Strong Compound-Risk Factors: Efficient Discovery through Emerging Patterns and Contrast Sets

Jinyan Li*

Institute for Infocomm Research

21 Heng Mui Keng Terrace, Singapore 119613

Email: jinyan@i2r.a-star.edu.sg

Qiang Yang

Department of Computer Science and Engineering

Hong Kong University of Science and Technology

Clearwater Bay, Kowloon, Hong Kong

* Correspondence Author: Tel. (65) 68748217, Fax. (65) 67748056

Abstract

Odds ratio, relative risk (risk ratio) and absolute risk reduction (risk difference) are biostatistics measurements that are widely used for identifying significant risk factors in dichotomous groups of subjects. In the past, they are often used to assess simple risk factors. In this paper, we introduce the concept of *compound-risk factors* to broaden the applicability of these statistical tests for assessing *factor interplays*. We observe that compound-risk factors with a high risk ratio or a big risk difference have an one-to-one correspondence to strong emerging patterns or strong contrast sets—two types of patterns which have been extensively studied in the data mining field. Such a relationship has been unknown to researchers in the past, and efficient algorithms for discovering strong compound-risk factors have been lacking. In this paper, we propose a theoretical framework and a new algorithm that unify the discovery of compound-risk factors that have a strong odds ratio, risk ratio or a risk differences. Our method guarantees that all patterns meeting a certain test threshold can be efficiently discovered. Our contribution thus represents the first of its kind in linking the risk and odds ratios to pattern mining algorithms, making it possible to find compound-risk factors in large scale data sets. In addition, we show that using compound factors can improve classification accuracy in probabilistic learning algorithms on several disease data sets, because these compound factors capture the inter-dependency between important data attributes.

1 Introduction

Odds ratio and risk ratio (relative risk) are two most often used ratios in the evidence-based biomedical research and epidemiological areas for comparative studies between two dichotomous groups of subjects [35, 37]. Risk is the chance, or probability, that a specific event occurs. Odds is an alternative measure for describing how likely an event is to happen—it is the risk of observing an event divided by the risk of not observing it. Therefore, both odds ratio and relative risks can be described as a likelihood change of an event between two dichotomous groups. In addition to comparing risks in relative terms, absolute risk reduction (risk difference) is used to capture the absolute risk difference of the factors between the two groups of subjects under comparison.

The conventional use of the above statistical measures is usually focused on simple risk factors. However, risk factors are seldom found in isolation. In this paper, complex interplays between different factors are conceptualized as *compound risk factors*. This idea is similar to many those that have been increasingly recognized as important issues in multivariate data analysis [16]. Identification of compound factors is also akin to *feature-subset selection* in the field of supervised machine learning [17] in the sense that both of them target on the dependence and interaction of a group of factors. In medical data analysis, how to find compound factors efficiently from large data sets is an open issue. In this paper, we present a data mining approach to solving this problem.

To highlight the importance of compound factors, consider the following two examples. The health state of adolescents was found [1] to be affected by the complex interplay of factors between a young person and his social environment that involves his family, peers, media and social norms, among others. One finding is that around 70% of the adolescent deaths and illness were caused by six categories of risk behaviors including alcohol use and drug abuse, tobacco use and unhealthy dietary behaviors, which in turn were the result of numerous factors within a young person's life—no single factor alone causes or explains a risk behavior [27]. Thus, the health state of a young person is determined by multiple factors that are inter-dependent together.

[1] In a study conducted by Minnesota Department of Health. See more details at `http://www.health.state.mn.us/divs/fh/mch/adolescent/actionplan/sect1/section1-6.html`.

The following example further strengthens this point, where if only individual factors are considered, some contradictory results may be derived. A survey [26] studied the relationships between satisfaction perceptions of care for dental patients and various variables (risk factors) including socio-demographic factors such as age, gender and economic status. The survey points out that findings by different authors are contradictory: If the age factor is considered alone, one study [33] observed that patients over 60 years old tended to be more satisfied with their dental care than younger patients. However, by another study [18], it was found that older patients were less satisfied. When combined with other factors, this contradiction disappears.

In this paper, these multiple interacting factors are referred to as *compound-risk factors*. By exploring the use of compound-risk factors, we can not only discover the outstanding single risk factors, but also the strong and subtle interplays of such factors even when they are weak individually. We define three types of compound-risk factors. One is defined in the context of relative risk, a second is defined in the context of risk difference and the third is defined in the context of odds ratio. We call the compound-risk factors with a high relative risk, a large risk difference, or a high odds ratio, the relative risk patterns, risk difference patterns, or odds ratio patterns, respectively.

Given the importance of compound factors, an open question is how to find such factors from a large data set efficiently. In this paper, we answer this question by presenting two linked contributions. First, we show that relative risk patterns are actually *emerging patterns* [13, 22], a data mining concept that captures sharp frequency growth change of patterns from one class to another class. Similarly, a counterpart of the extension for the absolute risk factors is the so-called *contrast sets*, which is another data-mining pattern concept [3, 4, 36] emphasizing on large frequency differences of patterns between the two classes of subjects. This correspondence makes it possible for efficiently discovering *strong* compound-risk factors from a database by exploiting data mining algorithms for strong contrast sets and emerging patterns.

A second contribution of our work is a new algorithm. As combinations of risk factors are considered here, identifying strong interplays is a computationally difficult problem because such combinations are many. In the past, the problem of discovering contrast sets and emerging patterns have been explored separately in [3, 4, 36] and in [24, 13, 14, 22, 21, 23, 5]. In this

paper, we introduce a unified approach to the discovery of the three types of compound-risk patterns. This unified approach integrates the two groups of subjects into one data set and then discover frequent closed patterns [28] and generators [2] from the combined data set. Then, strong compound-risk factors can be derived based on the *support* information of the discovered closed patterns and generators.

An additional contribution of the paper is that we provide a *2-D risk plane* to visually plot the two risk information of an event in a given pair of dichotomous groups. In the long-standing debate, some researchers are not in favor of odds ratio [1, 6, 10, 11, 31, 39] while others favor odds ratio as a gold standard [8, 32]. Our 2-D risk plane can help reconcile these two views, by explaining this discrepancy. This 2-D risk plane is also helpful to explain the computational difficulties in mining all compound-risk factors that have a wide range of values.

An added contribution of our work is showing that using strong compound-risk factors can be useful for improving the performance of Naive Bayes (NB) classifiers [19, 12], which operates under the independence assumption, particularly in applications where risk factors are strongly interacted. We take a leukemia disease diagnosis problem as an example to demonstrate that the classification accuracy can be indeed improved if in Naive Bayes, we replace the individual risk factors from single genes with compound-risk factors in gene groups. This increase in accuracy are also observed in some UCI machine learning data sets where traditional NB suffers dramatic accuracy loss, compared to other popular classifiers such as C4.5, SVM, and k-nearest neighbor. However, by using compound-risk factors, we can effectively overcome these problems.

