SUBGROUP VISUALIZATION

Petra Kralj(1), Nada Lavrač(1,2), Blaž Zupan (3,4)
(1) Jozef Stefan Institute, Jamova 39, Ljubljana, Slovenia
(2) Nova Gorica Polytechnic, Vipavska 13, Nova Gorica, Slovenia
(3) Faculty of Computer and Information Science, Tržaška 25, Ljubljana, Slovenia
(4) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX
e-mail: Petra.Kralj@campus.fri.uni-lj.si

ABSTRACT

This paper presents the state of the art of subgroup visualization methods. Visualization methods are evaluated by different criteria. A novel subgroup visualization method is proposed and its implementation as a part of an interactive interface for subgroup discovery is presented.

1 INTRODUCTION

Data visualization methods have been part of statistics and data analysis research for many years. This research concentrated primarily on plotting one or more independent variables against a dependent variable in support of explorative data analysis [3]. The visualization of analysis results, however, gained only recently attention with the proliferation of data mining [2]. The visualization of analysis results primarily serves four purposes:

- to better illustrate the model to the end user
- enable comparison of models
- increase model acceptance, and
- enable support for »what-if questions«

Subgroup discovery [4,5,6] aims at discovering individual patterns of interest. Formally the task of subgroup discovery is defined as follows: given a population of individuals and a specific property of the individuals that we are interested in, find population subgroups that are statistically 'most interesting', e.g. are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest.

Since subgroup discovery is a task of descriptive induction, the visualization of results is crucial for presenting the results to the end user. The subgroup visualization task is to visualize the subgroups detected by subgroup discovery algorithms.

Many subgroups visualization methods have been proposed so far. In this paper we evaluate them by their intuitiveness, attractiveness, correctness of displayed data, usefulness and ability to display the contents of data.

A novel subgroup visualization method that combines two other subgroup visualization methods is proposed and its implementation as a part of an interactive interface for subgroup discovery in the Orange data mining software [7] is presented.

This paper is organized as follows: Subgroup visualization methods are described in Sections 2-7. In Section 8 an implementation of subgroup discovery and visualization is presented.

2 SUBGROUP VISUALIZATION BY PIE CHARTS

Slices of pie charts are the most common way of visualizing parts of a whole. They are daily used and everybody, even a total laic, understands them.

Subgroup visualization by pie chart consists of a two level pie for each subgroup. The base pie represents the distribution of individuals in terms of the property of interest of the entire example set. The upper pie represents the size and the distribution of individuals in terms of the property of interest in a specific subgroup. An example of five subgroups visualized by pie chart is presented in Figure 1.

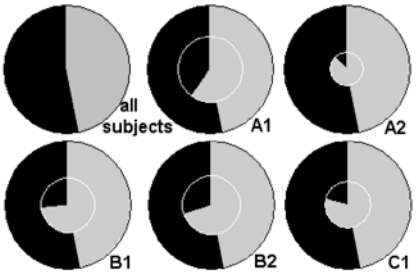

Figure 1: *Subgroup visualization by pie charts.*

The main weakness of this visualization is the misleading representation of the size of subgroups. The size of a subgroup is represented by the radius of the circle. The faultiness is that the surface of the circle increases with the square of its radius. A subgroup that covers 20% of examples would be represented by a circle that covers only 4% of the whole surface, and a subgroup that covers 50% of examples would be represented by a circle that covers 25% of the surface.

In terms of usefulness this visualization is not very handy because it is difficult to compare sizes of circles; neither the comparison of distributions is straightforward. This visualization does not show the contents of subgroups.

3 SUBGROUP VISUALIZATION BY BOX PLOTS

In the visualization by box plots each subgroup is represented by one box plot (all examples are also considered as one subgroup and are displayed in the top box). Each box shows the entire population; the hatched area on the left represents the positive examples and the white area on the right-hand side of the box represents the negative examples. The grey area within each box indicates the respective subgroup. The overlap of the grey area with the hatched area shows the overlap of the group with the positive examples. Hence, the more to the left the grey area extends the better. The less the grey area extends to the right of the hatched area, the more specific a subgroup is (less overlap with the subjects of the negative class). Finally, the location of the box along the X-axis indicates the relative share of the target class within each subgroup: the more to the right a box is placed, the higher is the share of the target value within this subgroup. The line (in Figure 2 at value 46.6%) indicates the default accuracy, i.e., the number of positive examples in the entire population. An example of five subgroups visualized by box plots is presented in Figure 2.

