Detecting Group Differences: Mining Contrast Sets

Stephen D. Bay and Michael J. Pazzani

Department of Information and Computer Science

University of California, Irvine

Irvine, CA 92697, USA

{sbay,pazzani}@ics.uci.edu

April 3, 2001

Abstract

A fundamental task in data analysis is understanding the differences between several contrasting groups. These groups can represent different classes of objects, such as male or female students, or the same group over time, e.g. freshman students in 1993 through 1998. We present the problem of mining contrast sets: conjunctions of attributes and values that differ meaningfully in their distribution across groups. We provide a search algorithm for mining contrast sets with pruning rules that drastically reduce the computational complexity. Once the contrast sets are found, we post-process the results to present a subset that are surprising to the user given what we have already shown. We explicitly control the probability of Type I error (false positives) and guarantee a maximum error rate for the entire analysis by using Bonferroni corrections.

Keywords: data mining, contrast sets, change detection, association rules

Contact Author: Stephen Bay

Email: sbay@ics.uci.edu

Phone: 949-824-3491

Fax: 949-824-4056

1 Introduction

A common question in exploratory research is: "How do several contrasting groups differ?" Learning about group differences is a central problem in many domains. For example, the US Census Bureau prepares many statistical briefs that compare groups such as the publication, "The Earnings Ladder: Who's at the Bottom? Who's at the Top?" which contrasts high and low income earners over the years 1979 to 1992. They report such facts as: "About 4 in 10 year-round, full-time workers aged 18 to 24 had low earnings in 1992, up 19 percentage points since 1979."

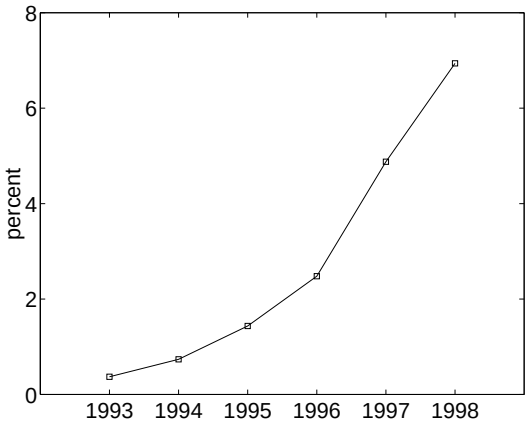
Contrasting specific groups of interest is especially important in social science research. For example, the Integrated Public Use Microdata Series (IPUMS) project (Ruggles & Sobek, 1997) has expended great effort standardizing federal census data to allow researchers to compare demographic groups over different time periods. Some of the research conducted with this data involves comparing different racial groups (Darity, 1998) or examining trends in divorce rates (Ruggles, 1997). As another example, the Department of Urban and Regional Planning at UCI conducts an annual survey of people in Orange County. Their goals are to compare "the quality of life and local government ratings in Orange County with Los Angeles County" and to "analyze the impact of changing demographics by contrasting survey responses of Latinos, Asians, and non-Hispanic whites."

Our goal is to automatically detect all differences between contrasting groups from observational multivariate data. We seek conjunctions of attributes and values, called contrast sets, that have different levels of support in different groups. For example, if we are comparing education groups, we might find that P(occupation = sales | PhD) = 2.7%, while P(occupation = sales | Bachelor) = 15.8%. Alternatively, groups could be based on time with groups falling into different years as shown in Figure 1.

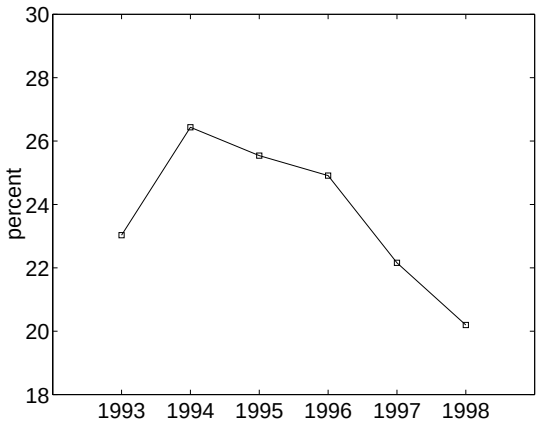
Our emphasis in this paper will be on *discovery* of contrast sets. Our goal is to develop an exploratory analysis tool so that users can gain insight into the differences between groups.

Our problem differs from those addressed by time series research or traditional statistical methods such as cohort or cross-sectional analyses (Glenn, 1977; Menard, 1991). In time series work, we typically have observations spaced through time with one observation per time point. In contrast, we have multiple observations at a few discrete points in time. For example, we could have thousands of observations for each of the 1970, 1980, and 1990 federal censuses. Cohort and cross-sectional analyses are typically guided by well defined prior hypotheses determined by domain knowledge. For example, a research question might be "How does aging affect political party affiliation?" In con-

Figure 1: Comparing UCI Applicants over 1993-1998



(a) Admitted ICS students with SAT Math $>$ 700 $\wedge$ SAT Verbal $>$ 700



(b) UCI applicants with Admit = Yes $\wedge$ Home Location = Los Angeles County $\wedge$ School Type = Public

trast, our work would take descriptions of young and old people and return all differences between them, one of which could be a difference in political affiliation.

Another approach to distinguishing two or more groups from each other is to use a rule learner or decision tree to learn a classification strategy. This has the advantage of being fast. However there are four major disadvantages: (1) Rule learners and decision trees are not complete. They achieve speed by using heuristics to prune large portions of the search space and thus they may miss alternative ways of distinguishing one group from another. (2) Rule learners and decision trees focus on discrimination ability and will miss group differences that are not good discriminators but are still important. For example, knowledge of Figure 1b will not give good classification accuracy in terms of placing students in their year of application.[1] Nevertheless the differences could be vitally important especially to a high school liaison officer at UCI. (3) Rules are usually interpreted in a fixed order where a rule is only applicable if all previous rules were not satisfied. This makes the interpretation of individual rules difficult since they are meant to be interpreted in context. Finally, (4) it is difficult to specify useful criteria such as minimum support or an acceptable false positive rate in the classification framework.

[1]The optimal classification strategy knowing only the information in the Figure 1b will yield an accuracy of 17.7% while random guessing gives 16.67% (assuming equal class priors).

1.1 Relation to Association Rule Mining

A closely related area to our work on contrast sets is association rule mining (Agrawal et al., 1993). Association rules are relations between variables of the form $X \rightarrow Y$. In market basket data X or Y would be items such as bread or milk. In categorical data X and Y are attribute-value pairs such as occupation = engineer or income $\geq$ \$50K.

Finding association rules and mining contrast sets both require search through a space of conjunctions of items or attribute-value pairs. In association rule mining, we look for sets that have support greater than a certain cutoff (these sets are then used to form the rules) and for contrast sets we seek those sets which represent substantial differences in the underlying probability distributions.

Because both techniques have a search element there are many commonalities. In fact we build on some of the search work developed for association rule mining to enhance our contrast set algorithms. However, our work on contrast sets differs substantially from association rule techniques because we are concerned with multiple groups and we have different search criteria. Trying to directly apply association rule mining algorithms to find contrast sets is a poor idea. For example, one approach would be to mine the large itemsets for each group separately and then compare them. However, mining the groups separately causes us to lose pruning opportunities and we show in Section 5.1 that these pruning opportunities can greatly improve efficiency.

Alternatively, since association rules find relations between variables, we could encode the group explicitly as a variable and let an association rule learner run on this representation. This will not, however, return group differences, and the results will be difficult to interpret. For example, we ran an association rule program on census data and obtained the results in Figure 2 (1% min-support, 80% confidence).

Examining these rules, it is extremely difficult to tell what is different between the two groups. First, there are too many rules to compare. Second, the results are difficult to interpret because the rule learner does not enforce *consistent contrast* (Davies & Billman, 1996) (i.e., using the same attributes to separate the groups). Clearly there are at least $26796 - 1674 = 25122$ rules that have no match. Finally, even with matched rules, we still need a proper statistical comparison to see if differences in support and confidence are significant.

Figure 2: Association rules for Bachelor and Ph.D. degree holders. Rules are in the form $X \rightarrow Y$ (support,confidence).

$\rightarrow$ Bachelors (93.1, 93.1)
Bachelors $\rightarrow$ CapitalLoss=0 (86.9, 93.4)
CapitalLoss=0 $\rightarrow$ Bachelors (86.9, 93.4)
Bachelors $\rightarrow$ United-States (83.4, 89.5)
United-States $\rightarrow$ Bachelors (83.4, 93.8)
Bachelors $\rightarrow$ CapitalGain=0 (82.3, 88.4)
CapitalGain=0 $\rightarrow$ Bachelors (82.3, 93.7)
Bachelors $\rightarrow$ race=White (81.6, 87.7)
race=White $\rightarrow$ Bachelors (81.6, 93.0)
Male $\rightarrow$ Bachelors (64.4, 92.0)

(a) First 10 of 26796 Association Rules for Bachelor holders

PhD $\rightarrow$ CapitalLoss=0 (6.1, 88.6)
PhD $\rightarrow$ United-States (5.5, 80.5)
PhD $\rightarrow$ CapitalGain=0 (5.5, 80.3)
PhD $\rightarrow$ race=White (6.1, 88.6)
PhD $\rightarrow$ Male (5.6, 81.0)
United-States $\wedge$ PhD $\rightarrow$ CapitalLoss=0 (4.9, 87.7)
CapitalGain=0 $\wedge$ PhD $\rightarrow$ CapitalLoss=0 (4.7, 85.7)
CapitalLoss=0 $\wedge$ PhD $\rightarrow$ race=White (5.4, 88.2)
race=White $\wedge$ PhD $\rightarrow$ CapitalLoss=0 (5.4, 88.2)
CapitalLoss=0 $\wedge$ PhD $\rightarrow$ Male (5.0, 81.7)

(b) First 10 of 1674 Association Rules for Ph.D. holders

1.2 Overview

In Section 2, we present a formal definition of the mining problem. Section 3 discusses our search algorithm for finding contrast sets and Section 4 discusses how we can filter the mined results to present a concise summary of the differences between groups. We evaluate our algorithms in Section 5 which analyzes objective measures of performance and in Section 6 which is a case study on UCI enrollment and measures empirically how useful the discovered knowledge is to an end user. We discuss related work in Section 7 and finally, we conclude in Section 8.

2 Problem Definition

Association rules typically deal with market basket data where the database $\mathcal{D}$ is a set of transactions with each transaction $T \subseteq I = \{i_1, i_2, \ldots, i_m\}$. Each member of I is a literal called an *item*, and any set of these literals is called an *itemset*. An important step for most mining algorithms is to find all itemsets whose support, the percentage of transactions in which the itemset occurs, is greater than a threshold min-support.

In this paper we generalize the data model to grouped categorical data. The data is a set of k-dimensional vectors where each component can take on a finite number of discrete values. The vectors are organized into n mutually exclusive groups $G_1, G_2, \ldots, G_n$, with $G_i \cap G_j = \emptyset \ \forall i \neq j$. The concept of an itemset can be extended to a contrast set for this model as follows:

> *Definition 1.* Let $A_1, A_2, \ldots, A_k$ be a set of k variables called attributes. Each A_i can take on values from the set $\{V_{i1}, V_{i2}, \ldots, V_{im}\}$. Then a **contrast set** is a conjunction of

attribute-value pairs defined on groups $G_1, G_2, \ldots, G_n$ with no A_i occurring more than once.

