Subgroup Visualization: A Method and Application in Population Screening

Dragan Gamberger[1] and **Nada Lavrač**[2] and **Dietrich Wettschereck**[3]

Abstract. This paper presents a method for the visualisation 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. Subgroup visualization is illustrated by graphs obtained for risk groups discovered in the problem of early detection of patient groups with risk for atherosclerotic coronary heart disease.

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 [1].

Some approaches to association rule induction can be used for subgroup discovery. For instance, the APRIORI-C algorithm [2], adapting the association rule induction algorithm APRIORI [3] 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. Similarly, the confirmation rule concept [4] used as a basis for the subgroup discovery algorithm whose results are used in this paper, utilizes the minimal support requirement as a measure which must be satisfied by every rule in order to be included in the induced confirmation rule set.

In this paper, subgroups were discovered by a new heuristic confirmation rule learning algorithm [5], adapting a confirmation rule learning algorithm [4] to subgroup discovery. One of the basic characteristics of the confirmation rule set induction concept is that separate rule sets are built for every target class. In a broader sense, the confirmation rule set concept introduces a decision model in which different rules can be incorporated: either rules induced by (one or more) inductive learning algorithms or even human encoded expert rules. If used for prediction, high predictive accuracy of such a decision model is expected only if, besides the high predictive accuracy of each individual rule, the whole set is as diverse as possible. As opposed to confirmation rules which cover only target class examples, the property of the heuristic confirmation rule learning algorithm, used for subgroup discovery in this paper, is that it enables the construction of rules that cover also a limited number of examples of the non-target class. 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. For details, please refer to [5].

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. The problem of early CHD detection is very important because CHD is one of the world's most frequent causes of mortality and a common problem in medical practice. The problem is known as a difficult one. Clinical studies have revealed plausible biological links between many risk factors and atherosclerosis. In addition, it was detected that coexistence of risk factors increases the disease rate. In many cases with significantly pathological test values (especially, for example, left ventricular hypertrophy, increased LDL cholesterol, decreased HDL cholesterol, hypertension, and intolerance glucose) the decision is not difficult. However, the problem of disease prevention is to decide in cases with slightly abnormal values and in cases when combinations of different risk factors occur. 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.

This paper presents our approach to subgroup visualisation (Section 2) and its application to visualise risk groups discovered in the problem of early detection of patient groups with risk for atherosclerotic coronary heart disease (Section 3). Some related work is outlined in Section 4.

2 SUBGROUP VISUALIZATION

The proposed 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.

In this section, subgroup visualization is illustrated by graphs obtained for risk groups discovered in the problem of early detection of patient groups with risk for atherosclerotic coronary heart disease.

Subgroup visualization, as described in this section, 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 1, for instance, the X-axis represents age, the Y^+-axis denotes class coronary heart disease (CHD) and Y^- denotes class non-CHD (or

[1] Rudjer Bošković Institute, Bijenička 64, 10000 Zagreb, Croatia
[2] J. Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[3] University of Applied Science, Bonn-Rhein-Sieg, 53757 Sankt Augustin, Germany

'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.

3 VISUALIZATION OF CHD RISK GROUPS

In this section, subgroup visualization is illustrated by graphs obtained for risk groups discovered in the problem of early detection of patient groups with risk for atherosclerotic coronary heart disease. Some interesting models of groups of CHD patients were constructed using the described methodology on the available patient data. There are three typical stages in the risk factor screening process, denoted by A, B, and C. Our goal was to construct at least one model for every stage.

Interesting properties of the induced models are presented in Figures 1–12 with respect to various attributes. For example, Figures 1 and 2 present patient distributions with respect to the patients' age for all five models, while Figure 4 uses exercise ST segment depression (used as a reliable sign of CHD) as the basis for model presentation. Each figure also shows the distribution of all CHD and healthy cases (thick lines) with respect to the selected base.

