Using Dependency/Association Rules to Find Indications for Computerized Tomography In A Head Trauma Dataset

Susan P. Imberman Ph.D. , Bernard Domanski Ph.D.
College of Staten Island, City University of New York
2800 Victory Blvd.
Staten Island, NY 10314 USA
Office 718-982-2850
Fax718-982-2856
imberman@postbox.csi.cuny.edu
domanski@optonline.net

Hilary W. Thompson Ph.D.
Louisiana State University Health Sciences Center,
Director, Clinical Trials and Biometry Unit,
Louisiana State University Eye Center
2020 Gravier Street, Suite B
New Orleans, LA 70112-2234
Office 504-412-1350
FAX 504-412-1315
hthomp2@lsuhsc.edu

Abstract

Analysis of a head trauma dataset was aided by the use of a new, binary-based data mining technique which finds dependency/association rules. With initial guidance from a domain user or domain expert, Boolean Analyzer (BA) is given one or more metrics to partition the entire data set. The weighted rules are in the form of Boolean expressions. To augment the analysis of the rules produced, we applied a probabilistic interestingness measure (PIM) to order the generated rules based on event dependency, where events are combinations of primed and unprimed variables. Interpretation of the dependency rules generated on the clinical head trauma data resulted in a set of criteria that identified minor head trauma patients needing computerized tomography (CT) scans. The criteria found by BA was smaller (5 variables vs. 7 variables) than that found using recursive partitioning of chi-square values. BA's 5 variable criteria was more sensitive and less specific than the 7 variable criteria. We believe that BA has broad applicability in the medical domain, and hope that this paper will stimulate other creative applications of the technique.

Keywords: association rules, dependency rules, interestingness, intelligent data analysis, head trauma, computerized tomography

Artificial Intelligence in Medicine, Elsevier publishers
Volume 26, Issues 1-2 September - October 2002

Introduction

Knowledge discovery in databases had been defined as, " The non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." [6] One method for identifying these patterns is through association rule algorithms [1, 2]. Association rule algorithms find rules that show the associations between a dataset's variables.

Given a set of items I, an association rule is an implication $X \Rightarrow Y$ where X and Y are subsets of I and $X \cap Y = \phi$ [2]. Interesting association rules are identified using two metrics, support and confidence. Support is a measure of a rule's significance with respect to the database. It is the number of times $X \cup Y$ occurs with respect to the number of observations in the database. Confidence is a measure of the rule's strength. Confidence denotes the number of times X occurs with respect to the occurrence of $X \cup Y$. Agrawal et al. [1, 2] have defined interesting association rules as having support above a minimum, *minsup,* and confidence above a minimum *minconf.* Note, the Boolean operation used to connect attributes is conjunction. Attributes in the implication are only allowed a value of one, hence only inclusion of variables is considered, not exclusion.

Silverstein et al. argue [13] that the support/confidence measured association rules that Agrawal et al. [1, 2] described could not discover rules that described negative dependencies. For example, these algorithms would not discover rules such as "*People who buy diapers do not buy prune juice*." Therefore association rules were not adequate for domains that need to mine these types of rules. Silverstein et al. [13] differentiate between variable dependency and event dependency. A dataset will have a set of variables or attributes. An event is an instantiation of these variables. Each variable can have a value of zero or one. A value of zero indicates the absence of the variable, and a value of one indicates its presence in the event. This algorithm tests for dependency between variables, and then looks at instantiations of the events that are defined by the variables to see which events define strong rules. For example, given that I_1 and I_2 are two

variables from the set of attributes $I = I_1, I_2, ..., I_m$, then the events associated with these two variables are $I_1 I_2$, $I_1 I_2'$, $I_1' I_2$, $I_1' I_2'$.

Dependency rules were defined as an implication, $X \Rightarrow Y$ where X and Y are subsets of I, the attributes of set X are dependent to the attributes of set Y, and $X \cap Y = \phi$. Instantiations of X and Y, such that any attribute $I_x \in X$ and any attribute $I_y \in Y$ can have values of one or zero, denoting presence or absence in the observation. One can think of dependency rules as being a superset of the association rules.

Dependency rules better describe relationships between variables in the medical domain than association rules. Most times physicians are interested in not only positive associations, i.e. high LDL cholesterol = yes $\Rightarrow$ risk of heart disease = yes., but combinations of positive and negative variable associations, i.e. high HDL cholesterol = yes $\Rightarrow$ risk of heart disease = no. In addition to finding these dependency rules, physicians are interested in which rules are more or less significant.

