Abstract

This paper proposes a new framework for rule induction methods, called “layered rule induction”, based on rule layers constrained by inequalities of statistical indices, such as confidence and support. The change of indices with an additional example reflects their sensitivity, and four patterns should be considered if confidence and support are focused on. Then, by using these two pairs of inequalities obtained by analysis, the proposed method classifies a set of formulae into four layers: the rule layer, subrule layer (in and out) and the non-rule layer. Using these layers, updates of probabilistic rules are equivalent to their movement between layers. Rules can be extracted from each rule layer. The proposed method was evaluated on datasets regarding headaches and meningitis, and the results show that the proposed method outperforms the conventional methods.

Keywords: rule induction; rough sets; probabilistic rules; rule layer

1. Introduction

Several symbolic inductive learning methods have been proposed, such as induction of decision trees [1, 2, 3], and AQ family [4, 5, 6]. These methods are applied to discover meaningful knowledge from large databases, and their usefulness is in some aspects ensured. However, most of the approaches induces rules from all the data in databases, and cannot induce incrementally when new samples are derived. Thus, we have to apply rule induction methods again to the databases when such new samples are given, which causes the computational complexity to be expensive even if the complexity is $n^2$.

Thus, it is important to develop incremental learning systems to manage large databases [7, 8]. However, most of the previously introduced learning systems have the following two problems: first, those systems do not outperform ordinary learning systems, such as AQ15 [6], C4.5 [9] and CN2 [4]. Secondly, those incremental learning systems mainly induce deterministic rules. Therefore, it is indispensable to develop incremental learning systems which induce probabilistic rules to solve the above two problems.

In this paper, we propose a simple but stronger framework for rule induction, which can be used both for ordinary and incremental rule learning, called called PRIMEROSE4 (Probabilistic Rule Induction Method based on Rough Sets for Incremental Learning Methods), which induces probabilistic rules in both contexts.

The important point is to consider four possibilities of incremental sampling, and calculates increase in given indices. Then with the updates of indices and threshold of rule selection inequalities, incremental rule selection
inequalities are obtained. In the case of combination of accuracy and coverage, we derive two inequalities for rule layer selection.

By using these two inequalities, the proposed method classifies a set of formulae into four layers: the rule layer, subrule layer (in and out) and the non-rule layer. Using these layers, updates of probabilistic rules are equivalent to their movement between layers. The proposed method was evaluated on datasets regarding headaches and meningitis, and the results show that the proposed method outperforms the conventional methods.

This paper is organized as follows: Section 2 briefly describe rough set theory and the definition of probabilistic rules based on this theory. Section 3 discusses problems in the incremental learning of probabilistic rules. Section 4 provides formal analysis of incremental updates of accuracy and coverage, where two important inequalities are obtained. Section 5 presents an induction algorithm for incremental learning based on the above results, which is then evaluated in Section 6. Finally, Section 7 concludes this paper.

2. Rough Sets and Probabilistic Rules

2.1. Rough Set Theory

Rough set theory clarifies set-theoretic characteristics of the classes over combinatorial patterns of the attributes, which are precisely discussed by Pawlak [10, 11]. This theory can be used to acquire some sets of attributes for classification and can also evaluate how precisely the attributes of database are able to classify data. One of the main features of rough set theory is to evaluate the relationship between the conditional attributes and the decision attributes by using the hidden set-based relations. Let a conditional attribute or conjunctive formula of attributes a decision attribute be denoted by $R$ and $D$. Then, a relation between $R$ and $D$ can be evaluated by each supporting sets ($[x]_R$ and $[x]_D$) and their overlapped region denoted by $R \land D ([x]_R \cap [x]_D)$. If $[x]_R \subset [x]_D$, then a proposition $R \rightarrow D$ will hold and $R$ will be a part of lower approximation of $D$. Dually, $D$ can be called a upper approximation of $R$. In this way, we can define the characteristics of classification in the set-theoretic framework. Let $n_R$, $n_D$ and $n_{RD}$ denote the cardinality of $[x]_R$, $[x]_D$ and $[x]_R \cap [x]_D$, respectively. Accuracy (true positive rate) and coverage (true positive rate) can be defined as:

$$\alpha_R(D) = \frac{n_{RD}}{n_R} \quad \text{and} \quad \kappa_R(D) = \frac{n_{RD}}{n_D} \quad (2)$$

It is notable that $\alpha_R(D)$ measures the degree of the sufficiency of a proposition, $R \rightarrow D$, and that $\kappa_R(D)$ measures the degree of its necessity. For example, if $\alpha_R(D)$ is equal to 1.0, then $R \rightarrow D$ is true. On the other hand, if $\kappa_R(D)$ is equal to 1.0, then $D \rightarrow R$ is true. Thus, if both measures are 1.0, then $R \leftrightarrow D$.