Example 1. $\mathsf{sex} = \mathsf{female} \wedge \mathsf{occupation} = \mathsf{manager}$.

Similar to the definition of support for an itemset, we define the support of a contrast set with respect to a group G as follows:

Definition 2. The **support** of a contrast set with respect to a group G is the percentage of examples in G where the contrast set is true.

Our goal is to find all contrast sets whose support differs meaningfully across groups. Formally, we want to find those contrast sets (cset) where:

$$\exists ij \; P(\mathsf{cset} = \mathsf{True} \mid G_i) \neq P(\mathsf{cset} = \mathsf{True} \mid G_j) \tag{1}$$

$$\max_{ij} |\mathsf{support}(\mathsf{cset}, G_i) - \mathsf{support}(\mathsf{cset}, G_j)| \geq \delta \tag{2}$$

and δ is a user defined threshold called the *minimum support difference*. We call contrast sets where Equation 1 is statistically valid *significant*, and contrast sets where Equation 2 is met *large*. If both requirements are met, then we call it a *deviation*.

The statistical significance criterion ensures that the contrast set represents a true difference between the groups. The second criterion measures the effect size and ensures that everything we report to the user is a big enough effect to be important.

One question that arises is why do we use a statistical test when the null hypothesis, that contrast sets have *exactly equal probabilities across groups*, is always false in the real world and thus a large enough sample will lead to its rejection (Cohen, 1990). We use the significance test in addition to the effect size test for two reasons. First, we may have limited data. Clearly with millions of data points, many statistically significant findings will be too small to be practically significant, but what if we have only 10000 or perhaps as little as 1000 data points? Even when data is abundant, we might have a rare group with very few observations. We need techniques that scale across a spectrum of data set sizes. Second, the number of hypotheses can grow exponentially with the number of variables and with multiple hypotheses we need stricter cutoffs to control the false positive error rate. However, it is not clear at what point effect size will be sufficient to eliminate statistically insignificant findings with multiple tests, thus we explicitly control for both factors.

Once we have found all significant and large contrast sets, we would like to present a subset which are "interesting" to the user. Deciding what is interesting is an open problem in data mining.

Clearly, any complete solution would consider the prior knowledge and the subjective viewpoint of the user (Silberschatz & Tuzhilin, 1996). However, this is beyond the scope of our paper; thus, we limit our system to reporting results that are surprising (in a statistical sense) given what we have already shown the user.

3 Search for Contrast Sets

We can find contrast sets that meet our criteria through search. We explore the space of all possible contrast sets and return only those sets that meet our criteria. Clearly this search space is huge (exponential in the number of attribute-value pairs) and cannot easily be explored.

Although we cannot change the fundamental complexity of the problem, we use admissible pruning rules in conjunction with heuristics to limit the complexity. We present STUCCO (Search and Testing for Understandable Consistent Contrasts), and in practice it runs efficiently and can mine at low support differences without being overwhelmed with the number of potential candidates (Section 5). STUCCO uses a breadth-first search framework which incorporates several techniques from work on efficiently mining large itemsets. To this search framework, our contributions are:

1. joint statistical significance and effect size testing to identify valid contrast sets
2. explicit control over search error to limit false discoveries
3. contrast set pruning rules for efficient mining

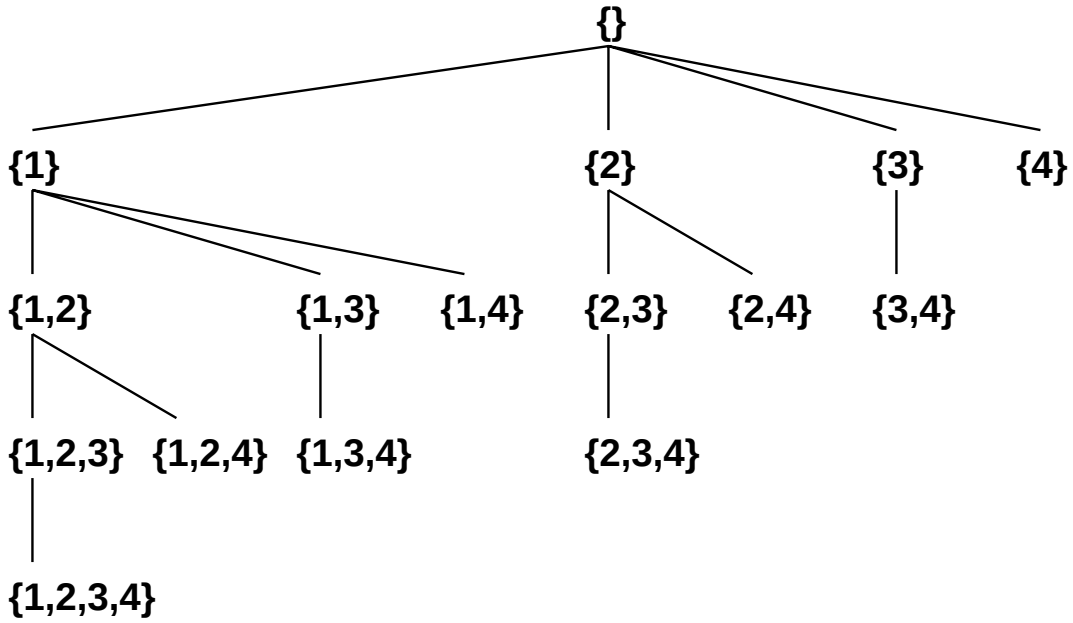
After presenting these contributions, we discuss how our work relates to algorithms that find large itemsets such as Apriori and its variants.

3.1 Framework

We organize the search for contrast sets using set-enumeration trees (Rymon, 1992; Bayardo, 1998) to ensure that we visit every node only once or not at all if nodes can be pruned. Figure 3 shows an example set-enumeration tree for four items (each item is an attribute-value pair). Note that set-enumeration trees are equivalent to canonical orderings (Riddle et al., 1994). A simple rule for creating the tree is to generate the children of a node by appending only those terms that follow all existing terms in a given ordering.

We can search the tree in Figure 3 in any manner we desire, using for example breadth first, depth first, or any other complete search algorithm. We use breadth-first search because it proceeds

Figure 3: Example search tree for four attribute-values pairs with ordering {1,2,3,4}



in a levelwise manner. We start searching the space of contrast sets with the most general terms first, i.e. those contrast sets with a single attribute-value pair such as gender = male or major = Arts. These sets are the easiest to understand and will have the largest support. We then progress to more complicated sets that involve conjunctions of terms, for example, gender = male $\wedge$ major = Arts. The levelwise nature allows us to present results in an anytime fashion.

At each level of the search, we scan the database and count the support of all nodes for each group. We examine the counts to determine which nodes meet our criteria and which nodes should be pruned. We then move to the next level.

During the counting phase we organize nodes into *candidate groups* (Bayardo, 1998; Srikant & Agrawal, 1996). We place nodes with the same parent (i.e. all nodes with a common prefix) in one group. Within the group we maintain two lists of items: the head, $h(g)$, which is the common prefix of all sets in the group, and the tail, $t(g)$, which is the set of all one item extensions to the prefix.

Candidate groups improve the search speed for two reasons. First, the common prefix allows us to quickly check if an observation matches an entire set of candidates (i.e. if the prefix doesn't match then we do not need to check the individual sets). Second, the head and tail organization allows us to specify all of the sets that may occur in that branch of the search tree. Every set is a superset of $h(g)$ and a subset of $h(g) \cup t(g)$. This allows us to bound the support for all sets in that

section of the tree and this is essential for pruning.

We count the support of $h(g) \cup i, \forall i, i \in t(g)$ and $h(g) \cup t(g)$. The first condition means that we count the support of all sets within the candidate group. It also provides an upper bound on the support in any children of a node since frequency counts can only decrease as we add additional terms. The second condition provides a lower bound because for any sets A, B, and C with $A \subset B \subset C$ then the $support(A) \geq support(B) \geq support(C)$. We use knowledge of these upper and lower bounds for our pruning methods described in Section 3.4. Example 2 highlights the use of candidate groups.

> *Example 2.* Consider the sets {1,2}, {1,3}, and {1,4} in Figure 3. We can represent these three sets with a candidate group that has as its head {1} (the common prefix) and as its tail {2,3,4}. During the counting phase we calculate the supports of all the three original sets $h(t) \cup i$ = {1,2}, {1,3}, {1,4} and the support of $h(g) \cup t(g)$ = {1,2,3,4}.

Finally the pruning techniques in Section 3.4 work best when the upper and lower bounds are close together. Thus we dynamically sort the items in the tail so that items that occur frequently are at the end (Brin et al., 1997; Bayardo, 1998). This ensures that the lower bound of support, $h(g) \cup t(g)$, is as large as possible.

3.2 Finding Significant Contrast Sets

We can check if a contrast set is significant by testing the null hypothesis that *contrast set support is equal across all groups* or, equivalently, *contrast set support is independent of group membership.*

The support counts from each group is a form of frequency data which can be analyzed in contingency tables. We form a $2 \times G$ contingency table where the row variable represents the truth of the contrast set, and the column variable indicates the group membership.

For example, consider the top admitted students at UCI as measured by SAT Verbal scores (SATV > 700) and their school of admission. Table 1 shows the contingency table and the counts from our data. If SATV and UCI School are independent variables, then we would expect the proportion of students with high SATV scores to be roughly equal across all groups. Clearly, the proportions are not equal and vary from a high of 12.0% for ICS to a low of 2.7% for Social Ecology. We need to determine if the differences in proportions represent a true relation between the variables, or if it can be attributed to random causes.

The standard test for independence of variables in contingency tables is the chi-square test. It

Table 1: 2×8 Contingency table enumerating high SAT Verbal scores and school: Arts, Biology, Engineering, Humanities, Information and Computer Science (ICS), Physical Science, Social Ecology, and Social Science.

	Arts	Bio	Eng.	Human	ICS	PhysSci	SocEc	SocSci
$SATV > 700$	45	142	85	70	60	34	11	102
$\neg(SATV > 700)$	583	2465	1523	733	502	738	414	1703

works by computing the statistic χ^2:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{3}$$

where O_{ij} is the observed frequency count for the cell in row i and column j. E_{ij} is the expected frequency count in cell ij given independence of the row and column variables and is calculated as follows: $E_{ij} = \sum_j O_{ij} \sum_i O_{ij}/N$ with N being the total number of observations. We then compare the result to the distribution of χ^2 when the null hypothesis is true. If the observed frequencies follow a multinomial distribution and the expected values are not too small, then the χ^2 statistic has an approximately chi-square distribution.

To determine if the differences in proportions are significant, we first pick a test α level. The choice of α sets the maximum probability of rejecting the null hypothesis when it is true. For a single test, α is commonly set to 0.05. We then calculate that $\chi^2 = 49.6$ with 7 degrees of freedom and has a p-value of 1.7e-8. Since the p-value is less than the 0.05 cutoff, we can infer that the null hypothesis is likely false and that contrast set support and group membership are not independent.

3.3 Controlling Search Error

Most data mining algorithms that use a statistical criterion to accept or reject candidates do not consider how search affects the overall error rate. One exception is the work by Megiddo and Srikant (Megiddo & Srikant, 1998) where they found that in association rule mining traditional support-confidence filtering did a good job of eliminating statistically insignificant rules. However their study was done in the context of market basket data, not in terms of dense multivariate data that we analyze. Other work (Silverstein et al., 1998; Liu et al., 1999a) suggests that in dense data there may be many false discoveries.

