# Visual Interactive Subgroup Discovery with Numerical Properties of Interest⋆

Alípio M. Jorge[1], Fernando Pereira[1], and Paulo J. Azevedo[2]

[1] LIACC, Faculty of Economics, University of Porto, Portugal
amjorge@liacc.up.pt
[2] Departamento de Informática, University of Minho, Portugal
pja@di.uminho.pt

**Abstract.** We propose an approach to subgroup discovery using distribution rules (a kind of association rules with a probability distribution on the consequent) for numerical properties of interest. The objective interest of the subgroups is measured through statistical goodness of fit tests. Their subjective interest can be assessed by the data analyst through a visual interactive subgroup browsing procedure.

## 1 Subgroup Discovery

Subgroup discovery is an undirected data mining task, first identified by Klősgen [3], and meanwhile studied by others. A subgroup is a subset of a population having interesting values w.r.t. a property of interest. For example, if the average level cholesterol for all the patients of an hospital is 190, we may find interesting that people who smoke and drink have a cholesterol of around 250. In this case, we have a property of interest (the level of cholesterol) and a subgroup of patients with a precise description. This subgroup can be regarded as relevant or interesting due to the fact that the mean of the property of interest is significantly different from a value of reference, such as the mean of the whole population.

**Definition -** Given a population of individuals $U$ and a criterion of interest, a *subgroup* $G \subseteq U$ is a subset of individuals that satisfies the criterion. Each subgroup has a description, (a set of conditions) satisfyed by all its members. ◇

We will define the interest of a subgroup w.r.t. a chosen property of interest (P.O.I.). We assume that the P.O.I. is one and is numerical, although in general we may consider other types of variables. In our work, the notion of interest of a subgroup is not limited to the value of particular measures such as mean. Instead, we compare the observed distribution of the values of the P.O.I. with the distribution of the whole population.

**Definition -** Let $y$ be a numerical property of interest, and $G$ a subgroup with description $desc_G$. The *distribution of the P.O.I.* $y$ for the individuals $x \in G$ is approximated by the observed $\Pr(y|desc_G)$ and is denoted by $D_{y|desc_G}$. ◇

The *a priori* distribution of the P.O.I. is the one for the whole population. Subgroup discovery methods work typically with categorical properties of interest.

⋆ Supported by POSI/SRI/40949/2000/ Modal Project (Fundação Ciência e Tecnologia), FEDER e Programa de Financiamento Plurianual de Unidades de I & D.

N. Lavrač, L. Todorovski, and K.P. Jantke (Eds.): DS 2006, LNAI 4265, pp. 301–305, 2006.
© Springer-Verlag Berlin Heidelberg 2006

Our method constructs the subgroups from discovered distribution rules [2], a kind of association rules with a statistical distribution on the consequent.

**Example** - Suppose we have clinical data describing habits of patients and their level of cholesterol. The distribution rule $smoke \wedge young \rightarrow chol = \{180/2, 193/4, 205/3, 230/1\}$ represents the information that, of the young smokers on the data set, 2 have a *cholesterol* of 180, 4 of 193, 3 of 205 and 1 of 230. This information can be represented graphically as a frequency polygon. The attribute *chol* is the property of interest. ◇

The objective interest of a subgroup is given by the unexpectedness of its distribution for the property of interest, which can be measured with existing statistical goodness of fit tests. We will define the interest of a subgroup as the deviation of the distribution of the property of interest with respect to the *a priori* distribution. In this sense, the interest of a subgroup is akin to the interest of an association rule as measured by lift, conviction, or $\chi^2$ [4]. However, in the current approach, we take into account the distribution of the possible values of the property of interest, instead of only one such value.

**Definition** - The *interest of a subgroup G* is given by the dissimilarity between the distribution of the property of interest for the subgroup $D_{y|descG}$ and a reference distribution $D_{y|ref}$. ◇

The reference is typically the *a priori* distribution. The degree of similarity can be measured using statistical goodness of fit tests. In this work we have used Kolmogorov-Smirnov. Given a dataset $S$, the task of subgroup discovery consists in finding all the distribution rules $A \rightarrow y = D_{y|A}$, where $A$ has a support above a determined mininum $\sigma_{min}$ and $D_{y|A}$ is statistically significantly different from the *a priori* distribution $D_y$ (the p-value of the K-S test is low).

## 2 The Visual Interactive Process

Given a population and a criterion of interest, the number of interesting subgroups/distribution rules can be very large. As in the discovery of association rules, for the data analyst to explore the discovered patterns it is useful that a post processing rule browsing environment exists [1].

