Generating Actionable Knowledge by Expert-Guided Subgroup Discovery

Dragan Gamberger[1] and Nada Lavrač[2]

[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

Abstract. This paper discusses actionable knowledge generation. Actionable knowledge is explicit symbolic knowledge, typically presented in the form of rules, that allows the decision maker to recognize some important relations and to perform an action, such as targeting a direct marketing campaign, or planning a population screening campaign aimed at targeting individuals with high disease risk. The disadvantages of using standard classification rule learning for this task are discussed, and a subgroup discovery approach proposed. This approach uses a novel definition of rule quality which is extensively discussed.

1 Introduction

In KDD one can distinguish between *predictive* and *descriptive* induction tasks. Classification rule learning [2,10] is a form of *predictive induction*. The distinguishing feature of predictive induction is the input data formed of labeled training examples (with class assigned to each training instance), and the output aimed at solving classification and prediction tasks. This paper provides arguments for actionable knowledge generation through recently developed *descriptive induction* approaches. These involve mining of association rules (e.g., APRIORI [1]), subgroup discovery (e.g., MIDOS [16]), and other approaches to non-classificatory induction. In this work we are particularly interested in subgroup discovery, where 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.

Actionable knowledge is explicit symbolic knowledge that allows the decision maker to perform an action, such as, for instance, select customers for a direct marketing campaign, or select individuals for population screening concerning high disease risk. The term *actionability* [14] denotes a *subjective* measure of

T. Elomaa et al. (Eds.): PKDD, LNAI 2431, pp. 163–175, 2002.

interestingness of a discovered pattern: "a pattern is interesting if the user can do something with it to his or her advantage" [13,14].[1]

This paper presents some shortcomings of actionable knowledge generation through predictive induction and proposes an approach to expert-guided knowledge discovery, where the induction task is to detect different, potentially important subgroups among which the expert will be able to select the patterns which are actionable. The paper is organized as follows. Section 2 discusses the types of induced knowledge and shortcomings of standard classification rule learning for actionable knowledge generation. Section 3 presents the advantages of subgroup discovery approaches for the formation of actionable knowledge, and proposes an approach to subgroup discovery, developed by adapting an existing confirmation rule learning algorithm. We conclude with some experimental evaluation results in Section 4 and lessons learned in Section 5.

2 Shortcomings of Classification Rule Learning for Actionable Knowledge Generation

In symbolic predictive induction, two most common approaches are rule learning and decision tree learning. The goal of rule learning is to generate separate models, one for each class, inducing class characteristics in terms of class properties occurring in the descriptions of training examples. Classification rule learning results in *characteristic descriptions*, usually generated separately for each class by repeatedly applying the covering algorithm. In decision tree learning, on the other hand, the rules which can be formed of paths leading from the root node to class labels in the leaves represent *discriminating descriptions*, formed of properties that best discriminate between the classes. Hence, classification rules serve two different purposes: characterization and discrimination.

An open question, discussed in this paper, is whether the knowledge induced by rule learning and decision tree learning is actionable in medical and marketing applications, outlined in this paper, whose goal is to uncover the properties of subgroups of the population which can guide a decision maker in directing some targeted campaign. The motivation for this work comes from two applications

- A medical problem of population screening aimed at spotting the individuals in a town or region with high risk for Coronary Heart Disease (CHD) [5]. In this application, the hard problem is to find suspect CHD cases with slightly abnormal values of risk parameters and in cases when combinations of different risk factors occur. The risk group models should help general practitioners to recognize CHD and/or to detect the illness even before the first symptoms actually occur.
- A marketing problem of direct mailing aimed at spotting potential customers of a certain product [3]. In this application. the problem is to select subgroups of potential customers that can be targeted by an advertising campaign. The

[1] The other subjective measure introduced in [14] is unexpectedness: "a pattern is interesting to the user if it is surprizing to the user".

specific task is to find significant characteristics of customer subgroups who do not know a brand, relative to the characteristics of the population that recognizes the brand.

