Data Mining to Improve Traffic Safety

By

Randy K. Smith and Huanjing Wang

Department of Computer Science,
The University of Alabama,
Box 870290,
Tuscaloosa, AL 35487-0290

Prepared by

UTCA

University Transportation Center for Alabama
The University of Alabama, The University of Alabama in Birmingham,
and The University of Alabama at Huntsville

UTCA Report 04107
May 26, 2005

Technical Report Documentation Page

1. Report No	**2. Government Accession No.**	**3. Recipient Catalog No.**

4. Title and Subtitle Data Mining to Improve Traffic Safety	**5. Report Date** May 26, 2005
	6. Performing Organization Code
7. Authors Dr. Randy K. Smith and Ms. Huanjing Wang	**8. Performing Organization Report No.** UTCA 04107
9. Performing Organization Name and Address University Transportation Center for Alabama P O Box 870205 University of Alabama Tuscaloosa, AL 35487-0205	**10. Work Unit No.**
	11. Contract or Grant No. DTSR0023424
12. Sponsoring Agency Name and Address University Transportation Center for Alabama P O Box 870205 University of Alabama Tuscaloosa, AL 35487-0205	**13. Type of Report and Period Covered** Final Report: 01/01/2004 – 12/31/2004
	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

The ever increasing size of datasets used for data mining and machine learning applications has placed a renewed emphasis on algorithm performance and processing strategies. This research addresses algorithms for ranking variables in a dataset, as well as for ranking values of a specific variable. We propose two new techniques, called Max Gain (MG) and Sum Max Gain Ratio (SMGR), which are well-correlated with existing techniques, yet are much more intuitive. MG and SMGR were developed for the public safety domain using categorical traffic accident data. Unlike the typical abstract statistical techniques for ranking variables and values, the proposed techniques can be motivated as useful intuitive metrics for non-statistician practitioners in a particular domain. Additionally, the developed techniques are generally more efficient than the more traditional statistical approaches.

17. Key Words Data mining, traffic safety	**18. Distribution Statement**

19. Security Class (of report)	**20. Security Class**. (Of page)	**21. No of Pages**	**22. Price**

Contents

List of Tables

List of Figures

Executive Summary

The ever increasing size of datasets used for data mining and machine learning applications has placed a renewed emphasis on algorithm performance and processing strategies. This research addresses algorithms for ranking variables in a dataset, as well as for ranking values of a specific variable. We propose two new techniques, called Max Gain (MG) and Sum Max Gain Ratio (SMGR), which are well-correlated with existing techniques, yet are much more intuitive. MG and SMGR were developed for the public safety domain using categorical traffic accident data. Unlike the typical abstract statistical techniques for ranking variables and values, the proposed techniques can be motivated as useful intuitive metrics for non-statistician practitioners in a particular domain. Additionally, the developed techniques are generally more efficient than the more traditional statistical approaches

Section 1
Introduction

Motivation

Data mining is the exploration and analysis of a large dataset in order to discover knowledge and rules. Data mining has been very successful as a technique for deriving new information in a variety of domains (Berry & Linoff, 1997). Data mining is typically conceptualized as a three-part process: preprocessing, learning (or training) and post-processing.

In the last decade, data that serves as a target for data mining has grown explosively. Data has been growing increasingly larger in both the number of rows (i.e., records) and columns (i.e., variables). The quality of data affects the success of data mining on a given learning task. If information is irrelevant or noisy, then knowledge discovery during training time can be ineffective (Hall & Smith, 1999). Variable selection is a process of keeping only useful variables and removing irrelevant and noisy information. It is always used as a data mining preprocessing step, particularly for high-dimensional data.

Variable selection can be used to select subsets of variables in terms of predictive power. Since variable selection effectively ranks set of variables according to importance, it may be referred to as *variable ranking*. There is also an analogous notion of *value ranking*, which refers to the idea of ranking the values of a particular variable in terms of their relative importance. In this research, we examine existing techniques for both variable and value ranking, and we propose new techniques in both categories. Our techniques were developed from traffic accident data utilized by public safety officials searching for efficient mechanisms to identify, develop and deploy appropriate countermeasures and enforcement regiments to lower traffic accident occurrences. We show that our proposed techniques give similar results to existing techniques, yet are conceptually simpler, and therefore of greater value to a practitioner using data mining in a particular domain. In particular, the proposed techniques are metrics that are meaningful to a practitioner beyond just their statistical implications. Because our proposed techniques are also relatively efficient, they can be used efficiently as conceptually simple substitutes for the more traditional and complex statistical approaches.

As a further investigation of the efficiency of our proposed techniques, we examine their performance under competing storage models. We show that when data are stored in column-major order, the performance of our proposed techniques is quite favorable. While column-major order is generally inappropriate for transactional systems, it has been shown to be superior to row-major order for non-transactional, statistical analysis systems that utilize categorical data (Parrish et al., 2005). Our proposed techniques confirm these results and further support the idea of using column-major storage for systems that support data mining.

Value and Variable ranking

Variable selection attempts to find a variable subset that describes the data for a learning task as good as or better than the original dataset. We note first that, historically, there has been an evolution in the terminology used to describe variable selection. Traditionally, attributes correspond to the columns in a database table. The process of selecting certain attributes or a smaller set of attributes to provide faster builds is called attribute selection. Over time, the terminology has evolved to also call this process variable selection and feature selection. Dating back to at least as early as the 1970's, feature selection (Fukunaga, 1970; Andrews, 1971) has been commonly used. More recently, some literature uses attribute selection (Baim, 1988; Caruana & Freitag, 1994; Pappa, Freitas, & Kaestner, 2002; Hall & Holmes, 2003) or variable selection (Viallefont, 2001; Foster & Stine, 2004). In 2001, the Neural and Information Processing Systems workshop on Variable and Feature Selection used both of the terms variable and feature selection. In this work, we use the *variable selection* in place of feature selection or attribute selection. Figure 1 illustrates the role of variable selection in the data mining process.

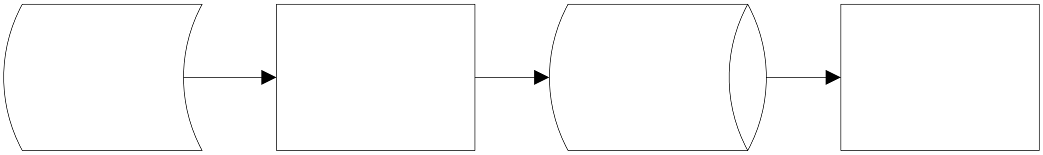


Figure 1-1. The data mining process

In data mining, variable selection generally falls into two categories (Kohavi, John, & Pfleger, 1994; Mladenic & Grobelnik, 1999):

- The filter model and
- The wrapper model.

The filter model selects a variable subset independently without involving any learning algorithm that will use the selected variables. The wrapper model selects a variable subset using the evaluation function based on the predetermined learning algorithm. This type of wrapper approach generally produces a better variable subset but tends to consume a lot of resources. When the number of variables is very large, filtering is generally the preferred approach.
As noted above, one filter model technique is commonly referred to as *variable ranking* (Kohavi & John, 1997; Guyon & Elisseeff, 2003). Variable ranking is a data mining preprocessing step where variables are selected independently of the learning algorithm that will use the selected variables. The procedure of variable ranking is to score each variable according to a particular method; the best *k* variables will then be selected. For example, using a given method, rank the 10 best variables in predicting alcohol related traffic accidents. The advantage of variable ranking is that it requires only the computation of scores of each variable individually and sorting the scores (Guyon & Elisseeff, 2003; Stoppiglia et al., 2003).

Also as noted above, *value ranking* is closely akin to variable ranking. For categorical data, a variable may take on one of a fixed set of possible values. For example, a variable labeled gender may take on values of “Male,” “Female,” or “Unknown.” Value ranking is the process of determining what values of a variable are most important or contribute most significantly to the variable selection process. For example, if the *variable* “Gender” is ranked significant in predicting alcohol related accidents, the *value* “Male” can be seen as the most important contributor in “Gender” being ranked.

Our approach in this research is to examine existing value and variable ranking techniques. We then propose new techniques that represent metrics that have intuitive utility to a practitioner, comparing our proposed techniques to the existing ones in terms of (a) consistency with the existing techniques and (b) performance. We show that the proposed techniques correlate well with the existing techniques, with favorable performance as well.

Value and variable ranking are statistical approaches that examine one or a small number of attributes that describe a record. Traditionally, these attributes are considered to be the columns in a two dimensional table with each record being a row in the table. Previous results indicate that statistical operations such as frequency and cross-tabulations are more efficient when the underlying data is stored and processed in column-major order (i.e., with each column stored contiguously on the disk) (Parrish et al., 2005). This is contrary to the vast majority of contemporary database systems that process their data in row-major order (i.e., with each row stored contiguously on the disk). In particular, rows are often inserted or deleted as a unit and updates tend to be row-oriented. Statistical processing which may be concerned with only a few attributes (columns) benefits from column-major order. Our results in evaluating Max Gain (MG), a new value ranking technique that we propose, and Sum Max Gain Ratio (SMGR), a new variable ranking technique that we propose, continue to support this research.

Section 2
The Problem Domain

Traffic Accident Analysis

The particular problem domain addressed in this study is the analysis of automobile traffic accident data. Given the appropriate tools and data, variable ranking allows traffic safety professionals to develop and adopt countermeasures to reduce the volume and severity of traffic accidents (Parrish et al., 2003). The University of Alabama has developed a software system called CARE (Critical Analysis Reporting Environment) for the analysis of traffic crash data. CARE provides a tool that allows transportation safety engineers and policy makers to analyze the data collected from traffic accident records.

CARE has been applied to traffic crash records from a number of states. The Alabama statewide crash database in CARE has records (rows) that contain 228 categorical variables (column); each variable contains attribute values varying from 2 to more than 600. CARE's analysis domain is restricted to categorical data, represented by nominal, ordinal and interval based variables. Nominal variables have attribute values that have no natural order to them (e.g., pavement conditions – wet, dry, icy, etc.). Ordinal variables do have a natural order (e.g., number of injured, day of the week). Interval variables are created from intervals on a contiguous scale (e.g., age of driver – 16-20, 20-25, 25-35, etc.).

The current system provides the users with filters to perform data analysis on particular subsets of the data that are of interest. Filters are defined by Boolean expressions over the variables in the database. A record satisfying a filter's Boolean expression is a member of the filter subset, while a record not satisfying the filter's Boolean expression is excluded from the filter subset. Common filters for crash data are filters for crashes within specific counties, filters defining crashes related to alcohol, filters defining crashes involving pedestrians, etc.

