Mining Medical Data using SQL Queries and Contingency Tables

Jan Zytkow†[1] and Shishir Gupta[1,2]

[1]Department of Computer Science, UNC-Charlotte, Charlotte
North Carolina, USA
{zytkow, shgupta}@uncc.edu
http://www.cs.uncc.edu
[2]Department of Information Technology, Catawba College, Salisbury
North Carloina, USA
shishir@shishir.net
http://www.catawba.edu

Abstract. This work presents the data mining effort to retrieve meaningful information from a Medical Database collected at the Chiba University Hospital, Japan for patients suffering from Collagen diseases. The data, after being examined and cleaned [1,2], was transformed and loaded into a Relational Database. SQL queries have been used to retrieve data in the form of contingency table, which is further used to extract patterns in the data. The results obtained using this technique have been discussed and compared to results obtained from previous work.

1. Introduction

In this paper we show a way to identify patterns in a dataset that resides in a Relational Database using contingency tables . A contingency table [5], also called a cross reference table, is a table showing the number of records for each value combination of two or more variables that constitute the table. It is a useful tool for pattern visualization that allows user to decide on the most useful description. Contingency tables have been used here to find relationships and patterns in the data. In case where the size of the contingency table is large, discretization techniques can be used to reduce the values to some small number per attribute

In this scenario, where the primary data is not available on a RDBMS, additional transformation of data is required. The raw data, after cleaning [1,2], needs to be further normalized [3] and partitioned to eliminate the numerous multi-valued attributes and to rearrange the attributes to reduce the number of null values, so that it can be used with a Relational Database Management System. Once loaded into RDBMS, Structured Query Language can be used to generalize and reduce the data and present the result into contingency tables.

The contingency table can be further analyzed to identify useful patterns. The reduction of data using SQLs substantially reduces the complexity of the discovery process and makes it feasible to spot a pattern once it is laid in the contingency table.

The following sections elaborate the above concepts. Section 2 describes the data preparation followed by section 3 where we describe the knowledge extraction and analysis tasks. Section 4 compares results obtained by this method with the results obtained from previous publications. Section 5 concludes with a discussion on limitations of using this process.

2 Data Preprocessing and Transformation

To understand the nature of data, for performing more meaningful analysis and to finally extract more meaningful knowledge, data preprocessing needs to be performed. The data preprocessing stage essentially involves data cleaning and data editing [4].

- In this scenerio data cleaning involved removing of unprintable characters and removing duplicate entries. Most of the unprintable characters were observed in the multivalued character fields like *Diagnosis* in table TSUM_A and *Symptoms* field in table TSUM_B. Duplicate entries cause Unique Constraint violation when loading in a RDBMS. In this particular scenario only one of the entries was retained and the other discarded by inspection.
- Data editing involved removing data inconsistencies like presence of an alphanumeric character in a numeric field like *>5000* or a name of the same disease spelt in different ways like *ANA(+)* and *ANA (+)* . Here the value was changed based on inspection and in accordance with the data source.

Data transformation involved further normalizing the data [4]. This helps representing the data and their relationships precisely in a tabular format that makes the database easy to understand and operationally efficient. This also reduces data redundancy and enhances performance. Applying normalization, tables TSUM_A & TSUM_B were further decomposed into several other tables [1]. Figure 1 lists all the tables derived from the above tables.

3. Knowledge Extraction using Contingency Tables

The process of analyzing data to extract useful knowledge involves searching for patterns in the data to predict an answer to the question one is looking for [9]. In this case it happens to be the prediction of occurrence of Thrombosis from the available patient data. Prediction is the process of defining, with high amount of certainty, that an attack is going to happen in the near future. This involves studying the levels of different anti-bodies and body fluids before and at the time of the attack. This can help find to patterns that can further help us predict the occurrence of the attack in the future.

To accomplish the above task the following steps have been followed:

– Preprocessed data is uploaded into a Relational Database tables defined in the previous section. This is a one time process.