In this paper we use an algorithm, *Boolean Analyzer* (BA), first defined by Orchard [9] and later extended by Domanski [5] and further by Imberman [7] to find indications for computerized tomography in a head trauma dataset. We show that dependency rules are effective in mining medical data. One of the reasons as to why our algorithm is suited for the medical domain is its use of variable partitioning which implicitly captures domain knowledge. By doing this we are able to essentially "query" the dataset for specific rule sets. In addition to finding dependency rules of the type defined above, BA uses a probabilistic interestingness measure (PIM) to order the rule set based on event dependency. Our approach encodes the original dataset into a matrix format that is at most as large as the original dataset, but usually much smaller.

This paper is organized as follows. Section 2 states the medical problem under investigation. Section 3 describes the data. BA is the data mining method used in this paper. It is discussed in section 4. Section 5 presents results from a clinical head trauma study where we have successfully used this algorithm to find

interesting and meaningful patterns. We discuss the algorithm and the results in section 6. Section 7 follows with our conclusions and directions for future research.

2. Statement of the Medical Problem

Haydel et al. [8] conducted an extensive clinical study to identify criteria for minor head trauma patients needing computerized tomography (CT) scans. Presently, all emergency room patients with minor head trauma receive CT scans. The goal of this study was to find a set of criteria that, when used, would reduce the number of patients receiving CT unnecessarily. Seven general criteria were identified using recursive partitioning of chi-square values. These included headache, nausea/vomiting, age over 60, drug or alcohol intoxication, deficiency in short-term memory (confusion), trauma above clavicles, and post-traumatic seizure. According to the data collected, when using these criteria to guide the use of CT scanning, unnecessary testing would have been reduced by 23%. In addition, Haydel et al. [8] contended that the seven criteria were 100% sensitive, i.e. there were no positive CT missed. Our goal was to use a data mining method that finds the dependency and association rules to identify a "better" criteria set.

Our new criteria set, in order to be preferred over the seven variable set, needed to be 100% sensitive with regard to positive CT. It also needed to reduce unnecessary CT. The new criteria should have fewer variables making it less cumbersome of an indicator. The criteria set needed to be general, in that it would be able to be accurate over "new" unseen data. Hence it should not overfit the dataset.

3. Head Trauma Data

The data was the same data that was collected by Haydel et al [8]. It contained ten variables. In addition to the seven criteria, the dataset included variables for sex (male/female), result (indicating a positive or negative CT) and for a history of blood coagulation problems. Some records had missing values for sex. Since our data mining algorithm cannot handle missing values we removed these records from

consideration. In addition, the number of positive CTs for blood coagulation problems (only one record) was not significant enough to be used in the analysis. Therefore the variable for blood coagulation problems was removed from consideration. The resulting dataset had 1,339 records representing that number of patients.

4. Data Mining for Dependency Rules

For our data mining method, we used the BA algorithm. This algorithm finds dependency rules. BA uses a probabilistic interestingness measure (PIM) to order the rule set generated by the algorithm, based on event dependency. The algorithm has been used extensively in the domain of computer performance evaluation. Recently, we have been applying BA to clinical medical data. BA has performed well in both domains. Interpretation of the dependency rules generated on clinical head trauma data resulted in a set of criteria that identified minor head trauma patients needing computerized tomography (CT) scans. The criteria found by BA was smaller (5 variables vs. 7 variables) than the original set derived from the same data, published in The New England Journal of Medicine [8].

BA takes Boolean data, and calculates the PIM of the rule, indicating the strength of events in a dependency relationship. Positive and negative values of this measure indicate if the relationship is direct or indirect, respectively. We summarize the algorithm as follows:

1. Create the State Occurrence Matrix (SO) from the dataset.
2. Derive the State Linkage Matrix (SL) from the SO
3. For the variable partition defined by the SL, generate all dependency rules that have at least minimum support, minimum confidence and minimum PIM.

