A Metric for Selection of the Most Promising Rules

Pedro Gago[1,2] and Carlos Bento[2]

[1] Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria
Morro do Lena, Alto Vieiro, 2400 Leiria
pgago@estg.iplei.pt
[2] CISUC - Centro de Informática e Sistemas da Universidade de Coimbra
Polo II, 3030 Coimbra
pgago,bento@eden.dei.uc.pt

Abstract. The process of Knowledge Discovery in Databases pursues the goal of extracting useful knowledge from large amounts of data. It comprises a pre-processing step, application of a data-mining algorithm and post-processing of results. When rule induction is applied for data-mining one must be prepared to deal with the generation of a large number of rules. In these circumstances it is important to have a way of selecting the rules that have the highest predictive power. We propose a metric for selection of the n rules with the highest average distance between them. We defend that applying our metric to select the rules that are more distant improves the system prediction capabilities against other criteria for rule selection. We present an application example and empirical results produced from a synthesized data set on a financial domain.

1 Introduction

The process of Knowledge Discovery in Databases pursues the goal of extracting useful knowledge from large amounts of data. It comprises a pre-processing step, application of a data-mining algorithm and post-processing of the results.

When rule induction is applied for data-mining one must be prepared to deal with the generation of a large number of rules. In these circumstances it is important to have a way of selecting the rules that have the highest predictive power.

The selection of interesting rules faces an essentially subjective problem. It is not likely that two different users will find the same rules to be interesting. Some authors support decisions on interestingness on subjective information provided by the user (Liu, Hsu and Chen [4], Klementinen, Mannila, Ronkainen, Toivonen and Verkamo [3]; Piatesky-Shapiro and Matheus [7]; Silberschatz and Tuzhilin [8]), other authors (Kamber and Shinghal [2]; Piatesky-Shapiro [6]; Major and Mangano [5]; Srikant and Agrawal [9]) choose to look for objective measures for rule interestingness. Some of those measures are simplicity, statistical significance, coverage and confidence.

Within our work the goal is not to decide on the individual interestingness of rules. Our aim is to build a set of rules that together gives a good coverage of the search space. We build on work on creativity (Gomes, Bento, Gago and Costa [1]) where a measure for the distance between cases is developed. We propose a metric for distance between two rules and use it to select the most heterogeneous set of rules that is possible in the assumption that this set has high predictive capabilities.

We present an example on a database of fictional data on bank clients. From the client database, rules are generated describing the 'good' and 'bad' clients. We use our metric on the rules to select n heterogeneous rules supposed to be the n rules that globally have the highest prediction power. Finally we present the results obtained using these rules and compare them with the use of the same number of rules selected randomly.

2 A Metric for Distance Between Two Rules

The rules are of the form 'IF *conditions* THEN *conclusion*'. *Conditions* is a conjunction of terms in the form $a_1 s_1 v_1 \wedge ... \wedge a_n s_n v_n$ with A=$\{A_1, ..., A_t\}$ a set of attributes and S=$\{<, <=, >, >=, =\}$ a set of comparison operators, $a_i \in$ A, $s_i \in$ S and v_i being a numeric value. *Conclusion* is a class $c_i \in$ C with C=$\{c_1, ..., c_m\}$, a set of classes.

The measure we propose for distance between two rules with the same conclusion is based on three factors:

1. the number of attributes in one rule and absent in the other,
2. the number of attributes in both rules with overlapping values,
3. the number of attributes in both rules with values slight or null overlapping.

Based on these features we propose the following distance metric:

$$dist\,(r_i, r_j) = \begin{cases} \frac{\alpha\ \#DA_{i,j} + \beta\ \#DV_{i,j} - \omega\ \#EV_{i,j}}{\#F_i + \#F_j} & \text{if } \#\mathrm{NO}_{i,j} = 0 \\ 2 & \text{otherwise} \end{cases}$$

In this metric we have:

$\#\mathrm{DA}_{i,j}$	Number of attributes in rule i and not in rule j plus the number of attributes in rule j and not in rule i.
$\#\mathrm{NO}_{i,j}$	Number of attributes both in rule i and in rule j but with non overlapping values.
$\#\mathrm{DV}_{i,j}$	Number of attributes both in rule i and in rule j, but with slightly overlapping values (we consider an overlapping bellow 66%).
$\#\mathrm{EV}_{i,j}$	Number of attributes in both rules, with overlapping values (we consider an overlapping above 66%).
$\#\mathrm{F}_i + \#F_j$	Number of attributes in rule i plus the number of attributes in rule j.

