Discovering and Describing Category Differences: What makes a discovered difference insightful?

Stephen D. Bay (sbay@ics.uci.edu)
Michael J. Pazzani (pazzani@ics.uci.edu)
Department of Information and Computer Science
University of California, Irvine
Irvine, CA 92697, USA

Abstract

Many organizations have turned to computer analysis of their data to deal with the explosion of available electronic data. The goal of this analysis is to gain insight and new knowledge about their core activities. A common query is comparing several different categories (e.g., customers who default on loans versus those that don't) to discover previously unknown differences between them. Current mining algorithms can produce rules which differentiate the groups with high accuracy, but often human domain experts find these results neither insightful nor useful. In this paper, we take a step toward understanding how humans interpret discovered rules by presenting a case study: we compare the responses of admissions officers (domain experts) on the output of two data mining algorithms which attempt to find out why admitted students choose to enroll or not enroll at UC Irvine. We analyze the responses and identify several factors that affect what makes the discovered rules insightful.

Introduction

Data collection is a daily activity of many organizations in business, science, education, and medicine. Large databases are routinely collected and with the advent of computers to process the information, these organizations want to analyze the data to gain insight and knowledge about the underlying process behind the data. The data usually represents information on their core business, and an important task is understanding the differences between various client groups. For example, bank loan officers may be interested in analyzing historical loan data to understand the differences between people who are good and poor credit risks. Admissions officers at UC Irvine (UCI) are interested in analyzing admissions data to understand the factors which influence an admitted student's choice to enroll at UCI. It is important that the discovered differences both be true and accurate descriptions of the data as well as being acceptable and understandable by the end users.

A common technique for discovering group differences from data is to apply a data mining algorithm to automatically find rules from the data. For example, after analyzing loan data we might find that people with graduate degrees are good loan risks (i.e. grad-degree $\rightarrow$ low-risk). There have been many studies which investigate the accuracy of rules that describe category differences, but very few which investigate how humans interpret the results.

In this paper, we focus on two issues relating to the interpretation of discovered rules by human domain experts: First, algorithms for automatically finding group differences can be categorized broadly into discriminative and characteristic (or informative) approaches (Rubinstein & Hastie, 1997). In discriminative approaches, the algorithms attempt to find differences that can be directly used to classify the instances of the groups. In characteristic approaches, the algorithms attempt to find differences in the class descriptions, some of which may also be highly predictive but are not necessarily so. We investigate if human domain experts have a preference for either strategy. Second, there are many objective measures of rule quality and typically mining algorithms seek rules that optimize these measures. For example, with if-then rules of the form $A \rightarrow C$ (antecedent implies consequent), many algorithms attempt to maximize the confidence which is the conditional probability of the consequent being true given the antecedent ($P(C|A)$). The assumption is that rules that score highly on the objective measure are useful to domain experts. The problem is that while there are many objectives measures of pattern quality, such as support (Agrawal, Imielinski, & Swami, 1993), confidence (Agrawal et al., 1993), lift (also known as interest) (Brin, Motwani, Ullman, & Tsur, 1997), conviction (Brin et al., 1997) and many others, none of the measures truly correlate with what human domain experts find interesting, useful, or acceptable. The reality is that most mined results are not useful at all. For example, Major and Mangano (1995) analyzed rules from a hurricane database and reduced 161 rules to 10 "genuinely interesting" rules. In a more extreme, but common case, Brin et al. found over 20000 rules on a census database from which they learned that "five year olds don't work, unemployed residents don't earn income from work, men don't give birth" and other uninteresting facts. Thus we investigate the relationship between human subjective measures of rule usefulness to objective measures of rule quality.