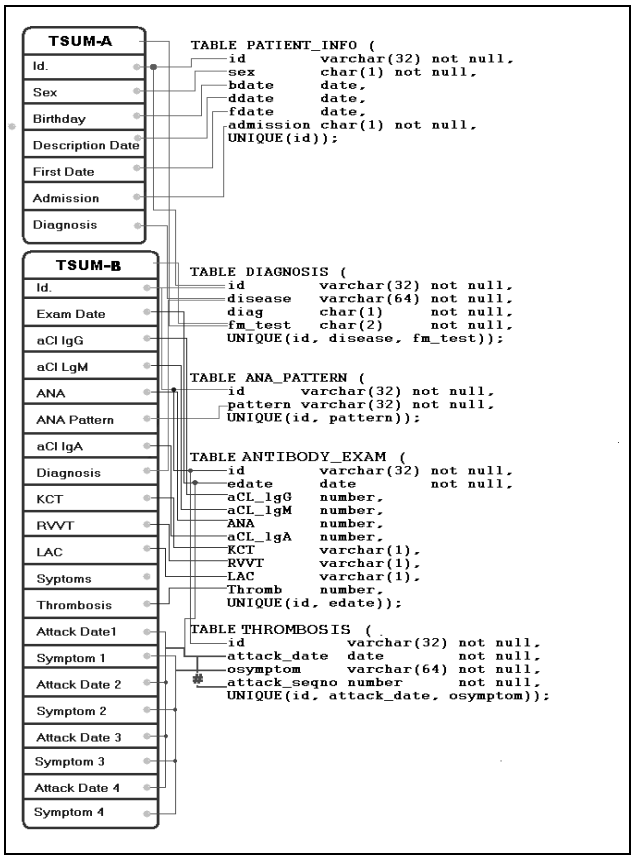

Fig. 1. Decomposition of tables Tsum_a and Tsum_b

– Data views are created from the database using SQL against the thrombosis attribute being analyzed. This process is repeated for each attribute
– The data summary is either the histogram of the attribute being analyzed or is a contingency table with respect to the different values of Thrombosis.
– For certain numerical attributes, where there are large number of groups, discretization has been applied and contingency tables created on the groups.
– The histograms and contingency tables are analyzed and conclusions regarding the effect of each predictor attribute on thrombosis are made.

Here attributes for tables Tsum_a, Tsum_b and Disease or for RDBMS tables Diagnosis, Ana_Pattern, Antibody_exam, Thrombosis and Disease have been considered. The following paragraphs describe the datamining tasks on individual attributes from these tables coupled with the target attribute *AntibodyExam.Thrombosis*. The two SQL statements are issued against the database against the attribute requesting a histogram of an attribute and requesting a cross tabulation of that attribute with *AntibodyExam.thrombosis*. The output is laid down as histograms and contingency tables. The tables are then analyzed for patterns.

It is worthwhile noting that the Antibody Exam that was done on the patient, which is represented by the table Tsum_b was said to be closely related to the time the patient underwent a thrombosis attack. Hence the values of different antibody levels and other attributes can be taken as values that can be expected during a thrombosis attack. In case when the value of thrombosis equal 0, the test was performed on patients when thrombosis was suspected [2].

The values of multivalued attributes in the tables Tsum_a & Tsum_b like *Diagnosis* or *Symptom*, we have assumed that the patient is suffering from all the disease or shows all the symptoms are listed. This fact is also reflected in the data where in the *Diagnosis* attribute when a disease is merely suspected, we have the word *'susp'* against it. Analysis of a few attributes is as follows

3.1 Analysis of Disease Type (disease.ds_type)

HISTOGRAM:

```
SELECT ds_type, count(ds_type)
FROM Disease
INNER JOIN Diagnosis
ON Disease.ds_name = Diagnosis.Diagnosis
GROUP BY ds_type
ORDER BY ds_type
```

CONTINGENCY TABLE:

```
SELECT ds_type, count(ds_type), Thrombosis
FROM Disease
INNER JOIN (Diagnosis
INNER JOIN AntibodyExam
ON Diagnosis.ID=AntibodyExam.ID)
ON Disease.ds_name = Diagnosis.diagnosis
GROUP BY ds_type, AntibodyExam.Thrombosis
ORDER BY ds_type, AntibodyExam.Thrombosis
```

DS-TYPE	Thromb(0)	Thromb(1)	Thromb(2)	Thromb(3)
C	676	110	39	11
D	18	0	0	0
N	5	0	0	0
O	21	1	0	0
U	33	16	2	0

Where:
C -> Collagen Disease D-> Non-Collagen Disease N -> No Diagnosis O -> Observation U -> Undefined

Fig. 2. Contingency table for disease type against different values of thrombosis.

