Lessons Learned from Data Mining of WHO Mortality Database

W. Paoin
Faculty of Medicine, Thammasat University, Pathumthani, Thailand

1. Introduction

1.1 Data Mining

Data mining is about solving problems by analyzing data already present in databases. It is defined as the process of discovering patterns in data. The process may be automatic or semiautomatic. The patterns discovered may be meaningful in the way that they lead to new knowledge.

Data mining functionalities include the discovery of concept/class descriptions, associations and correlations, classification, prediction, clustering, trend analysis, outlier and deviation analysis, and similarity analysis. They belong to a young interdisciplinary field grown from increasing demand for efficient and effective advanced data analysis tools [1, 2].

Classification is the process of finding a model (or a function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. Unlike classification, which analyzes class-labeled data objects, clustering analyzes data objects without consulting a known class label. In general, the class labels are not present in the data simply because they are not known to begin with. Association analysis or frequent pattern mining focuses on finding patterns such as item sets that frequently appear in a data set to find correlations between each item.

1.2 Application of Data Mining in Biomedicine

Recent research studies in biomedical data mining [3] showed that a data mining method could be used to develop clinical diagnostic and prognostic systems, to interpret biomedical signal and image data and to discover knowledge from biological and clinical databases in bio-surveillance and anomaly detection applications.

A mortality database contains data about each deceased person including sex, age, date and time of death, cause(s) of death and place where the death occurred. Usually it is included in the vital statistics of each country and data are collected from death registration systems of each country. Conventional analysis of mortality data used descriptive and analytic statistics, then was presented in each country health statistics in a tabular form, such as the top five or top ten leading causes of death, causes of death by age group, sex, region, etc. [4].

Correspondence to:
Wansa Paoin
Faculty of Medicine
Thammasat University
Rangsit Campus, Paholyotin Road
Pathumthani 12120
Thailand
E-mail: wansa@tu.ac.th

Keywords
Mortality statistics, data mining, classification, clustering, association analysis

Summary

Objectives: The objectives of this research were to test the ability of classification algorithms to predict the cause of death in the mortality data with unknown causes, to find association between common causes of death, to identify groups of countries based on their common causes of death, and to extract knowledge gained from data mining of the World Health Organization mortality database.

Methods: The WEKA software version 3.5.3 was used for classification, clustering and association analysis of the World Health Organization mortality database which contained 1,109,537 records. Three major steps were performed: Step 1 – preprocessing of data to convert all records into suitable formats for each type of analysis algorithm; Step 2 – analyzing data using the C4.5 decision tree and Naïve Bayes classification algorithm, K-means clustering algorithm and Apriori association analysis algorithm; Step 3 – interpretation of results and hypothesis testing after clustering analysis.

Results: Using a C4.5 decision tree classifier to predict cause of death, we obtained 440 leaf nodes that correctly classify death instances with an accuracy of 40.06%. Naïve Bayes classification algorithm calculated probability of death from each disease that correctly classify death instances with an accuracy of 28.13%. K-means clustering divided the data into four clusters with 189, 59, 65, 144 country-years in each cluster. A Chi-square was used to test discriminate disease differences found in each cluster which had different diseases as predominant causes of death. Apriori association analysis produced association rules of linkage among cancer of the lung, hypertension and cerebrovascular diseases. These were found in the top five leading causes of death with 99–100% confidence level.

Conclusion: Classification tools produced the poorest results in predicting cause of death. Given the inadequacy of variables in the WHO database, creation of a classification model to predict specific cause of death was impossible. Clustering and association tools yielded interesting results that could be used to identify new areas of interest in mortality data analysis. This can be used in data mining analysis to help solve some quality problems in mortality data.
Recent research on mortality data focused on several aspects. Some studies tried to identify socio-economic factors that increased premature death in different regions [5, 6]. Others tried to find factors that increased the risk of death in various groups of patients [7, 8]. Only a few studies used conventional data analysis to calculate the risks and identify preventable management processes to reduce mortality risks. Recent studies [9, 10] used data mining as a new tool for mortality data analysis. This showed trends of increasing usage of data mining in analysis of mortality data.

1.3 Analysis of WHO Mortality Database

The World Health Organization (WHO) mortality database [11] contains mortality statistics, i.e. number of deaths via country, year, sex, age group and causes of death, as far back as the 1950s. Included data were collected from countries that reported them properly according to the International Classification of Diseases (ICD) coding. MortIcd10.zip contained all mortality data according to the detailed 10th revision of the ICD, either with 3-character or 4-character ICD 10 codes with the total records of 1,109,537.