We answer our research questions, "Is a discriminative or characteristic approach more useful for describing group differences?" and "How do subjective and objective measures of rule interest relate to each other?" by reporting on an analysis of discovered rules by human domain experts. We analyzed UCI admissions data to understand the groups of students that decide to enroll or not enroll at UCI given an offer of admission. After discovering rules with two different algorithms, we then showed the rules to human domain experts and asked them to rate the rules according their *insightfulness*, i.e. did the rule expand their knowledge about the admission process? After obtaining experts results, we then analyzed the responses to compare and contrast discriminative and characteristic approaches as well as objective and subjective measures of rule quality.

In the remainder of this paper, we first highlight the differences between discriminative and characteristic approaches and describe the mining algorithms used. We then describe the knowledge discovery task: analyzing admissions data to improve the recruitment process at UCI. We examine domain experts responses and compare them to quantitative measures of rule quality. We conclude by discussing related work and examining possible directions for future work.

Background: Discovering Category Differences

Mining algorithms for finding category or group differences can be classified as discriminative or characteristic. Discriminative miners attempt to find differences that are useful for predictive classification with a high degree of accuracy. Characteristic miners attempt to find significant differences in the class descriptions. This can result in rules that are highly predictive as with discriminative mining, but predictiveness is not a requirement of the mined rules. Discriminative miners look for one key set of features that distinguish the categories while characteristic miners look for all important differences between the categories.

For example, a discriminative difference would be that students who do not enroll at UCI are much more likely to have a GPA greater than 4 and live more than 100 miles from UCI than students who do enroll. Ninety eight percent of these students (GPA greater than 4 and distance from UCI greater than 100 miles) reject UCI's admission offer and do not enroll. Knowing that a student has these characteristics allows us to classify them with high accuracy.

A characteristic difference would be that 39.8% of students that enroll at UCI are English native speakers compared with 47.9% of students who do not enroll. Clearly this difference affects many students, but knowing that a student is an English native speaker does not give us much information about whether the student will enroll or not. It contains information that is not useful for prediction, but nevertheless may be important to an analyst attempting to understand the two groups.

Formally, we can describe the two approaches as follows: Let X be the set of attributes and values with which we describe the differences. X can be a single attribute value pair such as $X = \{\mathsf{nativelanguage} = \mathsf{English}\}$ or it can be a conjunction, e.g. $X = \{\mathsf{GPA} > 4 \wedge \mathsf{UCIdistance} > 100\ \mathsf{miles}\}$. Let y be the class or category. Then discriminative approaches attempt to find X such that the following equation is maximized. [1]

$$|P(y = c_1|X) - P(y = c_2|X)| \quad (1)$$

Characteristic approaches attempt to find X such that

$$|P(X|y = c_1) - P(X|y = c_2)| \quad (2)$$

is maximized. Note that we can relate these two equations with Bayes Rule:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)} \quad (3)$$

[1]The exact form of the equation that is maximized can vary somewhat from this definition, but all discriminative approaches concentrate on finding large differences in $P(y|X)$.

Thus we can always convert from one to the other. Although the forms can be made equivalent, the difference is that the X that optimizes/maximizes Equation 1 is not necessarily the same as the X that is best for Equation 2.

We now describe two algorithms representative of the approaches. C5 (Quinlan, 1993) which is a discriminative approach and STUCCO (Bay & Pazzani, 1999) which is a characteristic approach.

A Discriminative Approach: C5

A discriminative approach to distinguishing two or more groups from each other is to use a rule learner or decision tree to learn a classification strategy. In this paper, we use the program C5 which is an updated version of C4.5 (Quinlan, 1993). It is a workhorse of the Machine Learning community and is a gold standard to which many new algorithms are compared.

Given two categories c_1 and c_2, C5 attempts to find sets of variables such that Equation 1 is maximized and so that as many examples in the database are covered by rules as possible. C5 performs greedy heuristic search to develop a decision tree. Starting at the root of the tree, C5 selects an attribute-value test to partition the feature space. Each partition is represented by a child node and is then recursively divided with more tests. The tests are chosen to create child nodes which tend to be mainly of one class.

