Chapter XV
Mining Conditional Contrast Patterns

Guozhu Dong
Wright State University, USA

Jinyan Li
Nanyang Technological University, Singapore

Guimei Liu
National University of Singapore, Singapore

Limsoon Wong
National University of Singapore, Singapore

ABSTRACT

This chapter considers the problem of "conditional contrast pattern mining." It is related to contrast mining, where one considers the mining of patterns/models that contrast two or more datasets, classes, conditions, time periods, and so forth. Roughly speaking, conditional contrasts capture situations where a small change in patterns is associated with a big change in the matching data of the patterns. More precisely, a conditional contrast is a triple (B, F_1, F_2) of three patterns; B is the condition/context pattern of the conditional contrast, and F_1 and F_2 are the contrasting factors of the conditional contrast. Such a conditional contrast is of interest if the difference between F_1 and F_2 as itemsets is relatively small, and the difference between the corresponding matching dataset of $B \cup F_1$ and that of $B \cup F_2$ is relatively large. It offers insights on "discriminating" patterns for a given condition B. Conditional contrast mining is related to frequent pattern mining and analysis in general, and to the mining and analysis of closed pattern and minimal generators in particular. It can also be viewed as a new direction for the analysis (and mining) of frequent patterns. After formalizing the concepts of conditional contrast, the chapter will provide some theoretical results on conditional contrast mining. These results (i) relate conditional

contrasts with closed patterns and their minimal generators, (ii) provide a concise representation for conditional contrasts, and (iii) establish a so-called dominance-beam property. An efficient algorithm will be proposed based on these results, and experiment results will be reported. Related works will also be discussed.

INTRODUCTION

This chapter formalizes the notions of conditional contrast patterns (C2Ps) and conditional contrast factors (C2Fs), and studies the associated data mining problem. These concepts are formulated in the abstract space of patterns and their matching datasets.

Roughly speaking, C2Ps are aimed at capturing situations or contexts (the conditional contrast bases or C2Bs) where small changes in patterns to the base make big differences in matching datasets. The small changes are the C2Fs and their cost is measured by the average number of items in the C2Fs. The big differences are the differences among the matching datasets of the

C2Fs; we use the average size of the differences to measure the impact (of the C2Fs). Combining cost and impact allows us to find those C2Fs which are very effective difference makers. In formula, a C2P is a pair $\langle B, \{F_1, ..., F_k\}\rangle$, where k >1, and B and F_i are itemsets; B is the C2B and the F_i's are the C2Fs.

For k=2, Figure 1 (a) shows that F_1 and F_2 are small itemset changes to B. Panel (b) shows that the matching datasets of $B \cup F_1$ and $B \cup F_2$ are significantly different from each other. The k>2 case is similar.[1]

We use the impact-to-cost ratio, defined as the impact divided by the cost, as well as other measures, to evaluate the goodness of C2Ps and C2Fs. Observe that one can also consider other factors involving class, financial benefit or utility in defining this ratio.

Example 1.1 *C2Ps can give new insights to many, especially medical/business, applications. We illustrate the concepts using a medical dataset. From a microarray gene expression dataset used in acute lymphoblastic leukemia subtype study [Yeoh* et al, *2002], we got a number of C2Ps, including the following*[2].

Figure 1. Conditional contrast patterns/factors: (a) F_1 and F_2 are small itemset changes to B, and (b) the matching dataset of $B \cup F_1$ is very different from that of $B \cup F_2$.

(a) Small Itemset Changes

(b) Big Matching Data Set Differences
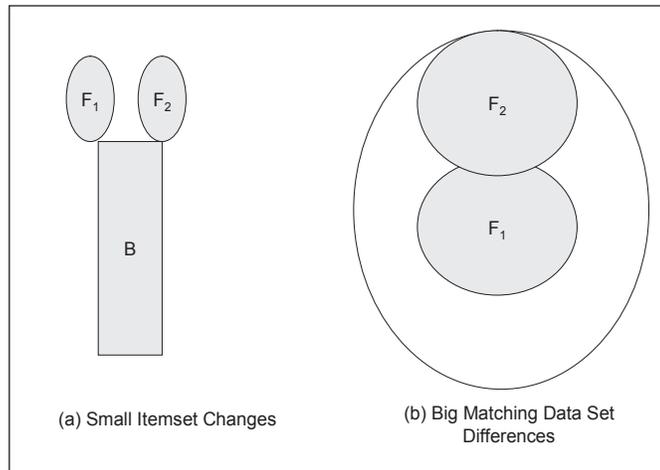
$P_L = \langle\{\text{gene-38319-at} \geq 15975.6\}, \{\{\text{gene-33355-at} < 10966\}, \{\text{gene-33355-at} \geq 10966\}\}\rangle$

Here {gene-38319-at $\geq$15975.6} *is the C2B,* {gene-33355-at < 10966} *is F_1, and* {gene-33355-at $\geq$ 10966} *is F_2. This C2P says that the samples that satisfy* gene-38319-at $\geq$ 15975.6 *(which are the samples of B-lineage type) are split into two disjoint parts: the first part are the E2A-PBX1 subtype (18 samples), and the other part are the other B-lineage subtypes (169 samples). Expressed as a rule, P_L says: Among the samples satisfying* gene-38319-at $\geq$ 15975.6*, if the expression of* gene-33355-at *is less than 10966, then the sample is E2A-PBX1; otherwise, it belongs to the other types of B-lineage.*

This C2P nicely illustrates how the regulation of gene-38319-at and gene-33355-at splits patients into different acute lymphoblastic leukemia subtypes.

Typically, an individual C2F of a C2P does not make the big differences between matching datasets; the differences are made by two or more C2Fs of the C2P. For example, in a C2P with two C2Fs F_1 and F_2, the set of items in $F_1 \cup F_2$ makes the differences.

The mining of C2Ps/C2Fs has several interesting applications. We highlight a few here. (a) The C2Fs of a C2P are difference makers, in a sense similar to issues that cause voters to swing in elections, or factors that cause customers to switch companies. It may be worthwhile to pay more attention to the most interesting C2Fs and the items in them in real world applications. (b) Given a dataset, it may be interesting to find the most significant C2Ps as important states of the dataset, and the most significant C2Fs as state transitions. (c) C2P/C2F mining can be used for unsupervised feature selection, for association mining, clustering, and other forms of knowledge pattern mining. (d) We note that one can also define some indices which can be used to identify important "distinguishing items," based on how frequently the items occur in C2Bs or C2Fs.

Besides formulating the concepts of C2Ps and C2Fs, we make the following contributions: (a) We present theoretical results on properties of C2Ps and C2Fs, concerning a so-called dominance beam property (DBP), relationship to closed and key (generator) patterns, and concise representation of C2Ps and C2Fs. The results are useful for expansion-based search for situations where anti-monotonicity does not hold. (b) We present an algorithm for mining C2Ps, called C2Pminer. It utilizes the theoretical results on the relationship of C2Ps/C2Fs with closed itemsets and keys, and the dominance beam property, for efficient mining. It produces all C2Ps under a special C2P representation. (c) We report experiment results performed on some data sets for cancer research and some datasets from the UCI repository.

Section 1.1 discusses related works. Section 2 formulates the main concepts. Section 3 presents the dominance beam property. Section 4 considers representation issues. Section 5 presents the C2Pminer algorithm. Section 6 gives an experimental evaluation. Section 7 concludes.

Related Work

This chapter is related six groups of previous studies.