For attributes appearing in both rules we check the intersection of their values. If there is no intersection we know the two rules cannot be applied to the same cases and thus we assign a value of two for the distance between the rules. We chose value two as it is a value higher than the one assigned in any other situation. In the case we have an intersection in less than 66% of the range of the rules values we consider the respective attributes to behave as if they were different attributes. We consider an attribute in two rules to be equal if the intersection of values in these rules is over 66% of the range of possible values.

To illustrate this concept of overlapping attributes consider that we have the attribute A appearing in two rules (Rule #1: A>=20 AND A<=70, Rule #2: A>=40): We can apply rule 1 when the values for attribute A are between 20 and 70 and can apply rule 2 if the values for attribute A are over 40 (see Fig. 1). Consider that the upper and lower limit for this attribute value is, respectively, 0 and 100.

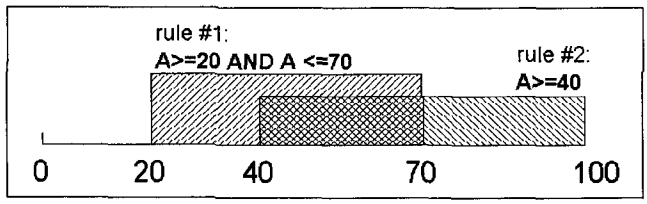

Fig. 1. Overlapping of an attribute in two rules.

We want to know if we can consider the conditions in both rules as equal. Suppose we have more rules with attribute A and that the highest value attribute A has in any of those rules is 100. For attribute A in rule 1 the range of possible values is 70-20 = 50. In rule 2 the range is 100-40 = 60 (100 is the maximum value for attribute A). The overlapping range for values in both rules is 70-40 = 30. This overlapping happens in 30/50 = 60% of the range of possible values for attribute A in rule 1 and in 30/60 = 50% of the range of possible values for attribute A in rule 2. If we average the intervals of overlapping in the two rules we get 55%. As this value is below 66% we consider attribute A in rule 1 and attribute A in rule 2 as if they were different attributes.

In the distance metric the characteristics which are strongly different between the two rules ($\#DA_{i,j}$ and $\#DV_{i,j}$) increase the value returned by the function. The ones appearing in both rules ($\#EV_{i,j}$) decrease the function value. The terms $\#DA_{i,j}$, $\#DV_{i,j}$ and $\#EV_{i,j}$ are weighed by constants α, β and ω. We use α=1, β=2 and ω=2, making the metric take values between minus one and one. A value of one relates to two rules that have few things in common whereas a value of minus one indicates that the rules overlap strongly. When we are absolutely sure that the rules do not overlap the metric returns a value of two.

3 Rule Selection

We use the metric described in the previous section for selection of the n rules that provide the highest coverage for a data set. We pursue this goal by looking for the rules most different from the ones selected till now. We believe those rules to be the ones that provide, in general, the best coverage for the data set.

Considering the original set of classification rules to be S we must apply the following algorithm in order to build the set of n rules which are more distant between each other. We use as distance criteria the one provided by the distance function described in the previous section. We name this set of rules S_R.

Algorithm RuleSelect(S, n)
  R $\leftarrow$ The rule with the highest average distance to the other rules in S;
  $S_R \leftarrow$ R;
  While $\#S_R <$ n **do**
    for each rule R' in S and not in S_R
      AV $\leftarrow$ The average distance of R' to the rules in S_R ;
    endfor
    $R_{max} \leftarrow$ The rule with the highest AV;
    $S_R \leftarrow S_R \cup \{ R_{max} \}$;
  endwhile
  return(S_R) .

If we look for the rules with the lowest average distance we get the set of the rules closest to each other and we assume this set to be the least representative ruleset S_L.

4 Domain Description

Our rule selection heuristic was tested in the domain of loan analysis. We synthesized a database with 3000 entries describing bank clients. The database was split in 2000 cases for training and 1000 for testing. For each client we know its yearly income, size of household, assets owned and amount to be paid per year. The database contains also information on whether the client presented a surety and on whether he is on a contract. The classification field in the database tells us if the client is a 'good' or 'bad' client. The following table shows the available information for three clients.

