DATA MINING AND STATISTICAL METHODS USED FOR SCANNING CATEGORICAL DATA

Murtadha M Hamad
Al-Anbar University-College of computers

Received: 1/5/2007 Accepted: 30/8/2007

ABSTRACT: It has been shown that data mining uncovers patterns in data using predictive techniques. These patterns play a critical role in decision making because they reveal areas for process improvement. Statistical techniques such as Chi-square test for association are widely used in the medical field. Yet, the interpretation of some of the results approached by the use of this statistical techniques is seems to be a very difficult task. The type of association is often non-linear and hence will mask the important part of the use of this technique.

In this research work a new approach is adopted by scanning the raw data for any possible association (linear or non-linear). More data mining methods and statistical inference were the base tools of this research work.

Keywords: Chi-square test, Data mining, Kappa measures, Linear and non-linear association, Categorical data.

1. INTRODUCTION
In the cross-classification of categorical data, researchers are always interested in the sense of searching the cross-classification table for any potential relationship between groups of the cross classified variables. Such a relationship is statistically denoted as an association.

Data mining often concerns with the meaning and quality of the information embedded in any given set of data. Ideas from information measurements and statistical analysis will be merged in order to establish a linkage between these two tools [1]. Such a linkage will allow the users to handle a straightforward interpretation as to unmask the type and degree of association, and hence will enhance the meaning of the results obtained.

Most analysts separate data mining software into two groups: data mining tools and data mining applications. Data mining tools provide a number of techniques that can be applied to any business problem. Regardless of whether we are aware of them, our daily lives are influenced by data mining
applications. For example, almost every financial transaction is processed by a data mining application to detect fraud. Both data mining tools and data mining applications are valuables, however. Increasingly organizations as data mining tools and data mining applications together in an integrated environment for predictive analytic.

Assume a concept of event patterns as an embodiment of information. Consider a set of mutually exclusive random variables \{X_i : i = 1 \ldots k\}. An instantiation of any sub set of variables in \(X\) referred to as an event pattern. In this research, light will be shed only on discrete, finite, multi-valued random variables. Using multi-valued discrete variables to represent a physical phenomenon, a concept, or an object, is common in a variety of fields such as medicine, business and economics. For example, in a medical diagnosis problem, gender and condition may be two variables of interest. A particular patient always has one and only one gender, meanwhile could be a located to none, one or more disease(s) [2].

The concept of data mining passed upon data patterns is to identify events patterns that are either statistically significant or not. One approach towards identifying statistical significant information is passed on event association [3]. Significant association may be determined by statistical hypothesis test passed on mutual information measure or residual analysis. As reported elsewhere, mutual information with regard to information theory is asymptotically distributed as chi-square distribution. This result has been extended elsewhere [4] to model residual analysis as a normally distributed random variable. In doing so, statistical hypothesis test passed on residual analysis may be used as a conceptual tool to discover data patterns with significant event associations. Another results discovered recently [5] is an algebraic linkage between information measure and statistical analysis that suggests yet another approach for detecting events association passed on symbol probability ratio.

2. Data Handling and Algorithm

As to clarify the algorithm with real set of data that were part of the data set collected by Holmquest et al [6] in an investigation into observer reliability in the histological classification of carcinoma in situ and related lesions of the uterine cervix [7], will be used (table 1).

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>negative</td>
</tr>
<tr>
<td>2</td>
<td>A typical squamous</td>
</tr>
<tr>
<td>3</td>
<td>Carcinoma in situ</td>
</tr>
<tr>
<td>4</td>
<td>Squamous carcinoma with early stromal invasion</td>
</tr>
<tr>
<td>5</td>
<td>Invasive carcinoma</td>
</tr>
</tbody>
</table>

The algorithm
Before they are plunged into the process of calculating the chi-square statistics and its relevant kappa [8] it is useful to state the gradual steps of the computer algorithm from the beginning till the handling of the decision statement. The following steps illustrate the detailed algorithm:

1. A database containing information about patients relevant to this research must be available. In this context a virtual medical database has been prepared to handle the real data considered in this research.