1. There are several interesting differences between conditional contrast mining and association mining [Agrawal et al., 1993]. In association mining one is interested in frequent itemsets or association rules. In conditional contrast pattern (C2P) mining one is interested in interactions among groups of patterns (namely the C2B and the C2Fs). Hence the C2P pattern type is quite complementary to association mining.

2. This work leads to new insights to the field of Formal Concept Analysis (FCA) [Ganter & Wille, 1999] and to the direction of closed pattern mining [Mineau & Ganter,

2000, Pasquier et al., 1999, Zaki & Hsiao, 2002, Wang et al., 2003]. It relates C2P/C2F mining with closed pattern and key mining. Moreover, a C2P can be considered as identifying important groups of concepts, and C2Fs can be viewed as small patterns that cause big data changes to the formal concepts.

3. This work is related to the direction on mining of contrast patterns (including patterns on changes and differences) [Dong & Li, 1999, Bay & Pazzani, 2001, Ganti et al., 1999, Liu et al., 2000, Webb et al., 2003, Ji et al, 2005], in a non-temporal, non-spatial, and unsupervised setting. The mining of emerging patterns and contrast (classification) patterns have been limited to situations where a number of classes are given. In this work, there is no need to have classes; the data cohorts corresponding to the C2B-C2F combinations can be viewed as dynamically discovered "classes," and the C2Fs can be viewed as the emerging/contrast patterns between those dynamically discovered "classes."

4. The notion of conditional contrast pattern is somehow related to the (generalized) disjunction free representation [Bykowski & Rigotti, 2001] for itemsets. Our study on representation issues of conditional contrast patterns is also related to other studies on concise representation of frequent itemsets and association rules [Kryszkiewicz, 2001, Calders & Goethals, 2002]. It is also related to the rough set approach [Pawlak, 1991].

5. This work is also related to interestingness [Tan et al., 2002] of association rules, especially the so-called "neighborhood based" interestingness [Dong & Li, 1998], and related to actionable rules [Jiang et al., 2005].

6. The mining of C2Ps/C2Fs can be used for unsupervised feature selection [Liu & Motoda, 1998]. Indeed, C2Fs and the items in them can be viewed as important features, since they participate in making big differences between data cohorts. Moreover, C2P/C2F mining does not depend on the existence of classes.

BACKGROUND: CONCEPTS OF CONDITIONAL CONTRAST PATTERNS

We formulate here the concepts of conditional contrast patterns (C2Ps) and conditional contrast factors (C2Fs), together with cost, impact, and (relative) impact-to-cost ratio.

Let I be a set of *items*. An *itemset*, or a *pattern*, is a set of items. A *transaction* is a non-empty set of items, which is also associated with a unique *transaction identity* (TID). A *dataset* is a non-empty multi-set of transactions. Following a popular convention, we write an itemset such as $\{a, b, c\}$ as *abc*, and an TID set such as *{1,2,3,5}* as *1235*. A transaction T is said to *contain* a pattern X if $X \subseteq T$. The *matching dataset* of an itemset X, denoted mat(X), is the set of transactions containing X. The *support* of a pattern X in a dataset D, denoted supp(X), is the number |mat(X)|. A pattern X is *frequent* w.r.t. *ms* in a dataset D if supp(X) $\geq$ *ms*, where *ms* is a given support threshold.

We now turn to the main concepts. We first define potential conditional contrast patterns, and then add restrictions to define the true conditional contrast patterns and conditional contrast factors. Intuitively, a conditional contrast pattern consists of a base and a set of conditional contrast factors, where small changes—the conditional contrast factors—make big differences.

Definition 2.1 *A* potential conditional contrast pattern[3] *(or* PC2P *for short) is an ordered pair P = $\langle B, \{F_1, ..., F_k\}\rangle$ where:*

*   *$k>1$, an integer, is the* arity *of the PC2P,*
*   *B, an itemset, is the* conditional contrast base *(C2B) of the PC2P,*

- $F_1 ,..., F_k$, *k distinct itemsets, are the* conditional contrast factors *(C2Fs) of the PC2P, and* $\{F_1,...,F_k\}$ *is the* conditional contrast factor set *(C2FS) of the PC2P.*

Example 2.2 The pair $P = \langle abc, \{d, ef\}\rangle$ is a PC2P; *abc* is the C2B, *d* and *ef* are the C2Fs, and $\{d, ef\}$ is the C2FS consisting of two C2Fs.

We are interested in the C2Fs' ability to effectively make big differences among their matching datasets, at a low cost. A C2B is like an efficient "watershed", where the C2Fs are "small" and they separate and direct the data into different valleys. The C2Fs can be considered as "tipping patterns/factors", since these small patterns are associated with big differences. Figure 1 illustrates these points.

To capture the intuition of "small changes making big differences," we need two functions on PC2Ps: cost measures how expensive the first change is, and impact measures how significant the second (induced) change is. They measure different properties: the former is focused on syntax (items) and the latter on the behavior (matching datasets, etc).

Definition 2.3 Given a PC2P $P=\langle B, \{F_1,...,F_k\}\rangle$, we define cost($P$) to be the average number of items used in the C2Fs, and impact(P) to be the average size of the matching dataset differences among the C2Fs; in formula[4].

- $\text{cost}(P) = \dfrac{\sum_{i=1}^{k}|F_i|}{k}$

- impact(P) =

$$\dfrac{\sum_{1\le i<j\le k}|\text{mat}(B\ F_i)\Delta\text{mat}(B\ F_j)|}{k(k-1)/2}$$

Combining the two, we define a ratio to measure "per-item ability of conditional contrast factors to make dataset changes".

Definition 2.4 The impact-to-cost ratio and *relative impact-to-cost ratio* of a PC2P P are respectively defined by:

- $\text{icr}(P) = \dfrac{\text{impact}(P)}{\text{cost}(P)}$

- $\text{ricr}(P) = \dfrac{\text{icr}(P)}{\text{supp}(B)}$

The ricr can be more useful than icr, since it is relative to (the size of) mat(B).

Example 2.5 Let D be a dataset which contains 8 transactions and which satisfies the following (for brevity, the data set itself is omitted): mat(*abc*) = 12345678, mat(*abcd*) = 12345, mat(*abcef*) = 13468, and mat(*abcg*) = 12678. For $P = \langle abc, \{d, g\}\rangle$, we have:

- cost(P) = (|d| + |g|)/2 =1
- impact(P) =|mat(*abcd*)Δmat(*abcg*)|
  = 6
- $\text{ricr}(P) = \dfrac{\text{impact}(P)/\text{cost}(P)}{|\text{mat}(B)|} = \dfrac{6/1}{8} = 0.75.$

By letting $B' = B\cup\cap_{i=1}^{k}F_i$ and $F_i' = F_i - B'$, a PC2P $P = \langle B, \{F_1, ..., F_k\}\rangle$ can be simplified into a PC2P $P' = \langle B', \{F_1', ..., F_k'\}\rangle$ where the following hold: icr(P')≥icr(P), $\cap_{i=1}^{k}F_i' = \varnothing$, $B'\supseteq B$, and for each i, $B'\cap F_i' = \varnothing$, $F_i'\subseteq F_i$ and mat($B\cup F_i$) = mat($B'\cup F_i'$) (C2Fs of P' describe the same datasets as C2Fs of P). Below we assume PC2Ps satisfy these conditions unless specified otherwise.

Definition 2.6 Given a threshold $\eta > 0$, an η-conditional contrast pattern is a PC2P $P = \langle B, \{F_1, ..., F_k\}\rangle$ such that ricr(P) $\geq \eta$.

Example 2.7 Continuing with Example 2.5, we have: $P = \langle abc, \{d, g\}\rangle$ is a 0.75-C2P, but it is not a 0.8-C2P.