	X_1	X_2	X_3	X_4	X_5	X_6
observation 1	0	1	1	1	0	1
observation 2	1	0	1	1	1	1
observation 3	1	1	1	0	1	0
•					•	
•					•	
•					•	

Figure 1 - Boolean Activity Matrix

4.1 The State Occurrence Matrix

BA views each row of data as an observation in a Boolean activity matrix. Each 0/1 vector represents a single entry in this matrix as can be seen in Figure 1. We partition the variables into two sets X and Y where $X \cap Y$ is $\varnothing$ and $X \cup Y$ is the set of all variables in the dataset. From this we calculate the SO where each entry is the support of $X \Rightarrow Y$, where X is a set of row variables, and Y is a set of column variables. An example of an SO for a six variable, 100 observation dataset is shown in Table 1.

	$X_4X_5X_6$	$X_4X_5X_6'$	$X_4X_5'X_6$	$X_4X_5'X_6'$	$X_4'X_5X_6$	$X_4'X_5X_6'$	$X_4'X_5'X_6$	$X_4'X_5'X_6'$
$X_1X_2X_3$	0	1	0	0	4	2	1	2
$X_1X_2X_3'$	2	0	1	0	3	0	2	2
$X_1X_2'X_3$	1	0	0	1	5	1	0	2
$X_1X_2'X_3'$	3	0	0	0	3	5	2	2
$X_1'X_2X_3$	1	0	1	1	3	1	2	1
$X_1'X_2X_3'$	0	2	0	1	0	3	5	4
$X_1'X_2'X_3$	0	0	0	1	5	3	4	2
$X_1'X_2'X_3'$	0	1	2	1	0	5	4	2

Table I. - State Occurrence Matrix

Primed variables indicate a negative dependency, unprimed variables a positive one. Using the SO, we can observe relationships between the variables. Assume ":" stands for "related to". If we wanted to view the relationship $X_1 : X_4$, we look at how many times X_1 and X_4 are both high, X_1 is high when X_4 is low, X_1 is low when X_4 is high, and both X_1 and X_4 are low. Using the SO above, we can form the following contingency table:

	X_4	X_4'
X_1	9	36
X_1'	11	44

The sum of the entries in the shaded area of the SO gives us the upper left hand entry in the contingency table for "X_1 and X_4 are both high". Similarly, we determine the other 3 values in the contingency table by summing the corresponding quadrants in the SO. You can also form contingency tables for more complex relationships such as $X_1X_3 : X_4X_6$.

4.2 Defining the Probabilistic Interestingness Measure

Assume we are given a 2 x 2 contingency table as described above, where X and X' stand for a group of rows and their complement respectively, and Y and Y' stand for a group of columns and their complement respectively. Assume that X and Y are independent.

Let the probability of event X be p(X). Therefore the probability of X' is 1 - p(X).

Let the probability of event Y be p(Y). Therefore the probability of Y' is 1 - p(Y).

Since X and Y are assumed to be independent, then the number of times X occurs with respect to Y is:

$p(X \wedge Y) = p(X)p(Y)$. We also have:

$$p(X \wedge Y') = p(X)\,(1- p(Y))$$

$$p(X' \wedge Y) = (1 - p(X))\, p(Y)$$

$$p(X' \wedge Y') = (1- p(X))\,(1 - p(Y))$$

Representing the above in a contingency table similar to the one discussed previously we get:

	Y	**Y'**
X	$p(X)p(Y)$	$p(X)\,(1- p(Y))$
X'	$(1 - p(X))\, p(Y)$	$(1- p(X))\,(1 - p(Y))$

We see that the column ratio of the number of times X occurs with respect to Y to the number of times X' occurs with respect to Y, is equal to the ratio of the number of times X occurs with respect to Y' to the number of times X' occurs with Y'.

$$\frac{p(X)p(Y)}{(1 - p(X))\, p(Y)} = \frac{p(X)\,(1 - p(Y))}{(1 - p(X))\,(1 - p(Y))} \qquad (1)$$

We find the same if we look at the row ratio.

Let a, b, c and d represent the values in each quadrant of the above contingency table. The values for a, b, c, and d are the occurrences of the joint events, which can be read from the SO.

	Y	**Y'**
X	a	b
X'	c	d

Then we have:

$$\frac{a}{c} = \frac{b}{d} \tag{2}$$

Therefore, when a, b, c and d are independent we find:

$$0 = ad - bc \tag{3}$$

Define events that are not independent as being dependent. Then using the above equation we define the PIM as having a value *m* where

$$m = ad - bc \tag{4}$$

The PIM is a measure on the type and extent of the dependency of the event *X* to *Y*. Large negative values show that *X* has a strong inverse dependency relationship to *Y*. Large positive values indicate a strong direct dependency relationship. Values close to or equal to zero indicate that the variables are independent and not related. For example, the measures of $X_1 : X_4$ is m =(9)(44) - (11) (36) = 0 (from the contingency table in section 2.1), and $X_1X_3 : X_4X_6$ is m = -120.