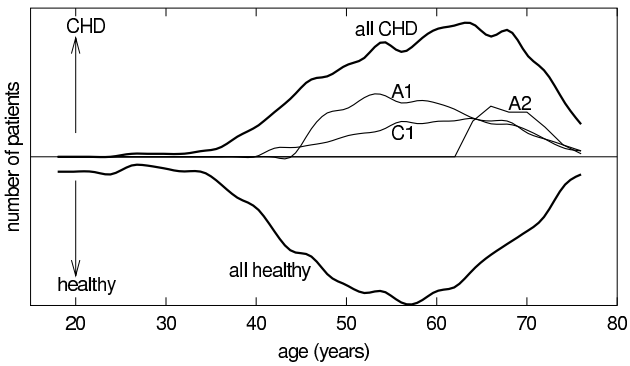

Figure 1. Distribution of CHD patients and healthy subjects with respect to age in years. Graphs A1, A2, and C1 present corresponding model properties. Model A1 is for men, model A2 is for women, and model C1 represents patients with left ventricular hypertrophy. About 60% of CHD patients detected by model C1 are also described by models A1 and A2. Healthy persons covered by models A1, A2, and C1 are not displayed.

At the first stage of patient examination (stage A, resulting in models A1 and A2), only anamnestic information and physical examination results are available. At this stage it was rather difficult to find models with a relatively small number of false positive predictions. The reason is a very restricted amount of available information about the patients. In order to make the problem easier, separate models were developed for male and female patient groups.

Model A1 for male:

CHD ← *positive family history* **AND** *age over 46 years*

Main supporting characteristic is psychosocial[4] stress, but cigarette smoking, hypertension, and overweight are also important. Both principal risk factors for this model are non-modifiable. The positive family history is a known important risk factor and it requires at least careful screening of other risk factors. The selected age margin in the second factor is rather low but it is in accordance with the existing medical experience. This low limit is good for prevention and early disease diagnosis although typical patients in this model are significantly older (Figure 1). The model has a rather high false positive rate as illustrated by the *A1 healthy* graph in the Y^- direction of Figure 8.

Model A2 for female:

CHD ← *body mass index over 25* kgm^{-2} *(typically 29)* **AND** *age over 63 years*

This simple model is very good for the female population with a sensitivity of about 50%. Supporting characteristics are positive family history and hypertension. Women in this risk group typically have slightly increased LDL cholesterol values and normal but decreased HDL cholesterol values. Body mass index over 25 (first principal risk factor) is exactly the generally accepted margin for overweight [6]. It

[4] *Principal* characteristics or risk factors are conditions of rules describing the subgroup. Positive family history and age over 46 years are principal risk factors for model A1. *Supporting* characteristics are determined by statistical significance analysis for the target population which consists of CHD patients correctly included into the model. The reference population are all the healthy subjects.

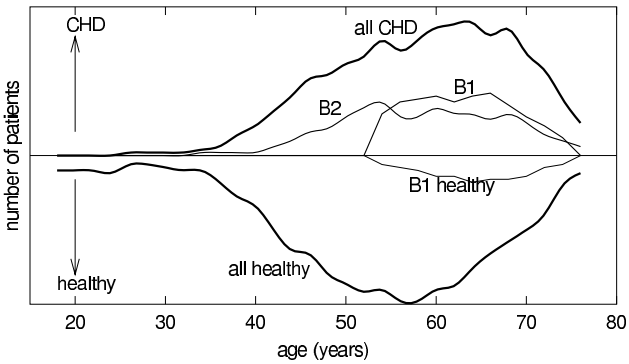

Figure 2. Same as Figure 1 but for models B1 and B2. Model B1 are elderly people with increased total cholesterol values while model B2 are patients with increased fibrinogen and total cholesterol values. The dashed line represents healthy people included into model B1. Models B1 and B2 have about 70% of patients in common.

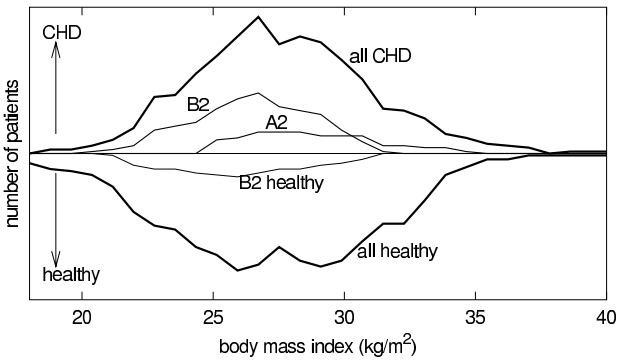

Figure 3. Distribution of CHD patients and healthy subjects with respect to the body mass index. Similarity in the distributions can be noticed. Models A2 and B2 have the greatest difference with respect to this risk factor. Overlapping of these two models is only about 25% .