With a single test, α sets the maximum probability of falsely rejecting the null hypothesis. However, with multiple tests, the probability of false rejection can be highly inflated. This is

especially true in data mining, where often thousands, or millions, of hypotheses are tested. For example, if the null hypothesis is always true (i.e., contrast set probability is independent of group membership) and we made 1000 tests each at $\alpha = 0.05$, we would obtain on average 50 "significant" differences. Falsely rejecting the null hypothesis, i.e., concluding that there is a difference when none exists, is known as a Type I error or false positive.

Type I error can be controlled for a family of tests by using a more stringent α cutoff for the individual tests. We can relate the α_i levels used for each individual test to a global α (the expected error rate) by using the Bonferroni inequality: given any set of events $e_1, e_2, \ldots, e_n$, the probability of their union $(e_1 \vee e_2 \vee \ldots \vee e_n)$ is less than or equal to the sum of the individual probabilities. Applied to hypothesis testing, we let e_i be the rejection of the ith hypothesis h_i. Then, in the simple Bonferroni method, we reject h_i if $p_i \leq \alpha_i$ where $\sum_i \alpha_i \leq \alpha$. Usually $\alpha_i = \alpha/n$, where n is the total number of tests.

This method controls the *error rate per family* (PFE), which is the expected number of false rejections ($PFE \leq \alpha$) for any combination of true or false hypotheses and holds even if the tests are dependent (Hochberg & Tamhane, 1987; Shaffer, 1995). Note that the *familywise error rate* (FWE), the probability of making at least one error, is always less than or equal to PFE (FWE $\leq$ PFE). If all the null hypotheses are true then $PFE = \alpha$.

There are two problems with applying this. First, if we are reporting results incrementally after we mine each level, we do not know how many tests we will make in total. Thus, n is unknown. Second, we use the same cutoff for testing a conjunction of size 1 as size 10. This is undesirable because as α_i gets smaller, we lose power and are less able to detect a difference if it exists. This is an unavoidable tradeoff, as power is related to Type I error. Since lower order conjuncts are more general, we would like more power on those tests.

The Bonferroni inequality holds as long as $\sum_i \alpha_i \leq \alpha$, so we can use different α_i for tests at different levels of the search tree as follows:

$$\alpha_l = \min(\frac{\alpha}{2^l}/|C_l|, \alpha_{l-1}) \tag{4}$$

where α_l is the cutoff for all tests at level l, and $|C_l|$ is the number of candidates at level l. This apportions 1/2 of the total α risk to tests at level 1, 1/4 to tests at level 2, and so on. The minimum requirement ensures that the test α levels always become more stringent and this is necessary for χ^2 based pruning (next section). We use this approach of partitioning risk whenever we make a series of tests.

3.4 Pruning

We can prune portions of the search space when we determine that the search will only lead to contrast sets that fail to meet our effect size or statistical significance criteria. We also prune when we determine that the search will lead only to uninteresting contrast sets.

3.4.1 Effect Size Pruning

We prune nodes based on effect size when we can bound the maximum support difference between the groups below δ.

THEOREM 1. Let $U[i]$ be an upper bound and let $L[j]$ be the lower bound of support for groups i and j. Then the following is an upper bound on the support difference between any two groups.

$$\delta_{\max} = \max_{ij, i \neq j} U[i] - L[j] \tag{5}$$

Proof. Consider groups i and j. The maximum support difference will occur when either the support of i is greater than j or vice versa. Clearly if the support of i is greater than j, then $U[i] - L[j]$ is an upper bound on the support difference between the two groups ($U[j] - L[i]$ if the support of i is less than j). The above equation considers all possible pairs of groups and takes the maximum, thus it is an upper bound. □

Apriori pruning is a special case where we have only a single group and the lower bound is zero. Thus we prune if the upper bound is below δ.

3.4.2 Statistical Significance Pruning

Nodes are pruned when either there are too few data points to have a valid chi-square test or the maximum value the χ^2 statistic can take is too small to be significant.

Validity of the chi-square test: The expected cell frequencies in the top row of the contingency table can only decrease as we specialize the contrast set. This is important because the validity of the chi-square test depends on approximating the distribution of the χ^2 statistic with the chi-square distribution. When the test is invalid, we prune the node because we cannot make valid inferences.

The approximation is made under the assumption that the expected cell frequencies in the contingency table are not "too small." Typically, expected values of 5 or more are considered satisfactory (Everitt, 1992). However, a number of researchers have pointed out that this may be overly conservative and that smaller expectations are sufficient. Lewontin and Felsenstein (Lewontin

& Felsenstein, 1965) suggest that for $2 \times c$ tables we can use the chi-square criterion even when the expected values are as low as 1. To be somewhat conservative, we set the limit at 3.

χ^2 **Maximum**: The χ^2 statistic is non-monotonic and thus cannot be used in any pruning method. However, since the actual cell counts in the top row of the contingency table must decrease as we specialize (the bottom row must increase), the maximum possible value of the χ^2 statistic in any child can be found. We can use this to prune candidates when it is no longer possible for specializations to meet the χ^2 cutoff implied by the test α_l value.

THEOREM 2. Let $U = \{u_1, u_2, \ldots, u_G\}$ be the upper bound on the counts in row 1 and let $L = \{l_1, l_2, \ldots, l_G\}$ be the lower bound. Note that the counts in row 2 are automatically determined once row 1 is specified since the columns must sum to the group totals. Then the following is the maximum value of the χ^2 statistic possible in any specialization.

$$\chi^2_{\max} = \max_{o_i \in \{u_i, l_i\}} \chi^2(o_1, o_2, \ldots, o_G) \tag{6}$$

where $\chi^2(o_1, o_2, \ldots, o_G)$ is the value of the χ^2 statistic for a contingency table with observations $o_1, o_2, \ldots, o_G$ in row 1 (i.e. $o_i = O_{1i}$).

Proof. χ^2 is a convex function (Appendix A), therefore the maximum value must occur at an *extreme* point (Bazaraa & Shetty, 1979) which is a corner of the feasible region. The above equation takes the maximum over all 2^G extreme points and thus is a maximum of the feasible region.□

We expect to compare only a small number of groups, say $G < 10$, so that the exponential number of extreme points we must evaluate is small. If G is large we can use χ^2 bounds (Bay & Pazzani, 1999) that can be found in linear time.

3.4.3 Interest Based Pruning

The previous pruning methods only eliminated deviations that could not meet the effect size or statistical significance criteria. In this section, we present pruning methods that may eliminate contrast sets that are deviations but are clearly not interesting. Contrast sets are not interesting when they represent no new information and this may occur when specializations of the contrast set have identical support or when the relation between groups is fixed.

Specializations with Identical Support: We believe that specializations with the same support as the parent are not interesting for two reasons. First, since both sets will cover the same instances in the database, targeting the parent with an action will be the same as targeting the child. Second, specializations with the same support often represent findings that are common knowledge. Consider

the following two examples.

Example 3. Consider the contrast sets race = Latino ∧ Income > $50K and race = Latino ∧ Income > $50K ∧ WorkedLastYear = Yes. Because of the high prevalence of WorkedLastYear = Yes, the support of both contrast sets will be nearly identical but given that we know the first contrast set adding the additional term is not interesting.

Example 4. Consider the contrast set marital-status = husband. This contrast set will have exactly the same support as marital-status = husband ∧ sex = male barring data errors. The addition of the sex = male term adds no information as by definition all husbands are male.

We require that specializations of a contrast set be different both statistically and in terms of an effect size measurement. Formally, let cset' be a specialization of the contrast set. Then if either of the following two conditions is not true, we prune the node.

$$\exists i \; P(\text{cset} = \text{True} \mid G_i) \neq P(\text{cset'} = \text{True} \mid G_i) \tag{7}$$

$$\max_i |\text{support}(\text{cset}, G_i) - \text{support}(\text{cset'}, G_i)| \geq \delta_s \tag{8}$$

We test these criteria in a similar fashion as Equations 1 and 2. We use the contrast set that represents L the lower bound in place of cset' as it is the maximally different descendent. Typically, we set δ_s to a very small number such as the minimum of 1% or $\delta/2$.

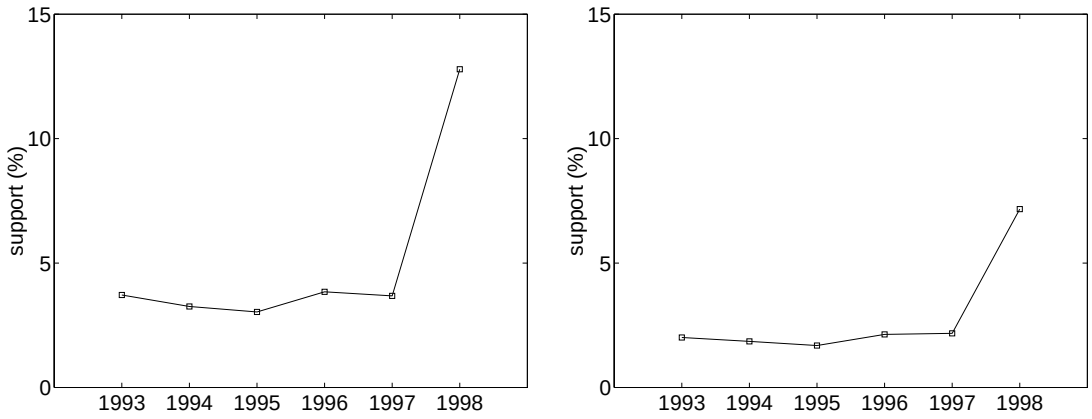
Fixed Relations: Often one group will have a much higher support level for a given contrast set than all other groups and this will be true no matter what additional terms we add. In this case we prune the node as the contrast set specializations do not add new information. For example, in Figure 4, the support for 1998 is much higher than the other years. This higher support continues no matter what terms we append. Formally, if m is the group with the highest support then we prune the node under following condition:

$$\min_{i, i \neq m} |L[m] - U[i]| \geq \delta_f \tag{9}$$

3.5 Relation to Itemset Mining

Our problem is related to finding all large itemsets because the minimum support difference criterion (Equation 2) implies constraints on the support levels in individual groups. For example, mining contrast sets at a support difference of 10% implies that the support in at least one group must be greater then or equal to this value.

Figure 4: Comparing admitted UCI students over 1993-1998



(a) Ethnicity = Decline to State

(b) Ethnicity = Decline to State $\wedge$ AdmitCode = Freshman $\wedge$ BornInUS = Y $\wedge$ HighSchoolType = public

Finding the large itemsets is a common step in association rule mining and most algorithms eliminate large portions of the search space based on the following properties:

If A and B are itemsets where $A \subset B$, then

1. If A is infrequent then B must be infrequent.
2. If B is frequent then A must be frequent.

These two properties allow efficient search for large itemsets. Property 1 is used by bottom up algorithms (Agrawal & Srikant, 1994; Mannila & Toivonen, 1997) (Apriori and its variants) and allows *subset infrequency* pruning where a candidate is eliminated if any of its subsets are eliminated. Property 2 is used by top down algorithms (Zaki et al., 1997) and allows *superset frequency* pruning where a candidate is eliminated if any superset is frequent. Algorithms such as Max-Miner (Bayardo, 1998) and Pincer-Search (Lin & Kedem, 1998) use both properties for efficient search.

Finding group differences is a more difficult task because these properties do not apply to contrasts sets: A can be a deviation while B is not and vice versa. However, we are still able to use A and B's support in our pruning methods as upper and lower bounds.

