Paul Helman†
Department of Computer Science
University of New Mexico
Albuquerque, New Mexico 87131

†This work was supported by a grant from the Safeguards Systems Group, Los Alamos National Laboratory, Los Alamos, New Mexico 87545.

LA-SUB--93-220

ABSTRACT

Further consequences of the inductive inference model of anomaly and misuse detection are presented. The results apply to the design of both probability models for the inductive inference framework and to the design of W&S rule bases. The issues considered include: the role of misuse models M_A, the selection of relevant sets of attributes and the aggregation of their values, the effect on a rule base of nonmaximal rules, and the partitioning of a set of attributes into a left hand and right hand side.

1. Introduction

We build on the research presented previously to establish a collection of basic principles for rule base design. The issues addressed include: the effect on the rule base of nonmaximal rules, the selection of relevant attributes, the aggregation of the values of the relevant attributes, the partition of a rule's attributes into left hand and right hand sides, abstraction of values, and the effect of competing probability models.

Section 2 presents an overview of design issues, focusing on the role of probability models in testing, and considers models for some specific types of misuse. Also considered in these terms is the relationship between anomaly and misuse detection, and the consequences of competing models. Section 3 generalizes the results of [Helm90a]. In [Helm90a] we presented a two field model which demonstrated that nonmaximal rules can always be forced to lead to inconsistencies in scoring. This result was derived for several specific M_A models, and the specific scoring function used in [Vacc89]. Section 3 generalizes this result in several important ways, demonstrating that the result holds for transactions over arbitrarily many fields, and for large and natural classes of M_A models and scoring functions. Section 4 considers the partioning of a rule's attributes into left and right hand sides. Section 5 explores a criterion for the selection of attributes and the partitioning of their values. Section 6 translates these results into a few suggested modifications and extensions of W&S.

2. An Overview of Rule Base Design Issues

In this section we present an overview of the rule base design issues considered in the remainder of the paper. Some of these issues are summarized as first presented in [Helm89,Helm90a], others are presented in refined form, and other issues are new. While this paper analyzes thoroughly several of the issues, we just scratch the surface of others; however, we feel that we have come a long way in identifying what issues are important and in developing techniques for addressing these issues.

Our primary vehicle for studying the rule base design issues of interest is the inductive inference-based hypothesis testing model for anomaly and misuse detection developed in [Helm89]. This model is useful because it provides a framework for studying rigorously design questions pertaining directly to W&S and similar systems.

The subsections that follow review and refine the components of our hypothesis testing model, consider the relationships between our model and W&S, and summarize what what we believe to be the fundamental design issues.

2.1 The Role of the Models M_N and M_A

As before [Helm89,Helm90a], we begin by assuming that one of two probabilistic processes, M_N (the "good", normal process) or M_A (the "bad", misuse process) has generated the transaction under consideration. The problem is to rank incoming transactions based on the likelyhood that a given transaction was generated by M_A.

Before reviewing the details of our testing procedure, we make two remarks reflecting recent research:

likely were generated by some specific process of concern. In Section 2.3 we view anomaly detection as a special case of misuse detection, a case in which the model M_A has special characteristics. We note that from this perspective, results in this paper and in [Helm89,Helm] address the general problem of misuse detection. As such, M_A is perhaps bad notation (the A originally was meant to stand for anomalous; M_M, for misuse, is now seen to be more appropriate); however, for consistency with our previous work, we shall continue to use the notation M_A.

2. To simplify our introductory discussions, we assume only two possible models, M_N and M_A. However, as the current paper argues, the concept of competing models (for both M_N and M_A) is extremely important. In the framework of competing models, we perform testing of each transaction under alternative models, and use techniques for combining the results which the models yield. For example, we might compute a linear combination of the results yielded by competing models, where the weights in the combination are adjusted by means of learning. More on this perspective in Section 2.4.

With these remarks in mind, we review now the details of the proposed transaction testing mechanism. Suppose we were able to compute, for each transaction t, $Pr\{t|M_N\}$ and $Pr\{t|M_A\}$. In this case, we could use Bayes' formula to test the hypothesis: "*The normal process M_N generated t.*"

$$Pr\{M_N|t\} = \frac{Pr\{t|M_N\}*Pr\{M_N\}}{Pr\{t|M_N\}*Pr\{M_N\}+Pr\{t|M_A\}*Pr\{M_A\}}$$

One issue that arises from the use of this formula is that the *a priori* values $Pr\{M_N\}$ and $Pr\{M_A\}$ are required. While in some contexts it is reasonable to estimate these values, in others, it's not. However, often it suffices to consider the ratio

$$R(t) = \frac{Pr\{t|M_N\}}{Pr\{t|M_A\}}.$$

Since the value $Pr\{M_N|t\}$ is monotonic in $R(t)$, $R(t)$ is all that is required to rank incoming transactions according to their level of alarm.

The above formulae highlight the importance of the model M_A to the hypothesis testing problem. In particular, when considering the general misuse (as opposed to anomaly) detection problem, it does not suffice to identify rare transactions. Intuitively, rare is a relative quality and, almost always, individual transactions will be rare. The transactions of interest are those that are rarer under M_N than M_A. Consequently, it seems highly desirable to hypothesize and study various models for M_A. Additionally, it seems highly desirable that a detection system be parameterable to the M_A models of interest (and to the *a priori* values we associate with them) for given applications.

We propose for consideration several simple, sample M_A models. These models include:

Independence model $M_{A(I)}$: Attributes take on values with the same distribution as in M_N, but associations break down. Thus:

$$Pr\{t[B_1]=v_1, \ldots, t[B_k]=v_k|M_{A(I)}\} = \prod_{i=1}^{k} Pr\{t[B_i]=v_i|M_N\}.$$

Constant value model $M_{A(C)}$: Probabilities are calculated as a function of domain sizes. Domain sizes can be assigned by the expert, or can be inferred from observations of the historical database. (See [Helm90a] for more details.) In either case, frequencies observed in the historical database play no role. This class of models is intended to reflect misuse characterized by the user performing seemingly (with respect to normal behavior) haphazard actions.

Masquerader model $M_{A(Q)}$: Associations between attributes in $M_{A(Q)}$ are like those in M_N, but with some values "crossed over." For example:

$$Pr\{t[User]='Smith' \wedge t[Port]='tty9' \wedge t[Process]='rm'|M_{A(Q)}\}$$
$$Pr\{t[User]='Jones' \wedge t[Port]='tty9' \wedge t[Process]='rm'|M_N\}$$

This model seems to hold a good deal of promise, and we propose to investigate it further in the future.

Negative correlation model $M_{A(A)}$: Probabilistic quantities in $M_{A(A)}$ are inversely related to corresponding quantities in M_N. In particular, this is the family of models characterized by the property that whenever

We now identify a general class of M_A models that is studied in Section 3 with regard to rule base design. We first need the following definitions.