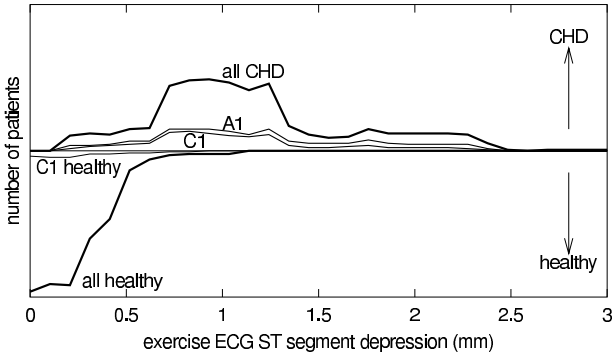

Figure 4. Distribution of CHD patients and healthy subjects with respect to exercise ECG ST segment depression in millimeters (1mm corresponds to 0.1 mV). Large difference between total healthy and ill populations can be noticed, but differences among models are very small. Models A1 and C1 are selected as extreme cases.

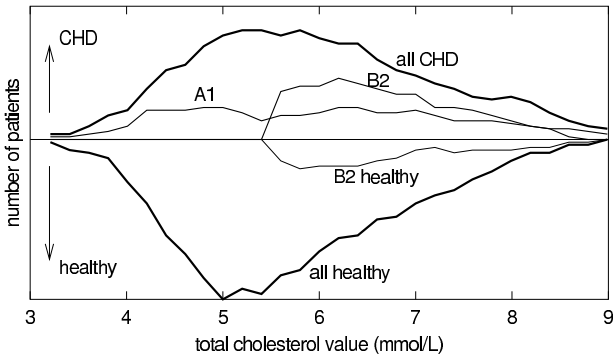

Figure 5. Distribution of CHD patients and healthy subjects with respect to total cholesterol value in $mmolL^{-1}$.

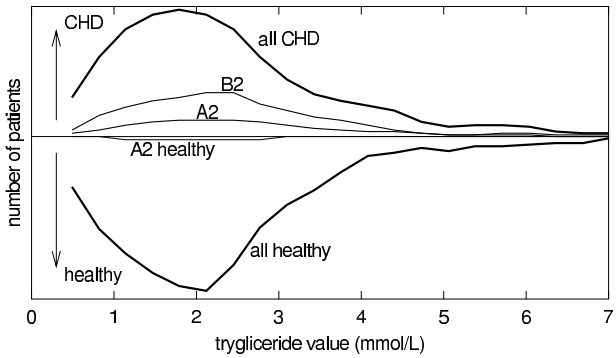

Figure 6. Distribution of CHD patients and healthy subjects with respect to trygliceride value in $mmolL^{-1}$.

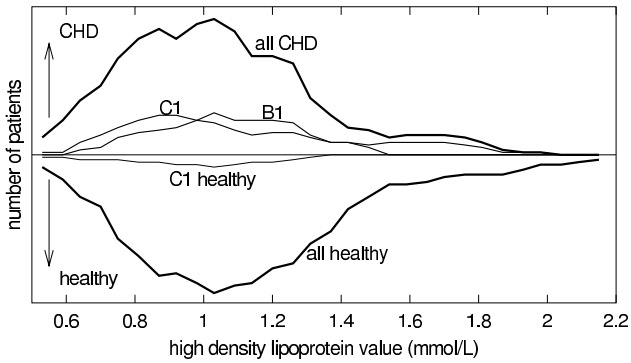

Figure 7. Distribution of CHD patients and healthy subjects with respect to high density lipoprotein value in $mmolL^{-1}$.

is well known that high body mass index strongly and positively correlates with the CHD rate. Typical values of the measured body mass index detected by CHD patients in this model are significantly over the margin of 25 (see Figure 3). Figures 6, 9, and 11 show dirtibutions of this model with respect to total cholesterol, uric acid, and left ventricular ejection fraction values, respectively. Figure 6 demonstrates in its Y^- part the very small false positive rate of this model.

