To appear in *Proceedings of The Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD-2001*), August 26-29, 2001, San Francisco, USA.

Discovering the Set of Fundamental Rule Changes

Bing Liu, Wynne Hsu, and Yiming Ma
School of Computing
National University of Singapore
3 Science Drive 2
Singapore 117543
{liub, whsu, maym}@comp.nus.edu.sg

ABSTRACT

The world around us changes constantly. Knowing what has changed is an important part of our lives. For businesses, recognizing changes is also crucial. It allows businesses to adapt themselves to the changing market needs. In this paper, we study changes of association rules from one time period to another. One approach is to compare the supports and/or confidences of each rule in the two time periods and report the differences. This technique, however, is too simplistic as it tends to report a huge number of rule changes, and many of them are, in fact, simply the snowball effect of a small subset of fundamental changes. Here, we present a technique to highlight the small subset of fundamental changes. A change is fundamental if it cannot be explained by some other changes. The proposed technique has been applied to a number of real-life datasets. Experiments results show that the number of rules whose changes are unexplainable is quite small (about 20% of the total number of changes discovered), and many of these unexplainable changes reflect some fundamental shifts in the application domain.

Keywords: Change mining, data mining.

1. INTRODUCTION

Detecting and adapting to changes are important activities for businesses and organizations. Companies and organizations want to know what are the changes in order to tailor their products and services to suit the changing needs of their customers. In many business environments, mining for changes can be more important than producing accurate models or classifiers for prediction, which has been the focus of existing data mining research. A model, no matter how accurate, in itself is passive because it can only predict based on patterns mined in the old data. It should not lead to actions that may change the environment because otherwise the model will cease to be accurate. Building models for prediction is more suitable in domains where the environment is relatively stable and there is little human intervention. However, in most business situations, constant human intervention is a fact of life. Companies simply cannot allow nature to take its course. They constantly need to perform actions in order to provide better products and services, to win more customers, and to resolve outstanding problems faced by the companies. For example, in a supermarket, there are always discounts and promotions to raise sale volume, to clear old stocks and to generate more sale traffic. Change mining is important in such situations as it allows the supermarket to compare results before and after promotions to see whether the promotions are effective, and to find interesting changes in customer behaviors.

In this paper, we study change mining in the context of association rules, i.e., to find rule changes that occur in the new time period as compared to the old time period. Association rule mining is defined as follows [2]: Let $I = \{i_1, \ldots, i_n\}$ be a set of *items*, and D be a set of transactions. Each transaction consists of a subset of items in I. An *association rule* is an implication of the form $X \rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \emptyset$. The rule $X \rightarrow Y$ holds in D with confidence c if c% of transactions in D that support X also support Y. The rule has support s in D if s% of transactions in D contains $X \cup Y$. The problem of mining association rules is to generate all association rules that have support and confidence greater than the user-specified minimum support (minsup) and minimum confidence (minconf).

In this work, we focus on association rules mined from a relational table, which consists of a set of tuples described by a number of attributes. An item in this context is an attribute value pair, i.e., (attribute = value) (numeric attributes are discretized). Mining in such data is typically targeted at a specific attribute, as the user normally wants to know how the other attributes are related to this *target* attribute (which has a number of discrete values) [3, 16]. With a target attribute, an association rule can be expressed as: $X \rightarrow y$, where y is an item (or a value) of the target attribute, and X is a set of items from the rest of the attributes.

On the surface, finding changes of association rules generated from data in two time periods is easy. We simply compare the supports and confidences of each rule from the two time periods, and then report the differences. This approach was taken by a number of researchers [e.g., 1, 5, 8]. However, it has two major problems. First, it often reports far too many changes and most of them are simply the snowball effect of some fundamental changes. Second, analyzing the difference in supports/confidences may miss some interesting changes. For example, rule $A,B \rightarrow y$ has similar supports in time period t_1 and t_2. However, the support of $A \rightarrow y$ and the support of $B \rightarrow y$ have both increased dramatically. This is an interesting phenomenon and should be reported.

A more useful approach is to identify the small set of fundamental changes (changes that cannot be explained by the presence of other changes) and report only the fundamental changes to the user. This makes sense because the fundamental changes are typically those that require user attention. This paper proposes a technique to identify the set of fundamental changes in two given datasets collected from two time periods. To the best of our knowledge, this is the first time that such a technique is proposed. We now use an example to illustrate the idea.