We argue that for such and similar tasks the models induced through classification rule learning and decision tree learning are not actionable. Besides subjective reasons [14] that can be due to the inappropriate choice of parameters used in induced descriptions, some objective reasons for the non-actionability of induced patterns that are due to the method used are listed below:

- Classification rules and decision trees could be used to classify all individuals of a selected population, but this is unpractical and virtually impossible.
- Rules formed of decision tree paths are discriminant descriptions, hence they are not actionable for the above tasks.
- Classification rules forming characteristic descriptions are intuitively expected to be actionable. However, the fact that they have been generated by a covering algorithm (used in AQ [10], CN2 [2], and most other rule learners) hinders their actionability. Only first few rules induced by a covering algorithm may be of interest as subgroup descriptions with sufficient coverage. Subsequent rules are induced from smaller and strongly biased example subsets, e.g., subsets including only positive examples not covered by previously induced rules. This bias prevents a covering algorithm to induce descriptions uncovering significant subgroup properties of the entire population.

A deeper analysis of the reasons for the non-actionability of patterns induced by decision tree and classification rule induction can be found in [9]. Our approach to dealing with the above deficiencies is described in this paper, proposing an approach to actionable knowledge generation where the goal is to uncover properties of individuals for actions like population screening or targeting a marketing campaign. For such tasks, actionable rules are characterized by high coverage (support), as well as high sensitivity and specificity[2], even if this can achieved only at a price of lower classification accuracy, which is a quality to be optimized in classification/prediction tasks.

3 Actionable Knowledge Generation through Subgroup Discovery

Subgroup discovery has the potential for inducing actionable knowledge to be used by a decision maker. The approach described in this paper is an approach to

[2] *Sensitivity* measures the fraction of positive cases that are classified as positive, whereas *specificity* measures the fraction of negative cases classified as negative. If TP denotes true positives, TN true negatives, FP false positives, FN false negatives, Pos all positives, and Neg all negatives, then $Sensitivity = TPr = \frac{TP}{TP+FN} = \frac{TP}{Pos}$, and $Specificity = \frac{TN}{TN+FP} = \frac{TN}{Neg}$, and $FalseAlarm = FPr = 1 - Specificity = \frac{FP}{TN+FP} = \frac{FP}{Neg}$. Quality measures in association rule learning are *support* and *confidence*: $Support = \frac{TP}{Pos+Neg}$ and $Confidence = \frac{TP}{TP+FP}$.

descriptive induction, but the underlying methodology uses elements and techniques from *predictive induction*. By basing the induction on labeled training instances, the induction process can be targeted to uncovering properties forming actionable knowledge. On the other hand, the standard assumptions like "induced rules should be as distinct as possible, covering different parts of the population" (which is the case in decision tree learning, as well as in rule learning using the covering algorithm) need to be relaxed; this enables the discovery of intersecting subgroups with high coverage/support, describing some population segments in a multiplicity of ways. This knowledge is redundant, if viewed purely from a classifier perspective, but extremely valuable in terms of its descriptive power, uncovering genuine properties of subpopulations from different viewpoints.

3.1 Algorithm *SD* for Subgroup Discovery

Algorithm SD is outlined in Figure 1. The algorithm is used in the Data Mining Server available on-line at `http://dms.irb.hr` and the reader can test it there. The algorithm assumes that the user selects one class as a *target class*, and learns subgroup descriptions of the form $TargetClass \leftarrow Cond$, where $Cond$ is a conjunction of features. The result is a set of best rules, induced using a heuristic beam search algorithm that allows for the induction of relatively general rules which may cover also some non-target class examples.

The aim of this heuristic rule learning algorithm is the search for rules with a maximal q value, where q is computed using the user-defined $TP/FP-tradeoff$ function. This function defines a tradeoff between true positives TP and false positives FP (see also Footnote 2). By searching for rules with high quality q, this algorithm tries to find rules that cover many examples of the target class and a low number of non-target examples. By changing a parameter of the tradeoff function the user can obtain rules of variable generality.

Typically, *Algorithm SD* can generate many rules of high quality q satisfying the requested condition of a minimal number of covered target class examples, defined by the *min_support* parameter. Accepting all these rules as actionable knowledge is generally not desired. A solution to this problem is to select a relatively small number of rules which are as diverse as possible. The algorithm implemented in the confirmation rule set concept [4] accepts as diverse those rules that cover diverse sets of target class examples. The approach cannot guarantee statistical independence of the selected rules, but ensures the diversity of generated models. Application of this algorithm is suggested for postprocessing of detected subgroups.