The rest of the paper is organized as follows: In Sections 2 and 3, we give definitions for odds, risk, odds ratio, relative risk, and absolute risk reduction, and explain discrepancies between odds ratio and relative risk using our 2-D risk plane. In Section 4, we define compound-risk factors, and relate them to contrast sets and emerging patterns in the context of relative risk and absolute risk reduction. Then we present a unified approach to the discovery of compound-risk factors that have a strong odds ratio, risk ratio and absolute risk reduction. Section 5 reports classification performance of our method in comparison to NB and other classic learning algorithms on a cancer diagnosis problem and some UCI data sets. Section 6 concludes this paper.

2 Background: Odds Ratio, Relative Risk, and Absolute Risk Reduction

The *odds* of an event happening is the probability that the event of interest occurs divided by the probability that the event does not happen. This is usually estimated by the ratio of the occurrence number of the event to the number of times the event does not happen. Moreover, the *risk* of an event happening is simply the number of occurrence the event happens divided by the total number of cases at risk of having the event. It is often expressed as a *percentage* or *frequency*.

As an example, consider the following data on the survival of the passengers on the Titanic tragedy.

	dead	alive	total
Female	154	308	462
Male	709	142	851

The death odds of the female is 1 to 2 (154/308 = 0.5); The death risk of the female is 33% (154/462 = 0.33)—This is exactly the probability of death in the female group. Similarly, the death odds of the male is 5 to 1 (709/142 = 4.99); The death risk of the male is 83% (709/851 = 0.83). In general, the value range for odds is from zero (event will never happen) to infinity (event is certain to happen). While for risk it is between 0 and 1.

Odds ratio (OR), relative risk (RR), and absolute risk reduction (ARR) are defined through risk or odds comparison between the two groups of subjects.

Definition 1 *An OR is calculated by dividing the odds of an event in one group by the odds in the other group; A RR is the ratio of the risk of the event in one group to the risk in the other group; An ARR is the risk difference between the two groups.*

Odds ratio, relative risk, and absolute risk reduction are most used in clinical trials [37] where treatment/drug effects are compared between patients receiving the new treatment/drug

and those not, and in epidemiological studies [35] where the risk of a disease are compared between people who had been exposed to some risk factor and people who not exposed to. As an example, we re-use data from a 11-year cohort follow-up study of 9510 male physicians [25] where the aim is to examine the association between baldness and coronary heart disease—whether there exists big risk difference of this disease between male physicians with normal hair and those balding. In that study, data show that 548 of those doctors with normal hair developed coronary heart disease during the 11 follow-up years, and 127 of those doctors with severe vertex balding developed the same disease. More detailed data are in the following table:

	heart disease	healthy	total
Balding	127	1224	1351
Hairy	548	7611	8159

The OR for the balding doctors to develop coronary heart disease over those hairy is

$$\frac{127/1224}{548/7611} = 1.44$$

The RR for the balding doctors to develop coronary heart disease over those hairy is

$$\frac{127/1351}{548/8159} = \frac{9.4\%}{6.7\%} = 1.40$$

The ARR is

$$\frac{127}{1351} - \frac{548}{8159} = (9.4 - 6.7)\% = 2.7\%$$

Therefore, a conclusion has been made in [25] that people with severe vertex balding are more likely to develop coronary heart disease than people with normal hair.

Observe that the above OR value is larger than the RR value. In fact, it is always true that OR is larger than RR when RR is larger than 1. In the Titanic example, the OR of death between male and female is 9.99—There is a ten fold greater odds of death for male than for female. But the RR of death is only 2.5—There is a 2.5 times greater probability of death for male than female. As OR is always larger than RR for the same level of risk, if the odds ratio is misunderstood as a relative risk, then the risk is overstated very much. Previous work [10, 39] have observed and discussed this discrepancy—when odds ratios can mislead. We discuss it

again, in the next section, under a new framework: our 2-D risk planes. The insight into these discussions can also help understand the computational challenges in the mining of all strong compound-risk factors.

3 A 2-D Risk Plane

A 2-D *risk plane* co-ordinates the risk information of an event happening between two groups. Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two classes of subjects (e.g., smokers vs non-smokers), and E be an event (e.g., lung cancer), the x-axis of the risk plane measures the risk of E in $\mathcal{D}_1$, and the y-axis measures the risk of E in $\mathcal{D}_2$. So, a point (x, y) in the risk plane plots the risk information of an event in the two classes as shown in Figure 1. Note that different events may share a same point in the risk plane. That is, two different events can have the same level of risk in $\mathcal{D}_1$ and in $\mathcal{D}_2$. However, some points in the risk plane do not have a corresponding event.

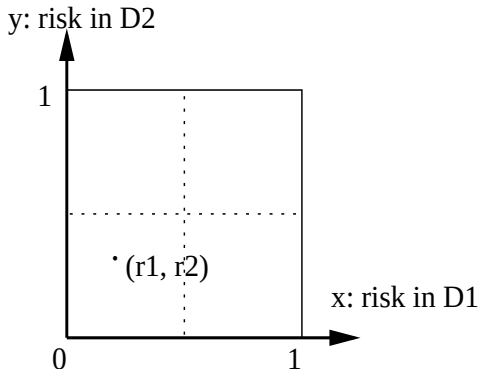


Figure 1: A 2-D risk plane for showing risk information of an event between two groups $\mathcal{D}_1$ and $\mathcal{D}_2$.

Given a point (r_1, r_2) in the risk plane, we can easily calculate the corresponding OR, RR, and ARR as follows:

$$OR = (r_1/(1 - r_1))\,/(r_2/(1 - r_2))$$

$$RR = r_1/r_2$$

and

$$ARR = r_1 - r_2$$

$OR = (r_1/(1-r_1))/(r_2/(1-r_2))$ can be re-written as $\frac{r_1}{r_2}\frac{1-r_2}{1-r_1} = RR * \frac{1-r_2}{1-r_1}$. Thus, if $RR = \frac{r_1}{r_2} > 1$, we have $\frac{1-r_2}{1-r_1} > 1$. Therefore $OR > RR$ if $RR > 1$. Similarly, we can see that OR is always less than RR if $RR < 1$. Furthermore, $OR \approx RR$ if both r_1 and r_2 are small (close to 0).