We propose a visual interactive subgroup discovery procedure that graphically displays the distribution of each subgroup and allows the navigation by the data analyst in a chosen continuous space of subgroups. To represent the continuous space of subgroups we propose an $x$-$y$ plot, where the coordinates $x$ and $y$ represent statistical measures of the distribution of the property of interest. A simple example is a mean-variance plot. Other subgroup spaces such as median-mode, skewness-kurtosis and mean-kurtosis have also been considered. Skewness and kurtosis are well known distribution shape measures, median, mode and mean are location measures and variance is a spread measure.

Given a two-dimensional plot, each subgroup is represented as a point. This plot will serve as a browsing device (Fig. 1). The data analyst can click on one of the points of that space and visualize the distributions (as frequency polygons) and definitions of the corresponding sugroups. In this phase the selected subgroup is also visually and statistically compared to a reference group. This process is iterated and interesting subgroups found can be saved.

The skewness-kurtosis space gives the data analyst an overall picture of the shapes of the distributions of the subgroups. The mean-kurtosis gives an idea of the location of the distributions as well as of how their shapes are more or less flat. The mean-standard deviation space identifies the subgroups that have their mean below and above the whole population. The mode-median space depicts the location of the distributions of the property of interest for the subgroups, both through mode and median.

## 2.1 Studying Algal Blooms

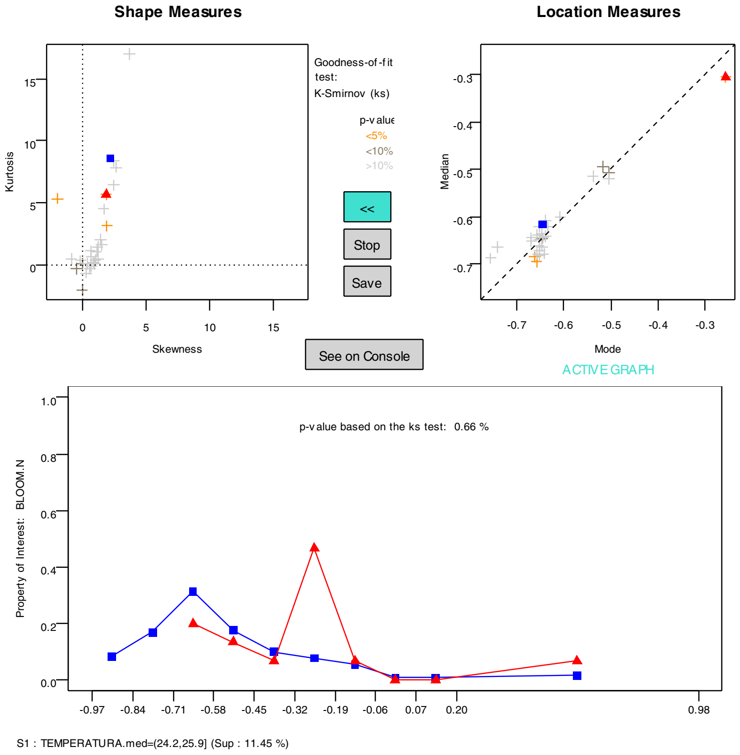
This subgroup discovery approach is being applied to study algae population dynamics in a river which serves as an urban water supply resource. The quantity and diversity of the algae are important for the quality of the water, which makes this an economically and socially critical eco-system. Blooms of these algae may reduce the life conditions in a river and cause massive deaths of fish, thus degrading water quality. The state of rivers is affected by toxic waste from industrial activity, farming land run-off and sewage water treatment [5]. Being able to understand and predict these blooms is therefore very important. This problem has been studied in the MODAL project (modys.niaad.liacc.up.pt/projects/modal) in collaboration with the local water distribution company.

The data were collected from 1998 to 2003. All attributes are continuous and are divided in three groups: *phytoplancton*, *chemical and physical* properties, and *microbiological* parameters. The phytoplancton attributes record the quantity of 7 micro-algae species, the chemical and physical attributes record the levels of various algae nutrients and other environmental parameters, the microbiological attributes record the quantities of some bacteria relevant for water quality.

The original data were pre-processed, so that each record stores the phytoplancton observed in one particular sample, and also each of the other descriptive attributes, aggregating the values observed between two samples of phytoplancton. Aggregating functions were maximum, minimum and median. Attributes with the values of the previous sample of the 7 phytoplancton species were also added, as well as two summary attributes measuring the Diversity and the Density of the algae. These attributes are important since a bloom of one of the species is characterized by a low diversity of species and a high density of algae. Three other attributes were added: Normalized Density and Normalized Diversity (normalized versions of Density and Diversity); and BLOOM.N calculated as the difference between normalized density and normalized diversity. High values of BLOOM.N indicate high possibility of a bloom. After pre-processing we have 72 input variables, 7+5 target variables and 131 cases. In the examples, variable names appear in Portuguese. The results were analysed by a biologist.