For further information on rough set theory, readers could refer to [10, 11, 12].

2.2. Probabilistic Rules

The simplest probabilistic model is that which only uses classification rules which have high accuracy and high coverage. $^1$ This model is applicable when rules of high accuracy can be derived. Such rules can be defined as:

$$R \xrightarrow{\alpha,\kappa} d \quad \text{s.t.} \quad R = \vee_i R_i = \vee_j [a_j = v_k],$$

$$\alpha_R(D) > \delta_{\alpha} \quad \text{and} \quad \kappa_R(D) > \delta_{\kappa},$$

where $\delta_{\alpha}$ and $\delta_{\kappa}$ denote given thresholds for accuracy and coverage, respectively. 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. We call these two inequalities rule selection inequalities.

It is notable that this rule is a kind of probabilistic proposition with two statistical measures, which is one kind of an extension of Ziarko’s variable precision model(VPRS) [11]. $^2$

$^1$ In this model, we assume that accuracy is dominant over coverage.

$^2$ In VPRS model, the two kinds of precision of accuracy is given, and the probabilistic proposition with accuracy and two precision conserves the characteristics of the ordinary proposition. Thus, our model is to introduce the probabilistic proposition not only with accuracy, but also with coverage.
3. Problems in Incremental Rule Induction

The most important problem in incremental learning is that it does not always induce the same rules as those induced by ordinary learning systems \(^3\), although an applied domain is deterministic. Furthermore, since induced results are strongly dependent on the former training samples, the tendency of overfitting is larger than in the ordinary learning systems.

The most important factor of this tendency is that the revision of rules is based on the formerly induced rules, which is the best way to suppress the exhaustive use of computational resources. However, when induction of the same rules as ordinary learning methods is required, computational resources will be needed, because all the candidates of the rules should be considered.

Thus, for each step, computational space for deletion of candidates and addition of candidates is needed, which causes the computational speed of incremental learning to be slow. Moreover, in the case when probabilistic rules should be induced, the situation becomes much severer, since the candidates for probabilistic rules become much larger than those for deterministic rules.

4. Incremental Updates of Statistical Indices

4.1. Four Possibilities

Usually, datasets will monotonically increase. Let \(n_R(t)\) and \(n_D(t)\) denote cardinalities of a supporting set of a formula \(R\) in given data and a target concept \(d\) at time \(t\).

\[
\begin{align*}
n_R(t + 1) &= \begin{cases} n_R(t) + 1 & \text{an additional example satisfies } R \\ n_R(t) & \text{otherwise} \end{cases} \\
n_D(t + 1) &= \begin{cases} n_D(t) + 1 & \text{an additional example belongs to a target concept } d. \\ n_D(t) & \text{otherwise} \end{cases}
\end{align*}
\]

Let \(\neg R\) and \(\neg D\) be the negations of \(R\) and \(D\), respectively. Then, the above two possibilities have the following two dual cases.

\[
\begin{align*}
n_{\neg R}(t + 1) &= \begin{cases} n_{\neg R}(t) & \text{an additional example satisfies } R \\ n_{\neg R}(t) + 1 & \text{otherwise} \end{cases} \\
n_{\neg D}(t + 1) &= \begin{cases} n_{\neg D}(t) & \text{an additional example belongs to a target concept } d. \\ n_{\neg D}(t) + 1 & \text{otherwise} \end{cases}
\end{align*}
\]

Thus, from the definition of accuracy (Eqn. (1)) and coverage (Eqn. (2)), accuracy and coverage may nonmonotonically change due to the change of the intersection of \(R\) and \(D\), \(n_{RD}\). Since the above classification gives four additional patterns, we will consider accuracy and coverage for each case as shown in Table 1, called incremental sampling scheme, in which 0 and +1 denote stable and increase in each value.

Since accuracy and coverage use only the positive sides of \(R\) and \(D\), we will consider the following subtable for the updates of accuracy and coverage (Table 2).

