# Mining Discrimination Patterns along Temporal Databases

André Magalhães and Paulo J. Azevedo

HASLab/ INESC TEC,
Departamento de Informática,
Universidade do Minho, Portugal
afgmagalhaes@gmail.com pja@di.uminho.pt

**Abstract.** In certain Data Analysis tasks, understanding the underlying differences between groups or classes is of the utmost importance. Contrast Set Mining relies on discovering significant patterns by contrasting two or more groups. A Contrast Set is a conjunction of attribute-value pairs that differ meaningfully in its distribution across groups. One technique proposed is Rules for Contrast Sets (RCS) which seeks to express each Contrast Set found in terms of rules. The main purpose of this work is to extend RCS to a Temporal Data Mining task. We define a set of temporal patterns in order to capture the significant changes in the contrasts discovered along the considered time line. To ascertain the proposal accuracy and ability to discover relevant information, two different real-life datasets were studied using this approach.

## 1 Introduction

Understanding the difference between contrasting groups is a fundamental Data Mining Task [4, 10]. This task can be used in many different domains. For example: census data collected this year can be compared to the data collected in a previous census activity, contrasting the data collected this year against the one collected thirty years ago. This comparison involves two groups (2011 vs 1981) and in this scenario it is fairly easy to infer some differences among these two groups in contrast: the mean number of children per couple should be lower nowadays but the income and education likely follow the reverse trend. This notion can easily be extrapolated to other domains.

Although Association Rule Mining captures the relations between items present in the data, it does not discriminate in regard to difference towards those same items. Even so, one proposal has shown that a commercial Association Rule Learner (Magnum_OPUS) with some tweaks could achieve this task fairly well [10]. Due to this latent inability, some techniques derived from Association Rule Mining have been proposed to tackle this problem. Contrast Set Mining (CSM) [4, 7, 3] has emerged as a data mining task whose goal is to effectively collect *contrast sets*, a formalism used to represent group differences. Rules for Contrast Sets (RCS) is a proposal that redesigns an association rule engine to derive rules that describe Contrast Sets [3].

By contrasting two or more groups, the aim is to obtain the attributes that distinguish them. Some proposals have been made in order to perform this task. However, none did consider further in order to contrast and differentiate groups along a time line. This permits to verify the contrast modification at different points in time.

Two certain groups being contrasted in a certain point in time could have just a few distinguishable features. Nothing guarantees that the relation between them has suffered a meaningful modification somewhere in another period either on the past or the future for a certain attribute or set of attributes. This proposal pretends, essentially, to look up on this matter. It tries to understand and identify the contrasts evolution along the defined periods, bridging together the areas of CSM and Temporal Data Mining (TDM).

The main contribution of this work is a proposal to represent group difference in a temporal database. In order to accomplish the objective of contrasting in a timely manner, we propose a set of temporal patterns. This set of patterns will allow to detect and represent situations of interest that mark a significant change in the contrasting behavior. Potentially, this can be considered highly valuable information by the end user.

The paper is organized as follows: section 2 briefly surveys Contrast Set Mining. The proposal with the patterns developed as well as the whole strategy to obtain and analyze them is described in section 3. Next section, presents the evaluation and application of the technique in two distinct case studies. Finally, conclusions are drawn regarding the work developed.

## 2 Contrast Set Mining

Contrast Set mining was first referred by Bay and Pazzani [4], as the problem of finding all contrast sets whose support differ meaningfully across groups. Association Rule Mining usually deals with market-basket data but for this problem the data model is grouped categorical data. The *itemset* concept present in Association Rules can be extended to contrast sets as defined by [4]:

**Definition 1.** *Let $A_1, A_2, ..., A_k$ be a set of k variables called attributes. Each $A_i$ can take on values from the set $\{V_{i1}, V_{i2}, ..., V_{im}\}$. Then a* ***contrast-set*** *is a conjunction of attribute-value pairs defined on groups $G_1, G_2, ..., G_n$.*

*Example: (sex = male) ∧ (occupation = manager)*

In this context, the support is considered in regard to the group and not to the whole dataset, meaning that the support of a contrast set *cs* is the percentage of examples in group $G$ where the contrast set is true.