In terms of our variable ranking techniques, the filter represents the target variable for learning with two values: "0" and "1." In particular, "0" corresponds to the records not satisfying the filter's Boolean expression, while "1" corresponds to the records satisfying the Boolean expression. Filters provide an effective conceptual framework for our value and variable ranking techniques in that they define two subsets: a *control subset* and an *experimental subset*. The idea that a filter defines an experimental group that is compared with a control group is fundamental to the value and variable ranking techniques discussed here. For example, in our traffic crash domain, an "alcohol" filter defines the subset of crashes where alcohol is involved. Our value and variable ranking techniques compare the alcohol crashes (experimental group) with all other crashes (control group) to conclude which values and variables are most important.

Section 3
Value Ranking

Value Ranking Fundamentals

Value ranking is the process of assigning a weight or score to a value based on the value's occurrence in the subset of interest (experimental subset) when compared to the comparison subset (control subset). Value ranking is described using the following notation. When considering a training dataset that has n input variables and one target variable, the input variables are noted as V_k (k = 1,…,n). For a particular variable V_k and each value i of variable V_k, a score $S(V_{k,i})$ of value i of the variable V_k is derived according to a particular value ranking method computed from a corresponding contingency table described as follows. It is assumed that a higher score of a value indicates a more valuable value than those with lower scores. For ranking purposes, these values are sorted in descending order of $S(V_{k,i})$.

Let the value of the target variable be "1" (i.e., experimental subset) or "0" (i.e., control subset), and the values of input variable V_k be $V_{k,0}$, …, $V_{k,r-1}$. Then a contingency table (frequency table) showing the responses of subjects to one variable as a function of another variable (Stockburger, 1996; Zembowicz & Zytkow, 1996) can be built up as described in Table 3-1.

Table 3-1. A contingency table of variable V_k and target variable

Input Variable V_k	Target Variable		Row Totals
	0	1	
$V_{k,0}$	$f_{0,0}(F_{0,0})$	$f_{0,1}(F_{0,1})$	$f_{0,*}$
…	…	…	
$V_{k,r-1}$	$f_{r-1,0}(F_{r-1,0})$	$f_{r-1,1}(F_{r-1,1})$	$f_{r-1,*}$
Column Totals	$f_{*,0}$	$f_{*,1}$	m

Where $f_{i,j}$ is the frequency for which the value of the variable V_k is i and the value of the target variable is j, $f_{*,j} = \sum_{i=0}^{r-1} f_{i,j}$, $f_{i,*} = f_{i,0} + f_{i,1}$, $F_{i,j} = \frac{f_{*,j} \times f_{i,*}}{m}$, i=0, 1, 2, …, r-1, j = 0 and 1, and m is the total number of records.

Existing Value Ranking Methods

The following sections describe existing value ranking techniques including Statistical Significance Z Value (SSZ) (Howell, 2001; Richards, 2002; Parrish et al., 2003), Confidence (CF) (Agrawal, Imielinski, & Swami, 1993), Support (SP) Agrawal, Imielinski, & Swami, 1993) and Improvement (IM) (Berry & Linoff, 1997).

Statistical Significance Z Value (SSZ)

To determine statistical significance of a particular value, a standard statistical significance Z test of proportions is performed. Any value with a difference in proportions which is significant at the 2 percent alpha level (the critical value is 2.33 or -2.33) is considered to be statistically significant, this also indicates that we can consider the likelihood that there is a difference in the population to be 98 percent (Parrish et al., 2003). The mathematical definition of Statistical Significance Z Value is described as follows (Howell, 2001; Richards, 2002). Eq:

$$SSZ(V_{k,i}) = \frac{\left(f_{i,1} \big/ \sum_{j=0}^{r-1} f_{j,1} - f_{i,0} \big/ \sum_{j=0}^{r-1} f_{j,0} \right)}{\sqrt{\left((f_{i,0}+f_{i,1}) \big/ (\sum_{j=0}^{r-1} f_{j,0} + \sum_{j=0}^{r-1} f_{j,1}) \right) \times \left(1 - (f_{i,0}+f_{i,1}) \big/ (\sum_{j=0}^{r-1} f_{j,0} + \sum_{j=0}^{r-1} f_{j,1}) \right) \times \left(1 \big/ \sum_{j=0}^{r-1} f_{j,0} + 1 \big/ \sum_{j=0}^{r-1} f_{j,1} \right)}} \quad \text{Eqn. 3-1}$$

where r is the number of values of variable V_k. The Statistical Significance Z Value considers the value's occurrence in both full dataset and in separate datasets (experimental subset and control subset), it also considers the contribution of the sample size.

Confidence (CF) and Support (SP)

The problem of mining association rules has been introduced previously (Srikant & Agrawal, 1995). Given a dataset and a filter (Y), we need rank attribute value (X) of a particular variable. The mining association rule can be X→Y. For example, Friday→ Alcohol where "Friday" is one attribute value of variable "Day of Week" and "Alcohol" is the filter. The intuitive meaning of such a rule is that records in the dataset which contain X tend to also contain Y (Agrawal, Imielinski, & Swami, 1993). The Confidence of the rule is the number of records where X and Y occurred together divided by the total number of records where X occurred. High confidence implies a strong relationship between X and Y. The formula is described as:

$$CF(V_{k,i}) = \frac{f_{i,1}}{f_{i,0} + f_{i,1}} \quad \text{Eqn. 3-2}$$

The Support of the rule is the number of records where X and Y happened together, divided by the total number of records. The formula is described as:

$$SP(V_{k,i}) = \frac{f_{i,1}}{\sum_{j=0}^{r-1} f_{j,0} + \sum_{j=0}^{r-1} f_{j,1}} \quad \text{Eqn. 3-3}$$

where r is the number of values of variable V_k.

Improvement (IM)

Given an if-then rule (Kautardzic, 2001), such as:

if (a special value occurs, called condition)

then

(experimental subset occurs, called result),

a measure called Improvement indicates how much better a rule is at predicting the result than just assuming the result in the entire dataset without considering anything. Improvement is defined as the confidence of the rule divided by the support of the result. The mathematical definition of Improvement is given by the following formula:

$$IM(V_{k,i}) = \frac{f_{i,1}/(f_{i,0} + f_{i,1})}{\sum_{j=0}^{r-1} f_{j,1}/(\sum_{j=0}^{r-1} f_{j,0} + \sum_{j=0}^{r-1} f_{j,1})} \qquad \text{Eqn. 3-4}$$

where r is the number of values of variable V_k. When Improvement is greater than 1, the value is better at predicting the result than random chance, otherwise the prediction is worse.

Section 4
New Value Ranking Methods

New Value Ranking Techniques

As detailed above, value ranking is the process of assigning a weight or score to a value based on its occurrence in the experimental subset under investigation when compared to the control subset. We propose three value ranking techniques: Over-Representation (OR), MG, and Max-Max Gain (MMG).

Over-Representation (OR)

Over-Representation is a simple extension of a frequency distribution. The degree of Over-Representation for a particular value is the value's occurrence in the experimental subset (the subset of interest) divided by the value's occurrence in the control subset (the subset for the comparison purpose) (Parrish et al., 2003). To calculate the degree of Over-Representation for a particular value one must first determine the value's occurrence in both the experimental class and control class (computed as percentage) and then divide these two values. It is possible to derive the Over-Represented values from a contingency table. Suppose variable V_k has value i, the Over-Representation of value i can be obtained by:

$$OR(V_{k,i}) = \frac{f_{i,1} / \sum_{j=0}^{r-1} f_{j,1}}{f_{i,0} / \sum_{j=0}^{r-1} f_{j,0}} \qquad \text{Eqn. 4-1}$$

where r is the number of values of variable V_k. Consider an example from the traffic safety domain. Suppose that 50 percent of the alcohol crashes occur on rural roadways, while only 25 percent of the non-alcohol crashes occur on rural roadways. Therefore, the degree of Over-Representation of alcohol crashes on rural roadways is 50 percent /25 percent = 2. Put simply, alcohol accidents are Over-Represented on rural roadways by a factor of 2. If the degree of Over-Representation is greater than 1, the value is an Over-Represented value. In the traffic safety domain, Over-Representation often indicates problems that need to be addressed through countermeasures (i.e., safety devices, sobriety checks, etc). As a simple ratio, OR is obviously a very intuitive quantity to a practitioner.

Max Gain (MG)

One important question a safety professional might ask is: What is the potential benefit from a proposed countermeasure? The answer to this question is that in the best case a countermeasure will reduce crashes to its expected value. It is unlikely to reduce crashes to a level less than what is found in the control group. A metric termed Max Gain (Parrish et al., 2003; Wang, Parrish, & Chen, 2003) is used to express the number of cases that could be reduced if the subset frequency (experimental subset) was reduced to its expected value (control subset). Max Gain can be defined by the value's occurrence in the experimental subset minus the experimental subset frequency times the probability the control class occurred. The formula is described as:

$$MG(V_{k,i}) = f_{i,1} - \sum_{j=0}^{r-1} f_{j,1} \times \frac{f_{i,0}}{\sum_{j=0}^{r-1} f_{j,0}} \qquad \text{Eqn. 4-2}$$

where r is the number of values of variable V_k. Max Gain is a powerful metric when designing countermeasures. If a choice must be made between two countermeasures, the countermeasure with higher Max Gain value has the higher potential benefit. If a particular value of a variable is Over-Represented, the value has positive Max Gain, otherwise the value has negative Max Gain.

Consider an example from traffic safety domain: Analysis indicates that "OFF ROADWAY" accidents demonstrate the highest Max Gain in an experimental subset. The Max Gain of "OFF ROADWAY" can be computed by: 3680 – 7743 * (18029 / 124883) = 2562.165. Table 4-1 shows the Max Gain of each attribute value for the variable "EVENT LOCATION".

Table 4-1. An example of Max Gain

V016: EVENT LOCATION	Experimental Subset	Control Subset	OR	MG
OFF ROADWAY	3680	18029	3.292	2562.165
MEDIAN	89	940	1.527	30.718
PRIVATE ROAD/PROPERTY	46	293	2.532	27.833
DRIVEWAY	10	50	3.226	6.9
INTERSECTION	1028	30641	0.541	-871.804
ON ROADWAY	2890	74930	0.622	-1755.81
Totals	7743	124883		

Max Gain is a metric that can be quoted by practitioners in the domain of interest. For example, in traffic safety, Max Gain is the reduction potential when a "countermeasure" achieves its highest potential. For the example given in Table 4-1, "Rumble Strips" are a countermeasure often used to reduce the number of "OFF ROADWAY" accidents. Assuming 100% success with Rumble Strips, the "OFF ROADWAY" crashes can be reduced by a total of 2562. One cannot expect a reduction that exceeds the Max Gain. Effectively, Max Gain then becomes an upper bound on crash reduction potential within this domain. Because Max Gain can be used in such an intuitive fashion, it becomes a very practical metric – much more practical than some of the more abstract statistical metrics, such as Confidence, Support and Statistical Significance Z value.

