Rule-Based Estimation of Attribute Relevance

Jerzy Błaszczyński[1], Roman Słowiński[1,2], and Robert Susmaga[1]

[1] Institute of Computing Science, Poznań University of Technology,
60-965 Poznań, Poland
{jblaszczynski,rslowinski,rsusmaga}@cs.put.poznan.pl
[2] Systems Research Institute, Polish Academy of Sciences,
01-447 Warsaw, Poland

Abstract. We consider estimation of relevance of attributes used for classification. This estimation takes into account the predictive capabilities of the attributes. To this end, we are using Bayesian confirmation measure. The estimation is based on analysis of rule classifiers in classification tests. The attribute relevance measure increases when more rules involving this attribute suggest a correct decision, or when more rules that do not involve this attribute suggest an incorrect decision in the classification test; otherwise, the attribute relevance measure is decreasing. This requirement is satisfied by a monotonic Bayesian confirmation measure. Usefulness of the presented measure is verified experimentally.

Keywords: attribute relevance, Bayesian confirmation, decision rule, classification, ensemble classifier.

1 Introduction

There are at least two possible goals of learning from classification data. One common goal is the discovery of relations between values of the condition attributes and the classes (i.e. values of the decision attribute). A second, less popular one, is the detection of attribute relevance, which involves identification of attributes that are crucial for correct prediction.

In this paper, we propose a method for eliciting useful information concerning the attributes from analysis of predictive capacity of a set of decision rules. This information may be useful for finding relevant subsets of attributes, i.e. subsets that, when fed to classifiers, improve the accuracy of classification. At the same time, the information may constitute good evaluation measure of individual attributes. The method that we present employs relevance measure that has the Bayesian confirmation property.

In practice, usefulness of subsets of attributes cannot be simply expressed as a sum of the usefulness of single attributes. As an illustration, consider two identical attributes with a high (and equal to each other) individual relevance, and a third one, independent of the two, of a lower individual relevance. Now, define a set consisting of the two identical, highly relevant attributes, and a set consisting of one highly relevant and the less relevant one. Owing to inevitable

J.T. Yao et al. (Eds): RSKT 2011, LNCS 6954, pp. 36–44, 2011.

duplication of information, the set consisting of two identical attributes carries potentially less information about the objects than the set consisting of two different attributes. In consequence, the set with two highly relevant attributes cannot be evaluated higher than the set with one highly relevant and one less relevant attribute. In the presented approach the elicited information constitutes a good trade-off between the relevance of individual attributes and the relevance of sets of attributes.

The rest of the paper is organized as follows. Section 2 presents related works. In Section 3, we describe the method for estimating attribute relevance from decision rules. In section 4, we present an experimental evaluation of the method. Final remarks and conclusions are contained in Section 5.

2 Related Works

In the following review, we briefly describe some works that are related to our proposal. We start with general methods that explain how individual attribute values contribute to predictions [15,16]. These methods can be applied to any classifier (and thus, also to a rule-based one) in a uniform way. All these methods use basically the same approach, in which attribute value's contribution is defined as a difference between the model's initial prediction and its average prediction across perturbations of the considered attribute. Such an approach can lead to some problems, explained and addressed in [16]. Nevertheless, possibility of application of these general methods to the evaluation of the contribution of whole attributes (i.e. all values of an attribute) seems to be an open question. On the other hand, the approach presented in this paper can be generalized in a way that allows to assess contribution of individual attribute values to predictions (e.g. by binarization of attribute values). Moreover, interest in general methods is well motivated in case of black-box models. In case of transparent models, like decision rules, this motivation is less important. Another method that allows to evaluate importance of conditions in a set of decision rules (i.e. individual attribute values) was presented in [10]. This method has a similar foundation (Shapley value). However, the evaluation of condition importance is made independently from predictive performance of rules.

An approach that permits to discover attributes that are important from classifier performance perspective was proposed in the context of Random Forests [7]. In this approach, contribution of an attribute to the performance of the classifier is derived by comparing the performance of a forest learned on the original data set to the performance of the forest that was learned on a data set with randomly permuted values of the attribute.