The independent attribute values were generated using the uniform distribution. The dependent attribute status is labeled 'Good' or 'Bad'. Status is assigned 'Good' or 'Bad' accordingly to two simple criteria: (1) if a client's available money (income minus loan) is over 500 units per household member then the client will honor his debts ((Income-Loan)/Household > 500); (2) 10% of the clients on a contract are 'Bad' clients. We do not consider noisy data.

C4.5 is used on this database for rule generation.

Table 1. Three cases

Income	Household	Assets	Loan	Surety	Contract	Status
7000	4	21000	3100	No	No	Good
4100	1	3000	400	No	No	Good
7500	4	8000	2300	No	Yes	Bad

5 An Example

We considered the 14 rules generated by C4.5 (see appendix) for illustration of our approach. We describe how the distance between two rules is calculated on three rules in the set of 14 initial rules and show the results returned by the distance function (see Fig. 2).

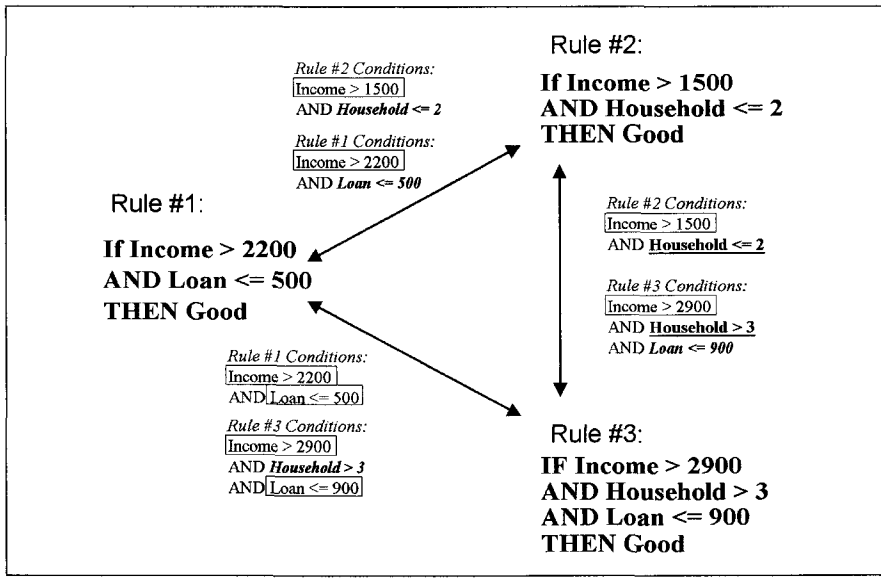

Fig. 2. Three rules generated by C4.5.

In order to determine the overlapping percentage for each attribute in the rules we need to know the range of values for this attribute. From the ruleset we have that the lowest value for *Income* is 1500, the highest being 5800. The *Loan* attribute ranges between 200 and 2900.

Rules 1 and 2 have two distinct attributes - *Loan* and *Household*. The attribute *Income* appears in both rules. We have to determine the intersection for the values for this attribute.

Rule 1 may be used when *Income* is between 2200 and 5800 (the upper limit). For this rule the range is 5800 - 2200 = 3600. For rule 2 the range is

5800 - 1500 = 4300. Rules 1 and 2 overlap from 2200 to 5800. The overlapping range of these rules is 5800 - 2200 = 3600. The rules overlap in all the range of the first rule. The overlapping percentage for attribute *Income* in rule 1 is 3600 / 3600 * 100 % = 100%. For rule 2 the overlapping takes place in 3600 / 4300 * 100% = 83.7 % of the range covered by the attribute *Income*. When we average these values we get (100 % + 83.7 %) / 2 = 91.85 %. As the value obtained is over 66% we consider the values in both rules to be equal.

We can now calculate the value returned by the distance function considering that $\#DA_{1,2} = 2$ (*Household* and *Loan*), $\#NO_{1,2} = 0$, $\# DV_{1,2} = 0$, $\#EV_{1,2} = 1$ and $\#F_1 + \#F_2 = 4$.