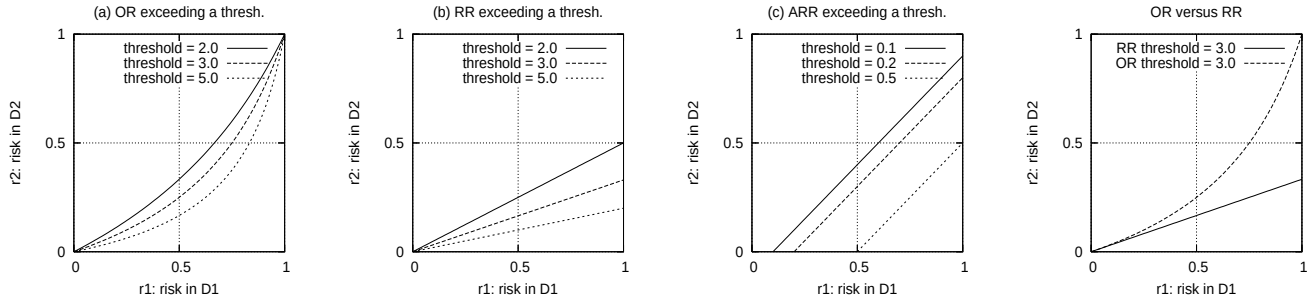


Figure 2: Regions (up-bounded by the different lines) in a 2-D risk plane where events have an OR, RR, or ARR exceeding some thresholds.

Figure 2 plots areas (all points below a solid, dashed, or a dotted line) in a 2-D risk plane where the event points have an OR, RR, or ARR value larger than a threshold. We note that:

- An event (e.g. developing coronary heart disease) may have a very large odds ratio and a very large relative risk even when the risk in the two groups are both small (close to 0). See the left bottom corner of Figure 2(a) and Figure 2(b).
- If the risk of the event in the two groups are very large (close to 1), its odds ratio can still be large. For example, let $r_1 = 0.99$ and $r_2 = 0.90$, then $OR = 11.0$. However in this situation, the relative risk is only 1.1—a value which is rarely significant. Thus, the OR information is not sufficient to infer whether the risk themselves in the two groups or the relative risk is high or low.
- Unlike the case of OR, a pair of small risks or a pair of large risks cannot produce a big ARR. Only the points in the right bottom corner of a 2-D risk plane are possible to produce a big ARR. See Figure 2(c).
- For the same threshold, events for which the corresponding OR values exceed this threshold are a superset of events that have a RR value that exceed this threshold (see Figure 2(d)).

These observations suggest that to discover a complete set of compound-risk factors with a strong OR or RR test value, small events (those with a pair of small risks in $\mathcal{D}_1$ and $\mathcal{D}_2$) have to be examined. Otherwise, some strong compound-risk factors will be missed. Our algorithms will carefully examine small events, usually comprising of many single factors, by using efficient data structures and search strategies. Actually, this is the most computationally challenging part in this problem as possible combinations of simple risk factors are exponentially large.

As shown in the top righthand corner of Figure 2(a), odds ratio is prone to misunderstanding. However, odds ratio is a gold standard for risk comparison in case-control (retrospective) studies and also in cohort studies [8, 32]. This is because both *disease odds ratio* and *exposure odds ratio* can be calculated based on data in a case-control study or in a cohort study. But, only *disease relative risk* or *exposure relative risk* (not both) can be calculated unless the sampling of the patients follows the distribution of the population. (Our explanation is provided later in this section.) However, such a sampling is not commonly used in practice due to high financial cost.

Let us explain these points using a case-control study as an example. The research design of a case-control study is to select patients on the basis of an outcome variable. The cases are a group of patients with a specific outcomes (e.g. lung cancer or coronary heart disease) and controls are those without that outcome. The goal of this type of study is to look backwards in time and to figure out what risk factors (e.g. smoking or balding) are more likely to cause the outcome. A synthetic data set of a case-control study is shown in the following table.

Exposure to a risk factor	Diseased cases	Controls	Total
Yes (+)	D_+	C_+	Y
No (–)	D_-	C_-	N
Total	D	C	T

The exposure OR is defined as the ratio of the odds of exposure for the cases' group to the odds of exposure in the control's group. That is

$$exposureOR = \frac{(D_+/D)/(D_-/D)}{(C_+/C)/(C_-/C)} = \frac{D_+/D_-}{C_+/C_-} \tag{1}$$

The exposure RR is defined as

$$exposureRR = \frac{D_+/D}{C_+/C} \tag{2}$$

The above two definitions are meaningful because D_+/D and C_+/C are indeed the probability (risk) of exposure for the diseased patients and controls respectively.

However, D_+/Y (or D_-/N) is not the probability of suffering a disease for people who had been exposed to some risk factors unless $D_+ : C_+$ (or $D_- : C_-$) is proportional to the whole population. Let z be a real number such that $D_+ : (C_+ * z)$ is the distribution of cases and controls of the population. Then $D_+/(D_+ + C_+ * z)$ is a risk, meaning that the probability of suffering a disease for people who had been exposed to some risk factor. Then

$$diseaseRR = \frac{D_+/(D_+ + C_+ * z)}{D_-/(D_- + C_- * z)} \tag{3}$$

is a true RR (called disease RR). Without the knowledge of z, the disease RR cannot be calculated in a case-control study.

The disease OR is defined as

$$diseaseOR = \frac{D_+/(C_+ * z)}{D_-/(C_- * z)} = \frac{D_+/C_+}{D_-/C_-} \tag{4}$$

which is always equal to the exposure OR as shown in (1). Thus, an OR can be always calculated in a case-control study or in a cohort study regardless of the distribution of the subject numbers in the two groups, and it is also invariant with regard to the arbitrary decision of deciding whether we concentrate on the relative odds of dying (disease OR) or the relative odds of exposure (exposure OR). This gold standard was also observed by [8, 32], where no mathematical details were given though.

4 Risk-Factor Discovery by Mining the Contrast Sets and Emerging Patterns

In this section, we explain the one-to-one correspondence between emerging patterns [13, 22] and compound-risk factors in the context of relative risks, and explaining the one-to-one correspondence between contrast sets [3, 4] and compound-risk factors in the context of absolute risk reduction. As previous algorithms for mining contrast sets [3, 4] and emerging patterns

[24, 13, 14, 22, 21, 23, 5] are disparate and computationally costly, we provide an efficient, unified method to discover compound-risk factors that have strong odds ratio, risk ratio and risk differences.

We first review the definition for contrast sets and a definition for emerging patterns. A *data set* is a set of *tuples*. A tuple is also called a *transaction* or a *subject*. Let $I = \{I_1, I_2, \cdots, I_k\}$ be a set of distinct *items* or *risk factors*. A tuple is defined as a non-empty set of *items*. A *pattern*, or called an *itemset*, is a set of items. A pattern P is said to *occur* or *be contained* in a tuple T if $P \subseteq T$. The *support* of a pattern P in a data set $\mathcal{D}$, denoted $sup(P, \mathcal{D})$, is the number of tuples in $\mathcal{D}$ that contain P divided by the total number of tuples in $\mathcal{D}$. The *negated* support of P in $\mathcal{D}$, denoted $\overline{sup}(P, \mathcal{D})$, is defined as $\overline{sup}(P, \mathcal{D}) = 1 - sup(P, \mathcal{D})$.