Moreover, measuring the relevance of rules has recently received much attention. Various quantitative measures of rule relevance (attractiveness) have been proposed and studied. The most commonly used measures of relevance include *support*, *confidence*, *lift* and *rule interest function* (see e.g. [8,14] for a survey on these measures). The proposed measures are, however, considered only with respect to identification of the most useful rules and filtering out the least useful

ones. A common conclusion resulting from this research is that there is no single way for assessment of rule relevance that would work best on every real-life problem. To qualify interestingness measures of decision rules, some desirable properties have been required for them, in particular the property of Bayesian confirmation [8]. This one, in turn, should possess other desirable properties, like special kinds of symmetry or monotonicity [9].

3 Attribute Relevance and Properties of Rules

We consider attribute relevance measures that satisfy the property of confirmation. These measures are taking into account interactions between attributes represented by decision rules. In this case, the property of confirmation is related to quantification of the degree to which presence of an attribute in the premise of a rule provides evidence for or against the conclusion of the rule. The measure increases when more rules involving an attribute suggest a correct decision, or when more rules that do not involve the attribute suggest an incorrect decision, otherwise it decreases.

Let us first give some basic definitions. A *rule* induced from a learning data set L can be denoted as $E \rightarrow H$, which reads as "*if* E, *then* H". A rule consists of a condition part (called also the premise or the evidence) E, and a conclusion (called also the prediction or the hypothesis) H. Considering a finite set of condition attributes $A = \{a_1, a_2, \ldots a_n\}$, we can define the condition part of the rule as a conjunction of elementary conditions on a particular subset of attributes:

$$E = e_{i_1} \wedge e_{i_2} \wedge \ldots \wedge e_{i_p}, \tag{1}$$

where $\{i_1, i_2, \ldots i_p\} \subseteq \{1, 2, \ldots, n\}$, $p \leq n$, and e_{i_h} is an elementary condition defined on the value set of attribute a_{i_h}, $h \in \{i_1, i_2, \ldots i_p\}$ (e.g., $a_{i_h} \geq 0.5$).

The set of rules R induced from data set L can be applied to objects from L or to objects from a testing set T. A rule $r \equiv E \rightarrow H$, $r \in R$, covers object x ($x \in L$ or $x \in T$) if x is satisfying the condition part E. We say that the rule is correctly classifying x if it both covers x and x satisfies the decision part H. In other words, we say that rule r is true for object x if it classifies this object correctly, and it is not true otherwise.

By $a_i \triangleright E$ we denote the fact that E includes an elementary condition e_i involving attribute a_i, $i \in \{1, 2, \ldots n\}$. An opposite fact will be denoted by $a_i \not\triangleright E$. Let us consider object x ($x \in L$ or $x \in T$), and set of rules R. We use the following notation throughout the paper:

$a = |H \wedge (a_i \triangleright E)|$ - the number of rules that correctly classify x and involve attribute a_i in the condition part,
$b = |H \wedge (a_i \not\triangleright E)|$ - the number of rules that correctly classify x and do not involve attribute a_i in the condition part,
$c = |\neg H \wedge (a_i \triangleright E)|$ - the number of rules that incorrectly classify x and involve attribute a_i in the condition part,
$d = |\neg H \wedge (a_i \not\triangleright E)|$ - the number of rules that incorrectly classify x and do not involve attribute a_i in the condition part.

Usually, we are more interested in how the induced rules are performing for a fixed set of objects than for one fixed object. Thus, the previously introduced parameters can be summed over the given set of objects. Nevertheless, in such a case, we will keep the same notation for the sake of simplicity. Observe that for a fixed set of objects T, a is interpreted as the number of all rules that correctly classify objects from T and involve attribute a_i. Interpretation of the remaining parameters is analogous. The values of a, b, c, and d can be also treated as frequencies that may be used to estimate probabilities, e.g. $\Pr(H \wedge (a_i \triangleright E)) = a/(a+b+c+d)$, or $\Pr(a_i \triangleright E) = (a+c)/(a+b+c+d)$.