After finding the tree, C5 can then convert it to rules and remove unnecessary terms imposed by the top down tree structure. It does this by following the path from the root to every leaf and constructing one rule for each path. The rules contain every term that appears in nodes along the path. C5 then tests each term that appears in a rule and removes terms that offer no predictive benefit.

A Characteristic Approach: STUCCO

Here we briefly review the STUCCO algorithm for mining contrast sets. The reader is directed to (Bay & Pazzani, 1999, 1999b) for a more detailed description.

STUCCO is a complete mining algorithm that searches for contrast sets, conjunctions of attribute-value pairs, that have substantially different probabilities across several distributions or groups. The goal is to find contrast sets where the value of Equation 2 is greater than a threshold δ.

STUCCO takes a two stage approach to mining. In the first stage, STUCCO searches for all possible contrast sets that meet the criteria. In the second stage, STUCCO summarizes the mined results to present only a small set of rules.

STUCCO organizes the search for contrast sets using set-enumeration trees (Rymon, 1992) to ensure that every node is visited only once or not at all if it can be pruned. Figure 1 shows an example set-enumeration tree for four attribute-value pairs. STUCCO searches this tree using breadth-first search; it starts with the most general terms first, i.e. those contrast sets with a single attribute-value pair such as $\mathsf{sex} = \mathsf{female}$ or $\mathsf{UCISchool} = \mathsf{Engineering}$. These sets are the easiest to understand and will have the largest support. It then progresses to more complicated sets that involve conjunctions of terms, for example, $\mathsf{sex} = \mathsf{female} \wedge \mathsf{UCISchool} = \mathsf{Engineering}$.

During search, STUCCO scans the database to count the support of all nodes for each group. It examines the counts

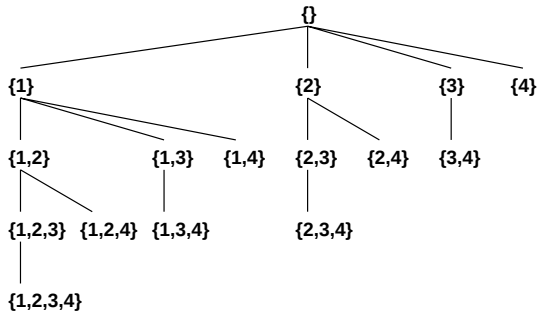

Figure 1: Example search tree for four attribute-values pairs $\{1,2,3,4\}$.

to determine which nodes meet the criteria and which nodes should be pruned. STUCCO also explicitly controls the search error to limit false discoveries by keeping careful track of the number of statistical tests made to verify Equation 2 and adjusting the α level for individual tests to control the overall Type I error rate.

In the second stage, STUCCO summarizes the mined results by showing the user the most general contrast sets first, those involving a single term, and then only showing more complicated conjunctions if they are surprising based on the previously shown sets. For example, we might start by showing the contrast sets $\mathsf{sex} = \mathsf{female}$, $\mathsf{UCISchool} = \mathsf{Engineering}$, and $\mathsf{GPA} > 4$. STUCCO would then move on to showing more complicated sets such as $\mathsf{sex} = \mathsf{female} \wedge \mathsf{UCISchool} = \mathsf{Engineering}$ or $\mathsf{UCISchool} = \mathsf{Engineering} \wedge \mathsf{GPA} > 4$, and finally $\mathsf{sex} = \mathsf{female} \wedge \mathsf{UCISchool} = \mathsf{Engineering} \wedge \mathsf{GPA} > 4$. The conjunctions are only shown if their frequencies could not be predicted from the subsets using a log-linear model (Everitt, 1992). This hierarchical approach eliminates many uninteresting results and can reduce the number of mined results by more than an order of magnitude leaving a small set of rules for a user to view.

Analysis of UCI Admissions Data

At UCI, the admissions office collects data on all undergraduate applicants. 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 that would be provided by a data mining algorithm.