The value returned for dist(r_1,r_2) is:

$$dist\,(r_1, r_2) = \frac{2 + 2 * 0 - 2 * 1}{4} = \frac{2 - 2}{4} = 0$$

For rules 1 and 3 we have $\#DA_{1,3} = 1$ (*Household*) and we must check to see if the values in the attributes *Income* and *Loan* should be considered equal or different. For *Income* in rule 1 the range is 3600 and in rule 3 the range is 2900. The two rules overlap from 2900 to 5800 (the overlapping range is 2900). The overlapping percentage for rule 1 is 80.5% (2900/3600 * 100 %) and for rule 3 is 100% (2900/2900 * 100 %). Averaging these values we get 90.25%. As the value is over 66% we consider the values to be equal.

For the other attribute shared by rules 1 and 3 (*Loan*) we have that the range for rule 1 is 500-200 = 300 and for rule 2 is 900-200 = 700 (200 is the lower limit for this attribute). The overlapping range is 300. For rule 1 the overlapping percentage is 100% and for rule 3 it is 42.8%. As the average for these values (71.4%) is once again over 66% we consider the values to be equal.

So, we have $\#DA_{1,3} = 1$, $\#DV_{1,3} = 0$, $\#EV_{1,3} = 2$, $\#OV_{1,3} = 0$ and $\#F_1 + \#F_3 = 5$, and the value returned by the function is:

$$dist\,(r_1, r_3) = \frac{1 + 2 * 0 - 2 * 2}{5} = \frac{1 - 4}{5} = -\frac{3}{5}$$

The attribute *Household* appears both in rule 2 and rule 3 and it is easy to see that these rules do not overlap. For these rules $\#NO_{2,3} \neq 0$ and the metric returns a value of two.

$$dist\,(r_2, r_3) = 2$$

Rules 1 and 2 share one attribute with values very much alike. Rules 1 and 3 have two attributes with almost equal values and only one that is different. These results agree with our intuitive knowledge that rule 2 is "much farther" from rule 3 than from rule 1 and that rules 1 and 3 are very close.

In our example, the rule with the highest average distance to the others is rule 2 (the average distance is one) and it will be the first in set S_R. The second rule in S_R will be rule 3 as it is the rule with the highest distance to the rule already in S_R (rule 2).

6 Empirical Results

We applied our algorithm for rule selection to 14 rules (see appendix) to get the set S_R of the most promising rules. We also considered a set with the least representative rules S_L and several sets of rules chosen at random S_{RA}.

Using our program we built sets S_R and S_L with sizes from 1 to 8. Using the 1000 cases for testing we measured the coverage for S_R and S_L function of the number n of rules in these sets. We also determined the average coverage of randomly chosen groups of rules denominated S_{RA}. The results are presented in the graph in Fig. 3.

When considering the sets S_R, S_{RA} and S_L with a number n of rules near half the number in the original ruleset we see that the rules in S_R have a higher prediction power than those in S_{RA} and in S_L. We also note that the rules in S_L are consistently the ones with the lowest predictive power.

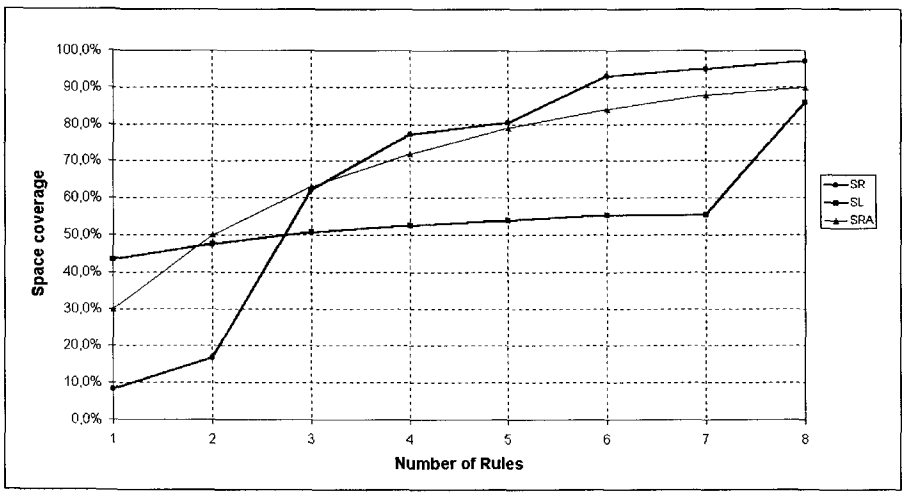

Fig. 3. Comparison of the sets coverage.