Figure 2 shows the SQL queries and their corresponding results on attribute Disease.ds_type and its effect on Thrombosis. Here it can be observed that a patient with Collagen disease has more chances of getting a thrombosis attack. However a patient can be diagnosed with several diseases per diagnosis and hence the result of the query does not present a clear picture due to multiple values of the Diagnosis attribute. A patient could suffer from a Collagen disease and a Non-Collagen disease or a disease from an undefined category. A thrombosis attack for this patient can be explained by the diagnosed Collagen disease. Hence it can be assumed that the Collagen Disease should have precedence as it is clearly linked with thrombosis.

If the values of disease are ordered as follows:

$$\mathbf{C > D > N > O > U} \quad (\mathbf{1})$$

records obtained from the following SQL are inspected one at a time and one output is generated against each patient thrombosis combination, eliminating the overlapping effect using programming techniques:

DS-TYPE	Thromb(0)	Thromb(1)	Thromb(2)	Thromb(3)
C	319	34	17	5
D	16	0	0	0
N	3	0	0	0
O	14	0	0	0
U	31	14	1	0

Table 1. Contingency table for disease type after removing overlapping values.

After removing the overlap there was not too much change as far as the undefined values, disease type 'U', go. The providers of the data need to provide the definitions of the undefined diseases. Besides the undefined category of diseases the following can be clearly observed from the above table.

- If a person has an attack of Thrombosis, he/she is suffering from a disease that is Collagen in nature

3.2 Analysis of ANA Pattern (anapattern.pattern)

Figure 3 shows the SQL queries and their corresponding results on attribute ANAPattern.Pattern and its effect on Thrombosis. Similarly there is an overlap among attributes and hence it is not a very clear picture. Again if precedence is defined like before, a more realistic picture can be observed. Defining the precedence as:

$$\mathbf{D > S > P > N} \quad (\mathbf{2})$$

we have the results as shown in table 2

HISTOGRAM:

```
SELECT Ana, count(Ana)
FROM AntibodyExam
GROUP BY Ana
ORDER BY Ana
```

CONTINGENCY TABLE:

```
SELECT Ana,count(Ana), Thrombosis
FROM Thrombosis
GROUP BY Thrombosis.AttackNumber, AntibodyExam.Thrombosis
ORDER BY Thrombosis.AttackNumber, AntibodyExam.Thrombosis
```

ANA	Thromb(0)	Thromb(1)	Thromb(2)	Thromb(3)
0	198	3	0	0
4	102	1	2	0
16	140	8	0	3
64	95	11	3	1
256	64	13	4	0
1024	44	9	2	0
4096	33	4	6	1

Fig. 3. Contingency table for ana pattern against thrombosis

PATTERN	Thromb(0)	Thromb(1)	Thromb(2)	Thromb(3)
D	7	2	0	1
N	1	0	0	0
P	160	12	3	2
S	295	33	14	2

Table 2. Contingency table for ana pattern after removing overlapping values

Here after removing the overlapping values we observe that there is no big difference than we had before. The results are not very conclusive. However the following can be observed :

- Pattern ‘N’, though occurs rarely among patients, indicates no thrombosis.
- Patients with pattern ‘D’ have more chance of getting thrombosis attack than any other ana pattern.
- Patients with pattern ‘S’ have a significantly bigger chance of getting thrombosis attack than patients with pattern ‘P’

3.3 Analysis of Symptom (thrombosis.symptom)

Figure 4 shows the SQL queries and their corresponding results on attribute Thrombosis.Symptom and its effect on Thrombosis.

The following can be observed from the table in figure 4:

HISTOGRAM:

```
SELECT Symptom, count(Symptom)
FROM Thrombosis
GROUP BY Symptom
ORDER BY Symptom
```

CONTINGENCY TABLE:

```
SELECT Symptom, count(Symptom), AntibodyExam.Thrombosis
FROM Thrombosis
INNER JOIN AntibodyExam
ON Thrombosis.ID=AntibodyExam.ID
GROUP BY Thrombosis.Symptom, AntibodyExam.Thrombosis
ORDER BY Thrombosis.Symptom, AntibodyExam.Thrombosis
```