Formally, an relevance measure $c(H, (a_i \triangleright E))$ has the property of Bayesian confirmation if and only if it satisfies the following conditions:

$$c(H, (a_i \triangleright E)) = \begin{cases} > 0 & \text{if} \quad \Pr(H|(a_i \triangleright E)) > \Pr(H), \\ = 0 & \text{if} \quad \Pr(H|(a_i \triangleright E)) = \Pr(H), \\ < 0 & \text{if} \quad \Pr(H|(a_i \triangleright E)) < \Pr(H). \end{cases} \tag{2}$$

The conditions of definition (2) thus equate the confirmation with an increase of the probability of the hypothesis caused by the evidence while disconfirmation with a decrease of the probability of the hypothesis caused by the evidence. Finally, neutrality is identified in case of lack of influence of evidence on hypothesis. Now, there are at least three logically equivalent ways to express the fact that $a_i \triangleright E$ confirms H [11] in the context of the Kolmogorov theory of probability [13]. Namely: $\Pr(H|(a_i \triangleright E)) > \Pr(H)$, $\Pr(H|(a_i \triangleright E)) > \Pr(H|(a_i \not\triangleright E))$, and $\Pr(H|(a_i \triangleright E)) > \Pr((a_i \triangleright E)|\neg H)$. The second way is especially interesting for our purposes. It allows for a redefinition of the relevance measure satisfying the Bayesian confirmation (2) to the form of the following conditions:

$$c(H, (a_i \triangleright E)) = \begin{cases} > 0 & \text{if} \quad \Pr(H|(a_i \triangleright E)) > \Pr(H|(a_i \not\triangleright E)), \\ = 0 & \text{if} \quad \Pr(H|(a_i \triangleright E)) = \Pr(H|(a_i \not\triangleright E)), \\ < 0 & \text{if} \quad \Pr(H|(a_i \triangleright E)) < \Pr(H|(a_i \not\triangleright E)). \end{cases} \tag{3}$$

When probabilities are estimated in terms of frequencies, (2) and (3) may be expressed in terms of a, b, c, and d. We use normalized confirmation measure c_1, defined in [11], since it has some desirable properties which are useful for our application. These properties guarantee that the measure handles properly extreme situations (no positive examples or no counterexamples for the hypothesis). Measure c_1, when applied to attribute confirmation, is defined as follows:

$$c_1(H, (a_i \triangleright E)) = \begin{cases} \alpha + \beta \frac{ad-bc}{(a+b)(b+d)} & \text{if} \quad \frac{a}{a+c} > \frac{b}{b+d} \wedge c = 0, \\ \alpha \frac{ad-bc}{(a+c)(c+d)} & \text{if} \quad \frac{a}{a+c} > \frac{b}{b+d} \wedge c > 0, \\ -\alpha + \beta \frac{ad-bc}{(b+d)(c+d)} & \text{if} \quad \frac{a}{a+c} < \frac{b}{b+d} \wedge a = 0, \\ \alpha \frac{ad-bc}{(a+b)(a+c)} & \text{if} \quad \frac{a}{a+c} < \frac{b}{b+d} \wedge a > 0, \end{cases} \tag{4}$$

where $\alpha > 0$, $\beta > 0$, and $\alpha + \beta = 1$.

4 Experiments

The purpose of the experiment was to select the most confirmatory attributes and to demonstrate that classifiers constructed on these attributes are better then the ones constructed on less confirmatory attributes. In other words, first, we wanted to check the distribution of attribute confirmation on different data sets, and second, we wanted to show that the distribution is meaningful with respect to the predictive performance of constructed classifiers. The experimental procedure was conceptually simple, but computationally expensive. In the main step of the procedure we estimated attribute confirmations by constructing multiple sets of rules on randomly selected subsets of the original objects. More precisely, we constructed bagging ensembles [4,5] with VC-DomLEM rule component classifiers [6].

4.1 Experimental Setup

The estimation of attribute confirmation $c_1(H, (a_i \rhd E))$ has been made in three distinct ways, depending on hypothesis H: Firstly, in single rule estimation (SRE), we test if a single rule (in the ensemble) involving attribute a_i assigns objects correctly. Secondly, in component classifiers estimation (CCE), we test if the component classifier including rules that involve attribute a_i assigns objects correctly. Finally, in ensemble estimation (EE), we test if the whole ensemble composed of rules that involve attribute a_i assigns objects correctly.

The following assumptions have been made in all the experiments. Values of α, and β were set to 0.5. Such an assumption leads to a sensible trade-off between the confirmation and disconfirmation cases detected by the measure. The number of rule component classifiers in each of the ensembles was 10. To get reliable estimates of confirmation the rules were constructed on training data sets resulting from 10-fold cross validation. For better reproducibility of the results we repeated the cross validation 10 times.

