How Healthy Is Your Agency?
Employing Data Mining in a Health Agency System

Richi Nayak
School of Information Systems
Queensland University of Technology
Brisbane, QLD 4001, Australia

David Warren
School of Computer and Information Science
University of South Australia
Adelaide, SA 5095, Australia

Abstract

With so many health organisations housing large data sources, data mining is becoming increasingly popular as the benefits are recognised in health industry. This paper aims to identify an application problem in health system and solve it using existing data mining methods. A Classification data mining technique has been chosen here to explore a health agency's data so we can project new cases by looking at past experience with known answers. The discovered knowledge can then be applied in the health agency to increase the working efficiency and improve the quality of decision making.

Keywords: Data Mining, Health Agency System, Classification task, Decision Tree

1 Introduction

The ability to accumulate large volumes of data has become more widespread with the extensive availability of low-cost powerful computers. These large volumes of data can be readily converted into useful information by employing data mining techniques. Data mining, in general, is the task of automatically extracting implicit, previously unknown, valid and potentially useful information from data [1].

Each data set is unique to its application and therefore holds potentially important information about the organisation. Due to the uncertainty of the results, it is difficult to know what to expect from a data mining process. This paper contributes to the understanding of how predictive modelling behaves in the mining of data sets from a health agency.

2 The Health Agency System

The system that has been chosen for this paper belongs to a health agency that provides various services related to home medical care. Health-care agents request services for their clients to the agency who in return finds and places a suitably qualified staff member to attend the client's needs. The agency has coordinators who attend the request for services, contact the suitable staff members and ask if they are available for the required dates, and finally enter jobs into the health agency system as required. The system records information about Clients, Agents, Staff, and Jobs (past and present). Currently, the coordinators manually match client details and requirements with the available staff's qualifications and preferences.

3 Data Mining

Data mining is the process of searching for trends and valuable anomalies in the entire data. The process involves (1) identifying a data mining problem and generating a subset of data according to the goals and interest of a data analyst, (2) pre-processing the data set to ensure a good quality by removing noise, handling missing information and transforming it to an appropriate format, (3) applying an appropriate data mining technique or a combination of techniques to the derived data set, and (4) evaluating and interpreting the discovered knowledge. There are various data mining techniques available with their suitability

dependent on the domain of application. We choose predictive modelling or classification data mining techniques to apply to the health care data. The goal of predictive modelling is to make predictions based on essential characteristics about the data. This is chosen as it offers the most precise description of why an event has taken place and its implementation has the greatest potential pay off by predicting the future cases.

3.1 Data Mining Process: Defining goals

Before we choose a data mining technique to uncover the interesting hidden data patterns from the data set, we must understand what the problem is and how we should approach it.

The health agency has identified two areas of interest that can be suitable for data mining. The first task is in relation to the manual allocation of suitable staff to jobs. Repeatedly the coordinators have to match the client's details (such as diagnosis, suburb, gender, age) with the staff's details (like suburb and qualifications). As there is already a history of who has done what job and where, we can use this existing knowledge to determine the staff member that is suitable for a job.

$f(x)$ = Staff Code where x is a subset of attributes from the database.

Another area of interest is in determining future markets based on past care. The agency would like to determine what kind of services they need to provide based on the client's age, gender, locality etc. After extracting this information from the data, the agency can then compare the results against population data from the Bureau of Statistics to determine the number of people that fit the defined categories within the areas that the agency covers. This may identify the need to expand into other areas that have a higher population of people in the determined categories.

$f(y)$ = Diagnosis where y is a subset of attributes from the database.

3.2 Data Mining Process: Data Pre-processing

The quality of data in a database cannot be guaranteed even with large corporate systems. Although data mining tools have some form of noise control, we still need to prepare the data for the mining process.

We started the process by extracting the raw data from the database in a de-normalised form where a single transaction (job) has all its information on a single tuple (record). Next, a statistical tool called MiniTab [4] is used to gain an understanding of the imported data and its quality. MiniTab identifies the number of different values for string type attributes and the count for each value. This includes a count for all unknown values where the data is either corrupt or missing. Non-character fields such as boolean can also be tallied and continuous number fields can be displayed in a boxplot.