The mortality database has been published on the WHO website since 2005 to “allow researchers to use them” [12]. However, few researchers have used this database to create new knowledge on the cause of death analysis and comparison between preventable mortality at an international level. In 2006, it was found that only one study [13] used this database to analyze trends of mortality among world regions. So, during the WHO Family of International Classification (WHO-FIC) meeting in Trieste, Italy in 2007, an announcement was made to identify voluntary researchers to do further analysis on the mortality database. The Thai Collaborating Center on WHO-FIC volunteered to do the analysis using three basic data mining algorithms, i.e. classification, association analysis and clustering. The results were presented in the annual WHO-FIC meeting in New Delhi, India in 2008.

Poor data quality in the mortality statistics is a current problem in many countries [12]. Some countries had a high proportion of unknown or ill-defined causes of death in their mortality data. Classification algorithms may solve the problem of unknown or ill-defined causes of death in the data by the creation of decision tree from the data with known causes of death to predict the cause of death from the data with unknown or ill-defined causes of death. An Association analysis algorithm may be useful to identify the association between causes of death, and a clustering algorithm could be used to identify groups of countries based on their common causes of death, for the purpose of mortality prevention planning in different groups of countries.

This paper highlights the findings and lessons learned from that mortality data mining.

Table 1: Samples of data from first input file after pre-processing

<table>
<thead>
<tr>
<th>“Country”, “Sex”, “Age”, “Cause”</th>
</tr>
</thead>
<tbody>
<tr>
<td>“3380”, “2”, “Adult”, “1–080”</td>
</tr>
<tr>
<td>“3380”, “1”, “YoungAdult”, “1–101”</td>
</tr>
<tr>
<td>“3380”, “1”, “YoungAdult”, “1–103”</td>
</tr>
<tr>
<td>“3380”, “2”, “YoungAdult”, “1–096”</td>
</tr>
<tr>
<td>“3380”, “2”, “YoungAdult”, “1–101”</td>
</tr>
<tr>
<td>“3380”, “2”, “Adult”, “1–038”</td>
</tr>
<tr>
<td>“3380”, “1”, “Adult”, “1–020”</td>
</tr>
<tr>
<td>“3380”, “1”, “Child”, “1–098”</td>
</tr>
</tbody>
</table>

Country data are shown via codes. Sex: 1 = male, 2 = female. Cause codes 1 – 080 represent group 80 (liver diseases) of the tabulation list 1 from ICD.

2. Objectives

The objectives of this study are 1) to test the ability of classification algorithms to predict the cause of death in the mortality data with unknown causes; 2) to find association between common causes of death; 3) to identify groups of countries based on their common causes of death; and 4) to extract knowledge gained from data mining the WHO mortality database.

3. Methods

The WHO mortality database was used in the data mining process including classification, association and clustering analysis. Three major steps are: 1) pre-processing of data, 2) analysis of data, and 3) interpretation of results.

The WEKA (Waikato Environment for Knowledge Analysis) data mining software version 3.5.3 (an open-source shareware developed by the University of Waikato in New Zealand) [14] was used as a tool for mining the WHO Mortality Database.

3.1 Pre-processing of Data

Data from the MortIcd10.zip file was converted to an appropriate format and then loaded into the Microsoft Access database software. The SQL command was then executed to extract country, year, sex, age, cause of death (in ICD10 code) and the number of deceased for each ICD code into a new table. Then a program was created to transform all the data into two suitable formats to be used as input files for the

Table 2: Samples of data from second and third input files after preprocessing

|----------------------------------|

Columns D3 to D103 represent presence (y) or absence (n) of diseases grouped as top 5 or top 10 leading causes of death for each country-year; for example D3 represents group 3 of tabulation list 1 from ICD (diarrhea), D5 represents group 5 (tuberculosis).
analysis by the WEKA software (Tables 1 and 2). The age value for each record was discretized into eight groups, comprised of infant (<1 year), small child (1 to <5 years), child (5 to <15 years), young adult (15 to <25 years), adult (25 to <45 years), middle age (45 to <60 years), senior (60 to <80 years), and elderly (≥80 years). The cause of death for each record was also discretized into 103 groups using the mortality tabulation list 1 from International Statistical Classification of Disease and Related Health Problem, 10 Revision – ICD-10 [15] (for example A09 was discretized to 1-003 etc.).

