There is no quality in this world that is not what it is merely by contrast. Nothing exists in itself.

–Herman Melville

University of Alberta

Data Mining Using Contrast-sets: A Comparative Study

by

Amit Satsangi

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

Spring 2011
Edmonton, Alberta

To Buddha: May All Beings Be Happy and Peaceful…

Abstract

Comparative analysis is an essential part of understanding how and why things work the way they do. How have the rich fared in comparison to the poor in the last decade? Why do we find more men in Science and Engineering as compared to women? Do postgraduate degree holders really earn more money than those with an undergraduate degree? What do some customers prefer to buy online vs. those that do not? What factors contribute to pre-term births? Why are some students more successful than others? All the above questions require comparison between various classes. Contrast-set mining was first proposed as a way to identify those attributes that significantly differentiate between various groups (or classes) for the case of discrete data. Contrast-set mining has now been applied in every conceivable field to find contrast-sets (conjunction of attribute-value pairs) that aid in differentiating between different groups; however no clear picture seems to have emerged regarding how to extract the contrast-sets that discriminate most between the classes. Various interestingness-measures and usefulness-measures have been proposed in the form of different contrast-set mining techniques claiming to find more meaningful contrast-sets than those found by the previous technique. It has been proven in literature that contrast-set mining **is a special case of** rule discovery task; in this thesis we try to address the problem of finding meaningful contrast sets by applying a methodology that is based on the foundations of contrast-set mining – Association Rule Mining. Amongst the many surprising results that we obtain, we also report a family of contrast-sets that were previously not known in the literature. We also show as to why we should expect contrast-sets of only a certain kind for any data. Finally, we present and compare the results of our experiments with the well known algorithm for contrast-set mining – STUCCO.

Acknowledgements

There are numerous people to whom I owe gratitude for having helped me in reaching this milestone.

To start with, my family: Mom and Dad, thanks for encouraging me to work hard, to know, to explore, and for inculcating in me an unending desire for knowledge. As they say: “Man is wiser than yesterday”, and I would be dead if I did not learn something new every single day. This research would not have been possible without the support from my immediate family: Little Jeea playing her part by breaking my all-important laptop, chewing up stuff, and tearing off some of my important papers that she could lay her hand on. Noopur, I cannot express my gratitude for all those tantrums – I guess one day I will become as patient as Buddha. Esha, thanks for all the support.

I have learnt so much from my supervisor Dr. Osmar Zaïane, and not all of it is about academics. There are several things that I admire about Osmar, and a couple of them really stand out: firstly the amount of work that he puts in every single day – to be able to provide guidance and support to each one of his ten students (or so) along with teaching and administrative duties is admirable. Secondly, Osmar’s patience never seems to run out (testified by the fact that he could put up with me).

.

I would like to thank the members if my defense committee, Dr. Nilanjan Ray and Dr. James Miller for reviewing my thesis, and offering valuable suggestions.

To all my friend and colleagues in the department – a big thank you. Pirooz, it was so kind of you to help me compile STUCCO; Abhishek, thank you for cheering me up whenever the chips were down. Vahid and Mojdeh I remember all those interesting discussions with you guys. Thank you everyone in the Database group for your support.

I could not have juggled between work, family and research without going insane – Vipassana meditation changed my life – thanks to the greatest man (in my view) who

ever set foot on planet Earth – Gautama the Buddha. The world would be such a wonderful place if each one of us gives his teachings a place in their life.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The problem of identification of significant differences between contrasting groups or classes has been well studied, and has been the focus of many statisticians and data miners. A commonly asked question for data analysis in any discipline is: “How can several contrasting groups be compared against each other?” Depending on the context this leads to specific questions like-which categories of students are more likely to accept an admission offer from a University? What are the specific characteristics that best differentiate between patients with a specific disease and normal patients? What distinguishes between the customers that buy more than some value and those that buy less than another threshold? What is the difference between male and female managers, all other things being equal? Do postgraduate degree holders fare better in their career than those who hold only an undergraduate degree?

Dong and Li proposed Emerging Patterns as a mechanism to identify differences between contrasting classes [1]. In their seminal paper Bay and Pazzani [2, 3] proposed contrast-sets to bring forth a contrast between different classes that led to the creation of a new sub-field in data mining — Contrast-set mining. Webb *et. al.* used techniques that had previously been applied in rule-discovery to successfully perform contrast-set mining tasks thus proving that contrast-set mining is a special case of the rule-discovery task [4].

Contrast-set mining is being applied in many diverse fields to identify attributes that provide greatest contrast between various classes. It has been successfully applied to predict patients with brain stroke from those with other severe neurological disorders with both the groups showing somewhat similar symptoms [5, 6, 7]. Contrast-set mining has been used to study the reasons behind the

disparity between successful students from those that are less successful using the data collected from a web based educational system [8, 9]. Contrast-sets have been applied to study time-series and multimedia data [10], and have also been used to attach a label to clusters obtained after the clustering process [11]. Contrast-sets have also been used to identify patterns of factors that result in aircraft accidents [12].

A number of variations and improvements have been proposed. An *et. al.* have applied a variation of contrast-sets for mining data in large databases [13]. Loekito and Bailey explore the second order differentiation – "contrast of contrasts" [14]. In a separate study Loekito and Bailey consider the issue of contrasting in case of dynamically changing data [15]. Simeon and Hilderman consider the issue of discretizing quantitative attributes for contrast-sets [16, 17], and introduce the concept of "jumping" contrast-sets. Wong and Tseng explore the problems associated with mining negative contrast-sets [18].

1.1 Background, Challenges and Approach

Contrast-set mining is valid only for categorical data, and is based on search-techniques of those attribute-value pairs that provide the best contrast/discrimination between various classes.

Attribute Name	Attribute Values
Income	low, high, medium
Job Profile	customer service rep, front-desk employee, manager
Sex	male, female, other
race	caucasian, south asian, oriental, african
age	<30 yr., 30-50 yr., >50 yr.

Table 1: A hypothetical Census Dataset

As an example consider a small hypothetical census dataset (Table 1) where one might be interested in finding out any long-term differences between the salaries of subjects with an undergraduate degree *vs.* those with a post-graduate degree.
In table 1 we list the (categorical) attributes in the first column, and the categorical values that they take in the second column. It is clear from the above description that we have two classes[1]: Undergraduate and Postgraduate that represent the undergraduate degree holders and the postgraduate degree holders, respectively. Let us assume that there is a **substantial** difference between the conditional probabilities for the two classes involving the conjunction of **same** attribute-value pairs for as follows:

P (Degree = Undergraduate | **Income = high ^ Job Profile = manager**) = 0.23,
and
P (Degree = Postgraduate | **Income = high ^ Job Profile = manager**) = 0.73

In the above case the difference between the probabilities is 0.50; let us assume that the minimum required probability difference between the two classes for the conjunct attributes to form a contrast-set is 0.3 (a user defined value). Then **Income = high ^ Job Profile = manager** is called a contrast-set; it is a conjunction of the two attributes: Income and Job Profile, and satisfies the probability difference condition. Note that for a contrast-set the values that the attributes can take (such as **"high**" and "**manager**") should be the same amongst the two classes. We discuss this example in more detail in the next section.

The fundamental question of what constitutes an **interesting** and **useful** contrast-set has received fair bit of attention in the field of contrast-set mining. Various interestingness-measures and usefulness-measures have been proposed in the form of alternate contrast-set mining techniques [4, 19, 20] that claim to find more meaningful contrast-sets than those obtained by STUCCO – the algorithm proposed in the seminal paper by Bay and Pazzani [3]; however no clear picture seems to have emerged.

[1] In the literaure the terms groups and classes are used interchangably

All the previous attempts made to find interesting and meaningful contrast-sets have adopted an approach based on applying statistical tests for the independence of variables – the only difference in these attempts lies in the fact that some focus on controlling Type I error while others focus on controlling type II error[2]. It has been shown in [4] that **contrast-set mining is a special case of rule discovery task**, and yet none of the contrast-set mining techniques that have been proposed so far (including the one proposed by Bay and Pazzani) employ rule-based analysis.

Hypothesis 1: Contrast-set mining based on Association Rule analysis can provide superior results as compared to the statistical techniques employed in the previous literature.

In this thesis we try to address the problem of finding meaningful contrast-sets by applying a methodology that is based on the foundations of contrast-set mining – Association Rule Mining.

We propose the following high-level algorithm in order to verify the hypothesis:

1. Divide the whole dataset into as many files as there are classes such that the data regarding each class should be in a separate file.
2. Apply association-rule mining for each class separately.
3. Compare the mined association-rules to find contrast-sets that satisfy the probability-difference condition.

In order to validate our hypothesis we run our code (that is based on the above algorithm) on various datasets, and then run STUCCO on the same datasets. We provide a ranking mechanism for the contrast-sets obtained using our algorithm,

[2] In common parlance Type I error is also called 'False Positives' while Type II error is called 'False Negatives'.

and we then compare our contrast-sets with those that were obtained using STUCCO. The findings of our experiments are as follows:

1. There exists a new family of contrast-sets that was hitherto undiscovered;
2. There seem to be two kinds of association-rules mentioned in the literature for contrast-set mining, however we prove that only one kind of association-rules can lead to contrast-sets;
3. All the interesting contrast-sets discovered by STUCCO are also found by the association-rule based technique, however there are several interesting contrast-sets that STUCCO seems to have missed.

