Expert Systems

Article

DOI: 10.1111/exsy.12080

Contrast set mining in temporal databases

André Magalhães and Paulo J. Azevedo*

HASLab/ INESC TEC, Departamento de Informática, Universidade do Minho, Braga, Portugal
E-mail: afgmagalhaes@gmail.com; pja@di.uminho.pt

Abstract: *Understanding the underlying differences between groups or classes in certain contexts can be 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. A previously proposed technique is rules for contrast sets, which seeks to express each contrast set found in terms of rules. This work extends rules for contrast sets 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 evaluate the proposal accuracy and ability to discover relevant information, two different real-life data sets were studied using this approach.*

Keywords: software engineering, artificial intelligence, knowledge acquisition, knowledge representation, knowledge base < system

1. Introduction

Understanding the difference between contrasting groups is a fundamental data mining task (Bay & Pazzani, 1999; Webb *et al*., 2003). This task can be used in many different domains. For example, census data collected this year can be compared with the data collected in a previous census activity, contrasting the data collected this year against the one collected 30 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 (Webb *et al*., 2003). Because of this latent inability, some techniques derived from association rule mining have been proposed to tackle this problem. Contrast set mining (CSM) (Bay & Pazzani, 1999; Hilderman & Peckham, 2005; Azevedo, 2010) 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 (Azevedo, 2010).

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. Such setting enables an analysis of how contrasts evolve along 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 in the past or in 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 behaviour. Potentially, this can be considered highly valuable information by the end-user.

The paper is organized as follows: Section 2 briefly surveys CSM. The proposal with the patterns developed as well as the whole strategy to obtain and analyse them is described in Section 3. Section 4 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 (1999), as the problem of finding all contrast sets whose support differ meaningfully across groups. Association rule mining usually deals with market basket data; however for this specific problem, the data are represented as relational

data. The *itemset* concept present in association rules can be extended to contrast sets as defined by (Bay & Pazzani, 1999).

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

Example $(\text{sex} = \text{male}) \wedge (\text{occupation} = \text{manager})$

In this context, the support is considered in regard to the group and not to the whole data set, 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 δ 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 represents 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 equation (1) is statistically valid are called *significant*, and those that met equation (2) are referred as *large*. If both criteria are met, they are considered as *deviations*.

2.1. Search and Testing for Understandable Consistent Contrasts

Presented in the first paper that introduced CSM (Bay & Pazzani, 1999), Search and Testing for Understandable Consistent Contrasts (STUCCO) is still widely used for mining contrast sets. It is based on Max-Miner (Bayardo, 1998) rule discovery algorithm and uses a breadth-first search framework.

In order to check for *significant* contrast sets (equation (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 represents 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 α 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 that 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 on the basis of some criteria (Bay & Pazzani, 1999, 2001) when there is a guarantee that a node and its own subtree will not contribute for finding deviations; for this reason, they do not need to be visited further.

After determining which contrast sets are interesting, STUCCO presents the results 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 three groups. In the second line, there are the absolute and relative values of support within each group and below the statistical values such as degrees of freedom, χ^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) (Azevedo, 2010) is a proposal that makes uses of an existing association rule engine (Azevedo, 2007) 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 such as STUCCO and CIGAR (Hilderman & Peckham, 2005) that do it in a breadth-first manner. This type of traversal does not fully exploit the downward closure property of support (Agrawal *et al.*, 1993) but still leads to an efficient rule-based algorithm (Azevedo, 2010).

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

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

The *p*-value is computed as follows:

$$p = \sum_{i=0}^{\min(b,c)} \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{(a+b+c+d)!(a+i)!(b-i)!(c-i)!(d+i)!} \tag{3}$$

where $\min(b, c)$ is the minimal value between values b and c. Fisher is an exact test and suitable for small samples. Being a directional test instead of a two-sided test, that is the case

of χ^2, a smaller p-value is generally obtained for data sets with two groups allowing more patterns to be found within the same cutoff level (Azevedo, 2010). The directionality of the test implies a slightly different null hypothesis:

$$H_0 : P(cs|G_i) \leq P(cs|G_j) \quad (4)$$

Because of this, instead of equation (1) it is used: $\exists\, i,j\ P(cs\,|\,G_i) > P(cs\,|\,G_j)$. Even being slightly different, the same principle applies as both derivations still capture the same *significant* patterns.