The conditional contrast factors (C2Fs) can be viewed as "actionable" patterns for the situation described by the C2B: By "making" certain C2Fs false or true through "item-changing actions" (such as financial incentives or medical treatments), certain objects in one C2F's match-

ing dataset may be switched into another C2F's matching dataset. As noted earlier, one can also use C2Fs to identify globally important individual items, which participate in many small C2Fs that make big differences.

We can also modify the icr/ricr definition to encourage the C2Fs to contain different classes or conditions on some (financial) utility attributes, and to consider the varying degree of difficulty in changing items over general attributes.[5]

One may use additional measures to help select interesting C2Ps, especially when there are globally rare items. Consider a dataset D containing a very rare item a. There can be many C2Ps P = <B, {a, {}}> where supp(B) and ricr(P) are high, due to the global rarity of a, not by a's interesting interaction with B; we are not interested in such C2Fs. We can use the minC2Fsupp threshold to fix the problem. We also use the minC2Bsupp threshold to help find C2Ps which cover relatively large number of tuples/transactions.

DOMINANCE BEAM RESULTS

We now consider properties of C2Ps, especially the "dominance beam" property (DBP), for efficient C2P mining. We motivate and define this property, and then establish the results concerning this property which are used in our algorithm.

The Dominance Beam Property

One might be tempted to try to adapt the frequently used anti-monotone property for efficient C2P mining. One may define a measure function f (e.g. ricr) to be *anti-monotone* over the C2Ps w.r.t. a partial order $\leq$ if $f(X) \geq f(Y)$ for all C2Ps X and Y such that $X \leq Y$. (We need to replace the usual $\subseteq$ for itemsets by $\leq$ for C2Ps, since C2Ps are no longer sets.) Unfortunately, it is not clear if there exist such partial orders $\leq$ over C2Ps for which ricr has the anti-monotone property.

Note: The f function discussed in the above paragraph has the C2Ps as its domain. Its range can be any type that has a partial order on it. For example, the range can be the real numbers when f is ricr.

Let us see what anti-monotonicity gives us. It is well known that it allows efficient search by join-based expansion, where one needs to examine a candidate Y only if every immediate subset X of Y is a valid result.

The "dominance beam property" introduced below can also be used for efficient search, and it is a dual of anti-monotonicity in some sense: we need to examine a candidate Y if at least one of its immediate predecessors X is a valid result; in other words, we only need to search along the branches of the search tree whose nodes are all valid results.

Definition 3.1 A function f has the dominance beam property (DBP) w.r.t. a partial order $\leq$ over C2Ps if, for each C2P P, there exists some C2P P' such that P' is an immediate predecessor of P under $\leq$ and $f(P') \geq f(P)$.

The DBP can be used for efficient mining when anti-monotonicity does not hold. This applies to C2P mining. Suppose f has the DBP, σ is a threshold, and we want to mine all C2Ps P such that $f(P) \geq \sigma$. The mining can proceed as follows: We start by constructing an initial set of C2Ps, to consist of all minimal C2Ps P (under $\leq$) such that $f(P) \geq \sigma$. Then we expand the search by recursively generating all immediate successors of the computed C2Ps P which satisfy $f(P) \geq \sigma$; the DBP ensures that this search will find all desired C2Ps.

One can define several natural partial orders over the C2Ps, including those based on C2B containment, C2FS containment, or C2F containment. Among them, the most useful for us is the one based on C2FS containment. Luckily, we can establish the dominance property for this partial order.

Dominance Beam w.r.t. Conditional Contrast Factor Set

We now present our most useful dominance beam results, for beams which link C2Ps that differ by exactly one C2F. Such results are especially useful because they imply that we only need to search by adding/replacing one C2F at a time when mining C2Ps.

The partial order for such dominance beams is $\leq_{C2FS}$ defined by: Given C2Ps $P1 = \langle B, FS_1 \rangle$ and $P2 = \langle B, FS_2 \rangle$, we say $P_1 \leq_{C2FS} P_2$ if $FS_1 \subseteq FS_2$. In this partial order, P_1 is an immediate predecessor of P_2, and P_2 is an immediate successor of P_1, if $FS_2 - FS_1$ contains exactly one itemset (viewed as a C2F). Moreover, only C2Ps with identical C2Bs are comparable.

Proposition 3.2 (ricr Dominance Beam w.r.t. C2FS) Let $P = \langle B, FS \rangle$ and $FS = \{F_1, .., F_k\}$ where $k > 2$. Then for each $1 < k' < k$, there is a $P' = \langle B, FS' \rangle$, where $FS' \subset FS$ and $|FS'| = k'$, such that $icr(P') \geq icr(P)$ and $ricr(P') \geq ricr(P)$. In particular, the above is true for $k' = k - 1$.

Proof: By induction, it suffices to consider $k' = k - 1$. For each $J \subset \{1, ..., k\}$ with $|J| = k - 1$, define $ssd(J) = \Sigma_{i,j \in J, i < j} |mat(B \cup F_i) \Delta mat(B \cup F_j)|$, and $sz(J) = \Sigma_{i \in J} |F_i|$.

Let us suppose, for a contradiction, that for all $J \subset \{1, ..., k\}$ where $|J| = k - 1$, we have

$$\frac{ssd(J)}{sz(J)}$$

$< icr(P)*(k - 2)/2$. So $ssd(J) < sz(J) * icr(P) * (k - 2)/2$. Summing over all possible J, we get $\Sigma_J ssd(J) < \Sigma_J sz(J)*icr(P)*(k - 2)/2$.

For each pair of distinct $i, j \in \{1, ..., k\}$, there are

$$\binom{k-2}{k-3} = k - 2$$

subsets J of $\{1,...,k\}$, where $|J| = k - 1$ and $\{i, j\} \subseteq J$. Hence $\Sigma_J ssd(J) = (k - 2) * \Sigma_{1 \leq i < j \leq k} |mat(B \cup F_i) \Delta mat(B \cup F_j)| = (k-2)*impact(P)*k*(k-1)/2$. For each $i \in \{1,...,k\}$, there are

$$\binom{k-1}{k-2} = k - 1$$

subsets J of $\{1,...,k\}$, where $|J| = k - 1$ and $i \in J$. Hence $\Sigma_J sz(J) = (k-1)*\Sigma_{1 \leq i \leq k} |F_i| = k*(k-1)*cost(P)$. Plugging these equalities into $\Sigma_J ssd(J) < \Sigma_J sz(J) *icr(P)*(k-2)/2$, we get $(k-2)*impact(P)*k*(k-1)/2 < icr(P)*cost(P)*k*(k-1)*(k-2)/2$. It follows that $icr(P) > icr(P)$, a contradiction.

So there is a $J \subset \{1,...,k\}$ where $|J| = k - 1$ and

$$\frac{ssd(J)}{sz(J)}$$

$\geq icr(P)*(k-2)/2$, proving the icr case. Since the conditional contrast bases for P and P' are identical, the statement for ricr follows.

Proposition 3.2 says that the impact-to-cost ratio can be increased by deleting C2Fs. Such increase can be achieved until the C2P is reduced to just two C2Fs. No more deletion is possible, since a C2P must have ≥ 2 C2Fs by definition.

One may wonder what happens when we further simplify a C2P with just two C2Fs. Here we note that the impact-to-cost ratio still can increase after we replace one of the C2Fs by $\{\}$.

Proposition 3.3 Let $P = \langle B, \{F_1, F_2\} \rangle$. Then there exists i such that $icr(P_i) \geq icr(P)$ and $ricr(P_i) \geq ricr(P)$, where $P_i = \langle B, \{F_i, \{\}\} \rangle$.