SYMPTOM	Thromb(0)	Thromb(1)	Thromb(2)	Thromb(3)
ABORTION	0	15	0	0
ALVEOLAR HEMORRHAGE	0	0	2	0
AMI	0	14	0	0
APO	0	1	0	0
BRAIN INFARCTION	0	6	0	0
BUDD-CHIARI	0	2	0	0
BURGER	0	4	0	0
CNS	0	5	37	0
CVA	0	17	0	0
DVT	0	15	0	0
EDEMA	0	3	0	0
EPILEPSY	0	11	0	0
FINGER ULCER	0	3	0	0
ITP	0	2	0	0
LEG ULCER	0	4	0	0
LUNG EMBOLI	0	2	0	0
MI	0	2	0	0
PH	0	8	0	0
PULMONARY EMBOLISM	0	2	0	0
RETINA ARTERY OCCLUSION	0	2	0	0
SPLEEN INFARCTION	0	1	0	0
THROMBOCYTEPENIA	0	0	0	7
THROMBOPHLEBITIS	0	10	0	0
THROMBOSIS	0	0	1	0
THROMBOSIS CENTRAL RETINAL	0	2	0	0
THROMBOSIS MULTIPLE	0	4	0	0
THROMBOSIS OF LOWER EXTREMITIES	0	2	0	0
THROMBOSIS PULMONARY	0	4	0	0
THROMBOSIS RENAL ARTERY	0	2	0	0
THROMBOSIS SVC	0	1	0	0

Fig. 4. Contingency table for symptom against thrombosis

- Most of the Symptoms observed are associated with severe thrombosis.
- Symptoms like Alveolar Hemorrhage, CNS, Thrombocytepenia is associate with milder attacks.

3.4 Analysis of Attack Number (thrombosis.attacknumber)

Figure 5 shows the SQL queries and their corresponding results on attribute Thrombosis.attacknumber with its effect on Thrombosis

HISTOGRAM:

```
SELECT AttackNumber, count(AttackNumber)
FROM Thrombosis
GROUP BY AttackNumber
ORDER BY AttackNumber
```

CONTINGENCY TABLE:

```
SELECT AttackNumber,count(AttackNumber),AntibodyExam.Thrombosis
FROM Thrombosis
INNER JOIN AntibodyExam
ON Thrombosis.ID=AntibodyExam.ID
WHERE AttackNumber > 0
GROUP BY Thrombosis.AttackNumber, AntibodyExam.Thrombosis
ORDER BY Thrombosis.AttackNumber, AntibodyExam.Thrombosis
```

ATTACK NO.	Thromb(0)	Thromb(1)	Thromb(2)	Thromb(3)
0	104	80	19	5
1	0	50 (35)*	18 (16)*	0
2	0	15 (9)*	2 (1)*	0
3	0	6 (2)*	1	0
4	0	4 (1)*	0	0
5	0	3	0	0

(N)* -> 'N' number of patients had exactly number of attack associated with them. It may be noted that a patient who had 5 attacks had the 4th attack too. Hence the number within the braces indicates the number of patients who had exactly the number of attacks against the Attack No. Listed on the corresponding left hand side.

Fig. 5. Contingency table for attack number against thrombosis

It may be noted from the tables that the attack numbers from 1 to 5 do not have the strength or degree of the thrombosis attack associated with them. The original table, Tsum_b, has only one value for the degree of thrombosis attack and that is associated with attack 0. Hence the values listed under various intensities of thrombosis may not be entirely true as there are no values listed against them. The above output, however, does show some interesting results;

- The milder the attack of Thrombosis against a patient, the lesser is the chance that it will relapse again in the future. Patients who were detected with the least degree of thrombosis, or attribute thrombosis = 3, had no relapse whatsoever.
- Conversely if the attack of Thrombosis is severe, there are more chances that it will relapse again in the future.
- If it can be assumed that most of the patients had survived the attacks, it can be inferred that a human body develops immunity against thrombosis after each attack

3.5 Analysis of ANA (antibodyexam.ana)

Figure 6 shows the SQL queries and their corresponding results on attribute AntibodyExam.ana and its effect on Thrombosis.