The data sets used in experiment were all real-life data sets of miscellaneous origin and nature, obtained from the UCI Repository [1]. The sets had been created for scientific purposes and were used in different experiments, which were, however, not necessarily related to experimental confirmation evaluation. All data sets were preprocessed in a way allowing their treatment in a general dominance-based rough set approach classification scheme [2,3]. This scheme allows to handle ordinal and non-ordinal classification problems in a uniform way by considering multiple realizations of original attributes. As a result, the number of continuous attributes in the preprocessed data sets doubled. Discrete attributes were binarized and the number of resulting binary attributes doubled. Consequently, the numbers of attributes specified in tables and figures is higher than in original data sets.

4.2 Experimental Results

For brevity of presentation, we present a chart of typical c_1 confirmation values in Figure 1. In most of the cases these values are clustered around zero

Table 1. Basic characteristics of the real-life data sets used in the experiments

Id	Short Name	Full Name	Nature of Cond. Attr.	#Continuous Cond. Attr.	#Discrete Cond. Attr.	#Objects
1	breast-w	Breast Cancer Wisconsin	discrete	0	18	699
2	colic	Horse Colic	mixed	14	106	368
3	diabetes	Pima Indians Diabetes	continuous	16	0	768
4	heart-statlog	Statlog (Heart)	mixed	16	36	270
5	ionosphere	Ionosphere	continuous	68	0	351
6	labor	Labor Relations	mixed	16	36	57
7	parkinsons	Parkinsons	continuous	44	0	195
8	promoters	Promoter Gene Sequences	discrete	0	456	106
9	spectf	SPECTF Heart	continuous	88	0	349
10	vote	Congressional Voting Records	discrete	0	32	435

(exceptions in this respect are data sets `labor` and `promoters`). This means that most of the attributes do not confirm or disconfirm the correct classification. It follows that only a small fraction of those present in the elementary conditions of the rules turns out to be more useful than others.

Kendall's τ correlation coefficients [12] were calculated in order to check the impact of the distinct ways in which the values of c_1 were estimated. Their values, presented in Table 2, indicate clearly that the most similar are the subsets of attributes identified by SRE and CCE. This was expected because SRE and CCE are more similar to each other than to EE. EE results in significantly more diversified sets of confirmatory attributes than the other two ways of estimation.

Table 2. Kendall's τ coefficients between ranks of c_1 attribute confirmation values estimated on different levels

Data Set	SRE/ CCE	SSR/ EE	CCE/ EE
breast-w	0.9216	0.8562	0.8301
colic	0.7808	0.5256	0.4927
diabetes	0.9216	0.8562	0.8301
heart-statlog	0.8235	0.8190	0.7572
ionosphere	0.5237	0.3033	0.1128
labor	0.7896	0.5747	0.5928
parkinsons	0.8499	0.6892	0.7040
promoters	0.6960	0.4069	0.4200
spectf	0.8417	0.5441	0.5572
vote	0.9657	0.9335	0.9173

The next step in our experiments consisted in checking whether the most confirmatory attributes (i.e., the attributes having highest values of c_1) produce better classifiers than the less confirmatory attributes. In Table 3 we present overall classification accuracy (OCA) of ensemble classifiers learned on the half of most confirmatory attributes (1/2 MC) according to: SRE, CCE, and EE, respectively. In the same table we also present OCA of the classifier learned on