Definition 2 *Let $\mathcal{D}$ be a data set consisting of a set $\mathcal{D}_p$ of positive tuples and a set $\mathcal{D}_n$ of negative tuples. Given a real number $\delta(> 0)$, a pattern X is called a δ-strong* **contrast set** *if*

$$CS_X^{\mathcal{D}} = sup(X, \mathcal{D}_p) - sup(X, \mathcal{D}_n) \geq \delta$$

where δ is called a difference threshold.

Definition 3 *Let $\mathcal{D}$ be a data set consisting of a set $\mathcal{D}_p$ of positive tuples and a set $\mathcal{D}_n$ of negative tuples. Given a real number $\rho(> 1)$, a pattern X is defined as a ρ-strong* **emerging pattern** *if*

$$EP_X^{\mathcal{D}} = \frac{sup(X, \mathcal{D}_p)}{sup(X, \mathcal{D}_n)} \geq \rho$$

where ρ is called a growth ratio threshold.

Contrast sets and emerging patterns defined above are patterns characterizing the positive class. Conversely, we can define patterns for characterizing the negative class by swapping $\mathcal{D}_p$ and $\mathcal{D}_n$ in the above definitions.

Consider why contrast sets are compound-risk factors in the context of absolute risk reduction, and why emerging patterns correspond to relative risks. We translate the concepts as follows.

positive class ($\mathcal{D}_p$)	$\rightarrow$	cases' group
negative class ($\mathcal{D}_n$)	$\rightarrow$	controls group
single risk factor	$\rightarrow$	an item
compound-risk factor	$\rightarrow$	an itemset
exposure risk to a risk factor	$\rightarrow$	support of an itemset

then the *risk difference* of a compound-risk factor between the two groups can be exactly said to be the *support difference* of an itemset between the positive and negative classes. In the same way, the *risk ratio* (the relative risk) can be said to be the *support growth ratio* between the two classes. Thus, δ-strong contrast sets are exactly compound-risk factors with a strong ARR test value, and ρ-strong emerging patterns are compound-risk factors with a strong RR test value.

The statistical test by odds ratio can be similarly applied to compound-risk factors, and thus we define a new type of pattern:

Definition 4 *Let $\mathcal{D}$ be a data set consisting of a set $\mathcal{D}_p$ of positive tuples and a set $\mathcal{D}_n$ of negative tuples. Given a real number $\beta(> 1)$, a pattern X is a β-strong* **odds ratio pattern** *if*

$$OR_X^{\mathcal{D}} = \frac{sup(X, \mathcal{D}_p)/\overline{sup}(X, \mathcal{D}_p)}{sup(X, \mathcal{D}_n)/\overline{sup}(X, \mathcal{D}_n)} \geq \beta$$

where β is called an odds ratio threshold.

As mentioned earlier, our computational problem is to find all compound-risk factors that have a strong odds ratio, relative risk, or/and absolute risk reduction test value from a data set $\mathcal{D}$ containing two classes of tuples. This problem is equivalent to finding all δ-strong contrast sets, ρ-strong emerging patterns, and β-strong odds ratio patterns from $\mathcal{D}$. The discovery of contrast sets is easy if the threshold δ is high, because mining frequent patterns of $\mathcal{D}_1$ with large support is an easy task (See Figure 2(c)). However, mining emerging patterns or odds ratio patterns is difficult no matter how large the thresholds ρ and β are. This is because a pattern with a small support value (i.e., risk) in both classes can be still a strong emerging pattern or a strong odds ratio pattern (see Figure 2 (a) and (b)). This represents a main contribution of this paper.

In fact, the discovery of emerging patterns is computationally difficult even for a subtype of emerging patterns called boundary-jumping emerging patterns [22]. A boundary jumping

Trans. id	items				
	a	b	c	d	e
1	1	1	1	0	1
2	0	0	1	1	0
3	1	1	0	0	1
4	1	1	0	1	1

Table 1: A data set $\mathcal{D}$ consisting of 4 transactions.

emerging pattern X is a pattern satisfying $sup(X, \mathcal{D}_p) > 0$, $sup(X, \mathcal{D}_n) = 0$ but $sup(X', \mathcal{D}_n) > 0$ where X' is any proper subset of X. These emerging patterns (events) are located only in the x-axis of Figure 2 (b). The complexity of mining boundary jumping emerging patterns is MAX NP-hard [34].

Here, we present a unified approach to the discovery of the three types of patterns. This approach is based on the concepts of equivalence classes [2], closed patterns [28], and generators [2].

Definition 5 *Let $\mathcal{D}$ be a data set, then an* **equivalence class** *in $\mathcal{D}$ is a set of those itemsets that always occur together in the same subset transactions of $\mathcal{D}$.*

In other words, for any two itemsets X and Y, they are in the same equivalence class iff $f_{\mathcal{D}}(X) = f_{\mathcal{D}}(Y)$ where $f_{\mathcal{D}}(Z) = \{T \in \mathcal{D} \mid Z \subseteq T\}$.

Example 1 *Table 1 shows a small data set $\mathcal{D}$. The following itemsets constitute an equivalence class: a, b, e, ab, ae, be, abe. This is because all of these itemsets occur in all and only transactions of 1, 3 and 4. Itemset cd is not in this equivalence class, as it occurs only in transaction 2.*

Proposition 1 *Let π be a support threshold, and $\mathcal{D}$ be a data set. Then the frequent patterns of $\mathcal{D}$ can be partitioned into frequent equivalence classes without any overlapping.*

Proof: *Let EC_1 and EC_2 be two π-frequent equivalent classes. Assume $P \in EC_1 \cap EC_2$. Then $f_{\mathcal{D}}(P) = f_{\mathcal{D}}(X)$, for all $X \in EC_1$, and $f_{\mathcal{D}}(P) = f_{\mathcal{D}}(Y)$, for all $Y \in EC_2$. Therefore*

$f_{\mathcal{D}}(X) = f_{\mathcal{D}}(Y)$, for all $X \in EC_1$, and all $Y \in EC_2$. This is a contradiction. Thus, we can see that the frequent patterns of $\mathcal{D}$ can be partitioned into frequent equivalence classes without any overlapping. □

Definition 6 *Let EC be an equivalence class in a data set $\mathcal{D}$. The maximal pattern of EC is defined as the* **closed pattern** *of $\mathcal{D}$, the minimal ones are defined as the* **generators** *(also called* **key patterns***) of $\mathcal{D}$.*

Proposition 2 *An equivalence class EC can be concisely represented in the form $[\mathcal{G}_{EC}, C_{EC}]$ where $\mathcal{G}_{EC}$ is the multi-set of generators of EC, C_{EC} is the closed pattern of EC, and $[\mathcal{G}, C] = \{X \mid X \subseteq C, X \supseteq Y, Y \in \mathcal{G}\}$.*

This property is known as convexity [28]. That is, the generators, which are the lower bound, and the closed patterns, which are the upper bound, can cover the whole EC in a lossless way.