Example 1: Assume that the following three rules are mined from the data of time period 1.

r1: Job = yes → Loan = approved [sup=3.2%, conf = 60%]

r2: Own_house = yes → Loan = approved [sup=4.4%, conf = 40%]

r3 Job = yes, Own_house = yes → Loan = approved [sup=2.6%, conf = 80%]

In time period 2, the three rules become:

r1': Job = yes → Loan = approved [sup=5.0%, conf = 65%]

r2': Own_house = yes → Loan = approved [sup=5.3%, conf = 45%]

r3': Job = yes, Own_house = yes → Loan = approved [sup=4.2%, conf = 89%]

Focusing on the changes in support, we observe the following facts. First, the proportion of people who have jobs and have their loans approved increased significantly in time period 2. Second, the proportion of people who own houses and have their loans approved has also increased. Third, the proportion of people who have jobs, own houses and have their loan approved increased significantly as well. The question that we would like to ask is: "is the increase in r3 (we use r3' and r3 in time period 2 interchangeably) the consequence of the increases in r1 and r2?" That is, can the increase in support of r3 be explained in terms of the changes in support of r1 and r2? If it can be explained, then we say the change observed in r3 is not a fundamental change. Intuitively, we can see that in this example the support increase in r3 could be somehow explained by the support increases in r1 and r2. If, however, we have the following rule in time period 2 instead of r3',

r3": Job = yes, Own_house = yes → Loan = approved [sup= 2.2%, conf = 89%].

The decrease in support of r3 (r3") is hard to explain because we would expect the support of r3 to increase given that the supports of r1 and r2 have both increased. In this case, we say that the change observed in r3 is unexplainable, and represents a fundamental change.

This paper proposes a technique to identify all such fundamental changes – changes that cannot be explained by other changes. Such rules are very interesting because they often represent some major shifts in the domain. The proposed technique has been evaluated using 10 real-life datasets. Experiment results show that only a small subset of the rules exhibit fundamental changes in support and/or confidence. They can be manually analyzed without much difficulty. The user can then focus his/her attention on those important/interesting aspects of changes, and selectively view those less important or explainable changes.

2. RELATED WORK

Mining or learning in a changing environment has been studied in both data mining and machine learning. In data mining, [1, 5, 8] address the problem of monitoring the support and confidence of association rules. Given an association rule, their techniques track the support and confidence variations of the rule over time. These techniques basically belong to the simple approach mentioned in Section 1. None of them aims to find fundamental changes.

A representative work of these techniques is the one reported in [1]. The paper proposes to monitor rules in different time periods. The discovered rules from different time periods are collected into a rule base. Ups and downs in support or confidence over time (called history) are represented and defined using shape operators. The user can then query the rule base by specifying some history specifications. Clearly, this is different from our work, as it does not mine fundamental rule changes.

[8] presents a general framework for measuring changes or differences in two models (e.g., two sets of association rules from two datasets).The framework does not mine fundamental changes.

[4, 9] study the maintenance of discovered association rules. The techniques aim to incrementally update the rules when data tuples are deleted or inserted. They do not mine explicit changes.

[6] proposes a method for spatial trend detection. A spatial trend is defined as a regular change of one or more non-spatial attributes when moving away from a given start object. This is different from our work, as it is not concerned with rule changes.

[11] proposes a technique to compare high-dimensional datasets. It first partitions the data into sections. Each section is summarized with a profile, a set of statistics. The profiles from the old data set and the new dataset are then compared using statistic tests for proportions and for change of means. This work is also different from ours, as it is not concerned with rule changes.

Our proposed technique is in spirit related to rule summarization in [16] and finding the minimal set of unexpected rules in [18]. Both techniques, however, do not deal with changes.

Another related research is the subjective interestingness in data mining. [19, 14, 18] propose a number of approaches for finding unexpected rules with respect to the user's expectations. These techniques also do not deal with rule changes.

In machine learning, many researchers have studied how to produce good classifiers (or concepts) in on-line learning of a drifting environment [e.g., 20, 13]. Their framework is different from ours, as they do not mine explicit drifts.