Max-Max Gain (MMG)

Max Gain shows the potential benefit of implementing a countermeasure. After applying the countermeasure, the experimental subset frequency is reduced and the control subset frequency is increased. A recalculation of Max Gain would then produce a different ordering based on these changed values. The reordering in Max Gain would then highlight another value. A traffic safety professional might ask: What is the potential reduction in accident numbers if I continue to focus my countermeasures on the original problem value? Max-Max Gain is a proposed measure to rank values based on that question.

Consider the attribute value of OFF ROADWAY which demonstrates the highest Max Gain in the above example. These particular accidents account for 3,680 crashes and demonstrate a Max Gain of 2,562. For this example, a countermeasure to reduce the OFF ROADWAY accidents to the same percentage as the control group would leave 1,118 “OFF ROADWAY” accidents in the experimental subset. A further calculation of Max Gain as described above would then identify another value (for example “Intersection”) as having the highest Max Gain. Instead of investigation or adopting a new countermeasure for the newly identified value, the traffic safety professional might be more interested in determining the maximum possible benefit of concentrating effort on the originally identified problem (OFF ROADWAY in this example). Max-Max Gain does this by repeatedly calculating and summing Max Gain for a value until the potential gain becomes less than 1.0. Table 4-2 shows the steps to compute the Max-Max Gain of attribute value OFF ROADWAY for variable “EVENT LOCATION”.

Table 4-2. An example of Max-Max Gain

Step	Subset Freq.	Subset Total	Other Freq.	Other Total	Max Gain
1	3680	7743	18029	124883	2562.17
2	1117.83	5180.83	20591.17	127445.17	280.77
3	837.06	4900.06	20871.94	127725.94	36.33
4	800.73	4863.73	20908.27	127762.27	4.78
5	795.95	4858.95	20913.05	127767.05	0.63
MMG 2562.17 + 280.77 + 36.33 + 4.78 + 0.63 = 2884.68					

Max-Max Gain can be calculated for all values of a variable providing the traffic safety engineer insight into the most important value that might be addressed by a countermeasure. In this regard, MMG provides the same benefits to the practitioner as MG.

Results Using New Value Ranking Methods

This research proposes three new value ranking methods, OR, MG and MMG. We choose to evaluate only MG and MMG here; OR’s contribution is principally to support the other two methods. To compare the performance of MG and MMG to the other ranking methods, we used the Alabama accident dataset for the year 2000. The dataset contains 132,626 records and 228 variables. The target variables were selected by applying the filters of Injury, Interstate, Alcohol, and Fatality. The corresponding occurrence percentage is 21.92%, 8.996%, 5.38% and 0.4853%, respectively. For each filter, the experimental class (those accidents indicated by the filter) was represented by a 1 and the control class (the remaining accidents) was represented by a 0.

Pearson’s Correlation With Existing Methods

Since the strength of the linear association between two methods is quantified by the correlation coefficient, we use Pearson’s correlation coefficient to test if there exits a relationship between

MG, MMG and the other methods. Since MG and MMG are conceptually simple, a strong or moderate correlation with an existing method would mean that MG and MMG could effectively substitute for that method.

Table 4-3 shows the number of variables and the strength of Pearson's correlation between each technique to Max Gain that will be used for value ranking. As described in Table 4-3, MG is strongly correlated to MMG and SSZ, since they have the most number of variables that are strongly correlated to MG. Similar results occur for MMG, as MMG is also strongly correlated to SSZ. Since MG and MMG are conceptually simple and are strongly correlated to SSZ, MG and MMG could effectively substitute for SSZ, with a likelihood of higher intuitive utility by a domain practitioner.

Table 4-3. Number of variables correlated to MG

Value Ranking	Number of variables Correlated based on Pearson's R		
	Strong	Moderate	Weak
MMG	175	48	1
SSZ	173	50	1
CF	24	70	131
SP	32	133	60
IM	23	70	131

Running Time Efficiency Compared With Existing Techniques

Table 4-4 shows the average run time cost over one thousand experiments for each value ranking method, using the Alabama crash dataset for the year 2000. For each experiment, we rank the attribute value for each variable and obtain a run time cost. Table 4-4 shows that MG and CF are the most efficient methods. MMG is not efficient since it needs to repeat to compute Max Gain.

Table 4-4. Comparison of run times for different value ranking techniques (in milliseconds)

Value Ranking	Filter			
	Injury	Interstate	Alcohol	Fatality
OR	0.8348	0.8	1.875	0.7782
MG	0.675	0.6814	0.9782	0.6938
MMG	1.775	2.1532	2.2032	1.0218
SSZ	0.8188	0.8126	1.8344	0.8124
CF	0.6782	0.6874	0.9532	0.675
SP	1.572	1.5812	1.6124	1.5814
IM	1.7906	1.8032	7.7094	1.7908

Section 5
Variable Ranking

The Variable Ranking Problem

The problem of variable ranking can be described in the notation provided by Guyon and Elisseeff (2003). We consider a training dataset with m records $\{x_t; y_t\}$ (t = 1,…,m), each record consists of n input variables $x_{t,k}$ (k = 1,…,n) and one target variable y_t. The input variables are noted as V_k (k = 1,…,n). Then we get a score $S(V_k)$ of each variable V_i according to a particular variable ranking method computed from the corresponding contingency table. We assume that a higher score of a variable indicates a valuable variable and that the input variables are independent of each other. The variables are then sorted in decreasing order of $S(V_k)$ allowing us to select the top most *x* variables of interest.

We can construct the contingency table for each variable V_k (k = 1, …, n) as described in section 3. Various variable ranking methods have been proposed. In the following, we present the ranking methods we applied.

Existing variable ranking methods

The following sections describe existing variable ranking methods including Chi-squared (CHI) (Lehmann, 1959; Hawley, 1996; Leonrd, 2000), Correlation Coefficient (CC) (Golub et al., 1999; Furey et al., 2000; Slonim et al., 2000; Guyon, Weston, Barnhill, & Vapnik, 2002), Information Gain (IG) (Mitchell, 1997, Yang & Pedersen, 1997) and Gain Ratio (GR) (Grimaldi, Cunningham, & Kokaram, 2003). The limitation of these methods will be discussed in section 6.

Chi-squared (CHI)

The Chi-squared test (Lehmann, 1959; Hawley, 1996; Leonard, 2000) is one of the most widely used statistical tests. It can be used to test if there is ‘no association’ between two categorical variables. ‘No association’ means that for an individual, the response for one variable is not affected in any way by the response for another variable. This implies the two variables are independent. The Chi-squared measure can be used to find the variables that have significant Over-Representation (association) in regard to the target variable.

The Chi-squared test can be calculated from a contingency table (Yang & Pedersen, 1997; Forman, 2003). The Chi-squared value can be obtained by the equation below:

$$\chi^2(V_k) = \sum_{i=0}^{r-1}\sum_{j=0}^{1}\frac{(f_{i,j}-F_{i,j})^2}{F_{i,j}} \qquad \text{Eqn. 5-1}$$

where k is the variable number (k=1,…,n), and r is the number of values of the variable.

Correlation Coefficient (CC)

If one variable's expression in one class is quite different from its expression in the other, and there is little variation between the two classes, then the variable is predictive. So we want a variable selection method that favors variables where the range of the expression vector is large, but where most of that variation is due to the class distribution (Slonim et al., 2000). A measure of correlation scores the importance of each variable independently of the other variables by comparing that variable's correlation to the target variable (Weston, Elisseeff, Scholkopf, & Tipping, 2003).

Let the value of the target variable be "1" (i.e., experimental class) or "0" (i.e., control class). For each variable V_k, we calculate the mean $\mu_{k,1}$ of the experimental class ($\mu_{k,0}$ of the control class) and standard deviation $\sigma_{k,1}$ of the experimental class ($\sigma_{k,0}$ of the control class). Therefore we calculate a score:

$$CC(V_k) = \left| \frac{\mu_{k,1} - \mu_{k,0}}{\sigma_{k,1} + \sigma_{k,0}} \right| \qquad \text{Eqn. 5-2}$$

which measures relative class separation. A large value of $CC(V_k)$ indicates a strong correlation between input variable and class distribution (Golub et al., 1999; Furey et al., 2000; Slonim et al., 2000; Guyon, Weston, Barnhill, & Vapnik, 2002). We then simply take the variables with the highest $CC(V_k)$ scores as our top variables. The correlation score is closely related to the Fisher criterion score (Bishop, 1995; Pavlidis, Weston, Cai, & Grundy, 2002; Weston et al., 2003).

Information Gain (IG)

Entropy measures the impurity of a set of data (Mitchell, 1997). If there is, at most, one class present, entropy is the lowest. And if the proportions of all present classes are equal, entropy is highest. Information Gain is a measure based on entropy. Information Gain measures the decrease of the weighted average impurity of the partitions, compared with the impurity of the complete set of data. Yang and Pederson (1997) reported Information Gain and Chi-squared test performed best in their multi-class benchmarks. Information Gain has a generalized form for nominal valued attributes (Forman, 2003). The Information Gain value can be obtained from the contingency table:

$$IG(V_k) = e(\sum_{i=0}^{r-1} f_{i,1}, \sum_{i=0}^{r-1} f_{i,0}) - \sum_{i=0}^{r-1} (\frac{f_{i,1} + f_{i,0}}{m} \times e(f_{i,1}, f_{i,0})) \qquad \text{Eqn. 5-3}$$

where $e(x,y) = -\frac{x}{x+y}\log_2\frac{x}{x+y} - \frac{y}{x+y}\log_2\frac{y}{x+y}$, k is the variable number (k=1,…,n), r is the number of values of the variable, and m is the total number of records.

Gain Ratio (GR)

Unfortunately, Information Gain is biased in favor of variables with more values. That is, if one variable has a greater numbers of values, it will appear to gain more information than those with fewer values, even if they are actually no more informative. Information Gain's bias is symmetrical toward variables with more values. Gain Ratio (GR) overcomes this problem by introducing an extra term. There are two methods to compute GR; the first method computes GR by:

$$GR(V_k) = \frac{IG(V_k)}{r}, \qquad \text{Eqn. 5-4}$$

where r is the number of values of variable V_k.
The second method computes GR (Grimaldi, Cunningham, & Kokaram, 2003) by:

$$GR(V_k) = \frac{IG(V_k)}{SP(V_k)}, \quad SP(V_k) = -\sum_{i=0}^{r-1} \frac{f_{i,1} + f_{i,0}}{m} \log_2 \frac{f_{i,1} + f_{i,0}}{m} \qquad \text{Eqn. 5-5}$$

where k is the variable number (k=1,…,n), r is the number of values of the variable, and m is the total number of records. Since the *SP* term can be zero in some special cases, the authors in (Grimaldi, Cunningham, & Kokaram, 2003) define:

$$GR(V_k) = IG(V_k) \text{ if } SP(V_k) = 0. \qquad \text{Eqn. 5-6}$$

In our study, we used the second method because it gives us more information about the distribution of accidents.