4.3 The State Linkage Matrix

Generalizing from the SO, the relationship matrix defined by a single row i and a single column j would be:

	column j	**(column j)'**
row i	a_{ij}	$r_i - a_{ij}$
(row i)'	$c_j - a_{ij}$	$N - r_i - c_j + a_{ij}$

where:

- a_{ij}= entry at row I, column j of the SO matrix
- N = total sample size of the activity matrix (the dataset)
- r_i = sum of the entries in row i of the SO matrix
- c_j = sum of the entries in column j of the SO matrix

Taking the PIM and combining terms we get:

$$m_{ij} = a_{ij} N - r_i c_j \qquad (5)$$

We can use the above PIM to derive a new matrix called the SL whose entries are the PIM for the simple dependency relationship of a single row with a single column. We show this in Table II.

	$X_4X_5X_6$	$X_4X_5X_6'$	$X_4X_5'X_6$	$X_4X_5'X_6'$	$X_4'X_5X_6$	$X_4'X_5X_6'$	$X_4'X_5'X_6$	$X_4'X_5'X_6'$
$X_1X_2X_3$	- 70	60	- 40	- 50	170	0	- 100	30
$X_1X_2X_3'$	130	- 40	60	- 50	70	- 200	0	30
$X_1X_2'X_3$	30	- 40	- 40	50	270	- 100	- 200	30
$X_1X_2'X_3'$	195	- 60	-60	- 75	- 45	200	-100	- 55
$X_1'X_2X_3$	30	- 40	60	50	70	- 100	0	- 70
$X_1'X_2X_3'$	- 105	140	- 60	25	- 345	0	200	145
$X_1'X_2'X_3$	-105	-60	- 60	25	155	0	100	- 55
$X_1'X_2'X_3'$	-105	40	140	25	- 345	200	100	- 5

Table II - State Linkage Matrix

From the SL we can determine the PIMs of more complex relationships. To calculate the PIM for a relationship X vs. Y, where X represents a set of rows and Y a set of columns, sum the PIMs of the entries that correspond to those rows and columns combined. This can be seen in the shaded portion of Table II. The ability to sum the PIMs can be proved by induction. We leave this proof out of this paper because of space constraints.

4.4 - How to Use The State Linkage Matrix

We can use the SL to find the highest PIM for any one variable such as X_5 and the row states. To do this we calculate m(row : X_5). From the SL we find that $X_1X_2'X_3'$: X_5 has the highest PIM, 290, for this relationship as shown in Table III.

$PIM(X_1X_2X_3 : X_5) = -70 + 60 + 170 + 0 = 160$
$PIM(X_1X_2X_3' : X_5) = 130 - 40 + 70 – 200 = -40$
$PIM(X_1X_2'X_3 : X_5) = 30 - 40 + 270 - 100 = 160$
$PIM(X_1X_2'X_3' : X_5) = 195 - 60 - 45 + 200 = 290$
$PIM(X_1'X_2X_3 : X_5) = 30 - 40 +70 - 100 = -40$
$PIM(X_1'X_2X_3' : X_5) = -105 + 140 - 345 + 0 = -310$
$PIM(X_1'X_2'X_3 : X_5) = -105 - 60 + 155 + 0 = -10$
$PIM(X_1'X_2'X_3' : X_5) = -105 + 40 - 345 + 200 = -210$

Table III - PIM values for row : X_5

If we combine with disjunction the relationships that have positive PIMs in the Table III, we form a composite state $X_1X_2X_3$ v $X_1X_2'X_3$ v $X_1X_2'X_3'$ with a PIM of 610. The PIM for this composite state is found by summing the measures of each component dependency relationship,

$$PIM(X_1X_2X_3 : X_5) + PIM(X_1X_2'X_3 : X_5) + PIM(X_1X_2'X_3' : X_5) = 160 + 160 + 290 = 610 \quad (6)$$

$X_1X_2X_3$ v $X_1X_2'X_3$ v $X_1X_2'X_3'$ is equivalent to $(X_1 (X_2' \text{ v } X_3))$. Therefore, the PIM of the more complex dependency relationship $(X_1 (X_2' \text{ v } X_3)) : X_5$ is 610. Taking a look at the values for this relationship from the SO we get:

	X_5	X_5'
$X_1 (X_2' \text{ v } X_3)$	25	10
$(X_1 (X_2' \text{ v } X_3))'$	29	36

It is more complicated to derive this contingency table from the SO. The SL gives a more direct way for finding complex relationships. Hence, by combining terms, and summing their PIMs, we can find disjunctive relationships as well as conjunctive ones. Often disjunctive relationships are interesting and useful because they can express relationships more succinctly. We have used this property to formulate a hill climbing phase that finds a disjunctive rule for the partition, that has a higher PIM than the conjunctive rules in the partition. Details can be found both in Domanski[5] and Imberman[7].

4.5 Rule Generation

We see that the PIM of a relationship is bi-directional because the PIM of X : Y is equal to the PIM of Y : X. We modify Silverstein et al.'s [13] definition of dependency rules to be the probabilistic implication $X \Rightarrow Y$ where X is some subset of row or column attributes and if X is a proper subset of row attributes then Y is a proper subset of column attributes. Also, if X is a proper subset of column attributes then Y is a proper subset of row attributes.

According to this definition, each relationship matrix represents two possible dependency rules, $X \Rightarrow Y$ and, $Y \Rightarrow X$. Obviously, $X \Rightarrow Y$ is not the same as $Y \Rightarrow X$. For example, the strength of the statement, "*People who buy dolls are likely to buy candy bar brand X.*", tells us nothing about people who buy candy

bar brand X and their affinity for dolls. We use the confidence metric defined previously to determine the strength of the rules. Those rules that have confidence above a minimum threshold are strong rules.

We also have to be mindful of the support for the generated rule. Rules that have the same PIM might have different levels of support in the dataset. Define an interesting dependency rule as one that has support, *minusp,* above a threshold *s*, confidence, *minconf,* above a threshold *c*, and a PIM, *minPIM,* above a threshold *m*.

5 Results On Clinical Head Trauma Data

BA was run on the cleaned dataset using a partition of result vs. sex, vomiting, age over 60, drug or alcohol intoxication, deficiency in short term memory (confusion), trauma above clavicles (trauma), and post-traumatic seizure (PTS). The generated rules were ordered by their PIMs. Rules that showed dependencies for a positive CT were examined. The highest PIM rule showed a dependency with positive drug or alcohol intoxication and negative confusion. Looking at the data, all but one patient having a positive drug or alcohol intoxication and a negative confusion value had positive CT results. This showed that the PIM was a good measure for finding significant dependencies. However, the goal of this study was to find criteria that had 100% sensitivity for a positive CT result. Further investigation of the top ten rules (Table IV) showed that positive drug or alcohol intoxication and negative confusion were also associated with age below 60, no vomiting, and being male. At this point we reduced the original seven criteria to five. The new criteria was 100% sensitive for a positive CT. Unfortunately, this criteria did not reduce unnecessary CT.

Positive Result	⇒	Positive Alcohol/Drug Negative Confusion
Positive Result	⇒	Positive Alcohol/Drug
Positive Result	⇒	Positive Alcohol/Drug Negative Age60 Negative Confusion
Positive Result	⇒	Positive Alcohol/Drug Negative Age60
Positive Result	⇒	Positive Sex Positive Alcohol/Drug Negative Confusion
Positive Result	⇒	Positive Sex Positive Alcohol/Drug
Positive Result	⇒	Negative Nausea/Vomit Positive Alcohol/Drug Negative Confusion
Positive Result	⇒	Negative Nausea/Vomit Positive Alcohol/Drug
Positive Result	⇒	Negative PTS Positive Alcohol/Drug Negative Age60 Negative Confusion
Positive Result	⇒	Negative PTS Positive Alcohol/Drug Negative Age60

Table IV - Top Ten Rules For Positive Result

We looked at rules with dependencies for negative CT and found dependencies with low headache, low post-traumatic seizure, and low trauma. Combining these dependencies and the ones found with positive CT we deduced a disjunctive set of five criteria that included positive drug or alcohol intoxication, male sex, positive headache, positive post-traumatic seizures, and positive trauma. This criteria were 100% sensitive for CT and reduced unnecessary CT by 11%

In order to compare BA to Haydel et al. [8], we needed to recalculate Haydel's performance on the cleaned data. Haydel's seven criteria, in addition to classifying all positive CT, was able to reduce unnecessary CT by 28%. BA's performance for lowering unnecessary CT was not as impressive as Haydel's. Notwithstanding, estimates indicate that even a 10% reduction in CT scans can result in savings of millions of health care dollars [10].