3. BASIC STEPS OF THE PROPOSED APPROACH

Let the dataset from time period t_1 be D_1 and the dataset from time period t_2 be D_2. The technique consists of the following two steps:

1. **Rule generation**: We first perform association rule mining on each sub-dataset D_i. Let the set of rules mined from D_i be R_i. We define R, the set of rules that we will use to find the fundamental changes, as the union of the R_i's:
$$R = \{r \mid r \in (R_1 \cup R_2)\}$$
Clearly, a rule r in R may appear in R_i but not in R_j ($i \neq j$) because r may not satisfy the minsup and/or minconf in D_j. However, in order to analyze the changes in rules, we need the supports and confidences of every rule in R for both time periods. Thus, the missing support and confidence information in either time periods for each rule needs to be obtained. This can be done easily. We can modify an association rule mining algorithm in [2] slightly so that it can mine rules from both D_1 and D_2 at the same time to produce R taking into consideration of the required supports and confidences of each rule. Note that we also perform rule pruning to remove those insignificant rules. See the details on pruning in [16].
2. **Identification of fundamental rule changes:** This step finds all fundamental rule changes. Recall we define a change as fundamental if it cannot be explained by other changes. We

say a change is explainable if the observed relationship in t_1 still holds in t_2 taking into account any increase or decrease in support or confidence of some related changes.

The rest of the paper will focus on step 2. Step 1 will not be discussed any further as it is fairly straightforward.

4. FINDING FUNDAMENTAL RULE CHANGES

This section presents the proposed technique for finding fundamental rule changes. We distinguish two types of changes, namely, those identified through quantitative analysis, and those identified through qualitative analysis. Below, we first present these two types of fundamental changes and the techniques for identifying them, and then discuss two sub-sets of changes that are of particular interest.

4.1 Finding Fundamental Rule Changes – Quantitative Analysis

In quantitative analysis, we use a statistical test for homogeneity to identify those unexplainable rule changes (or fundamental rule changes). Basically, the change in support (or confidence) of a rule is unexplainable or fundamental if the change is unexpected.

Definition 1 (**expected supports or confidences**): The expected support (or confidence) of a rule r in t_2 is defined as follows:

(1) If r is a 1-condition rule, the expected support (or confidence) of r in t_2 is the support (or confidence) of r in t_1 (i.e., r's support or confidence is expected to remain the same from t_1 to t_2 in the absence of prior knowledge).

(2) If r is a k-condition rule r ($k > 1$) of the form,

$$r: \quad a_1, a_2, \ldots, a_k \rightarrow y$$

we view r as a combination of 2 rules, a 1-condition rule r_{one} and a (k-1)-condition rule r_{rest}, with the same target y:[1]

$$r_{one}: \quad a_i \rightarrow y \qquad r_{rest}: \quad a_1, a_2, \ldots, a_j \rightarrow y$$

where $\{a_1, a_2, \ldots, a_j\} = \{a_1, a_2, \ldots, a_k\} - \{a_i\}$. Note that there are k such combinations. Let $sup_t(x)$ and $conf_t(x)$ be the support and confidence of rule x in time period t respectively. Let $E_{r_{one}}(sup_{t_2}(r))$ and $E_{r_{rest}}(sup_{t_2}(r))$ be the expected supports of r in t_2 with respect to r_{one} and r_{rest}. Let $E_{r_{one}}(conf_{t_2}(r))$ and $E_{r_{rest}}(conf_{t_2}(r))$ be the expected confidences of r in t_2 with respect to r_{one} and r_{rest}.

The expected supports and the expected confidences of r with respect to r_{one} and r_{rest} are computed as follows:

$$E_{r_{one}}(sup_{t_2}(r)) = \min\left(\frac{sup_{t_1}(r)}{sup_{t_1}(r_{one})} \times sup_{t_2}(r_{one}), 1\right)$$

$$E_{r_{rest}}(sup_{t_2}(r)) = \min\left(\frac{sup_{t_1}(r)}{sup_{t_1}(r_{rest})} \times sup_{t_2}(r_{rest}), 1\right)$$

$$E_{r_{one}}(conf_{t_2}(r)) = \min\left(\frac{conf_{t_1}(r)}{conf_{t_1}(r_{one})} \times conf_{t_2}(r_{one}), 1\right)$$

$$E_{r_{rest}}(conf_{t_2}(r)) = \min\left(\frac{conf_{t_1}(r)}{conf_{t_1}(r_{rest})} \times conf_{t_2}(r_{rest}), 1\right)$$

The computation of the expected supports and confidences are based on the Constant Proportion Assumption.

Constant Proportion Assumption[2]: The proportional relationships between r_{one} and r_{rest} to rule r should remain the same from time period t_1 to time period t_2, where $t_2 > t_1$.