Section 6
A New Variable Ranking Method

Sum Max Gain Ratio (SMGR)

The above variable ranking methods compute the score of each variable based on all the attribute values of a particular variable. This implies every attribute value has the same impact on a particular variable. We propose a new variable ranking method called Sum Max Gain Ratio (SMGR) (Wang, Parrish, Smith, & Vrbsky, 2005) that computes the score based on only a portion of the attribute values for a particular variable. In section 4, we introduced how to compute Max Gain. Those attribute values that have positive Max Gain will be Over-Represented.

Sum Max Gain (SMG) is the total number of cases that would be reduced if the subset frequency were reduced to its expected value for the attribute values that are Over-Represented (i.e., those values that have a Max Gain > 0). That is, SMG is the sum of all positive Max Gain values for a particular variable. More formally:

$$SMG(V_k) = \sum_{i=0}^{r-1} MG(V_{k,i}) \quad if \quad MG(V_{k,i}) > 0 \qquad \text{Eqn. 6-1}$$

where r is the number of values of variable V_k. As an example, Table 6-1 shows the distribution by “Day of Week” for alcohol accidents in Alabama’s Mobile County for the year 2000. Saturday and Sunday exhibit the positive Max Gains. The Sum Max Gain of variable of Day of Week (V_{008}) will be: SMG (V_{008}) = 1031.824 + 760.082 = 1791.906.

Table 6-1. Frequency table of variable of “Day of Week”

V_{008}: Day of Week	Subset Freq.	Other Freq.	Over Rep.	Max Gain
Saturday	*2012*	*15837*	*2.053*	*1031.824*
Sunday	*1474*	*11535*	*2.065*	*760.082*
Friday	1275	22868	0.901	-140.335
Tuesday	737	17750	0.671	-361.574
Thursday	872	19954	0.706	-362.983
Wednesday	693	18092	0.619	-426.741
Monday	667	18860	0.571	-500.274

One problem with SMG is its bias in favor of variables with fewer values. By dividing by the total number of cases in the subset frequency (experimental class), it is possible to factor out this issue. In particular, Sum Max Gain Ratio (SMGR) is the ratio of the number of cases that could potentially be reduced by an effective countermeasure (SMG) to the total number of cases associated with the Over-Represented values (i.e., those cases where the Max Gain > 0). That is:

Eqn. 6-2

$$SMGR(V_k) = SMG(V_k) / \sum_{i=0}^{r-1} f_{i,1} \quad if \quad MG(V_{k,i}) > 0$$

where r is the number of values of variable V_k. The SMG of variable “Day of Week” was computed from the above example to be 1791.906 (1031.824 + 760.082). Saturday and Sunday have positive Max Gain, and these two attribute values are Over-Represented. The total number of cases associated with the Over-Represented values will be 3486 (the sum of 2012 and 1474). The SMGR will be 0.514 (1791.906 / 3486 = 0.514).

SMGR is always in the range of 0 to 1 because SMG is always less than the corresponding subset frequency. A high score of SMGR is indicative of a valuable variable with a high degree of relevance to the filter subset (experimental group). When presented in decreasing SMGR order, the most relevant variables are at the top.

Results Using SMGR

To compare the performance of SMGR to the other ranking methods, we used the Alabama Mobile County accident dataset for the year 2000. The target variables were selected by applying the filters of Injury, Interstate, Alcohol, and Fatality, respectively.

The following sections compare the Pearson’s correlation between SMGR and existing variable ranking methods, the run time performance, predictive ability.

Pearson’s Correlation With Existing Methods

Since the strength of the linear association between two methods is quantified by the correlation coefficient, we use Pearson’s correlation coefficient to test if there exits a relationship between Sum Max Gain Ratio and any other method. Since SMGR is conceptually simple, a strong correlation with an existing method would mean that SMGR could effectively substitute for that method.

Table 6-2 shows the Pearson’s correlation coefficient between SMGR and the other variable ranking methods. SMGR is correlated to CHI, IG and GR. Table 6-3 shows the complexity of each method. For complexity comparison, n is the number of variables and m is the number of attribute values. SMGR may be preferable to other variable ranking methods because of its conceptual simplicity and less complexity.

Table 6-2. Pearson's correlation, SMGR and other methods

Variable Ranking	Filter			
	Injury	Interstate	Alcohol	Fatality
SMG	weak	moderate	strong	weak
CHI	moderate	moderate	strong	moderate
CC	weak	weak	weak	weak
IG	moderate	moderate	strong	moderate
GR	moderate	moderate	moderate	moderate

Table 6-3. Complexity analysis

Method	Complexity
SMGR	O(n □ m / 2)
SMG	O(n □ m / 2)
CHI	O(n □m)
CC	O(n □ 2m)
IG	O(n □ m)
GR	O(n □ m)

Running Time Comparison

SMGR is efficient since it computes a variable score based on the Over-Represented attribute values, while other methods compute a variable score based on all attribute values. To illustrate this, we ran experiments on a real-world traffic dataset. The results are given in Table 6-4.

Table 6-4. Comparison of run times for different variable ranking techniques (in milliseconds)

Variable Ranking	Filter			
	Injury	Interstate	Alcohol	Fatality
SMGR	1.542	1.504	1.465	1.407
SMG	1.560	1.498	1.484	1.413
CHI	2.834	2.851	2.769	2.840
CC	5.563	5.520	5.446	5.523
IG	5.296	4.990	4.336	3.491
GR	9.366	8.940	8.343	7.497

Table 6-4 shows the average run time cost over one thousand experiments. Each experiment uses the same dataset and variable ranking method. For each experiment, we rank variables and get a run time cost. From Table 6-4, we can see the average run time cost of SMGR and SMG is the lowest, while the running costs of CHI, CC, IG and GR are higher. The dataset examines contained 228 variables and each variable has attribute values ranging from 2 to 560. For this dataset, the execution time is relatively small for all tested approaches (less than 10 ms). However, the savings in execution cost (50%) of SMGR to the second closest approach will have significant implications for large datasets with tens or hundreds of thousands of variables, and when the attribute value of a variable is diverse.

Classification

In order to evaluate how our method of variable ranking affected classification, we employed the well known classification algorithm C4.5 (Quinlan, 1993) on the traffic accident data. We used C4.5 as an induction algorithm to evaluate the error rate on selected variables for each variable ranking method. C4.5 is an algorithm for inducing classification rules in the form of a decision tree from a given dataset. Nodes in a decision tree correspond to features and the leaves of the tree correspond to classes. The branches in a decision tree correspond to their association rule.

C4.5 was applied to the datasets filtered through the different variable ranking methods. We used the Injury filter as described above to define the target variable. The top 25 best variables were selected through different variable ranking methods. Table 6-5 shows the error rate and the size of the decision tree for each variable ranking method.

Table 6-5. Results for the C4.5 algorithm

Method	Error rate	Size of the tree (# of nodes)
SMGR	0.7%	32
SMG	1.0%	81
CHI	0.9%	159
CC	7.2%	13070
IG	0.9%	159
GR	0.7%	32

As seen in the experimental results, SMGR, CHI, IG and GR did not significantly change the generalization performance and these methods performed much better than CC. Table 6-5 also shows how variable ranking methods affects the size of the trees (the number of nodes in a tree) induced by C4.5. Smaller trees allow a better understanding of the decision tree. The size of the resulting tree generated by SMGR and GR showed a decrease from the maximum of 14438 nodes to 32 nodes, accompanied by a slight improvement in accuracy. To summarize, SMG and SMGR perform better than CHI, CC and IG.

Section 7
Using column-major storage to improve variable selection performance

Motivation

The increasingly large datasets available for data mining and machine learning tasks are placing a premium on algorithm performance. One critical item that impacts the performance of these algorithms is the approach taken for storing and processing the data elements. Typical applications, including machine learning applications, tend to store their data in row-major order (Parrish et al., 2005). This row-centric nature of storage is consistent with the needs of typical applications. In particular, rows are often inserted or deleted as a unit and updates tend to be row-oriented. This approach, while efficient for row-centric applications, may not be the most efficient for certain column-centric applications. In particular, many statistical analysis computations and variable selection approaches are column-centric.

In this section we measure the impact on processing time for row-major and column-major storage. For both approaches, we assume all of the data is stored in a single file. (We refer the reader to an analysis of the impact on disk I/O of column-major versus row-major storage.) The processing time advantage provided by column-major order is that column-centric statistical computations can be computed on-the-fly as all data for each column is read consecutively. This differs from the row-major order in which all of the data must be read before any statistical computations can be completed.   We next examine the performance tradeoffs between row-major order and column-major order in the context of variable and value ranking.

Results Under Value Ranking

In order to empirically test disk storage for value ranking, we ran experiments on the Alabama Mobile County dataset for the year 2000. We compute the average run time performance for the different value ranking techniques, including SSZ, CF, SP, IM, MG and MMG, under the same disk storage configuration. Figure 7-1 shows the average run time performance for these value ranking techniques. The chart compares row-major order versus column-major order for a varying number of columns. As seen in Figure 7-1, column-major order provided faster empirical performance for the value ranking techniques.