The first input file comprised of country, sex, age and cause fields was used for the classification algorithm. The second input file comprised of country-year and top-ten causes of death was used for the clustering algorithm. The third input file comprised of country-year and top-five causes of death was used for the association analysis. Parts of these input files are shown in Tables 1 and 2.

### 3.2 Analysis of Data

Two classification algorithms (C4.5 decision tree and Naïve Bayes) were used to classify causes of death based on country, sex, and age of the deceased.

A Scatter plot between country-year and the top ten leading causes of death roughly showed the distribution of data into four groups (Fig. 1). So K-means clustering with k-value set to four groups were used to cluster country-year based on the top-ten causes of death.

Apriori algorithm was chosen for association analysis because its standard methods are well known and it is one of the most popular association analysis algorithms.

### 4. Results

For classification analysis, the C4.5 algorithm generated a decision tree with 440 leaf nodes to predict causes of death. It correctly classified death instances at 40.06% and incorrectly classified them at 59.94%, while Naïve Bayes correctly classified death instances at 28.13%, and incorrectly classified them at 71.87%.


Clusters changed trend each country-year, which could be plotted into a graph as shown in Fig. 2.

The Chi-Square test was employed to test the different proportions of the top-ten causes of death between each cluster, as shown in Table 3 (using alpha value = 0.05).

The association analysis using the Apriori algorithm on the top five causes of death for each country found 20 association rules with confidence levels at 86 to 100%. The five best rules are shown in Table 4. It was obvious that if a country had CA lung included in the top five causes of death, that country will also have hypertension and cerebrovascular diseases with the confidence level at 99–100%.

5. Discussion

Data mining theory and techniques have been evolving for more than 20 years. These techniques are undoubtedly useful to medical and health data analysis [16–19]. However few studies have used data mining to analyze mortality data at national or international level. The study of Paoin et al. [20] used data mining to analyze mortality data of WHO mortality database and found some interesting results.

Table 3 A sample of the Chi-Square test to discriminate the difference proportions of the top ten causes of death between clusters

<table>
<thead>
<tr>
<th>DzName</th>
<th>no. reported as top 10 causes in</th>
<th>Chi-Square Prob. Test among cluster (alpha = .05)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C0</td>
<td>C1</td>
</tr>
<tr>
<td>Diarrhea</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Respiratory TB</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Septicemia</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Viral hepatitis</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HIV</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Other infectious</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CA Stomach</td>
<td>18</td>
<td>10</td>
</tr>
<tr>
<td>CA Colon, rectum</td>
<td>133</td>
<td>38</td>
</tr>
<tr>
<td>CA liver, biliary</td>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td>CA Breast</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>CA Ovary</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CA Prostate</td>
<td>18</td>
<td>7</td>
</tr>
<tr>
<td>Other CA</td>
<td>180</td>
<td>48</td>
</tr>
</tbody>
</table>

* Chi-square test for difference of proportion between each cluster with statistical significance (for example, in row diarrhoea at column 0123 = Chi-square test of difference between cluster 0, 1 and 2 with p < 0.001). na = Chi-square test not applicable due to number of diseases in all clusters = 0

Fig. 2 Trend of clusters of some countries which changed each country-year.
ternational levels. This research is the first to use data mining with the mortality database at an international level.

Common classification techniques are decision tree building, Bayesian classifier, and neural network. In this study, C4.5 decision tree and Naïve Bayes were used for classification. Since the mortality data is not numeric, Neural networks that require only numerical data could not be used.

The C4.5 algorithm uses the concept of information gained to build a decision tree for each problem of classification. In this study, the decision tree was built using the age group of the deceased as its first split point after the root node. The second split point was country, the last split point was sex. The Naive Bayes classification built a probability table of death from each disease. The Naive Bayes yielded the poorer result with 28.13% accuracy, while Naïve Bayes yielded the poorer result with 28.13% accuracy.

The reason why classification tools produce poor results with high error rates in this study could be explained by the availability of only a few attributes (country, sex, age, cause of death). This is a well known weakness of classification algorithms in data mining [20]. This problem may be solved by adding more attributes to the database and/or reducing the number of classes before conducting classification studies in the future. By doing so, classification tools could be used to predict the causes of death from datasets with ill-defined causes of death, a common problem with mortality data in some countries.

