Subgroup Visualization: A Method and Application to Population Screening

Dragan Gamberger[1], Nada Lavrač[2], Dietrich Wettschereck[3]

[1] Rudjer Bošković Institute, Bijenička 54, 10000 Zagreb, Croatia
dragan.gamberger@irb.hr
[2] Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
nada.lavrac@ijs.si
[3] University of Applied Sciences, Bonn-Rhein-Sieg, 53757 Sankt Augustin, Germany
dietrich.wettschereck@fh-bonn-rhein-sieg.de

Abstract. The paper presents a method for the visualization of subgroups, detected by a subgroup discovery algorithm. The main advantage and novelty of the method is that the visualized models can be used to illustrate the distributions of detected groups in terms of the percentages of true positive and false positive cases covered by the model.

1 INTRODUCTION

A subgroup discovery task can be defined as follows: given a population of individuals and a property of those individuals 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. An example subgroup discovery system is MIDOS [10].

Some approaches to association rule induction can be used for subgroup discovery. For instance, the APRIORI-C algorithm [4], adapting the association rule induction algorithm to classification rule induction, outputs classification rules with guaranteed high support and confidence. As such, each APRIORI-C rule represents a 'chunk' of knowledge about the problem, which is very important for knowledge discovery. In this paper, subgroups were discovered by a new heuristic rule learning algorithm [3]. The actual subgroup discovery algorithm is implemented in the on-line Data Mining Server, available at http://dms.irb.hr, whose description is out of the scope of this paper. The problem of population screening for early detection of atherosclerotic coronary heart disease (CHD) risk groups is used to illustrate the visualization of results obtained by applying our subgroup discovery methodology. To this end, the application of our subgroup discovery algorithm resulted in five models of patients with CHD risk which can be used for population screening.

The paper presents the detected risk groups and discusses approaches for their visualization (Section 2). A novel method for visualization of distributions of subgroups is described in Section 3, including a short review of related work.

2 VISUALIZATION OF CHD RISK GROUPS

Some interesting models of groups of CHD patients were constructed using the methodology of descriptive induction, using the available patient data, collected at the Institute for Cardiovascular Prevention and Rehabilitation, Zagreb, Croatia. There are three typical stages in the risk factor screening process, denoted by A, B, and C [7]. Our goal was to construct at least one model for every stage. Table 1 presents five induced models.

	Principal Factors	**Supporting Factors**
A1	positive family history age over 46 year	psychosocial stress cigarette smoking hypertension overweigth
A2	body mass index over 25 kgm^{-2} age over 63 years	positive family history hypertension slightly increased LDL cholesterol normal but decreased HDL cholesterol
B1	total cholesterol over 6.1 $mmolL^{-1}$ age over 53 years	increased triglycerides value
B2	total cholesterol over 5.6 $mmolL^{-1}$ fibrinogen over 3.7 $mmolL^{-1}$	positive family history
C1	left ventricular hypertrophy	positive family history hypertension diabetes mellitus

Table 1. Induced subgroup descriptions (principal factors) and their statistical characterizations (supporting factors). Subgroup A1 is for males, subgroup A2 for females, while subgroups B1, B2, and C1 are for both genders.

Figure 1 displays the respective coverages of subgroups A1, A2, B1, B2, and C1 in box plots. The figure shows the following information: the size of each subgroup, how it compares to the entire population and the distribution of the target values within each subgroup. Experience gained from working with non-technical end-users has shown that a pie chart visualization is more appealing to these users because they more closely resemble business charts. Pie charts, however, often mislead the perception of the user due to difficulties with relating the size of pie slices to actual values. Hence, the visualization with boxes is preferred. While these figures are more difficult to understand when first encountered, they allow for better comparison of the different subgroups and clearly display the size of each subgroup. This visualization technique can serve as an entry point to the more in depth visualization technique introduced in this paper.

3 VISUALIZATION OF SUBGROUP DISTRIBUTIONS

Data visualization methods have been part of statistics and data analysis research for many years. This research concentrated primarily on plotting one or

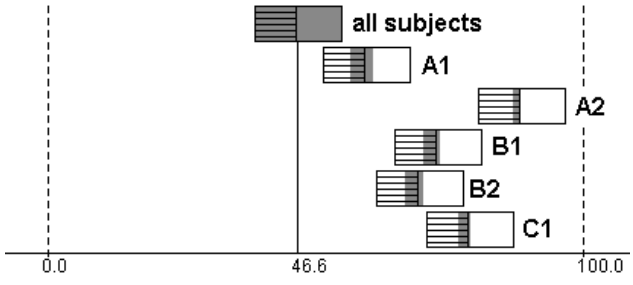

Fig. 1. Visualization by box plots. Each subgroup is represented in one box plot (all studied subjects are also considered one subgroup and are displayed in the top box). Each box shows the entire population. The gray area within each box indicates the respective subgroup. The overlap of the gray area with the hatched area shows the overlap of the group with the target (CHD). Hence, the farther to the left a gray area extends, the larger the overlap with the target (coverage). The lesser the gray area extends to the right of the hatched area, the more specific a subgroup is (less overlap with the non-target subjects). Finally, the location of the box along the X-axis indicates the relative share of the target CHD within each subgroup: the farther to the right a box is placed, the higher is the share of the target value within this subgroup. The line at 46.6% indicates default accuracy, i.e. the number of patients with CHD in the entire population.