False discoveries are controlled differently than STUCCO. The layered critical values (Webb, 2008) proposal is adapted for this context. The adjustment used in STUCCO (Bay & Pazzani, 2001) 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. Work carried out by Webb (2007) introduces a form to calculate the search space size before deriving any rule either with market basket data or attribute–value data that are easily adapted to this scenario. This is shown in more detail in these works (Webb, 2008; Azevedo, 2010).

Rules that describe the contrast sets whose support differs across groups are formally organized as follows: $G_1 >> G_2, \ldots, G_i >> G_j \leftarrow cs$, 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 > 50 K.
```

The rule is to be read as follows: *the occurrence of the contrast set 'working for the state government and making an income of more than 50 K' 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 50000), 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 evolved groups and direction in which difference between those groups occur are described.

3. Temporal data mining

Temporal data mining is concerned with data mining of large *sequential* data sets (Laxman & Sastry, 2006). Temporal data may be categorized in two main types: sequential data, a sequence composed by a series of nominal symbols from a particular alphabet (Antunes & Oliveira, 2001), and time series data, also a sequence but composed of continuous and real-valued element values where each event has uniform distance in the time window (Shahnawaz *et al*., 2011).

Time series analysis dates back longer than TDM. Stock market, medical care and weather forecasting are examples of the most common problems studied in this area (Antunes & Oliveira, 2001; Laxman & Sastry, 2006). TDM, however, has a different approach as the goals are somewhat distinct especially in the type of information expected to be retrieved (patterns versus predictions).

3.1. Sequence mining

To put it simple, sequence mining seeks to unearth all patterns of interest (Shahnawaz *et al*., 2011) from sequential data. To discover such patterns in a sequence of events, three steps are usually associated with this approach (Antunes & Oliveira, 2001): representation and modelling of the data into a suitable form, definition of a similarity measure to compare and distinguish sequences and mining operation suitable to solve the task at hand; the general problem of sequence mining was stated by Pujari (2001) as follows:

Definition 2 Let $\sum = \{i_1, i_2, \ldots, i_m\}$ be a set of distinct items comprising the alphabet. An event is a non-empty, disordered collection of items denoted as $(i_1, i_2, \ldots, i_k)$ where i_j is an item in $\sum$. A sequence $s = \{t_1, t_2, \ldots, t_n\}$ is an ordered set of events.

3.2. Frequent episodes

Another approach for unearthing temporal patterns is the frequent episode discovery framework (Mannila *et al*., 1997). In this framework, the objective is to find temporal patterns (designated here as *episodes*) that appear a sufficient number of times from the *event sequences* given.

Mannila and coauthors (Mannila *et al*., 1997) applied the framework in a telecommunication alarm setting. The main objective was to find relationships between alarms from the discovered *episodes* in order to better explain the problems that cause alarms to fire and to predict severe faults.

The sequence of events composes the input provided. Each event has an associated time of occurrence. Given a set E of *event types*, an *event* is a pair (A, t) where $A \in E$ and t is the time of the event. An *event sequence* **s** on E is three-tuple (s, T_s, T_e), where $s = \{(A_1, t_1), (A_2, t_2), \ldots, (A_n, t_n)\}$ is an ordered sequence of events such that $A_i \in E$ for all $i = 1, \ldots, n$, and $t_i \leq t_{i+i}$ for all $i = 1, \ldots, n-1$. T_s represents the starting time and T_e the ending time with $T_S \leq t_i < T_e$ for all $i = 1, \ldots, n$.

The concept of *interest* for a *frequent episode* is given by how close in time the events of an *episode* must arise. The user is able to define the width of the *time window* within which the *episode* must occur. Formally stated, a *window*

on an *event sequence* $\mathbf{s} = (s, T_s, T_e)$ is an event sequence $\mathbf{w} = (w, t_s, t_e)$, where $t_s < T_e$ and $t_e > T_s$, and w consists of those pairs (A, t) from s where $t_s \leq t < t_e$. The time elapsed $t_e - t_s$ is designated as the width of the window w.

An *episode* is a partially ordered collection of events occurring together (Mannila *et al.*, 1997).The notion of *frequency* of an *episode* assume a similar meaning as *support*. It is defined as the fraction of all fixed-width sliding windows over the data in which the episodes occur at least once (Mannila *et al.*, 1997).

3.3. Contrast and temporal mining

Some recent proposals exist in the literature that applies CSM in a temporal framework. For instance, (Wang *et al.*, 2013) uses contrast mining to address action recognition in videos by modelling spatial–temporal structure of human poses. They use a method to estimate human join locations from videos. Data are obtained from these estimations. The authors apply emerging patterns (Dong & Li, 1999) to these data to obtain spatial and temporal parts sets. These patterns assist in deriving body actions. No patterns are derived describing how contrast evolve along time.

