Rough Set Based K-Exception Approach to Approximate Rule Reduction

ZHANG Feng ZHANG Xianfeng QIN Zhiguang LIU Jinde
(School of Computer Science and Engineering University of Electronic Science and Technology of China, Sichuan, Chengdu 610054, P.R.China)

Abstract There are rules referring to infrequent instances after the procession of attribute reduction and value reduction with traditional methods. A rough set RS based k-exception approach (RSKEA) to rule reduction is presented. Its main idea lies in a two-phase RS based rule reduction. An ordinary decision table is attained through general method of RS knowledge reduction in the first phase. Then a k-exception candidate set is nominated according to the decision table. RS rule reduction is employed for the reformed source data set, which remove all the instances included in the k-exception set. We apply the approach to the automobile database. Results show that it can reduce the number and complexity of rules with adjustable conflict rate, which contributes to approximate rule reduction.

Key words rule; exception; k-exception set; approximate rule reduction

Early psychologist of cognitive science considered that human studied for the very target of attaining rules and applying these rules to new problem. Rules based learning model affects nowadays machine-learning researchers to some extension. Medin’s experiments resulted in CONTEXT model\[1\], which points out to what people learn and memorize are not rules but instances. AI researchers construct instance-based learning approach according to CONTEXT model. In the early 1990’s, Nosofsky found either of the above two learning models touched one side of learning, integrated them and proposed rule-plus-exception learning model\[2\]. Rule-plus-exception model shows great flexibility in terms of rule reduction\[3\]. Rules can always be reduced and simplified concerning removal of infrequent instances. Z. Pawlak, cooperating with other researchers, developed the rough set data analysis (RSDA)\[4\]. RSDA tries to discern internal characteristics of a data set without invoking external metrics and judgment. We propose a rough set based k-exception approach (RSKEA) for rule reduction and take account of rule overlay rate while nominating candidates for infrequency records. RS reduction gains better result over traditional methods, we use k-exception set to refine the result.

1 Related Work

Nosofsky proposes a rule-plus-exception (RULEX) model\[2\]. According to the model, observers learn categories by forming simple logical rules along single dimensions and by remembering occasional exceptions to those rules. The modeling analyses suggest that, when multiple rules are available for solving a problem, averaged classification data often represent a probabilistic mixture of idiosyncratic rule-plus-exception strategies.

Rough set contributes a lot to the applications of attribute reduction, rule generation and prediction. However, its main drawback lies in that rough set does not reduce the computation complexity of the tasks. Finding a smallest reduction of an arbitrary set is NP-hard\[5\]. Even finding all reduction can have exponential time. Heuristic algorithm for relatively smallest reduction with polynomial time complexity (O(n^k)) has been developed\[6,7\]. RSKEA is based on RS reduction algorithm which can’t reduce computation.
complexity but simplify the reduction result.

Wang Jun designed a new rule-plus-exception model for machine learning.[3] He defined an exception set as infrequent instance, and used an integrated form of rule-plus-exception to represent result rule set. We extend the exception set to \(k\)-exception set. \(K\) can be adjusted for performance reason. The result of rule reduction can be a precise one or an approximate one with exact conflict error rate. RSKEA removes the infrequent instance of the decision table. It makes use of alternative heuristic RS based algorithms to gain higher performance. Simpler and more readable result makes it useful for approximate reduction approach.

2 Rough Set Based \(k\)-Exception Approach

Let \(S = \langle U, C \cup D, V, f \rangle\) be a decision table, where \(U\) is the universe, \(V_i \in V\) is the value set of the attribute in \(C \cup D\), and \(f\) is an information function. \(C\) and \(D\) are condition and decision attributes, respectively[4].

