Micro and Macro Evaluation of Classification Rules

Yiyu Yao and Bing Zhou
Department of Computer Science
University of Regina
Regina, Saskatchewan, Canada S4S 0A2
E-mail: {yyao, zhou200b}@cs.uregina.ca

Abstract

Rule evaluation plays an important role in the rule learning and classification process. Many existing rule inductive learning algorithms are based on single rule evaluation measures. However, the overall rule induction system performance and the classification process are involved with the evaluation of a set of rules. This brings the need for studying the connections between single rule and rule set evaluation measures. The main objective of this paper is to introduce a general framework of classification rule evaluation which connects two types of evaluations, called micro and macro evaluation. Micro evaluation is based on single rules which can be measured by the common empirical measures. Macro evaluation is based on rule sets, depending on the relationships between rules in the set, different resolutions can be applied. By analyzing the relationships between these two types of evaluations, we suggest that under certain conditions, macro evaluation measures can be explicitly expressed by micro evaluation measures.

1. Introduction

Cognitive science [17] and cognitive informatics [19, 20] are both concerned with the current state of artificial intelligence research. Two related fields of artificial intelligence, namely, machine learning and data mining, explore the extraction of interesting information from data. Rules are commonly used for representing discovered knowledge [3, 14] and are usually expressed in an if-then type form. A classification rule indicates a classification relation between the left and right hand side of the rule. A typical example of classification rule may be stated as “if a product has three years warranty, then it belongs to class +.” There are two issues related to classification rule analysis. First, the number of rules generated from a massive data set can be very large, one needs to select the high performance rules and prune the less interesting ones. Second, if there are many rules involved with the classification of a new object, one needs to select the high-performance rules to make the decision.

Rule evaluation measures can be used in the discovery of interesting rules thereby classification can be made based on these selected rules. Numerous measures have been proposed in the past to evaluate the rule performance. Yao and Zhong analyzed quantitative measures associated with rules [22], such as the generality and support measure. Lavrac, Flash and Zupan provided a unified view of rule evaluation measures including classification accuracy, reliability, and sensitivity measure [10]. Geng and Hamilton surveyed the interesting measures for data mining [7]. Some of these measures look into the objective aspect of rule interestingness [15], such as coverage, precision and complexity. These measures are based on statistical, probabilistic or information theories. Other measures investigate the subjective aspect of rule interestingness [12]. They rely on user’s judgements and are usually domain specific, e.g., rule unexpectedness, usefulness and novelty. The selected evaluation measures in this paper are related to the objective aspect of rule interestingness.

An inductive rule learning algorithm is usually considered as a searching process over all possible rules in order to find a rule that satisfies some evaluation criterion. Most existing rule inductive learning algorithms have been developed based on single rule evaluation measures. In many cases, we also need to evaluate the performance of a rule set in order to make a decision. For example, the overall rule induction system performance relies on rule sets instead of single rules. For the classification of a new object, if the rules involved with the object conflict with each other, then we need to analyze the relationships between these rules in the set, find the conflict resolution and make the decision. Al-
though some resolutions has been proposed in the past for the evaluation of rule set [1, 2, 6, 11, 13, 21], but none of these resolutions have looked into the direct relationships between single rule and rule set evaluation measures.

The main objective of this paper is not to discuss the optimal evaluation measures for a single rule or a rule set, but to introduce a general framework which connects these two types of evaluations, called micro and macro evaluation. Micro (local) evaluation is based on single rules. The common evaluation measures, such as precision, coverage and support, are these types of evaluation measures. Macro (global) evaluation is based on rule sets. Depending on the relationships between rules in the set, different evaluation resolutions can be applied. By analyzing one of these relationships in details, we suggest that under certain conditions, macro evaluation measures can be explicitly expressed by micro evaluation measures.

The rest of the paper is organized as follows. In Section 2, we propose a general framework of classification rule evaluation. In Section 3, we analyze classification rule evaluation measures in the micro and macro level. In Section 4, we show the connections between micro and macro evaluation measures by analyzing the relationships between rules under a pre-defined condition. We conclude this paper and future works in Section 5.

2. A General Framework

In this section, we propose a general framework of rule evaluation as shown in Table 1. This framework is built based on four different classifications at different levels. We label each level by using a numeric labeling system called the Dewey Decimal Classification and proper indentation. Numbers representing units at different levels are connected together by a “·” and each lower level is indented.

At the first level, rule evaluation measures are classified into two different types according to the amount of rules being evaluated, called micro evaluation and macro evaluation.

1. Micro evaluation,
2. Macro evaluation.

Micro evaluation is based on single rules. Most existing rule evaluation measures are proposed for micro evaluation. They have been used to determine the stopping criteria for the rule generation and extract high quality rules for classification purpose. However, evaluation based on single rules might generate biased classification or overfitting results. For example, simply choosing the rule with the highest value of user-defined measure might benefit minority classes [11]. For the overall rule induction system performance, one also needs to consider the evaluation based on a rule set. Macro evaluation is based on rule sets. It is more complicated than micro evaluation since there are more than one rule involved in making a decision.