Stage B includes basic laboratory tests in addition to the anamnestic and physical examination results of stage A. Two different models were induced for this stage. Potentially interesting is the first one which includes only total cholesterol (Figure 5) from the laboratory tests because this risk factor can be easily and inexpensively measured. The second model is a combination of two risk factors based on blood tests. It demonstrates that the detection of values close to the generally accepted normal values for the risk factors may also be significant for prevention and early CHD diagnosis.

Model B1:

CHD $\leftarrow$ *total cholesterol over 6.1* $mmolL^{-1}$ *(typically 7.2, normal 3.6 to 5.0)*[5] **AND** *age over 53 years*

This model is characteristic for the older part of the population (Figure 2), especially for women. The typical age of people in this risk group is 65 years for females and 61 years for males. The only supporting risk factor is increased triglycerides value which is more often detected for men. Interesting is also that typical members of this model do not have problems with overweight and hypertension. The false positive rate of this model is about 15% and is illustrated in Figures 2, 9, and 12. Distributions of high density lipoprotein values (Figure 7) and uric acid values (Figure 9) is similar to other models.

Model B2:

CHD $\leftarrow$ *total cholesterol over 5.6* $mmolL^{-1}$ *(typically 6.6, normal 3.6 to 5.0)* **AND** *fibrinogen over 3.7* $mmolL^{-1}$ *(typically 4.4, normal 2.0 to 3.7)*

This is a CHD patient model with similar properties for the male and female population. Typical patients do not have problems with overweight (Figure 3), hypertension and cigarette smoking but often have positive family history. Very high body mass index is contraindicated for this CHD patient model. Although the main model properties are similar for both genders, a representative female in this risk group is about 66 years old while a male is 10 years younger (Figure 2). This model highly correlates with low density lipoprotein values (Figure 8). Its sensitivity is good (about 30%) and the false positive rate is less than 15% as illustrated in Figures 5 and 11.

Level C additionally includes ECG resting test. One of the acceptable models with a relatively low false positive rate is model C1.

Model C1:

CHD $\leftarrow$ *left ventricular hypertrophy*

This model is important both for males and females older than 55 (Figure 1). Left ventricular hypertrophy is a well known risk factor which includes many other known CHD risk factors like hypertension and obesity. The main supporting risk factor detected for this model is positive family history. Often the patients in this CHD risk group have problems with hypertension and diabetes mellitus. Prac-

[5] Normal values between 3.6 and 5.0 are reported in [6].

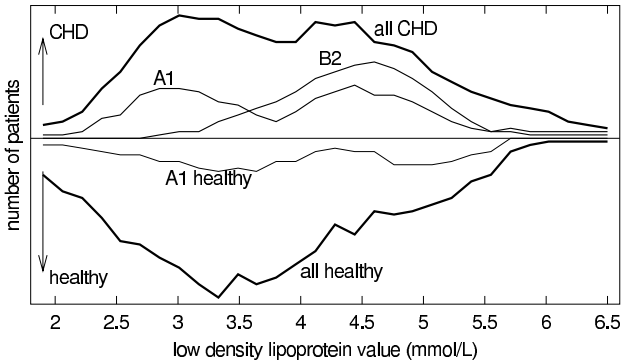

Figure 8. Distribution of CHD patients and healthy subjects with respect to low density lipoprotein value in $mmolL^{-1}$.

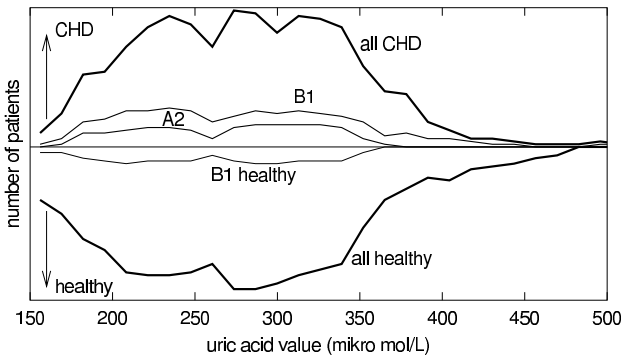

Figure 9. Distribution of CHD patients and healthy subjects with respect to uric acid value in $\mu molL^{-1}$.