3.2 Rule Quality Measures for Subgroup Discovery

Various rule evaluation measures and heuristics have been studied for subgroup discovery [7,16], aimed at balancing the size of a group (referred to as factor g in [7]) with its distributional unusualness (referred to as factor p). The properties of functions that combine these two factors have been extensively studied

Algorithm SD: Subgroup Discovery
Input: $E = P \cup N$ (E training set, P positive (target class) examples,
N negative (non-target class) examples)
L set of all defined features (attribute values), $l \in L$
$rule\_quality$ (user-defined $TP/FP - tradeoff$ function)
$min\_support$ (minimal support for rule acceptance)
$beam\_width$ (number of rules in the beam)
Output: $S = \{TargetClass \leftarrow Cond\}$
(set of rules formed of $beam\_width$ best conditions $Cond$)
(1) **for** all rules in the beam ($i = 1$ to $beam\_width$) **do**
initialize condition part of the rule to be empty, $Cond(i) \leftarrow \{\}$
initialize rule quality, $q(i) \leftarrow 0$
(2) **while** there are improvements in the beam **do**
(3) **for** all rules in the beam ($i = 1$ to $beam\_width$) **do**
(4) **for** all $l \in L$ **do**
(5) form a new rule by forming a new condition as a conjunction of the condition from the beam and feature l, $Cond(i) \leftarrow Cond(i) \wedge l$
(6) compute rule quality q defined by the $TP/FP - tradeoff$ function
(7) **if** $TP \geq min\_support$ **and** q is larger than any $q(i)$ in the beam **do**
(8) replace the worst rule in the beam with the new rule and reorder the rules with respect to their quality
(9) **end for** features
(10) **end for** rules from the beam
(11) **end while**

Fig. 1. Heuristic beam search rule construction algorithm for subgroup discovery

(the "p-g-space", [7]). Similarly, the weighted relative accuracy heuristic, used in [15], trades off generality of the rule ($p(Cond)$, i.e., rule coverage) and relative accuracy ($p(Class|Cond) - p(Cond)$).

In contrast with the above measures, in which the generality of a rule is used in the generality/unusualness or generality/relative-accuracy tradeoff, the measure used in *Algorithm SD* is aimed to enable expert guided subgroup discovery in the TP/FP space, in which FP (plotted on the X-axis) needs to be minimized, and TP (plotted on the Y-axis) needs to be maximized. The TP/FP space is similar to the ROC (Receiver Operating Characteristic) space [11] in which a point in the ROC space shows classifier performance in terms of false alarm or *false positive rate* $FPr = \frac{FP}{TN+FP}$ (plotted on the X-axis) that needs to be minimized, and sensitivity or *true positive rate* $TPr = \frac{TP}{TP+FN}$ (plotted on the Y-axis) that needs to be maximized. In the ROC space, an appropriate tradeoff, determined by the expert, can be achieved by applying different algorithms, as well as by different parameter settings of a selected mining algorithm. The ROC space and the TP/FP space are equivalent if a single problem is being analysed: in the ROC space the results are evaluated based on the TPr/FPr

tradeoff, and in the TP/FP space based on the TP/FP tradeoff - the "rate" is just a normalising factor enabling us intra-domain comparisons.

It is well known from the ROC analysis, that in order to achieve the best results, the discovered rules should be as close as possible to the top-left corner of the ROC space. This means that in the TPr/FPr tradeoff, TPr should be as large as possible, and FPr as small as possible. Similarly, in the TP/FP space, TP should be as large as possible, and FP as small as possible.

For marketing problems, for instance, we have learned that intuitions like "how expensive is every FP prediction in terms of additional TP's that should be covered by the rule" are useful for understanding the problem and directing the search. Suppose that some cost parameter c is defined that says: "For every additional FP, the rule should cover more than c additional TP examples in order to be better." Based on this reasoning, we can define a quality measure q_c, using the following TP/FP tradeoff: $q_c = TP - c * FP$. Quality measure q_c is easy to use because of the intuitive interpretation of parameter c. It also has a nice property for subgroup discovery: by changing the c value we can move in the TP/FP space and select the optimal point based on parameter c.

Consider a different quality measure q_g, using another TP/FP tradeoff: $q_g = TP/(FP + g)$. This quality measure is actually used in *Algorithm SD* for the evaluation of different rules in the TP/FP space, as well as for heuristic construction of interesting rules. Below we explain why this quality measure has been selected, and not some other more intuitive quality measure like the q_c measure defined above.

