

Mining California Vital Statistics Data

Du Zhang, Quoc Luan Ha and Meiliu Lu

*Department of Computer Science
California State University
Sacramento, CA 95819-6021
{zhangd, ha, mei}@ecs.csus.edu*

Abstract

Vital statistics data offer a fertile ground for data mining. In this paper, we discuss the results of a data mining project on the causes of death aspect of the vital statistics data in the state of California. A data mining tool called Cubist is used to build predictive models out of two million cases over a nine-year period. The objective of our study is to discover knowledge that can be used to gain insight into various aspects of mortality in California, to predict health issues related to the causes of death, to offer an aid to decision- or policy-making process, and to provide useful information services to the customers. The results obtained in our study contain valuable new information.

Keywords: *vital statistics data, causes of death, data mining, predictive models, Cubist.*

1 Introduction

Several types of data constitute what are commonly known as vital statistics data. These include births, deaths, fetal deaths, marriages, and divorces. The most commonly used types of vital statistics data in public health are data on births and deaths. Birth and death data are derived from the information reported on birth and death certificates sent to the offices of local and state registrars.

One of the most important public health functions is the monitoring of a population's health status. Vital statistics data provide a valuable source of information regarding the health status of a population.

In this paper, we discuss the results of a data mining project on the causes of death aspect of the vital statistics data in the state of California. A data mining tool called Cubist [8] is used to build predictive models out of nearly two million cases over a nine-year period. The objective of our study is to discover knowledge that can be used to gain insight into various aspects of mortality in California, to predict health issues related to the causes of death, to offer an aid to decision- or policy-making process, and to provide useful information services to the customers. So far we have not found any published work in the literature on mining vital statistics data.

2. Data Preparation

Identifying Sources of Data. There are two main sources of data used in our study, namely, Death Statistical Master File (DSMF) and Estimated Population File (EPF). The DSMF data files contain data from the death certificates registered in California [4]. Each file contains a year worth of death data and includes detailed information concerning the decedent, the place of death, and the medical data related to the death. The total size of DSMF files used in the project is 1,995,398 records for the period of 1989-1997. The EPF is prepared by the Demographic Research Unit of the California Department of Finance [3], and is used for the purpose of calculating the death rates for the training and test sets in the project.

Data preprocessing. This step is the most time consuming one. After acquiring 1.99 millions of records from DSMF and 9090 records from EPF, data preprocessing in our study consists of *data consolidation and cleaning, data reduction, and data integration and transformation.*

Data cleaning is carried out by eliminating or handling noisy, missing or inconsistent data. Data reduction process results in six attributes being used in our study (out of 59 attributes in DSMF).

The causes of death (COD) in our study are coded according to the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) [6].

To build predictive models from data, we need to have the death rate information in both the training and test data sets. Because the information is not explicitly contained in the original data files (DSMF, EPF), we derive the death rates from the data in DSMF and EPF through data integration and transformation.

Training/Test Data Sets Preparation. Before we can start the data mining process with Cubist, the *names, data and test* files must be in place.

3 Mining Process

The mining process is essentially COD-centric. It is geared toward generating models not only from the data of all-COD as a whole, but also from the data of each individual COD. There are several issues regarding how to

carry out the mining process. How are training and test data sets selected and what are their respective sizes? What type of models is to be generated, rule-based or composite? Is a model generated just for a particular COD or for all-COD?

Data Selection Strategies. We adopt two strategies in selecting training and test data for the mining process. The first strategy (S1) uses the data collection from the period of 1989-1996 as the training set and the data from 1997 as the test set. The second strategy (S2), on the other hand, randomly selects both the training set data and the test set data from the entire period of 1989-1997.

Training and Test Data Partitions. In S2, we further define five different ways (T1, ..., T5) of partitioning data into the training and test sets.

Model Generations. Of all the models generated in our study, 216 of them are defined according to the data selection strategies, training and test sets partitions, model types to be generated, and particular COD involved.

Committee Model. In addition to rule-based and composite models, Cubist can also generate what is referred to as the *committee* models out of several rule-based models. What a committee model does is that each member of the committee produces a target value for a case and the members' predictions are then averaged to yield a final prediction [8].

In our study, we select a best representative model (either rule-based or composite) for each COD based on considerations of prediction accuracy, average errors and relative errors. If a selected model is rule-based and its prediction accuracy is fairly high, we then turn on the committee model option to fine-tune the selected rule-based model for further performance improvements. Thus, 24 additional committee models were produced as a result of the fine-tuning process.

4 Result Analysis

Prediction Accuracy and Average Error. We compared the prediction accuracies and average errors under S1 and S2. There are some phenomena that depend on data selection strategies, and some that do not.

Impact of Training and Test Data Partitions. When data set sizes are small, the improvement of prediction accuracy is obvious as the training set size increases for some rule-based models.

Comparison of Rule and Composite Models. Composite models in general perform better than rule models (especially under S2). This is consistent with the observation that composite models are more effective when the number of attributes is small and all attributes are relevant to the prediction task.

Model Decomposition and Analysis. Once a model is generated, there are a number of issues pertaining to the rules in the model: their *objective* and *subjective*

interestingness [9], and reorganization for ease of analysis purpose [7]. In our study, we performed model decomposition and analysis on some obtained models to gain further insight.

Accuracy of Cubist Models. We compared the prediction results of Cubist models with those of the Vital Statistics of California Reports [1, 2]. Most of the Cubist results are consistent with the published reports.

Surprising Results. The models produced by Cubist also contain surprising results that are not found in the official published reports such as Vital Statistics of California [1, 2]. Most of those surprises represent valuable new information. Including marital status as an attribute during the mining process helped unearth valuable new information.

5 Conclusion

Our work pertains to a fertile ground for data mining. There is much to be done in the domain of vital statistics data. Future work can be pursued in several directions: generating and analyzing predictive models that include additional attributes such as individual underlying cause of death, place of occurrence, level of education, and place of residence, and that are from a subset of COD.

Acknowledgement. We would like to express our appreciation to Mike Quinn, Manager of VSS, Department of Health Services, State of California, for his help and comments

References

- 1 Center for Health Statistics, Department of Health Services, *Advance Report: Vital Statistics of California 1998*, February 2000.
- 2 Center for Health Statistics, Department of Health Services, *Vital Statistics of California 1997*, February 2000.
- 3 Department of Finance, State of California, <http://www.dof.ca.gov>.
- 4 Department of Health Services, Center for Health Statistics, State of California, <http://www.dhs.ca.gov/hisp/chsindex.htm>.
- 5 Q.L. Ha, *Knowledge Discovery on the State of California Health Statistical Data*, Master Degree thesis, Department of Computer Science, California State University, Sacramento, May 2001.
- 6 International Classification of Diseases 9th Revision, Clinical Modification (Volume One), <http://www.mcis.duke.edu/standards/termcode/icd9/>.
- 7 B. Liu, M. Hu and W. Hsu, Multi-Level Organization and Summarization of the Discovered Rules, Proceedings of the *Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 20-23, 2000, pp.208-217.
- 8 Rulequest web site, <http://www.rulequest.com/cubist-win.html>.
- 9 A. Silberschatz and A. Tuzhilin, What Makes Patterns Interesting in Knowledge Discovery Systems, *IEEE Transactions on Knowledge and Data Engineering*, Vol.8, No.6, December 1996, pp.970-974.