Then, Table 3 is obtained as the classification of four cases of an additional example.

---

\(^3\)Here, ordinary learning systems denote methods that induce all rules by using all the samples.
Table 1: Incremental Sampling Scheme

| $R$ | $D$ | $\neg R$ | $\neg D$ | $R \land D$ | $\neg R \land D$ | $R \land \neg D$ | $\neg R \land \neg D$ |
|-----|-----|--------|--------|-------------|------------------|-----------------|-------------------|
| 0   | 0   | +1     | +1     | 0           | 0                | 0               | +1                |
| 0   | +1  | +1     | 0      | 0           | +1               | 0               | 0                 |
| +1  | 0   | 0      | +1     | 0           | 0                | +1              | 0                 |
| +1  | +1  | 0      | 0      | +1          | 0                | 0               | 0                 |

Table 2: Four patterns for an additional example

| $t$:          | $[x]_R(t)$ | $D(t)$ | $[x]_R \land D(t)$ |
|---------------|------------|--------|---------------------|
| original      | $n_R$      | $n_D$  | $n_{RD}$            |

| $t$+1:        | $[x]_R(t+1)$ | $D(t+1)$ | $[x]_R \land D(t+1)$ |
|---------------|--------------|----------|-----------------------|
| Both negative (BN) | $n_R$       | $n_D$    | $n_{RD}$             |
| $R$: positive (RP)  | $n_R + 1$  | $n_D$    | $n_{RD}$            |
| $d$: positive (dP)  | $n_R$      | $n_D + 1$| $n_{RD}$            |
| Both positive (BP)  | $n_R + 1$  | $n_D + 1$| $n_{RD} + 1$       |

Table 3: Summary of change of accuracy and coverage

| Mode   | $\alpha(t+1)$ | $\kappa(t+1)$ | $\alpha(t)$ | $\kappa(t)$ |
|--------|----------------|---------------|--------------|--------------|
| BN     | $n_R$          | $n_D$         | $n_{RD}$     | $\alpha(t)$ | $\kappa(t)$ |
| RP     | $n_R + 1$      | $n_D$         | $n_{RD}$     | $\frac{\alpha(t)n_R}{n_R + 1}$ | $\kappa(t)$ |
| dP     | $n_R$          | $n_D + 1$     | $n_{RD}$     | $\alpha(t)$ | $\frac{\kappa(t)n_D}{n_D + 1}$ |
| BP     | $n_R + 1$      | $n_D + 1$     | $n_{RD} + 1$ | $\frac{\alpha(t)n_R}{n_R + 1}$ | $\frac{\kappa(t)n_D}{n_D + 1}$ |

4.2. Updates of Accuracy and Coverage

From Table 3, 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:

$$\alpha(t + 1) > \delta_\alpha, \quad \kappa(t + 1) > \delta_\kappa$$

Then, the conditions for updating can be calculated from the original datasets: when accuracy or coverage does not satisfy the constraint, the corresponding formula should be removed from the candidates. On the other hand, both accuracy and coverage satisfy both constraints, the formula should be included into the candidates. Thus, the following inequalities are important for inclusion of $R$ into the conditions of rules for $D$:

$$\alpha(t + 1) = \frac{\alpha(t)n_R + 1}{n_R + 1} > \delta_\alpha,$$

$$\kappa(t + 1) = \frac{\kappa(t)n_D + 1}{n_D + 1} > \delta_\kappa.$$

For its exclusion, the following inequalities are important:

$$\alpha(t + 1) = \frac{\alpha(t)n_R}{n_R + 1} < \delta_\alpha.$$
\[ \kappa(t + 1) = \frac{\kappa(t)n_D}{n_D + 1} < \delta_k. \]

Thus, the following inequalities are obtained for accuracy and coverage.

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

\[
\frac{\delta_a(n_R + 1) - 1}{n_R} < \alpha_R(D)(t) \leq \delta_a, \quad (4)
\]

\[
\frac{\delta_k(n_D + 1) - 1}{n_D} < \kappa_R(D)(t) \leq \delta_k. \quad (5)
\]

A set of \( R \) which satisfies the above two constraints is called **in subrule layer**.