Proof: Suppose to the contrary that $icr(P_i) < icr(P)$ for each i. Then $|mat(B \cup F_i) \Delta mat(B)| < icr(P)*|F_i|/2$. So $|mat(B \cup F_1) \Delta mat(B \cup F_2)| < \Sigma_{i=1}^2 |mat(B \cup F_i) \Delta mat(B)| < icr(P)*(|F_1| + |F_2|)/2$. This leads to $icr(P) < icr(P)$, a contradiction. Hence $icr(P_i) \geq icr(P)$ for some i. Since the conditional contrast base has remained the same, we have also $ricr(P_i) > ricr(P)$ for some i.

REPRESENTATION ISSUES AND RELATIONSHIP WITH CLOSED PATTERNS AND MINIMAL GENERATORS

A conditional contrast pattern is intended to capture changes in the underlying dataset under some given condition. It is possible for two distinct C2Ps to correspond to exactly the same changes. We consider here the issue of representing such changes and the corresponding C2Ps of such changes. We also discuss C2P/C2F's relationship with closed patterns and minimal generators.

We first define denotational equivalence on PC2Ps, then we take icr into consideration. Let D be a fixed data set.

Definition 4.1 The denotation $[|P|]$ of a PC2P $P = \langle B, FS \rangle$ is defined to be the collection $\{\text{mat}(B \cup F) \mid F \in FS\}$ of sets. The equivalence class $[P]$ of P is defined to be the set $\{ P' \mid [| P' |] = [|P|]\}$ of PC2Ps that have the same denotations as P. Moreover, we say that $P = \langle B, FS \rangle$ is redundant if $\|[|P|]\| < |FS|$—i.e., there are distinct $F, F' \in FS$ such that $\text{mat}(B \cup F) = \text{mat}(B \cup F')$.

Before discussing Definition 4.1, we give some background definitions. The *equivalence class* $[X]$ of an itemset X is defined as the set of patterns Y that occur in exactly the same transactions as X in the given dataset D—viz., $[X] = \{Y \mid \text{mat}(X) = \text{mat}(Y)\}$. The closed patterns are defined as the most specific patterns in these equivalence classes—viz., X is a closed pattern if $X \supseteq Y$ for all $Y \in [X]$. The key patterns are defined as the most general patterns (or equivalently, minimal patterns) in these equivalence classes—viz., X is a key pattern, if there is no pattern $Y \in [X]$ such that $Y \subseteq X$ and $Y \neq X$. It is well known that (i) $[X]$ is convex ($\forall Y_1, Y_2,$ and $Z, Z \in [X]$ holds if $Y_1, Y_2 \in [X]$ and $Y_1 \subseteq Z \subseteq Y_2$); (ii) $[X]$ has exactly one closed pattern and it has one or more key patterns.

The definition of $[|P|]$ is set theoretic and ignores the icr. There can be multiple PC2Ps in an equivalence class. The definition of redundancy is aimed at avoiding uninformative C2Fs. Consider

$P_1 = \langle B, \{F_1, \{\}\}\rangle$, $P_2 = \langle B, \{F_2, \{\}\}\rangle$, and $P_{12} = \langle B, \{F_1, F_2, \{\}\}\rangle$, where $B \cup F_1$ and $B \cup F_2$ are distinct key patterns of $[B \cup F_1]$. Then P_1, P_2 and P_{12} are in the same equivalence class. The C2Fs F_1 and F_2 in P_{12} are referring to the same underlying matching dataset, which is why P_{12} is redundant. The definition ensures that every C2F in a non-redundant PC2P refers to a distinct matching dataset. We focus on non-redundant PC2Ps from now on.

Equivalence classes have some nice structural properties.

Proposition 4.2 Let $P = \langle B, \{F_1,..., F_k\}\rangle$ be a PC2P. For each i, let cF_i be the closed itemset of $[B \cup F_i]$.

1. Let $BS([P]) = \{ B' \mid$ there is a FS' where $\langle B', FS' \rangle \in [P]\}$ be the set of all C2Bs that occur in $[P]$. Then $BS([P])$ is convex, $\{\}$ is its most general (minimum) itemset, and $\hat{B} = \cap_{i=1}^{k} cF_i$ is its most specific (maximum) itemset.

2. Given itemset $B' \in BS([P])$ and $i \in [1..k]$, let $FBS([P], B', i) = \{F' \mid$ there is a FS' such that $\langle B', FS' \rangle \in [P], F' \in FS',$ and $\text{mat}(B' \cup F') = \text{mat}(B \cup F_i)\}$. That is, $FBS([P], B', i)$ is the set of all C2Fs that can substitute for F_i for a fixed C2B B'. Then $FBS([P], B', i)$ is convex and cF_i is its most specific itemset.

Proof: To prove Part (1), note that $\langle \{\}, \{cF_1,...,cF_k\}\rangle$ and $\langle \hat{B}, \{cF_1,...,cF_k\}\rangle$ are clearly in $[P]$. Thus $\{\}$ and $\hat{B}$ are in $BS[P]$. It is obvious that $\{\}$ is minimum in $BS[P]$.

To show that $\hat{B}$ is maximum in $BS([P])$, let $P' = \langle B', FS' \rangle$ be an arbitrary PC2P in $[P]$. Then $[|P'|] = \{\text{mat}(B' \cup F) \mid F \in FS'\} = [|P|] = \{\text{mat}(B \cup F_1),..., \text{mat}(B \cup F_k)\}$. Since B' occurs in all transactions in the denotations of P', we know that B' occurs in all transactions in $\text{mat}(B \cup F_i)$, for $1 \leq i \leq k$. Then $B' \subseteq cF_i$ for $1 \leq i \leq k$, since cF_i is the closed pattern of $[B \cup F_i]$. Hence $B' \subseteq \hat{B} = \cap_{i=1}^{k} cF_i$. Therefore $\hat{B}$ is the maximum among all the C2Bs of $[P]$.

To show that $BS([P])$ is convex, suppose $X \subseteq Y \subseteq Z$ and $X, Z \in BS([P])$. Then $Y \subseteq Z \subseteq \hat{B} = \cap_{i=1}^{k} cF_i$. Then $\text{mat}(Y \cup cF_i) = \text{mat}(cF_i) = \text{mat}(B \cup F_i)$ for $1 \leq i \leq k$. Thus $\langle Y, \{cF_1, ..., cF_k\}\rangle \in [P]$. So $Y \in BS([P])$, proving Part (1).

To prove Part (2), note that for any $B' \in BS([P])$, it is the case that $\langle B', \{cF_1, ..., cF_k\}\rangle \in [P]$. Thus $cF_i \in FBS([P], B', i)$, for $B' \in BS([P])$ and $1 \le i \le k$. Suppose $F' \in FBS([P], B', i)$. Then $\mathrm{mat}(B' \cup F') = \mathrm{mat}(B \cup F_i) = \mathrm{mat}(cF_i)$ by construction. Since cF_i is the closed pattern of $\mathrm{mat}(B \cup F_i)$, we have $B' \cup F' \subseteq cF_i$. Thus $F' \subseteq cF_i$. So cF_i is the most specific itemset in $FBS([P], B', i)$.