3.3 Analysis of the q_g Quality Measure

The selected quality measure q_g and generalization parameter g used in it, enable that by changing parameter g, different optimal points (rules) in the TP/FP space can be selected as the final solution. Although large g means that more general solutions can be expected, sometimes we would like to know in advance

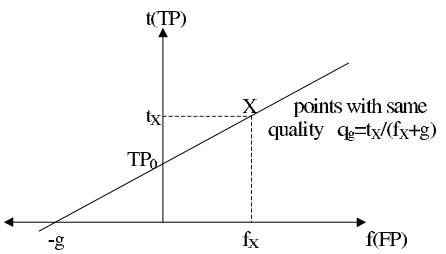

Fig. 2. Properties of quality q_g

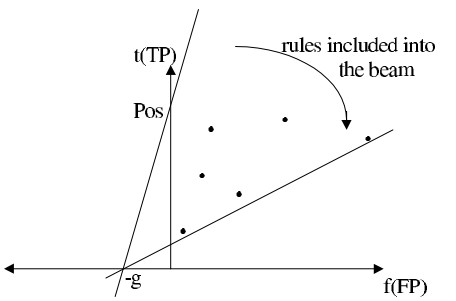

Fig. 3. Rules with highest quality included into the beam for $q_g = TP/(FP + g)$

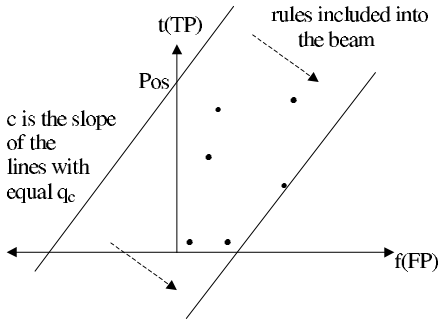

Fig. 4. Rules with highest quality included in the beam for $q_c = TP - c * FP$

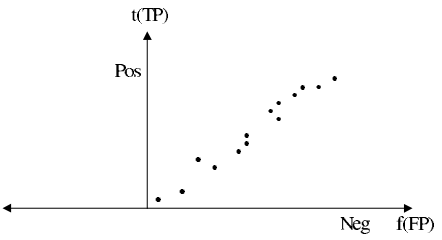

Fig. 5. Placement of interesting features in the TP/FP space after the first iteration

what properties of the selected rule can be expected for the selected g value, or, stated alternatively, by determining the desired properties of the rule under construction, what parameter value g should we select.

In *Algorithm SD*, increased generality (increasing g means moving to the right in the TP/FP space) results in more general subgroups discovered, covering more instances. If the value of g is low (1 or less) then covering of any non-target instance is made relatively very expensive and the final result are rules that cover only few target cases but also nearly no non-target class cases. This results in rules with high specificity (high confidence or low false alarm rate). If the value of parameter g is high (10 or higher) then covering of few non-target examples is not so expensive and more general rules can be generated. This approach is very appropriate for domains in which false positive predictions are not very expensive, like risk group detection in medical problems or detection of interesting customer groups in marketing, in which 'pure' rules would have too low coverage, making them unactionable.

If the algorithm employs exhaustive search (or if all points in the TP/FP space are known in advance) then there is no difference between the two measures q_g and q_c. Any of the two could be used for selecting the optimal point, only the values that must be selected for parameters g and c would be different. In this case, q_c might be even better because its interpretation is more intuitive.

However, since *Algorithm SD* is a heuristic beam search algorithm, the situation is different. Subgroup discovery is an iterative process, performing one or more iterations (typically 2–5) until good rules are constructed by forming conjunctions of features in the rule body. In this process, a rule quality measure is used for rule selection (for which the two measures q_g and q_c are equivalent) as well as for the selection of features and their conjunctions that have high potential for the construction of high quality rules in subsequent iterations; for this use, rule quality measure q_g is better than q_c. Let us explain why.