Here, we report on an analysis of the 1999-2000 enrollment data to identify differences between students who chose to enroll and those who did not for all students accepted at UCI. There were a total of 13344 students given admission offers, of which 3871 accepted and enrolled at UCI and 9473 who did not. For each student, the data contains information on variables such as ethnicity, UCI School (e.g. Arts, Engineering, etc.), sex, 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. Numeric variables, such as SAT scores and distances were manually converted into nominal variables at thresholds that are meaningful for the admissions office.

We ran STUCCO and C5 on the data to obtain contrast sets. For STUCCO we used the following parameter settings: $\delta = 1\%$ and global $\alpha = 1$. For C5 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 fail to find any rules distinguishing the two groups and would resort to a default strategy of always predicting that the students would not enroll (the more common class).

Both C5 and STUCCO produce results in their own particular format. To make interpretation easier and to eliminate any bias from the presentation format, we converted the results into an equivalent set of English sentences describing the differences using an identical sentence structure for both C5 and STUCCO. We translated the numeric results associated with the outputs of STUCCO and C5 into *yield* and *gain* which are meaningful quantities for the admissions officers. 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 results can be ordered by gain to highlight the differences that have the largest effect. Rule 1 shows a sample result converted automatically to English text. The full set of results are too big to be shown in this paper, but Appendix A presents a small subset and all examples used in this paper are actual findings.

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

We expressed yield relative to the average. In this example, the yield was 30% higher than average yield (25.6%) which is the percentage of all students who accepted UCI's offer. Thus the yield for this category is 55.6% = 30% + 25.6%. The gain of 66 students is a measure of the *effect size* of the rule, i.e. how many students does it affect? In general, we believe the effect size is a domain independent factor that contibutes to how insightful a rule is. For example, in a loan default problem the effect size would indicate how many additional loans that default can be attributed to customers that meet the conditions of the discovered rule.

The discriminative and characteristic approaches resulted in two very different rules sets describing the differences between students who enroll and do not enroll. Table 1 shows the size (as measured by the number of terms in the conjunction) and number of rules mined by C5 and STUCCO. It shows the results for all of the mined rules and the best 30 as measured by gain (we used only the best 30 rules for our experiment in the next section). 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. Table 2

summarizes the average gain (magnitude only) of the results and in Figure 2 we plot the discovered rules according to their yield difference (with respect to the average yield) and gain. We summarize the differences as follows:

- C5 tends to produce longer rules than STUCCO. The average number of terms in a C5 rule was 3 whereas for STUCCO the average is 1.7.
- Compared with STUCCO, C5 produced rules with higher yield differences but smaller effect sizes.
- C5 produces many rules that are clearly uninteresting because their effect size is very small. For example, on the full set of returned rules the median effect size was only 5 for C5. In contrast, STUCCO's median was 132.

Table 1: Summary of Results for C5 and STUCCO (S).

	All		Best 30	
Size	C5	S	C5	S
1	25	42	6	17
2	46	24	16	11
3	62	10	6	2
4	35	1	2	
5	21	1		
6	5			
total	194	78	30	30

Table 2: Summary of Effect Size Results for C5 and STUCCO (S). Values were calculated based on magnitude only.

	All		Best 30	
	C5	S	C5	S
median	5	132	125	273
mean	32	175	173	306
min	0	28	34	165
max	495	683	495	683

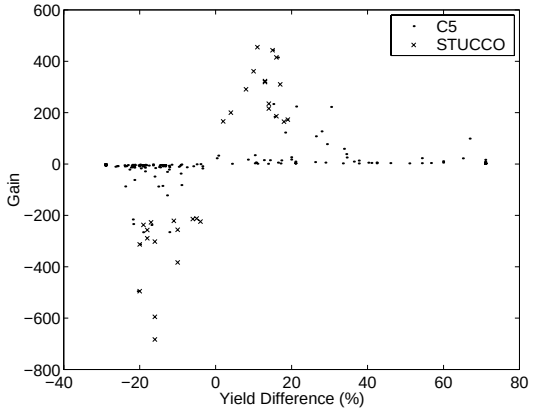

Figure 2: Comparing discriminative and characteristic rule sets.