HISTOGRAM:

```
SELECT Ana, count(Ana)
FROM AntibodyExam
GROUP BY Ana
ORDER BY Ana
```

CONTINGENCY TABLE:

```
SELECT Ana,count(Ana), Thrombosis
FROM Thrombosis
GROUP BY Thrombosis.AttackNumber, AntibodyExam.Thrombosis
ORDER BY Thrombosis.AttackNumber, AntibodyExam.Thrombosis
```

ANA	Thromb(0)	Thromb(1)	Thromb(2)	Thromb(3)
0	198	3	0	0
4	102	1	2	0
16	140	8	0	3
64	95	11	3	1
256	64	13	4	0
1024	44	9	2	0
4096	33	4	6	1

Fig. 6. Contingency table for ana against thrombosis

Since there are number of numerical values, detecting patterns in this table is difficult.

Hence we discretize the values of ANA so that some pattern can be made visible. Putting cut after ANA value of '4', by inspection, we get the following table.

ANA	Thromb(0)	Thromb(1)	Thromb(2)	Thromb(3)	%
0 - 4	300	4	2	0	1.5%
16 - 4096	376	45	15	5	17.3%

Table 3. Contingency table for ana pattern after grouping

The following can be observed from the above table:

- Patients with higher value of ANA antibody have a higher chance of getting a Thrombosis attack.

3.6 Results from analysis of all attribues

Similarly using the above process on all attribues we get the following results:

- If a person has a attack of Thrombosis, he/she is suffering from a disease which is Collagen in nature.
- Pattern ‘N’, though occurs rarely among patients, indicates no Thrombosis.
- Patients with Pattern ‘D’ have more chance of getting Thrombosis Attack than any other Ana Pattern.
- Patients with Pattern ‘S’ have slightly more chance of getting Thrombosis Attack than Patients with Pattern ‘P’.
- Most of the Symptoms observed are associated with severe thrombosis.
- Symptoms like Alveolar Hemorrhage, CNS, Thrombocytepenia are associated with milder attacks.
- The milder the attack of Thrombosis against a patient, the lesser is the chance that it will relapse again in the future. Patients who were detected with the least degree of thrombosis, or attribute thrombosis = 3, had no relapse whatsoever.
- Conversely if the attack of Thrombosis is severe, there are more chances that it will relapse again in the future.
- If it can be assumed that most of the patients did survive the attacks, it can be inferred that a human body develops immunity against thrombosis after each attack.
- Patients with higher value of ANA antibody have a higher chance of getting a Thrombosis attack.
- The chances of a thrombosis attack increases with the increase in concentration of aLCIgG antibody.
- The increased concentration of aCLIgG antibody has somewhat more effect on the most severe form of thrombosis.
- The chances of a thrombosis attack increases with the increase in concentration of aLCIgM antibody
- The increased concentration of aCLIgM antibody has somewhat more effect on the most severe form of thrombosis.
- The chances of a thrombosis attack increases with the increase in concentration of aLCIgA antibody.
- The increased concentration of aCLIgA antibody has somewhat more effect on the most severe form of thrombosis.
- A a positive value of KCT indicates a high chance of getting a thrombosis attack.
- A positive value of RVVT indicates a high chance of getting a thrombosis attack.

4. Comparision with other results

Three papers have been presented on this medical data set in the past. The following paragraphs describes the papers briefly and compares their result with what we have obtained using the Contingency table method.

In the paper Using WizSoft, presented at PKDD'99, by Levin B., Meidan A, Cheskis A, Gefen O and Vorobyov I., we found that only one rule comes from Tsum_b [6].

Using WizSoft *Rules obtained from TSUM_B only* *Presented in PKDD99*	**Using Contingency table** *Rules obtained from TSUM_B* *Presented in this Thesis*
If *Diagnosis* is **APS** than Thrombosis is **Yes**	Not Investigated
	More Rules listed on the previous section

Table 4. Comparision of results from past work WizSoft

In the paper "Using Other Measurements and Decision Tree", presented at PKDD 99, by Taylor C [7], all the 3 tables Tsum_a, Tsum_b and Tsum_c are taken into consideration and the rules obtained are presented as decision trees using multiple attributes. Hence the results presented in this paper cannot be compared to the ones obtained here as we have used only single attribute analysis against the decision attribute, thrombosis.

In the paper Using InfoZoom, presented in PKDD99, by Beilken C and Spenke M [8].