To show that $FBS([P], B', i)$ is convex, suppose $X \subseteq Y \subseteq Z$ and $X, Z \in FBS([P], B', i)$. Then $\mathrm{mat}(B' \cup X) = \mathrm{mat}(B' \cup Z) = \mathrm{mat}(B \cup F_i)$ by construction. Thus $B' \cup X \in [B \cup F_i]$ and $B' \cup Z \in [B \cup F_i]$. Then $B' \cup Y \in [B \cup F_i]$ by convexity of $[B \cup F_i]$. So $\mathrm{mat}(B' \cup Y) = \mathrm{mat}(B \cup F_i)$. It follows that $\langle B', \{cF_1, ..., cF_{i-1}, Y, cF_{i+1}, ..., cF_k\}\rangle \in [P]$. Thus $Y \in FBS([P], B', i)$. So $FBS([P], B', i)$ is convex as required.

We note that a C2B can influence the choice of patterns that can be used as C2Fs. Specifically, a C2B $B' \in BS([P])$ cannot be too specific, since it must leave some items in cFi for use in the C2Fs to uniquely identify $\mathrm{mat}(B \cup F_1)$, ..., $\mathrm{mat}(B \cup F_k)$; but it can be very general.

We now show that the equivalence classes of C2Ps can be represented by key and closed patterns.

Proposition 4.3 Let $P = \langle B, FS\rangle$ be a PC2P, where $FS = \{F_1, ..., F_k\}$. Let cF_i be the closed pattern of $[B \cup F_i]$ for $1 \le i \le k$. Let $\hat{B} = \cap_{i=1}^k cF_i$. Then:

1 $\langle \hat{B}, \{cF_i - \hat{B} | 1 \le i \le k\}\rangle \in [P]$.
2 $\langle \cap_{1 \le i \le k} kF_i, \{kF_j - \cap_{1 \le i \le k} kF_i | 1 \le j \le k\}\rangle \in [P]$, where kF_i is a key pattern of $[B \cup F_i]$ for $1 \le i \le k$.
3 $\langle \hat{B}, \{kF_i - \hat{B} | 1 \le i \le k\}\rangle \in [P]$, where kF_i is a key pattern of $[B \cup F_i]$.
4 $\langle B', FS' \rangle \in [P]$ iff (a) for each $F' \in FS'$, there is $F \in FS$ satisfying $kF \subseteq B' \cup F' \subseteq cF$; and (b) for each $F \in FS$, there is $F' \in FS'$ satisfying $kF \subseteq B' \cup F' \subseteq cF$, where kF is a key, and cF is the closed, pattern of $[B \cup F]$.

Proof: Parts (1), (2), and (3) follow from the fact that $[B \cup F_i] = [cF_i] = [kF_i]$ for $1 \le i \le k$. Part (4) follows from the fact that, for any pattern X and

Y, it is the case that $Y \in [X]$ iff $kX \subseteq Y \subseteq cX$, where kX is a key pattern of $[X]$ and cX is the closed pattern of $[X]$.

Consequently, for any PC2P $P = \langle B, \{F_1, ..., F_k\}\rangle$, the following are possible choices for a "canonical" representative of its equivalence class $[P]$ (where cF_i is the closed pattern of $[B \cup F_i]$ for $1 \le i \le k$):

• minCP is the (singleton) set of non-redundant PC2Ps in CP = $\{\langle \cap_{1 \le i \le k} cF_i, \{cF_j - \cap_{1 \le i \le k} cF_i | 1 \le j \le k\}\rangle\}$;

• minKP is the set of non-redundant PC2Ps in KP = $\{\langle \cap_{1 \le i \le k} kF_i, \{kF_j - \cap_{1 \le i \le k} cF_i | 1 \le j \le k\}\rangle | kF_i$ is a key pattern of $[B \cup F_i]\}$; and

• minCKP is the set of non-redundant PC2Ps in KCP = $\{\langle \cap_{1 \le i \le k} cF_i, \{kF_j - \cap_{1 \le i \le k} cF_i | 1 \le j \le k\}\rangle | kF_i$ is a key pattern of $[B \cup F_i]$ for $1 \le i \le k \}$.

The choice of minCP as a canonical representative of the equivalence class is nice in the sense that it is guaranteed to be a unique representative. But it has one weakness because it often has low—though not always the lowest[6]—impact-to-cost ratio in PC2Ps of its equivalence class.

On the other hand, PC2Ps in the set minKP generally have high—though not always the highest[7]—impact-to-cost ratios among the PC2Ps in [P]. Similar to minCKP, they do not guarantee a unique canonical representative. They are worse than minCKP is this aspect. All the PC2Ps in minCKP have exactly the same conditional contrast base, because of the uniqueness of closed patterns. In contrast, the PC2Ps in minKP may not have the same conditional contrast base.

As mentioned above, the PC2Ps in the set minCKP do not guarantee a unique canonical representative, even though they have the same conditional contrast base. Nevertheless, they are nice in a different way. Specifically, they have the highest impact-to-cost ratios among the non-redundant PC2Ps in [P]. For this reason, we recommend the PC2Ps in minCKP as canonical representatives, and think that the mining of conditional contrast patterns should be restricted to these canonical PC2Ps.

Proposition 4.4 (Optimality of minCKP) For each non-redundant PC2P $P'' \in [P]$, there is a $P' \in minCKP \subseteq [P]$, such that $icr(P') \geq icr(P'')$ and $ricr(P') \geq ricr(P'')$. Consequently, the PC2Ps in minCKP are among the non-redundant PC2Ps having the highest icr and ricr in $[P]$.

Proof: Let $P = \langle B, \{F_1, ..., F_k\} \rangle$. Let $P'' = \langle B'', \{F_1'', ..., F_k''\} \rangle$ be a non-redundant PC2P in $[P]$. By rearrangement if necessary, by Part (4) of Proposition 4.3, we can assume that $kF_i \subseteq B'' \cup F_i'' \subseteq cF_i$, where cF_i is the closed pattern of $[B \cup F_i]$ and kF_i is some key pattern of $[B \cup F_i]$, for $1 \leq i \leq k$.

Now, let $B' = \cap_{1 \leq i \leq k} cF_i$, $F_i' = kF_i' - B'$, and $P' = \langle B', \{F_1', ..., F_k'\} \rangle$. By Proposition 4.2, we have $B'' \subseteq B'$. For each $1 \leq i \leq k$, from $kF_i' \subseteq B'' \cup F_i''$, we get $F_i' = kF_i' - B' \subseteq (B' \cup F_i'') - B' \subseteq F_i''$. So (i) $cost(P') \leq cost(P'')$.

We know by construction that $[B' \cup F_i'] = [kF_i'] = [B'' \cup F_i'']$. So $mat(B' \cup F_i') = mat(B'' \cup F_i'')$, for $1 \leq i \leq k$. This implies (ii) $impact(P') \geq impact(P'')$. Combining (i) and (ii), we obtain (iii) $icr(P') \geq icr(P'')$.

By Proposition 4.2, we have $B'' \subseteq B'$. So $supp(B'') \geq supp(B')$. Combined with (iii), we get $ricr(P') \geq ricr(P)$.

Another nice property enjoyed by minCKP is that the unique C2B in these PC2Ps is a closed pattern. This special property is useful for the mining of conditional contrast patterns, as it allows us to anchor the mining process on closed patterns.

Proposition 4.5 Given an equivalence class $[P]$, the C2B of the PC2Ps in minCKP is unique and is a closed pattern.