Experimental Evaluation

In this experiment, we showed the results from data mining to admissions officers at UCI and asked them to rate the rules according to the insight they provide.

To give an example of an insightful rule, consider Rule 2. Admissions officers at UCI have 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. Rule 2 provides insight into this: it suggests that UCI competes fairly well for students with UCLA, UCSD and UC Riverside.

> *Rule 2.* 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.

As another example of what an admissions officer would find insightful, consider Rule 3. It 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 bright students who wanted to explore their options. Due to this discovery, UCI is changing the way it approaches recruiting undeclared students, particularly those with high GPAs.

> *Rule 3.* 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.

Subjects. The subjects were 4 faculty and staff at the University of California, Irvine who are actively involved in the admissions process and expressed an interest in viewing the results of a computer analysis of admissions data to find factors relating to student recruitment. The subjects did not receive any compensation.

Stimuli. The stimuli consisted of two sets of statements corresponding to the outputs of C5 and STUCCO. Each set consisted of 30 rules, 15 describing students with increased yield and 15 describing students with decreased yield. Appendix A shows the 15 increasing yield statements for STUCCO. Within the group of increasing or decreasing yield statements the rules were sorted by gain (largest first). The yield and gain values were rounded to the nearest integer. The subjects were not aware of the algorithm that generated each set of rules.

Procedures. Each subject was shown the two sets of rules and were asked to "consider the statements in the context of being an admissions officer whose goal is to improve the quality, quantity, and diversity of students that enroll." We then asked the subjects to rate each statement on its insightfulness using a scale from -3 to +3, with -3 corresponding to not insightful and +3 corresponding to insightful. After viewing both sets of statements, the subjects were asked to indicate which set they preferred overall.

Results & Discussion. Table 3 shows the mean ratings of the experts for both STUCCO and C5. It is clear from the values that the experts are using different scales and that the rules were not equally insightful to all. Experts 1, 2, and 4 rated STUCCO higher than C5. Using a group t test, the differences were significant for E2 and E4 at the 0.001 level. The difference in ratings were not significant for Experts 1 and 3.

Expert 1 indicated no preference for STUCCO or C5, but the remainder all stated that they preferred the rules which were learned by STUCCO.

Table 3: Mean Ratings for C5 and STUCCO

	Expert			
	E1	E2	E3	E4
STUCCO	0.7	-0.1	1.57	1.23
C5	0.43	-0.87	1.87	0.23
t(58)	0.69	3.48	-1.38	4.87

We pooled the STUCCO and C5 ratings for each expert and then calculated the correlation of the experts ratings with the objective measures of rule quality: yield difference, gain (effect size), and rule size (number of conjuncts). For our calculations we used the magnitude of the yield difference and gain. The results are shown in Table 4. For yield difference and rule size there were no significant correlations with insightfulness. For gain, we found a significant relation for Experts 1 and 2. With a t test of a correlation coefficient (H_0: $\rho = 0$), the results are significant at the 0.01 level or better. This suggests that the effect size is a factor in determining how insightful domain experts find the discovered differences. Since STUCCO finds many rules with large effect sizes, experts seem to prefer the STUCCO rules.