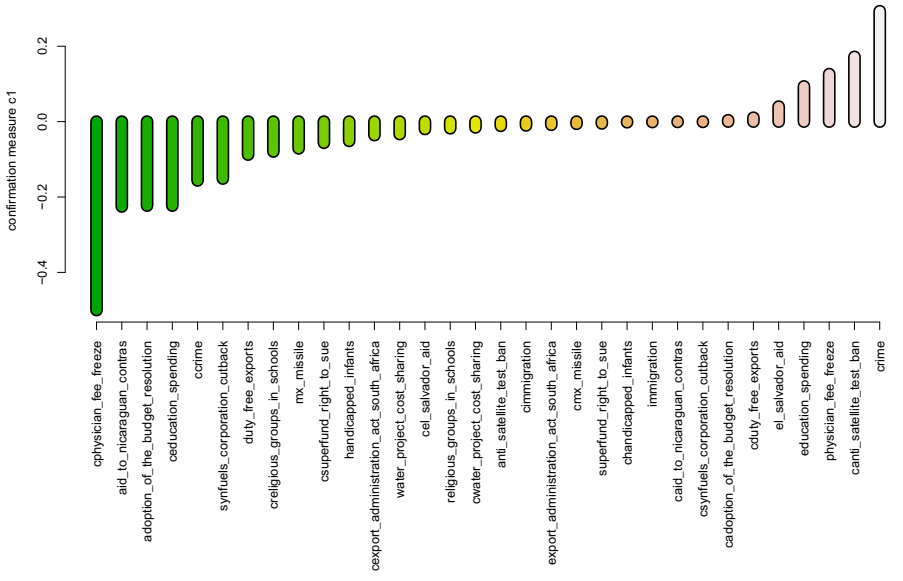

Fig. 1. Values of c_1 confirmation measure – single rule estimation on `vote` data set

the whole set of attributes (base) and OCA of the one learned on the half of the least confirmatory attributes (1/2 LC). The best among classifiers is marked with an asterisk. The better of classifiers learned on the half of the most confirmatory attributes and the rest of the attributes is marked by bold typeface.

Results in Table 3 show that the half of the most confirmatory attributes, according to SRE, almost always produces the best classifier – in all cases it outperforms the classifier learned on the rest of the attributes and in most cases it outperforms the base classifier learned on all the attributes (with the exception of: `breast-w`, `diabetes`, and `parkinsons`). This result can be attributed, partly, to the applied preprocessing of the data. The information encoded by each of the original attributes is rendered at least twice in the transformed data sets. Nevertheless, the results prove that the attributes identified as the most confirmatory are crucial for high OCA in prediction.

The results for CCE resemble the results obtained for SRE, although the OCA of the classifiers learned from the most confirmatory attributes is lower. In accordance with this observation, base classifier is better than the others for one additional data set: `labor`. The results for EE show the same trend that was observed for SRE and CCE. Ensemble estimation of c_1 leads to subsets of attributes that are worse in terms of predictive properties with respect to these selected by component classifiers estimation and single rule estimation. The classifiers learned on the most confirmatory attributes are still better than those learned on the rest of the attributes (with one exception: `parkinsons` data set), but worse, in most of the cases, than the base classifiers.

Table 3. Overall classification accuracy of classifiers constructed on the most confirmatory and the least confirmatory attributes

Id	Base	SRE		CCE		EE	
		1/2 MC	1/2 LC	1/2 MC	1/2 LC	1/2 MC	1/2 LC
1	96.2 ±0.31*	**96.1** ±0.26	00.0 ±0.00	**96.1** ±0.26	00.0 ±0.00	**96.0** ±0.29	86.6 ±0.92
2	84.0 ±0.58	**85.4** ±0.78*	68.5 ±0.81	**85.4** ±1.09	66.2 ±0.78	**83.7** ±0.69	70.9 ±1.53
3	75.5 ±0.64*	**68.4** ±0.77	17.5 ±0.60	**68.4** ±0.77	17.5 ±0.60	**68.4** ±0.77	17.5 ±0.60
4	81.3 ±1.00	**81.5** ±1.34*	44.0 ±1.73	**81.5** ±1.04*	42.6 ±1.68	**81.1** ±0.60	45.0 ±1.66
5	92.9 ±0.85	**93.9** ±0.37*	87.8 ±0.84	**93.0** ±0.84	87.7 ±1.01	**93.2** ±0.35	87.5 ±0.70
6	87.0 ±2.38	**87.9** ±2.54*	69.5 ±2.96	**87.0** ±1.95	73.3 ±1.72	**84.9** ±2.11	80.5 ±2.65
7	87.6 ±1.22*	**86.6** ±1.58	83.4 ±1.00	**86.4** ±1.52	83.7 ±1.38	86.2 ±1.61	**86.8** ±0.64
8	90.3 ±2.95	**93.8** ±2.03*	51.3 ±2.39	**93.5** ±1.49	55.9 ±2.67	**82.0** ±1.63	73.6 ±1.89
9	90.5 ±0.94	**91.2** ±0.94	86.6 ±1.38	**91.4** ±1.00*	86.1 ±1.70	**90.8** ±1.03	88.6 ±1.47
10	93.2 ±0.72	**94.1** ±0.81	20.7 ±3.34	**94.1** ±0.81	20.7 ±3.34	**94.5** ±0.68*	40.4 ±3.79

An otherwise interesting result was obtained for `breast-w` data set, which contains condition attributes that are monotonically correlated with the decision attribute. In our experiment, all the conditional attributes with value sets ordered in the same direction as the order of the decision classes were identified as the most confirmatory. In result, the classifier learned on the rest of the attributes was not able to make absolutely any correct prediction.