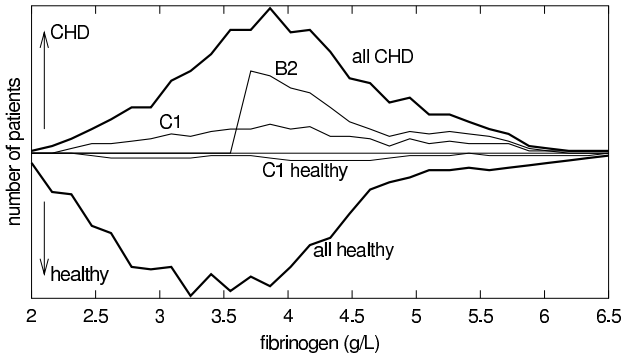

Figure 10. Distribution of CHD patients and healthy subjects with respect to fibrinogen value in gL^{-1}.

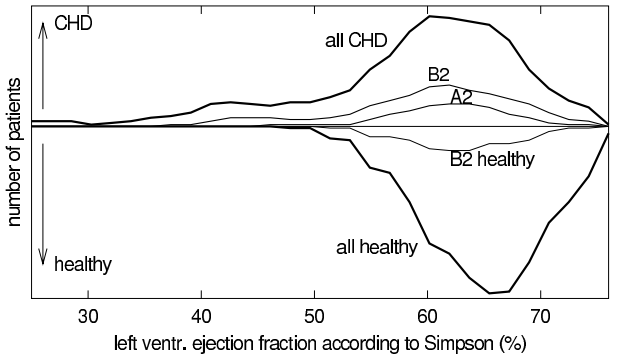

Figure 11. Distribution of CHD patients and healthy subjects with respect to left ventricular ejection fraction value in %.

tical importance of the model is that it has a relatively low false positive rate (Figures 4, 7, and 10) and that it does not correlate strongly with other previously described models (Figures 10 and 12).

4 BACKGROUND ON VISUALIZATION TECHNIQUES

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 [10,12]. The visualization of analysis results, however, gained only recently some attention with the proliferation of data mining [7,8,9,11]. 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".

Figures 13 and 14 are alternative means for displaying subgroup models A1, A2, B1, B2, and C1. Both figure types display identical 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 (Figure 14) was introduced. 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. These visualization techniques can serve as an entry point to the more in depth visualization technique presented in this paper.

5 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 (CHD patients, in this paper) 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 (patients without CHD). For instance, in the graph of subgroup $B1$ shown in Figure 2, the dashed line at the Y^- side represents the distribution of non-CHD subjects in subgroup $B1$.

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

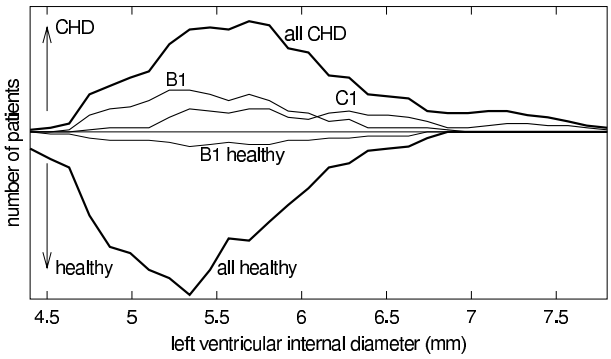

Figure 12. Distribution of CHD patients and healthy subjects with respect to left ventricular internal diameter value in mm.