more independent variables against a dependent variable in support of explorative data analysis [6, 8]. The visualization of analysis results, however, gained only recently some attention with the proliferation of data mining [1, 2, 5, 9]. This recent interest was spawned by the often overwhelming number and complexity of data mining results.

The visualization of analysis results primarily serves four purposes:

- better illustrate the model to the end user,
- utilize comparison of models,
- increase model acceptance, and
- enable model editing and support for "what-if questions".

The proposed novel visualization method can be used to visualize the output of any subgroup discovery algorithm, provided that the output has the form of rules with a target class in their consequent. It can also be used as a tool for visualizing standard classification rules. Its unique property is that it allows us to compare distributions of different 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 usually represents a class, or more precisely, the number of instances belonging to some target class. It must be noted that both directions of the Y-axis (Y^+ and Y^-) are used to indicate the number of instances. In Figure 2, for instance, the X-axis represents *age*, the Y^+-axis denotes class coronary heart disease (CHD) and Y^- denotes class

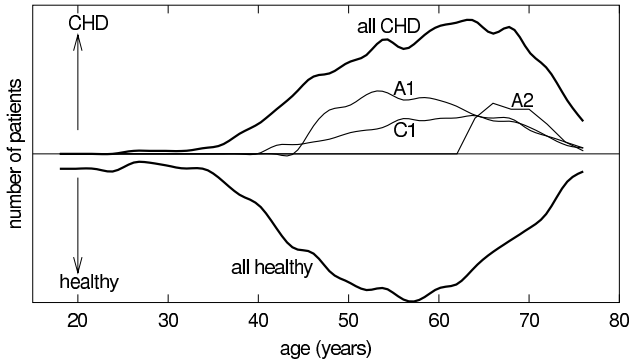

Fig. 2. Distribution of CHD patients and healthy subjects with respect to age in years. Graphs A1, A2, and C1 present corresponding model properties. Model A1 is for men, model A2 is for women, and model C1 represents patients with left ventricular hypertrophy. Healthy persons covered by models A1, A2, and C1 are not displayed.

non-CHD (or 'healthy'). Out of four graphs at the Y^+ side, three represent induced subgroups ($A1$, $A2$ and $C1$) of CHD patients, and the fourth shows the age distribution of the entire population of CHD (all CHD) patients. The graph at the Y^- side shows only the distribution of non-CHD (all healthy) patients in the training set. Note that the subgroups $A1$, $A2$ and $C1$ also cover some non-CHD patients, but the coverage of negative cases is not displayed for better viewing.

In general, it is not necessary that Y^+ and Y^- denote two opposite classes. If appropriate, they may denote any two classes, or even any two different attribute values, which the expert would like to compare.

Figures of this type can be drawn for any available numeric attribute and they are very valuable in the expert interpretation of the obtained results. For example, from Figure 2 it can be seen that there is no significant difference between CHD patients and healthy subjects in respect to their age, but that there are significant differences among detected models. From Figure 3 it can be noticed that there is a similar effect for the total cholesterol values although it is known that total cholesterol is an important risk factor for the CHD disease. This effect shows that the problem of CHD risk group detection typically can not be solved on the level of one feature and it demonstrates the importance of the descriptive induction methods which tries to describe models by a logical combination of a few correlated features. The advantage of the suggested visualization approach is that it makes such relations obvious. In this context, Figure 4 is interesting because it is different from the previous one. At first, it clearly demonstrates significant differences between all CHD and all healthy subjects in respect to ECG ST segment depression values, demonstrating that this measurement is an excellent disease indicator. But also, it shows that, although it is known that models A1-C1 cover various disease subpopulations, they behave very similarly in respect to the ECG ST segment depression property.

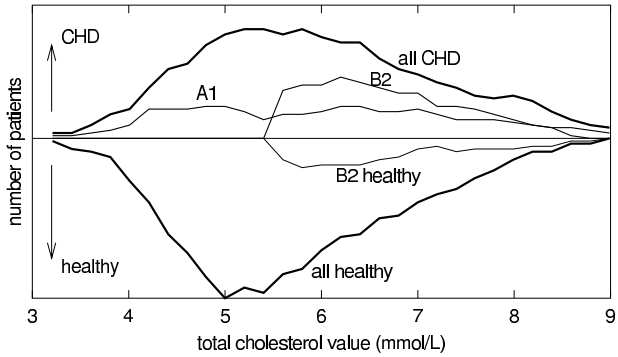

Fig. 3. Distribution of all CHD patients, CHD cases described by models A1 and B2, all healthy subjects, and healthy subjects erroneously included into model B2, with respect to total cholesterol value in $mmolL^{-1}$.