Definition: If DB is a historical database of transactions, $EPr\{S=t[S] \mid M_N \wedge DB\}$ denotes the estimate of $Pr\{S=t[S] \mid M_N\}$ obtained by sampling DB. That is,

$$EPr\{S=t[S] \mid M_N \wedge DB\}=\frac{number\ of\ transactions\ t\ in\ DB\ with\ t[S]=S}{number\ of\ transactions\ in\ DB}.$$

Similarly, $EPr\{S=t[S] \mid W=t[W] \wedge M_N \wedge DB\}$ denotes the estimate of $Pr\{S=t[S] \mid W=t[W] \wedge M_N\}$ obtained by sampling DB. That is,

$$EPr\{S=t[S] \mid W=t[W]\ M_N \wedge DB\}=\frac{number\ of\ transactions\ t\ in\ DB\ with\ t[S]=S \wedge W=t[W]}{number\ of\ transactions\ in\ DB\ with\ t[S]=S}.$$

When we say that a value is computed **against DB**, we shall mean that estimates of the above form are used in the computation. Quantities of the form $EPr\{S=t[S] \mid M_A \wedge DB\}$ and $EPr\{S=t[S] \mid W=t[W] \wedge M_A \wedge DB\}$ denote probabilities for M_A that are derived from probabilities computed against DB.

Definition: The model M_A is consistent with **piecewise sampling** (we shall say simply that M_A is a **piecewise model**) if it obeys the following condition.

Suppose that A is the set of transaction attributes, and that DB_1 and DB_2 are historical databases, such that for every $S \subseteq A$ we have that for transaction t:

$$EPr\{S=t[S] | M_N \wedge DB_1\} \leq EPr\{S=t[S] | M_N \wedge DB_2\}.$$

Then it is the case that

$$EPr\{S=t[S] | M_A \wedge DB_1\} \leq EPr\{S=t[S] | M_A \wedge DB_2\}.$$

It is fairly easy to demonstrate that independence and constant value models are piecewise. Negative correlation models clearly are not piece-wise. It remains an open question what types of masquerader models are piecewise and what types are not.

One reason for interest in piecewise models is that the results first presented in [Helm90a] and generalized in Section 3 of the current paper demonstrate a sense in which W&S is not consistent with any piecewise model, if W&S's rule base includes nonmaximal rules. There are several interpretations of this result which are explored at various points throughout the paper. At this juncture, we make the following general comments regarding what we hope to gain from the type of analysis performed in Section 3.

a) One of the primary goals of our research is the identification of criteria for pruning and searching the space of probability models (rule bases, in W&S terms). The results of Section 3 support the pruning of rule bases containing nested rules, whenever we are attempting to detect a type of misuse reflected in a piecewise model.

b) More generally, if W&S wishes to approximate the testing supported by the inductive inference framework, for some specific class of misuse models, the type of analysis performed in Section 3 can be used to "parameterize" W&S (i.e., by designing its rule base) so as to make it consistent with the models of interest.

c) Additionally, it has been established experimentally that W&S performs quite well in many environments. The type of analysis performed in Section 3 also can be used to discover characteristics of the M_A model that W&S is implicitly assuming. For example, it may be that W&S performs well in anomaly detection. If we can learn the details of the negative correlation model with which W&S is consistent, we could then add such a model to our collection of competing models to be used in the inductive inference framework.

2.2 Selection of Attributes and the Aggregation of their Values

The identification of models for M_A is one critical aspect of the design and application of transaction testing. A form of a rule base (i.e., rule bases containing nonmaximal rules) which should be avoided in many contexts. Section 4 proposes a further ning of the solution space based on an ana s of the effects of multiple rules that are different partionings o a left hand side (LHS) and a right hand s (RHS)) of the same collection of attributes and values.

In the current section, we introduce three additional aspects of the design problem: The selection of attributes to be used in the estimation of the probability distribution of transactions; the aggregation of attribute values; and, a proposed criterion for addressing these design problems. Section 5 presents a more rigorous analysis of the proposed criterion.

The Attribute Selection Problem

If we could compute accurately for every transaction t the ratio $R(t)$ (with respect to one or more models for M_A), we would have a method of ranking transactions by their level of suspiciousness. Unfortunately, $R(t)$ cannot be estimated well empirically from a finite historical database when t is rare. This is likely to be the case when transactions have many attributes, when attribute domains are large, and, of course, when one or more attributes form a key (making each transaction unique).

Our solution to this problem is to search for subsets $B = \{B_1, \ldots, B_k\}$ of the set of all attributes $A = \{A_1, \ldots, A_N\}$ such that:

(1) The quantity $Pr\{t[B_1], \ldots, t[B_k] | M_N\}$ is accurately reflected in the historical database, our sample population.

(2) The behavior of $R(t[B_1], \ldots, t[B_k])$ reflects that of $R(t)$.

In effect, we approximate the hypothesis test

$$Pr\{M_N | t\}$$

with the hypothesis test

$$Pr\{M_N | t[B_1], \ldots, t[B_k]\}.$$

Because of the exponential number of candidate subsets, we require heuristics to lead us to the most promising subsets.

Aggregation of Values

Additionally, it might be beneficial to apply **aggregation** (also called **abstraction**) to the values that the attributes can take on. This leads us to consider quantities of the form:

$$Pr\{t[B_1] \in X_1, \ldots, t[B_k] \in X_k | M_N\},$$

for *sets* X_i of values. We shall argue in Section 5 that using aggregations of attribute values, rather than single values, may be a reasonable method for obtaining "good" tests which are likely to apply often to random transactions.

Selection Criterion

Attribute selection and value aggregation are two dimensions of choice in the design of probability models and, of course, in the design of rule bases for W&S.

In Section 5 we propose as a heuristic measure of a candidate solution how well it separates the competing models (i.e., M_N from some model for M_A). If we expect a given probabilistic quantity to differ significantly under the two models of interest, it is worthy of consideration because it *potentially* can yield much information. As is explored later:

a) Methods are required (e.g., experimental feedback) to assess the quality of candidates chosen by this heuristic. In particular, model separation is not sufficient to ensure that a candidate solution provides good the measure are difficult (probably NP-hard). Hence, search heuristics are required to build the candidate solutions.

2.3 Design Issues in the Specific Context of Anomaly Detection

In order to present a simple illustration of the attribute selection and value aggregation problems, and in order to view anomaly detection in the framework we have presented, we summarize briefly some recent work with G. Liepins [Helm90b]. As observed earlier, [Liep90] defines anomaly detection, as opposed to misuse detection, as the identification of rare transactions. This is viewed as a special case of the misuse detection problem studied in this paper, where a negative correlation model is used for M_A.

We observe that we can describe anomaly detection, without explicit reference to an M_A model, simply as the identification of those transactions t such that $Pr\{t \mid M_N\}$ is smallest. Notice that, in the specific context of anomaly detection, rareness relative to M_N suffices, because the characterizing property of negative correlation models $M_{A(A)}$ is that $Pr\{t \mid M_N\} < Pr\{t' \mid M_N\}$ implies $Pr\{t \mid M_{A(A)}\} > Pr\{t' \mid M_{A(A)}\}$. Hence, "$t$ rarer than t' in M_N" implies "$R(t) < R(t')$ with respect to any negative correlation model."

Despite this simpler statement of anomaly detection, the phenomenon described in Section 2.2 which leads to the problems of attribute selection and value aggregation is present in anomaly detection as well. That is, $Pr\{t \mid M_N\}$ cannot be estimated well empirically from a finite historical database, when t is rare. Thus, while it should not be too surprising that the naive detector [Liep90] worked well relative to W&S when the empirical sampling is accurate (because of only two fields and two values), in the general case where transactions are expected to be rare, we require estimation techniques similar to those described in Section 2.2.