We also need to consider multiple groups. Although this may seem like an inherent disadvantage that would make search more difficult, it actually allows additional pruning opportunities not possible if we were to mine the groups separately and then combine the results in a post analysis. For example, both our statistical significance and interest based pruning require knowledge of all

group's supports to work. Without these methods, we would not be able to mine efficiently at low supports (Section 5.2).

Effect size pruning can be seen as a generalization of subset infrequency pruning extended to handle group differences and lower bounds. Our interest based pruning is similar to superset frequency pruning: we look ahead in the search (to find a lower support bound) and we use that knowledge to eliminate many sets. The difference is that we concentrate on eliminating uninteresting subspaces whereas superset frequency pruning eliminates non maximal sets that are large (the pruned sets are guaranteed to be large so they do not need to be explicitly tested).

Our statistical significance pruning does not have an analog in itemset mining because association rule programs have not incorporated statistical testing. Our pruning also differs from that used in correlation rules (Silverstein et al., 1998) because of our problem formulation. Our contingency tables are $2 \times G$ and the χ^2 statistic is non-monotonic whereas in correlation rules the contingency table is n-dimensional with 2^n cells where n is the number of items, and the χ^2 statistic is *upward closed* (i.e. if a set of items has a χ^2 value of S, then any superset will also have a χ^2 value of at least S).

4 Filtering for Summarizing Contrast Sets

Dealing with the large volume of data produced by data mining is a difficult problem. For example, association rule mining programs can produce thousands of results (Liu et al., 1999a) which are far too many for a user to view. We typically discover several thousand contrast sets, thus we need to summarize or reduce our results to present a small interesting subset.

Deciding what we should show to a user is a difficult task because of its subjective nature. Some past approaches have used constraints on the variables or items which appear to limit the rules shown (Klemettinen et al., 1994; Srikant et al., 1997; Ng et al., 1998; Bayardo et al., 1999). Another technique is to compare the discovered rules to an explicit list of rules or beliefs that the user already knows (Liu & Hsu, 1996; Liu et al., 1997; Padmanabhan & Tuzhilin, 1998; Silberschatz & Tuzhilin, 1996; Liu et al., 1999b) and then to show only those results that are unexpected. For temporal groups, time series shape matching could also be used (Agrawal et al., 1995; Keogh & Pazzani, 1998).

All of these methods could be used to filter the discovered contrast sets. Here we present two new methods that we have found useful. The first is an expectation based statistical approach which keeps track of what has been shown to the user and only presents results that are surprising

given the current context of results. It does not require the user to input information unlike several of the above approaches. The second allows us to identify and select contrast sets that are linear trends.

4.1 Statistical Surprise

We show the user the most general contrast sets first, those involving a single term, and then only show more complicated conjunctions if they are surprising based on the previously shown sets. For example, we might start by showing the contrast sets Gender = female, School = ICS, and GPA > 4. We would then move on to showing more complicated sets such as Gender = female $\wedge$ School = ICS or School = ICS $\wedge$ GPA > 4, and finally Gender = female $\wedge$ School = ICS $\wedge$ GPA > 4. This bottom up approach is similar to forward selection approaches (Agresti, 1990) and to Liu, Hsu, and Ma (1999a).

Figure 4.1 shows our algorithm for filtering contrast sets. The function model returns the expected value based on a log-linear model which we will describe next. The function isSurprising returns true if the expected counts are different (as in Equations 1 & 2) from the observed counts.

Figure 5: Filtering Algorithm for Finding Surprising Contrast Sets

Algorithm Filter
Input: D a queue of deviations sorted by size (smallest first)
Output: D_s a list of surprising deviations
Let pop(X) remove and return the head of queue X
Let model(X,Y) return the expected value for Y given knowledge of X
Let isSurprising(X,Y) return true if X is substantially different from Y
Begin
1. $D_s \leftarrow \{\}$ // initialize D_s to the empty set
2. **while** D is not empty
3. $C \leftarrow$ pop(D) // C is our current contrast set
4. $E \leftarrow$ model(D_s,C) // E holds the expected observations for each group
5. **if** isSurprising(C,E)
6. $D_s \leftarrow D_s \cup C$
7. **return** D_s

End

Our model estimates the probability of a conjunction based on its subsets and from this we obtain our expected frequency counts. This is a well studied problem and there are many algorithms, such as *Iterative Proportional Fitting* (IPF), (Bishop et al., 1975; Everitt, 1992) which can find the maximum likelihood estimates. These algorithms are equivalent to assuming a log-linear model of

the data (Agresti, 1990; Bishop et al., 1975; Everitt, 1992; Knoke & Burke, 1980) where "the log of the expected frequency is an additive function of a constant plus terms for each variable and their interactions" (Knoke & Burke, 1980). We believe that this model is accurate at identifying contrast sets that are not interesting because they can be explained by lower order interactions.

Our frequency expectations obtained under the log-linear model is the maximum likelihood estimate given the marginals seen. This assumes that the interaction term involving all variables is zero although this will not be true in general. For example, in calculating the expectation of Gender = female $\wedge$ School = ICS we ignore the interaction between gender and school and only look at the individual subsets even though these two are not independent. We deal with this by scaling the expected frequencies with a factor derived from all groups (Appendix B).

For computation speed, our model uses direct estimates (i.e. equations) when they exist, otherwise we use IPF.

4.2 Detecting Linear Trends

If our group variable is ordinal as in federal census data, our task of mining contrast sets is identical to finding changes over time. With temporal data, linear trend detection is important because they often represent simple relations between the variables which are easy to understand and may have good predictive power.

In Section 3.2 we showed that we could detect significant contrast sets by using the chi-square test to check for independence of contrast set support and group membership. However this test checks for all deviations from independence and does not measure a specific type of departure such as a linear trend. Since our contingency table is $2 \times G$ we can use regression techniques (Everitt, 1992) to find the portion of the χ^2 statistic that arises from a linear trend if the group variable is ordinal (such as data from consecutive years).

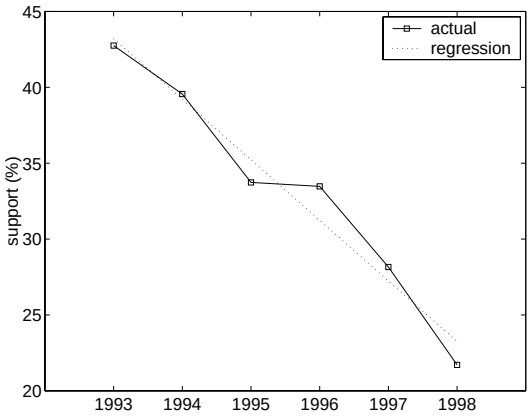
We fit a regression equation to the categorical data by encoding the group (column) as the independent variable (x) and the contrast set truth (row variable) as the dependent or target variable (y). The portion of the the χ^2 statistic that comes from linear trend is then $b_{yx}^2/V(b_{yx})$ where b_{yx} is the linear regression coefficient of y on x, and $V(b_{yx})$ is its variance. The portion of χ^2 explained by linear regression has 1 degree of freedom.

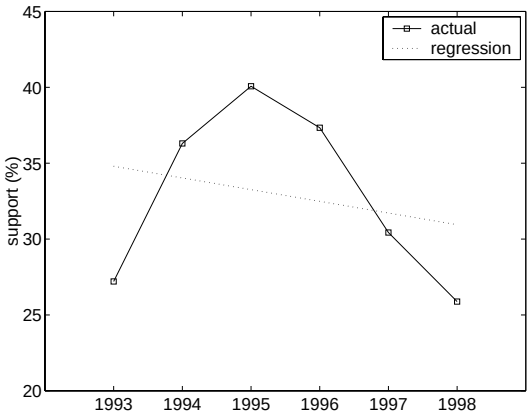
Figure 6 shows two different significant deviations found from 1998 UCI admissions data.[2] In part (a) we see that a linear trend is responsible for most of the total χ^2 value and that we have

[2]The reported p values are not adjusted for multiple tests.

an extremely low p value. The departure from regression is not significant. In part (b) we see that although the deviation is significant, a linear trend is not responsible.

Figure 6: Deviations from independence and linear trends: (a) significant linear trend in the percentage of admitted ICS students whose first language is not English; (b) no significant linear trend for intent to enroll of admitted Social Ecology students.





Source of Variation	d.f.	χ^2	p
linear regression	1	54.12	1.9e-13
departure	4	2.36	0.67
total	5	56.48	6.5e-11

(a)

Source of Variation	d.f.	χ^2	p
linear regression	1	2.64	0.10
departure	4	34.98	4.69e-07
total	5	37.62	4.51e-07

(b)

5 Evaluation

In this section, we evaluate STUCCO on objective measures of performance. We ask three research questions. (1) Can we efficiently mine contrast sets at low support differences (small δ)? (2) Which pruning rules are responsible for the speedups? (3) How effective are the filtering rules for reducing the number of items that need to be shown to the end user?

All experiments were performed on a Sun Ultra 5 computer with 128MB of memory. STUCCO was implemented in C++ and compiled with gcc version 2.7.2.1. To provide a comparison, we used C. Borgelt's implementation of Apriori, version 2.1, which was implemented in C.[3] This version of Apriori is highly optimized and uses prefix trees which implement set-enumeration search and can quickly count candidates in a similar manner to candidate groups. Note that Apriori does not perform the same task as STUCCO and it serves only as a foil to help understand performance

[3]This program is available from http://fuzzy.cs.Uni-Magdeburg.de/~borgelt/. Version 1.8 of his program is incorporated in the data mining tool Clementine.

issues. We used the following datasets which are summarized in Table 2.

- *Adult.* The Adult Census data contains information extracted from the 1994 Current Population Survey. There are variables such as age, working class, education, sex, hours worked, salary, etc.

- *Mushroom.* This data set describes mushrooms and their physical properties such as shape, odor, habitat, etc. Mushroom is not a true observational data set as the examples are not drawn from individual instances but rather are compiled from a field guide (Lincoff, 1981). It is a difficult data set for most mining algorithms because there are many frequent and long itemsets.

- *UCI Admissions Data.* See Section 6.

- *Integrated Public Use Microdata Series* (IPUMS). The IPUMS project (Ruggles & Sobek, 1997) is a large collection of federal census data which has standardized coding schemes to make comparisons across time easy. We obtained an unweighted 1 in 100 sample of responses from the Los Angeles – Long Beach area for the years 1970, 1980, and 1990. The household and individual records were flattened into a single table and we used all variables that were available for all three years.[4] When there was more than one version of a variable, we used the most general. For occupation and industry we used the 1950 basis. Continuous variables were discretized into roughly 5 to 10 equal sized divisions by frequency (e.g. income) or interval width (e.g. age). Finally, we further randomly sampled the data to obtain a 1 in 1000 sample. Federal Census data is one of the most difficult data sets to mine because of the long average record width coupled with the high number of popular attribute-value pairs which occur frequently in many records. These two factors combine to result in many long and frequent itemsets.

The Adult and Mushroom datasets are available from the UCI Repository of Machine Learning Databases (Blake & Merz, 1998). The IPUMS data is available from the UCI KDD Archive (Bay, 1999).

[4] Note that PUMS data is based on cluster samples, i.e. samples are made of households or dwellings from which there may be multiple individuals. Individuals from the same household are no longer independent and thus we violate the independence assumption. We ignore this as we are using the data only as a computational test (this violation is also ignored in (Silverstein et al., 1998; Dong & Li, 1999; Brin et al., 1997) etc.). Ruggles (Ruggles, 1995) suggests that even if the independence assumption is violated, because of stratification the standard errors on PUMS data may be similar to what we would expect of a true random sample.