**Theorem 2.** If accuracy and coverage of a formula \( R \) to \( d \) satisfies one of the following inequalities, then \( R \) may exclude from the candidates of formulae for probabilistic rules.

\[
\delta_a < \alpha_R(D)(t) < \frac{\delta_a(n_R + 1)}{n_R}, \quad (6)
\]

\[
\delta_k < \kappa_R(D)(t) < \frac{\delta_k(n_D + 1)}{n_D}. \quad (7)
\]

A set of \( R \) which satisfies the above two constraints is called **out subrule layer**.

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

Select all the formulae 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 a **rule layer** and a set of formulae which satisfies Eqn (4) and (5) is called a **subrule layer (in)**. Figure 1 illustrates the relations between a rule layer and a sublayer.

![Fig. 1: Intuitive Diagram of Rule and Subrule Layers](image)

Subrule Layer (out)
Subrule Layer (in)
Included when BP
Included when BP
Deleted when BP
Deleted when BP
Deleted when RP
Deleted when dP
Deleted when BN

Select all the formulae 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 a **rule layer** and a set of formulae which satisfies Eqn (4) and (5) is called a **subrule layer (in)**. Figure 1 illustrates the relations between a rule layer and a sublayer.
5. Rule Induction Algorithm

5.1. PRIMEROSE4

To provide the same classificatory power to incremental learning methods as ordinary learning algorithms, we introduce a rule induction method PRIMEROSE4, which is a non-learning extended version of PRIMEROSE-INC2 (Probabilistic Rule Induction Method based on Rough Sets for Incremental Learning Methods [13])

From the results in the above section, a rule induction algorithm is obtained as Figure 2. First, it picks up a formula \( R \) and calculate accuracy and coverage. Then, the ordinary rule selection inequalities, \( \alpha_R(D) > \delta_\alpha \) and \( \kappa_R(D) > \delta_\kappa \) are checked. If both inequalities are satisfied, \( R \) is included into a regular rule layer. Then, if one of the inequalities, Equation (6) and (7) is satisfied, \( R \) is also appended to a out subrule layer. If one of the ordinary inequalities is not satisfied, then Then, the inequalities, Equation (4) and (5) are checked. If one of the inequalities is satisfied, \( R \) is included into a in subrule layer. The process will be continued until the list of formulae is empty.

![Rule Induction Algorithm Diagram](image)

6. Experimental Results

PRIMEROSE4 \(^5\) was applied to headache and meningitis [14], whose precise information is given in Table 4. The proposed method was compared with the former version PRIMEROSE-INC, the non-incremental versions: PRIMEROSE \([15]\) and PRIMEROSE0 \(^6\), and the other three conventional learning methods: C4.5, CN2 and AQ15. The experiments were conducted using the following three procedures. First, these samples randomly split into pseudo-training samples and pseudo-test samples. Second, using the pseudo-training samples, PRIMEROSE4, PRIMEROSE-INC2, PRIMEROSE-INC, PRIMEROSE, and PRIMEROSE0 induced rules and the statistical measures \(^7\). Third, the induced results were tested by the pseudo-test samples. The performance

---

\(^4\)This is an extended version of PRIMEROSE-INC[14]

\(^5\)The program is implemented by using SWI-prolog.

\(^6\)This version is given by setting \( \delta_\alpha \) to 1.0 and \( \delta_\kappa \) to 0.0.

\(^7\)The thresholds \( \delta_\alpha \) and \( \delta_\kappa \) are set to 0.75 and 0.5, respectively in these experiments.
Table 4: Information about Databases

| Domain      | Samples | Classes | Attributes |
|-------------|---------|---------|------------|
| headache    | 1477    | 10      | 20         |
| meningitis  | 198     | 3       | 25         |

of PRIMEROSE-INC was measured both by rules and subrules. These procedures were repeated 100 times and each accuracy is averaged over 100 trials. Table 5 gives the comparison between PRIMEROSE-INC2 and other rule induction methods with respect to the averaged classification accuracy and the number of induced rules. The table compared four options: (In, Reg, Out), (In, Reg), (Reg, Out) and (Reg), where In, Reg and Out denote in-subrule, rule, out-subrule layers, respectively. For example, in the first case, three layers (in, out-subrule layers and regular region) were used. These results show that PRIMEROSE4 perform as good as PRIMEROSE-INC2 and outperformed all the other non-incremental learning methods, although this method needed a much larger memory space for run. Furthermore, it is notable that there exist differences among