This assumption is reasonable as it follows our intuition and allows us to find those unexpected or unexplainable changes that are important in practice.

Definition 2 (**fundamental rule change in support or confidence via quantitative analysis**): A change in support (or confidence) of rule r from t_1 to t_2 is said to be fundamental if:

(1) r is a 1-condition rule and its support (or confidence) is *significantly* different from its expected support (or confidence), or

(2) r is a k-condition rule r ($k > 1$), and for *all* r_{one} and r_{rest} combinations, $E_{r_{one}}(sup_{t_2}(r))$, $E_{r_{rest}}(sup_{t_2}(r))$ and $sup_{t_2}(r)$ are *significantly* different (or $E_{r_{one}}(conf_{t_2}(r))$, $E_{r_{rest}}(conf_{t_2}(r))$ and $conf_{t_2}(r)$ are *significantly* different).

Note that (2) of Definition 2 basically says that if the change in r can be explained by any one combination, it is not a fundamental rule change. This is reasonable because as long as there is one possible explanation for r's change, we cannot say that r is a fundamental change. Notice also that in this definition, we need a significance test. Here, we use the popular Chi-square test statistics for the purpose (see Section 5).

Let us use an example to illustrate the above definitions.

Example 2: We have the following rules in time period 1:

r1: Job = yes → Loan = approved [sup=5.2%, conf = 60%]
r2: Own_house = yes → Loan = approved [sup=6.0%, conf = 40%]
r3 Job = yes, Own_house = yes → Loan = approved [sup=2.1%, conf = 80%]

In time period 2, the three rules become:

r1': Job = yes → Loan = approved [sup=4.4%, conf = 65%]
r2': Own_house = yes → Loan = approved [sup=4.3%, conf = 45%]
r3': Job = yes, Own_house = yes → Loan = approved [sup=4.2%, conf = 89%]

Since r1 and r2 are 1-condition rules, we can use Chi-square test to check whether the support (or confidence) of each of them in t_2 is significantly different from that in t_1. If so, it is a fundamental rule change. We now focus on r3 (r3') in t_2. Here, we only evaluate the change in support (change in confidence can be evaluated in the same way). The expected supports of r3 in t_2 with respect to r1 and r2 are 0.018 and 0.015 respectively.

Now with these two expected supports of r3 in t_2, we can use Chi-square test for homogeneity to check whether $E_{r1}(sup_{t_2}(\text{r3}))$, $E_{r2}(sup_{t_2}(\text{r3}))$, and the actual support of r3 in t_2 (i.e., $sup_{t_2}(\text{r3})$) are significantly different. Suppose it turns out that they are statistically different. Then, the support of r3 in t_2 is unexplainable, and it is a fundamental rule change in support.

Figure 1 gives the algorithm for finding fundamental rule changes. The algorithm only shows the rule support evaluation. Confidence evaluation can be done in exactly the same way.

---

[1] We can also view the rule r as a combination of more than two shorter rules. The numbers of conditions in the shorter rules do not have to be 1 and k-1, but any possible partition. However, these alternatives require more computation. Conceptually, they are also more difficult to understand.

[2] We experimented with a number of other assumptions (ways to compute expectations), but their results were less satisfactory. They were also more difficult to understand.

	Case I	Case II	Case III	Case IV	Case V	Case VI	Case VII
r_{one}:	increase	drop	noChange	increase	noChange	drop	noChange
r_{rest}:	increase	drop	noChange	noChange	increase	noChange	drop
r:	increase	drop	noChange	increase	increase	drop	drop

Figure 2: 7 explainable change combinations – qualitative analysis

	Explainable		Fundamental change					
	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8
r_{one}	increase	drop	increase	drop	increase	increase	Drop	drop
r_{rest}	increase	drop	increase	drop	drop	drop	increase	increase
r	increase	drop	drop	increase	increase	drop	increase	drop

Figure 3: Various cases of explainable and fundamental changes – qualitative analysis