We have conducted several studies with different P.O.I., using a minimum support of 0.05. Here we provide a summary of results [1]. With DIVERSIDADE.N

[1] Please consult www.liacc.up.pt/~amjorge/docs/VSG-TR-06.pdf for more.



p-value based on the ks test: 0.66 %

S1 : TEMPERATURA.med=(24.2,25.9] (Sup : 11.45 %)

**Fig. 1.** Screen of the protoype showing two navigation plots (top), and the selected subgroup, where the P.O.I. BLOOM.N is affected by relatively high temperature (frequency polygon in triangles), and compared with the *a priori* distribution (in squares)

(normalized diversity) as the P.O.I., the median values of the microbiology and physical-chemical parameters as explanatory variables, we obtain 98 subgroups. Low diversity is a necessary condition for a bloom to occur. On the mode-median plot we can click on the subgroups with lowest mode and median. One of the rules obtained indicates that, for relatively low values of oxygen and low-medium values of iron there is a relatively high probability (when compared to the whole population) that DIVERSITY is low (between 0.3 and 0.4) or very low (below 0.1). While oxygen is necessary for phytoplancton primary production, the low quantity of one nutrient (iron) may reduce the phytoplancton to the species that live well under those conditions. This situation may lead to a bloom of one of the species. High values of BLOOM.N may indicate algae blooms (high density of micro-algae and low diversity). With this P.O.I., 27 subgroups were found. One of the rules relates relatively high temperatures (around 25 degrees Celsius) with a distribution of bloom values shifted right. This is a well known effect of high temperatures (Fig. 1).

Distribution rule generation is very fast (less than 5 seconds), and moving from subgroup to subgroup is made easy by the graphical interface. The mode-median plot is a very useful browsing device for these data. Looking for extremely skewed distributions of the target variables is facilitated with this subgroup space. By having immediate acces to all the generated subgroups, the data analyst can compare nearby subgroups and examine their descriptions.

The display of the distribution provides information that may be hidden by a summary measure such as mean. A distribution curve with two modes, for example, may indicate that a particular subgroup has two possible outcomes. If one of those outcomes is critical, than the antecedent of the subgroup may become an alarm trigger for water monitoring. This also implies that not only extreme values of median or mode indicate potentially interesting subgroups.

## 3 Conclusions

We have presented a visual interactive subgroup discovery approach for numerical properties of interest. Subgroups are discovered as distribution rules (DR) with sufficient support and having a distribution for the property of interest distinct from the whole population. The similarity between distributions is measured as the Kolmogorov-Smirnov statistical test's p-value. A large set of subgroups is presented to a data analyst as a two dimensional plot, corresponding to a space of subgroups. Each point on the plot is a different subgroup. The data analyst can inspect each of the subgroups by clicking on the respective point. Each subgroup is displayed with its definition, support, and the distribution of the P.O.I. The approach is being used in a project for monitoring the quality of water in a river. In this application, the properties of interest are the ones related with the control of algal blooms, which affect the quality of water.

**Acknowledgments.** We thank Luís Torgo and Rita Ribeiro for making available the problem and data, and also the biologist Dr. Catarina Magalhães, .

## References

1. A. Jorge, Poças, and P. J. Azevedo. Post-processing operators for browsing large sets of association rules. In S. Lange, K. Satoh, and C. H. Smith, editors, *Proceedings of Discovery Science, DS 02, Luebeck, Germany*, number 2534 in Lecture Notes in Computer Science, pages 414–421. Springer-Verlag, 2002.
2. A. M. Jorge, P. J. Azevedo, and F. Pereira. Distribution rules with numerical properties of interest. In *Proceedings of Principles of Data Mining and Knowledge Discovery (PKDD-06)*, LNAI. Springer-Verlag, 2006.
3. W. Klősgen. Explora: A multipattern and multistrategy discovery assistant. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park, CA, 1996.
4. B. Liu, W. Hsu, and Y. Ma. Pruning and summarizing the discovered associations. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 125–134, New York, NY, USA, 1999. ACM Press.
5. R. P. Ribeiro and L. Torgo. Predicting harmful algae blooms. In M. e. a. Pires, editor, *Proceedings of Portuguese AI Conference (EPIA'03)*, volume 2902 of *LNAI*, pages 308–312. Springer-Verlag, 2003.