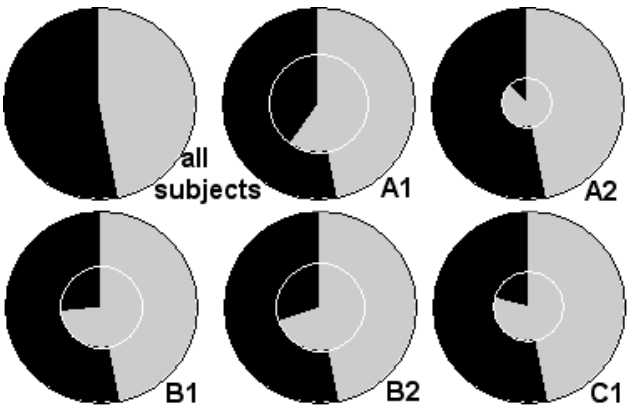

Figure 13. The distribution of CHD patients and healthy subjects for all studied subjects (top left) and for models A1 through C1. Gray pie chart slices show the number of patients with CHD. Black slices show the number of healthy people. Each inner pie shows the distribution of CHD patients and healthy people within the respective subgroup. The outer frame shows, for comparison, the distribution within the entire population. The area of the inner pie is proportional to the relative size of the subgroup.

Pos and Neg stand for the numbers of positive and negative cases in the entire population, respectively. For instance, in the visualization of subgroup $B1$ in Figure 2 the area under the dashed line on the Y^- side represents the numbers of misclassified training instances of subgroup $B1$.

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 visualizations presented in Figures 13 and 14 were developed by A. and G. Andrienko, AIS, FhG, Sankt Augustin, Germany.

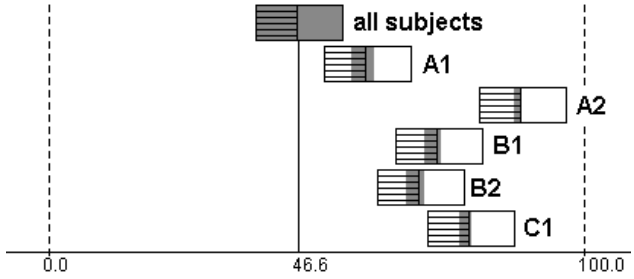

Figure 14. Alternative visualization by box plots. Each subgroup is represented in one box plot (all studied subjects are also considered one subgroup and are displayed in the top box). Each box shows the entire population. The gray area within each box indicates the respective subgroup. The overlap 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.

REFERENCES

1 S. Wrobel, An algorithm for multi-relational discovery of subgroups. *Proc. First European Symposium on Principles of Data Mining and Knowledge Discovery*, 78–87, Springer, (1997).

2 V. Jovanoski and N. Lavrač, Classification Rule Learning with APRIORI-C. *Proceedings of the Tenth Portuguese Conference on Artificial Intelligence, EPIA-2001*, Porto, Portugal, 44–51, (2001).

3 R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A.I. Verkamo, Fast discovery of association rules. In U.M. Fayyad, G. Piatetski-Shapiro, P. Smyth and R. Uthurusamy (Eds.) *Advances in Knowledge Discovery and Data Mining*, pp. 307–328. AAAI Press, (1996).

4 D. Gamberger and N. Lavrač, Confirmation rule sets. In *Proc. of 4th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD2000)*, pp.34–43, (2000).

5 D. Gamberger and N. Lavrač, 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.(2002)

6 D. Maron, P.M. Ridker, and A.T. Pearson, Risk factors and the prevention of coronary heart disease. In *A.R. Wayne, R.C. Schlant, and V. Fuster : HURST'S: The Heart*, 1175-1195. McGrawc Hill, NY, (1998).

7 S.K. Card, J.D. Mackinlay, and B. Shneidermann, Readings in information visualization. Morgan Kaufmann, (1999).

8 U.M. Fayyad, G.G. Grinstein, and A. Wierse, Information visualization in data mining and knowledge discovery. Morgan Kaufmann, (2002).

9 D.A. Keim and H.P. Kriegel, Visualization techniques for mining large databases: a comparison. In *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8:6, pp. 923-938, (1996).

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

11 Workshop on visual data mining, PKDD 2001, Freiburg, Germany, (2001). http://www-staff.it.uts.edu.au/~simeon/vdm_pkdd2001/

12 A. Unwin, Visualisation for data mining, (2000). http://www1.math.uni-augsburg.de/ unwin/