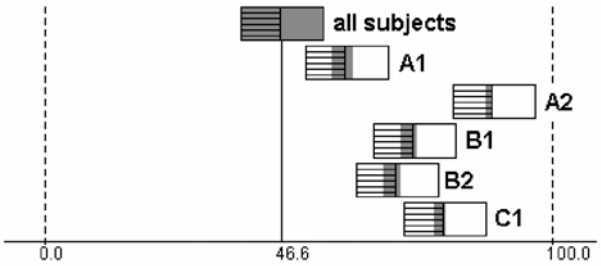

Figure 2: *Subgroup visualization by box plots.*

The intuitiveness of this visualization is scarce since interpretation is absolutely necessary for understanding it. It is illogical since the boxes that are placed more to the right and have more grey colour on the left-hand side represent the best subgroups. This visualization is not very attractive since most of the image is white. The grey area, (the part of the image that really represents the subgroups) is just a tiny portion of the entire image. All the displayed data is correct and the visualization is useful since the subgroups are arranged by their confidence. It is also easier to contrast the sizes of subgroups compared to the pie chart. This visualization does not display the contents of data. The general appraisal of this visualization is bad, even though it is useful.

4 VISUALIZING SUBGROUPS THROUGH DISTRIBUTIONS OF A CONTINUOUS ATTRIBUTE

The distribution of examples by a continuous attribute was first introduced as a visualization method in [1], and was often used in the medical domain. It is the only subgroup visualization method that offers an insight of the visualized subgroups.

The approach assumes the existence of at least one numeric (or ordered discrete) attribute of expert's interest for subgroup analysis. The selected attribute is plotted on the X-axis of the diagram. The Y-axis represents a target variable, or more precisely, the number of instances belonging (Y+) or not belonging (Y-) to the target class for a specific value of the attribute on the X-axis. It must be noted that both directions of the Y-axis are used to indicate the number of instances. The entire data set and two subgroups A1 and B2 are visualized by their distribution over a continuous attribute in Figure 3.

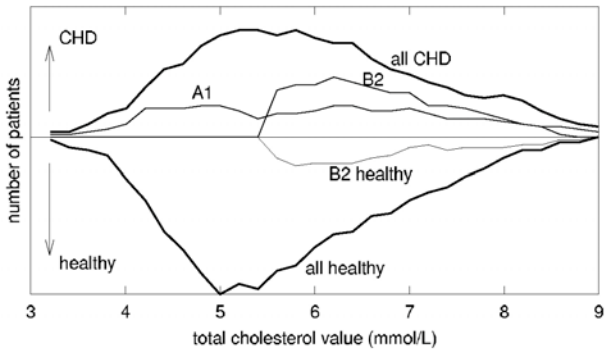

Figure 3: *Subgroup visualization by distribution over a continuous attribute. For clarity of the picture, only the positive (Y+) side of subgroup A1 is depicted.*

This visualization method is not completely automatic, since the automatic approach does not provide consistent results. The automatic approach calculates the number of examples for each value of the attribute on the X-axis by moving a sliding window and counting the number of examples in that window. The outcome is a smooth line. The difficulty appears when the attribute from the X-axis appears as a part of the condition that forms a subgroup. In such case a manual correction is needed for this method to be realistic.

This visualization method is very intuitive since it practically does not need any explanation. It is attractive and very useful to the end user since it offers an insight in the contents of displayed examples. However the correctness of displayed data is questionable.

5 REPRESENTATION IN THE ROC SPACE

The ROC (Receiver Operating Characteristics) space is a 2-dimensional space that shows classifier (rule/rule set) performance in terms of its false positive rate (FPr) plotted on the X-axis, and true positive rate (TPr) plotted on the Y-axis.

The ROC space is appropriate for measuring the success of subgroup discovery, since subgroups whose TPr/FPr tradeoffs are close to the main diagonal (line connecting the points (0, 0) and (1, 1) in the ROC space) can be discarded as insignificant [6]. The reason is that the rules with TPr/FPr on the main diagonal have the same distribution of covered positives and negatives (TPr = FPr) as the distribution in the entire data set. An example of five subgroups represented in ROC space is shown in Figure 4.