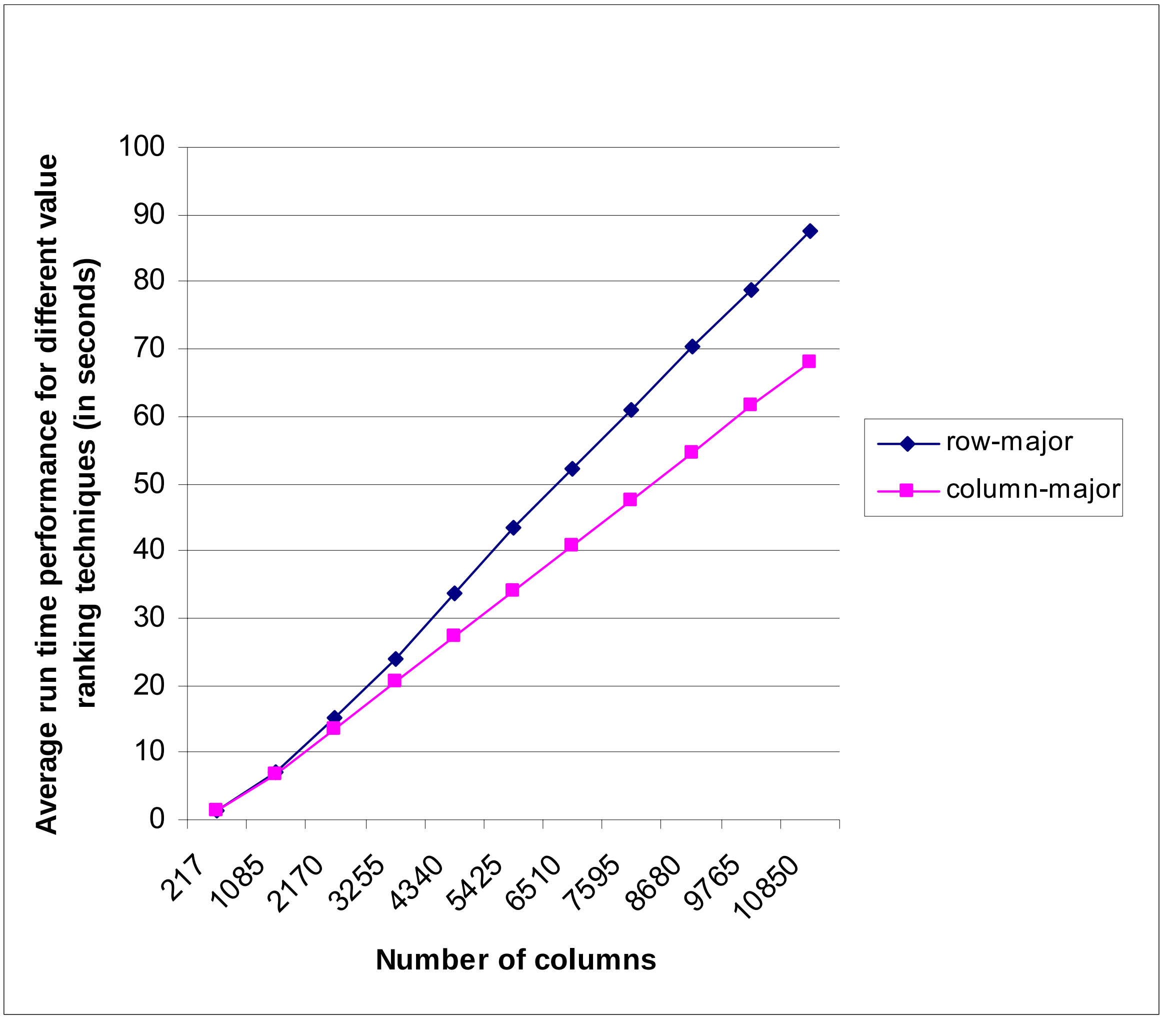


Figure 7-1. Average value ranking performance, column-major versus row-major

Results under variable ranking

To compare the performance of row-major order and column-major order in the context of variable ranking, we used the Alabama Mobile County accident dataset for the year 2000. The dataset contains 14,218 records (i.e., rows) and 228 variables (i.e., columns). The target variable was selected by applying the filter of Injury. The experimental class (those accidents indicated by the filter) was represented by a “1” and the control class (the remaining accidents) was represented by a “0.” To facilitate generating datasets with varying numbers of rows and columns, and to assist in creating row-major and column-major storage options, the original dataset from the CARE application was transferred to Microsoft Excel for manipulation. The manipulated datasets were stored from Microsoft Excel as ASCII files. The variable selection algorithms were developed in C++ and manipulated the datasets in the ASCII files.

For each particular variable selection approach on each different dataset, we run one experiment and get a run time cost. One thousand of the same experiments were run to obtain an average run time cost. Figure 7-2 and Figure 7-3 illustrate the performance differences between the row-major order and column-major order storage methods in the context of SMGR by utilizing specific values for the dataset size, number of rows, number of columns, etc. The figures show the average run time cost in seconds, for a varying number of data values accessed.

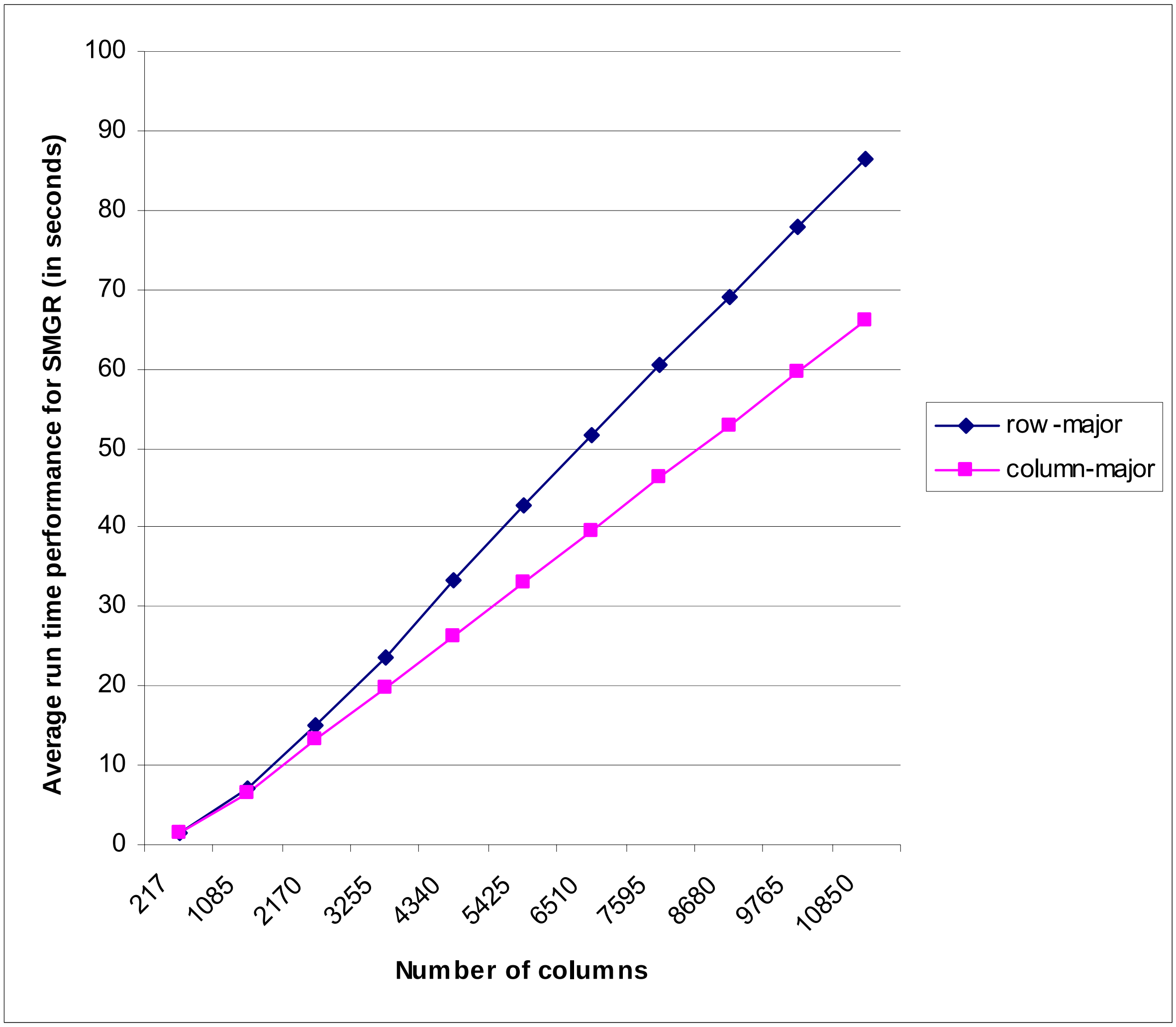


Figure 7-2. Performance comparison, row-major versus column-major, fixed number of rows, SMGR.

Figure 7-2 illustrates the number of seconds to perform the variable selection of SMGR for a varying number of columns but a constant number of rows. The figure illustrates the results for 14,218 records (i.e., rows). The diagonal lines illustrate the run time cost for variable selection using the row-major and column-major methods. The run time ranges from 1.415 to 86.3 seconds using row-major order, while the run time ranges from 1.356 to 66.1 seconds using column-major order. As illustrated in Figure 3, after approximately one thousand columns, the run time cost of row-major order processing increases faster than the column order processing.

In this experiment, column order processing outperforms row order regardless of the selection algorithm used.

Figure 7-3 illustrates the number of seconds to perform variable selection with a varying number of rows but a constant of columns.

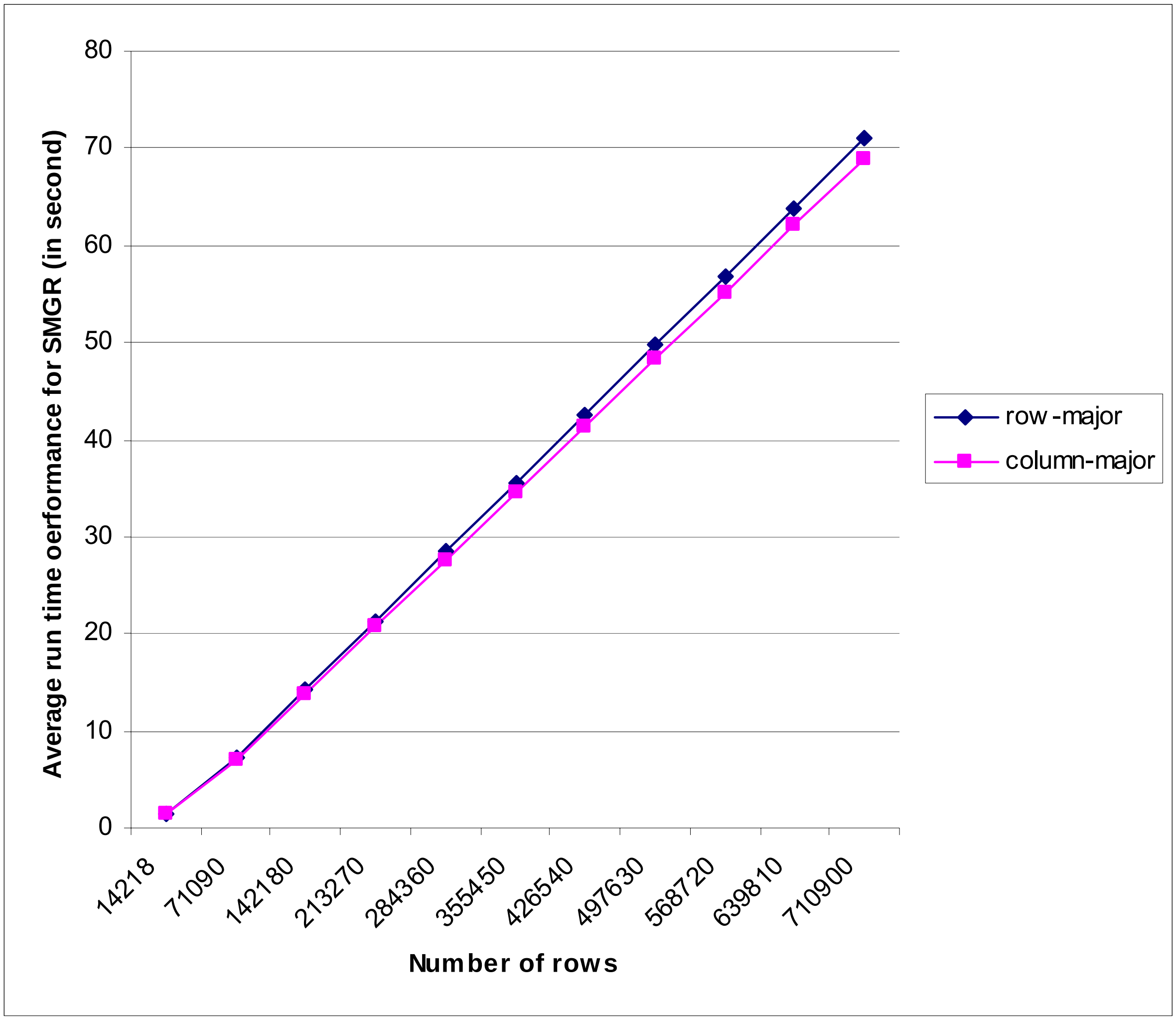


Figure 7-3. Performance comparison, row-major versus column-major, fixed columns, SMGR

Figure 7-3 illustrates the result for 217 columns (i.e., variables). The plot illustrates the run time cost to do variable selection using both row-major and column-major methods. As illustrated in the figure, the run time cost is almost the same using row order and column order until up to approximately 14,000 rows in the dataset. From that point, the run time cost for the row order increases slightly faster than using the column order.