Table 2: Description of Data Sets

Data Set	# Features	# Examples	Groups
Adult	13	8619	Doctorates, Bachelors
Mushroom	22	8124	Edible, Poisonous
UCI Admissions	19	100876	Applicants 1993–1998 (6 cohorts)
IPUMS Census	60	23348	1970, 1980, 1990

5.1 Mining Efficiency

Evaluating the efficiency of mining algorithms analytically is difficult without imposing strong distributional assumptions that make the analysis of limited worth. Thus we resort to experimental studies.

We ran both STUCCO and Apriori on the evaluation data sets and Figure 7 shows the results comparing CPU time to the minimum support difference. For Apriori we reported the sum of times to mine at the given support level for each group. Note that if we were to use Apriori to mine contrast sets this would grossly underestimate the computational effort because with multiple groups we need support counts for itemsets below the support difference threshold. For example, if we are mining at a support difference of 10% and group A has a support of 11% we still need to mine group B as long as its support is non-zero.

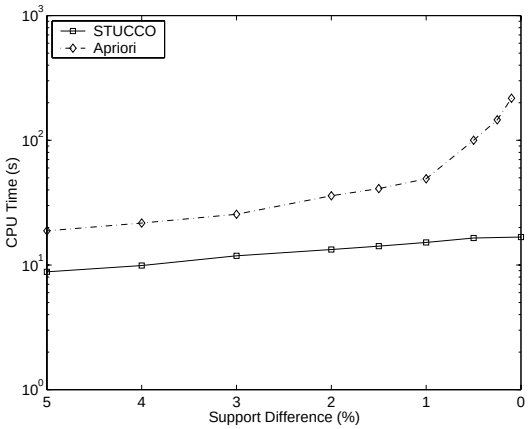
STUCCO was very fast and did well on all data sets, even on Mushroom and IPUMS which are among the most difficult data sets for mining algorithms. STUCCO was slower than Apriori on UCI Admissions by a factor of approximately three. This is probably because UCI Admissions has few high support attribute-value pairs thus STUCCO could not take advantage of lower bounds: Pruning is most effective when the lower and upper bounds are close. The number of candidates examined by STUCCO on UCI Admissions was similar to Apriori, but STUCCO does more work per candidate as it performs additional testing beyond simple size comparisons.

5.2 Pruning Strategies

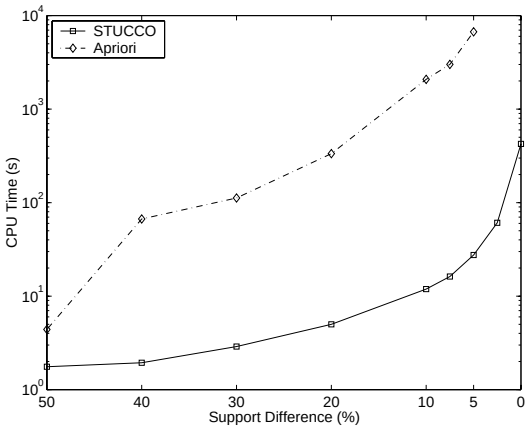
In this section, we examine the effectiveness of the three pruning strategies on STUCCO's complexity. We use a lesion study approach where we remove a single strategy and evaluate the algorithm's performance. This gives a good measure of the relative importance of the various parts and allows us to see the unique contribution of each pruning strategy.

We compared three methods by removing all pruning based on: (1) effect size, (2) statistical

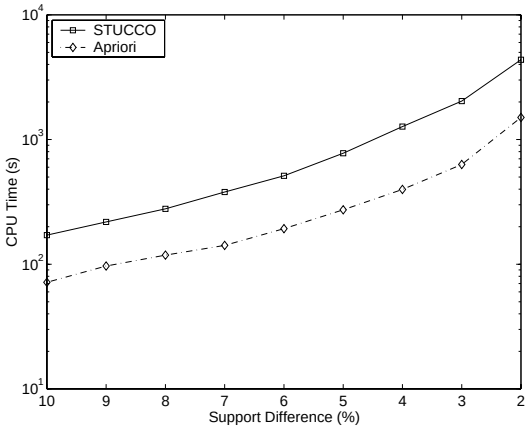
Figure 7: CPU Time versus Support Difference



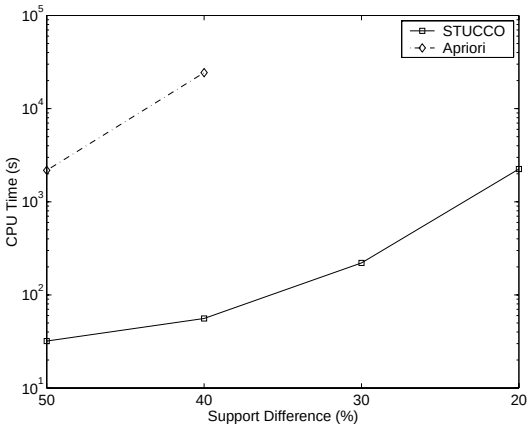
(a) Adult



(b) Mushroom



(c) UCI Admissions

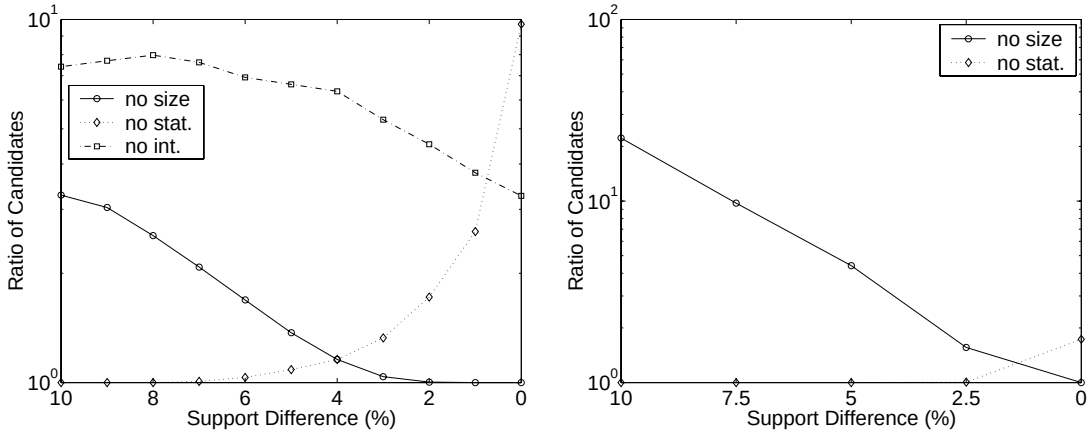


(d) IPUMS

significance, and (3) interest. We used the Adult and Mushroom data sets because we could mine these with the entire range of support differences (i.e. we could mine with $\delta = 0$). Figure 8 shows the ratio of candidates counted at different support levels with respect to the number of candidates counted using all of the pruning strategies. On the Mushroom data set it was impossible to mine without interest based pruning with $\delta \leq 20\%$.

The results indicate that effect size and statistical significance pruning are complimentary. As $\delta \to 0\%$, statistical significance pruning is more important. Conversely as $\delta \to 100\%$ effect size pruning is more important. Interest based pruning is essential for difficult data sets.

Figure 8: Lesion Studies of Pruning Effectiveness



(a) Adult

(b) Mushroom

5.3 Filtering Evaluation

In this section, we evaluate the filtering algorithm in terms of the number of contrast sets shown to the user. We realize that filtering techniques must also be measured in terms of the quality of the results and we address this in the next section.

Table 3 shows the number of deviations and surprising sets found by STUCCO for the four data sets. Most deviations were not surprising and thus we were able to drastically reduce the number of contrast sets by more than an order of magnitude.

It is not hard to see why filtering does a good job at removing unnecessary contrast sets. Consider a simple case with n variables all independent of each other but with different support levels for each group. All 2^n combinations of the variables should be deviations (assuming that when multiplied together the probabilities aren't exactly equal) but all combinations are uninteresting because there

are no higher order effects.

Table 3: Filtering Effectiveness

	Adult		UCI Admissions		Mushroom		IPUMS	
Level	Dev.	Surp.	Dev.	Surp.	Dev.	Surp.	Dev.	Surp.
1	34	34	32	32	69	69	16	16
2	234	7	295	61	627	354	179	35
3	699	4	1111	34	1769	382	695	27
4	1031	0	2008	11	2443	133	1610	86
5	1014	1	1606	1	1003	39	1411	11
6	610	0	459	0	196	0	503	6
7	87	0	29	0	6	0	64	0
8	4	0					0	0
total	3713	46	5540	139	6113	982	4478	181
time	93.4s		29.8m		336.7s		279.1s	

The performance of STUCCO is reasonable. The largest data set, UCI Admissions, took about 30 minutes and the other data sets took at most a few minutes. In the worst case, however the running time could be very bad as during iterative estimation with IPF we potentially need to sum over a contingency table of size 2^n where n is the number of variables.

6 Case Study: UCI Admissions Data

This section presents a case study that demonstrates how STUCCO is used in practice and provides empirical evidence that contrast sets are more useful for understanding group differences than other alternatives such as classification rules from C5.

At UCI, the admissions office collects data on all undergraduate applicants to UCI. The second author serves on a campuswide committee whose goal is to analyze this data to identify changes that could be made to admissions policies that would improve the quality, quantity, and diversity of students that enroll at UCI. Currently the admissions officers typically analyze the data by manipulating spreadsheets and thus they can only form simple summaries and do not perform detailed multivariate analyses.

We have access to 6 years of data from 1993–98 with about 17000 applicants for each year. The data contains information on variables such as ethnicity, UCI School (e.g. Arts, Engineering, etc.), if an offer of admission was made, gender, home location, first language, GPA, SAT scores, Selection Index Number (SIN) which is a composite score formed from GPA and SAT scores, statement of

intent to enroll, etc. We joined the data with a zipcode database and added fields for the distance to UCI and to other UC schools. From this data we selected the relevant tuples that corresponded to our groups. Numeric variables, such as SAT scores and distances were manually converted into nominal variables at thresholds that are meaningful for the admissions office.

Here, we report on an analysis of the 1997-98 enrollment data to identify differences between students who chose to enroll and those who did not for, Biology, ICS, undeclared students and all UCI students as a whole. We ran STUCCO and C5.0 Rules on the data to obtain contrast sets. For STUCCO we used the following parameter settings: $\delta = 1\%$ and $\alpha = 1$. For C5.0 Rules we used the default parameter settings except we set the misclassification costs to balance the different group sizes (typically only 30% of admitted students will enroll). This was necessary as without cost balancing C5 would sometimes fail to find any rules.

The contrast sets can easily be converted to English paragraphs describing the differences in a rule like format. With contrast sets we normally report the support for each group. However, admissions officers find this format difficult to understand and thus we translated the results into yield and gain. Yield is the percentage of students that enroll; gain is the difference in the number of students that would enroll if the yield was identical to the average yield. The contrast sets can be ordered by gain to see what changes might have the largest effect. Rule 1 shows a sample contrast set converted automatically to English text. Appendices C & D contain all results for STUCCO and C5 on Biology.

> *Rule 1.* Students who are Korean and have a SIN between 6000 and 6500 are more likely to enroll with a 30% higher yield than average. This represents a gain of 66 students.

Table 4 shows the size and number of sets mined by C5 and STUCCO. Examining the table we see that C5 returned far more results than STUCCO and that the individual sets tended to be larger and more complicated. While more complex results are undesirable, by itself, it is not an indication that one method is better than another. However when we examine individual sets, we found that C5 suffered from two problems that make it unsuitable for finding group differences.