2. A table containing the required information must be identified. In this context a virtual table has been designed to involve the following fields:
   - Patients number
   - Age
   - Sex
   - Date of admission
   - Complain
   - Symptoms
   - Mass
   - Pathologist 1
   - Pathologist 2

   The table contained information about the real patients as posed by the example mentioned in everitt B. S. [6].

3. Cases with negative mass findings has been ignored in the study. Only those cases with positive mass findings were involved. This procedure has been done by the use of filter statement available within the Microsoft database program.

4. A different procedure has been implemented to handle the cross-classification table of beliefs from pathologist 1 and 2 (Table 1).

5. A crystal report containing the cross classified table and the results of both chi-square statistics and kappa has been designed. The report also contained a decision statement the magnitude of kappa based on the comparison of the calculated value of kappa with its theoretical range of values.

   The detailed procedure for calculating chi-square test and kappa is given according to the following:

   a. The value of Chi-square statistics calculation:

   \[
   \chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}
   \]

   where m is the number of rows and k is the number of columns for the cross classified table: O is the observed frequency and E is the expected frequency calculated by multiplying the corresponding row and column totals and divide the result by the grand total.

   b. Calculation kappa :

   According to the data of table 1, the following components are going to be calculated:

   \[
   p_b = \left( \sum_{n} C_u \right) / T = \frac{22+7+36+7+3}{118} = 0.63
   \]
\[ P_0 = \left( \sum_{i=1}^{n} T_i T_i / T \right) \times 100\% = \left( \frac{26 \times 27 + 26 \times 12 + 38 \times 69 + 22 \times 7 + 6 \times 3}{118} \right) \times 100\% = 0.32 \]

Therefore

\[ Kappa = \frac{P_0 - P_e}{1 - P_0} = \frac{0.635 - 0.32}{1 - 0.32} = 0.47 \]

In order to calculate the variance of kappa, the proportion of each entry of table 1 will be divided by 118 (table 2)

\[
Var(Kappa) = \frac{1}{T(1-P_0)^4} \left[ \sum_{i=1}^{n} P_i \left[ (1-P_i) - (P_i - P_j)(1-P_0)^2 \right] \right] = 0.0036
\]

\[ Sd = \sqrt{Var(Kappa)} = 0.06 \]

95% confidence intervals = kappa ± 1.98sd / \[ \sqrt{T} = (0.33, 0.61) \]

3. Data model

In this research, a data table has been done which contains the various number of the different diagnosis cases and the number of the cases dealt with (118) as the shown in table (1). As figure 1 show:

1. Some statistical tools were used as helping devices, such as chi-square statistics and kappa metric in table (1) for obtaining table (2). Then, the standard deviation was calculated.

2. A database under the name (pathologists.mdb) was built and designed, which contains (118) records corresponding with the cases. The data base is expandable to include further number of records. Filtering of the data base records has been carried out in order to deal with the cases of positive mass under study.

The results of 1 and 2, and the use of the table below [8]:

\[ T = \frac{1}{1 - 0.32} \]
Were conducive to a report containing the essential information which has a role in making the decision that leads to diagnosis of the infection level of the studied cases.

4. Discussion Results

In this paper, 118 cases of factual data have been dealt with, as shown in table (1) and data base (pathologists.mdb).

After examining the statistical concepts, it has been noticed that using the weighted Kappa Metric is important in satisfactorily classifying and partitioning the data groups of the above data base according to the studied cases. It has also been noticed that Kappa Metric has an active role and important indicator relative to the data observed by (2 pathologists). Despite the complexities that accompany the calculation, the role of Kappa Metric is greater when the observes are more than two.

From algorithm (step 5) we got the following:

\[ Sd = \sqrt{\text{Var}(\text{Kappa})} = 0.06 \]

95% Confidence intervals =

\[ \text{Kappa} \pm 1.96 Sd / \sqrt{T} = (0.33, 0.61) \]

From table 3 we got the optimal value (0.47) of strength agreement:

Moderate

Kappa = 0.47 and hence a moderate linear association is detected

This messages one of six possible that the program may revealed according to the value of Kappa as stated in table 3. The message motional below the crystal report is the actual output of the program according to the value of Kappa (0.47).