5 Conclusions

In this paper we presented and evaluated rule-based estimation methods of attribute relevance. All these methods are based on computation of c_1 confirmation measure, which had been proven to be theoretically superior to other, similar measures. Our evaluation complements these previous results by showing that the measure is useful in the predictive perspective. More precisely, we demonstrated, in experiment, that the attributes identified by this measure as the most confirmatory are the ones that are crucial in constructing accurate classifiers. Moreover, we showed that the attributes that are the most confirmatory according to single rule estimation are the most useful in prediction. This result is concordant with the observation that this way of estimation is the most direct (i.e. it is not dependent on the effects of aggregation of rules, which disturb the other ways of estimation).

Acknowledgment. The authors wish to acknowledge financial support from the Polish Ministry of Science and Higher Education, grant N N519 314435.

References

1. Asuncion, A., Newman, D.: UCI machine learning repository (2007), `http://www.ics.uci.edu/~mlearn/MLRepository.html`

2. Błaszczyński, J., Greco, S., Słowiński, R.: Ordinal and non-ordinal classification using monotonic rules. In: 8th International Conference of Modeling and Simulation, MOSIM 2010 (May 2010)
3. Błaszczyński, J., Greco, S., Słowiński, R.: Inductive discovery of laws using monotonic rules. Engineering Applications of Artificial Intelligence (to appear)
4. Błaszczyński, J., Słowiński, R., Stefanowski, J.: Feature set-based consistency sampling in bagging ensembles. In: From Local Patterns To Global Models (LEGO), ECML/PKDD Workshop, pp. 19–35 (2009)
5. Błaszczyński, J., Słowiński, R., Stefanowski, J.: Variable consistency bagging ensembles. In: Peters, J.F., Skowron, A. (eds.) Transactions on Rough Sets XI. LNCS, vol. 5946, pp. 40–52. Springer, Heidelberg (2010)
6. Błaszczyński, J., Słowiński, R., Szeląg, M.: Sequential covering rule induction algorithm for variable consistency rough set approaches. Information Sciences 181(5), 987–1002 (2011)
7. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
8. Fitelson, B.: Likelihoodism, Bayesianism, and relational confirmation. Synthese 156, 473–489 (2007)
9. Greco, S., Słowiński, R., Pawlak, Z.: Can Bayesian confirmation measures be useful for rough set decision rules? Engineering Applications of Artificial Intelligence 17(4), 345–361 (2004)
10. Greco, S., Słowiński, R., Stefanowski, J.: Evaluating importance of conditions in the set of discovered rules. In: An, A., Stefanowski, J., Ramanna, S., Butz, C.J., Pedrycz, W., Wang, G. (eds.) RSFDGrC 2007. LNCS (LNAI), vol. 4482, pp. 314–321. Springer, Heidelberg (2007)
11. Greco, S., Słowiński, R., Szczęch, I.: Properties of rule interestingness measures and alternative approaches to normalization of measures. IEEE Transactions on Knowledge and Data Engineering (to appear)
12. Kendall, M.G.: A new measure of rank correlation. Biometrika 30(1–2), 81–93 (1938)
13. Kolmogorov, A.: Foundations of Probability. AMS Chelsea publishing, Providence (1956)
14. McGarry, K.: A survey of interestingness measures for knowledge discovery. The Knowledge Engineering Review 20(1), 39–61 (2005)
15. Robnik-Šikonja, M., Kononenko, I.: Explaining classifications for individual instances. IEEE Trans. on Knowl. and Data Eng. 20, 589–600 (2008)
16. Strumbelj, E., Kononenko, I.: An efficient explanation of individual classifications using game theory. J. Mach. Learn. Res. 11, 1–18 (2010)