Algorithm findFdmChanges(R) /* R is the set of input rules */
```
1   FdmChanges = ∅;
2   for each r in R from the shortest rule to the longest rule do
3       for each pair, r_one (X_i → y) and r_rest (X_rest → y) of r,
                        where X_i ∈ X, and X_rest = X − {X_i} do
4           compute E_r_one(sup_t2(r)), and E_r_rest(sup_t2(r));
5           flag = homoTest(E_r_one(sup_t2(r)), E_r_rest(sup_t2(r)), sup_t2(r));
6           if flag = true then
                /* E_r_one(sup_t2(r)), E_r_rest(sup_t2(r)) and sup_t2(r) are homogenous */
7               go-to line 2   /* r is explainable, and we go to test the next rule */
8           endif
9       endfor
10      if flag = false then /* r's support has changed fundamentally */
11          FdmChanges = FdmChanges ∪ {r}
12      endif
13  endfor
```

Figure 1. Finding fundamental rule changes - quantitative analysis

Line 1 initializes the variable *FdmChanges* for storing the set of fundamental rule changes. Line 2 sets the loop to start evaluation of each rule from the shortest rule to the longest rule. Line 3 produces every r_{one} and r_{rest} pair from r. Lines 4 & 5 compute $E_{r_{one}}(sup_{t_2}(r))$ and $E_{r_{rest}}(sup_{t_2}(r))$, and perform the homogeneity test (see Section 5) to determine whether the support of r in t_2 can be explained by r_{one} and r_{rest}. In lines 6-12, if r can be explained by any of the r_{one} and r_{rest} pairs, it is not a fundamental change (lines 6 & 7). Those fundamental rule changes (line 11) are stored in *FdmChanges*. In lines 6-12, if r can be explained by any of the r_{one} and r_{rest} pairs, it is not a fundamental change (lines 6 & 7). Fundamental rule changes (line 11) are stored in *FdmChanges*.

Complexity: The time complexity of the algorithm is linear in the number of discovered rules. See [15] for details.

4.2 Finding Fundamental Rule Changes – Qualitative Analysis

We now identify fundamental changes via qualitative analysis. Here, the magnitude of change is ignored. Instead we only focus on the direction of change, i.e., *increase*, *drop* or *noChange* from t_1 to t_2. This set of rules is less interesting, but it enables us to find two sub-sets of interesting rules, as we will see later.

Definition 3 (**fundamental rule change in support/confidence via qualitative analysis**): The support (or confidence) change in rule r from t_1 to t_2, where r is a k-condition rule ($k > 1$) [3], is said to be a fundamental support (or confidence) change if for ***all*** r_{one} and r_{rest} combinations, the directions of support (or confidence) changes of r_{one}, r_{rest} and r in time period 2 do not belong to any of the 7 cases in Figure 2.

The algorithm for finding fundamental rule changes in this case can be obtained easily by modifying the algorithm in Figure 1.

4.3 Some Interesting Subsets of Fundamental Rule Changes

Note that Definition 2 and Definition 3 may result in different sets of fundamental rule changes. They are both useful in practice. We now discuss some interesting situations. We produce all the cases of Definition 3. To simplify our discussion, we omit *noChange*, as it is unlikely to occur in practice (but it can be easily included). Without *noChange*, the total number of possible cases is reduced to 8 (see Figure 3). Note that Case 7 and Case 8 are symmetric to Case 5 and 6 respectively with respect to r_{one} and r_{rest}. We use support change in our following discussion.

Qualitatively, we observe that r's increase or drop in support in Case 1 and Case 2 are intuitively sound. If two things increase or drop, we would expect their combination to increase or drop too. However, Cases 3, 4, 5, 6, 7 and 8 are hard to explain.

Quantitative analysis, however, could conform to or conflict with qualitative analysis. For instance, if a particular rule belongs to Case 1 in Figure 3, it is thus explainable qualitatively. However, if we take into consideration the magnitude of changes, it becomes non-trivial. If the supports of r_{one} and r_{rest} increase a little and the support of r also increases a little, then quantitative analysis may also show the rule is explainable. However, if the actual support of r increases drastically, it may suggest that r is a fundamental change, as the drastic increase cannot be explained by the small increases in r_{one} and r_{rest}.