Formally, the objective is to find all the contrast sets ($cs$) that meet the following criteria:

$$\exists ij \; P(cs|G_i) \neq P(cs|G_j) \tag{1}$$

$$max_{i,j}|sup(cs, G_i) - sup(cs, G_j)| \geq \delta \tag{2}$$

where $\delta$ is a user-defined threshold named *minimum support difference*. These two equations albeit different represent the same goal, finding contrast sets whose support differ meaningfully across groups. Equation 1 guarantees that the contrast set represent a true difference between at least a pair of groups (i.e. the basis of a statistical test of meaningful) and equation 2 ensures that only contrast sets whose difference is big enough to be considered relevant are obtained. The contrast sets that Eq.1 is statistically valid are called *significant* and those that met Eq.2 are referred as *large*. If both criteria is met, they are considered as *deviations*.

### 2.1 STUCCO

Presented in the first paper that introduced Contrast Set Mining [4], STUCCO (Search and Testing for Understandable Consistent Contrasts) is still widely used for mining contrast sets. It is based on Max-Miner [6] rule-discovery algorithm and uses a breadth-first search framework.

In order to check for *significant* contrast sets (Eq. 1) a statistical test is required. The null hypothesis to be tested is: *contrast set support is equal across all groups*. The support counts needed for this are organized in a $2 \times G$ contingency table where the row variable represent the truth of the contrast set and the columns represent each group considered. STUCCO uses a standard test for testing independence of variables in a contingency table, the *chi-square* test.

A test $\alpha$ level has to be selected in order to check if the differences are significant. This sets the maximum probability of falsely rejecting the null hypothesis for each test. In a case where multiple tests have to be applied, this probability quickly rises. Incorrectly rejecting the null hypothesis (concluding that a difference exists when it does not is known as a *Type I error* or *false positive*. To reduce the chance of obtaining a false positive, STUCCO uses a specific Bonferroni adjustment which reduces the probability of false discoveries at lower levels but it also decreases the number of contrast sets at these levels (those with a significant number of items).

Regarding pruning, STUCCO prunes away all nodes that are not *deviations*. Nodes of the search tree are pruned based on some criteria [4, 5] when there is a guarantee that a node and its own subtree won't contribute for finding deviations, for this reason they do not need to be visited further.

After determining which contrast sets are interesting, they are presented to the user in the following form:

```
hours_per_week = ]20.6:40.2]
2880 857 161 | 0.537815 0.497388 0.389831
============================
d.f.    chi^2    pvalue
2       38.37    4.65e-09
============================
```

This is a contrast set with just one item (hours per week) in a domain with 3 groups. In the second line there is the absolute and relative values of support within each group and below the statistical values like degrees of freedom, $\chi^2$

statistic value and its p-value. This representation has an intrinsic flaw because it does not show in which combination of groups there is a significant difference in support.

### 2.2 Rules for Contrast Sets

Rules for Contrast Sets (RCS) [3] is a proposal that makes uses of an existing association rule engine [2] redesigned to mine contrast sets that are expressed in form of rules. Rules are known by their ease of interpretation and expressive power making this representation easier to read than the one STUCCO adopted. Like a frequent itemset algorithm, search space traversal is performed in a depth-first manner contrasting to other proposals like STUCCO and CIGAR [7] that do it in a breadth-first manner. This type of traversal does not fully exploit the downward closure property of support [1] but still leads to an efficient rule-based algorithm [3].

Contrast Sets mined by this algorithm have to meet Eq.1 and 2. Although not specifically introduced as a equation, a minimum support criteria is also used here.

This implementation, like CIGAR, uses $2 \times 2$ contingency tables. This allows to perceive between which exactly groups the differences are significant but unlike STUCCO a $\chi^2$ test is replaced by a Fisher-exact test that is directional (one-sided) to determine if the frequencies observed are significant.

False discoveries are controlled differently than STUCCO. The layered critical values [9] proposal is adapted for this context. The adjustment used in STUCCO [5] considers the number of hypothesis evaluated rather than the size of the search space. That is, only accounts for candidate patterns that met a set of constraints. However, candidates are the patterns most likely to pass the statistical test. Thus, the critical value should be adjusted by the number of patterns from which those to be tested are selected instead of the number of times the statistical test is applied.

