Dependency Mining on the 2005-06 National Health and Nutrition Examination Survey Data

Jun won Lee Yao Huang Lin Matthew Smith
Advisor: Christophe Giraud-Carrier

Data Mining Lab, Brigham Young University, Provo, Utah

Abstract

The National Center for Health Statistics (NCHS) provides data and statistical information in an effort to improve public health. Data mining approaches provide practical methodologies to exploit meaningful information from raw data. Therefore, data mining tools may be used to derive implicit and useful hidden information from health data. In this paper, we carry out two experiments using decision tree and association rule modeling to report informative relations between attributes extracted from the 2005-2006 National Health and Nutrition Examination Survey data. The goal of this paper is to deliver informative relations and association rules that are not trivial but have potential in that our discoveries can provide valuable insights to clinical psychologists and people in medicine. According to our experimental results, we disclose several implicit relations such as the association between high blood pressure and hearing problem, as well as breathing problems and diabetes. We believe that these discoveries provide interesting insights that can explain the prevalence of major diseases and risk factors for diseases.

1 Introduction

Each year, the National Center for Health Statistics (NCHS) conducts the National Health and Nutrition Examination Survey (NHANES) to assess the health and nutritional status of adults and children in the United States. The 2005-2006 NHANES is a collection of data consisting of physical examinations and responses to health-related questions. Mining this large collection of data (over 10,000 respondents) presents opportunities for discovering new insights that could benefit public health.

In this paper, we limit our focus to the *laboratory* and *questionnaire* sets of the 2005-2006 NHANES data collection. The laboratory data is more quantitative in nature, while the questionnaire data is more qualitative. Combined together, these two sets provide a rich set that that plugs nicely into the data mining methodology presented in Section 2. The findings, presented in Section 3, aim to trigger further investigations into the behavior of discharged patients.

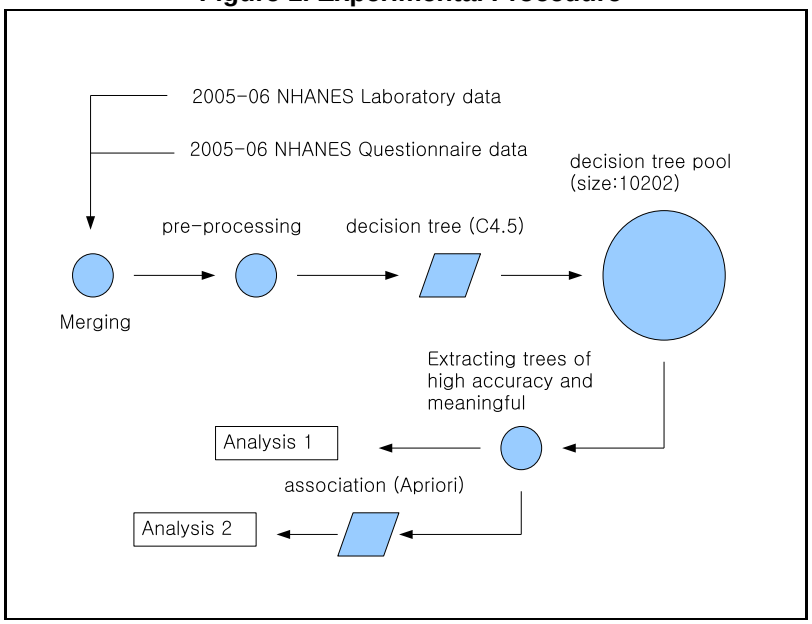
2 Experimental Setup

The process of collecting data, extracting and discovering meaningful and informative knowledge in a large database is called the knowledge discovery process. This process is largely comprised of three steps:

1. collecting and preprocessing data,
2. applying data mining methodologies, and
3. extracting informative results and performing analysis.

Figure 1 shows our instantiation of that process in the context of the present analysis. Details about each step are given below.