Helman and Liepins [Helm90b] have proposed that, in the context of anomaly detection, attribute selection and value aggregation be stated as finding probabilistic quantities (e.g., $Pr\{t[B_1], \ldots, t[B_k] \mid M_N\}$) that accentuate the tails of the distribution under sampling (i.e., by peaking the distribution), while preserving the relative frequencies of the transactions. This appears to be a special case of the model separation criterion proposed in Section 2.2, and we propose to investigate the relationship between heuristics for the more general problem and the anomaly detection problem.

2.4 Competing Models and Pruning the Search Space

Observe that if we somehow knew that a given selection of attributes and aggregation of their values were optimal, and if we knew the M_A model we were attempting to detect, we would implement a single test of each transaction, based on a single ratio $R(t)$. Similarly, under this assumption, W&S's rule base would contain a single rule.

In reality, however, we have only conjectures as to good selections of attributes, aggregation of values, and models of M_A. To handle this uncertainty, we propose a general framework in which a single transaction t is analyzed simultaneously under many competing models for M_N and M_A. That is, we compute for t many $R(t)$ based on different, selections, aggregations, and models for M_A. We then (for example) take a linear combination of the results. The learning problem, which does require some expert feedback, adjusts the weights attached to the existing models, and creates new models when appropriate. From this perspective, a W&S rule base is a collection of competing probability models, and its scoring functions (e.g., TFOM) are methods of combining the results yielded by the competing models.

The main thrust of our research is to develop techniques for determining which competing models to build and include in our tests. Because of the enormity of the search space, we require: (1) Heuristics for pruning the search space before any optimization is performed; (2) Heuristic measures which can be evaluated very quickly and indicate which candidates should be explored further; and (3) Search heuristics which explore, based on the heuristic measures, the solution space resulting from the pre-search pruning. Examples of heuristic pruning are suggested by the results of Sections 3 and 4. Note that such pruning eliminates entire classes of models, based on heuristic criteria. Section 2.2 introduced informally a heuristic measure which is explored more thoroughly in Section 5. We are just beginning to investigate search heuristics in the framework of the problem that has resulted from this work.

3. The Effect of Nonmaximal Rules

A recent change in the implementation of W&S was to disallow rules over attribute sets which are nested by el-

In particular, suppose that, through experimentation and learning, we discover that the Bayesian hypothesis test should be performed with respect to a specific set A of attributes, and that one or more piecewise models should be used for M_A. We demonstrate that when a W&S rule base contains any rule which does not reference all the attributes in A, W&S scoring is not consistent with the hypothesis testing framework outlined in the previous section. We interpret this result to mean that: (1) If a W&S rule base is to approximate a final, static probability model, and if M_A is piecewise, then the rule base should contain only rules which reference all the relevant attributes since, otherwise, W&S scoring in some cases will be inconsistent with our hypothesis testing model; (2) While we cannot say that it is always detrimental to have nested rules in an initial rule base that is to undergo modification in response to learning, the fact that the rules over nested sets necessarily give conflicting information is construed as strong evidence for pruning from the search space rule bases that contain nested rules.

Suppose that $A=\{A_1, \ldots, A_N\}$ is the set of attributes that is used to approximate the distribution of transactions in M_N and M_A. That is, we compute

$$Pr\{M_N | t\}$$

as

$$Pr\{M_N | t[A_1],...,t[A_N]\}.$$

A is said to be the **relevant** set of attributes. Suppose the W&S rule base RB contains a rule

$$R: (B_1=v_1) \cdots (B_k=v_k) \rightarrow (B=r), \text{ where}$$

$\{B, B_1, \ldots, B_k\}$ is a proper subset of A. Rule R is said to be **nonmaximal**.

We shall demonstrate that, relative to piecewise M_A models, "W&S-like" scoring functions exhibit certain types of inconsistencies whenever the rule base contains one or more nonmaximal rules. To this end, let R be the nonmaximal rule above, and B' a relevant attribute not appearing in R. Consider a fixed transaction $\bar{t}$ which fails the nonmaximal rule R (i.e., $\bar{t}[B_i]=v_i$, $1 \le i \le k$, $\bar{t}[B]=v \ne r$). We shall establish a sense in which $\bar{t}$ is scored in a manner inconsistent with the Bayesian hypothesis test. The proof strategy is as follows. In the next subsection we introduce pairs (DB_1, DB_2) of historical databases, such that DB_2 is obtained from DB_1 by means of a simple transform, and compare the Bayesian test of a transaction against DB_1 with the Bayesian test against DB_2. The following subsection exhibits the inconsistency by performing the same analysis on a class of W&S-like scoring functions, against a rule base containing nonmaximal rules.

3.1 A Class of Database Transforms and their Effect on the Bayesian Measure

We consider a class of simple transforms to historical databases.

> **Definition:** Let t be a fixed transaction, and A_i, A_j a pair of relevant attributes. Historical database DB_2 is obtained from historical database DB_1 by means of a **simple** (t, A_i, A_j)**-transform** if every t' that differs between the DB_1 and DB_2 is such that in DB_1 t' agrees with t everywhere but on A_i, A_j, and in DB_2 t' agrees with t everywhere but on A_j.

In the following analysis, we shall be concerned exclusively with pairs (DB_1, DB_2) of databases such that DB_2 is obtained from DB_1 by means of a simple $(\bar{t}, B, B')$ transform, where $\bar{t}, B$, and B' are as defined previously. In this case, we shall say simply that (DB_1, DB_2) is a simple **transform pair.**

> **Theorem 3.1:** Let (DB_1, DB_2) be a simple transform pair, and suppose that the Bayesian hypothesis test is performed with respect to some piecewise model for M_A. Then
>
> $$(Pr\{M_N | \bar{t}\} \text{ computed against } DB_1) \ge (Pr\{M_N | \bar{t}\} \text{ computed against } DB_2).$$
>
> That is, $\bar{t}$ is ranked as being more suspicious in DB_2 than in DB_1.
>
> **Proof:** Observe first that
>
> $$(Pr\{\bar{t} | M_N\} \text{ computed against } DB_1) = (Pr\{\bar{t} | M_N\} \text{ computed against } DB_2),$$

measure is monotonic in the ratio ... □

The significance of this ɔrem is that when RB contains one or mc onmaximal rules, transform pairs (DB_1, DB_2) can be constructed so that any scoring function in the class defined in the following subsection ranks $\bar{t}$ such that:

$$(Score(\bar{t}) \text{ computed against } DB_1) > (Score(\bar{t}) \text{ computed against } DB_2).$$

That is, contrary to the Bayesian hypothesis test (assuming a piecewise model for M_A), such scoring functions rank $\bar{t}$ as being more suspicious relative to historical database DB_1 than relative to historical database DB_2.

3.2 A Class of Rule Base Scoring Functions

We now define a large class of transaction scoring functions which includes most imaginable variations to the W&S scoring function of [Vacc89].

Generally speaking, a scoring function is a function of the transaction t being scored, the rule base RB, and the database DB of historical transactions. Property 1 restricts the manner in which a scoring function may depend on DB.

Property 1: $Score(t)$, computed against RB and DB, is a function

$$f(t, RB, (R_1, a_1, b_1), \ldots, (R_n, a_n, b_n)),$$

where $R_1, \ldots, R_n$ are exactly the rules in RB fired by t, a_i is the number of transaction in DB which fire R_i, and b_i is the number of transactions in DB which fail rule R_i.