Rules that describe the contrast sets whose support differ across groups are formally organized as follows: $G_1 >> G_2, ..., G_i >> G_j \leftarrow cs$, where $cs$ represents the contrast set and $G_i$ each group. The pairs $G_i >> G_j$ indicate the direction to where the support differs ($G_i$ has bigger support than $G_j$). Consider the following example:

```
Gsup = 0.17191 | 0.04121  p = 1.1110878451E-017 education=Doctorate >> education=Masters
Gsup = 0.17191 | 0.01681  p = 3.0718399575E-040 education=Doctorate >> education=Bachelors
Sup(CS) =  0.03097                                          <---   workclass=State-gov  &
                                                                   class > 50K.
```

The rule is to be read as: *the occurrence of the contrast set "working for the state government and making an income of more than 50K" is significantly larger within people holding a PhD than a MSc. The same occurs between PhD and BSc holders.* Gsup refers to the support of the contrast set in the group (for example, 17,91% of the PhD holders are State-Governers and have a salary superior to 50,000) and Sup(CS) is the support of the contrast set in the entire

database. The p-value of the Fisher-exact test is also shown. This approach is much more readable than STUCCO output because of the rule format and the way envolved groups and direction in which difference between those groups occurs are described.

## 3 Proposal

The method can be summed up in a 3-step process that occur in a serialized manner. The output that is produced at each stage serves as input for the next step.

RCS is the first operator in the chain. It is the algorithm included in CAREN [2] for discovering Contrast Sets in any given dataset. This will be used in order to obtain the contrasts at each observation (single period of time considered). Having individual datasets for each period, it will be executed as many times as the number of periods. For each execution, CSV output file that contains all the contrasts found and related information like group support values, *p-values*, among others will be produced.

Post-Processing Contrast Sets (PPCS) was developed in order to process the set of output files produced by RCS at each period and to yield the temporal patterns that occur in them. An additional file is produced in order to be used by the *PPCS Viewer*.

The Viewer emerges as an optional but recommended manner to interpret the output given in the last step. Due to fact that there is some inability to interpret the results given in a textual format (at least not in an easy and intuitive manner), a graphical tool ( *Viewer*) was developed in order to surpass those difficulties. It makes use of graphical representations like Histograms, filtering and searching features that significantly improve readability and increase the user interactivity.

### 3.1 Comparing Contrast Sets from different periods

Despite being a self-sufficient form of rule, Contrast Sets require that a measure of interest is defined. This will enable the contrast comparison from different periods and will be able to contribute for patterns finding. That measure would ideally capture the contrast own strength at each period. Its evolution along the time line would reveal if that specific contrast got "stronger" or instead got "weaker".

*Supdif* (difference in group support) remains as the only and obvious choice and it indeed answers successfully the questions posed before. It can act as a measure of interest to gauge the strength of a contrast (bigger the difference, stronger the contrast). Observations regarding how the contrast evolved along the time can be done with ease and some patterns involving this notion will be presented next.

### 3.2 Patterns of interest

From the contrasts present at each single point in time, the goal is to find patterns that can somehow express how some contrast has evolved along time. Those results are presented in the form of temporal patterns.

**Growth and Shrink** The first patterns relate to the widening and narrowing of the support difference of the involved groups in consecutive periods. The issue here revolves around the quantification of how much does the contrast needs to grow or shrink to consider it a significant change. It is clearly dependent on the context involved which definitely imposes the requirement of some input from the user who should be able to suggest the adequate value due to its domain dependency. This threshold is called the *sigdif* and operates much like *support* and *confidence*. If the difference from one period to the next is greater (in module) than the *sigdif* value then the variation is deemed as significant and should be reported. With this threshold, the first two patterns appear and they are called *Growth* and *Shrink*. They are the dual of each other with the first referring to the situation where some contrast has its *supdif* grow bigger than the sigdif value from point $N$ to point $N + 1$. The second is the exact opposite.

These two patterns assume a important role, since they alert the end-user to a relevant spike in a contrast by moving to the next period. This change highlight that the groups being contrasted suffered some kind of modification for that specific antecedent and that change might be potentially informative for the end-user. This might help to locate some specific contextual phenomenon that occurred at that time and thus enable him to establish some possible relation of cause-effect.

**Spring Up and Fade Out** Two other patterns came up one opposite to another, much like the two listed above. This time, the goal here solely involves the appearance and disappearance of a contrast in the periods considered.

Consider an example where a contrast is found for period $N$ and $N + 2$ but not for period $N + 1$. This "hole" should trigger the analyst to query what happened at that moment. Knowing exactly why there is no contrast might entail strategical and valuable information. From point $N$ to point $N + 1$ there is the disappearance of the contrast. This kind of pattern is referred as *Fade Out*. That same contrast arises again from period $N + 1$ to period $N + 2$, an example of an occurrence of a *Spring Up*.