2.1 Define \(k\)-Exception Set

Let \(W\) be source data sets. \(R\) is the rule set generated after RS reduction. Check all the records in the source data sets according to \(R\) and count the frequency of each \(r\) in \(R\). \(h(r, Q)\) is the frequency function. \(Q\) is the instance record using rule \(r\). \(k\)-exception set can be defined as

\[
e_k = \bigcup \{Q : h(r, Q) \leq k \quad k = 1, 2, 3, \ldots\}
\]

where \(e_k\) is the instance set, within which \(k\) instances matches rule \(r\). \(k\) must be chosen as an absolutely smaller integer than the number of records in the source data set.

2.2 RSKEA

It is a two-phase rule reduction approach. In the two phases, RS based attribute reduction and value reduction are used for progressively refined rule set.

Phase 1:

Input the source data set \(W\). Common RS reduction algorithm is used to process the source record. Output the initial rule set \(R^{(0)}\).

Phase 2:

Step 1, \(k \leftarrow k_0\); construct \(e_k\) by traversing records in \(W\) using \(R^{(0)}\) generated in Phase 1.

Step 2, Reform \(W\) by removing of all the instances in \(e_k\). RS reduction algorithm again is used to process the reformed \(W\). Output the refinery rule set.

2.3 Considerations

RSKEA is useful for rule reduction of large-scale source data set. For the very reason that decision table will change dramatically by removal of \(k\)-exception (in section 4, we will demo the great changes). RSKEA is also useful for approximate rule reduction. In some fields, users want to make decisions on features or trends corresponding to the majority. It would be better to remove exceptions considering the exception percent is down to zero (\(k\) must be far more smaller than the number of records in source data set). One thing should be pointed out that when RSKEA is used as an approximate rule reduction method, there may be exceptions in source data set conflicting with the reduction rule set \(R^{(0)}\). Generally, there are more exceptions while setting larger \(k\). Value of \(k\) can be adjusted to meet different requirement of precision on rules reduction, where \(k=0\), RSKEA can be used in precise rule reduction.

3 Example

We get magic result when applying RSKEA to approximate rule reduction. To demo the approach, we cite a simple example of decision table[8].

Given the decision table shown in Tab 1, \(U=\{1, 2, \ldots, 21\}\), condition attributes set \(C=\{\text{size, cyl, turbo, fuelsys, displace, comp, power, trans, weight}\}\), decision attribute set \(D=\{\text{mileage}\}\). In the phase 1, we apply RS reduction to the above decision table. Tab 2 shows the generated rule set \(R^{(0)}\). \(R^{(0)}\) is used to compute \(e_k\). Let \(k\) be 1, and we get

\[
e_1 = \bigcup \{Q : h(r, Q) = 1\}.
\]

Calculate \(e_1\) as follows

\[
h(0, \{1, 2, 3, 5, 13, 16, 17, 18\}) = 8
\]
\[
h(1, \{4, 8\}) = 2
\]
\[
h(2, \{6, 7\}) = 2
\]
\[
h(3, \{11, 12, 14, 15, 19\}) = 5
\]
\[
h(4, \{9, 10, 21\}) = 3
\]
\[
h(5, \{20\}) = 1
\]
When counting the frequency of \( r_k \) rule with lower number has higher priority to match the instance. We know \( e_1(0) = 20 \). Removing the No.20 instance reforms the decision table of Tab.1. Applying RS reduction again to the reformed decision table, we get a more readable and simpler rule set shown in Tab. 3. We can do approximate rule reduction by RSKEA with different conflict error rate (CER). Define CER as percent of conflict instance in the total instances set of the source decision table. Conflict instance doesn’t match any rule of the result rule set.