Since the size of the dataset was not very large, and contained a relatively low number of positive CT (93 out of 1429 in the original dataset), there was a good chance for overfit. This meant these sets of criteria might not perform well on unseen data. A valid method for comparing BA's five criteria to Haydel's seven was to analyze them using a discriminant function derived from discriminant analysis of the data. This type of analysis takes into account the problem of overfit. Sensitivity and specificity values based on a discriminant function, statistically, provide a realistic and practical comparison between criteria of this type.

Discriminant analysis is a multivariate statistical method for classifying or separating distinct sets of observations. The goal is to find a set of multipliers or weights that, when multiplied by the values of the variables in an observation, result in a discriminant function that enables the calculation of a discriminant score. This score can then be used to determine whether an observation lies in one group or another. The groups to be discriminated between here are the need for a head CT or not, as indicated by the positive head CT. For a conservative estimate of the performance of the discriminant function, we report the sensitivities and specificities based on classification of a 25% hold out data subset, or test data. The test data was randomly chosen from the cleaned data set and not used in discriminant function creation. These estimates are more likely to describe how a rule would perform if it were applied to other datasets not used in the computation of the discriminant function.

A discriminant function was derived for the seven Haydel criteria (the 25% random test data, as described above, was used for all evaluations that follow). The results for the seven criteria were, sensitivity= 42%, specificity= 79%. Specificity measures are defined as the proportion of true negatives = (number of true negatives) / (number of true negatives + number of false positives). Whereas sensitivity measures are defined as the proportion of true positives = (number of true positives / (number of true positives + number of false negatives). The results of the same analysis on the criteria identified by BA was sensitivity= 79%, specificity= 58%. Specificity and sensitivity results from the discriminant function are summarized in Table V.

	Alc/ Drug	Nausea/ Vomiting	Confusion	Trauma	PTS	Age>60	headache	Sex	Sensi-tivity	Specifi -city	Reduce CT
Haydel	X	X	X	X	X	X	X		42	79	28%
BA	X			X	X		X	male	79	58	11%
Discrimin-ant Analysis	X	X		X		X	X	male	71	61	9%
CARTI	X					X	X	male	75	58	13%
CARTII	X			X		X	X	male	79	58	10%

Table V- Summary of Head Trauma

Besides the chi-square partitioning used by Haydel, and BA's methodology, another statistical method that could identify criteria that would be less likely to overfit is stepwise discriminant analysis. This method is useful in selecting variables that may be effective in creating discriminant functions for distinguishing between classes of observations. It works by choosing variables that maximize the squared partial correlation of each variable and the class variable (positive CT), controlling for the effects of the variables already selected for inclusion in the model. We applied a stepwise discriminant analysis on the entire dataset. This method chose 6 criteria: trauma, drug or alcohol intoxication, vomiting, headache, age over 60, and sex. When these 6 criteria were used on the same random 25% test data subset as described above, the test data gave sensitivity = 71%, specificity = 61%. The 6 criteria only reduced unnecessary CT by 9% (Table IV).

Additional experiments were performed using Salford Systems CART™. The CART decision tree identified four criteria needed to classify positive CT. These criteria were drug or alcohol intoxication, headache, age over 60, and male sex. When applied to the dataset, two positive CT's were missed by this criteria. Inspection of the two records showed that trauma was the only criteria that both records were positive for. We thus included trauma into the criteria set. The resulting criteria was 100% sensitive for positive CT and reduced unnecessary CT by 10%. Sensitivity and Specificity measures based on the discriminant function were 79% and 58%, respectively. CART and BA produced similar results. Sensitivity and specificity measures were equivalent with BA performing slightly better in reducing unnecessary CT. Table V summarizes all these results.