1.2 Dissertation Organization

The rest of the thesis is organized as follows. The terminology and the definition of the problem are introduced in chapter 2. A review of previous literature is carried out in chapter 3. In chapter 4 we discuss the drawbacks with previous approaches to contrast-set mining, and present a new mechanism based on association-rule mining; we also introduce a new family of contrast-sets. A discussion of the datasets, experiments carried out on the datasets, and an analysis of the results of the experiments is presented in chapter 5. We present a comprehensive summary and analysis of our research in chapter 6.

The research work here was presented in the form of a paper [21] at IDEAS-2007 Conference.

Chapter 2

Background, Problem Definition and Terminology

First we provide a background of the association-rule mining problem. Then we discuss the terminology related to contrast-sets. Finally we provide a formal definition of the problem.

2.1 The Association Rule Mining Problem

We consider the hypothetical census dataset (table 1) once again:

Attribute Name	Attribute Values
Income	low, high, medium
Job Profile	customer service rep, front-desk employee, manager
Sex	male, female, other
Race	aucasian, south asian, oriental, african
Age	<30 yr., 30-50 yr., >50 yr.

Table 1: A hypothetical Census Dataset

For the attributes and attribute values considered above we have ten "transactions" in Table 2 (below). The data in each individual row represents a different subject/person, and each row is called a transaction. This is a terminology that has been borrowed from the context of market-basket data analysis.

Subject Number	Degree		Income	Job Profile	Sex	Race	Age
S1	Postgraduate		high	manager	male	caucasian	<30 yr.
S2	Undergraduate		high	csr	female	oriental	30-50 yr.
S3	Postgraduate		high	manager	male	oriental	<30 yr.
S4	Postgraduate		high	manager	male	african	<30 yr.
S5	Undergraduate		high	manager	female	oriental	30-50 yr.
S6	Postgraduate		high	manager	female	caucasian	30-50 yr.
S7	Undergraduate		low	front-desk	male	caucasian	<30 yr.
S8	Postgraduate		high	manager	male	oriental	<30 yr.
S9	Postgraduate		low	csr	female	caucasian	>50 yr.
S10	Undergraduate		low	csr	female	south asian	>50 yr.

Table 2: Transactions (hypothetical) in the Census Dataset

Association rules are relations between variables of the form A → B where A can either be a single attribute-value pair, or it can be a conjunction of attribute-value pairs. There are two terms that are used frequently in the context of association-rules:

- **Support:** Support of an association-rule (A → B) is the percentage of transactions that contain both the itemsets A and B. **The support of a rule is a measure of how often does that rule occur in the dataset.**

  $$\text{Support } (A \rightarrow B) = \text{Probability } (A \cup B) = \frac{\#(A \cup B)}{n}$$

  where n is the total number of transactions in the dataset.
- **Confidence**: The confidence of an association-rule (A → B) is the ratio of number of transactions that contain both itemsets A and B to the number of itemsets that contain itemset A. **The confidence of a rule is a measure of the strength of the rule.**

$$\text{Confidence } (A \rightarrow B) = \text{Probablility } (B/A) = \frac{\#(A \cup B)}{\#A}$$

Here we provide an example for calculating the Support and Confidence values for the rule: **Income = high ^ Job profile = manager** → Degree = Postgraduate

The support-value for the above association rule = 5/10 = 0.5, or 50%
The confidence-value for the above association-rule = 5/6 = 0.833, or 83.3%

2.2 Terminology

We define contrast-sets by borrowing the terminology introduced by Bay and Pazzani [3]. The dataset *D* is considered to be a set of transactions (rows) with each transaction having m attributes (columns). One of the attributes is used to divide the dataset *D* into *n* mutually exclusive groups G_1, G_2 **…**G_n such that $G_i \cap G_j = \phi$ where n is the number of groups for the class attribute in *D*.

Definition 1. Let A_1, A_2 …A_m be a set of m variables that form the attributes for the dataset *D*. Each attribute A_i can take values from the set $\{A_{i1}, A_{i2} \ldots A_{ik}\}$. Then a contrast-set is a conjunction of attribute value pairs defined on groups G_1, G_2 …G_n such that no A_i occurs more than once.

Example 1. A **potential** contrast set: **Income = high ^ Job Profile = manager** is defined on the two groups: Degree = Undergraduate and Degree = Postgraduate

It is important to emphasize that the above potential contrast-set is equivalent to the association-rules (for the two groups):

Income = high ^ Job Profile = manager → Degree = Undergraduate **(0.1, 0.17) … (1)**

and

Income = high ^ Job Profile = manager → Degree = Postgraduate **(0.5, 0.83)**

… (2)

We have added the values for support and confidence in equations 1 and 2 from the calculations of the previous section. Equations 1 and 2 can also be mathematically expressed in terms of conditional probabilities:

P (Degree = Undergraduate | **Income = high ^ Job Profile = manager**) = 0.1, and
P (Degree = Postgraduate | **Income = high ^ Job Profile = manager**) = 0.5

It is clear from the following definition that the value of support for a **contrast-set** *w.r.t.* a particular group is equivalent to the support for the **corresponding association-rule** w.r.t. the same group.

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.

In light of the above definition the question of whether the **potential** contrast-set in example 1 actually forms a contrast-set depends on the satisfiability of the following conditions:

$$\exists ij P(cset = True \mid G_i) \neq P(cset = True \mid G_j) \quad \text{…} (3)$$

$$\max_{ij} |\text{support } (cset, G_i) - \text{support } (cset, G_j)| \geq \delta \quad \text{…}(4)$$

where δ is a user defined threshold called minimum support difference.

Condition 3 is called the Significance condition while condition 4 is called the Largeness condition. If both the conditions are satisfied then it is called a Deviation. As mentioned earlier the value of support used in the Largeness condition for the contrast-sets would be the same as the value of support for their corresponding association-rules in the same group.

2.3 Problem Definition

Given a dataset *D* with *n* groups we generate association-rules for each of the groups separately, and then compare the association-rules to find potential contrast-sets that satisfy the largeness condition. The foundations of contrast-set mining lie in association-rule analysis; we believe that building up contrast-sets from association-rules will not only be able to replicate the results from statistical techniques that have previously been used in the literature, but can also find many other interesting contrast-sets. Many questions have been raised in the literature around what constitutes interesting and useful contrast-sets – we believe that our method will provide better insight into this issue.

Chapter 3

Related Works

3.1 Statistical Techniques

The contrast-set mining problem was proposed by Bay and Pazzani [2, 3] wherein they apply statistical techniques to validate their hypothesis that the support for every contrast-set is independent of group membership (or alternately that the support-value is equal across all groups). As mentioned earlier the terminology discussed in section 2.2 was borrowed from [2, 3]. In the discussion below we will be using the same terminology.

Bay and Pazzani look for those contrast-sets (cset) that satisfy the two conditions (equations 3 and 4 that were discussed in the previous chapter):

$$\exists ij P(cset = True \mid G_i) \neq P(cset = True \mid G_j) \quad \ldots (3)$$

$$\max_{ij} |\text{support}\ (cset, G_i) - \text{support}\ (cset, G_j)| \geq \delta \quad \ldots (4)$$

δ is the *minimum support difference*, which is a user-defined threshold. Contrast-sets for which equation (3) is **statistically** valid is called *significant*, and those for that satisfy equation (4) are called *large*. If both the conditions are met then it is called a deviation.

We present a scenario based on [20] to show that it is possible that a contrast set may satisfy the largeness condition, but may not satisfy the significance condition.

	Degree = undergraduate	**Degree = postgraduate**	Σ Row
Income = high	194	355	549
¬(Income = high)	360	511	871
Σ Column	554	866	1420

Table 3: Contingency table for χ^2-testing (based on [20])

Largeness Condition

Using the values for the support from Table 3 we calculate the support-difference for the contrast-set: **Income = high.** It is not easy to visualize one-dimensional contrast-sets, hence one may look at them in terms of their corresponding association-rules:

Income = high → Degree = undergraduate, and
Income = high → Degree = postgraduate

Support-difference = | Support (**Income = high | undergraduate**) – Support (**Income = high | postgraduate**) | = | 194/554 – 355/866| = 0.06. If the value of δ is 0.05 then the support-difference is greater than the threshold, and the contrast-set is considered to be sufficiently "**large**".

Significance Condition

The significance condition for a contrast-set tests, **statistically**, whether the support for that contrast-set is "significantly different" across the different groups. In other words one needs to determine whether the contrast-set support is independent of group membership. Bay and Pazzani start with the null hypothesis that contrast-set support is equal across all groups, and form a 2 X *G* contingency table (with *G* being the number of groups), and apply the standard test of the independence of variables – chi-square test. The χ^2 statistic tests the null

hypothesis that the row and column totals are not related (*i.e.* are independent), and is given by:

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

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 assuming that the rows and the column variables are independent calculated as: $E_{ij} = \sum_j O_{ij} \sum_i O_{ij} / N$, where N is the total number of observations.

$X^2 = 5.08$ using the data in the contingency table (Table 3). If we evaluate results at 5% significance ($\alpha = 0.05$) then $\chi^2 = 3.84$. As $5.08 > 3.84$ the null hypothesis is rejected meaning that Degree and Income are related.

Thus we see that even when the largeness condition is satisfied it is possible that the significance condition may not be satisfied. It is mainly because of the role of α which controls maximum probability of falsely rejecting the null hypothesis in a single χ2 test (also known as Type I error, or false positives)

As discussed below Bay and Pazzani convert the contrast-set search problem into a tree-search problem with more generalized contrast-sets (single attribute-value pairs) forming the top-nodes, and the more specialized contrast-sets (conjunctions of attribute-value pairs) forming the nodes below. As is clear the size of the search-space will depends not only on the total number of attributes, but also on the number of unique values that each attribute can take. Thus multiple hypotheses are required to test for the deviations for all the attributes both at the same level and also at the different levels in the tree. For $\alpha = 0.05$ one can expect that the null hypothesis will be wrongly rejected on account of pure chance once in 20 tests – type I error, of false positive. In order to control Type I errors Bay and Pazzani make use of the Bonferroni inequality whereby they use a different value of α at different levels in the search space. The modified value of α is given by:

$\alpha_i = \min((\alpha/2^i)/|C_i|, \alpha_{i-1})$, where $|C_i|$ is the number of candidates at level i.