Table 7-1 provides the relative performance of the different selection algorithms for row-major datasets for a range of column sizes. For row-major order, Chi-squared (CHI) consistently performed the worst. No individual approach distinguished itself as best in row-major order. Table 7-2 gives the relative performance of the different selection algorithms for column-major datasets for a range of column sizes. Under column-major order, Information Gain (IG) was consistently the worst performer while Sum Max Gain Ratio (SMGR) was consistently the top performer.

Table 7-1. Variable selection, row-major order

Number of columns	Performance	
	Best	Worst
217	SMGR	CHI
1085	CC	CHI
2170	CC	CHI
3255	GR	CHI
4340	SMGR	CHI
5425	GR	CHI
6510	GR	CHI
7595	IG	CHI
8680	CC	CHI
9765	CC	CHI
10850	SMGR	CHI

Table 7-2. Variable selection, column-major order

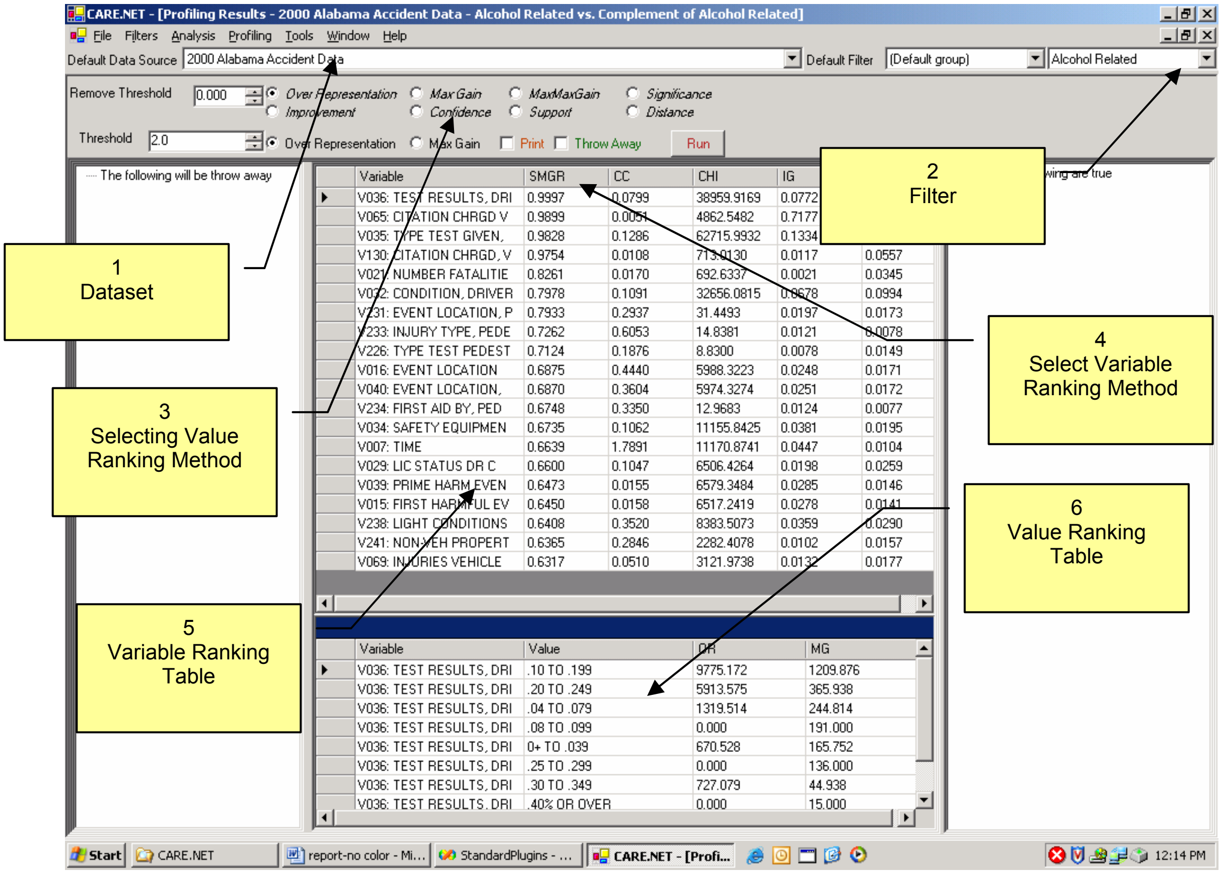
Number of columns	Performance	
	Best	**Worst**
217	CHI	IG
1085	SMGR	IG
2170	SMGR	IG
3255	SMGR	IG
4340	SMGR	IG
5425	SMGR	IG
6510	SMGR	IG
7595	SMGR	IG
8680	SMGR	IG
9765	SMGR	IG
10850	SMGR	IG

This research has shown that column-major order performs better than row-major order in the context of heuristic variable selection. Chi-squared performs the worst when using row-major order and Sum Max Gain Ratio performs the best when using column-major order.

Section 8
Integration with CARE

Improved Query Interface

A primary goal of the CARE application is the ability to provide traffic safety professionals with quick, intuitive and easy to use interface to the rich traffic accident data. The incorporation of the identified variable and value ranking algorithms into the CARE application could provide traffic safety professionals with another tool to assist in identifying causal factors and evaluating deployed counter measures. Figure 8-1 shows an initial prototype for incorporating these data mining techniques into CARE.



CARE.NET - [Profiling Results - 2000 Alabama Accident Data - Alcohol Related vs. Complement of Alcohol Related]

Default Data Source: 2000 Alabama Accident Data
Default Filter: (Default group) | Alcohol Related

Remove Threshold: 0.000
● Over Representation ○ Max Gain ○ MaxMaxGain ○ Significance
○ Improvement ○ Confidence ○ Support ○ Distance

Threshold: 2.0
● Over Representation ○ Max Gain □ Print □ Throw Away

The following will be throw away

...wing are true

Variable	SMGR	CC	CHI	IG	
V036: TEST RESULTS, DRI	0.9997	0.0799	38959.9169	0.0772	
V065: CITATION CHRGD V	0.9899	0.0051	4862.5482	0.7177	
V035: TYPE TEST GIVEN,	0.9828	0.1286	62715.9932	0.1334	
V130: CITATION CHRGD, V	0.9754	0.0108	713.0130	0.0117	0.0557
V021: NUMBER FATALITIE	0.8261	0.0170	692.6337	0.0021	0.0345
V032: CONDITION, DRIVER	0.7978	0.1091	32656.0815	0.0678	0.0994
V231: EVENT LOCATION, P	0.7933	0.2937	31.4493	0.0197	0.0173
V233: INJURY TYPE, PEDE	0.7262	0.6053	14.8381	0.0121	0.0078
V226: TYPE TEST PEDEST	0.7124	0.1876	8.8300	0.0078	0.0149
V016: EVENT LOCATION	0.6875	0.4440	5988.3223	0.0248	0.0171
V040: EVENT LOCATION,	0.6870	0.3604	5974.3274	0.0251	0.0172
V234: FIRST AID BY, PED	0.6748	0.3350	12.9683	0.0124	0.0077
V034: SAFETY EQUIPMEN	0.6735	0.1062	11155.8425	0.0381	0.0195
V007: TIME	0.6639	1.7891	11170.8741	0.0447	0.0104
V029: LIC STATUS DR C	0.6600	0.1047	6506.4264	0.0198	0.0259
V039: PRIME HARM EVEN	0.6473	0.0155	6579.3484	0.0285	0.0146
V015: FIRST HARMFUL EV	0.6450	0.0158	6517.2419	0.0278	0.0141
V238: LIGHT CONDITIONS	0.6408	0.3520	8383.5073	0.0359	0.0290
V241: NON-VEH PROPERT	0.6365	0.2846	2282.4078	0.0102	0.0157
V069: INJURIES VEHICLE	0.6317	0.0510	3121.9738	0.0132	0.0177

Variable	Value	OR	MG
V036: TEST RESULTS, DRI	.10 TO .199	9775.172	1209.876
V036: TEST RESULTS, DRI	.20 TO .249	5913.575	365.938
V036: TEST RESULTS, DRI	.04 TO .079	1319.514	244.814
V036: TEST RESULTS, DRI	.08 TO .099	0.000	191.000
V036: TEST RESULTS, DRI	0+ TO .039	670.528	165.752
V036: TEST RESULTS, DRI	.25 TO .299	0.000	136.000
V036: TEST RESULTS, DRI	.30 TO .349	727.079	44.938
V036: TEST RESULTS, DRI	.40% OR OVER	0.000	15.000

Figure 8-1. Alcohol related filter, rank by SMGR

Figure 8-1 shows the ranking of variables using the alcohol filter for the year 2000 Alabama Accident Dataset. Table 8-1 describes the highlighted boxes in the diagram:

Table 8-1. Description of featured elements in Figure 8-1

Box	Description
1	The dataset selected
2	The filter selected
3	The value ranking method to be used
4	The variable ranking method to be used for sorting the table
5	The variable ranking table sorted by the appropriate method
6	Values for the selected variable sorted by the appropriate value ranking method