The first problem is that C5 tended to find many rules with a gain or loss of very few students such as in Rule 2. For UCI, *the median number of students affected by C5 contrast sets was only 3 whereas for STUCCO the median was 103.* Of the 235 sets found by C5 for UCI, 189 affected less than 10 students. The minimum number of students affected by a STUCCO rule was 22. Of course, this occurs because C5.0 Rules tries to find one way to distinguish with high accuracy those students that will enroll in UCI and divides the data into very small sets with high concentrations of students that do or do not enroll.

Table 4: Summary of Results for C5 and Stucco

	UCI		Biology		ICS		Unaffiliated	
Size	C5	Stucco	C5	Stucco	C5	Stucco	C5	Stucco
1	29	42	13	26	22	6	19	32
2	54	34	20	12	14	0	50	15
3	58	16	17	7	3	1	32	10
4	48	9	15			1	20	1
5	33	1	3				9	
6	10		1					
7	2							
8	1							
total	235	102	69	45	39	8	130	58

Rule 2. Students who declined to state their ethnicity, are from Los Angeles County, have a SIN between 6500 and 7000, have a parental income greater than $80000, and live between 30 and 100 miles away from UCI are more likely to enroll with a 34% higher yield than average. This represents a gain of 2 students.

The second problem of C5 is that it missed several important rules, even when the rules were simple and obvious. Rule 3, the most important factor found by STUCCO, did not show up as a factor for C5.

Rule 3. Students who live within 30 miles of UCI are more likely to enroll with a 11% higher yield than average. This represents a gain of 432 students.

We now present an actionable rule which identifies students that could be directly targeted. Rule 4 was found by STUCCO and suggests that UCI does an extremely poor job of recruiting bright students who have not yet declared a major. This is probably because recruiters treated non-declared majors as confused students who needed help rather than as students who wanted to explore their options. Rule 4 passes the filtering mechanism because its yield is much greater than one might expect given the subsets (Rules 5 and 6).

Rule 4. Students who have a GPA greater than 4, and are undeclared majors are less likely to enroll with a 15% lower yield than average. This represents a loss of 123 students.

Rule 5. Students who have a GPA greater than 4 are less likely to enroll with a 9% lower yield than average. This represents a loss of 76 students.

Rule 6. Students who are undeclared majors are less likely to enroll with a 2% lower yield than average. This represents a loss of 57 students.

C5 did not find the contrast set in Rule 4. Due to this discovery, UCI is changing the way it approaches recruiting undeclared students, particularly those with high GPAs.

UCI has been uncertain of the effect that the proximity to UCI and other UC campuses plays in student's college choice. Students who live at home with their parents substantially reduce the cost of higher education. It was well known that students who live close to UCI are more likely to accept offers, but little was understood about how this interacts with other variables. The following STUCCO rules provide insight into this. Rule 7 suggests that UCI competes fairly well for students with UCLA, UCSD and UC Riverside. Rule 8 suggests that UCI does a poor job of recruiting highly qualified first generation students who live near other UC campuses.

> *Rule 7.* Students who live within 30 miles of UCI and live within 30 miles of another UC school are more likely to enroll with a 10% higher yield than average. This represents a gain of 329 students.

> *Rule 8.* Students who have a Selection Index Number greater than 7000, are not born in the US, live within 30 miles of another UC school are less likely to enroll with a 18% lower yield than average. This represents a loss of 139 students.

Our results here have shown that there are problems with using C5 to perform the descriptive task of mining interesting contrast sets and that these problems do not occur with STUCCO. This is not surprising as C5 was primarily intended to be a classification tool (and it does this job well).

7 Related Work

We restrict our discussion of related work to general change detection algorithms and to filtering algorithms for reducing and summarizing mining results.

7.1 Change Detection Algorithms

Concurrent with our work, Dong and Li (1999) worked on the problem of discovering *emergent patterns* (EP). An EP is defined as an itemset X where

$$growthrate(X) = \frac{support_{D_2}(X)}{support_{D_1}(X)} > g \tag{10}$$

and D_1, D_2 are two different data sets and g is growth limit such as 2. Their algorithm represents the EPs by using borders. For example, an EP could be $< \{1,2,3\}, \{1,2,3,6,9,10\} >$. This means

that all sets that are contained in the border (superset of $\{1,2,3\}$ and subset of {1,2,3,6,9,10}) would have a growth rate of at least g between the two data sets.

To calculate the EPs they first find the border of large itemsets using Max-Miner for each of the two data sets. For example, if comparing Texas and Michigan at $g > 1.2$ they might enumerate all itemsets with support greater than 25% in Texas and 30% in Michigan. They then use an operation called border-diff to compare these sets and find the EPs as above.

This is an interesting and very promising approach in that it attempts to calculate all possible EPs after finding the large borders with Max-Miner. It is reminiscent of the highly influential version space algorithm(Michell, 1977). However, for the problems we are addressing, there would be several drawbacks with this. First, their algorithm must mine the data multiple times for different base supports. For example, if they were trying to find all EPs in census data between Texas and Michigan with $g > 1.2$ they would have to enumerate with Max-Miner all itemsets with support greater 25% in Texas and 30% in Michigan, 40% Texas and 50% Michigan, etc. Second, it is not clear if the method can be extended to handle more than two groups. Third, so far there is no method of verifying the statistical significance of discovered EPs. Consider that if an itemset occurs once in D_2 and never occurs in D_1, its growth rate is considered ∞. Their algorithm would find and report this. Finally, there is the problem of displaying the large volume of results. For example, on the Mushroom data set they found 299811 borders, each representing about 2^{18} sets. This is far too many results to show to an end user.

Explora (Hoschka & Klösgen, 1991; Klösgen, 1996) searches for subgroups of cases with unusual distributions of a target variable with respect to a parent population. For example, the target variable could be the mean salary which is larger for the subgroups gender = male, education > 15 years, and race = white (Klösgen, 1993). In contrast, given high and low income groups, our goal would be to find the differences between them which could be in gender, race, or education. Explora controls the search complexity by using redundancy filters to prune the search space. For example, a redundancy rule used is *if a node is true than its successor is false*. In the given example, this would manifest itself as not searching for any subgroups involving conjunctions with the term male; i.e. the successors are sets such as gender = male $\wedge$ education > 15 years.

Chakrabarti, Sarawagi, and Dom (1998) tackle the problem of finding surprising temporal patterns in boolean market basket data: i.e. finding itemsets whose support varies over time and cannot be explained by changes in the support of the itemset's component subsets. They use a Minimum Description Length approach where surprising patterns are those with long encoding costs. The data is segmented into distinct time periods and then a model is fit to each period so that encoding

costs can be calculated. Our work has similar goals to theirs, to find changes in data, but is fundamentally different: We find differences between two or more probability distributions, whereas they find changes in a single distribution as it varies through time. Thus a query such as "How does group A differ from B" has no meaning in their data model as different groups (distributions) do not exist. Conversely, in our model, asking for what has changed without reference to a group is nonsensical.

Ganti, Gehrke, Ramakrishnan, and Loh (1999) work on detecting the differences between datasets by examining differences between models induced on the data. They represent each model with a *structure* component which identifies regions in the feature space and a *measure* component which summarizes the data mapped to the region (e.g. fraction of examples). They aggregate and compare the measure components over the regions and at the end of the analysis they have a single number which quantifies the dissimilarity of the data sets.

7.2 Filtering Algorithms

Chakrabarti, Sarawagi, and Dom (1998) also calculate the expectation of an itemset based on all proper marginals. Our work differs in how we use these expectations. Once they have the expected supports, they use this to encode differences in an MDL framework. We use the expected values to directly compare differences in the observed probabilities to measure the size of the effect and in statistical tests to see if the expected probabilities could possibly generate the observed values.

Liu, Hsu, and Ma (1999a) summarize a rule set by only showing *direction setting* (DS) rules. The direction of a rule $X \rightarrow y$ is the sign of the correlation between X and y (the sign is 0 if the X and y are independent). A DS rule then is a rule whose direction cannot be explained by its subsets. For example, if we have the rules $X \rightarrow y$ (sign -1) and $Z \rightarrow y$ (sign -1) then if $X \wedge Z \rightarrow y$ has a sign of 1 it is surprising. This method is very fast at pruning rules, however it has a number of limitations. First, the combination rules are ad hoc and there is no method of handling situations like finding the expected sign for positive and negative combinations. Second, the method is limited to determining expected directions but not expected counts. Thus it will not be able to find combinations that have correlations that are much bigger than the individual components if the signs are the same.

Srikant and Agrawal (1996) filter quantitative association rules unless the confidence and/or support are greater than expected. For example, they would filter the rule $age(25..30) \rightarrow married$ unless it was significantly different from $age(20..30) \rightarrow married$. Their work applies to numerical

ranges which are subsets of each other.

8 Conclusion

We introduced the problem of detecting differences across several contrasting groups as that of finding all contrast sets, conjunctions of attributes and values, that have meaningfully different support levels. This allows us to answer queries of the form, "How are History and Computer Science students different?" or "What has changed from 1993 through 1998?"

We combined statistical hypothesis testing with search to develop the STUCCO algorithm for mining contrast sets. It has (1) pruning rules which allow efficient mining at low support differences, (2) guaranteed control over false positives, (3) linear trend detection, and (4) compact summarization of results.

Acknowledgments

This research was funded in part by the National Science Foundation grant IRI-9713990. We thank Nira Brand and Wagner Truppel for their comments.

Appendix A: χ^2 is convex

THEOREM 3. *The χ^2 statistic is a convex function of the observed values.*

Proof: We prove that χ^2 is convex for 2×2 contingency tables (the proof, although cumbersome, generalizes easily to $2 \times c$ tables). Expanding the summations in Equation 3 we know that for 2×2 tables

$$\chi^2 = \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{21} - E_{21})^2}{E_{21}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \frac{(O_{22} - E_{22})^2}{E_{22}} \tag{11}$$

Let $C_1 = \sum_{i=1}^{r} O_{i1}$ and $C_2 = \sum_{i=1}^{r} O_{i2}$ be the column sums in our contingency tables. These values are fixed as the column sum is constant. Then because of constraints we know that

$$O_{21} = C_1 - O_{11} \tag{12}$$

$$O_{22} = C_2 - O_{12} \tag{13}$$

$$E_{21} = C_1 - E_{11} \tag{14}$$

$$E_{22} = C_2 - E_{12} \tag{15}$$

Thus leaving us with

$$\chi^2 = \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(-O_{11} + E_{11})^2}{C_1 - E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \frac{(-O_{12} + E_{12})^2}{C_2 - E_{12}} \tag{16}$$

substituting

$$E_{11} = \frac{C_1(O_{11} + O_{12})}{N} \tag{17}$$

$$E_{12} = \frac{C_2(O_{11} + O_{12})}{N} \tag{18}$$

where $N = C_1 + C_2$, and simplifying our notation to let $o_1 = O_{11}$ and $o_2 = O_{12}$ leaves

$$\chi^2 = \frac{(o_1 - (o_1 + o_2)C_1/N)^2 N)}{(o_1 + o_2)C_1} + \frac{(-o_1 + (o_1 + o_2)C_1/N))^2}{C_1 - (o_1 + o_2)C_1/N} + \frac{(o_2 - (o_1 + o_2)C_2/N)^2 N}{(o_1 + o_2)C_2} + \frac{(-o_2 + (o_1 + o_2)C_2/N)^2}{C_2 - (o_1 + o_2)C_2/N} \tag{19}$$