Table 5: Experimental Results: Accuracy (Headache)

| Method                   | Headache    | Meningitis  |
|--------------------------|-------------|-------------|
| PRIMEROSE4 (In, Reg, Out)| 89.9 ± 2.4% | 82.5 ± 1.2% |
| PRIMEROSE4 (In, Reg)     | 87.5 ± 6.3% | 83.5 ± 2.7% |
| PRIMEROSE4 (Reg, Out)    | 89.5 ± 5.3% | 81.5 ± 3.2% |
| PRIMEROSE4 (Reg)         | 84.5 ± 3.7% | 75.5 ± 2.2% |
| PRIMEROSE-INC2           | 89.9 ± 2.4% | 82.5 ± 1.2% |
| PRIMEROSE-INC            | 89.5 ± 5.4% | 77.3 ± 3.0% |
| PRIMEROSE                | 84.5 ± 5.4% | 75.5 ± 3.0% |
| PRIMEROSE0               | 79.9 ± 1.7% | 67.1 ± 4.1% |
| C4.5                     | 85.8 ± 2.4% | 81.5 ± 3.2% |
| CN2                      | 87.0 ± 3.9% | 74.0 ± 2.1% |
| AQ15                     | 86.2 ± 2.6% | 69.0 ± 1.8% |
|                         | Out Dominant| In Dominant |

7. Conclusion

By extending concepts of rule induction methods based on rough set theory, called PRIMEROSE-INC2 (Probabilistic Rule Induction Method based on Rough Sets for Incremental Learning Methods), we have introduced a new approach to knowledge acquisition, called PRIMEROSE4 which induces probabilistic rules both in an ordinary and an incremental ways.

The method classifies elementary attribute-value pairs into four categories: a rule layer, in/out subrule layers and a non-rule layer by using the inequalities obtained from the proposed framework. This system was evaluated on clinical datasets regarding headache and meningitis. The results show that PRIMEROSE4 outperforms previously proposed methods.

---

8The performance of PRIMEROSE-INC2 was equivalent to that of PRIMEROSE-INC.
Acknowledgements

This research is supported by Grant-in-Aid for Scientific Research (B) 15H2750 from Japan Society for the Promotion of Science (JSPS).

References

[1] L. Breiman, J. Freidman, R. Olshen, C. Stone, Classification And Regression Trees, Wadsworth International Group, Belmont, CA, 1984.
[2] B. Cestnik, I. Kononenko, I. Bratko, Assistant 86: A knowledge-elicitation tool for sophisticated users, in: EWSL, 1987, pp. 31–45.
[3] J. Quinlan, Induction of decision trees, Machine Learning 1.
[4] P. Clark, T. Niblett, The cn2 induction algorithm, Machine Learning 3.
[5] R. S. Michalski, A theory and methodology of inductive learning, Artif. Intell. 20 (2) (1983) 111–161.
[6] R. Michalski, I. Mozetic, J. Hong, N. Lavrac, The multi-purpose incremental learning system aq15 and its testing application to three medical domains, in: Proceedings of the fifth National Conference on Artificial Intelligence, AAAI Press, Menlo Park, 1986, pp. 1041–1045.
[7] N. Shan, W. Ziarko, Data-based acquisition and incremental modification of classification rules, Computational Intelligence 11 (1995) 357–370.
[8] P. E. Utgoff, Incremental induction of decision trees, Machine Learning 4 (1989) 161–186.
[9] J. Quinlan, C4.5 - Programs for Machine Learning, Morgan Kaufmann, Palo Alto, 1993.
[10] Z. Pawlak, Rough Sets, Kluwer Academic Publishers, Dordrecht, 1991.
[11] W. Ziarko, Variable precision rough set model, Journal of Computer and System Sciences 46 (1993) 39–59.
[12] S. Tsumoto, Automated induction of medical expert system rules from clinical databases based on rough set theory, Information Sciences 112 (1998) 67–84.
[13] S. Tsumoto, S. Hirano, Incremental induction of probabilistic rules based on subrule layers, Fundam. Inform. (accepted).
[14] S. Tsumoto, Incremental rule induction based on rough set theory, in: M. Kryszkiewicz, H. Rybinski, A. Skowron, Z. W. Ras (Eds.), ISMIS, Vol. 6804 of Lecture Notes in Computer Science, Springer, 2011, pp. 70–79.
[15] S. Tsumoto, H. Tanaka, Primerose: Probabilistic rule induction method based on rough sets and resampling methods, Computational Intelligence 11 (1995) 389–405.