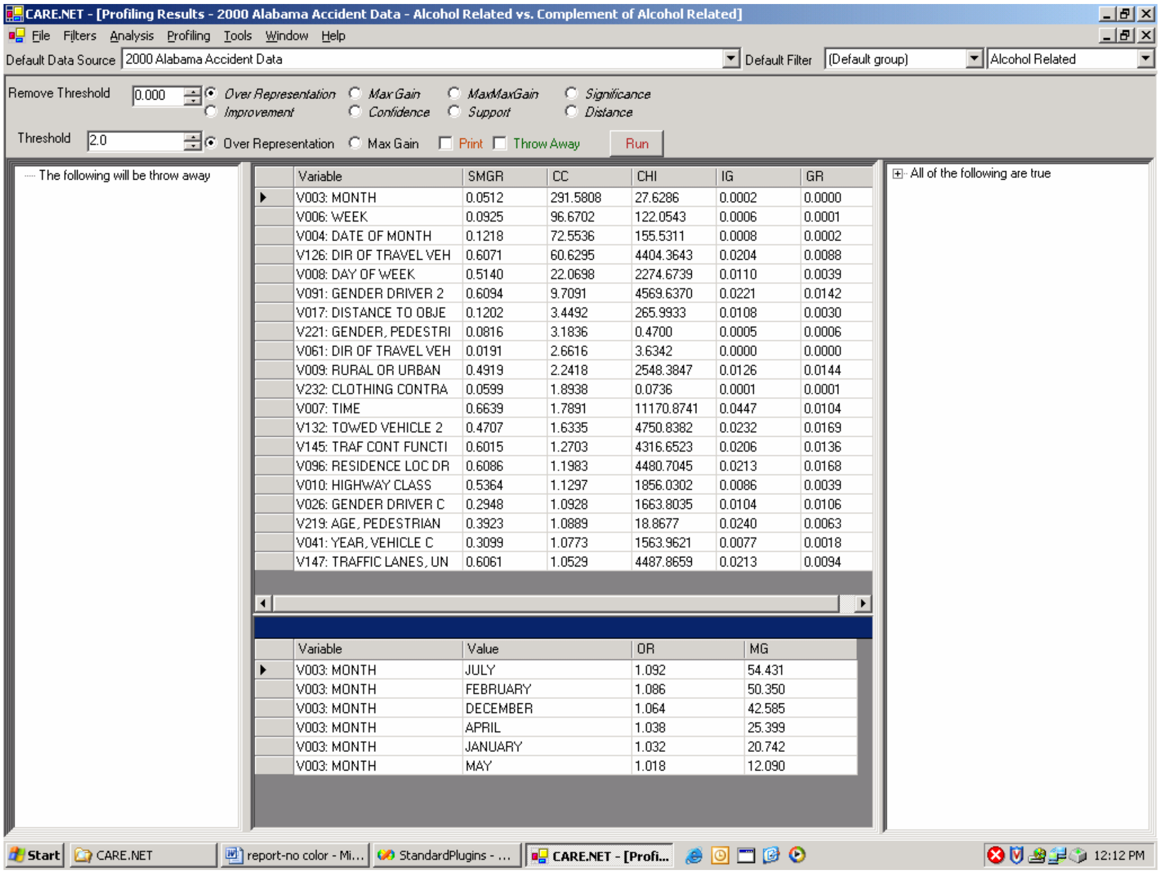
Table 8-2 summarizes the top five variables selected by SMGR as shown in Figure 8-1.

Table 8-2. Summary of top five variables from Figure 8-1

Ranking	Variable	Label	Description
1	V0036	Test Results Driver C	Indicates the results for the Blood Alcohol Test given.
2	V0065	Citation Charged, Vehicle C	Indicates that a citation was given to the causal driver
3	V0035	Type Test Given Driver C	Indicates the type of test given the causal driver
4	V0130	Citation Charged, Vehicle 2	Indicates that a citation was given a second vehicle
5	V0021	Number Fatalities	Indicates the number of fatalities in the accident

The value ranking (indicate by box 6 in Figure 8-1) is for a single variable. For the case in Figure 8-1, the value ranking is for variable V0036 as indicated by the black triangle (▶) next to the variable in the variable ranking table. For this case, the value "0.10 – 0.199" was the most over-represented value for the variable "Test Results, Driver C."

For comparison purposes, Figure 8-2 shows the screen when using CC as the variable ranking method using the same dataset and filter as Figure 8-1. Table 8-3 provides a summary of the top five variables from Figure 8-2.


CARE.NET - [Profiling Results - 2000 Alabama Accident Data - Alcohol Related vs. Complement of Alcohol Related]
File Filters Analysis Profiling Tools Window Help
Start
CARE.NET
report-no color - Mi...
StandardPlugins - ...
CARE.NET - [Profi...
12:12 PM

Default Data Source: 2000 Alabama Accident Data
Default Filter: (Default group) | Alcohol Related

Remove Threshold: 0.000 ● Over Representation ○ Max Gain ○ MaxMaxGain ○ Significance ○ Improvement ○ Confidence ○ Support ○ Distance

Threshold: 2.0 ● Over Representation ○ Max Gain □ Print □ Throw Away Run

The following will be throw away

All of the following are true

Variable	SMGR	CC	CHI	IG	GR
V003: MONTH	0.0512	291.5808	27.6286	0.0002	0.0000
V006: WEEK	0.0925	96.6702	122.0543	0.0006	0.0001
V004: DATE OF MONTH	0.1218	72.5536	155.5311	0.0008	0.0002
V126: DIR OF TRAVEL VEH	0.6071	60.6295	4404.3643	0.0204	0.0088
V008: DAY OF WEEK	0.5140	22.0698	2274.6739	0.0110	0.0039
V091: GENDER DRIVER 2	0.6094	9.7091	4569.6370	0.0221	0.0142
V017: DISTANCE TO OBJE	0.1202	3.4492	265.9933	0.0108	0.0030
V221: GENDER, PEDESTRI	0.0816	3.1836	0.4700	0.0005	0.0006
V061: DIR OF TRAVEL VEH	0.0191	2.6616	3.6342	0.0000	0.0000
V009: RURAL OR URBAN	0.4919	2.2418	2548.3847	0.0126	0.0144
V232: CLOTHING CONTRA	0.0599	1.8938	0.0736	0.0001	0.0001
V007: TIME	0.6639	1.7891	11170.8741	0.0447	0.0104
V132: TOWED VEHICLE 2	0.4707	1.6335	4750.8382	0.0232	0.0169
V145: TRAF CONT FUNCTI	0.6015	1.2703	4316.6523	0.0206	0.0136
V096: RESIDENCE LOC DR	0.6086	1.1983	4480.7045	0.0213	0.0168
V010: HIGHWAY CLASS	0.5364	1.1297	1856.0302	0.0086	0.0039
V026: GENDER DRIVER C	0.2948	1.0928	1663.8035	0.0104	0.0106
V219: AGE, PEDESTRIAN	0.3923	1.0889	18.8677	0.0240	0.0063
V041: YEAR, VEHICLE C	0.3099	1.0773	1563.9621	0.0077	0.0018
V147: TRAFFIC LANES, UN	0.6061	1.0529	4487.8659	0.0213	0.0094

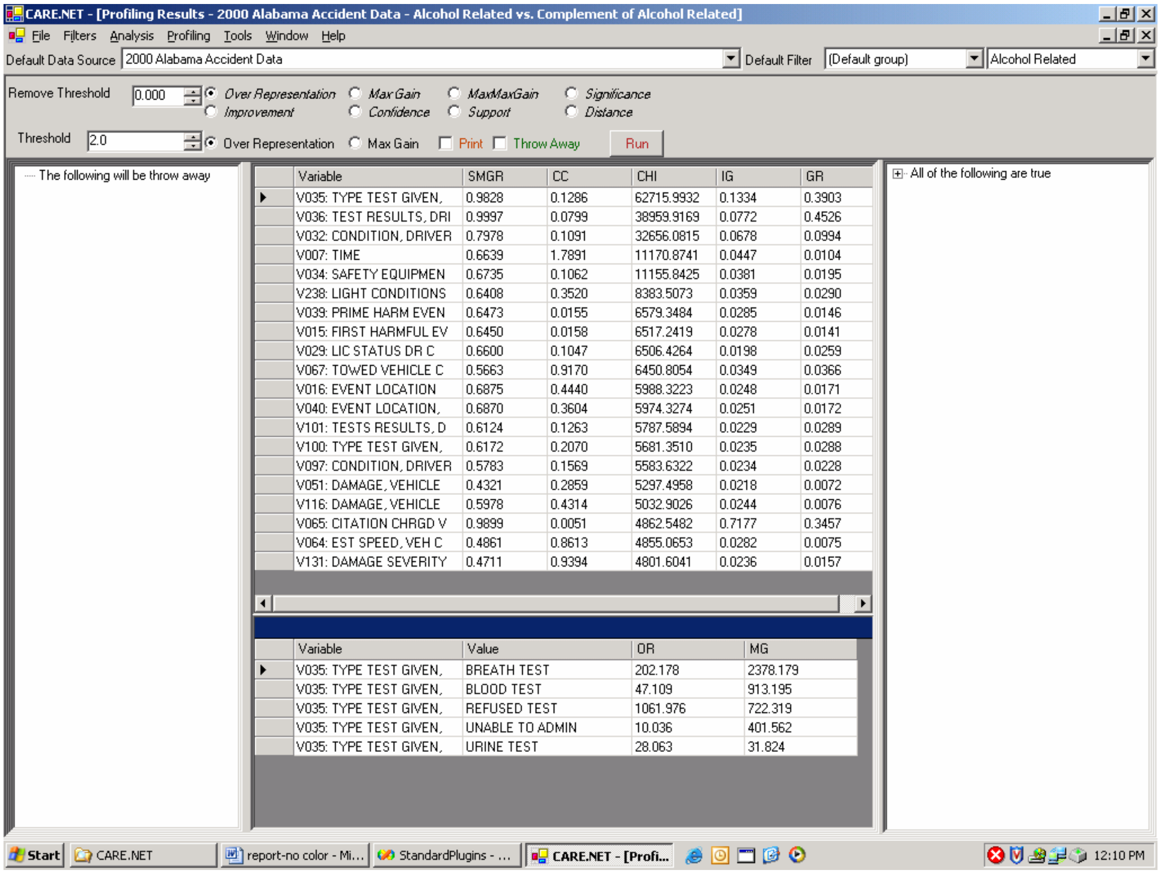
Variable	Value	OR	MG
V003: MONTH	JULY	1.092	54.431
V003: MONTH	FEBRUARY	1.086	50.350
V003: MONTH	DECEMBER	1.064	42.585
V003: MONTH	APRIL	1.038	25.399
V003: MONTH	JANUARY	1.032	20.742
V003: MONTH	MAY	1.018	12.090

Figure 8-2. Alcohol related filter, rank by CC

Table 8-3. Summary of top five variables from Figure 8-2

Ranking	Variable	Label	Description
1	V0003	Month	The month the accident occurred
2	V0006	Week	The week of the accident beginning with January 1 – January 7 as Week 1
3	V0004	Date of Month	The date of the month the accident occurred (between 1 and 31)
4	V0126	Dir of travel Vehicle 2	Indicates the direction of travel for a second vehicle involved in the accident. (North, South, East, West, Unknown)
5	V0008	Day of Week	The day of the week of the accident.