7 Conclusions and Future Work

We developed a method for rule selection that consistently outperforms random choice. Using this method one can reduce the time spent analyzing uninteresting rules and concentrate on the best ones. Unlike previous work (Kamber and Shinghal [2]; Piatesky-Shapiro [6]; Major and Mangano [5]; Srikant and Agrawal [9]) we do not try to measure the interestingness of a rule. We use our metric to measure the distance between any two rules in order to build a set with the rules with the highest predictive power.

One of the problems found within our framework is the difficulty in the assignment of weights for our metric. For now this selection is guided by the fact that these values make the metric take values between minus one and one. We believe our approach is domain independent as long as the underlying data are uniformly distributed.

This method for rule selection works well if the underlying data follow a uniform distribution. If the data follow other distributions one must adjust the algorithm that allows us to calculate the range of the interception between two rules. One of the next steps is to adapt the metric to make it work with normally distributed data. It would also be interesting to explore the behavior of the metric when it has available domain knowledge concerning the upper and lower limits for the values of each attribute without having to guess them from the induced rules. For now our program looks at the rules in order to find values it may use as limits. We will probably get better results if we ask the user to provide us with an upper and lower limits for the attributes values.

References

1. Gomes, P., Bento, C., Gago, P., Costa, E.: Towards a Case-Based Model for Creative Processes. *Proceedings of the 12 th European Conference on Artificial Intelligence* (1996) 122-126.
2. Kamber, M., Shingal, R.: Evaluating the Interestingness of Characteristic Rules. *Proceedings of the Second International Conference on Knowledge Discovery & Data Mining* (1995) 263-266.
3. Klementinen, M., Mannila, H., Ronkainen, P., Toivonen, H., Verkamo, A.: Finding Interesting Rules from Large Datasets of Discovered Association Rules. *Proceedings of the Third International Conference on Information and Knowledge Management* (1994) 401-407.
4. Liu, B., Hsu, W., Chen, S.: Using General Impressions to Analyse Discovered Classification Rules. *Proceedings of the Third International Conference on Knowledge Discovery & Data Mining* (1997) 31-36.
5. Major, J.A., Mangano, J.: Selecting Among Rules Induced from a Hurricane Database. *Proceedings of the AAAI-93 Workshop on Knowledge Discovery in Databases* (1993) 28-44.
6. Piatesky-Shapiro, G.: Discovery, Analysis and Presentation of Strong Rules. *G. Piatesky-Shapiro & W.J. Frawley, eds., Knowledge Discovery in Databases. Menlo Park, CA: AAAI/MIT Press.* (1991) 229-248.
7. Piatesky-Shapiro, G., Matheus, C.J.: The Interestingness of Deviations. *Proceedings of the AAAI-94 Workshop on Knowledge Discovery in Databases* (1994) 25-36.
8. Silberschatz, A., Tuzhilin, A.: What Makes Paterns Interesting in Knowledge Discovery Systems. *IEEE Trans. On Know. and Data Eng. 8(6)* (1996) 970-974.
9. Srikant, R., Agrawal, R.: Mining Generalized Association Rules. *Proceedings of the 21st VLDB conference* (1995) 407-419

Appendix - The rules generated by C4.5.

Rule #1:
if Income > 5600
and Contract = 0
then Good

Rule #2:
if Income > 4100
and Assets <= 28000
and Surety = 0
and Contract = 0
then Good

Rule #3:
if Income > 5800
then Good

Rule #4:
if Income > 3600
and Loan <= 1800
then Good

Rule #5:
if Income > 4500
and Loan <= 2900
then Good

Rule #6:
if Income > 3100
and Loan <= 1300
then Good

Rule #7:
if Income > 2900
and Household <= 3
then Good

Rule #8:
if Income > 2900
and Household > 3
and Loan <= 900
then Good

Rule #9:
if Income > 1500
and Household <= 3
and Loan <= 200
then Good

Rule #10:
if Income > 1500
and Household <= 2
then Good

Rule #11:
if Income > 2300
and Household <= 3
and Loan <= 1100
then Good

Rule #12:
if Income > 1900
and Household <= 3
and Loan <= 700
then Good

Rule #13:
if Income > 2200
and Loan <= 500
then Good

Rule #14:
if Household <= 1
then Good