Example 2 *Following Example 1, we can see that the closed pattern is abe, while the generators are a, b, and e. The equivalence class can be represented as $[\{a, b, e\}, abe]$. It is concise as it uses only 4 itemsets to represent 7 itemsets in this equivalence class.*

Proposition 3 *Let EC be an equivalence class in a data set $\mathcal{D}$ consisting of two classes of tuples $\mathcal{D}_p$ and $\mathcal{D}_n$. Let X and Y be two patterns in EC, then $CS_X^{\mathcal{D}} = CS_Y^{\mathcal{D}}$, $EP_X^{\mathcal{D}} = EP_Y^{\mathcal{D}}$, and $OR_X^{\mathcal{D}} = OR_Y^{\mathcal{D}}$. That is, all patterns in EC are at the same level of significance in terms of risk reduction, growth ratio, or odds ratio.*

Proof: *It is straightforward to prove this proposition by definition, as only the support information are needed to calculate the ratio or difference values.* □

By Propositions 1, 2, and 3, we only need to discover closed patterns and generators from a data set. Then we can find out which patterns are strong by ranking the risk reduction, growth ratio, or odds ratio values of these closed patterns. The pseudo code of our algorithm, called CRF, is described by Figure 3.

Algorithm 1 CRF: Mining strong compound-risk factors

Input: $\mathcal{D} = \mathcal{D}_p \cup \mathcal{D}_n$, π, δ, ρ, and β.

Output: The complete set of δ-strong CS, ρ-strong EP, and β-strong OR.

Method:

1: Discover π-frequent closed patterns C and generators from $\mathcal{D}$;
2: **for** each c in C **do**
3: $x_c = sup(c, \mathcal{D}_p) - sup(c, \mathcal{D}_n)$;
4: **if** $x_c \geq \delta$ **then**
5: output c and its corresponding generators;
6: **end if**
7: $y_c = sup(c, \mathcal{D}_p)/sup(c, \mathcal{D}_n)$;
8: **if** $y_c \geq \delta$ **then**
9: output c and its corresponding generators;
10: **end if**
11: $z_c = \frac{sup(c,\mathcal{D}_p)/\overline{sup}(c,\mathcal{D}_p)}{sup(c,\mathcal{D}_n)/\overline{sup}(c,\mathcal{D}_n)}$;
12: **if** $z_c \geq \delta$ **then**
13: output c and its corresponding generators;
14: **end if**
15: **end for**

Figure 3: Our algorithm (CRF) for discovering three types of strong compound-risk factors.

Subject id	Risk factors							Class
	a	b	c	d	e	f	g	pos./neg.
1	1	1	1	1	1	0	0	p
2	1	1	1	1	1	0	0	p
3	1	1	1	1	0	1	0	p
4	1	1	1	1	0	0	1	p
5	1	0	0	0	1	1	1	n
6	0	1	0	0	1	1	1	n
7	0	0	1	0	1	1	1	n
8	0	0	0	1	1	1	1	n

Table 2: A data set containing strong compound-risk factors but no strong single risk factors.

We have implemented this algorithm using a depth-first search strategy on the set-enumeration tree [30] for mining both generators and closed patterns. However, by the minimum description length (MDL) principle, generators are preferable to closed patterns [20]. Thus, in the implementation, we change the codes (in lines 5, 9, 13) of the algorithm in Figure 3 to output only frequent generators. The performance is often 10-100 times faster [20] than a previous algorithm [28]. The implementation and data structure details for mining frequent generators can be found in our earlier work [20].

We have also observed that top-ranked patterns usually are itemsets having a risk ratio of infinity. That is, the support of these itemsets is non-zero (sometimes very big) in one class but zero in the other class. Of course, these itemsets have an odds ratio of infinity as well. We denote them by $P_{\mathcal{D}}^{\infty}$. For two patterns $X, Y \in P_{\mathcal{D}}^{\infty}$ satisfying $X \subseteq Y$, then $f_{\mathcal{D}}(X) \supseteq f_{\mathcal{D}}(Y)$. Thus, Y is a redundant pattern in $P_{\mathcal{D}}^{\infty}$. Therefore, only minimal patterns of $P_{\mathcal{D}}^{\infty}$ are non-redundant. To rank these non-redundant patterns, we refer to their non-zero support levels. These ideas can be implemented as we adopt a depth-first search strategy. In the following section, for classification applications, we only used these non-redundant patterns that have a strong relative risk. The efficiency of our this implementation is usually two orders of magnitude faster than previous emerging pattern mining algorithms [24, 13, 14, 22] for this same mining task. See appendix for the details.

Next, we show an example to illustrate how equivalence classes are used in mining strong compound-risk factors that are also non-redundant.

Table 2 shows a simple data set $\mathcal{D}$ consisting of two classes (p/n) of 8 subjects, where 1s in a subject mean the subject was exposed to the risk factors, otherwise 0s mean the subject was not.

Let the support threshold π be 2, then the following equivalence classes are frequent. The singleton equivalence classes (followed by their support in $\mathcal{D}$) are: e:6, a:5, b:5, c:5, d:5, f:5, g:5, ae:3, be:3, ce:3, de:3, af:2, bf:2, cf:2, df:2, ag:2, bg:2, cg:2, and dg:2. The other 3 multi-member equivalence classes are:

$$[\{ab, ac, ad, bc, bd, cd\}, abcd] : 4$$

$$[\{ef, eg, fg\}, efg] : 4$$

and

$$[\{abe, ace, ade, bce, bde, cde\}, abcde] : 2$$

The risk ratio (p/n) of the risk factor a is $100\%/25\% = 4.0$; its odds ratio is $\frac{4/0}{1/3} = +\infty$; and its risk reduction is $100\% - 25\% = 75\%$. The other single risk factors such as b, c, and d have the same level of risk ratio, odds ratio, and risk reduction as a's.

The risk ratio (p/n) of the compound risk factor *abcd* is $100\%/0\% = +\infty$; its odds ratio is $\frac{4/0}{0/4} = +\infty$; and its risk reduction is $100\% - 0\% = 100\%$. In the equivalence class of *abcd*, the generators (ab, ac, ad, bc, bd, cd) have the same level of risk ratio, odds ratio, and risk reduction as abcd. Observe that all of these compound-risk factors have stronger test values than the single risk factors.

The test values of risk factors characterizing the negative class such as *efg*, can be similarly calculated.

As discussed above, our algorithm outputs only minimal patterns of $P_{\mathcal{D}}^{\infty}$ (that are also non-redundant). So in this example, our algorithm outputs only ab, ac, ad, bc, bd, and cd—the generators of the equivalence class of *abcd*. However, the generators of the equivalence class of abcde are not in the output as they are redundant.