If sensitivity is our major consideration, then the rule set produced by BA is superior or equivalent to the other methods when evaluated with a discriminant function. Also, the specificity of the discriminant function produced by BA selected variables is only slightly less (58%) than that produced by stepwise discriminant analysis (61%). By this discriminant function criteria then, the criteria chosen by the BA method

are equal or superior in sensitivity and specificity to criteria chosen by stepwise discriminant analysis, and superior to the criteria chosen by recursive partitioning of chi-square values.

6. Discussion

Most association rule programs such as Apriori[1, 2] and dependency rule algorithms[13] use support to reduce the search space. In BA the partition step explicitly uses the domain knowledge of an expert to split the set of variables into column variables and row variables. By doing this we reduce our search space considerably. To get a sense of how reduced our search space becomes, for a dataset of 6 variables, the number of possible variable partitions is equal to C(6,3) + C(6,2) + C(6,1) = 41, where C(n, k) is the combination of n items taken k at a time. This represents the creation of 3 X 3, 2 X 4, and 5 X 1 partitions respectively. By recognizing that this dataset has one significant partition of interest, we have reduced our search space by a factor of forty.

Although missing values did not pose a significant problem in this dataset, medical data is known to suffer from the missing value problem. BA does not directly address the handling of missing values. Missing values must be handled as a preprocessing step before implementation of the algorithm. The method for handling missing values depends on the nature of the missing values. If data is randomly missing, then there are effective methods for estimating these values. One can use methods that employ maximum likelihood estimates such as the EM algorithm described by Dempster, Laird, and Rubin [4]. One can also substitute a mean or median value for the missing data. Substituting mean or median works best when the missing data is truly random, otherwise biased estimates can occur. If the missing values are associated with some property of the observations, as when one gender will not give certain information, or subjects are less willing to describe certain disease symptoms, then there are no methods that deal with such correlated missing values effectively.

The number of positive CT in the cleaned head trauma data was 88 out of 1,339 records. For medical data, this skew is not unusual. Most people are negative for a disease or a set of symptoms. Using support-based algorithms on data with this skew would require the user to set *minsup* at a near zero value. The closer *minsup* is to zero, the larger the search space is. When *minsup* equals zero, we search the entire space. Variable partitioning, as implemented in BA, allows one to find interesting patterns in skewed data, while pruning the search space.

Although variable partitioning works well with datasets where there is a finite, small number of partitions, in datasets where there is no delineation, it is necessary to look at all partitions. This can be a computationally expensive procedure, and is in fact NP-complete.

In this paper we compared sensitivity and specificity measures between methods based on one representative test and train set where the "new" data tested was a 25% holdout set. The test and train paradigm dictates that this should be done multiple times. Sensitivity and specificity comparisons between BA and Haydel were repeated ten times with 10 different randomly generated holdout sets. In 5 trials out of 10, BA showed higher sensitivity than Haydel, 3 trials out of 10 showed lower sensitivity and in two trials both methods were equally sensitive. In 9 out of 10 trials BA was less specific than Haydel. Therefore, over multiple test and train sessions, BA was superior to Haydel with respect to both sensitivity and specificity.

BA makes only one pass through the dataset in order to create the SO. For this paper the SO is shown fully expanded. In reality it is a sparse vector that is at most as large as the dataset, but most times smaller than the dataset. The SL tends to have more entries then the SO and is therefore more bound by the dimensionality of the dataset. Therefore in BA there is a tradeoff between size and dimensionality. If the number of variables is n, then BA is most efficient when the number of records in the dataset is larger than 2^n.

The PIM is an objective interestingness measure that imposes an order on the set of rules. Objective interestingness measures are calculated based on a rule's structure and the underlying data used [12]. The

PIM associated with each rule is contiguous within the interval - $(n/2)^2 < 0 < (n/2)^2$, where n is the number of observations in the dataset. Since the PIM is dependent on the distribution of the underlying data, we cannot generalize about the form of this contiguous distribution. Therefore the types of statistics used with the PIM are usually non-parametric.

Dependency/association rule algorithms tend to generate a lot of rules. Even though BA looks at one data partition, for the head trauma data, BA generated 13,120 rules which is orders of magnitude larger that the dataset! The PIM rule order allowed us to find a significant criteria set by only examining 60 of these rules.

Other methods for ranking rules have been cited in the literature. These include metrics such as confidence, added value, mutual information, [11] and conviction measures [3]. Sahar and Mansour [11] showed that objective interestingness measures seem to cluster into three groups when support and confidence levels are low. Interestingness measures in the same cluster produce similar rule orders. For future work, to see how the PIM compares to these interestingness measures, an empirical study similar to that done by Sahar and Mansour is indicated.