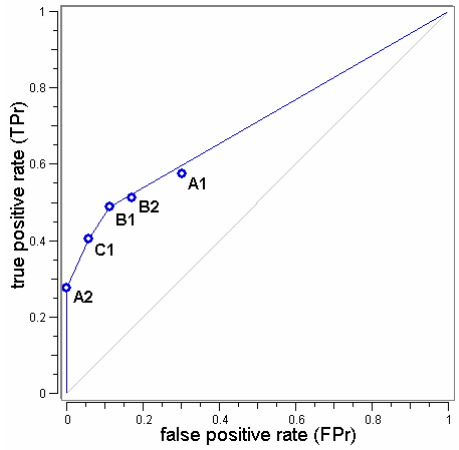

Figure 4: *Representation of subgroups in ROC space.*

Even though the ROC space is a good-looking visualization, it is more often used for evaluation of discovered rules. The ROC convex hull is the line connecting the potentially optimal subgroups. The area under the ROC convex hull (AUC, area under curve) is a measure of the quality of the result.

This visualization method is not intuitive to the end user, but is absolutely clear to any machine learning expert. The displayed data is absolutely correct, even though there is no content displayed. An advantage of this method compared to the others is that it allows the comparison of outcomes of different algorithms at the same time.

6 BAR CHARTS VISUALIZATION

The new visualization method we propose combines the good properties of the pie chart and the box plot visualization. It is simple, understandable and shows all the data correctly. An example of five subgroups visualized by bar charts is shown in Figure 5.

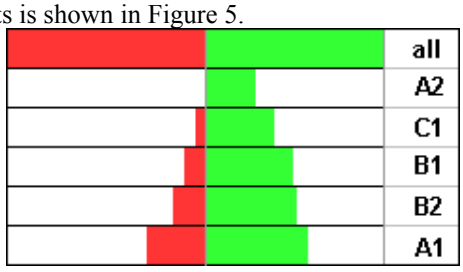

Figure 5: *Subgroup visualization by bar charts.*

In the visualization by bar charts the first line's purpose is to visualize the distribution of the entire example set. The area on the right represents the positive examples and the area on the left represents the negative examples. Each following line represents one subgroup. The positive and the negative examples of each subgroup are drawn below the positive and the negative examples of the entire example set. Subgroups are sorted by the relative share of positive examples.

This visualization method allows simple comparison between subgroups and is therefore useful. It is very intuitive and attractive enough. All the displayed data is correct and not misleading. It does not display the contents of data.

7 SUMMARY OF SUBGROUP VISUALIZATION METHODS

We now discuss the five different subgroup visualization methods by considering their intuitiveness, attractiveness, correctness of displayed data, usefulness and their ability to display the contents of the data. The summary of the evaluation is presented in Table 1.

	intuitiveness	attractiveness	correctness	usefulness	contents
pie chart	+	+	o	-	-
box plot	-	o	+	+	-
continuous	+	+	-	+	+
ROC	+ , -	o	+	+	-
bar chart	+	+	+	+	-

Table 1: *Summary evaluation of subgroup visualization methods.*

8 THE INTERACTIVE INTERFACE

In this section we present an implementation of the bar visualization and the ROC representation of subgroups in the Orange data mining software [7]. Orange goes beyond static visualization, by allowing interaction of the user and combination of different visualization techniques. A screen shot displayed in Figure 6 shows a use case of this tool.

In Figure 6 an example of a visual program in the Orange visual programming tool Orange Canvas is shown. The first widget from the left (File) loads the dataset (in this example we load the Brain Ischemia dataset). The following two widgets (Build Subgroups Apriori-SD and Build Subgroups SD) are two instances of the same widget Build Subgroups that performs subgroup discovery by one of the selected algorithms: SD [4], Apriori-SD [6] or CN2-SD [5]. The rest of the program performs visualization.

The outputs of subgroup discovery are connected to the ROC Visualization widget. In this widget the subgroups discovered by the algorithm Apriori-SD are displayed as blue circles and the subgroups discovered by the algorithm SD are shown as red circles (the window top right). The output of Build Subgroups SD is also connected to the Bar visualization widget (window bottom left). In the bar visualization we can select one or more subgroups (in Figure 6 subgroup D_Fibr>4.30 is selected.) The output of the Bar visualization widget is connected to the ROC Visualization widget causing the circle of the subgroups selected in the Bar visualization to be filled.