4 CONCLUSIONS

Subgroup visualization, described in this paper, allows us to compare distributions of different subgroups in terms of the selected attribute, plotted on the X-axis of the diagram. In medical domains we typically use the Y^+ side to represent the number of positive cases in order to reveal properties of induced models for subgroups of these patients. On the other hand, the Y^- side is reserved to reveal properties of these same models (or other models) for the negative cases. One of the advantages of using Y^+ and Y^- as proposed above is that in binary classification problems the comparison of the area under the graph of a subgroup and the graph of the entire population visualizes the fractions of $\frac{TP}{Pos} = \frac{TP}{TP+FN}$ at the Y^+ side (sensitivity TPr), and $\frac{FP}{Neg} = \frac{FP}{TN+FP}$ at the Y^- side (false alarm FPr), where Pos and Neg stand for the numbers of positive and negative cases in the entire population, respectively. For instance, in the visualization of subgroup $C1$ in Figure 4 the area under the thin line on the Y^- side represents the numbers of misclassified training instances of subgroup $C1$.

Acknowledgment

This work has been supported in part by the Croatian Ministry of Science and Technology, the Slovenian Ministry of Education, Science and Sport, and the EU funded project Data Mining and Decision Support for Business Competitiveness: A European Virtual Enterprize (IST-1999-11495). We are grateful to Goran Krstačić from the Institute for Cardiovascular Prevention Rehabilitation, Zagreb, Croatia for his involvement in the experiments in the CHD risk domain. The visualization presented in Figures 1 was developed by A. and G. Andrienko, AIS, FhG, Sankt Augustin, Germany.

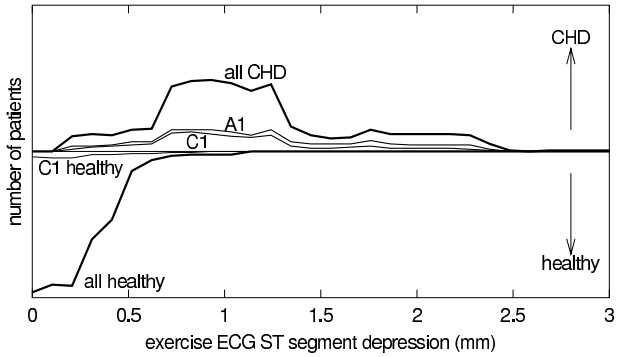

Fig. 4. Distribution of CHD patients and healthy subjects with respect to exercise ECG ST segment depression in millimeters (1mm corresponds to 0.1 mV). Large difference between total healthy and ill populations can be noticed, but differences among models are very small. Models A1 and C1 are selected as extreme cases. The thin line on the Y^- side represents the misclassified cases by subgroup $C1$.

References

1. Card, S.K., Mackinlay, J.D., & B. Shneidermann, B. (1999) Readings in information visualization. Morgan Kaufmann.
2. Fayyad, U.M., Grinstein, G.G., & Wierse, A. (2002) Information visualization in data mining and knowledge discovery. Morgan Kaufmann.
3. Gamberger, D. & Lavrač, N. (2002) Descriptive induction through subgroup discovery: a case study in a medical domain. In *Proc. of 19th International Conference on Machine Learning (ICML2002)*, Morgan Kaufmann, in press.
4. Jovanoski, V. & Lavrač, N. (2001) Classification Rule Learning with APRIORI-C. In *Proceedings of the Tenth Portuguese Conference on Artificial Intelligence, EPIA-2001*, Porto, Portugal, pp.44–51.
5. Keim, D.A & Kriegel, H.P. (1996) Visualization techniques for mining large databases: a comparison. In *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8:6, pp. 923–938.
6. Lee, H.Y, Ong, H.L., and Quek, L.H. (1995) Exploiting visualization in knowledge discovery. In *Proc. of the First Inter. Conference on Knowledge Discovery and Data Mining*, pp. 198–203.
7. Maron, D, Ridker, P.M., & Pearson, A.T. (1998) Risk factors and the prevention of coronary heart disease. In *A.R. Wayne, R.C. Schlant, V. Fuster : HURST'S: The Heart*, 1175-1195. McGrawc Hill, NY.
8. Unwin, A. (2000) Visualisation for data mining, http://www1.math.uni-augsburg.de/ unwin/
9. Workshop on visual data mining, PKDD 2001, Freiburg, Germany. http://www-staff.it.uts.edu.au/~simeon/vdm_pkdd2001/
10. Wrobel, S. (1997) An algorithm for multi-relational discovery of subgroups. In *Proc. First European Symposium on Principles of Data Mining and Knowledge Discovery*, pp.78–87, Springer.