**Flip** The last pattern is the *Flip*. The name selected is well representative of its nature because of the "180 turn" notion that this pattern entails. Let's consider that for some antecedent there are two groups being contrasted, $A$ and $B$. At some point in time, the contrast $A >> B$ exists but a few periods later this contrast disappears and gives place to $B >> A$. Hence the name Flip because the contrast directionality was turned around. Due to its specific nature, the *Flip* is the less frequent temporal pattern.

### 3.3 Stability measure

Apart from the patterns developed and described before, the lack of a global mean to evaluate a contrast motivated the development of a measure. The patterns introduced with exception of *Flip* operate in consecutive periods (i.e. locally) and do not allow to categorize or obtain the general behavior of a specific contrast in its whole lifetime.

The existence of a numerical value that could gauge the variability of a contrast would provide an easy and intuitive manner to understand how the contrast evolved. For instance, it would enable to verify whether it suffer frequent abrupt changes or instead it has remained relatively stable in all considered periods.

To achieve its purpose, this stability measure will be based on the following two premises:

- The maximum score or value will be given to a contrast that appeared in all the periods considered and did not suffer any significant variations (no *Growth* or *Shrink* patterns);
- Any pattern found will contribute to lower the score since they translate significant variations that affect what we consider contrast stability;

The proposed formula for this measure that abides by the remarks stated above is the following:

$$Stability = \frac{T - \frac{P}{2}}{N - 1} \tag{3}$$

$N$ represents the number of periods considered. $T$ stands for the number of consecutive periods with contrasts found. $P$ is the number of *Growth* and *Shrink* patterns found in the whole time line. In the denominator, $N - 1$ simply represents the number of transitions present in the periods considered. Best case scenario, $T = N - 1$, which means contrasts have been found for every single period. Thus, it becomes evident that Stability varies from 0 to 1. If there are never two contrasts found in consecutive periods then $T = 0$. Consequently $Stability = 0$, which seems adequate since there is absolutely no consistency as contrasts that appear in one period immediately disappear in the next one.

### 3.4 PPCS Implementation

The application developed can be summarized in a high-level, simplified pseudocode listed in algorithm 1

The set of files ($F$) produced by CAREN will be read and its contents inserted into the data structure in a multi-threaded fashion having each thread processing one file. Then, the list of antecedents ($A$) and contrasts ($C$) will be traversed in order to find Patterns and calculate Stability. This algorithm has a time complexity of $\Theta(|F| + |A| \times |C|)$.

**input** : files $F$, sigdif $S$
**output**: results $R$

```
1  R := ∅;
2  D := ∅;
3  validateUserInput();
4  foreach file ∈ F do
5      D += insertIntoDataStructure(file);
6  end
7  foreach antecedent ∈ D do
8      R += findFlip(antecedent);
9      foreach contrast ∈ antecedent do
10         R += findPatterns(antecedent, contrast, S);
11         R += calculateStability(antecedent, contrast);
12     end
13 end
14 Return R;
```

**Algorithm 1**: PPCS pseudo-code

### 3.5 PPCS Results Viewer

This application intends to be an alternative to the results expressed in a textual form. It makes use of visual representations and some other features to perform different tasks. By using graphical features it will enrich the analysis, increasing the readability and understanding of the Contrast Sets previously found.

After loading the output file from PPCS, the main frame containing all the features available is constructed and it is represented in figure 1.

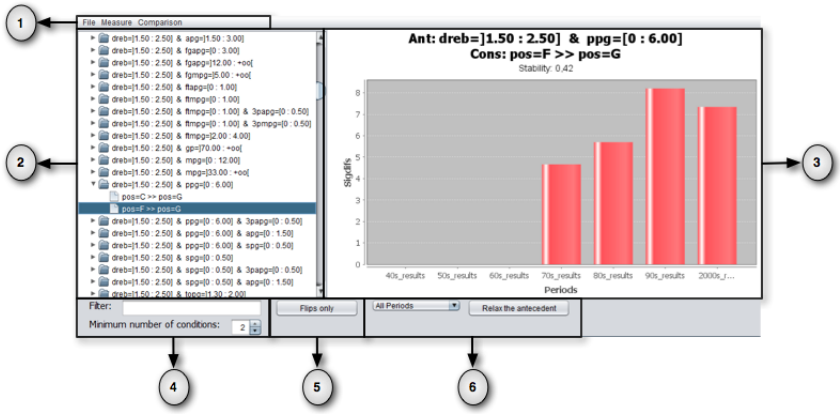


**Fig. 1.** *Viewer* main frame decomposed by areas

Every Contrast Set found is contained in a tree-like representation as seen in figure 1, area 2. It follows a 2-level approach allowing each antecedent to be expanded. It will show which contrasts were found for that same antecedent. This allows for a better organization of the Contrast Set and makes navigation more intuitive. By clicking in a antecedent or contrast, the graphical pane (area 3) is updated. What is represented is dependent on what is selected in the *Tree model*. This entails the close relation between both components and how they depend on each other.