Bonferroni inequality ensures that α becomes more restrictive as one descends down the tree decreasing by a factor of half as one descends down the tree.

As mentioned earlier Bay and Pazzani convert the contrast-set mining problem into a tree-search problem. They use a canonical ordering of nodes in the search space by employing set-enumeration trees thereby ensuring that each node is visited only once, or not at all if it can be pruned. Here is an example of a search tree borrowed from [2]

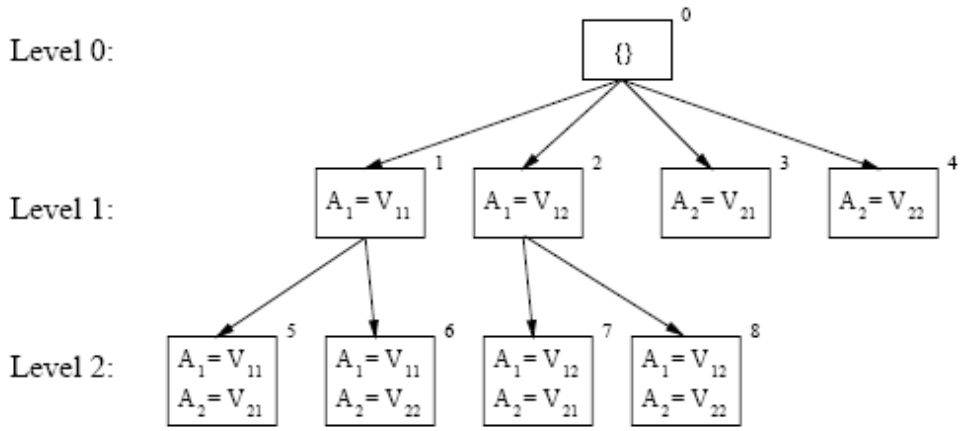

Figure 1: Set-enumeration Trees (from [2])

Pruning Strategies

Pruning of nodes is done when no specializations of that node can result in a deviation. Nodes are pruned based on the following strategies:

Minimum Deviation Size Pruning: If for a node the maximum of the support-difference between any two groups is less than the minimum threshold then that node can be pruned; this is because it is guaranteed that if minimum deviation size condition is not met for the parent node then it will not be met for any of the child

nodes. This is similar to the subset-infrequency pruning that is employed in case of Association-rule mining.

Expected Cell Frequencies Pruning: If the expected cell frequencies in the top row of the contingency table are too small (ranging from 1 to 5) then the node can be pruned because no valid inferences can be made in such a case, and also one needs to take into account the fact that the expected cell frequencies as we go down the tree.

X^2 — value pruning: It is obvious that as one descends down the set-enumeration tree the number of attribute-value pairs in a contrast-set increases leading to a bound on the support for the contrast-set on all descendents. As an example consider the association-rule/contrast-set at node ***j***: **Degree = undergraduate => Income = high** (25%). The value in the bracket is the support for the association-rule/contrast-set. The node associated with this contrast-set may have a child node ***j+1*** such as: **Degree = undergraduate => Income = high ^ Job Profile = manager.** We know for a fact that the support for this association-rule/contrast-set cannot be greater than 25%. Thus the support for the parent rule becomes an upper bound on the support for all descendents in the search space. Inversely, the support for the child rule becomes a lower bound for all ancestors for all ancestors in the search space. These bounds on the support values lead to bounds on the observed frequencies in the contingency table leading further to an upper bound on the value of χ^2, which can then be used for pruning.

Interest-based Pruning: If the addition of an attribute value-pair to an existing contrast-set (*i.e.* specialization of the contrast-set from the parent) does not lead to a change in the support value then STUCCO considers it to be non-interesting, and prunes the specialized rule from the search space. The logic behind this pruning is that in the absence of any change in the support-value the specialized rule does not provide any more information than what the parent does.

Statistical Surprise Pruning: A contrast-set is considered statistically surprising when the observed frequencies are different from the expected frequencies. We use an example from [20] to illustrate this point. The expected frequency is the product of the observed frequencies for the case where two variables are involved (for complex cases iterative proportional fitting is advised [22]). If P (Income = high | Degree = undergraduate) = 40% and P (Income = high | Job profile = manager) = 65%, then the expected frequency is given by: P(Income = high ^ Job profile = manager | Degree = undergraduate) = 26%. Bay and Pazzani prune all those contrast-sets for which the expected frequencies lie within a certain threshold of the observed frequencies because they consider such contrast-sets to be providing no new information – not surprising, and hence not useful.

Bay and Pazzani proposed an algorithm called STUCCO (**S**earch and **T**esting for **U**nderstandable **C**onsistent **Co**ntrasts) for mining contrast-sets. Here we reproduce their algorithm (borrowed from [2]).

Algorithm STUCCO
Input: data $\mathcal{D}$
Output: $D_{surprising}$
Begin
Set of Candidates $C \leftarrow \{\}$
Set of Deviations $D \leftarrow \{\}$
Set of Pruned Candidates $P \leftarrow \{\}$
Let prune(c) return true if c should be pruned
1. **while** C is not empty
2. scan data and count support $\forall c \in C$
3. **for each** $c \in C$
4. **if** significant(c) $\wedge$ large(c) **then** $D \leftarrow D \cup c$
5. **if** prune(c) is true **then** $P \leftarrow P \cup c$
6. **else** $C_{new} \leftarrow C_{new} \cup$ GenChildren(c, P)
7. $C \leftarrow C_{new}$
8. $D_{surprising} \leftarrow$ FindSurprising(D)

Figure 2: Algorithm for STUCCO (from [2])

Hilderman and Peckham [19] provide an alternate approach for mining contrast-sets. Their algorithm CIGAR (**C**ontrast**I**ng **G**rouped **A**ssociation **R**ules) provides a variation on the popular algorithm STUCCO. Hilderman and Peckham provide a variation to the *2xn* contingency table approach of STUCCO by breaking it down into a series of *2x2* contingency tables. They employ three additional constraints to the STUCCO framework:

$$\text{support}(X, G_i) \geq \beta \quad \ldots (5)$$

$$\text{correlation}(X, G_{i,} G_{j)} \geq \lambda \quad \ldots (6)$$

$$|\text{correlation}(X, G_{i,} G_{j)} - \text{correlation}(\text{child}(X, G_{i,} G_j))| \geq \gamma \quad \ldots (7)$$

where: X is a contrast-set, G_k is a group, β is a user-defined minimum threshold, λ is the user-defined minimum correlation threshold, and γ is the user-defined minimum correlation difference.

Contrast-sets satisfying condition I are called *frequent*. Contrast-sets satisfying condition (D) are called *strong*. Contrast-sets that satisfy the two constraints set by STUCCO and constraints C and D that are imposed by CIGAR are called *deviations*, however those contrast-sets that do not satisfy condition E are called spurious, and pruned from the search space. The fact that CIGAR employs additional constraints implies that in addition to the pruning strategies of STUCCO there are more pruning strategies that are based on these new constraints.

Also, the phi correlation coefficient is employed to measure the degree of association between the variables in the *2x2* contingency tables. CIGAR seeks to control Type II error (or false negatives) in contrast to STUCCO that focuses on controlling Type I error.

Hilderman and Peckham use three datasets: Mushroom, GSS Social and Adult Census, and run both STUCCO and CIGAR on these datasets. They then compare the results to find that in most of the cases the contrast-sets match for the two algorithms. For the cases where the contrast-sets do not match for the two

algorithms the authors provide reasons as to why either those contrast-sets were wrongly pruned by STUCCO, or why they were (rightly) pruned by CIGAR (and thereby failed to be pruned by STUCCO). Hilderman and Peckham conclude that while STUCCO and CIGAR are based on different statistical philosophies and assumptions both the approaches can be used to generate interesting contrast-sets.

3.2 Non-Statistical Techniques

In order to capture emerging trends in time stamped databases Dong and Li propose a new kind of pattern called Emerging Patterns (EP's). EP's can also be used to provide contrast between different classes in a dataset [1]. The authors provide EP-mining algorithms that manipulate only the "borders" of datasets – as discussed in the description later. Given the fact that Apriori property does not hold in case of EP-mining, and the fact that there may be too many candidates for the case of large datasets the border manipulation technique of Dong and Li provides an efficient pruning mechanism in the search space. EP's are defined as item-sets whose support increases significantly from one dataset D_1 to another D_2. A more precise definition of EP's is that these are item-sets whose growth rate – the ratio of their support values in D_2 over D_1 (or vice versa) – is larger than a user supplied threshold value ρ.

For data-sets D_1 and D_2 with X as an item-set with $supp_i$ (X) denoting the support of item-set X in D_i (the definition for support remains the same as was introduced earlier in case of Association-rule mining) the growth rate of an item-set X, denoted as GrowthRate(X) is defined as:

$$\begin{cases} 0, \ldots if \ \sup p_1(X) = 0 \wedge \sup p_2(X) = 0 \\ \infty, \ldots if \ \sup p_1(X) = 0 \wedge \sup p_2(X) \neq 0 \\ \dfrac{\sup port_2(X)}{\sup port_1(X)}, otherwise \end{cases}$$

Also, an item-set X is said to be an emerging pattern if for a given growth rate threshold $\rho > 1$ if GrowthRate (X) $\geq \rho$. For a given growth rate threshold ρ the EP-mining problem is to find all ρ-EP's.