This observation leads us to identify some interesting subsets of fundamental rule changes. Let the set of fundamental rule changes obtained from quantitative analysis be F_{qt}, and the set of fundamental rule changes obtained from qualitative analysis be F_{ql}, i.e., Cases 3-8. Two interesting subsets of fundamental rule changes can be defined (these two subsets partition F_{qt}):

1. $\boldsymbol{F_{qt} - F_{ql}}$**:** These rules are either 1-condition fundamental change rules, or fall in Case 1 or Case 2 (i.e., their directions of change are the same, but their magnitudes of change cannot be explained). For example, for a particular rule r, its $sup_{t_2}(r_{one})$ increases to 5% from $sup_{t_1}(r_{one})$ = 4.8% and $sup_{t_2}(r_{rest})$ increases to 6% from $sup_{t_1}(r_{rest})$ = 5.6%. However, $sup_{t_2}(r)$ increases to 4% from $sup_{t_1}(r)$ = 1.5%. This situation falls into Case 1 above. However, $sup_{t_1}(r)$ = 4% is hard to explain as its increase is too

[3] For 1-conditional rules, qualitative analysis does not apply. They are evaluated by quantitative analysis.

drastic. These rules are interesting because they represent some relationship of disproportional increases. In applications, if we want to achieve a big increase in r, we only need to perform some actions to generate a small increase in r_{one} and r_{rest}.

2. $F_{qt} \cap F_{ql}$**:** These are rules (with 2 or more conditions) that have changed dramatically, both qualitatively and quantitatively. They belong to Cases 3-8 and their magnitudes of change are also significant. These rules are most unexplainable, and they warrant further analysis.

5. CHI-SQUARE TEST

For the proposed technique, we need a statistical test for homogeneity of a set of proportions (support and confidence are both proportions). Chi-square (χ^2) test [17] is a popular choice.

Test of homogeneity: A test of homogeneity involves testing the null hypothesis (H_0) that the proportions, $p_1, p_2, \ldots, p_k$, in two or more populations are the same against the alternative hypothesis (H_1) that these proportions are not the same. That is,

H_0: $p_1 = p_2 = \ldots = p_k$

H_1: the population proportions are not all equal.

We assume that the data consists of independent random samples of size $n_1, n_2, \ldots, n_k$ from k populations. The data is arranged in a $2\times k$ contingency table (Figure 4). The numbers $x_1, \ldots, x_k, n_1\text{-}x_1, \ldots, n_k\text{-}x_k$ listed inside the $2k$ cells are called *observed frequencies* of the respective cells.

	1	2	…	k
Successes	x_1	x_2	…	x_k
Failures	$n_1\text{-}x_1$	$n_1\text{-}x_2$	…	$n_1\text{-}x_k$

Figure 4: A $2\times k$ contingency table

Let O be an observed frequency, and E be an expected frequency for a cell in the above table. The statistic defined as

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

has a χ^2 distribution with: $df = (Row - 1)(Column - 1)$, degrees of freedom, where *Row* and *Column* are the number of rows and the number of columns, respectively, in the given contingency table.

Under the null hypothesis in a test of homogeneity, we would expect the frequency E (expected frequency) for a cell to be as follows [17]: (Row total)×(Column total)/Sample size.

Thus, χ^2 represents a normalized deviation from expectation. It works as follows: if all values were really homogeneous, the χ^2 value would be 0. If it is higher than a threshold value (5.99, at the 95% significant level with two degrees of freedom) we reject H_0, and state that the population proportions are significantly different. Let us see an example.

Example 3: We use the three rules in Example 2. Assume the dataset in time period 2 has 1000 tuples. Here, only support is used as an example. The expected supports of r3 (r3') in time period 2 with respect to r1 and r2 are:

$E_{r1}(sup_{t_2}(r3)) = 0.018$, and $E_{r2}(sup_{t_2}(r3)) = 0.015$.

We want to test the null hypothesis that $E_{r_1}(sup_{t_2}(r3))$, $E_{r_2}(sup_{t_2}(r3))$, and $sup_{t_2}(r3)$ are the same statistically using the significance level of 95% (a commonly used level [17]).

All the support information can be presented in a 2x3 contingency table (Figure 5) containing 6 cells. The expected frequencies are included in parentheses next to the observed frequencies within the corresponding cells.

	E_{r_1}	E_{r_2}	$r3$	Col. Total
satisfy r3	**18** (25)	**15** (25)	**42** (25)	75
Do not satisfy r3	**982** (975)	**985** (975)	**958** (975)	2925
Row total	1000	1000	1000	

Figure 5: The 2x3 contingency table of supports

For the values in the table of Figure 5, the test statistic χ^2 can be computed [4]. The observed χ^2 value is 18 (as computed). If we use the significance level of 95%, the critical value for χ^2 is 5.99 with 2 degree of freedom (df = (2-1)(3-1) = 2). The observed χ^2 value is much larger than the critical value. Thus, we reject the null hypothesis, and conclude that the supports are significantly different. We say that r3 shows an fundamental change in support.