The 95% confidence intervals indicated that the value of Kappa will never be out of range (0.33, 0.61).

5. Conclusion

The study has reached the following conclusions:

1- Using the statistical concepts and tools as supporting tools in dealing with data mining has a role in uncovering data which are not easily revealed by normal methods.

2- Several interesting results are found in this research. First, the concept of data patterns allows us to visualize any early step of data mining has being a process of finding significant event associations. This process lends itself to a set of data patterns for discovering an inference model; where such a model encapsulates significant behavior of the data as measured by statistical analysis as well as information measure.

3- Data mining uncovers patterns in data using predictive techniques. These patterns play a critical role in decision making because they reveal areas for processes improvement. Using data mining, organizations can increase the profitability of their interactions with customers, detect fraud, and improve risk management. The
patterns uncovered using data mining help organizations make better and timelier decisions.

6. References


Figure 1: Suggested Data Model for current work
Table 1: observed frequencies of biopsy slides classified by two pathologist according to most involved lesion of the uterine cervix.

<table>
<thead>
<tr>
<th>Pathologist</th>
<th>11</th>
<th>22</th>
<th>33</th>
<th>44</th>
<th>55</th>
<th><strong>Total T</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>222</td>
<td>72</td>
<td>22</td>
<td>00</td>
<td>00</td>
<td>26</td>
</tr>
<tr>
<td>2</td>
<td>55</td>
<td>27</td>
<td>114</td>
<td>00</td>
<td>00</td>
<td>26</td>
</tr>
<tr>
<td>3</td>
<td>00</td>
<td>12</td>
<td>336</td>
<td>00</td>
<td>00</td>
<td>38</td>
</tr>
<tr>
<td>4</td>
<td>00</td>
<td>01</td>
<td>114</td>
<td>77</td>
<td>00</td>
<td>22</td>
</tr>
<tr>
<td>5</td>
<td>00</td>
<td>00</td>
<td>33</td>
<td>00</td>
<td>33</td>
<td>6</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>227</td>
<td>112</td>
<td>669</td>
<td>77</td>
<td>33</td>
<td>118</td>
</tr>
</tbody>
</table>

Table 2: proportion of each entry as compared to the grand total.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th><strong>T</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.186</td>
<td>0.017</td>
<td>0.017</td>
<td>0.000</td>
<td>0.000</td>
<td>0.220</td>
</tr>
<tr>
<td>2</td>
<td>0.042</td>
<td>0.059</td>
<td>0.119</td>
<td>0.000</td>
<td>0.000</td>
<td>0.220</td>
</tr>
<tr>
<td>3</td>
<td>0.000</td>
<td>0.017</td>
<td>0.305</td>
<td>0.000</td>
<td>0.000</td>
<td>0.322</td>
</tr>
<tr>
<td>4</td>
<td>0.000</td>
<td>0.008</td>
<td>0.119</td>
<td>0.059</td>
<td>0.000</td>
<td>0.186</td>
</tr>
<tr>
<td>5</td>
<td>0.000</td>
<td>0.000</td>
<td>0.025</td>
<td>0.000</td>
<td>0.025</td>
<td>0.051</td>
</tr>
<tr>
<td><strong>T</strong></td>
<td>0.229</td>
<td>0.102</td>
<td>0.0585</td>
<td>0.059</td>
<td>0.025</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Evaluation of observed Kappa values

<table>
<thead>
<tr>
<th>Weighted_kappa</th>
<th>Strength of agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>Poor</td>
</tr>
<tr>
<td>0.00-0.20</td>
<td>Slight</td>
</tr>
<tr>
<td>0.21-0.40</td>
<td>Fair</td>
</tr>
<tr>
<td>0.41-0.60</td>
<td>Moderate</td>
</tr>
<tr>
<td>0.61-0.80</td>
<td>Substantial</td>
</tr>
<tr>
<td>0.81-1.00</td>
<td>Almost perfect</td>
</tr>
</tbody>
</table>
طريق التحري والإحصاء المستخدمة في مسح البيانات الفنية

مرتضى محمد حمد
E. mail : mortadha61@yahoo.com