Notice that a scoring function obeys this property if $Score(t)$ depends on DB as a function of only the conditional probability

$$Pr\{B=v \mid (B_1=v_1) \wedge \cdots \wedge (B_k=v_k)\},$$

for each rule

$$(B_1=v_1) \wedge \cdots \wedge (B_k=v_k) \rightarrow (B=v)$$

that t fires.

Property 2 stipulates that the scoring function interprets in the natural direction a change in any rule's (#fired/#failed) ratio. We first need the following definition.

Definition: A collection of #fired,#failed values for some or all of the rules in a rule base is **consistent** if there exists a historical database yielding these values. A collection of values that is inconsistent, but in which #fired ≥ #failed for each rule, is called a **pseudo-database**.

In what follows, as a definitional convenience, we assume that a scoring function is defined (as an abstract function) over pseudo-databases, as well as over consistent #fired,#failed values.

Property 2: $Score(t)$ is **monotonic** in the following sense. Let R be any rule and t any transaction. If the (#fired R,#failed R) values are varied while the #fired,#failed values for all other rules are held fixed, then:

If t fails R, $Score(t)$ is nondecreasing as the ratio (#fired R / #failed R) is increased.

If t passes R, $Score(t)$ is nonincreasing as the ratio (#fired R / #failed R) is increased.

In the special case where #failed R is 0, $Score(t)$ is nondecreasing as #fired R is increased, if t fails R; if t passes R, $Score(t)$ is nonincreasing.

In the special case where #failed R = #fired R, $Score(t)$ is nonincreasing as #fired R and #failed R are increased by the same amount, if t fails R; if t passes R, $Score(t)$ is nondecreasing.

...ed) values of any particular rule. Before presenting the condition formally, we require the following definition.

Definition: Let V be a vector of fixed (#fired,#failed) values for some collection C of the rules in RB. A **δ-neighborhood** $\delta(V)$ is the set of vectors V' specifying (#fired,#failed) values to this same collection C of rules such that if the triple (R,a,b) appears in V then the triple (R,a',b') appearing in V' is such that

$$|a'-a|<\delta \text{ and } |b'-b|<\delta$$

The required sensitivity condition is as follows.

Property 3:

3.1: *Sensitivity to change within a constant-size neighborhood is bounded uniformly.*

Let A_i and A_j be any attributes in the relevant set A and let t be any transaction. Let V be a fixed vector of consistent values for all rules that t fires which contain A_i in the LHS and A_j in the RHS. Then for every constant δ, there exists an ε such that for every vector X of values for the remaining rules,

$$|Score(t,V,X)-Score(t,V',X)|<\varepsilon,$$

for all $V' \in \delta(V)$.

3.2: *Sensitivity to unbounded change is unbounded.*

Let A_i be any attributes in the relevant set A, t be any transaction, and Y be a fixed vector of consistent values for some subset of the rules with A_i in the RHS, such that this subset does not include at least one such rule failed by t. Let $RP_1, \ldots, RP_p$ and $RF_1, \ldots, RF_f$ $(f \geq 1)$ be the rules with A_i in the RHS respectively passed and failed by t which are *not* assigned a value by Y. Then for every constant Δ, there exists a constant δ such that for every assignment Z to the rules not containing B in the RHS:

$$Score(t,(RF_1,a,b),\ldots,(RF_f,a,b),(RP_1,a,a-b),\ldots,(RP_p,a,a-b),Y,Z) - Score(t,(RF_1,a',b'),\ldots,(RF_f,a',b'),(RP_1,a',a'-b'),\ldots,(RP_p,a',a'-b'),Y,Z) > \Delta,$$

whenever $\frac{a}{b} - \frac{a'}{b'} > \delta$ and $a,b,a',b' > 0$.

Theorem 3.2: The W&S scoring function TFOM in [Vacc89] obeys Properties 1-3.
Proof: It is obvious that the function depends on the historical DB only as permitted and is monotonic. We now verify that the two sensitivity conditions are obeyed.

3.1. Sensitivity to change within a constant-size neighborhood is bounded uniformly.

Let B and B' be the attributes in question, t the transaction in question, $RP_1, \ldots, RP_n$ the rules with B in LHS and B' in RHS that t passes, $RF_1, \ldots, RF_m$ the rules with B in LHS and B' in RHS that t fails, $OP_1, \ldots, OP_p$ rules without B in LHS and with B' in RHS that t passes, $OF_1, \ldots, OF_q$ the rules without B in LHS and with B' in RHS that t fails.

By inspection of the scoring function, the values associated with these rules influence only the $FOM_{B'}$ component. $FOM_{B'}$ can be written as:

$$\frac{\sum_{i=1}^{m} Score(RF_i)}{\left(\sum_{i=1}^{n} Score(RP_i)+\sum_{i=1}^{m} Score(RF_i)+\sum_{i=1}^{p} Score(OP_i)+\sum_{i=1}^{q} Score(OF_i)\right)^{1/2}} + \frac{\sum_{i=1}^{q} Score(OF_i)}{\left(\sum_{i=1}^{n} Score(RP_i)+\sum_{i=1}^{m} Score(RF_i)+\sum_{i=1}^{p} Score(OP_i)+\sum_{i=1}^{q} Score(OF_i)\right)^{1/2}}.$$

greater than

$$|\frac{\sum_{i=1}^{m} Scor\quad\;_i)}{(\sum_{i=1}^{n} Score(RP_i)+\sum_{i=1}^{m} Score(RF_i))^{1/2}} - \frac{\sum_{i=1}^{m} Score(RF_i) \;+}{(\sum_{i=1}^{n} Score(RP_i)+\sum_{i=1}^{m} Score(RF_i) \;+k_2)^{1/2}}|,$$

where k_1 and k_2 are constants depending only on V and δ. Since the values $Score(RF_i)$ and $Score(RP_i)$ depend only on V and not X, there is a bound on the term's change that is valid for all values X.

(b) The contribution to the difference between $Score(t,V,X)$ and $Score(t,V',X)$ of the second term is no greater than

$$|\frac{\sum_{i=1}^{q} Score(OF_i)}{(\sum_{i=1}^{q} Score(OF_i))^{1/2}} - \frac{\sum_{i=1}^{q} Score(OF_i)}{(\sum_{i=1}^{q} Score(OF_i) \;+k)^{1/2}}|,$$

where k is a constant depending only on V and δ. While $\sum_{i=1}^{q} Score(OF_i)$ can be made arbitrarily large by varying X, both $\frac{\sum_{i=1}^{q} Score(OF_i)}{(\sum_{i=1}^{q} Score(OF_i))^{1/2}}$ and $\frac{\sum_{i=1}^{q} Score(OF_i)}{(\sum_{i=1}^{q} Score(OF_i) \;+k)^{1/2}}$ approach $(\sum_{i=1}^{q} Score(OF_i))^{1/2}$ as $\sum_{i=1}^{q} Score(OF_i)$ grows. Hence, the sensitivity of this term to a change from V to V' decreases as $\sum_{i=1}^{q} Score(OF_i)$ is increased. Therefore, there is a bound on the term's change that is valid for whatever value X induces in $\sum_{i=1}^{q} Score(OF_i)$.