In practice, one may want to be able to easily test whether an arbitrary PC2P P' is in the equivalence class of another PC2P P. Representing an equivalence class by a single canonical PC2P does not facilitate such tests. We suggest that the equivalence class $[P]$ of a (non-redundant) PC2P $P = \langle B, \{F_1, ..., F_k\} \rangle$ be represented by a set of borders $\langle B', \{\langle K_1', C_1' \rangle, ..., \langle K_k', C_k' \rangle\} \rangle$, where C_i is the closed pattern of $[B \cup F_i]$, K_i is the set of key patterns of $[B \cup F_i]$, $B' = \cap_{1 \leq i \leq k} C_i$, $C_i' =$ $C_i - B'$, and $K_i' = \{K - B' \mid K \in K_i\}$, for $1 \leq i \leq k$. This representation allows us to readily test if a PC2P is in a particular equivalence class as per Part (4) of Proposition 4.3. It also let us quickly enumerate all minCKP, which are the PC2Ps having the highest impact-to-cost ratios in $[P]$.

THE C2PMINER ALGORITHM

In this section, we present our C2PMiner algorithm for mining conditional contrast patterns. The algorithm uses four thresholds: minC2Pricr, maxC2Fcost, minC2Bsupp and minC2Fsupp. The first two thresholds ensure that only interesting conditional contrast patterns with low cost and high ricr are discovered. The last two parameters ensure that the mined conditional contrast patterns have big absolute changes, and the big changes are not caused by conditional contrast factors alone but caused by adding conditional contrast factors to conditional contrast bases. A conditional contrast pattern $P = \langle B, \{F_1, F_2, .., F_k\} \rangle$ is called desired if P satisfies $supp(B) \geq minC2Bsupp$, $ricr(P) \geq minC2Pricr$, $|F_i| \leq maxC2Fcost$ and $supp(F_i) \geq minC2Fsupp$ for all $i \in [1, k]$.

The C2PMiner algorithm first mines frequent closed itemsets and keys with respect to minC2Bsupp simultaneously using the GcGrowth algorithm [Li et al., 2005]. The frequent closed itemsets are used as candidate conditional contrast bases. We modified the GcGrowth algorithm to mine those closed itemsets X and their keys kX such that $|kX - B| \leq maxC2Fcost$, $supp(kX - B) \geq minC2Fsupp$ and $B \subset X$, where B is some frequent closed itemset. We build inverted files on closed itemsets to facilitate subsequent superset search. Next C2PMiner generates all C2Ps containing only one non-empty conditional contrast factor, and then uses these C2Ps as starting patterns to generate all the C2Ps containing more than one non-empty C2Fs based on the dominance-beam properties given in Section 3. The pseudo-code of the C2PMiner algorithm is given in Algorithm 1.

Box 1. Algorithm 1 C2PMiner

Input: A dataset D and four thresholds: minC2Bsupp, minC2Fsupp, maxC2Fcost, minC2Pricr.

Output:

All desired C2Ps satisfying the thresholds, under the minCKP representation.

Description:

1: Use a modified GcGrowth to mine a) the set of frequent closed itemsets CS_b wrt minC2Bsupp and all of their keys, and b) the set of closed itemsets CS_f and their keys such that for each closed itemset $X \in CS_f$, there exists a closed itemset $Y \in CS_b$ such that $X \supset Y$, $|kX - Y| \leq maxC2Fcost$ and $supp(kX - Y) \geq minC2Fsupp$. Build inverted files on $CS_b \cup CS_f$ to facilitate subsequent superset search.

2: For each closed itemset $B \in CS_b$ do:

2.a) Use the inverted files to find all the supersets of B. For each superset X, test and generate C2Ps of the form $P = \langle B, \{kX - B, \{\}\}\rangle$, where kX is a key of X satisfying $|kX - B| \leq maxC2Fcost$ and $supp(kX - B) \geq minC2Fsupp$. Maintain all the generated C2Fs $kX - B$ in a list $CandC2Fs$. Output P if $ricr(P) \geq minC2Pricr$.

2.b) Call DFSMineC2P(B, $CandC2Fs$, $|CandC2Fs|$) to generate all the C2Ps with B as conditional contrast base and containing more than one non-empty C2Fs. This procedure uses the dominance beam properties to search for C2Ps meeting the thresholds.

It calls a procedure DFSMineC2P(B, $CandC2Fs$, L) to generate all the C2Ps containing more than one non-empty C2Fs. This procedure uses the dominance beam properties to search for C2Ps meeting the thresholds.

Now we describe how DFSMineC2P(B, $CandC2Fs$, L) works. Besides using the dominance-beam properties, C2PMiner uses the relative impact of conditional contrast factors with respect to conditional contrast bases to prune the search space. The relative impact of a conditional contrast factor F with respect to a conditional contrast base B is defined as $rimp(F,B) = (supp(B) - supp(B \cup F))/supp(B)$.

Lemma 5.1 Let $P = \langle B, \{F_1, \cdots, F_k\}\rangle$. If $ricr(P) \geq minC2Pricr$, then we have $\sum_{i=1}^{} rimp(F_i, B) \geq k'* minC2Pricr/2$, where k' is the number of non-empty conditional contrast factors in P.

Proof:

$$minC2Pricr \leq ricr(P)$$

$$= \frac{\sum_{1 \leq i < j \leq k} supp(B \cup F_i) + supp(B \cup F_j) - 2supp(B \cup F_i \cup F_j)}{supp(B) * (k-1)/2 * \sum_{i=1}^{k}|F_i|}$$

$$\leq \frac{\sum_{1 \leq i < j \leq k} (supp(B) - supp(B \cup F_i) + supp(B) - supp(B \cup F_j))}{supp(B) * (k-1)/2 * \sum_{i=1}^{k}|F_i|}$$

$$\leq \frac{(k-1)\sum_{i=1}^{k}(supp(B) - supp(B \cup F_i))}{supp(B) * (k-1)/2 * k'}$$

$$= \frac{2\sum_{i=1}^{k} rimp(F_i, B)}{k'}$$

Hence we have $\sum_{i=1}^{k} rimp(Fi, B) \geq k'* minC2Pricr/2$.

When generating the C2Ps containing only one non-empty conditional contrast factor, C2PMiner also maintains the list of candidate conditional contrast factors for each conditional contrast base B, denoted as $CandC2Fs(B)$, and calculates the relative impact of the candidate conditional contrast factors. Here we say a conditional contrast factor F is a candidate conditional contrast factor of B if $|F| \leq maxC2Fcost$ and $supp(F) \geq minC2Fsupp$. Any combination of the candidate conditional contrast factors can form a conditional contrast factor set of B, so the search space of conditional contrast factor set wrt B is the power set of $CandC2Fs(B)$. The C2PMiner algorithm explores the search space in the depth-first order. It sorts the candidate conditional contrast factors of B in ascending order of relative impact. The candidate extensions of a conditional contrast fac-

tor include all the conditional contrast factors that are before it in the ascending order. During the mining process, C2PMiner maintains the accumulated relative impact of the C2Fs on the current path, denoted as *rimp_sum*. Let *F* be the current candidate C2F to be appended. If *rimp_sum<k'* minC2Pricr/*2 and *rimp*(*F*)<*minC2Pricr*/2, where *k'* is the number of non-empty C2Fs on the current path, then there is no need to explore the current branch further based on Lemma 5.1.

The C2PMiner algorithm uses the dominance beam property with respect to C2FS as follows. It explores the search space in ascending order of relative impact of the conditional contrast factors, and the candidate extensions of a conditional contrast factor includes all the conditional contrast factors that are before it in the ascending order of relative impact. Therefore, the subsets of a conditional contrast factor set *FS* are always

discovered before *FS*. C2PMiner maintains the maximal number of conditional contrast factors contained in the desired conditional contrast patterns that have been generated, denoted as k_{max}. If the current exploration depth is greater than k_{max}, it means that none of the immediate subsets of the current conditional contrast factor set *FS* satisfies the minC2Pricr threshold, so there is no need to explore further based on Proposition 3.2.

