DATA MINING AND STATISTICAL METHODS USED FOR SCANNING CATEGORICAL DATA

Murtadha M Hamad

Al-Anbar University-College of computers

ARTICLE INFO

Received: 1 / 5 /2007
Accepted: 30 / 8 /2007
Available online: 14/6/2012
DOI: 10.37652/juaps.2007.15361

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

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.

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 relation ship 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

* Corresponding author at: Al-Anbar University-College of computers, Iraq.E-mail address: mortadha61@yahoo.com

data mining applications together in a integrated environment for predictive analytic.

Assume a concept of event patterns as an embodiment of information. Consider a set of mutually exclusive random variables {Xi : i = 1 …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).

1: negative

2: A typical squamous

3: Carcinoma in situ

4: Squamous carcinoma with early stromal invasion.

5: Invasive carcinoma

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 in formation 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-sqaure statistics calculation:

$$x^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(O_y - E_y)^2}{E_y}$$

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:

$$p_0 = \left(\sum_{i=1}^{n} C_u\right) / T = \frac{22+7+36+7+3}{118} = 0.635$$

$$p_0 = \left(\sum_{i=1}^{n} T_i T_n / T\right) \times 100\% =$$

$$\left(\frac{26*27}{118} + \frac{26*12}{118} + \frac{38*69}{118} + \frac{22*7}{118} + \frac{6*3}{118}\right)$$

$$\times 100\% = 0.32$$

Therefore

$$Kappa = \frac{P_O - P_0}{1 - P_0} = \frac{0.635 - 0.32}{1 - 0.32} = 0.47$$

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

$$Var(Kappa) = \frac{1}{T(1-P_0)^4}$$

$$\left[\sum_{i=1}^{n} \begin{matrix} P_u\left[(1-P_c) - (P_i - P_j)(1-P_0)\right]^2 \\ + (1-P_0)^2 \sum_{i=1}^{n} \sum_{j=1}^{n} P_n (P_i + P_j)^2 - (P_0 P_c - 2P_c + P_0)^2 \end{matrix}\right]$$

$$= 0.0036$$

$$Sd = \sqrt{Var(Kappa)} = 0.06$$

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
2. standard deviation was calculated.
3. 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]:

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:

$$Sd = \sqrt{Var(Kappa)} = 0.06$$

$$95\% Confidence\, inmtervals =$$
$$Kappa \pm 1.98 Sd / \sqrt{T} = (0.33, 0.61)$$

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

[1] Sy B. K., "Pattern-based Inference Approach for Data Mining", Queen College, Department of Computer Science, 1999.

[2] Sy B. K. & D. Sher, "An abstraction Theory framework for Probabilistic Inference", Proc. Of the workshop on spatial and temporal Interaction, Nov. 1994.

[3] Goldstein D, Ghosh D, Conlon E, "Statistical issues in the clustering of gene expression data", 2002, 12:219-241.

[4] Wrong A.K.C., "High order pattern discovery from Discrete –valued Data", IEEE Trans. On Knowledge and Data Engineering, 9(6):877-839, 1997.

[5] Sy B. K., "Information-theoretical and statistical approaches for independence Test ", Proc. Of the international S-PLUS Users Conference, published by Mathsoft Inc., Washington D.C., and Oct 1998.

[6] N.D. Holmquist & O.D. Williams "Variabililty in classification of carcinoma in Situ of the Uterine Cervix", Archives of pathology, 1967.

[7] B.S. Everitt, "statistical methods for medical Investigations", first published in Great Britain 1989.

[8] S. Swift, "Consensus clustering and functional interpretation of gene-expression data ", Genome Biology, Nov. 2004.