A chart like an *histogram* expresses the contrasts evolution along the time line. It permits a quick identification of periods with and without contrasts as well as the patterns found.

*Tree Model* usually has a considerable number of items present and locating some specific antecedent might involve some scrolling effort which is not desirable. The features present in area 4 have been implemented in order to improve that situation. The first one relates to a typical filter as indicated by the label and a *text box* in which the user can type. This works as a filter over antecedents if the introduced text matches. The other feature discards antecedents which number of items are inferior to the number present in the *spinner*. This is useful for finding the complex antecedents which can reveal interesting relations.

*Flips* usually appear just a few times and a mechanism to quickly spot them was developed in the form of a toggle button (area 5). When pressed on, the *Tree Model* shows only the Contrast Sets that contain at least one *Flip* pattern.

In area 6 the antecedent relaxation feature is present. It attempts to help the user in finding a possible explanation to why some specific contrast was not discovered in a specific period. If an antecedent with minus one item than the antecedent being analyzed, has a contrast in that specific period, one might conclude that the item removed may be the main reason (or at least a contributer) for that event.

For a given antecedent, all its one step above generalizations are considered. For the selected period and contrast, a list of antecedent generalizations is constructed. The option *"All periods"* widens the test not only to one period but to all periods without contrasts. If the antecedent being tested has at least one contrast in a period where the more specific one did not, the more general antecedent is added to the list. Still, there is the possibility of another outcome. This happens when no antecedent with one less item has a contrast for the selected period (or periods). In that a case, a dialog pops up querying the user whether he intents to find a generalization of that antecedent (of any size) that has a contrast for that specific period.

## 4 Case Studies & Experimentation

In order to ascertain the accuracy of the proposal, two distinct datasets were studied. First scenario involved the study of data collected from the Portuguese Ministry of Labor and Social Security for all employed individuals in the private sector ranging from 1986 until 2009. The main goal was to check how the gen-

der of an individual affects attributes like salary, education, etc. Each year was considered as an observation, comprising a total of 24 time points.

The results obtained were highly discriminative in regard to gender and some early suspicions were confirmed. For the higher tiers of income, the contrast of $sex = male >> sex = female$ was always present regardless of the period considered. This confirms common sense believe that, in average, men earn more than women. One attribute that displayed effective modification in the years considered was the workers education. The current trend is that women pursue higher education more than male counterpart with the gap between them increasing at a steady pace. Figure 2 corroborates this situation.

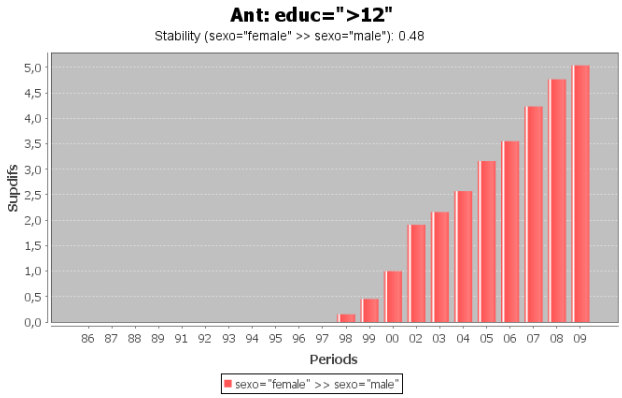


**Fig. 2.** Contrasts found for individuals with higher education

The other case study was related to sports, more specifically basketball. The analysis aims to understand how each position on the field affected the typical statistical contribution and how it evolves over the years. The data obtained ranged from 1946 until 2010 with every player totals from each regular season in the NBA. Each period was defined as a decade. Three groups were considered according to a broader set of positions: *Guards* (G), *Forwards* (F) and *Centers* (C).

The results pointed towards an increasingly positional discrepancy, where players tend to have a more specific skillset regarding their position on the field. In the early days, the difference between players playing different roles were not as significant as it is in nowadays NBA seasons. Another attribute which marked a clear change in the sport was the height of the players along each position. Labeled as a big men sport, this tendency is observed along time with growing emphasis. Figure 3 reveals the contrasts found for players that are 6'6" tall (1,98cm). In early days, the contrast $pos = C >> pos = G$ is present which states that players with less than 2 meters were tall enough to be *Center* players (usually the biggest player in the team). Nowadays, players with that height

play the *Guard* position (smaller players in the field), justified by the contrast $pos = G >> pos = C$ in later periods. This emerged a *Flip* pattern.

