Analyzing company growth data using genetic algorithms on binary trees

Gerrit K. Janssens\textsuperscript{1}, Kenneth Sørensen\textsuperscript{2}, and Koen Vanhoof\textsuperscript{1}

\textsuperscript{1}Limburg University Centre, Faculty of Applied Economics, Data Analysis and Modelling Research Group
\textsuperscript{2}University of Antwerp, Faculty of Applied Economics

December 16, 2003

Abstract
In this extended abstract, we discuss an application of the genAID data mining tool to company growth data. genAID uses a genetic algorithm to evolve a set of high-quality AID decision trees.

1 Data mining with AID

The technique used in this research is called the Automatic Interaction Detection (AID). For a detailed overview of data mining techniques from a database perspective, we refer to Chen et al. [1]. The original idea for this technique was formulated by Morgan and Sonquist [5]. They were interested in developing a statistical analysis technique that would more accurately describe “real world” social and economic events than was possible using standard statistical regression techniques.

The AID technique is a mechanical, or automatic, system that mimics the steps taken by an experienced data analyst to determine strong data interaction effects. Its basic principle is to explain the variance of a dependent variable through an exhaustive search of all possible relations between predictors and the dependent variable. The results of the search are represented as a binary tree. The nodes represent predictor variables for which a binary split explains most. In a first step, every possible predictor variable is tested to see which one has the strongest predictive power. The population is then split into two classes according to this predictor variable. This process is repeated for the descendent classes, until some stopping criterion is met. The strength of the predictor is measured by the value of

\[ P = \frac{n_1(\bar{x}_1 - \bar{x}) + n_2(\bar{x}_2 - \bar{x})}{n s^2} \]  

where \( \bar{x}_1 \) and \( \bar{x}_2 \) are the averages in subgroup 1 and 2, \( n_1 \) and \( n_2 \) are the number of subjects in subgroup 1 and 2, \( \bar{x} \) is the population average, \( n \) is the total number of subjects in the population, \( s^2 \) is the population variance. Both of the subgroups formed are then candidates for a new split.
The result is a series of splits, or branches in the data. Where each split produces a new data set that is in turn split. The result is a classification tree. The process of splitting ends when (1) a class contains a number of members that is too small to split, (2) the maximum number of splits set by the user of the algorithm have been reached or (3) a class is so homogenous that no split is statistically significant.

One of the main criticisms of the original AID technique was that it tends to be overly aggressive at finding relationships and dependencies in data and that it could not discriminate meaningful from meaningless rela-tionships. AID would often find relationships that were the result of chance. Later versions of the AID tech-nique improved on some of its deficiencies Einhorn [2]. The AID technique was combined with statisti-cal hypothesis testing methods. Branches of the AID technique can be tested at a certain significant level and insignif-icant splits can be disregarded when showing the classification tree. CHAID and THAID are some of the techniques that were developed and employed a statistical framework to discover more meaningful results [4].

Regardless of the fact that AID is a strong technique, it has its disadvantages Einhorn [2], Kass [3].

1. The AID technique requires large sample sizes, since it possibly divides the population of subjects into many categories. If some statistical significance is required, the number of subjects in the sample must be large.

2. The technique does not take correlated variables into consideration. This means that spurious relation-ships may be discovered by the technique.

3. An asymmetric variable disturbs the performance of the AID technique. When coping with an asymmet-ric dependent variable, the technique tends to repeatedly split off or separate smaller groups. Asymmetric predictors on the other hand decrease their predictive power and their probability to appear in a tree.

4. The explanatory tree structures are not stable if multiple samples are taken from the same population. These samples from the same population may lead to different trees.

5. Stopping rules tend to be not very clear.

2 \textit{genAID}

To solve some of the shortcomings of AID, \textit{genAID} was developed by Sörens en and Janssens [6]. This technique uses a GA to develop a population of diverse AID decision trees. The fitness of these trees is determined by a formula that borrows from one-way ANOVA:

\[
f_r(A) = \frac{\sum_{i=1}^{K} n_i (\bar{x}_i - \bar{x})^2}{\sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2},
\]

\[\text{(2)}\]
where \( n_i \) is the number of observations in class \( i \), \( K \) is the number of classes (equal to \( w(A) \), the number of leaves of tree \( A \)), \( x_{ij} \) is the \( j \)-th observation in class \( i \), \( \bar{x}_i \) is the class \( i \) sample mean and \( \bar{x} \) is the overall sample mean. Calculating the fitness of a classification tree is a computationally difficult process because each subject in the population has to be classified into one of the \( w(A) \) classes. After that, the sum of squares in each class has to be calculated.

Taking the computational burden of the fitness calculation into account, genetic operators have been devised in such a way that the number of fitness function evaluations is reduced to a minimum. Macro-operators exchange sub-trees within and between trees. Micro-operators only change the node labels.

3 Analysing company growth data

3.1 Independent variables

The data analysed with the genAID procedure are based on a survey into the growth rate of 840 companies. To determine which factors influence company growth, these companies were asked such questions as “To which extent did you have trouble finding suitable personnel”, “To which extent did you experience trouble with government regulations”, etc. The answer to most these questions required an answer on a scale from 1 to 5.

A number of financial determinants were also entered into the analysis. E.g. the solvency and earning potential of the company.

Finally, a number of marketing variables were also used in the analysis, e.g. marketing image.

All variables are converted to binary variables. This is done using the following method. For a variable \( x \) that can take \( n \) values, \( n - 1 \) binary variables are created. Binary variable \( x_i \) is set to 1 if \( x > i + 1 \). E.g. if \( x \) can take 4 values, 3 variables are created.

<table>
<thead>
<tr>
<th>( x )</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The main motivation for this encoding is the fact that every binary variable splits the population in a high group and a low group, whereas other binary encodings do not.

3.2 Dependent variable

The dependent variable is a continuous variable, indicating the growth rate of a company during a certain period. Several periods are taken into account.

3.3 Some results

Figure 1 shows the result of 1000 generations of the algorithm. As can be seen from this figure, the algorithm quickly improves the quality of the solutions in
the population. After some generations, a diversification phase starts and the algorithm starts to generate a diverse set of high-quality solutions.

The best 5 trees found with this experiment are schematically shown in figure 2. Each tree is preceded by its fitness value.

Figure 1: Result for 1000 generations

Figure 2: Best 5 trees found

4 Conclusions

References