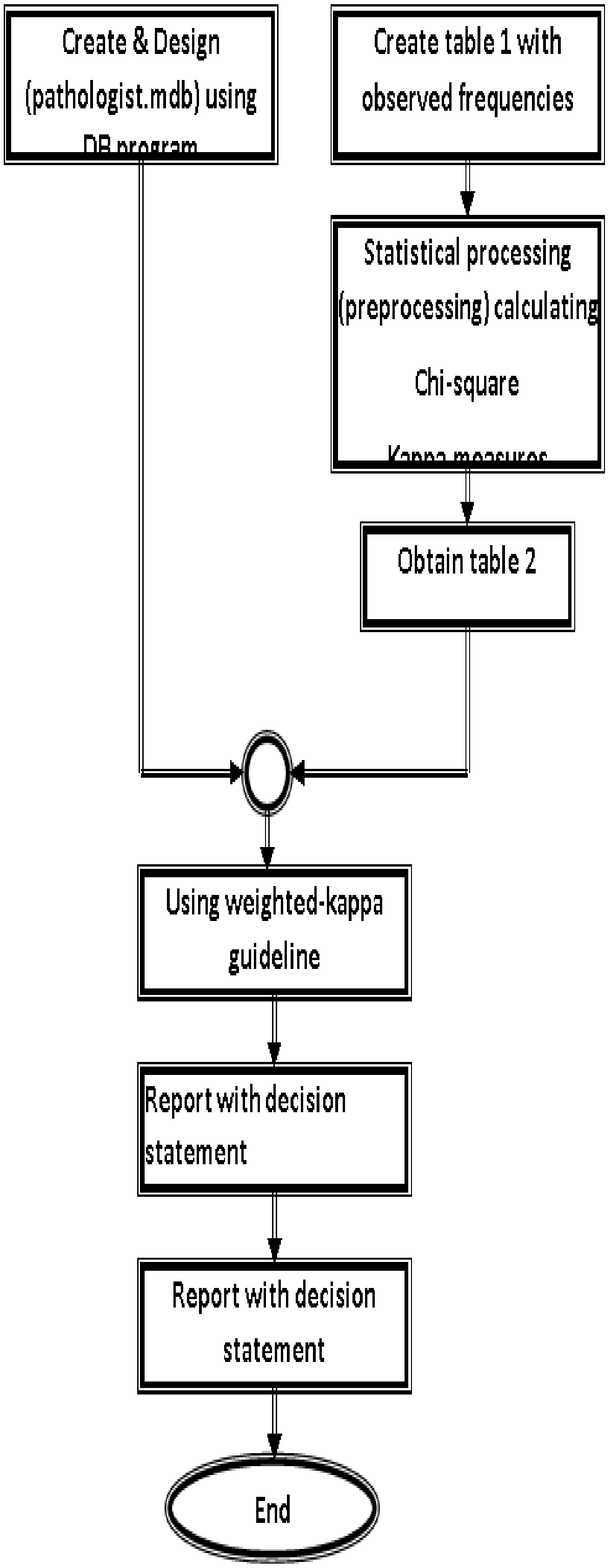

Figure 1: Suggested Data Model for current work

Table 1: observed frequencies of biopsy slides classified by two pathologist according to most involved lesion of the uterine cervix.

Total T_i	5	4	3	2	1	Pathologist1	
227	00	00	00	55	222	11	Pathologist 2
112	00	01	12	27	72	22	
669	33	114	336	114	22	33	
77	00	77	00	00	00	44	
33	33	00	00	00	00	55	
118 T	6	22	38	26	26	Total T_i	

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

T_j	5	4	3	2	1	
0.229	0.000	0.000	0.000	0.042	0.186	1
0.102	0.000	0.008	0.017	0.059	0.017	2
0.0585	0.025	0.119	0.305	0.119	0.017	3
0.059	0.000	0.059	0.000	0.000	0.000	4
0.025	0.025	0.000	0.000	0.000	0.000	5
	0.051	0.186	0.322	0.220	0.220	T_i

Table 3: Evaluation of observed Kappa values

Weighted_kappa	Strength of agreement
0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost perfect

طرق التحري والإحصاء المستخدمة في مسح البيانات الفئوية

مرتضى محمد حمد

E. mail : mortadha61@yahoo.com

الخلاصة

لقد تبين ان اسلوب التحري عن البيانات يكشف لنا الكثير من الانماط الغير المعروفة والتي لها دورها الهام في عملية صنع القرار . ان هذه الانماط يمكن ان تلعب دورا مهما في ازالة الستار عن المساحات التي يمكن من خلالها تحسين عملية صنع القرار . التقنية الاحصائية كأختبار chi-square لتحديد نوع الاقتران كثيرة الاستعمال في الحقل الطبي. رغم ذلك، تفسير البعض من النتائج التي جرى حسابها باستعمال هذه التقنية الاحصائية تبدو مهمة صعبة جدا. ان نوع الاقتران في اغلب الاحيان يكون لاخطيا ولذلك سيخفي الجزء المهم لاستعمال هذه التقنية. في هذا البحث تم التعامل باسلوب جديد من خلال مسح البيانات الاولية لاي اقتران محتمل (خطي او لاخطي). طريقة التحري عن البيانات والاستدلال الاحصائي تعتبر من الادوات الاساسية لهذا البحث.