At the second level, we analyze the relationships between rules in a set which is an essential step for macro evaluation. There are two basic relationships between rules, called non-overlapping and overlapping.

2.1. Non-overlapping rules,
2.2. Overlapping rules.

Rules are non-overlapping if an object \( x \) of the universe \( U \) satisfies at most one rule in the set. Rules are overlapping if an object \( x \) of the universe \( U \) satisfies more than one rule in the set.

At the third level, we further analyze two cases for overlapping rules. First, if an object \( x \) of the universe \( U \) satisfies one or more rules with the same class in the set, we call the rules are consistent. Second, if an object \( x \) of the universe \( U \) satisfies more than one rules with at least two different classes, we call the rules are conflict.

2.2.1. Consistent rules,
2.2.2. Conflict rules.

The fourth level is based on the evaluation purpose of different rule measures. If the measures are designed for evaluating the complexity of rules, we call them complexity measures. If the measures are designed for evaluating the performance of rules, we call them performance measures.

2.2.1. Complexity measures

2.2.1.1. Number of attributes
2.2.1.1.2. Number of rules
2.2.1.1.3. Degree of overlapping
2.2.1.1.4. ... ...

2.2.1.2. Performance measures

2.2.1.2.1. Precision
2.2.1.2.2. Coverage
2.2.1.2.3. Generality
2.2.1.2.4. ... ...

Examples of complexity measures are, number of attributes on the left hand side of a rule, number of rules in the set (2.1.1.2. and 2.2.1.1.2.), degree of overlapping (2.2.1.1.3), and degree of conflict (2.2.2.1.3.). Examples of performance measures are, precision, coverage and generality. Each of these measures captures different characteristics of a rule. This categorization fits both micro and macro evaluations, but differs in the details of the measures. The complexity and performance measures of single rules can be applied directly to rule set
1. Micro evaluation
   1.1. Complexity measures
      1.1.1. Number of attributes
      1.1.2. ... ...
   1.1.2. Performance measures
      1.1.2.1. Precision
      1.1.2.2. Coverage
      1.1.2.3. Generality
      1.1.2.4. ... ...

2. Macro evaluation
   2.1. Non-overlapping rules
      2.1.1. Complexity measures
         2.1.1.1. Number of attributes
         2.1.1.2. Number of rules
         2.1.1.3. ... ...
      2.1.2. Performance measures
         2.1.2.1. Precision
         2.1.2.2. Coverage
         2.1.2.3. Generality
         2.1.2.4. ... ...
   2.2. Overlapping rules
      2.2.1. Consistent rules
         2.2.1.1. Complexity measures
            2.2.1.1.1. Number of attributes
            2.2.1.1.2. Number of rules
            2.2.1.1.3. Degree of overlapping
            2.2.1.1.4. ... ...
         2.2.1.2. Performance measures
            2.2.1.2.1. Precision
            2.2.1.2.2. Coverage
            2.2.1.2.3. Generality
            2.2.1.2.4. ... ...
      2.2.2. Conflict rules
         2.2.2.1. Complexity measures
            2.2.2.1.1. Number of attributes
            2.2.2.1.2. Number of rules
            2.2.2.1.3. Degree of conflict
            2.2.2.1.4. ... ...
         2.2.2.2. Performance measures
            2.2.2.2.1. Precision
            2.2.2.2.2. Coverage
            2.2.2.2.3. Generality
            2.2.2.2.4. ... ...

Table 1. A Classification of Rule Evaluation Measures
3. Classification Rule Evaluation Measures

In this section, we analyze the classification rule evaluation measures in the micro and macro level.

3.1. A Formal Definition of Classification Rules

Typically, a classification rule in the studies of machine learning and data mining is expressed in an if-then form, written as:

\[ X \rightarrow D_i, \]

stating that an object \( x \) satisfies the attributes in \( X \) belongs to the class \( D_i \).

Rule evaluation measures can be derived by analyzing the relationships between \( X \) and \( D_i \) in a \( 2 \times 2 \) contingency table. Table 2 is such a contingency table, where \( | \cdot | \) denotes the cardinality of a set, \( X \) and \( \overline{D_i} \) denote the set complement of \( X \) and \( D_i \).

From this table, some common micro and macro evaluation measures based on single rules can be reviewed as follows.

<table>
<thead>
<tr>
<th>( )</th>
<th>( D_i )</th>
<th>( \overline{D_i} )</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X )</td>
<td>(</td>
<td>X \cap D_i</td>
<td>)</td>
</tr>
<tr>
<td>( \overline{X} )</td>
<td>(</td>
<td>\overline{X} \cap D_i</td>
<td>)</td>
</tr>
<tr>
<td>Totals</td>
<td>(</td>
<td>D_i</td>
<td>)</td>
</tr>
</tbody>
</table>

Table 2. A \( 2 \times 2 \) Contingency Table

If the rules are non-overlapping or consistent. For the conflict relationship, it is necessary to analyze the connections between micro and macro evaluation and find the conflict resolutions. In the past, the most common way to resolve the rule conflicts is to choose the rule with the highest value of user-defined measure, e.g., the maximum-precision value. Such a resolution may not always be correct, and it might generate biased or overfitting prediction since the classification is only based on a single rule. Another resolution is to divide the rules into different subsets according to the class label, then computer user-defined criterion value of each rule set, the rule set with the highest value decides the class label of the new object. However, none of these resolutions are based on the analysis of direct connections between micro and macro evaluations.