Dong and Li then introduce the concept of borders – they define a border as an ordered pair $< L, R >$ such that (a) (b) of definition 3.2 where L and R are called the left hand bound of the border and the right hand bound of the border respectively. Essentially a border is a collection of large item-sets that satisfies the previous conditions. They also show that if certain conditions are satisfied then every collection S of sets has a unique border $< L, R >$ where L is a collection of minimal sets in S, and R is a collection of maximal sets in S. They then provide three algorithms, each one an improvement over the previous one, to illustrate the use of borders in pruning the search for Emerging Patterns – their algorithms take borders (collection of large item-sets) as input, manipulate the borders, and then again produce borders as output – representing EP's. More details can be found in [1].

Webb *et. al.* [4] use a commercial rule discovery system called Magnum Opus, and also use C4.5rules, which is a classification-rule discovery system; they apply these rule discovery systems on two-day sales data of a departmental store. Their objective is to contrast the retail activity on two different days in order to identify the effect of specific market promotions. Finally, they run STUCCO on the same data, and carry out a comparison of the contrast-sets produced by the trio.

While Magnum Opus is a commercial rule discovery system that looks for rules of the form *antecedent* → *consequent*, it is important to note that it differs from association-rule like systems in that it does not employ frequent-itemset strategy. Unlike other association-rule discovery systems Magnum Opus does not require minimum support value for extracting the rules – instead it requires parameters called strength (same as confidence for association-rules), maximum number of rules to be returned, lift, coverage and leverage. Coverage is merely the support of the antecedent and hence if the consequent is a group then the antecedent might be the contrast-set, and coverage the support value for the contrast-set. Leverage measures the degree to which the "observed joint frequency of the antecedent and consequent differ from the joint frequency that would be expected if the

antecedent and the consequent were independent of each other" [4]. While STUCCO uses the χ^2 test of significance Magnum Opus uses a binomial sign test, and unlike STUCCO Magnum Opus does not make corrections for multiple comparisons – according to the authors this is to avoid the risk of increasing type II errors.

For the departmental store data (discussed earlier) Magnum Opus produced 83 rules, STUCCO discovered 19 contrast-sets, and C4.5 produced 24 rules. It was found that the rules produced by classification-rules discovery – C4.5 system had two major problems: Firstly, they were missing some key contrasts, and secondly there were many negative rules that are very hard to interpret. Hence the rules obtained from C4.5 were dropped from further analysis. The authors report that Magnum Opus had produced rules corresponding to all the contrast-sets found by STUCCO. The authors then go on to prove that the constraints used by STUCCO and Magnum Opus are equivalent (a point that is disputed in [23], and we discuss that further), and they hold the view that the main respect in which the two systems differ is the application of the filters. The authors are of the view that Magnum Opus' filter is much more lenient than that of STUCCO thus leading Magnum Opus to find many more rules as compared to STUCCO. It was also found that many more rules found by Magnum Opus were considered surprising by the experts as compared to the contrast-sets discovered by STUCCO. The authors conclude that some of the rules that were not found by STUCCO, but were discovered by Magnum Opus (and were considered potentially useful) support the lenient filters applied by Magnum Opus. On the other hand there were some rules discovered by Magnum Opus that were spurious because of the leniency of the filter, and such rules were pruned by STUCCO leading the authors to conclude that neither STUCCO nor Magnum Opus is applying a perfect filter. They advocate further studies to find a better filter that finds a middle ground between the two systems.

Chapter 4

An Alternate Approach

4.1 Drawbacks of Previous Approaches

In their concluding remarks in [4] the authors mention that "neither STUCCO nor Magnum Opus is applying a perfect filter", and that while STUCCO seemed to discard some contrasts of potential value, Magnum Opus appears to include contrast-sets that were probably spurious, thus highlighting the inadequacy of the two approaches. In [23], the authors discuss how Magnum Opus actually does a within-groups comparison rather than a between-groups comparison and thus generates only a subset of the contrast-sets generated by STUCCO. Given that Magnum Opus does only a within-groups comparison the claim of the authors in [4] seems to be surprising that Magnum Opus could produce **all** the contrast-sets (an exercise that requires a between-groups comparison) that were generated by STUCCO and a few more interesting ones that STUCCO failed to produce. Hilderman and Peckham apply a completely different filter (that minimizes type II errors) than STUCCO, and claim that their method is able to find all the contrast-sets discovered by STUCCO. Thus there seems to be no consensus on the kind of filter to be used to prune the search space. All of these divergent views and results seem to be adding more confusion to the field. In light of the above issues it appears that all the approaches so far have been unable to tackle the root of the problem, and this issue is far from closed.

In [2, 3, 19, 20] the contrast-sets are reported as belonging to the association-rules such as Group → Contrast-set (for brevity we will, hence forth, refer to these kinds of contrast-sets as the "first kind"). The authors do not consider other kind of contrast-sets (henceforth referred to as the "second kind") that come from the

rules of the type Contrast set → Group. In [4], the authors consider only the second kind of contrast-sets. Later, we show that only the second kinds of contrast-sets can exist – our experimental results show that where the first kind of contrast-sets do exist, they cannot have more than one attribute.

4.2 Association-rule based approach

As quoted in [4] Bay and Pazzani oppose the association-rules based approach because

> [association rule discovery will not] return group differences, and the results will be difficult to interpret. [with reference to an example]. First, there are too many rules to compare. Second, the results are difficult to interpret because the rule learner does not enforce consistent contrasts (i.e. using the same attributes to separate groups) … Finally, even with matched rules, we still need a proper statistical comparison to see if differences in support and confidence are significant.

Even though Webb *et. al.* use Magnum Opus – a rule-based discovery system (although not an association-rule based approach) for contrast-set mining they tend to be critical of using association-rule mining for contrast-set discovery:

> To assess the importance of [the] difference [between using statistical-filters and association-rule mining] we applied Christian Borgelt's implementation of Apriori to the data … There was no obvious way, however, to configure the system to filter the most useful rules for this application.

From the above we observe that the biggest criticism against using association-rule mining for generating contrast-sets is the lack of knowledge in how it could be possible. Bay and Pazzani point have a number of objections:

1. If association-rule mining does not return group-differences how will one generate contrast-sets
2. There are too many rules to compare
3. The rules are difficult to interpret because we need to isolate rules of the kind group → precedent, which is not an easy task.
4. In some cases, there are rules in one group that have no match amongst any of the other groups. How does one interpret those?
5. Even for the case of matched rules one needs to do a statistical comparison to check if the differences in the support and confidence are significant;

We concede that the first and the fifth assertions are true – there indeed could be too many rules, however, that only affects the total run-time and not the accuracy. Regarding the last issue we argue that one does not need to do statistical comparisons if one is not creating a tree. The significance condition can be applied mathematically (in terms of probabilities) – the advantage with statistical comparisons comes in case of a tree-setup where one can do a lot of statistical pruning.

The argument laid down by Bay and Pazzani regarding a particular association-rule existing for only one group, and missing in the rest of the groups has implications on the accuracy of the results, however, having found a way to overcome this issue we decided to use association-rules to investigate the problem of finding contrast-sets because of the inadequacies of the earlier mentioned techniques, and their conflicting conclusions. Association-rules form the backbone of all the previously mentioned approaches, and hence the accuracy of the results obtained by this approach cannot be questioned, even if this approach might be slower. Our hypothesis was that association-rules, being the foundation of this problem, will generate all the "interesting" and "useful" contrast-sets that were generated by STUCCO and potentially many more. While our approach still aims at identifying the contrast-sets that satisfy the deviation conditions of STUCCO (i.e. to find the significant and large contrast-sets), it does so using

association-rules, and in the process does not suffer from the shortcomings that other techniques do.

4.3 Finding Deviations

First we will provide a brief introduction as to how we extract the association-rules – this is important in the understanding of the modifications that we make to the largeness condition. Given a dataset *D* with the class (group) being an attribute we break it up into *n* datasets $D_1 \ldots D_n$ where *n* is the number of distinct groups in *D* such that a dataset D_i corresponds to the group G_i. Then we run Christian Borgelt's association-rule mining code [24] on each of the D_i's separately; this requires a minimum support value and a minimum confidence value to be provided as an input parameter to the code. For the example below let us assume that the (min-support, min-confidence) values are: (10%, 7%). All association-rules for which the support and confidence values exceed the respective user-input-minimum-values are extracted, and the rest are pruned[3]. This implies that it is possible that an association- rule may satisfy the minimum support and confidence thresholds for one group such as: Degree = Postgraduate → **Income = high ^ Job Profile = manager** (16%, 19%) with the values in the brackets standing for (min-support, min-confidence) respectively. It is possible that the corresponding association-rule may be absent in **all** other groups because the minimum-support and minimum-confidence conditions may not be satisfied. As an example for the group of undergraduate degree holders we may have:

Degree = Undergraduate → **Income = high ^ Job Profile = manager** (9%, 3%)

As 9<10, and 7<3 the above rule will not be present in the group Degree = Undergraduate. If there were more groups assume that the contrast-set is absent from all of those as well.

[3] Note that we choose very low values of minimum support and minimum confidence so that the set of **pruned** association rules is small; this also ensures that the association rules that are discarded are of low "significance".

Now that we know that it is possible that a contrast-set may be present in only of the groups, and may be absent in the rest of the groups we are in a position to talk about the modification that we had to make in the largeness condition for the case of single contrast-sets.