Suppose that we have a point (a rule) x in the TP/FP space, where t_x is its TP value and f_x its FP value, respectively. For a selected g value, q_g can be

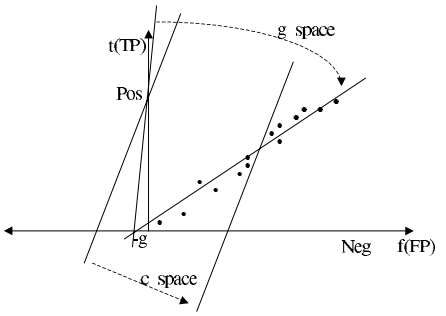

Fig. 6. The quality q_c employing the c parameter tends to select points with small TP values, while quality q_g employing the g parameter will include also many points with large TP values (from the right part of the TP/FP space) that have a chance to take part in building conjunctions of high quality rules

determined for this point x. It can be shown that all points that have the same quality q_g as the point (rule) x lie on a line defined by the following function:

$$t = \frac{t_x * f}{f_x + g} + \frac{t_x * g}{f_x + g} = \frac{t_x * (f + g)}{f_x + g}.$$

In this function, t represents the TP value of the rule with quality q_g which covers exactly $f = FP$ negative examples. By selecting different f's, corresponding t's can be determined by this function.

The line, determined by this function, crosses the $t(TP)$ line at point $t = t_x * g/(f_x + g)$ and the $f(FP)$ line at point $f = -g$. This is shown in Figure 2. The slope of this line is equal to the quality of point X, which equals $t_x/(f_x + g)$.

In the TP/FP space, points with higher quality than q_g are above this line, in the direction of the upper left corner. Notice that in the TP/FP space the top-left is the preferred part of the space: points in that part represent rules with the best TP/FP tradeoff. This reasoning indicates that points that will be included in the beam must all lie above the line of equal weights q_{beam} which is defined by the last point (rule) in the beam.

If represented graphically, first *beam_width* number of rules, found in the TP/FP space when rotating the line from point $(0, Pos)$ in the clockwise direction, will be included in the beam. The center of rotation is point $(-g, 0)$. This is illustrated in Figure 3. On the other hand, for the q_c quality measure defined by $q_c = TP - c * FP$ the situation is similar but not identical. Again points with same quality lie on a line, but its slope is constant and equal to c. Points with higher quality lie above the line in the direction of the left upper corner. The points that will be included into the beam are the first *beam_width* points in the TP/FP space found by a *parallel* movement of the line with slope c, starting from point $(0, Pos)$ in the direction towards the lower right corner. This is illustrated in Figure 4.

Let us now assume that we are looking for an optimal rule which is very specific. In this case, parameter c will have a high value while parameter g will have a very small value. The intention is to find the same optimal rule in the TP/FP space. At the first level of rule construction only single features are considered and most probably their quality as the final solution is rather poor.

See Figure 5 for a typical placement of potentially interesting features in the TP/FP space.

The primary function of these features is to be good building blocks so that by conjunctively adding other features, high quality rules can be constructed. By adding conjunctions, solutions generally move in the direction of the left lower corner. The reason is that conjunctions can reduce the number of FP predictions, but they reduce the number of TP's as well. Consequently, by conjunctively adding features to rules that are already in the left lower corner, the algorithm will not be able to find their specializations nearer to the left upper corner. Only the rules that have high TP value, and are in the right part of the TP/FP space, have a chance to take part in the construction of interesting new rules.

Figure 6 illustrates the main difference between quality measures q_g and q_c: the former tends to select more general features from the right upper part of the TP/FP space (points in the so-called 'g space'), while the later 'prefers' specific features from the left lower corner (points in the so-called 'c space'). In cases when c is very large and g is very small, the effect can be so important that it may prevent the algorithm from finding the optimal solution even with a large beam width. Notice, however, that *Algorithm SD* is heuristic in its nature and no statements are true for all cases. This means that in some, but very rare cases, the quality based on parameter c may result in a better final solution.

4 Experimental Evaluation

We have verified the claimed properties of the proposed rule quality measure in the medical Coronary Heart Disease (CHD) problem [5]. The task is the detection of subgroups which can be used as risk group models. The domain includes three levels of descriptors (basic level A with 10, level B with 16, and level C with 21 descriptors) and the results of subgroup discovery are five models (A1 and A2 for level A, B1 and B2 for level B, and C1 for level C), presented in [5]. Algorithm SD with the q_g measure was used for subgroup detection, with the goal of detecting different, potentially relevant subgroups. The algorithm was used iteratively many times with different g values. In each iteration few best solutions from the beam were shown to the domain expert. The selection of subgroups which will be used as model descriptions was based on the expert knowledge. The position of the expert selected subgroups in the TP/FP space is presented in Figures 7–9. It can be noticed that the selected subgroups do not lie on the ROC curves: this means that expert-selected actionability properties of subgroups were more important than the optimization of their TP/FP tradeoff.