(Yang *et al.*, 2008) proposed a new methodology in the context of classification learning that carries out contrast mining by measuring the degree of conceptual equivalence between groups. The authors did not consider a temporal dimension along the used data.

In (Langohr *et al.*, 2012), the authors proposed to extend subgroup discovery, where interesting subsets of objects of a given class are found, by a second subgroup discovery step to find interesting subgroups of objects specific for a class in one or more contrasting classes. They applied this method to gene potate database with three time points of observation. This method does not derive patterns representing contrasts evolving along time.

One of the main sources of failure in concurrent systems is unforeseen encompassed interleavings. (Leue & Befrouei, 2013) proposed to apply sequential mining methods for revealing unforeseen interleavings in the form of sequences of actions derived from counterexamples. The used data are sequential data representing sequences of actions, that is, traces. The proposed method is based on contrasting the patterns of a set of counter examples with the patterns of a set of correct traces that do not violate a desired property. Although sequential data carry a temporal dimension, this work does not consider how contrasting patterns evolve along time.

4. Proposal

The method can be summed up in a three-step process that occurs in a serialized manner. Figure 1 represents the whole process and how each individual step is related, showing the output that is produced at each stage that serves as input for the next step. The output that is produced at each stage serves as input for the next step.

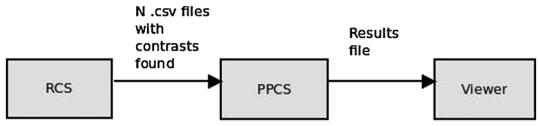

Figure 1: *Process overview.*

The RCS is the first operator in the chain. It is the algorithm included in CAREN (Azevedo, 2007) for discovering contrast sets in any given data set. This will be used in order to obtain the contrasts at each observation (single period of time considered). Having individual data sets for each period, it will be executed as many times as the number of periods. For each execution, comma-separated values output file that contains all the contrasts found and related information such as group support values and *p*-values, among others, will be produced.

Post-processing contrast set (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. Because 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 such as histograms, filtering and searching features that significantly improve readability and increase the user interactivity.

4.1. 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's 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.

4.2. 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.

4.2.1. Increase 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 that 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 *increase* 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 an important role, because they alert the end-user to a relevant spike in a contrast by moving to the next period. This change highlights 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 or her to establish some possible relation of cause–effect.

4.2.2. Spring up and fade out Two other patterns came up one opposite to another, much like the two listed earlier. 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*.

4.2.3. 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 of the contrast directionality was turned around. Because of its specific nature, the *flip* is the less frequent temporal pattern.

4.3. 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 behaviour 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 suffers 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 *increase* or *shrink* patterns).
- Any pattern found will contribute to lower the score because they translate significant variations that affect what we consider contrast stability.

The proposed formula for this measure that abides by the remarks stated earlier is as follows:

$$\text{Stability} = \frac{T - \frac{P}{2}}{N - 1} \tag{5}$$

N represents the number of periods considered. T stands for the number of consecutive periods with contrasts found. P is the number of *increase* and *shrink* patterns found in the whole time line. These are the patterns that affect stability, but only a factor of 0.5 is considered to diminish the impact in the computed stability value. In the denominator, $N-1$ simply represents the number of transitions present in the periods considered. Best case scenario is $T = N-1$, which means that contrasts have been found for every single period. Thus, it becomes evident that stability varies from 0 to 1. If there are no two contrasts found in consecutive periods, then $T = 0$. Consequently, stability $= 0$, which seems adequate because there is absolutely no consistency as contrasts that appear in one period immediately disappear in the next one.

4.4. Post-processing contrast set implementation

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

Algorithm 1. PPCS pseudo-code