Consider the two association-rules assuming that we have the two groups: Degree = Undergraduate and Degree = Postgraduate

Degree = Undergraduate → **Income = high ^ Job Profile = manager**
and
Degree = Postgraduate → **Income = high ^ Job Profile = manager**

If the support values for the above association-rules satisfy the two deviation conditions then the association-rules are equivalent to the contrast-set: **Income = high ^ Job Profile = manager.** We coined a term for such contrast-sets – **β-contrast-sets** – those contrast-sets for which the association-rules exists for **at least** two groups. These are the normal contrast-sets that satisfy the deviation conditions given by Bay and Pazzani.

Consider a situation where we have association-rule corresponding to only one group[4]:

Degree = Postgraduate → **Income = high ^ Job Profile = manager**

If the corresponding rule does not exist for the other group: Degree = Undergraduate then while such a situation does not violate any conditions for it to be a deviation, however it cannot be handled by the deviation conditions given by Bay and Pazzani. **We call such contrast-sets α-contrast-sets**. In the case of α-contrast-sets the largeness condition cannot be applied owing to the absence of a support-value for all the other groups. The normal largeness condition is:

[4] As discussed earlier

$$Max_{ij} \mid support\ (cset, Gi\ ) - support(cset, G\ j\ )\mid \geq min\_dev$$

With the support-value available for only one group the normal deviation condition will not work:

$$Max_{ij} \mid support\ (cset, G_i) - support\ (absent\_cset, G_j) \mid \geq min\_dev$$

Notice that we are looking for the maximum value on the left-hand-side. As the contrast-set does not exist in other groups we have no idea for its support-value (or equivalently the support-value for the corresponding association-rule). If the minimum-support-value at which the association-rules were extracted is 10% it is possible that the support value for the absent contrast-set may be 9.9%, because of which it was pruned, or it may be close to 0% – we have no idea as to what is the case. However an upper-bound for this value will certainly be 10% (the minimum-support-value at which the association-rules were extracted – call it sup-min-G_j_ub) because if the value was 10% (or higher) it would form a contrast-set.

We propose that the upper bound sup-min-G_j_ub should instead be used in the largeness condition in place of support (absent_cset, G_j). The justification for using the upper bound is that it will minimize the quantity: |support (cset, G_i) – support (absent_cset, G_j)|, and thus will make the condition more stringent. If we find that even if the minimum value of |support (cset, G_i) – support (absent_cset, G_j)| is greater than min_dev then it implies that the support for the contrast-set in the lone group is quite large, and in our opinion it makes a strong case for being considered a contrast-set even when the contrast-set is absent in the other groups. Thus the modified largeness condition is given by:

$$Max_{ij} \mid support\ (cset, Gi\ ) - \text{sup-min-}G_j\text{\_ub} \mid \geq min\_dev$$

The above represents the worst case analysis in the largeness condition, and if the potential α-contrast set (cset) satisfies this condition then it should be considered to satisfy the largeness condition. By employing this condition we were able to keep a significant number of contrast-sets that would have been wrongly pruned. Later on we discuss about the following ratio for the datasets that we used for our experiments:

$$\eta = \frac{\text{Number of potential } \alpha \text{ contrast - sets}}{\text{Number of potential } \beta \text{ contrast - sets}}$$

For the datasets considered we found the value of η to be close to 100, meaning that for every single potential β-contrast-set in the data there are 100 α-contrast-sets. We emphasize that prior to our publication, α-contrast-sets had never been considered in the literature – on the contrary Bay and Pazzani had hinted on the problems of comparison of association rules where a rule exists only in a single group.

Note that the assumption that the support for a contrast set is zero because it does not appear in the set of association-rules (Ã) would be wrong; consider the case that the actual support in the dataset for a particular association-rule was 1.9% (for e.g.) while the min_support used in the Apriori code (association-rule generator) was 2.0%, and hence that association-rule did not appear in Ã. This does not imply a support of 0% for that association-rule – had we used 1.9% as the value for min_support we would have found that particular association-rule in Ã.

4.4 Contrast-sets – First and Second Kind

As mentioned earlier we coined a term for the contrast-set that comes from association-rules of the type: Group → Contrast-set: Contrast-set of the "first kind", while the contrast-set that comes from association-rules of the type Contrast set → Group is called a contrast-set of the "second-kind". In the literature we found both types of contrast-sets being used (sometimes

interchangeably). We ran an association-rule program[5] on our data-sets and discovered that the number of association-rules generated for the first kind of contrast-sets was far too less (always less than 1%) than the number of association-rules corresponding to the second kind of contrast-sets. Also, the first kinds of contrast-sets were **always** composed of only single attributes. Initially we found these results quite puzzling, however, on a deeper examination we found that this result made perfect sense. Consider attributes A, B and C and assume that we have two groups in the data set. Also assume that they occur approximately equally in the data-thus 50% of the records belong to group 1 while the other 50% belong to group 2. The support value for the "first kind" of association-rule: Group1 → A, B, C is:

$$P (A \cap B \cap C \cap Group1)/P(Group1)$$

Given that P(Group1) is very high (~0.5) the support for it will be very low. In order to extract association-rules, and the corresponding contrast-sets, we need to run Apriori with a very low value for minimum support. We used a value of 1% for the minimum support and found that only single item-sets on the right hand side such as A, or B, or C are able to meet these conditions. On the other hand for the rules of the second kind: A, B, C → Group, the support-value is:

$$P (A \cap B \cap C \cap Group1)/ P(A \cap B \cap C)$$

The quantity in the denominator: $P(A \cap B \cap C)$ is small and hence the support-value for the contrast-set is higher than that of the "first kind" of contrast-set.This implies that the minimum support that goes into Apriori code can be relatively high. We believe that the above analysis will provide clarity around the issue of different kinds of contrast-sets being reported in the literature – for our analysis we decided to consider only contrast-sets (and hence association-rules) of the "second kind". Having laid the theoretical groundwork we discuss our experimental results in the next section.

[5] Christian Borgelt's implementation of Apriori version 4.28 [24]

Chapter 5

Contrast-set Mining Experiments

5.1 Background Information

In this chapter we present our algorithm, and the results of our experiments. For the first set of experiments we compare the contrast-sets obtained from STUCCO and those obtained from our approach based on association-rules. STUCCO is implemented in C++ and was compiled using g++ (version 3.4.4) run on Linux (2.6.9-42.0.3Elsmp). We implemented our code in Java 1.4.1 and ran it on a Linux Kernel (version 2.6.9- 42.0.3Elsmp) PC with a 2.4 GHz. AMD 64 bit Processor (4000+) and 2 GB of memory. The Apriori code [24] is based on the Apriori Algorithm [25], and is written in C. Our Java code encapsulates Apriori, and compares the association-rules extracted by Apriori to find those contrast-sets that satisfy the deviation conditions.

In the next set of experiments we look for ways in which an association-rule based contrast-set mining approach can help in increasing the accuracy of different classification techniques. We use a dataset of historic maternal and newborn records to predict pre-term births. First, we apply various classification techniques such as Naïve Bayes, Decision trees, SVM, logistic regression, associative classifier *etc.* on our dataset to find the performance measures such as Precision, Recall, F-measure, AUC *etc.* We then apply the contrast-set mining technique to select features from the dataset that can provide maximum contrast between the two classes – normal birth, and pre-term birth. The first set experiments are repeated again with only the selected features from the previous step being used this time. This work was carried out in association with Yavar Naddaf and Mojdeh Jalali Heravi.

5.2 Properties of Datasets

We ran STUCCO and our code on several datasets, and we present results for three datasets here: Mushroom, Breast Cancer and Adult Census. The Mushroom dataset describes characteristics of gilled mushrooms; it is available from the UCI Machine Learning Repository (www.ics.uci.edu/~mlearn/MLRepository.html). The Adult Census dataset is a small subset of the Census Income (1994/1995) dataset – a survey dataset from the U.S. Census Bureau. The Breast Cancer dataset, again obtained from the UCI Machine Learning Repository, was collected by physicians and the data contains two groups: recurring and non-recurring. The characteristics of the datasets are shown in Table 4 where the first column describes the number of tuples in the dataset, the second column describes the number of attributes, the third column describes the number of unique values contained in the attributes, and the last column describes the number of distinct groups – as defined by the number of unique values in the grouping attribute.