To show that χ^2 is convex we must show that the Hessian (the matrix of second order partial

derivatives) $H = \nabla^2\chi^2$ is positive semi-definite (Bazaraa & Shetty, 1979). Thus,

$$H = \nabla^2\chi^2 = \begin{pmatrix} \frac{\partial^2\chi^2}{\partial {o_1}^2} & \frac{\partial^2\chi^2}{\partial o_1 \partial o_2} \\ \frac{\partial^2\chi^2}{\partial o_2 \partial o_1} & \frac{\partial^2\chi^2}{\partial {o_2}^2} \end{pmatrix} \tag{20}$$

where,

$$\frac{\partial^2\chi^2}{\partial {o_1}^2} = 2\frac{No_2^2}{C_1(o_1+o_2)^3} - 2\frac{N(o_2-C_2)^2}{C_1(-N+o_1+o_2)^3} + 2\frac{No_2^2}{C_2(o_1+o_2)^3} - 2\frac{N(o_2-C_2)^2}{C_2(-N+o_1+o_2)^3} \tag{21}$$

$$\begin{aligned} \frac{\partial^2\chi^2}{\partial o_1 \partial o_2} = \frac{\partial^2\chi^2}{\partial o_2 \partial o_1} &= -2\frac{No_1o_2}{C_1(o_1+o_2)^3} + 2\frac{N(o_2-C_2)(-C_1+o_1)}{C_1(-N+o_1+o_2)^3} - 2\frac{No_1o_2}{C_2(o_1+o_2)^3} + \\ &\quad 2\frac{N(o_2-C_2)(-C_1+o_1)}{C_2(-N+o_1+o_2)^3} \end{aligned} \tag{22}$$

$$\frac{\partial^2\chi^2}{\partial {o_2}^2} = 2\frac{No_1^2}{C_1(o_1+o_2)^3} - 2\frac{(-C_1+o_1)^2N}{C_1(-N+o_1+o_2)^3} + 2\frac{No_1^2}{C_2(o_1+o_2)^3} - 2\frac{(-C_1+o_1)^2N}{C_2(-N+o_1+o_2)^3} \tag{23}$$

Recall that H is positive semi-definite if $\mathbf{x}^T H\mathbf{x} \geq 0$ for all $\mathbf{x}$. Expanding and factoring $\mathbf{x}^T H\mathbf{x}$ gives

$$\begin{aligned} \mathbf{x}^T H\mathbf{x} &= 2\frac{N(x_1o_2 - x_2o_1)^2}{C_1(o_1+o_2)^3} + 2\frac{N(x_1(o_2-C_2) - x_2(o_1-C_1))^2}{C_1(N-o_1-o_2)^3} + 2\frac{N(x_1o_2 - x_2o_1)^2}{C_2(o_1+o_2)^3} + \\ &\quad 2\frac{N(x_1(o_2-C_2) - x_2(o_1-C_1))^2}{C_2(N-o_1-o_2)^3} \end{aligned} \tag{24}$$

Terms 1 and 3 are always positive. Terms 2 and 4 are also always positive as $N - o_1 - o_2 \geq 0$ (recall that $N = C_1 + C_2$ and that $C_1 \geq o_1$ and $C_2 \geq o_2$). Thus H is positive semi-definite and therefore χ^2 is convex.□

Appendix B: Estimating the Scale Factor

We estimate the scale factor (ρ) used in Section 4.1 with maximum likelihood under a binomial model.

Let n_i be the total number of observations from group i and let o_i be the number of observations meeting the contrast set criteria from group i. Let λ_i be the expected probability for group i from our initial log-linear model (for convenience we treat this as a fixed value). Then the probability of

obtaining the observed counts as a function of ρ is:

$$P(\rho) = \prod_{i=1}^{G} \left(\begin{array}{c} n_i \\ o_i \end{array} \right) (\rho\lambda_i)^{o_i} (1 - \rho\lambda_i)^{n_i - o_i} \tag{25}$$

Taking logarithms to get the log-likelihood:

$$l(\rho) = \log P(\rho) = \sum_{i=1}^{G} \log \left(\begin{array}{c} n_i \\ o_i \end{array} \right) + o_i \log(\rho\lambda_i) + (n_i - o_i) \log(1 - \rho\lambda_i) \tag{26}$$

Taking partial derivatives with respect to ρ

$$\frac{\partial l(\rho)}{\partial \rho} = \sum_{i=1}^{G} \frac{o_i}{\rho} - \frac{\lambda_i (n_i - o_i)}{1 - \rho\lambda_i} \tag{27}$$

For two groups, we solve for the maximum ($\frac{\partial l(\rho)}{\partial \rho} = 0$) exactly by using the quadratic formula. For more than two groups, exact solution becomes cumbersome and thus we use an iterative solver.

Appendix C: STUCCO Results for Biology 1997-98

This appendix contains the rules found by STUCCO which differentiate the 1997-98 Biology students who chose to, or not to enroll at UCI. There were 700 students who enrolled and 1954 who did not resulting in an average yield of 26.4%.

Positive Yield Rules

1. Students who live within 30 miles of UCI are more likely to enroll with a 13% higher yield than average. This represents a gain of 125 students.
2. Students who scored less than 500 on their SAT Verbal are more likely to enroll with a 18% higher yield than average. This represents a gain of 104 students.
3. Students who have a SIN between 6000 and 6500 are more likely to enroll with a 16% higher yield than average. This represents a gain of 99 students.
4. Students who scored between 500 and 600 on their SAT Math are more likely to enroll with a 11% higher yield than average. This represents a gain of 86 students.
5. Students who are from Orange County are more likely to enroll with a 17% higher yield than average. This represents a gain of 85 students.
6. Students who are from Orange County and live within 30 miles of UCI are more likely to enroll with a 17% higher yield than average. This represents a gain of 82 students.
7. Students who have a GPA between 3.5 and 4 are more likely to enroll with a 5% higher yield than average. This represents a gain of 62 students.
8. Students who have a GPA between 2.75 and 3.5 are more likely to enroll with a 23% higher yield than average. This represents a gain of 62 students.

9. Students who have a SIN between 6000 and 6500 and scored between 500 and 600 on their SAT Math are more likely to enroll with a 19% higher yield than average. This represents a gain of 61 students.
10. Students who left the ethnicity blank are more likely to enroll with a 18% higher yield than average. This represents a gain of 54 students.
11. Students who scored less than 500 on their SAT Verbal and have a SIN between 6000 and 6500 are more likely to enroll with a 20% higher yield than average. This represents a gain of 53 students.
12. Students who have a SIN between 5000 and 6000 are more likely to enroll with a 23% higher yield than average. This represents a gain of 52 students.
13. Students who are not born in the US are more likely to enroll with a 5% higher yield than average. This represents a gain of 47 students.
14. Students who scored less than 500 on their SAT Math are more likely to enroll with a 17% higher yield than average. This represents a gain of 42 students.
15. Students who scored less than 500 on their SAT Verbal, are not native English speakers, and are not born in the US are more likely to enroll with a 15% higher yield than average. This represents a gain of 34 students.
16. Students who scored less than 500 on their SAT Math and scored less than 500 on their SAT Verbal are more likely to enroll with a 20% higher yield than average. This represents a gain of 31 students.
17. Students who live between 30 and 100 miles away from another UC school and live within 30 miles of UCI are more likely to enroll with a 26% higher yield than average. This represents a gain of 31 students.
18. Students who have a SIN between 5000 and 6000, scored less than 500 on their SAT Verbal, and live within 30 miles of another UC school are more likely to enroll with a 28% higher yield than average. This represents a gain of 30 students.
19. Students who have a SIN between 5000 and 6000 and have a GPA between 2.75 and 3.5 are more likely to enroll with a 27% higher yield than average. This represents a gain of 30 students.
20. Students who have a SIN between 5000 and 6000 and scored less than 500 on their SAT Math are more likely to enroll with a 26% higher yield than average. This represents a gain of 26 students.
21. Students who have a GPA between 2.75 and 3.5 and have a SIN between 6000 and 6500 are more likely to enroll with a 22% higher yield than average. This represents a gain of 26 students.
22. Students who are Filipino are more likely to enroll with a 8% higher yield than average. This represents a gain of 22 students.
23. Students who scored less than 500 on their SAT Verbal, scored between 500 and 600 on their SAT Math, and have a parental income of less than $35000 are more likely to enroll with a 17% higher yield than average. This represents a gain of 22 students.
24. Students who scored between 650 and 700 on their SAT Math, have a SIN between 6500 and 7000, and live within 30 miles of UCI are more likely to enroll with a 26% higher yield than average. This represents a gain of 18 students.
25. Students who have a SIN between 5000 and 6000, have a GPA between 2.75 and 3.5, and scored less than 500 on their SAT Verbal are more likely to enroll with a 26% higher yield than average. This represents a gain of 17 students.
26. Students who scored between 650 and 700 on their SAT Math, are from Orange County, and have a SIN between 6500 and 7000 are more likely to enroll with a 38% higher yield than average. This represents a gain of 16 students.

Negative Yield Rules

1. Students who have a SIN greater than 7000 are less likely to enroll with a 16% lower yield than average. This represents a loss of 142 students.
2. Students who have a GPA greater than 4 are less likely to enroll with a 11% lower yield than average. This represents a loss of 124 students.
3. Students who have a SIN greater than 7000 and have a GPA greater than 4 are less likely to enroll with a 15% lower yield than average. This represents a loss of 103 students.
4. Students who live more than 100 miles from UCI are less likely to enroll with a 12% lower yield than average. This represents a loss of 80 students.

5. Students who scored greater than 700 on their SAT Math and have a SIN greater than 7000 are less likely to enroll with a 15% lower yield than average. This represents a loss of 55 students.
6. Students who scored between 600 and 650 on their SAT Verbal are less likely to enroll with a 11% lower yield than average. This represents a loss of 55 students.
7. Students who scored greater than 700 on their SAT Math are less likely to enroll with a 12% lower yield than average. This represents a loss of 53 students.
8. Students who scored between 650 and 700 on their SAT Math are less likely to enroll with a 8% lower yield than average. This represents a loss of 49 students.
9. Students who are born in the US are less likely to enroll with a 3% lower yield than average. This represents a loss of 48 students.
10. Students who scored between 650 and 700 on their SAT Verbal are less likely to enroll with a 14% lower yield than average. This represents a loss of 46 students.
11. Students who live between 30 and 100 miles away from UCI are less likely to enroll with a 5% lower yield than average. This represents a loss of 44 students.
12. Students who scored between 650 and 700 on their SAT Verbal and have a SIN greater than 7000 are less likely to enroll with a 16% lower yield than average. This represents a loss of 40 students.
13. Students who have a parental income greater than $80000 are less likely to enroll with a 6% lower yield than average. This represents a loss of 36 students.
14. Students who have a SIN greater than 7000, are not born in the US, and are from Los Angeles County are less likely to enroll with a 24% lower yield than average. This represents a loss of 32 students.
15. Students who are Chinese are less likely to enroll with a 7% lower yield than average. This represents a loss of 30 students.
16. Students who are from Santa Clara County are less likely to enroll with a 19% lower yield than average. This represents a loss of 22 students.
17. Students who scored greater than 700 on their SAT Verbal are less likely to enroll with a 12% lower yield than average. This represents a loss of 21 students.
18. Students who scored greater than 700 on their SAT Verbal and have a SIN greater than 7000 are less likely to enroll with a 13% lower yield than average. This represents a loss of 21 students.
19. Students who are from San Diego County are less likely to enroll with a 11% lower yield than average. This represents a loss of 20 students.