5 Case Study: Compound-risk Factors for Accurate Diagnosis in Leukemia Disease

The Naive Bayes (NB) classifier [19, 12] assumes that all risk factors (the attributes) are independent. According to Bayes Rule, the probability of a subject $E = (rf_1 = a_1, rf_2 = a_2, \cdots, rf_n = a_n)$ being class c is

$$p(c|E) = \frac{p(E|c)p(c)}{p(E)}$$

where rf_i, $i = 1, \cdots, n$, are the n attributes. As NB assumes the independence among all the attributes, $p(E|c)$ can be rewritten as:

$$p(E|c) = \prod_i p(rf_i = a_i|c)$$

Then a class label $g(E)$ predicted by NB is:

$$g(E) = argmax_{c \in C} \;\; p(c) \times \prod_i p(a_i|c)$$

where C is the set of class labels.

The concept of compound-risk factors does not assume the independence among the individual risk factors, instead it emphasizes on their interplays. A class label $g'(E)$ predicted by using k compound-risk factors is:

$$g'(E) = argmax_{c \in C} \;\; p(c) \times \prod_i^k p(crf_i = b_i|c)$$

where crf_i, $i = 1, \cdots, k$, are k top-ranked compound-risk factors. Suppose crf_i consists of three attributes, say, rf_1, rf_2 and rf_3, then $b_i = (a_1, a_2, a_3)$. We term this modified NB as crf-NB. In this paper, we set k as 10. We also tried other choices such as $k = 20, 30$. There were no much difference in the classification performance.

We now discuss how compound-risk factors are concerned in a real-life application. One of important problems in in-silico cancer diagnosis is the subtype classification of childhood leukemia and identification of gene groups that are responsible for this disease [38]. Actually, this problem can be stated as finding out which compound-risk factors (gene groups) that cause those specific

Number of subjects	Subtypes of the disease							Total
	T-ALL	E2A-PBX1	TEL-AML1	BCR-ABL	MLL	Hyperdip> 50	misc.	
Training	28	18	52	9	14	42	52	215
Test	15	9	27	6	6	22	27	112

Table 3: Numbers of training and test subjects for each subtype of the disease.

subtypes of this disease, and which of them have strong prediction power for diagnosis. In that study [38], the whole data consists of gene expression profiles of 327 patients samples. These profiles were obtained by hybridization on the Affymetrix U95A GeneChip containing probes for 12558 genes. Table 3 shows the number of training subjects and the number of test subjects [38] for each of the subtypes of the disease. All the training data are discretized and only top 20 χ^2 ranked genes are selected for the discovery of strong compound-risk factors.

For this diagnosis classification problem, we examine which of the two methods, NB or our crf-NB method, can make a clear distinction between anyone specific subtype against all other subtypes. Table 4 shows that our method reduces the *test errors* of NB by half. (By the test errors of a classifier, we mean the number of mistakes made by the classifier in the test subjects.) Our performance is also comparable to those achieved by the classic non-linear classifiers such as SVM [9] and 3-nearest neighbor (3-NN). We note that the two non-linear classifiers cannot produce any compound-risk factors explicitly. The performance by C4.5 [29], Bagging [7] and Boosting [15] are much worse than our method in this application. (The main software package used in the experiments is Weka version 3.2, its Java-written open source are available at `http://www.cs.waikato.ac.nz/ml/weka/` under the GNU General Public Licence.)

Our compound-risk factors responsible for the BCR-ABL subtype are found to contain 3 or 4 genes. For example, one group contains 3 genes corresponding to probe numbers 40698_at, 39730_at, and 1211_s_at in the Affymetrix microarry gene chips. This compound-risk factor has a support of 89% in the BCR-ABL class, but no occurrence in other subtypes. Thus, this compound-risk factor has an OR or RR value of infinity and an ARR of 0.89. Note that when we decompose this compound-risk factor, then the three single genes do not have good OR, RR, or ARR values. Perhaps this is a direct reason why NB made 5 more mistakes than our method for separating BCR-ABL from other subtypes. More importantly for biological use, the three

Datasets	Numbers of Test Errors						
	Our crf-NB	NB	C4.5	Bagging	Boosting	SVM	3-NN
BCR-ABL	1:0	0:6	1:4	2:0	1:4	1:1	1:0
E2A-PBX1	0:0	0:0	0:0	0:0	0:0	0:0	0:0
HyperL50	2:2	0:5	4:5	4:2	1:4	0:3	1:4
MLL	0:0	0:1	1:1	0:0	1:1	0:0	0:0
T-ALL	0:0	0:0	0:1	0:1	0:1	0:0	0:0
TEL-AML1	2:0	0:3	3:1	1:0	1:0	1:1	2:0
Total	7	15	21	10	14	7	8

Table 4: Error numbers of 7 classifiers on the test data for the subtype classification of childhood leukemia. Here, the symbol $x : y$ means x number of errors made in the subtype specified in the first column, and y number of errors made in all other subtypes, by a classifier.

genes should be investigated in an interactive way, not in a separate way.

For some UCI data sets [2] such as the tic-tac-toe, sick and mushroom data sets, we have observed an interesting thing. We found that NB suffers from a dramatic accuracy loss: 95% as opposed to 100% obtained by other classifiers including SVM, Bagging and Boosting on the mushroom data set, 93% as opposed to 98% by the other three classifiers on the sick data set, and 70% as opposed to 98% by the three on tic-tac-toe (see detailed comparison in Table 5). However, our method crf-NB did not suffer from this problem—It matches the performance of SVM, Bagging, or Boosting on these data sets. Note that for the tic-tac-toe data set, strong compound-risk factors consist at least 3 attributes. However, none of the 9 single attributes have a good relative risk. This indicates that strong compound-risk factors have to be used, since otherwise the probability independence rule, as exploited by NB, may not work well. We also note that NB and our crf-NB have comparable performance on some other UCI data sets such as breast cancer, cleve, heart, hepatitis, hiv, hypo, and lymph, where SVM and Boosting did not achieve high classification accuracy.

As discussed above, by replacing single risk factors with compound-risk factors, the Naive Bayesian classifier can improve its accuracy in most cases. However, there are several problems that require further investigation. For example, a problem is how to deal with data sets having more than two decision classes. Another problem is how to handle missing/noise data, and how

[2] `http://www.cs.uci.edu/~mlearn/MLRepository.html`

Table 5: Performance comparison between our method, NB and other classifiers by 10-fold cross validation.

Data sets	Accuracy (%)					
	Our crf-NB	NB	SVM	C4.5	Bagging	Boosting
mushroom	100	95.5	100	100	100	100
sick	98.4	92.9	93.9	98.6	98.9	99.2
tic-tac-toe	98.8	**69.5**	98.3	85.8	92.6	96.1

to evaluate the robustness of the approach. Although addressing these future issues is not the main purpose of this paper, readers are still referred to [14] for partial answers.