```
   input : les 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 += _ndPatterns(antecedent, contrast, S;
11         R += calculateStability(antecedent; contrast);
12     end
13 end
14 Return R;
```

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 $\Theta(|F| + |A| \times |C|)$.

4.5. *Post-processing contrast set 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 2.

Every contrast set found is contained in a tree-like representation as seen in Figure 2, area 2. It follows a two-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 an 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 that 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 analysed has a contrast in that specific period, one might conclude that the item removed may be the main reason (or at least a contributor) 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 also 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 case, a dialogue pops up querying the user whether he or she intents to find a generalization of that antecedent (of any size) that has a contrast for that specific period.

5. Case studies and experimentation

In order to ascertain the accuracy of the proposal, two distinct data sets 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 to 2009 (except years 1990 and 2001 where data were not available). The main goal was to check how the gender of an individual affects attributes such as salary and education. Each year was considered as an observation, comprising a total of 22 time points.

The results obtained were highly discriminative in regard to gender, and some early suspicions were confirmed. In

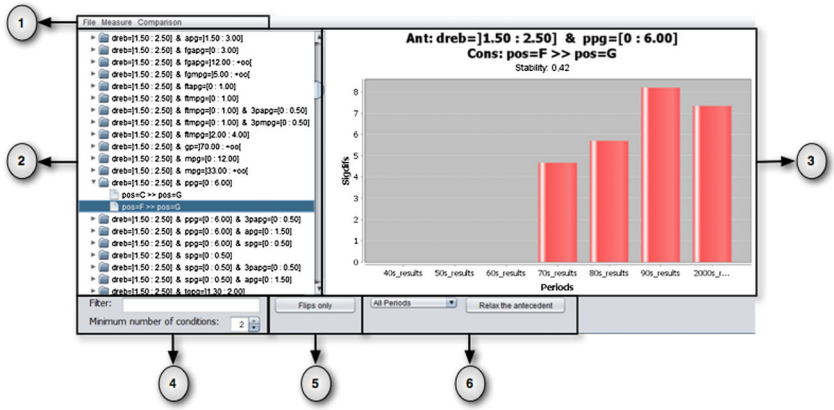

Figure 2: *Viewer main frame decomposed by areas.*

Figure 3, 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 believing 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 4 corroborates this situation.

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 to 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 *centres* (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 was not as significant as it is nowadays in NBA seasons. Another attribute that marked a clear change in the sport was the height of the players along each position. Labelled as a big men sport, this tendency is observed along time with growing emphasis. Figure 5 reveals the contrasts found for players that are 6 ft 6 in. tall (198 cm). In early days, the contrast $pos = C > > pos = G$ is present, which states that players with less than 2 m were tall enough to be *centre* 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.

The obtained patterns enable to categorize each positional contribution in terms of the considered attributes. For *guards*, it was evident that the better *three point*, *free throw percentage*, the more *steals* and *assists* than players from other positions and with smaller height. *Centre* players exhibit better shooting percentage, the ability to get more

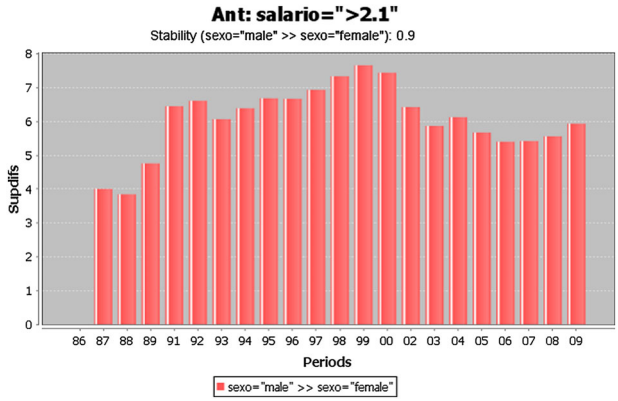

Figure 3: *Contrast found for employees whose income is over 2.1 (in natural log scale).*

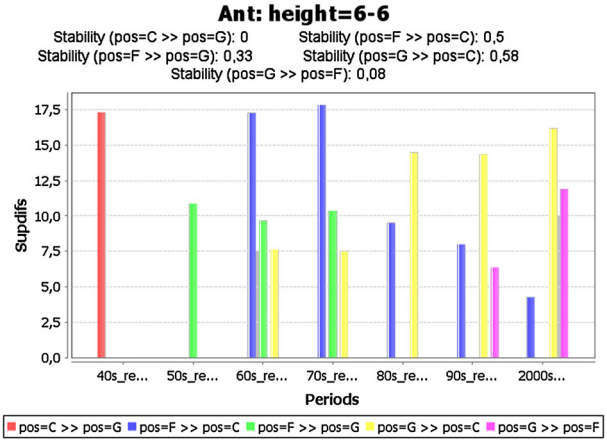

Figure 5: *Contrasts found for height = 6 ft 6 in.*

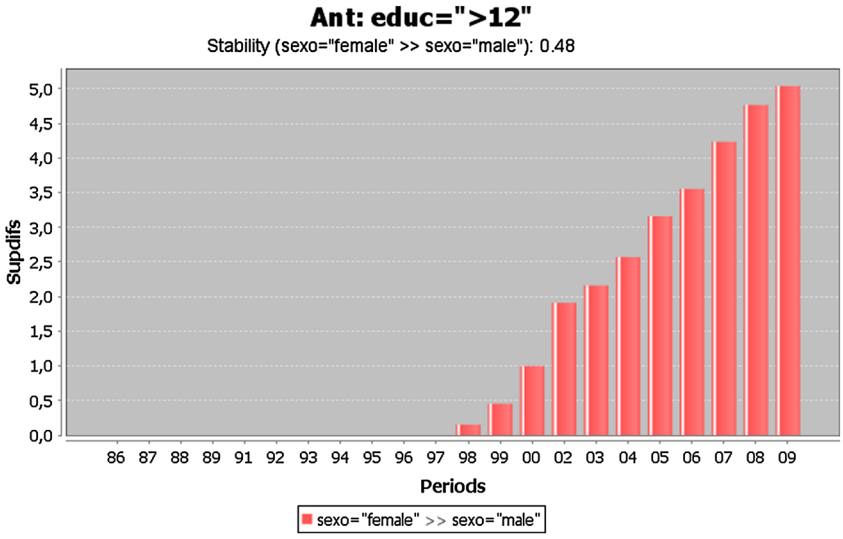

Figure 4: *Contrasts found for individuals with higher education.*