Algorithm 2 shows the pseudo-codes of the DFSMineC2P(*B*, *CandC2Fs*, *L*) procedure. During the depth first exploration, C2PMiner maintains the set of C2Fs on the current path, denoted as *FS*, the accumulated relative impact of the C2Fs in *FS*, denoted as *rimp_sum*, the number of non-empty C2Fs in *FS*, denoted as *k'* and the maximal number of conditional contrast factors contained in the conditional contrast factor sets that have been generated, denoted as k_{max}.

Box 2. Algorithm 2 DFSMineC2P(B, CandC2Fs, L)

Input:
B is a conditional contrast base, *CandC2Fs* is the set of candidate C2Fs of *B* and *L* is the size of *CandC2Fs*
Output:
All desired C2Ps containing more than one non-empty C2Fs with *B* as base.
Description:
1: **for** *i*=1 to *L* **do**
2:   *FS* = *FS* ∪ *CandC2Fs*[*i*];
3:   *rimp_sum* = *rimp_sum* + *rimp*(*CandC2Fs*[*i*], *B*);
4:   **if** *CandC2Fs*[*i*] ≠ ∅ **then**
5:     *k'*= *k'*+1;
6:   **if** *k'*≥ 2 AND *ricr*(*P*=⟨*B*, *FS*⟩) ≥ *minC2Pricr* **then**
7:     Output *P* = ⟨*B*, *FS*⟩;
8:     **if** k_{max}<|*FS*| **then**
9:       k_{max} = |*FS*|;
10:   **if** *i*>0 AND |*FS*|≤ k_{max} AND (*rimp_sum* ≥ *k'* minC2Pricr*/2 OR *rimp*(*CandC2Fs*[*i*−*1*], *B*) ≥ *minC2Pricr*/2) **then**
11:       DFSMineC2P(*B*, *CandC2Fs*, *i*−*1*));
12:   *FS* = *FS*− *CandC2Fs*[*i*];
13:   *rimp_sum* = *rimp_sum* − *rimp*(*CandC2Fs*[*i*],*B*);
14:   **if** *CandC2Fs*[*i*] ≠ ∅ **then**
15:     *k'*= *k'*−1;

Initially, *rimp_sum* and k' are set to 0, k_{max} is set to 2 and *FS* is set to {}.

The correctness and completeness of the C2P-Miner algorithm is guaranteed by Proposition 3.2, Proposition 3.3 and Lemma 5.1.

EXPERIMENTAL EVALUATION

This section describes an experimental evaluation of the performance of the C2Pminer algorithm. We will show that the algorithm is effective in C2P mining; there is of course still room for further improvement. Since this is the first paper on C2P mining, there are no previous algorithms to compare against. The program was written in C++. The experiments were performed on a machine running Microsoft Windows XP professional, with a 3.00GHz CPU and 2GB memory.

Datasets Used: In this paper (here and Section 1) we consider four datasets: two microarray gene expression datasets (one for acute lymphoblastic leukemia sub-type study [Yeoh et al, 2002] and another for prostate cancer [Singh et al., 2002]), and two datasets from the UCI repository. All are dense datasets and frequently used in data mining evaluations. An entropy-based method [Fayyad & Irani, 1993] was used to discretize continuous attributes into ≥2 bins. (-23, 24] represents an in-terval; <11 represents the interval of $(-\infty,11)$. Each gene has an ID of the form 36533_at. The other two datasets are available at the UCI repository. The adult dataset was extracted from the 1994 U.S. Census. It was originally collected to predict whether an individual's income exceeds $50K per year based on census data. The attributes are concerned with personal economical, educational, and family conditions etc. Each sample contains 15 features. The mushroom dataset consists of 8124 hypothetical mushroom samples. Each sample has 23 features.

Runtime Performance: The first experiment evaluates C2PMiner's efficiency w.r.t. the four thresholds. We conducted this experiment on datasets adult and mushroom. Figures 2–5 show the results. In the figures, a question mark "?" indicates that the threshold is the varying one.

Figure 2 reports the runtime behavior of C2P-miner when varying minC2Pricr. The thresholds are fixed at (10%, 30%, 1,?) for adult, and at (10%, 50%, 1,?) for mushroom. The figure shows that execution time grows at roughly a linear speed when ricr decreases.

Figure 3 reports the runtime behavior of C2P-Miner when varying minC2Bsupp. The thresholds are fixed at (?, 30%, 1, 0.8) for adult, and at (?, 50%, 1, 0.8) for mushroom.
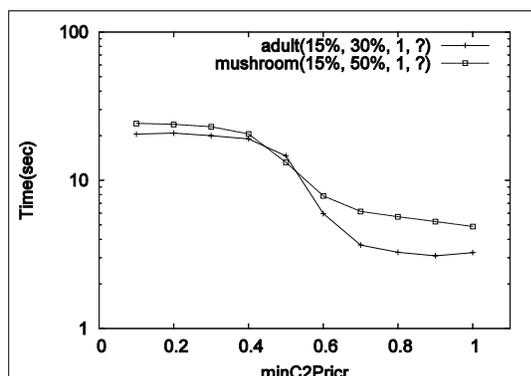
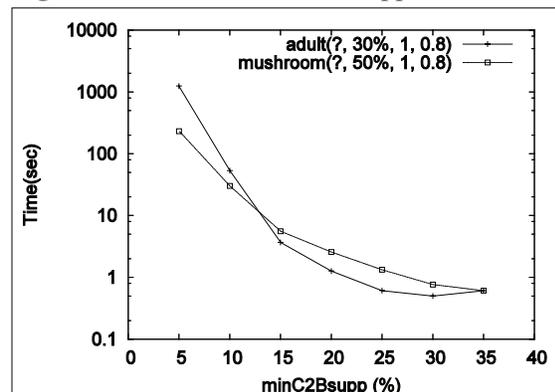*Figure 2. Runtime vs ricr*

Figure 3. Runtime vs minC2Bsupp

Figure 4. Runtime vs minC2Fsupp
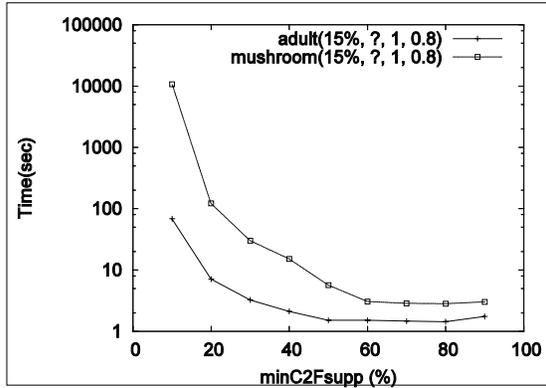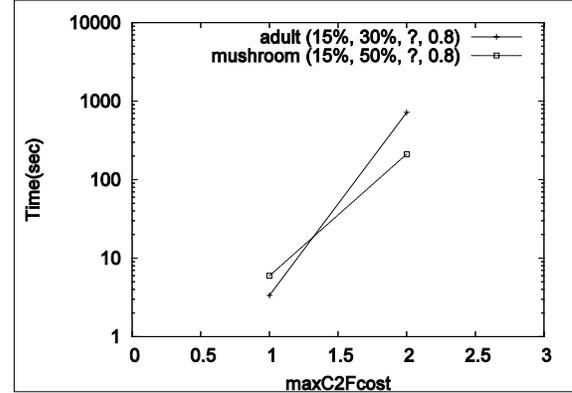

Figure 5. Runtime vs maxC2Fcost

Figure 4 reports the runtime behavior when varying minC2Fsupp. The thresholds are fixed at (10%, ?, 1, 0.8) for both datasets.