Figure 1. Experimental Procedure



2.1 Step 1: Collecting and Preprocessing Data

All NHANES data is available in SAS format. The data is comprised of 10,348 total data instances 10,202 nominal attributes, and 146 continuous attributes. During the preprocessing step, we merged laboratory and questionnaire data by the primary key (i.e., SEQN) representing each respondent. Both missing entries and responses with only a question mark (i.e., '?') in nominal attributes were treated as a distinct, perhaps artificial, response. These entries were replaced with the values "missing" and "question-mark," respectively. The data set had 865 attributes containing missing values. Lastly, we discretized non-nominal attributes to accommodate the association rule learner.

2.2 Step 2: Applying Data Mining Methodologies

Following the data preprocessing, we applied the following data mining methodology: (1) discover a set of attributes that relate and contribute to predict our pre-designated target attribute, and (2) identify the strongest associations between these attributes. To begin with, we built a series of decision trees (C4.5 algorithm) for each attribute as a target. Decision trees classify data by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. Each node in the tree specifies a test on some attribute of the data [2, 4]. The motivation behind this approach is based on how the decision tree structure identifies the set of attributes that contribute most to predict the target attribute. This implies a stronger associativity between attributes in the tree rather than those not included. Next, we ran an association rule miner (Apriori algorithm) with attributes that were carefully selected from trees we built in the first phase. Association rules provide information about how particular attributes group together and causal relations in the form of if-then statements [2].

2.3 Step 3: Extracting Informative Results and Performing Analysis

Using each attribute in turn as the target attribute, we used the other attributes to build decision trees predicting that target attribute (i.e., 10,212 trees in all). Among them, we eliminated trees that had under 80% predictive accuracy or over 95% predictive accuracy. This elimination process removes trees having low predictive accuracy, because they are not very predictable. In

the opposite case, we found that many of the trees that had high accuracy were very skewed towards the target attribute. For example, a decision tree for predicting the CDQ009A attribute, "whether one has pain in arms," was highly accurate, however a single value dominated 90% of the values. Therefore, trees making such predictions provide little information. Trees were also discarded if the target attribute itself was deemed not to be very interesting. Since we built a decision tree model for every attribute in an automatic manner, there were numerous trees that predict uninteresting values. For example, we eliminated the tree that predicted the SMD410 attribute, i.e., "does anyone smoke at home?" (regardless of its accuracy).

3 Experimental Results

3.1 Task 1: Predicting Attribute with the Decision Tree Learner C4.5

In this section, we report on a decision tree, obtained with the C4.5 algorithm, that exhibited high accuracy and was informative. The selected decision tree predicts the BPQ050A attribute, i.e., whether one is taking medicine for high blood pressure. Attributes in the derived tree provide useful information of what are most significant factors that contribute to high blood pressure. Figure 2 shows the summarized description of the decision tree structure and Table 1 explains the target attribute and list the most prevalent attributes in this tree.[1] As can be seen in Table 1, factors causing hearing problem seem most strongly related with high blood pressure since attribute PFD069H is in the root of decision tree. Also, ALQ120Q, ALQ101, and ALQ140Q are three other dominant attributes in that tree, which are all capturing information about alcohol consumption. Therefore, we suggest that components in alcohols can affect the level of blood pressure.

[1]Because of space restriction, we do not show the entire tree, only attributes that are dominant in the decision tree. The full results can be found at http:
dml.cs.byu.edu
amia