If approximate rule reduction is required, we get result rule set without taking account of exception rule after reforming the source decision table by removing of \( K \)-exception set. Let \( k \) be different value to get approximate reduction with different CER. When \( k = 1 \), the result of RSKEA approximate approach is shown in Tab.3. There are 4 rules and 3 attributes in the result rule set. The value of CER is 1/21, which means there is totally one instance in the source decision table conflicting with result rule set and any rules in the result rule set at least matches 2 instances of the source decision table. When \( k = 2 \), the result of RSKEA approximate approach is shown as ordinary rule in Tab.5. There are only 2 rules and 2 attributes in the result rule set but more conflict instance. CER is 5/21, which means there are 5 instance in the source decision table conflicting with result rule set and any rules in the result rule set at least matches 3 instances of the source decision table. Obviously, when \( k \) is bigger, the result rule set is smaller and CER is increasing.

If precise rule reduction is required, we integrate the exception rule and ordinary rule set. Exception rule set is attained in phase 1 of RSKEA. When calculating \( e_{k} \), we get rule frequency table mapping the matching frequency of each rules shown in Fig.1. Considering the value of \( k (k \geq 1) \), exception rule set is made up of
rules whose matching frequency is equal or less than \( k \).
We get ordinary rule set by the same method used in approximate rule reduction. The finally result rule set consists of exception rule and ordinary rule. Such adjustment as attribute reduction should be taken when integrating exception rule and ordinary rule set for the mismatch of the number of attribute. Tab.4 and Tab.5 show the integrated rule set of RSKEA precise approach of \( k=1 \) or \( k=2 \) respectively. In this example, we find that the integrated rule set is unnecessarily smaller when assigning bigger \( k \). Let \( k \) be 1 or 2, the integrated rule set consists of 5 rules and 5 attributes, but their rules are different and the rule set in Tab. 5 is simpler. CER of RSKEA precise approach is zero because each instance can match a rule and there is no conflict.

| Tab. 4  | \( R(n), k=1 \) integrated rule set |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| Size | fuelsys | displace | weight | mileage |
| exception | * | EFI | small | * | high |
| ordinary rule | compact | * | * | medium | medium |
| rule | light | high |
| subcompact | * | * | * | high |

| Tab. 5  | \( R(n), k=2 \) integrated rule set |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| Size | fuelsys | displace | weight | mileage |
| exception | * | * | * | light | high |
| ordinary rule | EFI | small | * | * | high |
| rule | compact | * | * | medium |
| subcompact | * | * | * | high |

| Tab. 6  | Comparison |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| \( k \) | rules | attributes | conflict error rate (CER) |
| Common approach | \( X \) | 6 | 5 | 0 |
| RSKEA precise approach | 1 | 5 | 5 | 0 |
| RSKEA approximate approach | 1 | 4 | 3 | 1/21 |
| | 2 | 2 | 2 | 5/21 |

Tab.6 compares the common approach and RSKEA precise approach with RSKEA approximate approach on rule reduction from the view of simplicity (the number of rules and attributes) and CER. RSKEA is adaptive to precise and proximate rule reduction, and have simpler result on approximate rule reduction. RSKEA take advantages over large-scale decision tables. If the exception ratio \( (K/N, N \) is the number of objects included in \( U \)) is relatively small, \( k \) can be adjusted to a larger number such as 2 or 3 to achieving better performance.

4 Conclusions

Based on RS reduction, RSKEA uses \( k \)-exception set to fulfill the refinery of progressive rule reduction. The result of rule-set is smaller and simpler over ordinary reduction approaches. Rules are easy to identify and apply. This very feature makes it suitable for database mining. RSKEA has the flexibility that reduction algorithm used in the two phases can be easily replaced by newly built RS attribute reduction and value reduction algorithm for better performance. RSKEA is an error rate manageable approach to approximate rule reduction. \( K \) can be assigned different value to meet various requirement of error rate.

We haven’t analyzed the quantitative relationship between exception ratio \( k/N \) and the conflict error rate. Increasing exception ratio for the sake of a larger \( k \) leads to more competitive result while more instances will be removed from the initial decision, and that will affect conflict error rate. There should be a tradeoff adaptive to different scale of decision table. All these tasks are included in the future work.

References


(Continued on page 62)