rebounds and *blocks*, than other players and those with bigger height. As for *forwards*, they tend to stay in the middle ground between *guards* and *centres*. This seems to sustain their known versatility and mixed characteristics from the other positions.

6. 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 because it was effortless to import the contrasts found by RCS usage into the application developed (PPCS).

However, another approach was also possible that 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 comma-separated values files creation and import.

A frequent setback involved the presence of continuous, numerical attributes. This always called for a discretization treatment on this type of attributes. Research on CSM generally overlooks this situation and assumes that the data are composed exclusively of categorical attributes with a finite set of values, which is not always the case. The two data sets used required numerical discretization for a considerable number of attributes. There has been a proposal for the incorporation of the discretization process in the contrast set mining algorithm (Simeon & Hilderman, 2007) but has a significant drawback of increasing the size of the search space. This matter could be looked upon in a future iteration, as the results produced are directly affected by the specific chosen discretization method.

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 behaviour 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 behaviour, 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 4 could be one of those cases. From 1998 onwards, the *supdif* is steadily increasing, but because it never increases more than the 1% *sigdif* threshold defined in each passing period, there are no *increase* patterns. Despite this, one of these sequences of continuous expansion 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 Mannila *et al.* (1997) could serve this purpose by defining a number of periods that could reveal a persistent trend.

The graphical user interface application developed to inspect the obtained results, the *Viewer*, although being something not considered in an initial phase assumed a prominent role by significantly improving the analysis compared with the previous form (i.e. the results presented in a simple fashion, by just displaying them in a plain text file). Its features also allow to reduce the set the contrasts presented, detect the patterns found easily and provide visual elements and aspects that aid the comparison effort.

References

Agrawal, R., T. Imieliński and A. Swami (1993) Mining association rules between sets of items in large databases, *SIGMOD Record*, **22**, 207–216.

Antunes, C. M. and A. L. Oliveira (2001) Temporal data mining : an overview. *Lecture Notes in Computer Science*, 1–15.

Azevedo, P. J. (2007) CAREN - class project association rule engine, http://www.di.uminho.pt/~pja/class/caren.html [28 July 2014].

Azevedo, P. J. (2010) Rules for contrast sets, *Intelligent Data Analysis*, **14**, 623–640.

Bay, S. D. and M. J. Pazzani (1999) 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, New York, NY, USA: ACM, 302–306.

Bay, S. D. and M. J. Pazzani (2001) Detecting group differences: mining contrast sets, *Data Mining and Knowledge Discovery*, **5**, 213–246.

Bayardo, R. J. Jr. (1998) Efficiently mining long patterns from databases, in Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, SIGMOD '98, New York, NY, USA: ACM, 85–93.