The first step to ensure the quality of the data is to address the missing values. We can delete them, fill them in manually or use a tool to automatically identify a suitable value. Our decision is based on the quantity of missing data and whether the attribute of the missing data is a key field or not. As a small quantity of data is missing from some of the non-critical fields we can easily delete them. There is a problem with the *client age* field as this is seen as a critical field and needs to be filled in.

The tool Cubist [5] is used to predict the value of the client age based on the other fields. Cubist replaces the client age with the most probable value, which in this case is the mean of the known values for the attribute. The following is the output from Cubist on the data.

Replacing unknown attribute values: `CLIENT AGE' by 64.2 , ` STAFF SEX' by `FEMALE', `STAFF AGE' by 44.5

As Cubist provides one value to fit all missing values there is now a bias towards that value, especially when there is about 15% of client age fields with missing values. It is unfortunate that Cubist does not automatically fill in the missing values of the data set with values that are predicted from the known cases. This could be done if the data set had the tuples of the missing values removed and then a prediction model created on the data set with the results being used to populate the missing fields.

To aid in the completion of this paper it was decided that the tuples with missing values would be deleted. This also removes the bias of making all missing values the same.

Another pre-processing area to look at is corrupt data. Examples include:

- bad date formatting such as using '.' instead of '/' as a divider

- dates entered as 1/1/80 incorrectly converted to 1/1/2080 by system
- spelling errors of suburbs or abbreviating parts of, therefore ending up with two names for same suburb

Data transformation is considered, in the way of converting creation dates to a period that identifies how long clients, agents and staff have been with the agency as well as dates of b irths being converted to ages. A large number of records did not have a gender for the client. This could be determined from the first name of th e client but would be very time consuming to do. So as not to waste too much time, only records those were definitely useable had their genders input. We al so discount any attribute s that are redundant to the task at hand.

The initial raw data contained approximately 65000 records. This has been reduced to approximately 36000 records after pre-processing, which seems to be quite a large prune. It was expected that there would be a fair amount of data removal because of the current system not being a tru e relational database and that it does not support referential integrity thus creating many orphan records during archiving and deletion of records. There is also no documentation on the design of the system, which reflects on its poor structure.

Some o f the key attributes that were selected for the data mining process are Item Code (categorical), Creator Id (categorical), Agent Code (categorical), Agent suburb (categorical), Agent Years (Quantitative), Client Code (categorical), Client suburb (categoric al), Client Years (Quantitative), Client Diagnosis (categorical), Client Age (Quantitative), Staff Code (categorical), Staff suburb (categorical), Staff Years (Quantitative), Staff Sex (categorical), Staff Age (Quantitative).

3.3 Data Mining Process: Data Modelling

The predictive modelling or classification task of data mining builds a model to map (or classify) a data item into one of several predefined classes. Usually, the model is given some already known facts with correct answers, from which the mod el learns to make accurate predictions. Mainly three techniques namely neural induction, tree induction and bayesian classifiers are used for classification data mining tasks [2, 3].

Tree in duction classifiers or d ecision trees have been chosen because t hey fit the task at hand of classification. While constructing a decision tree from top to bottom, a ttributes are evaluated at each step to form descendant nodes. The attribute selection is based on a `statistical test' to determine how well it classifies the training examples. Classification of unknown samples is made by tracing a path through the decision tree until a leaf node holding the class prediction is reached. We use the modelling tool called C5.0 [6] that generates decision trees as well as classification rules.

3.4 Data Mining Process: Analysing the Results

The result is poor when all the staff and client related attributes are included in the mining process to det ermine the staff member suitable for a job. The resulting model focuses on the *staff suburb* , whic h is like stating the obvious. Of course a *staff suburb* will identify a staff member, especially if they are the only employee in that area. To improve the results we need to remove fields that prejudice the results and then redo the analysis.