Dataset	Tuples	Attributes	Values	Groups
Mushroom	8142	23	130	2
Adult Census	826	13	129	2
Breast Cancer	286	9	53	2

Table 4: Properties of the datasets

The dataset related to maternal and foetal data was collected by Northern and Central Alberta Perinatal Outreach Program between 1992 and 2003. It contains maternal and newborn data for 243948 cases, including 21193 preterm cases. There are 244 attributes, containing "maternal demographic information, medial history such as pre-existing chronic illness, lifestyle information such as smoking and alcohol use, past reproductive history including previous [preterm] or [small for gestational age] delivery, and history with the current pregnancy such as

presence of hypertension or toxaemia. As quite a few were collected during or after delivery, and thus cannot be used for predicting preterm birth. We list the 46 attributes that were collected before delivery in Appendix A. The attributes "Group B Step", "Maternal Hepatitis B", and "Steroid During Pregnancy" have a high ratio of missing values, and were therefore discarded. Also, there are 2107 records with missing class labels, which are also omitted

5.3 Algorithm ARCS

Herein we present a high level view of our algorithm – ARCS (**A**ssociation **R**ule-based **C**ontrast **S**ets). We assume that the dataset *D* is divided into n datasets D_1 … D_n where n is the number of groups in dataset *D.* Also, as mentioned earlier, we only test for contrast-sets of the second kind *i.e.* Group → Contrast-set

Algorithm ARCS (min_support, min_confidence)

Begin

/* n is the total number of groups (classes) in the dataset */

for i=1, n
  Run Apriori (min_support, min_confidence) on each D_i separately;
  Store association-rules (contrast-sets) in file A_i;
end;

Dictionary-sort the contrast-sets in each (file) A_i separately;
Read first line (contrast-set) from each (file) A_i;

While (not-end-of-all-files)

{

  If (contrast-sets do not match) /* A contrast-sets exists in only one group */

    Test for the Deviation conditions for α-contrast-sets;
    Read a new line from the corresponding file;
  Else /* A contrast-sets exists in at least two groups */

    Test for the Deviation conditions for β-contrast-sets;
    Read a new line from the corresponding files**;**

}

End ARCS

5.4 Effect of Maximal Number of Item-sets

The experiments mentioned in this section clearly show the significant differences that exist between the α-contrast-sets and the β-contrast-sets. In order to carry out the experiments mentioned in section 5.5 we needed to use Christian Borgelt's code for generating association-rules using Apriori analysis; the code requires an input value called maximal number of items per set (m), which corresponds to the maximum number of attribute-value pairs in a discovered contrast set. As we did not have a good approximation for an optimal value of m we decided to vary this number, and we tracked the number of association rules generated by Apriori and also the number of α and β-contrast-sets generated.

The results of our experiments related to maximal number of item-sets are presented in Figures 3 and 4 for two data sets. For the sake of efficiency the plot for the Mushroom dataset was clipped to 7 maximal items per rule otherwise with the huge amount of association rules generated the code would have taken a long time to process those. However, we were still able to capture the trend. There are several inferences that can be drawn from Figures 3 and 4:

- As the maximal number of items per rule increase, the number of contrast-sets (both α-contrast-sets and the β-contrast-sets) increase, and eventually the curve becomes flat. A similar behaviour is observed for the curves for association-rules for the two the groups.
- In Figure 3, the maximal number of items per rule at which the curve for the number of association-rules become almost flat is approximately 11. While the curve for the number of β-contrast-sets becomes flat at 8 maximal items per association-rule, the corresponding number for the α-

contrast-sets is 10, showing that the behaviour of the α-contrast-sets is different from that of β- contrast-sets.

- The number of α-contrast-sets is higher than the number of β-contrast-sets by a factor of two orders of magnitude (approximately).

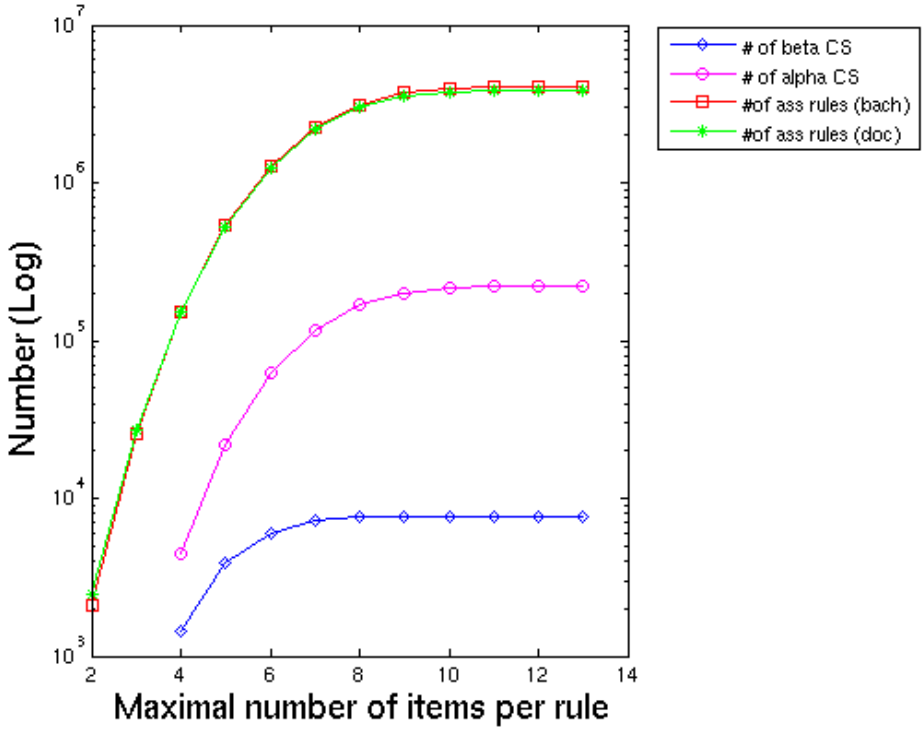

Figure 3: Behaviour of association-rules with maximal number of items for the Adult dataset ("bach" stands for bachelor; "doc" stands for doctorate – the two groups)

For the case of Mushroom data set it is clear that the curve for the number of α-contrast-sets seems to follow the curve for the number of association-rules for both the groups, and also the fact that the curve for β-contrast-sets seems to be close to flattening out while the other curves still have a rising trend. Figures 3 and 4 **clearly show that there is a marked distinction between the α-contrast-sets and the β-contrast-sets, and ignoring α-contrast-sets amounts to throwing away useful information.** We repeated the above mentioned experiments for several datasets; however in all cases the trend remained the same.

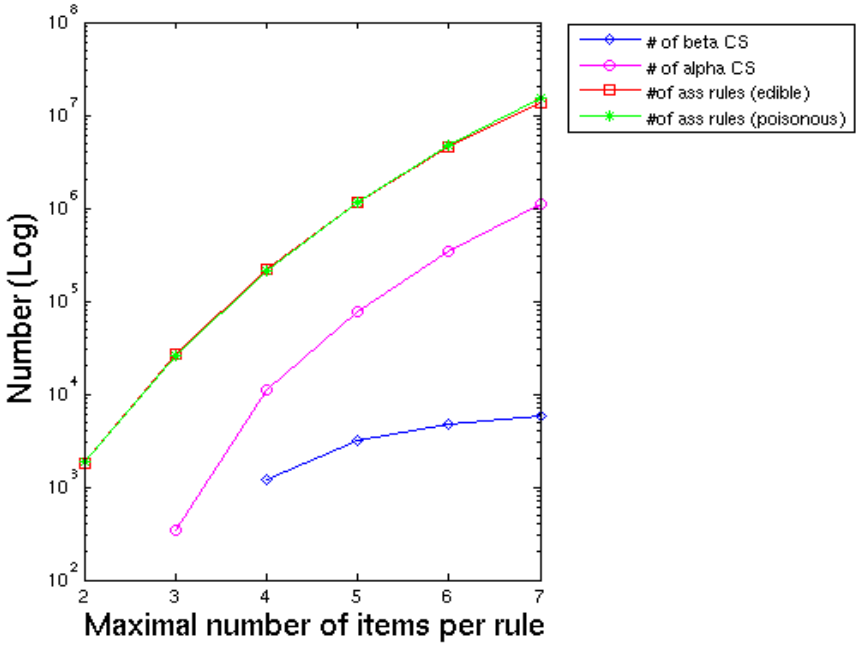

Figure 4: Behaviour of association-rules with contrast-sets for the Mushroom dataset

5.5 Comparison of results with STUCCO

A comparison of the results based on our approach, and those based on STUCCO is summarized in Table 5

Dataset	No. of STUCCO contrast-sets	No. of contrast-set from our approach	All CS of STUCCO found?
Mushroom	**—**	**Top 50**	**Not Applicable**
Adult Census	**24**	**Top 50**	**Yes**
Breast Cancer	**5**	**Top 50**	**Yes**

Table 5: Comparison of results with STUCCO

For the Adult data set STUCCO found 24 interesting contrast-sets out of a total number of 919 identified deviations. All the contrast-sets found by STUCCO

were also present in the list of Top 50 contrast-sets that were generated using our approach. We ranked the contrast-sets from our code in terms of their interestingness (*i.e.* support differential). There seemed to be many interesting contrast-sets in our list that STUCCO missed – STUCCO's result did not include any the α-contrast-sets, and as can be seen from Table 6, α-contrast-sets can constitute (approximately) anywhere from 10% to 20% of all contrast-sets.

Dataset	Top 10	Top 20	Top 50
Adult Census	10%	10%	18%
Breast Cancer	0	15%	14%

Table 6: Percentage of α-contrast-sets found in our Top N contrast-sets

Table 7 shows the highest ranked STUCCO contrast-set in our list of Top 50, and also the lowest ranked contrast-set in the same list – for the adult census dataset the highest ranked STUCCO contrast-set in our Top 50 list lies at number 3, while the lowest ranked contrast-set lies at number 43. It is clear that STUCCO missed a number of contrast-set with a high value of support differential (the ones that provide the best contrast).

Dataset	No. of STUCCO contrast-sets	Highest Ranked STUCCO contrast-set in our Top 50 list	Lowest Ranked STUCCO contrast-set in our Top 50 list
Adult Census	24	3	43
Breast Cancer	5	5	37

Table 7: Spread of STUCCO's contrast-sets in our Top 50 list

For the Breast Cancer data set STUCCO found only 18 deviations and 5 interesting contrast-sets, again all of these belonged to our list of discovered interesting contrast-sets. STUCCO did not produce any α-contrast-sets while our code was able to extract them. The highest ranked STUCCO-contrast-set was at

number 5 in our list of Top 50 while the lowest ranked contrast-set was at number 37; clearly this is a fairly wide spread, and STUCCO seems to have missed all our top four contrast-sets besides missing several other top contrast-sets. A comparative analysis for the Mushroom data set could not be performed because STUCCO did not output any results in a legible format for that dataset while our approach pinpointed many relevant contrast-sets.