Table 4: Correlation of Ratings to Objective Rule Measures

	Expert			
	E1	E2	E3	E4
Yield Diff.	0.1406	-0.0827	0.1664	-0.1139
Gain	0.5598	0.4316	0.2469	0.0848
Rule Size	-0.2449	-0.2224	0.0787	-0.2620

We believe there are other factors that influence the ratings of insight given to a rule. In particular, some rules are already well known to the admissions officers. In addition, some admission officers have a particular focus (e.g., minority students) and would be more interested in rules of that type. We tabulated the inter-correlation of the experts in Table 5 using the pooled C5 and STUCCO responses. The results are surprisingly in that the correlation between experts is very low. Figure 3 plots the ratings of E2 and E4, the experts with the highest correlation. This suggests that insight is very subjective and there are important individual differences, possibly relating to the prior knowledge of the task.

Table 5: Correlation of Ratings Between Experts

	Expert		
	E2	E3	E4
E1	0.2748	0.1028	-0.1329
E2		0.1894	0.2778
E3			-0.1613

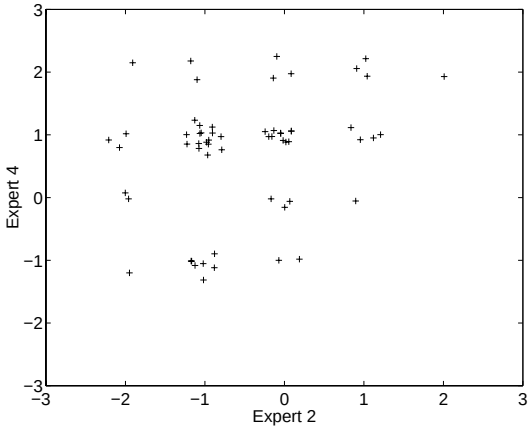

Figure 3: Expert 2 versus Expert 4. [2]

Related Work

Insightfulness is an extremely difficult notion to capture, and this work has only begun to investigate this concept. Blake and Pazzani (2000) have taken an orthogonal approach to understanding when a rule is insightful. They examined how background knowledge encoded in an electronic knowledge base could be used to remove uninteresting rules from a set. In contrast this work examines how the discovery strategy (discriminative or characteristic) affects the insight of rules found and how insightfulness correlates with objective measures of rule quality.

Clearly, before a rule can be insightful to a person, it must be understood and considered valid. In the past, researchers have considered understandable as synonymous with "short" and thus designed mining algorithms with a strong bias towards rules that are short and accurate under this assumption (Karalic, 1996; Craven, 1996). While this makes intuitive sense, there have been no studies which quantitatively confirm this. There have been two studies which indicate that perceived validity of the mined rules affects the credibility and willingness to use mined results: Pazzani, Mani, and Shankle (1997) examined the effect of monotonicity relationships on rule acceptance in diagnosing potential Alzheimer patients. They found that regardless of rule accuracy neurologists were unwilling to use rules which violated the intent of the diagnostic test. Pazzani and Bay (1999) looked at the effect of incorrect signs on the credibility of regression equations and likewise found that equations where the sign of a variable differed from subjects expectations were rated poorly. An interesting result of their study was that longer regression equations were more credible than shorter equations.

Silberschatz and Tuzhilin (1996) suggested that interestingness is a subjective quality that depends on the individual. However, they did not test this theory quantitatively with human subjects. Our results with inter-expert agreement support their theory.

[2] A small amount of random jitter has been added to the points.

Conclusions and Future Work

We asked the following two questions in our paper: "Is a discriminative or characteristic approach more useful for describing group differences?" and "How do subjective and objective measures of rule interest relate to each other?"

We answered these questions by conducting a study of admissions officers and their responses to the outputs of the data mining algorithms C5 and STUCCO. Our main findings are that (1) characteristic differences are more useful to domain experts than purely discriminative differences. (2) Many objective measures of rule quality correlate poorly with expert opinions on what is insightful, but there is some evidence that effect size is important. (3) Rule insightfulness is highly subjective as even experts examining the rules for the same task do not correlate well with each other.