6. EMPIRICAL EVALUATION

We experimented the proposed technique on 3 real-life datasets from three different domains. These datasets were collected by our user organizations over a number of years. The first dataset is from the education domain. We have 4 years of data. The second set of data is from the insurance domain. We also have 4 years of data. The last set is from the medical domain. We have 5 years of patients' screening data. We group all the datasets into pair groups (each pair group consists of data from two successive years from their respective domains). This gives us 10 pair groups of data. We mine the fundamental changes from each pair group.

Table 1 shows the summary of our experiment results using quantitative analysis. Each column is explained below (the average value for each column is shown in the last row):

Column 1 gives the name of each data group (with two years of data). Column 2 gives the total number of rules in R from the two years (see Section 3 for the definition of R) that satisfy the minimum support and minimum confidence requirements. Column 3 gives the number of rules after pruning for each data group. Pruning removes those non-significant rules. We apply the pruning technique in [16]. Column 4 gives the number of explainable confidence rules for each group, and column 5 gives the number of significant rules that exhibit fundamental changes in confidence through the quantitative analysis in Section 4.1. We observe that the number of fundamental rule changes is quite small. It is possible to manually analyze them. Column 6 shows the ratio of column 5 vs. column 3. On average, the number of rules that exhibit fundamental changes in confidence is only 23% of the total number of rules. It is also interesting to note that the education domain is very stable as the number of fundamental rule changes is consistently small over the years. The insurance domain is the most volatile. Columns 7-9 show the same set of results as columns 4-6, but for rule support changes. We can see that only 34% of the rules exhibit fundamental changes. Column 10 gives the number of fundamental rule changes in both support and confidence. Column 11 shows the rule generation time for each data group. Column 12 gives the time used for finding fundamental rule changes, i.e., the proposed technique in this paper. It includes the time for both confidence and support change evaluations through quantitative analysis and for finding (F_{qt}–F_{ql}) and ($F_{qt} \cap F_{ql}$). We can see that these operations can be done extremely efficiently. Columns 13 and 14 give the number of

[4] See [17, 15] for the assumptions made by the chi-square distribution, and for how to deal with them when the assumptions are not satisfied.

Table 1: Main experiment results (quantitative analysis, significance level for $\chi^2 = 95\%$)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
	Dataset	rules	sig. rules	Confidence			Support			conf & sup	Execution time			
				expl.	fund.	ratio (%)	expl.	funda.	ratio (%)	fund.	rule gen.	find fund.	t_2 data size	t_1 data size
1	EDU1	27626	534	478	56	10.49%	500	34	6.37%	31	3.384	0.058	3556	3472
2	EDU2	29572	565	515	50	8.85%	549	16	2.83%	13	3.397	0.062	3430	3556
3	EDU3	29525	625	565	60	9.60%	598	27	4.32%	20	3.451	0.059	3794	3430
4	INSUR1	2595	570	410	160	28.07%	277	293	51.40%	135	17.038	0.074	40246	39141
5	INSUR2	2022	569	282	287	50.44%	282	287	50.44%	200	17.031	0.072	45018	40246
6	INSUR3	1694	566	370	196	34.63%	354	212	37.46%	152	17.078	0.069	33746	45018
7	MED1	5489	264	173	91	34.47%	63	201	76.14%	81	20.348	0.059	19115	10507
8	MED2	6258	274	179	95	34.67%	149	125	45.62%	69	20.328	0.061	19743	19115
9	MED3	5569	298	256	42	14.09%	190	108	36.24%	23	20.357	0.062	14618	19743
10	MED4	6523	291	271	20	6.87%	206	85	29.21%	12	20.361	0.057	11017	14618
	Avg.	11687	456	350	105	23.22%	317	139	34.00%	73.6	14.276	0.063	19428.3	19884.6

tuples in time period 2 and 1 respectively. All our experiments were conducted on a PII-350 PC with 128MB RAM.

Table 2 gives the results of the two interesting subsets of fundamental rule changes, i.e., $F_{qt}–F_{ql}$ and $F_{qt} \cap F_{ql}$. The table has two parts. The first part (columns 2-4) gives the results for fundamental confidence changes. The second part (columns 5-7) gives the results for fundamental support changes.