6 Conclusion

In this paper, we have introduced three types of compound-risk factors to broaden the applicability of the statistical ideas behind odds ratio, relative risk, and absolute risk reduction. To efficiently discover the three types of strong and subtle interplays of the risk factors, we have presented a unified method that make use of the closed patterns and generators of the data sets from the dichotomous classes. Our algorithm can be easily adapted to output non-redundant compound-risk factors from $P_{\mathcal{D}}^{\infty}$ that have good potential in classification. Our algorithm is found to be two orders of magnitude faster than previous algorithms for this mining task. We have also observed that classification by using compound-risk factors can be more accurate than Naive Bayesian when attributes are not independent. In our case study, compound-risk factors have been shown to be useful for leukemia cancer diagnosis and for the identification of gene groups responsible for the disease. The same idea can be extended to analyzing other types of cancer data. As future work, we believe that much room is still there for new methods to combine compound-risk factors for classification. We also plan to explore the relationships between odds ratio patterns, emerging patterns, and other statistical patterns such as chi-squared patterns.

Appendix

A recent work epMiner [24] is designed for mining emerging patterns. Given two classes of data $\mathcal{D}_{pos}$ and $\mathcal{D}_{neg}$, the epMiner algorithm mines those minimal emerging patterns occurring frequently ($> \alpha\%$) in the positive class and infrequently ($< \beta\%$) in the negative class. This algorithm has been shown to outperform another recent algorithm proposed by [14]. Our experiment here is to study the efficiency of our algorithm, in comparison to the epMiner algorithm [24]. For a fair comparison, our algorithm takes the option of mining non-redundant minimal emerging patterns by setting $\pi = (|\mathcal{D}_{pos}| \cdot \alpha\%)/(|\mathcal{D}_{pos}| + |\mathcal{D}_{neg}|)$, and $\delta = |\mathcal{D}_{neg}| \cdot \beta\%$. Figure 4 shows the running time of the two algorithms. From this figure, we can see that our algorithm is constantly faster than epMiner with multiple orders of magnitudes on the two high-dimensional gene expression data sets when varying α and β. (The experiments on the ALL-AML data set was terminated when the time reached 50000 seconds; on the LungCancer data set was terminated when reached 10000 seconds; our running time included that of mining closed patterns.)

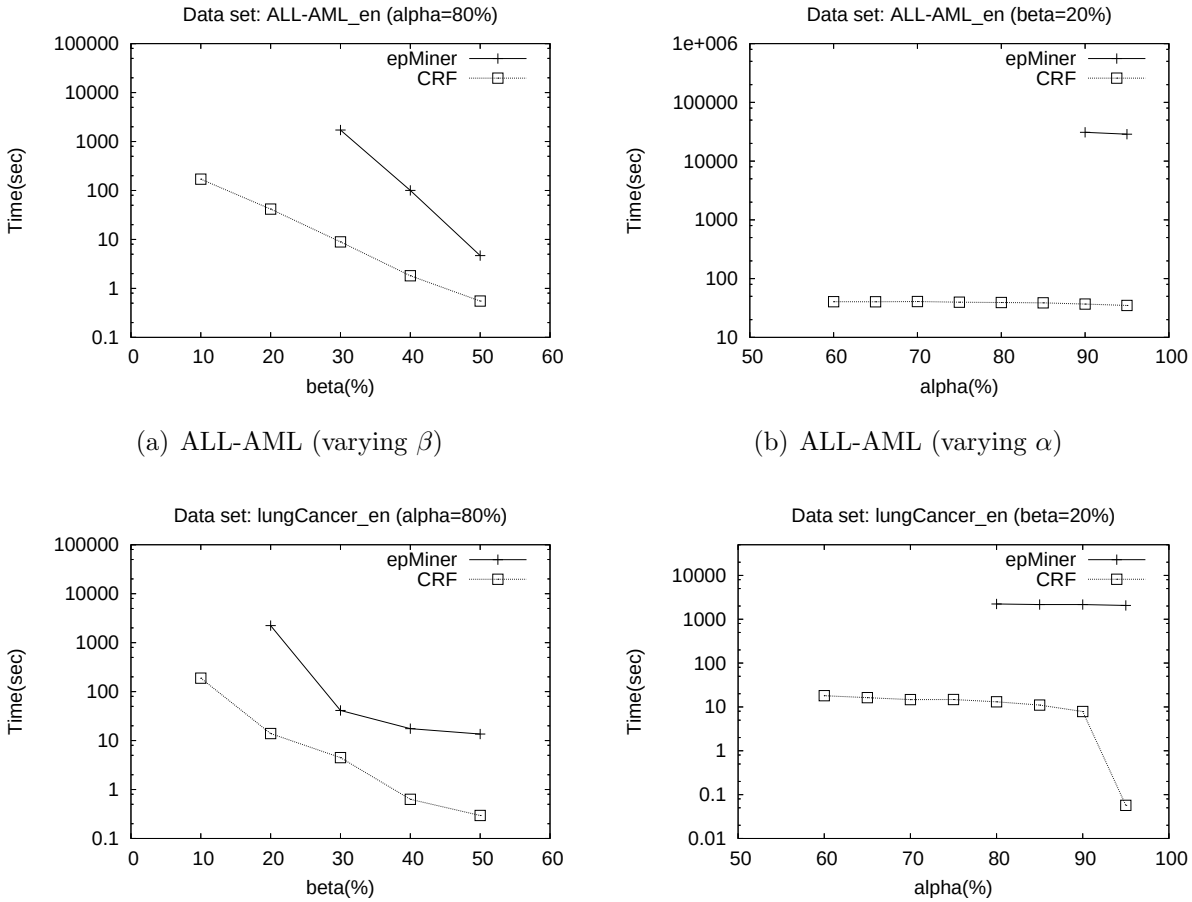
The ALL-AML and LungCancer data sets (available at `http://research.i2r.a-star.edu.sg/rp/`) contain gene expression levels of some leukemia and lung cancer patients. They are challenging, as seen in [24], because every transaction contains hundreds (865 in ALL-AML) or even thousands (2172 in LungCancer) of items. The experiments were conducted on a 3.60Ghz Pentium IV with 2GB memory running Fedora core. All codes were compiled using g++.

References

[1] Douglas G Altman, Jonathon J Deeks, and David L Sackett. Odds ratios should be avoided when events are common. *British Medical Journal*, 317:1318, 1998.

[2] Yves Bastide, Rafik Taouil, Nicolas Pasquier, Gerd Stumme, and Lotfi Lakhal. Mining frequent patterns with counting inference. *SIGKDD Explorations*, 2(2):66–75, 2000.

[3] Stephen D. Bay and Michael J. Pazzani. Detecting change in categorical data: Mining contrast sets. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge*



(a) ALL-AML (varying β)

(b) ALL-AML (varying α)

(c) Lung Cancer (varying β)

(d) Lung Cancer (varying α)

Figure 4: Time for mining emerging patterns by our method and epMiner [24]

Discovery and Data Mining, pages 302–306, 1999.