Dong, G. and J. Li (1999). Efficient mining of emerging patterns: discovering trends and differences, in Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '99, New York, NY, USA: ACM, 43–52.

Hilderman, R. J. and T. Peckham (2005) A statistically sound alternative approach to mining contrast sets, in Proceedings of the 4th Australasian Data Mining Conference (AusDM), Gold Coast, Queensland, Australia: AusDM 2007, 157–172.

Langohr, L., V. Podpecan, M. Petek, I. Mozetic and K. Gruden (2012) Contrast mining from interesting subgroups, in *Bisociative Knowledge Discovery*, Volume 7250 of *Lecture Notes in Computer Science*, Berthold, M. (ed.), Berlin, Heidelberg: Springer, 390–406.

Laxman, S. and P. Sastry (2006) A survey of temporal data mining, *Sadhana*, **31**, 173–198.

Leue, S. and M. Befrouei (2013) Mining sequential patterns to explain concurrent counterexamples, in *Model Checking Software*, Volume 7976 of *Lecture Notes in Computer Science*, Bartocci, E. and C. Ramakrishnan (eds), Berlin Heidelberg: Springer, 264–281.

Mannila, H., H. Toivonen and A. Inkeri Verkamo (1997) Discovery of frequent episodes in event sequences, *Data Mining and Knowledge Discovery*, **1**, 259–289.

Pujari, A. (2001) Data Mining Techniques. India, Hyderguda, Hyderabad: Universities Press.

Shahnawaz, M., A. Ranjan and M. Danish (2011) Temporal data mining: an overview, *International Journal of Engineering and Advanced Technology*, **1**, 20–24.

Simeon, M. and R. J. Hilderman (2007) Exploratory quantitative contrast set mining: a discretization approach, in ICTAI (2)'07, Patras, Greece: ICTAI 2007, 124–131.

WANG, C., Y. WANG and A. YUILLE (2013) An approach to pose-based action recognition, in Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, Portland, Oregan, 915–922.
WEBB, G. I. (2007) Discovering significant patterns, *Machine Learning*, **68**, 1–33.
WEBB, G. I. (2008) Layered critical values: a powerful direct-adjustment approach to discovering significant patterns, *Machine Learning*, **71**, 307–323.
WEBB, G. I., S. BUTLER and D. NEWLANDS (2003) On detecting differences between groups, in Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '03, New York, NY, USA: ACM, 256–265.
YANG, Y., X. WU and X. ZHU (2008) Conceptual equivalence for contrast mining in classification learning, *Data & Knowledge Engineering*, **67**, 413–429.

The authors

André Magalhães

André Magalhães did his MSc (2012) at Departamento de Informática of Universidade do Minho in Data Mining. His research interests are pattern mining and contrast sets mining. He is currently at Deloitte (Portugal) enrolled in data warehousing projects.

Paulo J. Azevedo

Paulo J. Azevedo obtained his PhD from Imperial College at University of London (1995) and is now an assistant professor (with habilitation) at the Department of Informatics University of Minho. He is a member of the High-Assurance Software Lab (HASLab) – INESC Tec L. A. His research interests are knowledge discovery in databases (data mining) and machine learning, in particular association rules, subgroup discovery, motif discovery in time series, graph mining, bioinformatics and recommendation systems. His most recent research interests include distribution learning and social networks. He was a PC member of several data mining conferences such as DS, PKDD, ECML, ECAI and DaWak. He was also a reviewer for several data mining journals.