Table 2: Experiment results of $F_{qt}–F_{ql}$ and $F_{qt} \cap F_{ql}$

	1	2	3	4	5	6	7
	Dataset	Confidence			Support		
		fund.	$F_{qt}–F_{ql}$ (cases 1-2)	$F_{qt} \cap F_{ql}$ (cases 3-8)	fund.	$F_{qt}–F_{ql}$ (cases 1-2)	$F_{qt} \cap F_{ql}$ (cases 3-8)
1	EDU1	56	29	27	34	26	8
2	EDU2	50	31	19	16	11	5
3	EDU3	60	29	31	27	21	6
4	INSUR1	160	65	95	293	166	127
5	INSUR2	287	216	71	287	176	111
6	INSUR3	196	95	101	212	82	130
7	MED1	91	63	28	201	119	82
8	MED2	95	50	45	125	68	57
9	MED3	42	12	30	108	59	49
10	MED4	20	6	14	85	43	42
	Avg.	105	60	46	139	77.1	61.7

7. CONCLUSION

Detecting fundamental shifts in the environment is of crucial importance to businesses. Companies constantly need to know what are changing in the market place. This allows them to produce the right products and/or services to suit the changing market needs. Fundamental changes often require immediate attention and action. In this paper, we studied this issue in the context of association rules, and proposed an efficient technique to identify fundamental rule changes in support and in confidence. Empirical evaluation showed that only a small percentage of rules exhibit fundamental changes. This enables the users to analyze them to obtain those actionable changes without much difficulty.

8. REFERENCES

[1]. Agrawal, R. and Psaila, G. "Active data mining." *KDD-95*.

[2]. Agrawal, R. and Srikant, R. “Fast algorithms for mining association rules.” *VLDB-94*, 1994.

[3]. Bayardo, R., Agrawal, R, and Gunopulos, D. "Constraint-based rule mining in large, dense databases." *ICDE-99*, 1999.

[4]. Cheung, D. Han, J, V. Ng, and Wong, C.Y. “Maintenance of discovered association rules in large databases: an incremental updating technique.” *ICDE-96*, 1996.

[5]. Dong, G. and J. Li. “Efficient mining of emerging patterns: discovering trends and differences.” *KDD-99*, 1999.

[6]. Ester, M., Frommelt, A., Kriegel, H-P., and Sander, J. "Algorithms for characterization and trend detection in spatial databases." *KDD-98*, 1998.

[7]. Fayyad, U. M. and Irani, K. B. “Multi-interval discretization of continuous-valued attributes for classification learning.” *IJCAI-93*, 1993.

[8]. Ganti, V., Gehrke, J., and Ramakrishnan, R. “A framework for measuring changes in data characteristics.” *POPS-99*, 1999.

[9]. Ganti, V., Gehrke, J., and Ramakrishnan, R. “DEMON: mining and monitoring evolving data.” *ICDE-2000*, 2000.

[10]. Han, J. and Fu, Y. “Discovery of multiple-level association rules from large databases.” *VLDB-95*, 1995.

[11]. Johnson T. and Dasu, T. "Comparing massive high-dimensional data sets," *KDD-98*, 1998.

[12]. Kohavi, R., John, G., Long, R., Manley, D., and Pfleger, K. “MLC++: a machine learning library in C++.” *Tools with artificial intelligence*, 740-743, 1994.

[13]. Lane, T. and Brodley, C. "Approaches to online learning and concept drift for user identification in computer security." *KDD-98*, 1998.

[14]. Liu, B., Hsu, W., and Chen, S. “Using general impressions to analyze discovered classification rules.” *KDD-97*, 1997.

[15]. Liu, B., Hsu, W. and Ma, Y. *Discovering the set of fundamental rule changes*. NUS Technical Report, 2001.

[16]. Liu, B., Hsu, W. and Ma, Y. “Pruning and summarizing the discovered associations." *KDD-99*, 1999.

[17]. Mann, P. *Introductory statistics*. John Wiley & Sons, 1998.

[18]. Padmanabhan, B., and Tuzhilin, A. “Small is beautiful: Discovering the minimal set of unexpected patterns.” *KDD-2000*, 2000.

[19]. Silberschatz, A., and Tuzhilin, A. “What makes patterns interesting in knowledge discovery systems.” *IEEE Trans. on Know. and Data Eng.* 8(6), 1996.

[20]. Widmer, G. "Learning in the presence of concept drift and hidden contexts." *Machine learning*, 23:69-101, 1996.