For the purpose of comparing the q_g and q_c measures we have constructed one ROC curve for each of the two measures. The procedure was repeated for all levels A–C. The ROC curve for the q_g measure was constructed so that for g values between 1 and 100 the best subgroups lying on the convex hull in the TP/FP space were selected: this results is the thick lines in Figures 7–9. The thin lines represent ROC curves obtained for subgroups induced by the q_c measure for c values between 0.1 and 50.

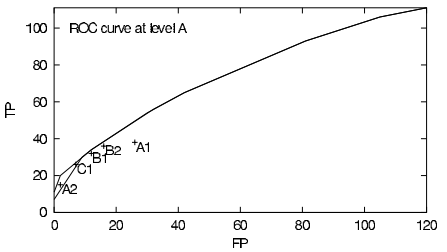

Fig. 7. TP/FP space presenting the ROC curves of subgroups induced using quality measures $q_g = TP/(FP + g)$ (thick line) and $q_c = TP - c * FP$ (thin line) at data level A. Labels A1–C1 denote positions of subgroups selected by the medical expert as interesting risk group descriptions [8,5]

Figure 7 for level A demonstrates that both curves agree in the largest part of the TP/FP space, but that for small FP values the q_g measure is able to find subgroups covering more positive examples. According to the analysis in the previous section, this was the expected result. In order to make the difference more obvious, for levels B and C, only the left part of the TP/FP space is shown in Figures 8 and 9. Similar curve properties can be noticed for different data sets.

The differences between the ROC curves for q_g and q_c measures may seem small and insignificant, but in reality it is not so. The majority of interesting subgroups (this claim is supported also by models A1–C1 selected by the domain expert) are subgroups with a small false positive rate which lie in the range in which q_g works better. In addition, for subgroups with $FP = 0$ the true positive rate in our examples was about two times larger for subgroups induced with q_g than with q_c. Furthermore, note that for levels A and B there are two out of five subgroups (A2 and C1) which lie in the gap between the ROC curves. If the q_c measure instead of q_g measure were used in the experiments described in [5], at least subgroup A2 could not have been detected.

5 Conclusions and Lessons Learned

This work describes actionable knowledge generation in the descriptive induction framework, pointing out the importance of effective expert-guided subgroup discovery in the TP/FP space. Its main advantages are the possibility to induce knowledge with different generalization levels (achieved by tuning the g parameter of the subgroup discovery algorithm) and the measure that ensures high quality rules also in the heuristic environment. In addition, the paper argues that expert's involvement in the induction process is substantial for successful actionable knowledge generation.

The presented methodology has been applied to different medical and marketing domains. In the medical problem of detecting and describing of Coronary Heart Disease risk groups we have learned a few important lessons. The main is that in this type of problem, there are no predefined specificity or sensitivity levels to be satisfied. The actionability of induced models, based on the detected subgroups, largely depends on the applied subgroup discovery method, but also on (a) whether the attributes used in the induced model can be easily and reliably measured, and (b) how interesting/unexpected are the subgroup descriptions in

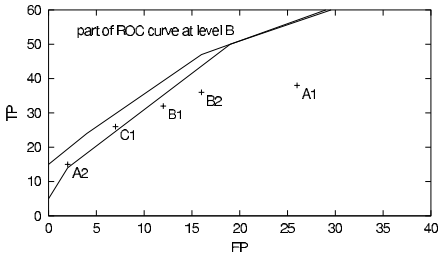

Fig. 8. The left part of the ROC curves representing subgroups induced at data level B

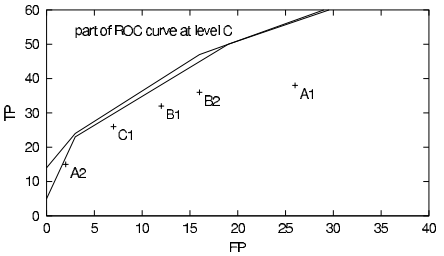

Fig. 9. The left part of the ROC curves representing subgroups induced at data level C

the given population. Evaluation of such properties is completely based on expert knowledge and the success of the search depends on expert involvement. The aim of machine learning based subgroup detection described in this work is thus to enable the domain expert to effectively search the hypothesis space, ranging from very specific to very general models.