Figure 2. C4.5 Decision Tree for BPQ050A

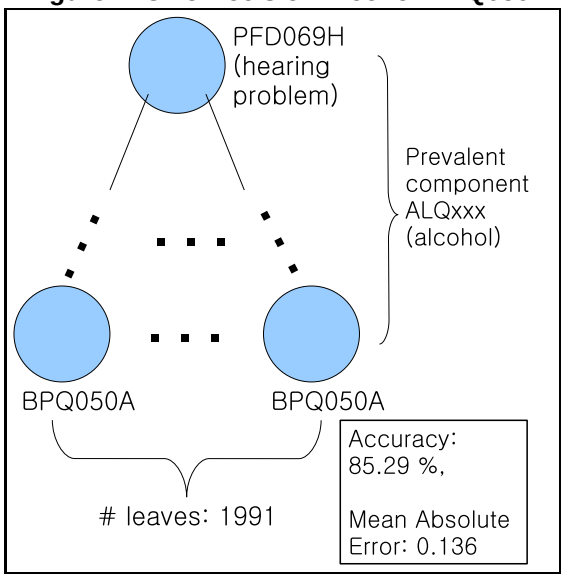


Table 1. Description for Figure 1

Target class	
BPQ050A	Taking medicine for ↑ blood pressure
Dominant attributes inside the tree	
PFD069H	Hearing problem (root node)
ALQ120Q	Frequency of alcohol over a year
ALQ101	Had $\geq$ 12 alcohol over a year
ALQ140U	# days have $\geq$ 5 alcohol over a year
PFQ063A	Health problem causing difficulty

3.2 Task 2: Learning Rules with the Association Rule Learner Apriori

In data mining, association rule learning is a popular and well researched method for discovering interesting relations among attributes in large databases [3]. It discovers hidden regularities among attributes in the form of if-then rules. The strength of generated rules is usually summarized by the two measurements *support* and *confidence*. The *support* of a particular association rule, $A \rightarrow B$, is the portion of data records that contain both A and B. The *confidence* of the association rule, $A \rightarrow B$ is a measure of the accuracy of the rule, as determined by the portion of data records that contain A, that also contain B. In

Table 2. Two Strongest Association Rules

precondition	consequent
HUQ050 = 3, MCQ053= No	DIQ050= No
support	*confidence*
0.90	0.99
PHQ020 = No, PHQ030 = No	PHQ050 = No
support	*confidence*
0.85	1.0

Table 3. Attributes Descriptions

Attributes	Brief Descriptions
HUQ050	♯ times receive health-care last year
MCQ053	Taking treatment for anemia
DIQ050	Taking insulin now?
PHQ020	Coffee or tea with cream or sugar?
PHQ030	Taking alcohol or liquor?
PHQ050	Taking antacids, or anti-diarrheals?
DIQ010	Diagnosed as diabetes
RDQ070	Wheezing or whistling in chest
SLQ190	Difficulty eating when tired
HOQ080	Water treatment devised used?
FSD170N	♯ of family can receive food stamp

other words,

$$support = P(A \wedge B) \text{ and } confidence = P(B|A)$$

Strong rules meet certain minimum support and confidence criteria. Apriori is a popular association rule learner, implemented in Weka software [5]. Given a set of itemsets (i.e., attributes), the algorithm attempts to find subsets where their value co-occur at least a minimum number C (the cutoff, or confidence threshold). Apriori uses a bottom-up approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found [3, 1].

To begin with, we seek association rules whose support is at least 0.45 (4,657 data records out of a total of 10,348) and confidence is at least 0.9, which demonstrates the strength of associativity between the precondition and consequent in the rules.

Table 2 describes the list of association rules with minimum support of 0.45 and Table 3 describes attributes appearing in the rules. The first rule says that one who receives three health-care treatments in the last year and is not taking treatment for anemia is not taking insulin. The high confidence of this rule implies that there is some potential associativities between anemia and some illnesses related with insulin even though there are usually unrelated. The second rule says that one who does not take coffee or tea with cream or sugar and does not take alcohol or liquor is not taking of antacids or anti-diarrheals. This rule should not be over-estimated but it suggests that avoiding drinks including alcohol or coffee may help avoid problems with stomach acidity and diarrhea.