3.2. Sensitivity to unbounded change is unbounded.

Let B be the attribute in question, t the transaction in question, and Y the fixed set of triple values. By inspection of the scoring function, changes to the a/b ratios of the rules RF_i and RP_i influence only the FOM_B component. Let c be the largest ratio in the given Y. Write FOM_B as a sum of terms: one term for each failed B-rule assigned a value by Y and a single term for the remaining failed B-rules. (A B-rule is any rule with B in the RHS.)

The largest term among those for the failed B-rules assigned a value by Y is no larger than

$$\frac{c}{(sum\ grades\ of\ all\ B{-}rules)^{1/2}}.$$

Hence, when we change values, the largest change to the overall function contributed by each of these terms is not more than c (the denominator is never less than 1), a constant depending only on Y (and not on Z,a,b,a',b').

Let f be the number of failed B-rules not assigned a value by Y and p the number of passed B-rules not assigned a value by Y. When these rules are assigned the (fired, failed) values of (a,b) and $(a,a-b)$ respectively, the contribution of these rules to FOM_B is:

$$\frac{f*\frac{a}{b}}{(f*\frac{a}{b}+p*\frac{a}{(a-b)}+k)^{1/2}},$$

where k is a constant depending only on Y (and not on Z). For sufficiently large values of $\frac{a}{b}$ this can be made arbitrarily close to $(f*\frac{a}{b})^{1/2}$. When values (a',b') and $(a',a'-b')$ are used, the contribution of this large enough to make $\frac{\ldots}{(f^*\frac{a}{b}+p^*\frac{a}{(a-b)}+k)^{1/2}}$ arbitrarily close to $(f^*\frac{\ldots}{b})^{1/2}$, and to make $(f^*\frac{a}{b})^{1/2} - (f^*\frac{a'}{b'})^{1/2}$ arbitrarily large. □

Observation: It appears, based on an informal description of the scoring function used in the current implementation of W&S, that this function obeys Properties 1-3 as well. However, more detailed information on this function is required before a formal analysis can be performed.

3.3 An Inconsistency Result

In order to establish inconsistency when a nonmaximal rule is present, we consider 3 classes of rules. Nonmaximal rule R, attributes B and B', value v, and transaction $\bar{t}$ are as defined at the beginning of Section 3.

Rule Classes

WR(a): $\bar{t}$ fires rule, $(B=v)$ appears in LHS, B' does not appear
WR(b): $\bar{t}$ fires rule, B appears in RHS, B' does not appear
RR: $\bar{t}$ fires rule, $(B=v)$ appears in LHS, B' appears in RHS

Observe that by the construction of $\bar{t}$, rule R is in class WR(b). Observe also that of the 3 classes, only RR can contain maximal rules.

Theorem 3.3: If DB_2 is obtained from DB_1 by a simple transform, then every rule fired by $\bar{t}$ whose (#fired,#failed) values differ between DB_1 and DB_2 is in one of the 3 classes WR(a), WR(b), or RR.
Proof: Consider any rule whose (#fired,#failed) values differ between DB_1 and DB_2.

(a) If B does not appear anywhere in the rule, it is clear that the rule's (#fire,#failed) values do not change between DB_1 and DB_2. Hence, B appears in the rule.

(b) If any $(C\neq\bar{t}[C])$ appears in the LHS, then $\bar{t}$ does not fire the rule. Hence, if B appears in the LHS, it must appear as $(B=\bar{t}[B])$.

(c) By the previous observation, if B' appears in the LHS, it must appear as $(B'=\bar{t}[B'])$. But if $(B'=\bar{t}[B'])$ appears in the LHS, no t that changes between DB_1 and DB_2 can fire the rule, either before or after t is changed. Hence, the (#fired,#failed) values are unchanged for such a rule. Thus, if B' appears in the rule, it must appear in the RHS. □

Consider intuitively the effect of a change from DB_1 to DB_2 on the #fired/#failed ratio of rules in the 3 classes. When DB_1 is changed to DB_2 by a simple transform, the ratio associated with each rule in WR(a) and WR(b) increases (or is unchanged) for rules passed by $\bar{t}$ and decreases (or is unchanged) for rules failed by $\bar{t}$. For the class of scoring functions defined in Section 3.2 (actually for any function obeying Properties 1 and 2), this implies that such a rule's influence on the scoring of $\bar{t}$ is to rank $\bar{t}$ less suspicious with respect to DB_2 than DB_1, a conclusion inconsistent with the Bayesian hypothesis test. On the other hand, the ratio for rules in RR behave in the opposite manner, and hence such a rule's influence on the scoring of $\bar{t}$ is consistent with the Bayesian hypothesis test.

In order to demonstrate that an inconsistency in the scoring of $\bar{t}$ over all rules is always possible, we show that simple transform pairs (DB_1,DB_2) always can be constructed so that the effects of WR(a) and WR(b) is arbitrarily stronger than that of RR. Many constructions suffice to demonstrate this; we exhibit here a construction that is easy to describe and analyze, though the distribution of values in the database is perhaps a bit unnatural. It should not be construed, however, that only unnatural distributions exhibit these inconsistencies.

We consider a family of historical databases (that will play the role of DB_1 in transform pairs (DB_1,DB_2)) defined by three distribution parameters n_1,n_2 and n_3, whose meaning is given as follows. For every $S\subseteq A'$, where $A'=A-\{B,B'\}$:

$|DB_1|=n_1$
$\#(t \supset (t[S]=\bar{t}[S]))=n_1$
$\#(t \supset (t[B]=\bar{t}[B])$ and $(t[S]=\bar{t}[S]))=n_2$

$\#(t \Rightarrow (t[B] \neq \bar{t}[B])$ and $(t[B'] \neq \bar{t}[B'])$ and $(t[S] = \bar{t}[S])) = n_1 - (n_2 + n_3)$

Figure 1. Form of the family of historical databases.

The following theorem is this section's main result.

Theorem 3.4: Suppose that scoring function Score satisfies Properties 1-3 and that RB contains one or more nonmaximal rules. Then for every constant D, there exists transactions $\bar{t}$ and transform pairs (DB_1, DB_2) such that

$$(Pr\{M_N | \bar{t}\} \text{ computed against } DB_1) \geq (Pr\{M_N | \bar{t}\} \text{ computed against } DB_2),$$

yet

$$Score(\bar{t}, RB, DB_1) - Score(\bar{t}, RB, DB_2) > D,$$

where the Bayesian hypothesis test is computed with respect to any piecewise model for M_A.
Proof: Let the nonmaximal rule (with respect to relevant attribute set A) in RB be

R: $(B_1=v_1)...(B_k=v_k) \rightarrow (B=r)$.

Let $\bar{t}$ be any transaction which fires and fails R, and let B' be an attribute that does not appear in R. As before, define $A' = A - \{B, B'\}$. Distribution parameters n_1, n_2, and n_3 have the meaning given above.

The construction of the required pair DB_1 and DB_2 is specified in the following steps.

A. Let DB be the historical database defined by any integral values $\bar{n}_1, \bar{n}_2, \bar{n}_3$ for the distribution parameters n_1, n_2, n_3 such that $0 < \bar{n}_3 < \bar{n}_2 < \bar{n}_1$.

In DB, the values for the rules fired by $\bar{t}$ thus are as follows, where S is any subset of A'. (The values are listed as (#fired,#failed) pairs.)