In the marketing problems where the task is to find significant characteristics of customer subgroups who do not know a brand compared to the characteristics of the population that recognizes the brand, the main lesson learned is that the ROC space is very appropriate for the comparison of induced models. Only subgroups lying on the convex hull may be optimal solutions and all other subgroups can be immediately discarded. When concrete parameters of the mailing campaign are known, like marginal cost per mailing and the size of the population, they define the slope of the lines with equal profit in the ROC space. Movements in the ROC space along these lines will not change the amount of total profit while movements upward or downward will increase or decrease the profit, respectively. The optimal subgroup in a concrete marketing situation is the point on the convex hull which has an equal profit line as its tangent. In terms of actionability, however, the appropriate parameters for subgroup discovery need to be determined in data preprocessing.

Acknowledgements

This work was supported by the the Croatian Ministry of Science and Technology, Slovenian Ministry of Education, Science and Sport, and the EU funded project Data Mining and Decision Support for Business Competitiveness: A European Virtual Enterprise (IST-1999-11495). We are grateful to Miro Kline, Bojan Cestnik, and Peter Flach for their collaboration in marketing domains, and to Goran Krstačić for his collaboration in the experiments in coronary heart disease risk group detection.

References

1. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo, A. I. (1996) 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. 163
2. Clark, P. & Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning*, 3(4):261–283. 163, 165
3. Flach, P. & Gamberger, D. (2001) Subgroup evaluation and decision support for direct mailing marketing problem. *Integrating Aspects of Data Mining, Decision Support and Meta-Learning* Workshop at ECML/PKDD 2001 Conference. 164
4. Gamberger, D. & Lavrač, N. (2000) Confirmation rule sets. In *Proc. of 4th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD2000)*, pp.34–43, Springer. 166
5. 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. 164, 171, 172
6. Kagan, T. & Ghosh, J. (1996) Error correlation and error reduction in ensemble classifiers. *Connection Science*, 8, 385–404.
7. Klösgen, W. (1996) Explora: A multipattern and multistrategy discovery assistant. In U. M. Fayyad, G. Piatetski-Shapiro, P. Smyth and R. Uthurusamy (Eds.) *Advances in Knowledge Discovery and Data Mining*, pp. 249–271. MIT Press. 166, 167
8. Krstačić, G., Gamberger, D., & Šmuc, T. (2001) Coronary heart disease patient models based on inductive machine learning. In *Proc. of 8th Conference on Artificial Intelligence in Medicine in Europe (AIME 2001)*, pp.113–116. 172
9. Lavrač, N., Gamberger, D., & Flach, P. (2002) Subgroup discovery for actionable knowledge generation: Defiences of classification rule learning and lessons learned. *Data Mining Lessons Learned* Workshop at ICML 2002 Conference, to be printed. 165
10. Michalski, R. S., Mozetič, I., Hong, J., & Lavrač, N. (1986) The multi-purpose incremental learning system AQ15 and its testing application on three medical domains. In *Proc. Fifth National Conference on Artificial Intelligence*, pp. 1041–1045, Morgan Kaufmann. 163, 165
11. Provost, F. & Fawcett, T. (2001) Robust classification for imprecise environments. *Machine Learning*, 42(3), 203–231. 167
12. Rivest, R. L. & Sloan, R. (1988) Learning complicated concepts reliably and usefully. In *Proc. Workshop on Computational Learning Theory*, 69–79, Morgan Kaufman.
13. Piatetshy-Shapiro, G. & Matheus, C. J. (1994) The interestingness of deviation. In *Proc. of the AAAI-94 Workshop on Knowledge Discovery in Databases*, pp. 25–36. 164
14. Silberschatz, A. & Tuzhilin, A. (1995) On Subjective Measure of Interestingness in Knowledge Discovery. In *Proc. First International Conference on Knowledge Discovery and Data Mining (KDD)*, 275-281. 163, 164, 165
15. Todorovski, L., Flach, P., & Lavrač, N. (2000) Predictive Performance of Weighted Relative Accuracy. In Zighed, D. A., Komorowski, J. and Zytkow, J., editors, *Proc. of the 4th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD2000)*, Springer-Verlag, 255–264. 167

16. 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. 163, 166