Only the *staff suburb* is removed in next analysis, which provides us with a better result. It is noted that there is a huge increase in the mean size of the decision tree from 255 to 1165.9 and the mean error rate from 3.5% to 0.2%. However, as the standard error of the mean is less than 1 we can accept the result as being useful. The result now focuses on the *client suburb* allowing us to identify in most cases a staff member by this attribute. It was good to see that some of the sub -branches that were produced relied on the *client diagnosis* to determine which staff member to choose for a job in the selected suburb.

The *staff age* and *staff years* also play a part in the result and initial thoughts would suggest that perhaps they should be removed. The *staff years* could be important for identifying long -term employees that have a longer working knowledge of company procedures for care of a difficult client or the *staff age* may be important if specified as a preference by the client and are therefore kept in the analysis. The *staff sex* and *client sex* are also used in some rules which may also be useful if requested by the client.

It is difficult to report on all the confidence levels and statistical summaries in the rule sets because of the large number of rules. An overall view shows that most rules are well over the requested 50% confidence level, however some of the rules have slipped below and this is reflected in the number of training cases that did not fit the rule.

We also predicted the *client diagnosis*. We only use the client attributes because we want to compare them against statistical information put out by the Bureau of Statistics to identify future markets. The run shows a bias against the *client suburb*, which is doubted that it would determine a client's diagnosis. As a result the *client suburb* is removed for the next run.

Due to the low number of attributes that are represented here, one must be cynical about the results produced. For all folds, the first decision in the tree is *client age > 58* . With a mean size of 115.1 and only ages in the data set between 2 and 112 it can quickly be seen that almost each year has its own rule. There is also a high level of unknown *client* diagnosis, which is putting a strong bias on some of the rules. Further analysis of the results does not seem appropriate as the lack of attributes seems to be hampering the data mining process.

4 Conclusion and Future Directions

This paper outlines the data mining method followed to process the health care data. From the results, we can conclude that the extracted information and knowledge is limited to the quality of the source data and its suitability for the data mining techniques used.

It appears that the chosen data set does not have enough depth and therefore a lot of the results are based on a small number of occurrences. There is also a need for more attributes to be included in the analysis. Also, there are some attributes those need to be refined such as suburbs, work times, etc. Suburbs can be generalised to council zones or Messenger paper zones to reduce the number of different values for the field. This would also have more meaning in a spatial aspect as staffs are more likely to work in the surrounding area to where they live. At present the relationship between two suburbs cannot be determined. Employees' preference of work times was left out for simplicity. This is a multi-valued field, which requires generalisation to reduce the large combination of 35 fields (7 days x 5 time slots). We could have grouped the preferences into Mon-Fri 9-5, before this time, after this time and weekends. This would require extra analysis and possibly software to convert. Approximately 23000 of the lost records could have been saved by utilising the above mentioned techniques.

So far our studies have been based on the classification data mining task. There needs to be some analysis on the bigger picture which is 'What is a job made up of?', 'How does one job differ from another?' and 'How can we group similar jobs together?'. These questions are ideally suited to *clustering*, as there are no pre-existing classifications for a job. One may argue that the *item code* is the job classification but it probably better fits the description of the classification of the service provided and not the overall job which includes more details. Further experiments of this kind may provide us with new knowledge.

Nonetheless, analysis of health agency data set has provided us some important information that was unidentified otherwise. This will certainly help the health agency to make its decision better in the future with the incorporation of these results.

References

[1] Fayyad, U M, Piatetsky-Shapiro, G and Smyth, P (1995) 'From Data Mining to Knowledge Discovery: An Overview.' Advances in Knowledge Discovery and Data Mining. Fayyad, U M, Piatetsky-Shapiro, G, Smyth, P and Uthurusamy, R. Menlo Park, AAAI Press: 1 - 34.

[2] Han J, Kamber M 'Data Mining – Concepts and Techniques' Morgan Kaufman Publishers Inc. – 2000

[3] Lim T, Loh W 'A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms' Kluwer Academic Publishers, Boston – May 1995.

[4] http://www.minitab.com

[5] http://www.rulequest.com/cubist-info.html

[6] http://www.rulequest.com/see5-info.htm