$(S=\bar{t}[S]) \rightarrow (B=\bar{t}[B])$: $(\bar{n}_1, (p-\bar{n}_2))$
$(S=\bar{t}[S]) \rightarrow (B \neq \bar{t}[B])$: $(\bar{n}_1, (\bar{n}_2))$

$(S=\bar{t}[S])(B'=\bar{t}[B']) \rightarrow (B=\bar{t}[B])$: $(2*\bar{n}_3, \bar{n}_3)$
$(S=\bar{t}[S])(B'=\bar{t}[B']) \rightarrow (B \neq \bar{t}[B])$: $(2*\bar{n}_3, \bar{n}_3)$

$(S=\bar{t}[S])(B=\bar{t}[B]) \rightarrow (B'=\bar{t}[B'])$: $(\bar{n}_2, (\bar{n}_2-\bar{n}_3))$
$(S=\bar{t}[S])(B=\bar{t}[B]) \rightarrow (B' \neq \bar{t}[B'])$: $(\bar{n}_2, \bar{n}_3)$

$(S=\bar{t}[S]) \rightarrow (B'=\bar{t}[B'])$: $(\bar{n}_1, (\bar{n}_1-\bar{n}_3))$
$(S=\bar{t}[S]) \rightarrow (B' \neq \bar{t}[B'])$: $(\bar{n}_1, \bar{n}_3)$

For any C different from B and B',

$(S=\bar{t}[S])(B=\bar{t}[B]) \rightarrow (C=\bar{t}[C])$: $(\bar{n}_2, 0)$
$(S=\bar{t}[S])(B=\bar{t}[B]) \rightarrow (C \neq \bar{t}[C])$: $(\bar{n}_2, \bar{n}_2)$

(Omitted here and from future discussions are the values of some additional rules not containing B' in the RHS (i.e., rules containing neither B nor B', rules containing both B' and B' on the LHS, and rules containing B' on the LHS and not containing B). It is easy to see that the values of these omitted rules do not change between DB_1 and DB_2 and have no effect on the relationships established in what follows.)

B. Select any integer constant $\delta_1 > 0$. Let V be the values specified in DB for the collection of rules fired by $\bar{t}$ containing B in the LHS and B' in the RHS. Find a bounding ε with respect to neighborhood $\delta_1(V)$, whose existence is guaranteed by Property 3.1.

C. Let Y be the triple values assigned in DB to B-rules containing B' on the LHS and let $\Delta > D + \varepsilon$. Find

$$- Score(t,(RF_1,a',b'),\ldots,(RF_f,a',b'),(RP_1,a',a'-b'),\ldots,(RP_p,a',a'-b'),Y,Z) > \Delta,$$

whenever $a/b - a'/b'$. Find integer p such that $p/\bar{n}_2 - p/(\bar{n}_2+\delta_1$ δ_2. Note that such an integer p always exist, since

$$p/\bar{n}_2 - p/(\bar{n}_2+\delta_1) = \frac{p\,\delta_1}{\bar{n}_2+\bar{n}_2\delta_1},$$

and hence can be made arbitrarily large by selecting p sufficiently large.

D. Let DB_1 be the historical database with n_2 and n_3 equal to $\bar{n}_2$ and $\bar{n}_3$ as in DB, and $n_1=p$. In DB_1, the values for the rules fired by $\bar{t}$ thus are as follows, where S is any subset of A'.

$$(S=\bar{t}[S])\rightarrow(B=\bar{t}[B]) \ : (p,(p-\bar{n}_2))$$
$$(S=\bar{t}[S])\rightarrow(B\neq\bar{t}[B]) \ : (p,(\bar{n}_2))$$

$$(S=\bar{t}[S])(B'=\bar{t}[B'])\rightarrow(B=\bar{t}[B]) \ : (2*\bar{n}_3,\bar{n}_3)$$
$$(S=\bar{t}[S])(B'=\bar{t}[B'])\rightarrow(B\neq\bar{t}[B]) \ : (2*\bar{n}_3,\bar{n}_3)$$

Note that the last two rules have the same values as in DB, i.e., values induced by Y.

$$(S=\bar{t}[S])(B=\bar{t}[B])\rightarrow(B'=\bar{t}[B']) \ : (\bar{n}_2,(\bar{n}_2-\bar{n}_3))$$
$$(S=\bar{t}[S])(B=\bar{t}[B])\rightarrow(B'\neq\bar{t}[B']) \ : (\bar{n}_2,\bar{n}_3)$$

Note that the last two rules have the same values as in DB, i.e., values induced by V.

$$(S=\bar{t}[S])\rightarrow(B'=\bar{t}[B']) \ : (p,(p-\bar{n}_3))$$
$$(S=\bar{t}[S])\rightarrow(B\neq\bar{t}[B]) \ : (p,\bar{n}_3)$$

For any C different from B and B',

$$(S=\bar{t}[S])(B=\bar{t}[B])\rightarrow(C=\bar{t}[C]) \ : (\bar{n}_2,0)$$
$$(S=\bar{t}[S])(B=\bar{t}[B])\rightarrow(C\neq\bar{t}[C]) \ : (\bar{n}_2,\bar{n}_2)$$

Strategy: We now consider the DB_2 obtained from DB_1 by application of a simple transform in which δ_1 transactions are changed in accordance with the definition of a $(\bar{t},B,B')$-transform. The result is DB_2 in which the distribution parameters are:

$n_1=p$
$n_2=\bar{n}_2+\delta_1$
$n_3=\bar{n}_3$

It follows that the values for rules in DB_2 are as follows.

$$(S=\bar{t}[S])\rightarrow(B=\bar{t}[B]) \ : (p,(p-\bar{n}_2-\delta_1))$$
$$(S=\bar{t}[S])\rightarrow(B\neq\bar{t}[B]) \ : (p,(\bar{n}_2+\delta_1))$$

$$(S=\bar{t}[S])(B'=\bar{t}[B'])\rightarrow(B=\bar{t}[B]) \ : (2*\bar{n}_3,\bar{n}_3)$$
$$(S=\bar{t}[S])(B'=\bar{t}[B'])\rightarrow(B\neq\bar{t}[B]) \ : (2*\bar{n}_3,\bar{n}_3)$$

$$(S=\bar{t}[S])(B=\bar{t}[B])\rightarrow(B'=\bar{t}[B']) \ : (\ (\bar{n}_2+\delta_1),(\bar{n}_2-\bar{n}_3+\delta_1)\ )$$
$$(S=\bar{t}[S])(B=\bar{t}[B])\rightarrow(B'\neq\bar{t}[B']) \ : (\ (\bar{n}_2+\delta_1),(\bar{n}_3)\ )$$

$$(S=\bar{t}[S])\rightarrow(B'=\bar{t}[B']) \ : (p,(p-\bar{n}_3))$$
$$(S=\bar{t}[S])\rightarrow(B\neq\bar{t}[B]) \ : (p,\bar{n}_3)$$

For any C different from B and B',

$$(S=\bar{t}[S])(B=\bar{t}[B])\rightarrow(C=\bar{t}[C]) \ : (\bar{n}_2+\delta_1,0)$$
$$(S=\bar{t}[S])(B=\bar{t}[B])\rightarrow(C\neq\bar{t}[C]) \ : (\bar{n}_2+\delta_1,\bar{n}_2+\delta_1)$$

on two pseudo-database PDB and PDB', a collections of rule values which are inconsistent in that no distribution parameters y these values.