[4] Stephen D. Bay and Michael J. Pazzani. Detecting group differences: Mining contrast sets. *Data Mining and Knowledge Discovery*, 5:213–246, 2001.

[5] Anne-Laure Boulesteix, Gerhard Tutz, and Korbinian Strimmer. A CART-based approach to discover emerging patterns in microarray data. *Bioinformatics*, 19:2465–2472, 2003.

[6] Michael B Bracken and John C Sinclair. When can odds ratios mislead? avoidable systematic error in estimating treatment effects must not be tolerated. *British Medical Journal*, 317:1155–1157, 1998.

[7] Leo Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[8] Thomas D. Cook. Up with odds ratios! a case for odds ratios when outcomes are common. *Academic Emergency Medicine*, 9:1430–1434, 2002.

[9] C. Cortez and V. Vapnik. Support vector networks. *Machine Learning*, 20:273–279, 1995.

[10] Huw Talfryn Oakley Davies, Iain Kinloch Crombie, and Manouche Tavakoli. When can odds ratios mislead? *British Medical Journal (BMJ)*, 316:989–991, 1998.

[11] Jon Deeks. When can odds ratios mislead? odds ratios should be used only in case-control studies and logistic regression analyses (letter). *British Medical Journal*, 317:1155–1157, 1998.

[12] P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130, 1997.

[13] Guozhu Dong and Jinyan Li. Efficient mining of emerging patterns: Discovering trends and differences. In Surajit Chaudhuri and David Madigan, editors, *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 43–52, San Diego, CA, 1999. ACM Press.

[14] Hongjian Fan and Kotagiri Ramamohanarao. Fast discovery and the generalization of strong jumping emerging patterns for building compact and accurate classifiers. *IEEE Transaction on Knowledge and Data Engineering*, 18(6):721–737, 2006.

[15] Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In Lorenza Saitta, editor, *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 148–156, Bari, Italy, July 1996. Morgan Kaufmann.

[16] Joseph Hair, Bill Black, Barry Babin, Rolph Anderson, and Ronald Tatham. *Multivariate Data Analysis (6th Edition)*. Pearson Prentice Hall, 2005.

[17] Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

[18] S. Lahti, H. Hausen, and R. Kriinen. Patients'expectations of an ideal dentist and their views concerning the dentist they visited: Do the views conform to the expectations and what determines how well they conform? *Community Dent Oral Epidemiol*, 24:240–244, 1996.

[19] P. Langley, W. Iba, and K. Thompson. An analysis of Bayesian classifier. In William R. Swartout, editor, *Proceedings of the Tenth National Conference on Artificial Intelligence*, pages 223 – 228. AAAI Press, 1992.

[20] Jinyan Li, Haiquan Li, Limsoon Wong, Jian Pei, and Guozhu Dong. Minimum description length principle: Generators are preferable to closed patterns. In *The Twenty-First National Conference on Artificial Intelligence (AAAI-06)*, Boston, USA, 2006.

[21] Jinyan Li, Huiqing Liu, James R. Downing, Allen Eng-Juh Yeoh, and Limsoon Wong. Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients. *Bioinformatics*, 19:71 –78, 2003.

[22] Jinyan Li, Kotagiri Ramamohanarao, and Guozhu Dong. The space of jumping emerging patterns and its incremental maintenance algorithms. In *Proceedings of the Seventeenth International Conference on Machine Learning, Stanford, CA, USA*, pages 551–558, San Francisco, June 2000. Morgan Kaufmann.

[23] Jinyan Li and Limsoon Wong. Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns. *Bioinformatics*, 18:725–734, 2002.

[24] Elsa Loekito and James Bailey. Fast mining of high dimensional expressive contrast patterns using zero-suppressed binary decision diagrams. In *Proceedings of The Twelfth ACM SIGKDD International Conference On Knowledge Discovery and Data Mining (KDD 2006)*, pages 307–316, Philadelphia, USA, 2006.

[25] Paulo A. Lotufo, Claudia U. Chae, Umed A. Ajani, Charles H. Hennekens, and JoAnn E. Manson. Male pattern baldness and coronary heart disease the physicians' health study. *Archive of Internal Medicine*, 160:165–171, 2000.

[26] P.R.H. Newsome and G.H. Wright. A review of patient satisfaction—dental patient satisfaction: an appraisal of recent literature. *British Dental Journal*, 186:166–170, February 1999.

[27] D. Nuemark-Sztainer. The social environments of adolescents: Associations between socioenvironmental factors and health behaviors during adolescence. *Adolesc Med State Arts, Rev.*, 10:41–55, 1999.

[28] Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal. Discovering frequent closed itemsets for association rules. In *Proceedings of the 7th International Conference on Database Theory (ICDT)*, pages 398–416, 1999.

[29] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.

[30] Ron Rymon. Search through systematic set enumeration. In *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning*, pages 539–550, Cambridge MA, October 1992.

[31] David L Sackett, Jonathon J Deeks, and Douglas G Altman. Down with odds ratios! *Evidence-Based Medicine*, 1(9/10):164–166, 1996.

[32] Stephen Senn. Rare distinction and common fallacy [letter]. *e-British Medical Journal*, Website:`http://bmj.com/cgi/eleters/317/7168/1318`, 1999.

[33] P. Stege, S. Handelman, J. Baric, and M. Espeland. Satisfaction of the older patient with dental care. *Gerodontics*, 2:171–174, 1986.

[34] Lusheng Wang, Hao Zhao, Guozhu Dong, and Jianping Li. On the complexity of finding emerging patterns. *Theoretical Computer Science*, 335(1):15–27, 2005.

[35] Sylvia Wassertheil-Smoller. *Biostatistics and Epidemiology*. Springer Verlag, 2004.

[36] Geoff I. Webb, Shane Butler, and Douglas 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. ACM Publisher, 2003.

[37] K. M. Weiss. *Genetic Variation and Human Disease: Principles and Evolutionary Approaches*. Cambridge University Press, 1993.

[38] Eng-Juh Yeoh, Mary E. Ross, Sheila A. Shurtleff, W. Kent Williams, Divyen Patel, Rami Mahfouz, Fred G. Behm, Susana C. Raimondi, Mary V. Relling, Anami Patel, Cheng Cheng, Dario Campana, Dawn Wilkins, Xiaodong Zhou, Jinyan Li, Huiqing Liu, Ching-Hon Pui, William E. Evans, Clayton Naeve, Limsoon Wong, and James R. Downing. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1:133–143, 2002.

[39] Jun Zhang and Kai F. Yu. What's the relative risk? a method of correcting the odds ratio in cohort studies of common outcomes. *JAMA*, 280:1690–1691, 1998.