Table 4. Diabetes Rules With Support $\geq$ 0.45

precondition	consequent
RDQ070= No, SLQ190= No	DIQ010= No
RDQ070= No, HOQ080= No	DIQ010= No
RDQ070= No	DIQ010= No
RDQ070= No, FSD170N= $\geq$ 7	DIQ010= No

Table 5. Extra Diabetes Rules With Support $\geq$ 0.30

precondition	consequent	confidence
RDQ070= No	DIQ010= No	0.94
DIQ010= No	**RDQ070= No**	0.87

Tables 4 and 5 demonstrate the rules related with diabetes. The minimum *support* and minimum *confidence* of Table 4 are 0.45 and 0.90, respectively. The minimum *support* and minimum *confidence* of Table 5 are 0.30 and 0.90, respectively. The reason we lowered the support value and obtained extra rules is that we wanted extra evidence to support our hypothesis that some breathing conditions are related with diabetes according to our association rules. In Table 4, we can see that the rules having DIQ010 (Diagnosed as diabetes) as a consequent have a common attribute in their preconditions, namely RDQ070 (Wheezing or whistling in chest). In particular, the third line stresses the direct association between RDQ070 and DIQ010. It indicates that one

who does not suffer from wheezing or whistling in chest would likely not suffer from diabetes. This is not a very interesting statement in itself. Unfortunately, we can not deduce the contrapositive rule (i.e., that one who has diabetes would suffer from wheezing or whistling in chest) since the confidence of the original rule is not 1.

To be able to collect more rules related with diabetes, we lower our support threshold to 0.3. This allows us to highlight the two rules shown in Table 5. The first rule is the same as the third rule in Table 4. Interestingly, the second rule is the converse of the first one, with confidence 0.87, which is lower than the first rule but still sufficiently high. Therefore, we conjecture that, from Table 4 and Table 5, there should be some connection between wheezing or whistling in chest and diabetes.

According to the National Library of Medicine, wheezing or whistling in chest is a breathing problem related with a number of lung diseases, such as asthma.[2] Furthermore, we found out that there are lung diseases that are related with (some form of) diabetes, such as sarcoidosis and pulmonary brosis.[3] These results seem to confirm the evidence gathered by our data analysis, that there may exist a link between wheezing or whistling in chest and diabetes.

4 Conclusion

Data mining is the process of sorting through large amounts of data and picking out relevant information. With the 2005-2006 NHANES data, decision tree and association rule learners were applied in order to obtain some valuable and insightful discoveries. Through the experimental process, we obtained several results.

First, there is a certain dependencies between a hearing problem and alcohol consumption with blood pressure. The accuracy of this is about 85.29%. Second, we obtained some associative rules; anemia with insulin, coffee and alcohol with antacid or anti-diarrheals, and most importantly, breathing issues with diabetes. The last one was confirmed with the results of studies in the medical literature. Note that these discoveries are derived from the 2005 and 2006 Laboratory and Questionnaire data. Therefore, it is obvious that our results are limited in terms of finite data size and finite data mining tools (decision tree and association rule learner). In spite of these limitations, our results provide an interesting insight that is worthwhile for further researches and investigations.

[2] http://www.nlm.nih.gov/medlineplus/ency/article/003070.htm

[3] http://archinte.ama-assn.org/cgi/content/summary/95/6/823

5 Future Work

In this paper, we selected the most dominant attributes in the tree as prominent ones that are strongly related with the target attribute. In other words, we provided all attributes with equal importance weight regardless of the position in the tree. As an alternative or better way, we could give different weight to attributes depending on the level of the tree where they are located, since attributes in a higher level have a higher impact than ones in a lower level. Furthermore, we could include examination data for further investigation, which can be included in this report since this data became available in the middle of the competition.

References

[1] R. Agrawal, T. Imielinski, and A.N. Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD*, pages 207–216, 1993.

[2] M. Berry and G. Linoff. *Data Mining Techniques*. Wiley Publishing, Inc, 2004.

[3] D.T. Lose. *Discovering Knowledge in Data: An Introduction to Data Mining*. John Wiley Sons, Inc, 2005.

[4] T. Mitchell. *Machine Learning*. McGraw-Hill, 1998.

[5] I.H. Witten and F. Eibe. *Data Mining: Practical Machine Learning Tools with Java Implementations*. Morgan Kaufmann, San Francisco, CA, 2000.