This paper presented an initial study of how experts with a particular task in mind evaluated data mining discoveries. It is important to use experts because they represent the intended users of the mining programs and will have a distinct purpose in mind when evaluating results. However, the limitation is that experts are inherently rare and thus we were only able to obtain the responses from four people. We plan on conducting a larger study on a less specialized domain so that we can involve more subjects.

Acknowledgments

This research was funded in part by the National Science Foundation grant IRI-9713990.

Appendix A

We show here the 15 positive yield rules with the largest gain found by STUCCO.

1. 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 455 students.
2. Students who have a Selection Index Number between 6000 and 6500 are more likely to enroll with a 15% higher yield than average. This represents a gain of 443 students.
3. Students who have a GPA between 2.75 and 3.5 are more likely to enroll with a 16% higher yield than average. This represents a gain of 415 students.
4. 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 361 students.
5. Students who are from Orange County are more likely to enroll with a 13% higher yield than average. This represents a gain of 323 students.
6. Students who are from Orange County and live within 30 miles of UCI are more likely to enroll with a 13% higher yield than average. This represents a gain of 319 students.
7. Students who scored less than 500 on their SAT Verbal are more likely to enroll with a 17% higher yield than average. This represents a gain of 310 students.
8. Students who scored between 500 and 600 on their SAT Math are more likely to enroll with a 8% higher yield than average. This represents a gain of 291 students.
9. Students who have a Selection Index Number between 6000 and 6500 and scored between 500 and 600 on their SAT Verbal are more likely to enroll with a 14% higher yield than average. This represents a gain of 235 students.
10. Students who have a Selection Index Number between 6000 and 6500 and scored between 500 and 600 on their SAT Math are more likely to enroll with a 14% higher yield than average. This represents a gain of 216 students.
11. Students who scored between 500 and 600 on their SAT Verbal are more likely to enroll with a 4% higher yield than average. This represents a gain of 200 students.
12. Students who have a GPA between 2.75 and 3.5 and have a Selection Index Number between 6000 and 6500 are more likely to enroll with a 16% higher yield than average. This represents a gain of 186 students.
13. Students who have a Selection Index Number between 5000 and 6000 are more likely to enroll with a 19% higher yield than average. This represents a gain of 173 students.
14. Students who live within 30 miles of another UC school are more likely to enroll with a 2% higher yield than average. This represents a gain of 166 students.
15. Students who have a GPA between 2.75 and 3.5 and scored between 500 and 600 on their SAT Math are more likely to enroll with a 18% higher yield than average. This represents a gain of 165 students.

References

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

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

Bay, S. D., & Pazzani, M. J. (1999). Detecting Group Differences: Mining contrast sets. *under review*.

Blake, C., & Pazzani, M. J. (2000). Identifying Insightful Association Rules. *under review*.

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

Craven, M. W. (1996). *Extracting Comprehensible Models from Trained Neural Networks*. Ph.D. thesis, University of Wisconsin-Madison.

Everitt, B. S. (1992). *The Analysis of Contingency Tables* (second edition). Chapman and Hall.

Karalic, A. (1996). Producing more comprehensible models while retaining their performance. In *Proceedings of Information, Statistics and Induction in Science*, pp. 54–65.

Major, J. A., & Mangano, J. J. (1995). Selecting among rules induced from a hurricane database. *Journal of Intelligent Information Systems*, *4*, 39–52.

Pazzani, M. J., & Bay, S. D. (1999). The independent sign bias: Gaining insight from multiple linear regression. In *Proceedings of the Twenty First Annual Conference of the Cognitive Science Society*, pp. 525–530.

Pazzani, M. J., Mani, S., & Shankle, W. R. (1997). Comprehensible knowledge-discovery in databases. In *Program of the Nineteenth Annual Conference of the Cognitive Science Society*.

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

Rubinstein, Y. D., & Hastie, T. (1997). Discriminative vs informative learning. In *Proceedings Third International Conference on Knowledge Discovery and Data Mining*, pp. 49–53.

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

Silberschatz, A., & Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, *8*(6).