Using InfoZoom *On Table TSUM_B* *Presented during PKDD-99*	**Using Contingency Tables** *On Table TSUM_B*
Thrombosis is 3 if *Symptom* is **thrombocytopenia** *Thrombosis* is 2 iff *Symptom* is **CNS** *Thrombosis* is 1 if any other *Symptom* is present	**Alveolar Hemorrhage** implies Thrombosis = 2 **Thrombocytopenia** implies Thrombosis = 3 **CNS** is associated with Thrombosis = 2 Most of the Symptoms observed are associated with most severe Thrombosis or Thrombosis = 1.
KCT, RVVT **positive** implies *LAC* **positive**	Not investigated
With positive values of *LAC* the chance of Thrombosis increases.	Positive values of *LAC, RVVT, KCT* indicate a high chance of getting a thrombosis attack.
With high values of *ANA* the changes of thrombosis increases	Patients with higher value of *ANA* antibody have a higher chance of getting a thrombosis attack.
High values of *aCL IgG, aCL IgA, aCL IgM* are good indicators of *Thrombosis*	The chances of Thrombosis increases with increase in concentration of *aCL IgG, aCL IgA, aCL IgM* antibodies
If *LAC* is not measured at all the chances of Thrombosis is only 3.5%	
	Increase in concentration of *aCL IgG, aCL IgA, aCL IgM* antibodies has somewhat more effect on the most severe form of Thrombosis.

	If a patient has an attack, he/she is suffering from a Collagen disease.
	Ana Pattern '**N**' occurs rarely among patients and indicates no Thrombosis. *Ana Pattern* '**D**' has more chance of getting thrombosis attack than any other Pattern. Patients with Pattern '**S**' have significantly higher chances of getting Thrombosis attack than patients with Pattern '**P**'.
	The milder the attack of Thrombosis, the lesser are the chances it will relapse again in the future. If the Thrombosis attack is severe, there are more chances it will relapse again. If it can be assumed that most patients did survive the attacks, it can be inferred that a human body develops immunity against thrombosis after each attack.

Table 5. Comparison of results from past work InfoZoom

5. Conclusions

From the above comparisons it can be seen that by using Contingency Tables we have been able to extract almost all the rules that were obtained using other data_mining Systems like InfoZoom plus some more.

Contingency Tables are generally easier to examine and can be applied in a simple uniform way what can be clearly observed from our data_mining efforts. However if the Contingency table is large, it becomes difficult to interpret the results. Here the data reduction techniques like discretization become necessary to reduce the size of the table. Although we have chosen the intervals for discretization merely by inspection, there are ways to select the intervals where the cuts are not very obvious.

Acknowledgement : The author likes to thank Dr. Zbigniew Ras for taking his valuable time to edit this paper and offer useful remarks and suggestions.

6. References

1. Zytkow J., Tsumoto S., Takabayashi K., Medical (Thrombosis) Data description In:Berka P., Siebes A. *4th European Conference on Principles and Practice of Knowledge Discovery in Databases.*
2. Tsumoto. S 1999 Guide to Medical Data Set In: Berka P. Ed. *Workshop Notes on Discovery Challenge, PKDD-1999* Sep 15-18 Univ of Economics, Prague, p45-47.
3. Elmasri R., Navathe S. 1999 *Fundemental of Database Systems* Part-I and Part-II.
4. Famili A., Shen W., Weber R.. and Simoudis E., Paper 1996, *Data Preprocessing and Intelligent Data Analysis* In: http://www-east.elsevier.com/ida/browse/96-1/ida96-1.htm.

5. *Two-variable Contingency Tables* In: http://www-eksl.cs.umass.edu/eis/pages/techniques /table-const.html.
6. Beilken. C 1999 Visual, Interactive Data Mining with Infozoom - Medical Data Set In: Berka P. Ed. *Workshop Notes on Discovery Challenge , PKDD-1999* Sep 15-18 Univ of Economics, Prague, p49-54.
7. Meidan A., Cheskis A., Gefen O., Levin B., Vorobyov I, The WizWhy analysis of the PKD2000 Discovery Challenge Medical Domain In:Berka P., Siebes A. *4th European Conference on Principles and Practice of Knowledge Discovery in Databases.*
8. AhmedY., Strickland K, Mining Medical Data for Casual and Temporal Patterns In:Berka P., Siebes A. *4th European Conference on Principles and Practice of Knowledge Discovery in Databases.*
9. Zytkow J, Zembowicz R 1997 From Contingency table to other forms of knowledge In Fayyad U, Shapiro G Smyth P and Uthurswamy R. *Advances in Knowledge Discovery and Data Mining.*