For further comparison, Figure 8-3 shows the screen when using CHI as the variable ranking method using the same dataset and filter as used in Figure 8-1 and Figure 8-2. Table 8-4 provides a summary of the top five variables from Figure 8-3.



Default Data Source: 2000 Alabama Accident Data

Default Filter: (Default group) | Alcohol Related

Remove Threshold: 0.000 ● Over Representation

Threshold: 2.0 ● Over Representation

The following will be throw away

All of the following are true

	Variable	SMGR	CC	CHI	IG	GR
▶	V035: TYPE TEST GIVEN,	0.9828	0.1286	62715.9932	0.1334	0.3903
	V036: TEST RESULTS, DRI	0.9997	0.0799	38959.9169	0.0772	0.4526
	V032: CONDITION, DRIVER	0.7978	0.1091	32656.0815	0.0678	0.0994
	V007: TIME	0.6639	1.7891	11170.8741	0.0447	0.0104
	V034: SAFETY EQUIPMEN	0.6735	0.1062	11155.8425	0.0381	0.0195
	V238: LIGHT CONDITIONS	0.6408	0.3520	8383.5073	0.0359	0.0290
	V039: PRIME HARM EVEN	0.6473	0.0155	6579.3484	0.0285	0.0146
	V015: FIRST HARMFUL EV	0.6450	0.0158	6517.2419	0.0278	0.0141
	V029: LIC STATUS DR C	0.6600	0.1047	6506.4264	0.0198	0.0259
	V067: TOWED VEHICLE C	0.5663	0.9170	6450.8054	0.0349	0.0366
	V016: EVENT LOCATION	0.6875	0.4440	5988.3223	0.0248	0.0171
	V040: EVENT LOCATION,	0.6870	0.3604	5974.3274	0.0251	0.0172
	V101: TESTS RESULTS, D	0.6124	0.1263	5787.5894	0.0229	0.0289
	V100: TYPE TEST GIVEN,	0.6172	0.2070	5681.3510	0.0235	0.0288
	V097: CONDITION, DRIVER	0.5783	0.1569	5583.6322	0.0234	0.0228
	V051: DAMAGE, VEHICLE	0.4321	0.2859	5297.4958	0.0218	0.0072
	V116: DAMAGE, VEHICLE	0.5978	0.4314	5032.9026	0.0244	0.0076
	V065: CITATION CHRGD V	0.9899	0.0051	4862.5482	0.7177	0.3457
	V064: EST SPEED, VEH C	0.4861	0.8613	4855.0653	0.0282	0.0075
	V131: DAMAGE SEVERITY	0.4711	0.9394	4801.6041	0.0236	0.0157

	Variable	Value	OR	MG
▶	V035: TYPE TEST GIVEN,	BREATH TEST	202.178	2378.179
	V035: TYPE TEST GIVEN,	BLOOD TEST	47.109	913.195
	V035: TYPE TEST GIVEN,	REFUSED TEST	1061.976	722.319
	V035: TYPE TEST GIVEN,	UNABLE TO ADMIN	10.036	401.562
	V035: TYPE TEST GIVEN,	URINE TEST	28.063	31.824

Figure 8-3. Alcohol related filter, rank by CHI

Table 8-4. Summary of top five variables from Figure 8-3

Ranking	Variable	Label	Description
1	V0035	Test Results Driver C	Indicates the results for the Blood Alcohol Test given.
2	V0036	Citation Charged, Vehicle C	Indicates that a citation was given to the causal driver
3	V0032	Condition, Driver C	Indicates the officer's assessment of the causal drivers condition to include apparently asleep, ill, fatigued, etc.
4	V0007	Time	Time of day of the accident
5	V0034	Safety Equipment, Driver C	Indicates the officer's assessment of the type of safety equipment available and used by the causal driver. This includes lap and shoulder belts and air bags.

Section 9
Conclusions

In this research, we have presented a value ranking method called Max Gain (MG). MG is a very intuitive method, in that it serves as a metric that has definitive meaning to a practitioner, particularly within the traffic safety domain discussed here. In particular, MG gives the maximum potential for reduction in crashes, given the application of a countermeasure designed to reduce crashes. Thus, MG allows the practitioner to make resource tradeoffs among countermeasures, based on real numbers.

MG is not only useful as a metric, but is also useful as a value ranking method. In particular, MG is strongly correlated with SSZ and outperforms most of the previous value ranking techniques, making it a conceptually simpler proxy for SSZ. It is also moderately correlated with SP, making it a potential proxy for SP as well.

We also present a variable ranking method called Sum Max Gain Ratio (SMGR). SMGR is derived from MG and uses Over-Represented attribute values as the primary contributing factor in variable ranking. The experiments have shown SMGR performs well at variable ranking with less run time cost than more traditional approaches, such as Chi-squared and Information Gain. In certain cases, it was empirically shown to provide a faster run time with similar variable rankings. The findings suggest that SMGR is more sensitive to the number of variables (columns) than to the number of records (rows).

In addition, we examined the performance tradeoffs between row-major order and column-major order in the context of heuristic variable selection. The research has shown that column-major order performs better than row-major order in the context of heuristic variable selection and value ranking.

Section 10
References

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of ACM SIGMOD Conference on Management of Data, Washington DC*, 207-216.

Andrews, H. C. (1971). Multidimensional rotations in feature selection. *TC, 20*(9), 1045.

Baim, P. W. (1988). A method for attribute selection in inductive learning systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 10*(6), 888-896.

Berry, M., & Linoff, G. (1997). *Data mining techniques: For marketing, sales, and customer support*. John Wiley & Sons, Inc.

Bishop, C. (1995). *Neural networks for pattern recognition*. Oxford University Press, Oxford.

Caruana, R., & Freitag, D. (1994). Greedy attribute selection. *Proceedings of 11th International Conference on Machine Learning*, 28-36.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research, 3*(Mar):1289-1305.

Foster, D. P., & Stine, R. A. (2004). Variable selection in data mining: building a predictive model for bankruptcy. *Journal of the American Statistical Association, 99*(466), 303-313.

Fukunaga, K., & Koontz, W. L. G. (1970). Application of karhunen-loeve expansion to feature selection and ordering. *TC, 19*(4), 311.

Furey, T., Cristianini, N., Duffy, N., & Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics, 16*(10), 906-914.

Golub, T. R., et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science, 286*(5439), 531-537.

Grimaldi, M., Cunningham, P., & Kokaram, A. (2003). An evaluation of alternative feature selection strategies and ensemble techniques for classifying music. *The 14th European Conference on Machine Learning and the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Dubrovnik, Croatia.

Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning, 46*(1-3), 389–422.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*(Mar), 1157-1182.

Hall, M. A., & Smith, L. A. (1999). Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper. *Proceedings of the Florida Artificial Intelligence Symposium*, Orlando, Florida, 235-239.

Hall, M. A., & Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering, 15*(6), 1437-1447.

Hawley. W. (1996). *Foundations of statistics*. Harcourt Brace & Company.

Howell, D. C. (2001). *Statistical methods for psychology* (5th ed.). Duxbury: Belmont, CA.

Kautardzic, M (2001). *Data mining – concepts, models, methods and algorithms*. IEEE Press.

Kohavi, R., John, G., & Pfleger, K. (1994). Irrelevant features and the subset selection problem. *Proceedings of 11th International Conference on Machine Learning*, New Brunswick, NJ, 121-129.

Kohavi, R., & John, G. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97*(1-2), 273-324.

Lehmann, E. L. (1959). *Testing statistical hypothesis*. Wiley, New York.

Leonard, T. (2000). *A course in categorical data analysis*. Chpman & Hall CRC.

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Mladenic, D., & Grobelnik, M. (1999). Feature selection for unbalanced class distribution and naïve bayes. *Proceedings of the 16th International Conference on Machine Learning*, 258-267.

Pappa, G. L., Freitas, A. A., & Kaestner, C. A. A. (2002). A multiobjective genetic algorithm for attribute selection. *Proceedings of 4th International Conference on Recent Advances in Soft Computing (RASC)*, 116-121.

Parrish, A., Dixon, B., Cordes, D., Vrbsky, S., & Brown, D. (2003). CARE: A tool to analyze automobile crash data. *IEEE Computer 36*(6), 22-30.

Parrish, A., Vrbsky, S., Dixon, B., & Ni, W. (2005). Optimizing disk storage to support statistical analysis operations. *Decision Support Systems Journal, 38*(4), 621-628.

Pavlidis, P., Weston, J., Cai, J., & Grundy, W. N. (2002). Learning gene functional classifications from multiple data types. *Journal of Computational Biology, 9*(2), 401-411.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Los Altos, California: Morgan Kaufmann.

Richards, W. D. (2002). *The Zen of empirical research (quantitative methods) in communication*. Hampton press, Inc.

Slonim, D., Tamayo, P., Mesirov, J., Golub, T., & Lander, E. (2000). Class prediction and discovery using gene expression data. *Proceedings of 4th Annual International Conference on Computational Molecular Biology Universal Academy Press*, Tokyo, Japan, 263-272.

Srikant, R., & Agrawal, R. (1995). Mining generalized association rules. *Proceedings of 21th VLDB Conference*, Zurich, Swizerland.

Stockburger, D. W. (1996). *Introductory statistics: Concepts, Models, and Applications*.

Stoppiglia, H., Dreyfus, G., Dubois, R., & Oussar, Y. (2003). Ranking a random feature for variable and feature selection. *Journal of Machine Learning Research, 3*(Mar), 1399-1414.

Viallefont, V., Raftery, A. E., & Richardson, S. (2001). Variable selection and Bayesian model averaging in epidemiological case-control studies. *Statistics in Medicine, 20*, 3215-3230.

Wang, H., Parrish, A., & Chen, H. C. (2003). Automated selection of automobile crash countermeasures. *Proceedings of 41th ACM Southeast Regional Conference*, 268-273.

Wang, H., Parrish, A., Smith, R., & Vrbsky, S. (2005). Variable selection and ranking for analyzing automobile traffic accident data. *Proceedings of the 20th ACM Symposium on Applied Computing, Santa Fe, New Mexico*, 36-41.

Weston, J., Elisseff, A., Scholkopf, B., & Tipping, M. (2003). Use of the zero-norm with linear models and kernel methods. *Journal of Machine Learning Research, 3*(Mar), 1439-1461.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. *Proceedings of the 14th International Conference on Machine Learning*, 412-420.

Zembowicz. R., & Zytkow, J. (1996). From contingency tables to various forms of knowledge in databases. *Advances in Knowledge Discovery and Data Mining*, AAAI Press/The MIT Press, 328-349.