The K-means Clustering algorithm produced satisfactory results in this study and could divide country-year according to patterns of the top-ten causes of death into four clusters with different patterns. Since clustering algorithms only divided data into clusters, they gave no clues about characteristic of each cluster. This study used a Chi-square test to discriminate the difference of proportion of the top 10 leading causes of death among each cluster.

Chi-square tested the discrimination difference of the proportion from diseases found in each cluster. The null hypothesis found no difference between proportion disease clusters. The alpha value was set to 0.05. If the p-value from the test was less than 0.05, we would reject the null hypothesis. For example, in table 3, the Chi-square test showed the HIV disease proportion between 4 clusters (column 0123) was different (p < 0.001), but when compared among clusters 0 and 1 and 2 (column 012), the HIV disease proportion was not different (p=0.148). We also noted that the number of HIV diseases found in cluster 3 was 34, which seemed to be much more than the number found in clusters 0, 1 and 2. Thus, we could conclude that the HIV disease was a predominant cause of death in the cluster 3 data. Using this inference method, we could conclude the findings as follows:

Cluster 0 (189 country-years) had cancer, cardiovascular disease and cerebrovascular disease as predominant causes of death.

Cluster 1 (59 country-years) had cancer, cardiovascular disease and cerebrovascular diseases as common causes as found in cluster 0. Three more common causes found in only cluster 0 and 1 are CA colon, Alzheimer’s disease, and ischemic heart disease.

Cluster 2 (65 country-years) had infectious disease, malnutrition and traffic accidents as the predominant causes of death, similar to cluster 3, with other external causes that were only found in clusters 2 and 3.

Cluster 3 (144 country-years) had infectious disease, malnutrition and traffic accidents as predominant causes of death.

The K-means clustering algorithm with Chi-square was used to test discrimination differences of proportion of the top 10 leading causes of death. This displayed an overview picture of different cause patterns. This could be used in the development of different models and tools for the improvement of a country’s mortality data. For example, verbal autopsy tools could be refined to four different versions to be used in each cluster. The trend analysis of cluster changes for each country such as in figure 3. This could also be performed in countries with trend changes. For example, Thailand should do further analysis to find the reason of cluster change between clusters 3 and 2 during the years 1998–2003. This analysis should include socio-economic factors into the study.

The Apriori association analysis also showed interesting results. It displayed the association rule of linkage among cancer of the lung, hypertension and cerebrovascular diseases that could be found together in the top five causes of death with a 99–100% confidence level. This means any country that had cancer of the lung in the top five leading causes of death surely had hypertension and cerebrovascular diseases in the top five leading causes of death as well. This could lead to further research on patho-physiologic processes of diseases leading to death.

Among the three major tools in data mining, classification produced the poorest results due to low accuracy while clustering tools and association tools found new data that had never been found in mortality analysis using conventional mortality statistical methods.

6. Conclusions

Several lessons were learnt from mining the mortality database. Firstly, we found an alternative way to analyze the mortality database as opposed to conventional statistical models. Secondly, results from data mining uncovered some interesting patterns including the clustering of countries with similar characteristic in other groups. Finally, we found an alternative way of analysis, which can probably be used to solve some quality problems in mortality data. For example, we can use clustering knowledge to develop different models and tools to improve mortality data for each cluster.

### Table 4

<table>
<thead>
<tr>
<th>Rule</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. CA lung=y Cerebrovascular disease=y</td>
<td>0.99</td>
</tr>
<tr>
<td>2. CA lung=y Hypertension=y Cerebrovascular disease=y</td>
<td>0.99</td>
</tr>
<tr>
<td>3. CA lung=y Hypertension=y</td>
<td>0.99</td>
</tr>
<tr>
<td>4. CA lung=y Cerebrovascular disease=y Hypertension=y</td>
<td>0.99</td>
</tr>
<tr>
<td>5. CA lung=y Hypertension=y Cerebrovascular disease=y</td>
<td>0.99</td>
</tr>
</tbody>
</table>
Thus, in countries with a high proportion of ill-defined death in their mortality data, an appropriate classification model (with high accuracy prediction value) should be selected followed by building the decision tree from well known causes of death data to predict the cause of death. This approach can be used for developing countries with a high percentage of ill defined causes of death [21], thus improving the quality of mortality data could be achieved globally.

Acknowledgments
This research was supported by the Thai Health Standard Coding Center (THCC), Ministry of Public Health, Thailand and Thai Collaborating Center for WHO-Family of International Classification. The author would like to thank anonymous reviewers for helpful and useful comments and suggestions.

References