5.6 Experiments for Preterm Birth Prediction

Pre-term birth is one that takes place after at least 20 completed weeks of gestation, but is less than 37 completed weeks of gestation. Pre-term births can cause moderate to severe disability, and extreme complications in infancy and childhood. With more than two-thirds of all prenatal deaths caused due to pre-term births decision support tools are needed for doctors to be able to predict pre-term births early on.

The objective of this experiment was to determine whether contrast-set can help in increasing the accuracy of a classifier. The experiments were carried out in two phases, such that in the first phase various classification methods were applied to the dataset, and measures such as precision, recall, AUC *etc.* were calculated for the results. In the second phase contrast-set mining was carried out on the dataset to carry out feature selection – determine a few features (attributes) that discriminate most between the two classes – natural birth and pre-term birth. The classification experiments were re-run with only a reduced set of attributes in the hopes of achieving better accuracy.

As mentioned earlier this work was carried out in association with Yavar Naddaf and Mojdeh Jalali Heravi – for the sake of completeness and understanding we will describe the whole experimental process, and will mention the part that was carried out specifically by us. In the first step the dataset was converted into a transactional database with each transaction being a record of a unique patient. Eclat – a depth first search algorithm developed by Zaki *et. al.* [26], and

implemented by Christian Borgelt [27] was used because the transaction volume is extremely large (241841 transactions with and average of 41 items per transaction).

Measure → / Algorithm ↓	**True +ve Rate**	**False -ve Rate**	**Preci —sion**	**Recall**	**F-Measure**	**AUC Weka**	**AUC Haneley And McNeill**
Naïve Bayes	**0.281**	0.036	0.564	**0.281**	**0.375**	0.716	**0.622**
Logistic Regression	0.207	0.014	0.713	0.207	0.321	**0.724**	0.597
SVM Linear Kernel	0.155	**0.008**	**0.757**	0.155	0.257	0.573	0.573
ZC 4.5 Decision Tree	0.197	0.013	0.708	0.197	0.308	0.666	0.592
Neural Networks	0.228	0.020	0.657	0.228	0.338	0.711	0.604
Associative Classifier	0.218	0.029	0.419	0.218	0.286	N/A	0.594

Table 8: Performance of various classification methods

It was hoped that Eclat would be more efficient than Apriori however, it was found that for the volumes of transactions involved (241841) the algorithm takes about a month to generate all the rules. Due to time limitations we started with a very small subset of the dataset (1% of the original dataset), and then increased the size gradually. An associative classifier was used for classification of the records into the two classes: normal birth and pre-term birth. After trying a number of confidence values the minimum confidence chosen was 85%, and the minimum support was set to 1%. A number of other classifiers (mentioned in

table 5) were also used; these classifiers are available in Weka [28] – an open-source data-mining package – the default optimized values of parameters were used. Various performance measures are recorded, and the best value amongst all these measures is highlighted in bold (see Table 8). AUC is the area under the Receiver Operating Characteristic (ROC) curve; it provides an estimate of the false positive and false negative error rate. For estimating the AUC value we use a method provided by Haneley and McNeil [29], and also the method provided in the Weka package. All the classification methods perform poorly – within the classification methods tried Naïve Bayes provides the best value for the AUC (0.662).

Measure → **Algorithm ↓**	**True +ve Rate**	**False -ve Rate**	**Preci —sion**	**Recall**	**F-Measure**	**AUC Weka**	**AUC Haneley And McNeill**
Naïve Bayes	0.181	0.014	0.678	0.181	0.286	0.676	0.584
Logistic Regression	0.171	0.012	0.698	0.171	0.275	0.676	0.579
SVM Linear Kernel	0.133	**0.009**	0.711	0.133	0.224	0.562	0.562
ZC 4.5 Decision Tree	0.164	0.011	**0.712**	0.164	0.266	0.629	0.576
Neural Networks	**0.195**	0.020	0.617	**0.195**	**0.296**	**0.677**	**0.587**
Associative Classifier	0.137	0.013	0.513	0.137	0.216	N/A	0.562

Table 9: Performance of classification methods after feature selection

Next we tried contrast-set mining on the dataset for selecting those attributes that provide the best contrast between the two classes. The approach mentioned in Satsangi and Zaiane [21] was applied to the association rules generated earlier. The original program was modified to accommodate for the different format, and a different logic used for pruning the rules. This exercise was carried out by us. After feature selection all the classification experiments performed earlier were repeated, but with a reduced set of attributes (obtained from contrast-set mining) in the hopes of an improvement in the performance. In Table 9 we present the classification results obtained after feature selection based on contrast-set mining.

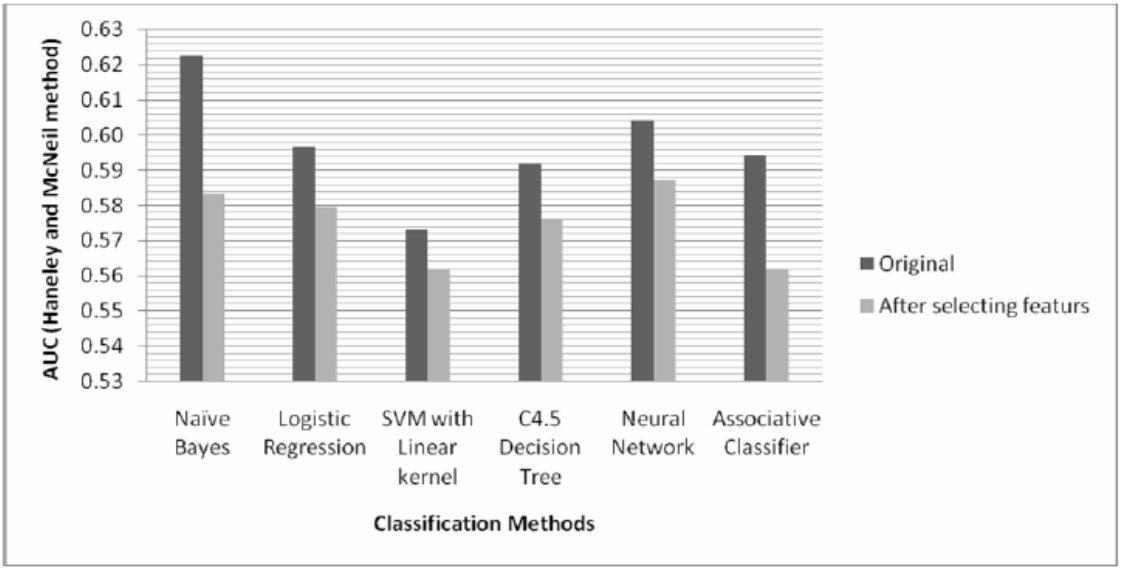

Figure 5: Comparison of classification results before and after feature selection

In figure 5 we show a comparison of the classification results before and after the feature selection was implemented. As is clear from figure 5 the prediction performance worsens after feature selection.

5.7 Analysis of Feature Selection Results

We went back to analyze the causes behind the worsened performance after feature selection. We tried changing parameters in our code as well as Weka, and

re-running the experiments to analyze the results behind the worsening of performance after feature selection, however there was no improvement in the results. We could not find any works related to **contrast-set mining** for maternal and foetal data in the literature; however we did come across research on using classifiers to predict pre-term birth; one such work was by Goodwin and Maher [30] wherein the authors use five different classification techniques: Neural Networks, Logistic Regression, CART Decision Trees, and two custom software Packages: PVRuleMiner and FactMiner. They use 32 demographice variables and 393 clinical variables – each classification technique is applied once on the demographic variables only, and once on all the variables. The performance of the various classifiers is evaluated by the area under the ROC curve. Their prediction performance is comparable to ours results – for e.g. by applying Neural Networks to the demographic variables only the AUC found was 0.64 while with the addition of 393 additional variables the value of AUC was 0.66. The best performance came from the custom software package: FactMiner with an AUC of 0.725 when applied on demographic variables only, and an AUC of 0.757 when applied on all the variables. Thus it seems that including 393 clinical variables does not make any significant improvement in the results that are obtained from the 32 demographic variables. Given the fact that the attributes on which the data was collected in our dataset is very similar to their attributes it seems that most of our clinical variables are redundant too. Also Goodwin and Maher analyze some of the leading risk assessment tools, and they report that the ability of these tools to predict pre-term birth is very poor (17% - 38%). The authors quote Dr. Creesy, who is a leading expert on pre-term risk scoring tools, as acknowledging that the pre-term risk scoring tools have not worked.

Thus it seems that most of the datasets available for pre-term births, including ours, seem to be collecting data for the parameters that may not have much significance while missing out on some other important parameters – one such possibility could be related to the genetic makeup (gene profile) and genetic

analysis of the subject; we did not come across any study that looked at the role of genes and complex genetics in this problem in a wholesome manner.

In our opinion the absence of some important parameters and very poor accuracy of the classification methods is the reason behind feature selection giving poor results in every case. Feature selection works by considering the attributes that provide the best contrast between the groups while dropping the rest of the attributes, but in case of the pre-term birth datasets available so far it seems that the attributes that provide the best contrast are missing from the datasets – one reason why the accuracy of the classification methods is so poor.