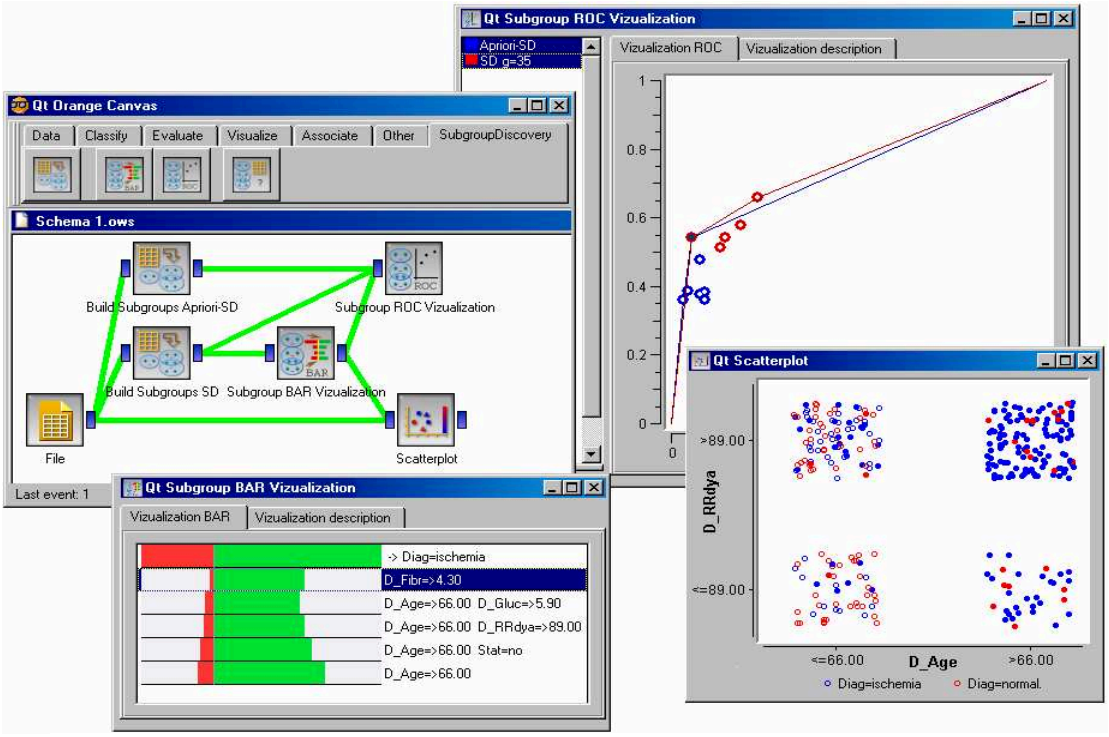

Figure 6: *A screen shot of the usage of the subgroup discovery tool implemented in Orange.*

The last part of the visual program allows us to see the contents of the selected subgroups. The Scatterplot widget takes as its input the example set from the widget File (the empty circles) and a set of examples covered by the selected subgroup(s) in Bar Visualization (full circles). The examples are arranged by the attribute D_Age on the X-axis and by the attribute D_RRdya on the Y-axis.

This visual program is just one example of what can be done by using the Subgroup discovery tool implemented in Orange. Subgroup evaluation and different method for visualizing the contents of subgroups are also available.

9 CONCLUSIONS

This paper presents the state of the art of subgroup visualization methods. The visualization methods are compared and evaluated by different criteria and a new visualization method is proposed. The implementation and usage of the Subgroup Discovery tool in Orange data mining software is presented.

In our view, the implemented capabilities offer new possibilities for understanding and usage of the subgroup discovery process.

Acknowledgements

The authors acknowledge the support of the Slovenian Ministry of Higher Education, Science and Technology and the 6FP EU project Inductive Quesries for Mining Patterns and Models.

References

[1] D. Gramberger, N. Lavrač, D. Wettschereck. Subgroup Visualization: A Method and Application in Population Screening. In Proceedings of the International Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP'02), pages 31–35, 2002.

[2] U.M. Fayyad, G. Grinstein, A. Wierse (2002). Information Visualization in Data Mining and Knowledge Discovery. Morgan Kaufmann Series in Data Management Systems

[3] H.Y. Lee, H.L. Ong, H.L. Quek (1995). Exploiting visualization in knowledge discovery. In Proc. of the First Inter. Conference on Knowledge Discovery and Data Mining, pp 198-203

[4] D. Gramberger, N. Lavrač. Expert-Guided Subgroup Discovery: Methodology and Application. Journal of Artificial Intelligence Research, 17:501–527, 2002

[5] N. Lavrač, B. Kavšek, P. Flach, L. Todorovski. Subgroup Discovery with CN2-SD. Journal of Machine Learning Research, 5: 153–188, 2004

[6] B. Kavšek, N. Lavrač. APRIORI-SD: Adapting Association rule Learning to Subgroup Discovery. Proceedings of the 5th International Symposium on Intelligent Data Analysis, pages 230–241, Springer, 2003.

[7] J. Demšar, B. Zupan, G. Leban. (2004). Orange: From Experimental Machine Learning to Interactive Data Mining. White Paper (www.ailab.si/orange), Faculty of Computer and Information Science, University of Ljubljana.