E. Create a pseudo-database PDB from DB_1 by changing the values for the RF_i rules from $(RF_i, p, \bar{n}_2)$ to $(RF_i, p, \bar{n}_2+\delta_1)$ and the values for the RP_i rules from $(RP_i, p, p-\bar{n}_2)$ to $(RP_i, p, p-\bar{n}_2-\delta_1)$.

By the way p was chosen in step C (observe that the values in DB_1 for the remaining B rules are as specified by Y),

$$Score(\bar{t}, DB_1) - Score(\bar{t}, PDB) > \Delta.$$

F. Create PDB' from PDB by changing the values for all rules Q which $\bar{t}$ passes containing B in LHS and B' in RHS from

$$(Q, \bar{n}_2, (\bar{n}_2-\bar{n}_3))$$

to

$$(Q', (\bar{n}_2+\delta_1), (\bar{n}_2-\bar{n}_3+\delta_1)),$$

and changing the values for rules Q' of this form which $\bar{t}$ fails from

$$(Q, \bar{n}_2, \bar{n}_3)$$

to

$$(Q', (\bar{n}_2+\delta_1), (\bar{n}_3))$$

Since the starting values are V and the resulting $V' \in \delta_1(V)$, we have that

$$|Score(\bar{t}, DB_2) - Score(\bar{t}, PDB)| < \varepsilon.$$

G. Create DB_2 from PDB' by changing the values for all rules W which $\bar{t}$ passes which contain B on the LHS and do not contain B' from

$$(W, \bar{n}_2, 0)$$

to

$$(W, \bar{n}_2+\delta_1, 0),$$

and changing the values for all rules W' of this form which $\bar{t}$ fails from

$$(W, \bar{n}_2, \bar{n}_2)$$

to

$$(W, \bar{n}_2+\delta_1, \bar{n}_2+\delta_1).$$

By the special cases of the monotonicity condition,

$$Score(\bar{t}, PDB') \leq Score(\bar{t}, PDB).$$

Putting together our sequence of changes, we have that

$$Score(\bar{t}, DB_1) - Score(\bar{t}, DB_2) > D. \quad \Box$$

4. Further Reducing the Search Space: The Equivalence of Joint and Conditional Probabiliti with Respect to Hypothesis Testing

The previous sections focused on the problem of selecting probability models for the process that generates transaction and on the effects of including nonmaximal rules in a rule base. In this section we consider another question that arises in rule base and model design, by addressing the following question raised in [Helm89]:

> For a given collection C of attributes and a set X of value vectors for these attributes, how do we choose between the joint probability $Pr\{C \in X\}$ and the many possible conditional probabilities $Pr\{C_1 \in X_1 | C_2 \in X_2\}$ (where C_1 $[X_1]$ and C_2 $[X_2]$ partition C $[X]$) as the best quantities to include in M_N and M_A.

Observe how this question relates to the problem of partitioning a W&S rule into an LHS and an RHS. The main results of this section are surprisingly simple observations that allow us to eliminate this question as a dimension of choice, thus yielding an enormous reduction in the size of the search space proposed in [Helm89].

The following theorem demonstrates that under the Bayesian hypothesis testing procedure we have proposed, the choice of joint versus conditional (and the possible partitionings into conditionals) is immaterial. The theorem states simply that the hypothesis tests supported by the joint quantity and *any* corresponding conditional quantity are identical in that the tests apply to the same transactions and yield identical results.

Theorem 4.1: Let C be a set of attributes, X any set of value vectors for these attributes, and C_1, C_2 and X_1, X_2 any partitions of C and X. The quantities $Pr\{C \in X\}$ and $Pr\{C_1 \in X_1 | C_2 \in X_2\}$ support the testing of the hypothesis $Pr\{M_N | t\}$ for exactly the same transactions t, and yield exactly the same result.
Proof: The inclusion in the models M_N and M_A of either of these quantities supports the test $Pr\{M_N | t\}$ of exactly those transactions t such that $t[C_1] \in X_1$ and $t[C_2] \in X_2$, i.e., $t[C] \in X$. If the models include the joint probability, then the hypothesis test approximates $Pr\{M_N | t\}$ of such a transaction by computing:

$$Pr\{M_N | t\} \cong Pr\{M_N | t[C_1] \in X_1 \wedge t[C_2] \in X_2\}$$

$$= \frac{Pr\{t[C_1] \in X_1 \wedge t[C_2] \in X_2 | M_N\} * Pr\{M_N\}}{(Pr\{t[C_1] \in X_1 \wedge t[C_2] \in X_2 | M_N\} * Pr\{M_N\} + Pr\{t[C_1] \in X_1 \wedge t[C_2] \in X_2 | M_A\} * Pr\{M_A\})}.$$

If, instead of the joint probability, the models include the conditionals $Pr\{t[C_1] \in X_1 | M_N \wedge t[C_2] \in X_2\}$ and $Pr\{t[C_1] \in X_1 | M_A \wedge t[C_2] \in X_2\}$, we would approximate $Pr\{M_N | t\}$ by computing:

$$Pr\{M_N | t\} \cong Pr\{M_N | t[C_1] \in X_1 \wedge t[C2] \in X_2\}$$

$$= \frac{Pr\{t[C_1] \in X_1 | M_N \wedge t[C_2] \in X_2\} * Pr\{M_N \wedge t[C_2] \in X_2\}}{(Pr\{t[C_1] \in X_1 | M_N \wedge t[C_2] \in X_2\} * Pr\{M_N \wedge t[C_2] \in X_2\} + Pr\{t[C_1] \in X_1 | M_A \wedge t[C_2] \in X_2\} * Pr\{M_A \wedge t[C_2] \in X_2\})}$$

In each case, the computed quantity is equal to

$$\frac{Pr\{t[C_1] \in X_1 \wedge t[C_2] \in X_2 \wedge M_N\}}{Pr\{t[C_1] \in X_1 \wedge t[C_2] \in X_2\}},$$

i.e., the two forms of the test yield identical results. □

This result has several interpretations. First, when designing probability models to support the Bayesian hypothesis testing approach, we can reduce significantly the search space defined in [Helm89]; reduction is by a factor proportional to number of partitions of each selection of attributes and aggregation of their values. Second, assuming that W&S is indeed based on the Bayesian model (or approximates it), the rule base design decision of how to partition a potential rule into a LHS and RHS should be immaterial. Since the above theorem is valid regardless of the model assumed for M_A or of the scoring function employed, its result imply there is no mathematical reason to estimate and test against conditional rather than joint probabilities. On the other hand, W&S's structuring of rules in the form of conditional probabilities may be desirable for reasons of an efficient implementation. Our results imply that this is perfectly valid mathematically; however, they imply also that there is no reason to consider different partitions into LHS and RHS of a given set of attributes and their values.