Association rules do well in many domains. There are some domains, such as the medical domain, where dependency rules are a better choice. As evidenced in this paper, the most interesting information was gleaned from those rules that could not have been found using traditional association rule algorithms. In these domains, the probabilistic interestingness measure was significant for finding interesting dependency rules. The PIM was able to yield good results in a clinical medical study.

7. Conclusions and Future Work

Boolean Analyzer like all algorithms of this genre is an exponential algorithm. It gains efficiency by pruning the search space by using domain knowledge to partition the variable set into row variables and column variables. BA uses a probabilistic interestingness measure to order the rules according to the event

dependencies between the variables. We have shown that in a clinical head trauma dataset, the PIM was able to find significant rules allowing us to find a five variable criteria set for identifying patients needing CT.

If there is no good way to partition the variables, all partitions need to be searched. For the future we want to identify a heuristic for finding partitions that might yield high PIM dependent relationships. In addition, we intend to apply BA to a clinical ophthalmology dataset. This dataset has 10,000 records and 40 variables. We will be using BA and other techniques to find risk factors for surgery.

Acknowledgements

We wish to thank Miriam Tausner for her helpful comments and criticisms during the writing of this paper. We would also like to acknowledge Robert Orchard for his pioneering work on the BA algorithm.

References

1. Agrawal R, Mannila,S., Ramakrishnan H, Toivonen H, and Verkamo AI. Fast Discovery of Association Rules. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P. and Ramasamy U, eds. *Advances In Knowledge Discovery and Data Mining*. AAAI/MIT Press, Cambridge Mass, 1996; pp 307 - 328.

2. Agrawal R, Imielinsk T, and Swami A.. Mining Association Rules between Sets of Items in Large Databases. *Proceedings of the ACM SIGMOD International Conference on the Management of Data,* 1993: pp 207-216

3. Brin S, Motwani R, Ullman J., and Tsur S. Dynamic Itemset Counting and Implication Rules for Market Basket Data. *Proceedings of the ACM SIGMOD International Conference on Management of Data.* 1997 pp 255 - 264.

4. Dempster, A.P., Laird, N.M., and Rubin, D.B. Maximum Likelihood From Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 1977 (B) 39(1) 1-38.

5. Domanski B. Discovering the Relationships Between Metrics. *The Proceedings of the 1996 Computer Measurement Group.* December 1996, San Diego California, 1996; pp 309 – 313.

6. Fayyad UM, Piaetsky-Shapiro G, and Smyth P. From Data Mining to Knowledge Discovery: An Overview. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P. and Ramasamy U, eds. *Advances In Knowledge Discovery and Data Mining*. AAAI/MIT Press, Cambridge Mass, 1996; pp 1 - 34.

7. Imberman S. Comparative Statistical Analyses Of Automated Booleanization Methods For Data Mining Programs (Doctoral dissertation, City University of New York, 1999). UMI Microform, 9924820.

8. Haydel MJ, Preston CA, Mills TJ, Luber S, Blaudeau E, DeBleiux PMC. Indications for Computed Tomography in Patients with Minor Head Injury. *The New England Journal of Medicine,* 2000 343(2): pp 100-105.

9. Orchard RA., 1975. On the Determination of Relationships Between Computer System State Variables. *Bell Laboratories Technical Memorandum,* January 15, 1975

10. Reinus WR, Wippold FJ, Erickson KK, Practical Selection Criteria For Noncontrast Cranial Computed Tomography In Patients With Head Trauma, *Annals of Emergency Medicine,* 1993 (5): pp 134-40.

11. Sahar S and Mansour Y. An Emirical Evaluation of Objective Interestingness Criteria. *SPIE Conference on Data Mining and Knowledge Discovery,* 1999 pp 63 - 74.

12. Siberschatz A, Tuzhilin A. What Makes Patterns Interesting Knowledge Discovery Systems. *IEEE Transactions on Knowledge Discovery and Data Engineering,* 1996 8(6):pp 970 - 974.

13. Silverstein C, Brin S, and Motwani R. Beyond Market Baskets: Generalizing Association Rules to Dependence Rules. *Data Mining and Knowledge Discovery,* 1998 2(1): pp 39-68.