Chapter 6

Conclusion and Future Work

Our analysis on various datasets show that Association-rule based analysis is more correct; it is able to find all the contrast-sets that are found by STUCCO, and some more potentially interesting ones that STUCCO fails to discover. We have also shown that only one kind of Association-rules make sense – the second kind. We were able to discover a new family of contrast-sets – α-contrast-sets that are based on a novel approach for calculating support values. **A-contrast-sets had always been pruned in the literature due to the absence of a satisfactory method to handle them.** Our experiments clearly show that α-contrast-sets have markedly different properties and behaviour from that of β-contrast-sets. Also, the number of potential α-contrast-sets exceeds the number of potential β-contrast-sets by a factor of 100 (approximately) for most datasets. **Thus we see no justification in throwing away such useful information.**

While our work on maternal and foetal data did not produce any results that were better than the previous efforts we argue that none of the works in literature have been able to get satisfactory results implying that some essential attributes are missing, and hence contrast-set mining is not able to select features that can improve upon the accuracy of the results obtained from various classification methods.

Our research has proved beyond doubt that contrast-set mining that is based on association-rules mining can not obtain all the contrast-sets that are obtained from other methods that are based on statistical techniques, however we do concede that it is based more on a "proof-of-concept" scenario rather than worrying about issues such as efficiency *etc.* A well developed code that uses a tree structure for efficient pruning, does not generate association rules for each and every attribute,

but generates α-contrast-sets and β-contrast-sets directly for only the group (class) attribute can speed up the whole process. It will also allow contrast-set mining on very large datasets – a problem that exists today.

We believe that our work also has implications both for clustering using the contrast-sets obtained from the data, and analyzing the quality of clustering that is carried out by any of the known methods. Contrast-sets can discriminate among clusters and thus help describe and label clustering results.

Bibliography

[1] G. Dong and J. Li. “Efficient mining of emerging patterns: Discovering trends and differences”. Proc. Fifth ACM SIGKDD international conference on Knowledge discovery and data mining *(KDD ’99).* ACM, New York, NY, USA, 43-52

[2] S.D. Bay and M.J. Pazzani. “Detecting change in categorical data: Mining contrast sets”. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’99), pages 302-306, San Diego, U.S.A., August 1999

[3] S.D. Bay and M.J. Pazzani. “Detecting group differences: Mining contrast sets”. Data Mining and Knowledge Discovery, 5(3): 213-246, 2001

[4] G.I. Webb, S. Butler, and D. Newlands. “On detecting differences between groups”. In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’03), pages 256-265, Washington, D.C., U.S.A., August 2003

[5] P. Kralj, N. Lavrac, D. Gamberger, A. Krstacic. “Contrast Set Mining Through Subgroup Discovery Applied to Brain Ischaemina Data”. PAKDD 2007: 579-586

[6] P. Kralj, N. Lavrac, D. Gamberger, A. Krstacic. “Contrast Set Mining for Distinguishing Between Similar Diseases”. AIME 2007: 109-118

[7] P. Kralj, N. Lavrac, D. Gamberger, A. Krstacic. CSM-SD. “Methodology for contrast set mining through subgroup discovery”. Journal of Biomedical Informatics 42(1): 113-122 (2009)

[8] D. Perera, J. Kay, I. Koprinska, K. Yacef, O. R. Zaiane. “Clustering and Sequential Pattern Mining of Online Collaborative Learning Data”. IEEE

Transactions on Knowledge and Data Engineering, 21(6):759-772, 2009

[9] J. Kay, N. Maisonneuve, K. Yacef, O. R. Zaiane. "Mining Patterns of Events in Students' Teamwork Data". Proceedings of Educational Data Mining Workshop, held in conjunction with Intelligent Tutoring Systems (ITS), Taiwan, June 26, 2006

[10] J. Lin, E. J. Keogh. "Group SAX: Extending the Notion of Contrast Sets to Time Series and Multimedia Data". PKDD 2006: 284-296

[11] F. Alqadah, R. Bhatnagar. "Discovering Substantial Distinctions among Incremental Bi-Clusters". SDM 2009: 197-208

[12] Z. Nazeri, D. Barbara, K. De Jong, G. Donohue, and L. Sherry. "Contrast-Set Mining of Aircraft Accidents and Incidents". Lecture Notes in Computer Science, 2008, Volume 5077/2008, 313-322

[13] A. An, Q. Wan, J. Zhao, X. Huang. "Diverging patterns: discovering significant frequency change dissimilarities in large databases". CIKM 2009: 1473-1476

[14] E. Loekito, J. Bailey. "Mining influential attributes that capture class and group contrast behaviour". CIKM 2008: 971-980

[15] J. Bailey, E. Loekito. "Efficient incremental mining of contrast patterns in changing data". Inf. Process. Lett. 110(3): 88-92 (2010)

[16] M. Simeon, R. J. Hilderman. "Improving Contrast Set Mining". Proceedings of the Doctoral Consortium 2008, 7th Australasian Data Mining Conference (AusDM'08), Glenelg, Australia, November, 2008: 1-4.

[17] M. Simeon, R. J. Hilderman. "Exploratory Quantitative Contrast Set Mining: A Discretization Approach". ICTAI (2) 2007: 124-131

[18] Tzu-T. Wong, Kuo-L. Tseng. "Mining negative contrast sets from data with discrete attributes". Expert Syst. Appl. 29(2): 401-407 (2005)

[19] R.J. Hilderman and T. Peckham. “A Statistically Sound Alternative Approach to Mining Contrast Sets”. In Proceedings of the 4th Australasian Data Mining Conference (AusDM), pages 157-172, Sydney, Australia, December, 2005

[20] R. J. Hilderman and T. Packham. “Statistical Methodologies for Mining Potentially Interesting Contrast Sets”, Studies in Computational Intelligence (SCI) 43, 153-177 (2007)

[21] A Satsangi, O. R. Zaïane. “Contrasting the Contrast Sets: An Alternative Approach”. IDEAS 2007: 114-119

[22] B.S. Everitt. “The Analysis of Contingency Tables”. Chapman and Hall, 1992

[23] Terry Peckham. “Contrasting interesting grouped association-rules”. Master’s thesis, University of Regina, 2005

[24] C. Borgelt, Fast Implementation of the Apriori Algorithm available at: http://fuzzy.cs.uni-magdeburg.de/~borgelt/apriori.html

[25] R. Agrawal and R. Srikant. "Fast Algorithms for Mining Association Rules", 20th Int'l Conf. on Very Large Data Bases, Santiago, Chile, Sept. 1994

[26] M. Zaki, J. S. Parthasarathy, and W. Li. “A localized algorithm for parallel association mining”. In 9th ACM Symp. Parallel Algorithms and Architectures, 1997

[27] C. Borgelt. “Efficient implementations of apriori and éclat”. Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations

(FIMI'03), Florida, USA, 19. November 2003

[28] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten. "The WEKA Data Mining Software: An Update". SIGKDD Explorations, Volume 11, Issue 1 (2009)

[29] J. A. Hanley, B.J. McNeil. "The meaning and use of the area under a receiver operating characteristic (ROC) curve". Radiology.,143:29-36, 1982

[30] L. Goodwin, S. Maher. "Data mining for preterm birth prediction". In Proceedings of the 2000 ACM Symposium on Applied Computing - Volume 1, ACM, New York, NY, 46-51, 2000.

Appendix A

Table A.1: Attributes in the pre-term birth dataset

1	Pregnancy type. Singleton/multiple gestation	Numeric	0%
2	The woman's age in year at delivery time	Numeric	1.24%
3	Total number of pregnancies including the current pregnancy	Numeric	0.02%
4	Total number of babies born, excluding the current pregnancy	Numeric	2.22%
5	Total number of pregnancy losses	Numeric	2.22
6	Mother's weight	Numeric	2.22%
7	Mothers' height less than 152 cm	Binary	2.21%
8	Diabetes controlled by diet	Binary	2.22%
9	Diabetes documented retinopathy	Binary	2.22%
10	Insulin dependent diabetes	Binary	2.22%
11	Heart disease - asymptomatic	Binary	2.23%
12	Heart disease - symptomatic	Binary	2.23%
13	Hypertension 140/90 or greater	Binary	2.23%
14	Anti hypertensive drug used	Binary	2.22%
15	Chronic renal disease	Binary	2.23%
16	Other medical disorders	Binary	2.23%
17	Past neonatal death	Binary	2.23%
18	Past stillbirth	Binary	2.23%
19	Past abortion	Binary	2.23%
20	Past preterm	Binary	2.23%
21	Previous cesarean section	Binary	2.16%
22	Previous small for gestational age	Binary	2.23%
23	Previous large for gestational age	Binary	2.24%
24	previous RH isoimmunization - unaffected infant	Binary	2.24%
25	previous RH isoimmunization - affected infant	Binary	2.24%
26	previous major congenital anomaly Downs, Heart, CNS defect etc	Binary	2.24%
27	current large for gestational age	Binary	2.24%
28	current small for gestational age	Binary	2.24%
29	current polyhydramnios or oligohydramnios	Binary	2.24%
30	current malpresentation	Binary	2.24%
31	bleeding < 20 weeks gestation	Binary	2.23%
32	bleeding >= 20 weeks gestation	Binary	2.23%
33	pregnancy induced hypertention	Binary	2.24%
34	proteinuria >= 1+	Binary	2.24%
35	gestational diabetes	Binary	2.24%
36	blood antibodies	Binary	2.24%
37	Anemia	Binary	2.24%
38	poor weight gain	Binary	2.24%
39	smoker anytime during pregnancy	Binary	2.22%
40	major fetal anomaly	Binary	22.42%
41	acute medical disorder	Binary	22.43%
42	alcohol >= 3 drinks on any one occasion during pregnancy	Binary	22.43%
43	alcohol >= 1 drink per day throughout pregnancy	Binary	22.43%
44	group B strep	Binary	65.75%
45	maternal hepatitis B	Binary	95.99%
46	Sex	M or F	0.51%
47	steroids were given during pregnancy	Binary	63.13%