A question left open by the above theorem is whether its results apply when the probabilities are obtained via empirical sampling. That is, could it be that the joint and conditional tests differ as a result of differences in the sampling? If so, it might be reasonable to consider alternative quantities since, in this case, they could yield appearing in Theorem 4.1. That is (sampling from DB),

$$EPr\{t[C_1]=V_1 \wedge \quad _2]=V_2 \mid M_N \wedge DB\}$$
$$= EPr\{t[C_1]=V_1 | t[C_2]=V_2 \wedge M_N \wedge DB\} * EPr\{t[C_2]=V_2 \wedge M_N \wedge DB\}.$$

Proof: The result follows from a simple combinatorial argument. Let

$N=|D|$
N_1 = #transactions in DB with $t[C_1]=V_1$
N_2 = #transactions in DB with $t[C_2]=V_2$
N_3 = #transactions in DB with $t[C_1]=V_1$ and $t[C_2]=V_2$

Then,

$$EPr\{t[C_1]=V_1 \wedge t[C_2]=V_2 \mid M_N \wedge DB\}=\frac{N_3}{N}$$
$$EPr\{t[C_1]=V_1 \mid t[C_2]=V_2\} * EPr\{t[C_2]=V_2 \mid M_N \wedge DB\} = (\frac{N_3}{N_2})*(\frac{N_2}{N})=\frac{N_3}{N} \quad \Box$$

5. The "Distance from Unity" Optimization Criterion

In this section, we consider refinements to the "distance from unity criterion" proposed in [Helm89] as the objective function value to be used in determining which quantities should be included in the models M_N and M_A.

Consider again the attribute selection and value aggregation problems discussed in Section 2.2. The attribute selection problem requires us to find subsets $\{B_1, \ldots, B_k\}$ of the set of all attributes $\{A_1, \ldots, A_N\}$ such that:

(1) The quantity $Pr\{t[B_1], \ldots, t[B_k] | M_N\}$ is accurately reflected in the historical database, our sample population.

(2) The behavior of $R(t[B_1], \ldots, t[B_k])$ reflects that of $R(t)$.

The only way we know of to test whether these conditions are satisfied for candidate subsets of attributes is experimentally. In particular, this is one aspect of the problem for which at least a limited amount of expert feedback seems essential. With feedback, the conceptual solution to the problem would be simply to experiment with each subset of attributes and determine which does the best job of detection (i.e., for what subsets B is the ratio $R(t[B])$ a good detector).

However, it is not computationally feasible to test these conditions experimentally as we attempt to design the model, even if feedback were readily available. What we propose is heuristics for constructing subsets B that are *potentially* interesting. We then would compute probability distributions for the candidate subsets, include these quantities in one or more probability models, and test the models experimentally.

Consider how we might determine quickly if a given subset B of attributes is potentially interesting. Observe that when the distribution of

$$Pr\{t[B_1]=v_1, \ldots, t[B_k]=v_k | M_N\}$$

is, for most values v_i, similar to that of

$$Pr\{t[B_1]=v_1, \ldots, t[B_k]=v_k | M_A\},$$

the ratio $R(t[B_1], \ldots, t[B_k])$ often is near 1. This implies $Pr\{M_N | t\}$ is often calculated to be near $Pr\{M_N\}$ and hence the test has little potential of yielding information. That is, if such a collection of attributes is included in one of the competing tests $R(t)$, the test will often contribute information approximating the *a priori* value $Pr\{M_N\}$. It seems reasonable to conclude that such subsets B are not potentially interesting. In contrast, subsets that maximize the separation for most values v_i should be considered good candidates, because such subsets provide the greatest discrimination between the models in the sense that a transaction to which the test applies will score either very much higher or lower than the *a priori* value and, hence, the test yields much information.

We therefore propose as a criterion the distance from unity measure. For example, if we are considering

$$\overline{Pr\{t[B_1]=v_1|M_N\}*...*Pr\{t[B_k]=v_k|M_N\}}$$

is far from 1, for most value. .

We emphasize the two comments made in Section 2.2:

a) Methods are required (e.g., experimental feedback) to assess the quality of candidates chosen by this heuristic. In particular, model separation is not sufficient to ensure that a candidate solution provides good approximation to the hypothesis test $Pr\{M_N|t\}$.

b) Even if the heuristic measure were a perfect criterion, the problem of building solutions which optimize the measure are difficult (probably NP-hard). Hence, search heuristics are required to build the candidate solutions.

In order to focus on the attribute selection problem, the previous discussion simplified away the related problem of value aggregation. That is, once we have a candidate subset B, there remains the question of how to partition the values of B's attributes. For a given aggregation $X_1, \ldots, X_k$ of values, the corresponding test is based on the ratio

$$R(t[B])=\frac{Pr\{t[B_1]\in X_1, \ldots, t[B_k]\in X_k \mid M_N\}}{Pr\{t[B_1]\in X_1, \ldots, t[B_k]\in X_k \mid M_A\}}.$$

Since the number of possible attribute aggregations for a fixed subset B is exponential in the size of B, we desire heuristics of similar spirit to the one discussed above.

In [Helm89] we propose to apply the distance from unity criterion to this problem as well. While this criterion of maximizing information *when the test applies* is valid, the it is easy to see that the partition cell X_i (of each attribute B_i) which maximizes this measure will always contain only a single value. However, it is quite possible that the test corresponding to such an aggregation will apply to only a very small percentage of transactions. It therefore seems reasonable that the measure of a given aggregation be based on the expected separation, computed by weighting the separation of each potential test by the probability that the test will apply to a random transaction.

6. Summary and Conclusions

The analysis performed in the previous sections is an important first step in the development of a methodology for designing the probability models to be used in misuse and anomaly detection. This research has addressed three important aspects of the design problem: The role and selection of probability models, the identification of classes of models (and W&S rule bases) that can be pruned heuristically before the search begins, and heuristic criteria to be used in the search.

The main results of this paper translate to the following basic principles for the design of a W&S rule base.

1. Knowledge of the M_A models of concern is required in order to analyze the quality of W&S rule bases. It appears that current W&S rule bases perform anomaly, rather than general misuse, detection as they appear to be configured for negative correlation models. It may be possible in the future to configure W&S rule bases for other misuse models, such as piecewise and masquerader models, by applying model-specific principles to the design of the rule base.

2. One model-specific principle implied by our results is that if the M_A models of concern are piecewise, then rule bases containing nested rules should be pruned from the search space. In particular, any rule which does not contain all relevant attributes, in some cases, influences in an incorrect direction the scores of certain transaction.

3. Independent of the M_A models of concern, it appears that rule bases containing rules which are different partitions into LHS and RHS of the same attributes and values should be pruned from the search space.

In addition to proposing a trimmed search space for the design problem, we have refined our criteria for evaluating candidate models. When searching for subsets of attributes and aggregations of their value, we propose considering only candidates that are potentially interesting, in that they are likely to discriminate between the models of concern. We point out, however, that this criterion alone is not sufficient to identify probabilistic quantities that should appear in the model. Experimentation with feedback is necessary to verify that a given

References

[Helm89] Helman, P., "Anomaly detection as inductive inference," Los Alamos National Laboratories Technical Report, 1989.

[Helm90a] Helman, P., "A mathematical analysis of the inductive learning model of anomaly detection and its relation to Wisdom and Sense's scoring conventions, Los Alamos National Laboratories Technical Report, 1989.

[Helm90b] Helman, P. and Liepins, G., working paper.

[Liep90] Liepins, G. and Vaccaro, H., "Artificial Intelligence for Computer Security," unpublished manuscript, 1990.

[Vacc89] Vaccaro, H. and Liepins, G., "Detection of anomalous computer activity," IEEE Symposium on Research in Security and Privacy, pp. 280-289, 1989.