Appendix D: C5 Results for Biology 1997-98

This appendix contains the rules found by C5 which differentiate the 1997-98 Biology students who chose to, or not to enroll at UCI. Rules which affected 5 or fewer students are not shown for space reasons.

Positive Yield Rules

1. Students who attended a public high school and have a SIN between 6000 and 6500 are more likely to enroll with a 19% higher yield than average. This represents a gain of 94 students.
2. Students who have a SIN between 6500 and 7000 and live within 30 miles of UCI are more likely to enroll with a 19% higher yield than average. This represents a gain of 59 students.
3. Students who have a SIN between 5000 and 6000 are more likely to enroll with a 23% higher yield than average. This represents a gain of 52 students.
4. Students who are from Los Angeles County and have a SIN between 6000 and 6500 are more likely to enroll with a 18% higher yield than average. This represents a gain of 46 students.
5. Students who are not born in the US, have a SIN between 6500 and 7000, and live within 30 miles of UCI are more likely to enroll with a 28% higher yield than average. This represents a gain of 42 students.

6. Students who are from Orange County, and have a SIN between 6000 and 6500 are more likely to enroll with a 43% higher yield than average. This represents a gain of 42 students.
7. Students who are male, are Filipino, and have a SIN between 6000 and 6500 are more likely to enroll with a 40% higher yield than average. This represents a gain of 11 students.
8. Students who are from Los Angeles County, have a SIN between 6500 and 7000, and have a parental income between $35000 and $55000 are more likely to enroll with a 8% higher yield than average. This represents a gain of 8 students.
9. Students who left the ethnicity blank, are born in the US, have a SIN between 6500 and 7000, and live within 30 miles of UCI are more likely to enroll with a 49% higher yield than average. This represents a gain of 8 students.
10. Students who are Chicano, are from Los Angeles County, are born in the US, and live within 30 miles of UCI are more likely to enroll with a 14% higher yield than average. This represents a gain of 7 students.
11. Students who stated their ethnicity as "other", have a SIN between 6500 and 7000, and live within 30 miles of UCI are more likely to enroll with a 55% higher yield than average. This represents a gain of 6 students.

An additional 16 rules affect 5 or fewer students.

Negative Yield Rules

1. Students who have a SIN greater than 7000 are less likely to enroll with a 16% lower yield than average. This represents a loss of 142 students.
2. Students who have a GPA greater than 4 are less likely to enroll with a 11% lower yield than average. This represents a loss of 124 students.
3. Students who have a SIN between 6500 and 7000 and live more than 100 miles from UCI are less likely to enroll with a 18% lower yield than average. This represents a loss of 47 students.
4. Students who have a GPA between 3.5 and 4 and scored between 650 and 700 on their SAT Verbal are less likely to enroll with a 15% lower yield than average. This represents a loss of 17 students.
5. Students who have a SIN between 6500 and 7000, have a parental income of less than $35000, and live between 30 and 100 miles away from UCI are less likely to enroll with a 16% lower yield than average. This represents a loss of 16 students.
6. Students who are from San Diego County and have a SIN between 6500 and 7000 are less likely to enroll with a 20% lower yield than average. This represents a loss of 13 students.
7. Students who are from Santa Clara County, and scored between 650 and 700 on their SAT Math are less likely to enroll with a 26% lower yield than average. This represents a loss of 10 students.
8. Students who are from Alameda County, and have a GPA between 3.5 and 4 are less likely to enroll with a 24% lower yield than average. This represents a loss of 9 students.
9. Students who are from Sacramento County are less likely to enroll with a 20% lower yield than average. This represents a loss of 7 students.
10. Students who are East Indian or Pakistani, and are from Los Angeles County are less likely to enroll with a 11% lower yield than average. This represents a loss of 6 students.
11. Students who are from San Francisco County are less likely to enroll with a 26% lower yield than average. This represents a loss of 6 students.
12. Students who are Caucasian, are from San Diego County, and attended a public high school are less likely to enroll with a 18% lower yield than average. This represents a loss of 6 students.
13. Students who are from San Mateo County are less likely to enroll with a 12% lower yield than average. This represents a loss of 6 students.
14. Students who are African American, are from Los Angeles County, and have a SIN between 6000 and 6500 are less likely to enroll with a 26% lower yield than average. This represents a loss of 6 students.

An additional 24 rules affect 5 or fewer students.

References

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining associations between sets of items in massive databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data* (pp. 207–216).

Agrawal, R., Psaila, G., Wimmers, E., & Zait, M. (1995). Querying shapes of histories. *Proceedings of the 21st International Conference on Very Large Databases.*

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of the 20th International Conference on Very Large Databases.*

Agresti, A. (1990). *Categorical data analysis.* John Wiley & Sons.

Bay, S. D. (1999). The UCI KDD archive. [http://kdd.ics.uci.edu/]. Irvine, CA: University of California, Department of Information and Computer Science.

Bay, S. D., & Pazzani, M. J. (1999). Detecting change in categorical data: Mining contrast sets. *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 302–306).

Bayardo, R. J. (1998). Efficiently mining long patterns from databases. *Proceedings of the ACM SIGMOD Conference on Management of Data.*

Bayardo, R. J., Agrawal, R., & Gunopulos, D. (1999). Constraint-based rule mining in large, dense databases. *Proceedings 15th International Conference on Data Engineering.*

Bazaraa, M. S., & Shetty, C. M. (1979). *Nonlinear programming: Theory and algorithms.* John Wiley & Sons.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice.* The MIT Press.

Blake, C., & Merz, C. J. (1998). UCI repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science. [http://www.ics.uci.edu/~mlearn/MLRepository.html].

Brin, S., Motwani, R., Ullman, J. D., & Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. *Proceedings of the ACM SIGMOD International Conference on Management of Data* (pp. 255–264).

Chakrabarti, S., Sarawagi, S., & Dom, B. (1998). Mining surprising patterns using temporal description length. *Proceedings of the 24th International Conference on Very Large Databases.*

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304–1312.

Darity, W. A. (1998). Intergroup disparity: Economic theory and social science evidence. *Southern Economic Journal, 64,* 805–826.

Davies, J., & Billman, D. (1996). Hierarchical categorization and the effects of contrast inconsistency in an unsupervised learning task. *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society* (p. 750).

Dong, G., & Li, J. (1999). Efficient mining of emerging patterns: Discovering trends and differences. *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Everitt, B. S. (1992). *The analysis of contingency tables.* Chapman and Hall. second edition.

Ganti, V., Gehrke, J. E., Ramakrishnan, R., & Loh, W. (1999). A framework for measuring changes in data characteristics. *Proceedings of Eighteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems.*

Glenn, N. D. (1977). *Cohort analysis.* Newbury Park, CA: Sage Publications.

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures.* John Wiley & Sons.

Hoschka, P., & Klösgen, W. (1991). A support system for interpreting statistical data. In G. Piatetsky-Shapiro and W. J. Frawley (Eds.), *Knowledge discovery in databases*, 325–346. AAAI Press.

Keogh, E., & Pazzani, M. J. (1998). An enhanced representation of time series that allows fast and accuracte classification, clustering, and relevance feedback. *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining.*

Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., & Verkamo, A. I. (1994). Finding interesting rules from large sets of discovered association rules. *Proceedings of the Third International Conference on Information and Knowledge Management* (pp. 401–407).

Klösgen, W. (1993). Explora user documentation: A support system for discovery in databases.

Klösgen, W. (1996). Explora: A multipattern and multistrategy discovery assistant. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (Eds.), *Advances in knowledge discovery and data mining*, 249–271. AAAI Press / The MIT Press.

Knoke, D., & Burke, P. J. (1980). *Log-linear models.* Newbury Park, CA: Sage Publications.

Lewontin, R. C., & Felsenstein, J. (1965). The robustness of homogeneity in $2 \times n$ tables. *Biometrics*, *21*, 19–33.

Lin, D., & Kedem, Z. M. (1998). Pincer-search: A new algorithm for discovering the maximum frequent set. *Proceedings of the Sixth European Conference on Extending Database Theory.*

Lincoff, G. H. (1981). *The audubon scoiety field guide to north american mushrooms.* Random House.

Liu, B., & Hsu, W. (1996). Post-analysis of learned rules. *Proceedings of the Thirteenth National Conference on Artificial Intelligence* (pp. 828–834).

Liu, B., Hsu, W., & Chen, S. (1997). Using general impressions to analyze discovered classification rules. *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining* (pp. 31–36).

Liu, B., Hsu, W., & Ma, Y. (1999a). Pruning and summarizing the discovered associations. *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Liu, B., Hsu, W., Mun, L., & Lee, H. (1999b). Finding interesting patterns using user expectations. *IEEE Transactions on Knowledge and Data Engineering, 11.*

Mannila, H., & Toivonen, H. (1997). Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery, 1*, 241–258.

Megiddo, N., & Srikant, R. (1998). Discovering predictive association rules. *Proceedings of the 4th International Conference on Knowledge Discovery in Databases and Data Mining.*

Menard, S. (1991). *Longitudinal research.* Newbury Park, CA: Sage Publications.

Michell, T. M. (1977). Version spaces: A candidate elimination approach to rule learning. *Proc. of the 5th Int'l Joint Conf. on Artificial Intelligence.*

Ng, R., Lakshmanan, L. V. S., Han, J., & Pang, A. (1998). Exploratory mining and pruning optimizations of constrained associations rules. *Proceedings of the ACM SIGMOD Conference on Management of Data.*

Padmanabhan, B., & Tuzhilin, A. (1998). A belief-driven method for discovering unexpected patterns. *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining.*

Riddle, P., Segal, R., & Etzioni, O. (1994). Representation design and brute-force induction in a boeing manufacturing domain. *Applied Artificial Intelligence, 8*, 125–147.

Ruggles, S. (1995). Sample designs and sampling errors. *Historical Methods, 28*, 40–46.

Ruggles, S. (1997). The rise of divorce and separation in the united states, 1880-1990. *Demography, 34*, 455–466.

Ruggles, S., & Sobek, M. (1997). Integrated public use microdata series: Version 2.0. [http://www.ipums.umn.edu/].

Rymon, R. (1992). Search through systematic set enumeration. *Third International Conference on Principles of Knowledge Representation and Reasoning.*

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review Psychology, 46*, 561–584.

Silberschatz, A., & Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering, 8.*

Silverstein, C., Brin, S., & Motwani, R. (1998). Beyond market baskets: Generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery, 2*, 39–68.

Srikant, R., & Agrawal, R. (1996). Mining quantitative association rules in large relational tables. *Proceedings of the ACM SIGMOD Conference on Management of Data.*

Srikant, R., Vu, Q., & Agrawal, R. (1997). Mining association rules with item constraints. *Proceedings of the 3rd International Conference on Knowledge Discovery in Databases and Data Mining.*

Zaki, M. J., Parthasarathy, S., Ogihara, M., & Li, W. (1997). New algorithms for fast discovery of association rules. *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining.*

Stephen D. Bay received his B.A.Sc. and M.A.Sc. from the Department of Systems Design Engineering at the University of Waterloo. He is currently completing his Ph.D. in the Department of Information and Computer Science at the University of California, Irvine. His research interests include data mining and machine learning.

Michael J. Pazzani is a Professor and former chair of the Information and Computer Science Department at the University of California, Irvine. He received his M.S. and B.S. in Computer Engineering from the University of Connecticut and his Ph.D. in Computer Science from UCLA. He is a member of AAAI and the Cognitive Science Society. His research interests include data mining and intelligent agents.