2000 GENERAL INSURANCE CONVENTION
25-28 OCTOBER

BAYESIAN NETWORKS AND DATA MINING

Presenters: James Orr, Dr Peter England, Dr Robert Cowell and Duncan Smith

Bayesian Networks and Data Mining

James Orr, Dr Peter England, Dr Robert Cowell, Duncan Smith

Data mining means finding structure in large-scale databases. Although not a new activity, it is becoming more popular as the scale of databases increases. Holders of data are keen to maximise the value of information held, and companies offering software to help analyse such data have been able to raise huge amounts of capital more on the prospect of future earnings than past performance.

Data mining is a subject of interest in statistics, engineering and computer science, and a variety of techniques exist for the purpose, including standard statistical methods such as generalised linear modelling and multivariate analysis (such as principal components analysis and cluster methods), tree-based methods, neural networks, near-neighbour methods and Bayesian belief networks.

Bayesian belief networks (BBNs) will be the main focus in this workshop. Typically in a BBN, relationships between variables in a dataset can be viewed pictorially through a so-called *directed acyclic graph* (DAG), which shows local dependencies between variables. Expert opinion might be used to build the model initially, or the model might be found according to some algorithm which "learns" the structure and estimates its associated probability tables. Given new data, Bayes' rule can be used to make inferences using the model, and if desired to improve the model's probability tables.

Many applications of BBNs have been in building expert systems to help with, for example, medical diagnosis or the trouble-shooter of a computer package, where the probability of a "diagnosis" can be estimated conditional on information provided. In this context, the model structure is typically built from expert opinion providing relationships between variables and paths of causality. After building the model, probabilities of outcomes can be assigned again through expert opinion, or from data, if available.

In a data mining context, the model structure is "learnt" by implementation of an algorithm which searches for the most likely relationships between variables. This could be achieved "blind" by just letting the computer get on with it, or with a limited degree of help from the user who might suggest an initial structure, or some constraints that have to be obeyed.

A data set on motor insurance renewals has been selected to evaluate the usefulness of Bayesian networks in an insurance context, and initial results will be reported at the workshop. The primary variable of interest was whether a

policyholder renewed their policy or not (a binary 0/1 outcome), with other variables providing information on policyholder characteristics and price relativities. The data set was provided "blind", with variable names hidden to prevent knowledge of the modeller influencing the effectiveness of the methodology. The results will be compared to a "standard" statistical analysis based on logistic regression (generalised linear model with binomial error structure and logit link function). Similarities between logistic regression and certain types of Bayesian network will be highlighted and the advantages and disadvantages of the two approaches will be compared.

Freely available software downloadable from the internet will be demonstrated using a sample of the data mentioned above to help explain the concepts.

Other areas where the techniques might prove useful will also be suggested, such as identification of possible fraudulent claims, which are a huge burden on the insurance industry, the costs ultimately being passed on to policyholders.