way of identifying areas for further research.

Our longer-term goal is to produce a suite of data mining codes, supported by a web interface, to allow social scientists to mine their datasets in a straightforward way and hence, gain new insights into their data. The knowledge gained from our data mining case studies will inform the development of these data mining tools. It will show us the nature of the datasets that are being studied and the type of algorithms that will be required. We intend to run these codes on national/regional grid services in the U.K.

References

Bielenski, H. and Riedmann, A. (2005): Establishment Survey on Working Time and Work-Life Balance in 21 EU Member States (ESWT 2004/2005) Technical Report: Methodology, Questionnaire Development and Fieldwork.

Dunham, M. (2003): *Data Mining Introductory and Advanced Topics*, Prentice Hall, Upper Saddle River, NJ.

Hand, D., Mannila, H. and Smyth, P. (2001): *Principles of Data Mining*, MIT Press, Cambridge, MA.

Hochbaum, D., and Shmoys, D. (1985): 'A best possible heuristic for the k-center problem', *Mathematics of Operations Research*, vol. 10, no. 2, 1985, pp.180-184.

Tan, P., Steinbach, M. and Kumar, V. (2006): *Introduction to Data Mining*, Addison Wesley.

Witten, I. and Frank, E. (2005): *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*, Morgan Kaufmann, San Francisco.