Incremental Rule Induction based on Rough-Set based Rule Layers

Shusaku Tsumoto and Shoji Hirano

Department of Medical Informatics, School of Medicine, Faculty of Medicine
Shimane University
89-1 Enya-cho Izumo 693-8501 Japan
{tsumoto,hirano}@med.shimane-u.ac.jp

Abstract. This paper proposes a new framework for incremental learning based on rough-set based rule layers. According to classification of addition of an example into four cases, two inequalities for accuracy and coverage are obtained in terms of conditions of these two indices for rule induction. The proposed method classifies a set of formulae into three layers: rule layer, subrule layer and non-rule layer by using the inequalities obtained. Then, subrule layer plays a central role in updating rules.

Keywords: incremental rule induction; rough sets; accuracy; coverage; subrule layer

1 Rough Sets and Probabilistic Rules

1.1 Accuracy and Coverage

In the framework of rough set theory, the classification of training samples $D$ can be viewed as a search for the best set $[x]_R$ which is supported by the relation $R$. We can define the characteristics of classification in the set-theoretic framework. For example, accuracy and coverage, or true positive rate can be defined as:

$$\alpha_R(D) = \frac{|[x]_R \cap D|}{|[x]_R|} \quad \text{and} \quad (1)$$

$$\kappa_R(D) = \frac{|[x]_R \cap D|}{|D|}, \quad (2)$$

where $|A|$ denotes the cardinality of a set $A$, $\alpha_R(D)$ denotes an accuracy of $R$ as to classification of $D$, and $\kappa_R(D)$ denotes a coverage, or a true positive rate of $R$ to $D$, respectively.

1.2 Probabilistic Rules

The simplest probabilistic model is that which only uses classification rules which have high accuracy and high coverage. This model is applicable when rules of high accuracy can be derived. Such rules can be defined as:

In this model, we assume that accuracy is dominant over coverage.
\[ R \xrightarrow{\alpha, \kappa} d \text{ s.t. } R = \bigvee_i R_i = \bigvee \bigwedge_j [a_j = v_j], \]
\[ \alpha_{R_i}(D) > \delta_\alpha \text{ and } \kappa_{R_i}(D) > \delta_\kappa, \]
where \( \delta_\alpha \) and \( \delta_\kappa \) denote given thresholds for accuracy and coverage, respectively.

## 2 Theory for Incremental Learning

Table 1 gives the classification of four cases of an additional example. From

<table>
<thead>
<tr>
<th>( n_R )</th>
<th>( n_D )</th>
<th>( n_{RD} )</th>
<th>( \alpha(t) )</th>
<th>( \kappa(t) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n_R + 1 )</td>
<td>( n_D )</td>
<td>( n_{RD} )</td>
<td>( \frac{\alpha(t)}{n_R + 1} )</td>
<td>( \frac{\kappa(t)}{n_D + 1} )</td>
</tr>
<tr>
<td>( n_R )</td>
<td>( n_D + 1 )</td>
<td>( n_{RD} )</td>
<td>( \frac{\alpha(t)}{n_D + 1} )</td>
<td>( \frac{\kappa(t)}{n_R + 1} )</td>
</tr>
<tr>
<td>( n_R + 1 )</td>
<td>( n_D + 1 )</td>
<td>( n_{RD} + 1 )</td>
<td>( \frac{\alpha(t)}{n_R + 1} )</td>
<td>( \frac{\kappa(t)}{n_D + 1} )</td>
</tr>
</tbody>
</table>

Table 1, updates of Accuracy and Coverage can be calculated from the original datasets for each possible case. Since rules is defined as a probabilistic proposition with two inequalities, supporting sets should satisfy the following constraints:

\[ \alpha(t + 1) > \delta_\alpha, \; \kappa(t + 1) > \delta_\kappa \]

Then, the conditions for updating can be calculated from the original datasets:

**Theorem 1.** If accuracy and coverage of a formula \( R \) to \( d \) satisfies one of the following inequalities, then \( R \) may exclude or include into the candidates of formulae for probabilistic rules.

\[ \frac{\delta_\alpha(n_R + 1) - 1}{n_R} < \alpha_{R_i}(D)(t + 1) < \frac{\delta_\alpha(n_R + 1)}{n_R}, \quad (3) \]
\[ \frac{\delta_\kappa(n_D + 1) - 1}{n_D} < \kappa_{R_i}(D)(t + 1) < \frac{\delta_\kappa(n_D + 1)}{n_D}. \quad (4) \]

It is notable that the lower and upper bounds can be calculated from the original datasets.

Select all the formula whose accuracy and coverage satisfy the above inequalities. They will be a candidate for updates. A set of formulae which satisfies the inequalities for probabilistic rules is called rule layer and one which satisfies Eqn (3) and (4) is called subrule layer.
3 An Algorithm for Incremental Learning

3.1 Algorithm

In order to provide the same classificatory power to incremental learning methods as ordinary learning algorithms, we introduce an incremental learning method PRIMEROSE-INC2 (Probabilistic Rule Induction Method based on Rough Sets for Incremental Learning Methods)\(^2\).

From the results in the above section, a selection algorithm is defined as follows, where the following four lists are used. \(List_c\) stores a formula which satisfies the above inequalities shown in Eqn (3) nad (4). This \(List_c\) is called subrule layer. \(List_a\) is a list of formulae probabilistic rules which satisfy the condition on the thresholds of accuracy and coverage, which called rule layer. Finally, \(List_r\) stores a list of formulae which do not satisfy the above condition.

1. Apply rule induction to the initial data.
2. Store formulae in rules into \(List_a\) and others in \(List_r\).
3. Calculate upper and lower bounds for accuracy and coverage from accuracy, coverage and given thresholds for rules.
4. Check the above inequalities for attributes in \(List_a\) and \(List_r\). If an attribute satisfies one of the inequalities, then it includes into \(List_c\).
5. From an additional example, classify addition from four cases.
   (a) An additional example satisfies a target and a formula in \(List_c\). Move a formula from \(List_r\) to \(List_a\).
   (b) An additional example only satisfies a formula in \(List_c\). Move a formula from \(List_a\) to \(List_r\).
   (c) An additional example only satisfies a target in \(List_c\). Move a formula from \(List_a\) to \(List_r\).
   (d) An additional example neither satisfies a target nor a formula in \(List_c\). No movement.
6. Generate rules from \(List_a\).
7. Return to (3).

For further information, please refer to [2].

Acknowledgements

This research is supported by Grant-in-Aid for Scientific Research (B) 24300058 from Japan Society for the Promotion of Science (JSPS).

References

2. Tsumoto, S., Hirano, S.: Rough-set based criteria for incremental rule induction (submitted)

\(^2\) This is an extended version of PRIMEROSE-INC[1]