Figures 3 and 4 show that the running time of C2PMiner grows much faster than linear when minC2Bsupp or minC2Fsupp decreases. The reason is that when minC2Bsupp decreases, the number of candidate conditional contrast bases increases significantly, and when minC2Fsupp decreases, the number of candidate conditional contrast factors increases greatly. Both cases expand the search space dramatically. Nevertheless the C2PMiner algorithm can finish the mining within 10-20 minutes for the most challenging parameter settings in the figures.

Figure 5 reports the runtime behavior of C2P-Miner when varying maxC2Fcost. The thresholds are fixed at (10%, 50%, ?, 0.8) for adult, and at (10%, 60%, ?, 0.8) for mushroom. When individual C2F size of 2 is allowed, the execution time becomes much longer than when it is limited to 1. The reason is: The most expensive step in the C2PMiner algorithm generates all the C2Ps containing more than one non-empty conditional contrast factors, and the search space of conditional contrast factor sets is exponential to the number of candidate conditional contrast factors; moreover, when maxC2Fcost increases from 1 to 2, the number of potential candidate conditional contrast factors increases sharply.

CONCLUDING REMARKS

This chapter introduced the concepts of conditional contrast patterns and conditional contrast factors, as pattern types for data mining. These concepts capture small patterns that make big matching dataset differences. The paper presented theoretical results on the dominance beam property (which allows expansion-based search), on representation issues of conditional contrast patterns, and on relationship of conditional contrast patterns/conditional contrast factors with closed itemsets and keys/generators. It also designed an algorithm called C2PMiner based on those results. Experimental results demonstrated the performance of the algorithm, and produced interesting patterns from datasets on cancer research and from UCI.

ACKNOWLEDGEMENT

REFERENCES

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. *In Proc. ACM-SIGMOD Int. Conf. Management of Dat*a (pp. 207–216).

Bay, S. D., & Pazzani, M. J. (2001). Detecting Group Differences: Mining Contrast Sets. *Data Mining and Knowledge Discovery*.

Bykowski, A., & Rigotti, C. (2001). A condensed representation to find frequent patterns. *In Proc. of PODS*.

Calders, T., & Goethals, B. (2002). Mining all non-derivable frequent itemsets. *In Proc. of PKDD*. (pp. 74–85).

Dong, G., & Li, J. (1998). Interestingness of Discovered Association Rules in terms of Neighborhood-Based Unexpectedness. *In Proc. Pacific Asia Conf. on Knowledge Discovery and Data Mining*.

Dong, G., & Duan, L. (2008). *Mining Class Converters for Converting Undesirable Objects into Desirable Ones*. Submitted for publication.

Dong, G., & Li, J. (1999). Efficient mining of emerging pat-terns: Discovering trends and differences. *In Proc. of the 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*.

Fayyad, U., & Irani, K. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. *In Proceedings of the 13th International Joint Conference on Artificial Intelligence*.

Ganter, B., & Wille, R. (1999). *Formal Concept Analysis: Mathematical Foundations*. Springer, Heidelberg.

Ganti, V., Gehrke, J., Ramakrishnan, R., & Loh, W. Y. (1999). A Framework for Measuring Changes in Data Characteristics. *In Proc. of ACM PODS* (pp. 126–137).

Ji, X., Bailey, J., & Dong, G. (2005). Mining Minimal Distinguish-ing Subsequence Patterns with Gap Constraints. *In Proc. of ICDM*.

Jiang, Y., Wang, K., Tuzhilin, A., & Fu, A. (2005). Mining Patterns That Respond to Actions. *In Proc. of IEEE International Conference on Data Mining*.

Kryszkiewicz, M. (2001). Concise Representation of Frequent Patterns Based on Disjunction-Free Generators. *In Proc. of ICDM* (pp. 305–312).

Li, H., Li, J., Wong, L., Feng, M., & Tan, Y.-P. (2005). Relative Risk and Odds Ratio: A Data Mining Perspective. In Proceedings of 23rd ACM Symposium on Principles of Database Systems (pp. 368–377).

Liu, B., Hsu, W., Han, H.-S., & Xia, Y. (2000). Mining Changes for Real-Life Applications. *In Proc. of DaWaK* (pp. 337–346).

Liu, H., & Motoda, H. (1998). *Feature Selection for Knowl-edge Discovery and Data Mining*. Kluwer Academic Publishers.

Mineau, G., & Ganter, B. (Eds.) (2000). Proceed-ings of International Conference on Conceptual Structures, 2000. *LNCS 1867*, Springer.

Tan, P. N., Kumar, V., & Srivastava, J. (2002). Selecting the right interestingness measure for association patterns. *In Proceedings of KDD*.

Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L. (1999). Discovering Frequent Closed Itemsets for Association Rules. *In Proc. of ICDT 1999* (pp. 398–416).

Pawlak, Z. (1991). *Rough Sets: Theoretical Aspects of Reasoning About Data*. Kluwer Academic Publishers.

Pei, J., Han, J., & Lakshmanan, L. V. S. (2001). Mining Frequent Item Sets with Convertible Constraints. *In Proceedings of the 17th International Conference on Data Engineering* (pp. 433–442).

Singh, D., et al (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell, 1*, 203–209.

Wang, J., Han, J., & Pei, J. (2003). CLOSET+: Searching for the Best Strategies for Mining Frequent Closed Itemsets. *In Proceedings of KDD.*

Webb, G. I., Buttler, S., & Newlands, D. (2003). On Detecting Differences Between Groups. *In Proceedings of KDD.*

Yeoh, E. J., et al (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene ex-pression profiling. *Cancer Cell, 1*, 133–143.

Zaki, M., & Hsiao, C. (2002). CHARM: An efficient algorithm for closed itemset mining. *In Proceedings of SDM.*

ENDNOTES

[1] There is no $k=1$ case. To contrast the matching dataset of $B \cup F_1$ against that of B, one can use the C2P of $\langle B, \{F_1, \{\}\}\rangle$.

[2] We also found similar C2Ps from other datasets, including a prostate cancer microarray gene expression dataset [Singh et al., 2002] and some datasets from the UCI repository.

[3] One may also define a PC2P as a set $\{X_1,...,X_k\}$ of itemsets, and define the C2B and C2Fs as $B = \cap_{i=1}^{k} X_i$ and $F_i = X_i - B$.

[4] We denote the symmetric set difference as Δ. Given two sets X and Y, the set $X \Delta Y$ is defined to be *(X-Y) $\cup$ (Y-X)*.

[5] Reference [Jiang et al., 2005] considers changing utility attributes for actionable rules. Reference [Dong & Duan, 2008] considers the mining of converter sets (each of which is a set of attribute changes) to convert undesirable objects into desirable ones.

[6] To see that the unique PC2P in minCP does not have the lowest icr among PC2Ps in [P], consider $D = \{abcdk, abek, abfgjk, bg, ah\}$ and $P = \langle b, \{ak, fg\}\rangle$. Then minCP = $\{P'\}$ = $\{\langle abk, \{\{\}, fgj\}\rangle\}$ and icr(P)=1<4/3=icr(P').

[7] To see that PC2Ps in minKP do not have the highest icr among PC2Ps in [P], consider $D = \{abcdk, abek, abfk, bg, ah\}$. Let P = $\langle b, \{a, f\}\rangle$. Then one of PC2Ps in minKP is P'= $\langle\{\}, \{ab, f\}\rangle$. Clearly, icr($P'$)<icr(P).