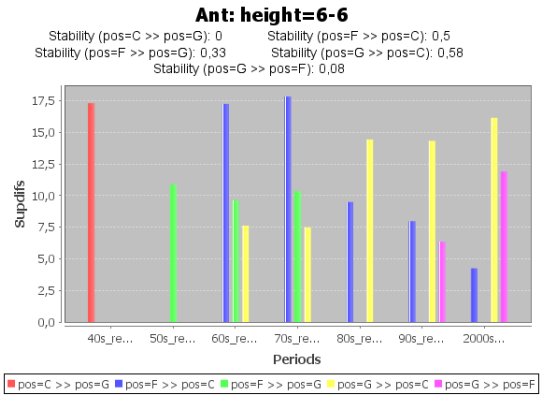


**Fig. 3.** Contrasts found for height = 6'6"

The obtained patterns enable to categorize each positional contribution in terms of the considered attributes. For *Guards*, it was evident the better *Three Point*, *Free Throw Percentage*, more *steals* and *assists* than players from other positions and smaller height. *Centers* players exhibit better Shooting Percentage, the ability to get more *rebounds* and *blocks* than other players and bigger height. As for *Forwards*, they tend to stay in the middle ground between *Guards* and *Centers*. This seems to sustain their known versatility and mixed caracteristics from the other positions.

## 5 Conclusion

This paper aimed to bring the concept of discrimination pattern to a temporal setting in order to check how group differences evolved along time. The post-processing scheme employed has its merits since it was effortless to import the contrasts found by RCS usage into the application developed (PPCS).

However, another approach was also possible which would integrate the role of RCS and PPCS into a single application. This could probably relax the process from the user standpoint. It would imply the reduction of the number of performed steps. There was also the possibility to obtain a faster execution time by removing certain steps like CSV files creation and import.

The patterns developed for this effect had a significant impact in revealing intriguing situations and on others that contain the so called common knowledge. Still, they tend to rely solely on the stepping from one period to the next, not focusing in a more global form of behavior that considers a set of periods. *Stability* arose as a way to tackle this. Despite being able to characterize the contrast

evolution in terms of its general behavior, there are some situations that could benefit from a special emphasis given by a new pattern (or set of patterns).

The example present in figure 2 could be one of those examples. From 1998 onwards, the *supdif* is steadily increasing but since it never increases more than the 1% *sigdif* threshold defined in each passing period, there are no *Growth* patterns. Despite this, one of these sequences of continuous growth could be meaningful and future work could look upon this matter, obtaining new patterns to stress this potentially intriguing situations. A time window concept like the one used in the work by Manilla et al. [8] could serve this purpose by defining a number of periods that could reveal a persistent trend.

## References


1. R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. *SIGMOD Rec.*, 22:207–216, June 1993.
2. P. J. Azevedo. Caren - class project association rule engine. http://www.di.uminho.pt/∼pja/class/caren.html.
3. P. J. Azevedo. Rules for contrast sets. *Intell. Data Anal.*, 14:623–640, November 2010.
4. S. D. Bay and M. J. Pazzani. Detecting change in categorical data: mining contrast sets. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '99, pages 302–306, New York, NY, USA, 1999. ACM.
5. S. D. Bay and M. J. Pazzani. Detecting group differences: Mining contrast sets. *Data Min. Knowl. Discov.*, 5:213–246, July 2001.
6. R. J. Bayardo, Jr. Efficiently mining long patterns from databases. In *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, SIGMOD '98, pages 85–93, New York, NY, USA, 1998. ACM.
7. R. J. Hilderman and T. Peckham. A statistically sound alternative approach to mining contrast sets. In *Proceedings of the 4th Australasian Data Mining Conference (AusDM*, pages 157–172, 2005.
8. H. Mannila, H. Toivonen, and A. Inkeri Verkamo. Discovery of frequent episodes in event sequences. *Data Min. Knowl. Discov.*, 1(3):259–289, Jan. 1997.
9. G. I. Webb. Layered critical values: a powerful direct-adjustment approach to discovering significant patterns. *Mach. Learn.*, 71:307–323, June 2008.
10. G. I. Webb, S. Butler, and D. Newlands. On detecting differences between groups. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pages 256–265, New York, NY, USA, 2003. ACM.