In general, the proposed framework provides a classification of rule evaluation measures by focusing on four different levels. More importantly, we want to show that micro and macro evaluations are both indispensable in the process of rule induction and classification. It is important to study the connections between them and discover more efficient rule evaluation resolutions. Throughout this paper, we are trying to investigate these connections by analyzing the properties of single rules and properties of rule sets. Furthermore, we want to show this connection by expressing macro evaluation measures directly with micro evaluation measures.

3.2. Micro Evaluation of a Single Rule

From this table, some common micro and macro evaluation measures based on single rules can be reviewed as follows.

Macro evaluation focus on the overall rule induction system performance in stead of the performance of each single rule of the system. The correspondence evaluation measures can be analyzed as follows.

3.2.1. Performance Measures

The precision of the rule \( X \rightarrow D_i \) can be defined as:

\[ \text{precision}(X \rightarrow D_i) = \frac{|X \cap D_i|}{|X|}, \]

where \( 0 \leq \text{precision}(X \rightarrow D_i) \leq 1 \). In set-theoretic terms, precision is the degree to which \( X \) is included in \( D_i \). Obviously, \( \text{precision}(X \rightarrow D_i) = 1 \) if and only if \( X \subseteq D_i \). In classification problems, precision measure is defined as the ratio of the number of objects in \( X \) that are correctly classified as decision class of \( D_i \) and the number of objects in \( X \). Other alternative names for precision measure are confidence \([5]\), accuracy \([8, 18]\), strength \([4, 9, 16]\) and consistency \([1]\). They are essentially refer to the same measure.

The coverage of the rule \( X \rightarrow D_i \) can be defined as:

\[ \text{coverage}(X \rightarrow D_i) = \frac{|X \cap D_i|}{|D_i|}, \]

where \( 0 \leq \text{coverage}(X \rightarrow D_i) \leq 1 \). Coverage reflects the applicability of the rule. In classification problems, coverage measure is defined as the ratio of the number of objects in \( X \) that are correctly classified as decision class of \( D_i \) and the number of objects satisfying \( D_i \).
3.2.2. Complexity Measures

The complexity of the rule \( X \rightarrow D_i \) can be defined as:

\[
generality(X \rightarrow D_i) = \frac{|X|}{|U|},
\]

which indicates the relative size of the subset \( X \). A subset is more general if it covers more instances of the universe. Clearly, we have \( 0 \leq \text{generality}(X \rightarrow D_i) \leq 1 \).

3.2.2. Complexity Measures

The complexity of the rule \( X \rightarrow D_i \) can be defined by \( \text{number of attributes} \) on the left hand side of the rule:

\[
\text{number of attributes}(X \rightarrow D_i) = \# \text{ of attributes in the rule}.
\]

A rule is more complex if it contains more attributes.

3.3. Macro Evaluation of a Rule Set

Macro evaluation focus on the overall rule induction system performance in stead of the performance of each single rule of the system. The correspondence evaluation measures can be analyzed as follows.

3.3.1. Performance Measures

The \( \text{precision} \) of the rule set can be interpreted as the ratio of number of correctly classified objects and the number of classified objects covered by all the rules in the set, written as:

\[
\text{precision}(RS) = \frac{\# \text{ of correctly classified objects by } RS}{\# \text{ of objects in } U}.
\]

The \( \text{coverage} \) of the rule set can be interpreted as the ratio of number of correctly classified objects covered by the rule set and the number of objects in \( U \), written as:

\[
\text{coverage}(RS) = \frac{\# \text{ of correctly classified objects by } RS}{\# \text{ of objects in } U}.
\]

The \( \text{generality} \) of the rule set can be interpreted as the ratio of number of objects covered by the rule set and the number of objects in \( U \), written as:

\[
\text{generality}(RS) = \frac{\# \text{ of objects covered by } RS}{\# \text{ of objects in } U}.
\]

3.3.2. Complexity Measures

The \( \text{complexity} \) of the rule set can be interpreted as the number of attributes or the number of rules in the set, written as:

\[
\text{number of attributes}(RS) = \# \text{ of attributes in } RS,
\]

and

\[
\text{number of rules}(RS) = \# \text{ of rules in } RS.
\]

The \( \text{complexity} \) of the rule set can also be interpreted as the the degree of overlapping and degree of conflict in the rule set depending on the relationships between rules. Suppose we have two rules in the set: \( X_1 \rightarrow D_i \) and \( X_2 \rightarrow D_i \). These two rules have the same classification label on the right hand side but overlapping on their left hand sides. The degree of overlapping can be written as:

\[
\text{degree of overlapping}(RS) = \frac{|X_1 \cap X_2|}{|X_1 \cup X_2|}.
\]

Suppose we have another pair of rules in the set: \( X_1 \rightarrow D_i \) and \( X_2 \rightarrow D_j \). These two rules are overlapping on the left hand sides and have different classification labels on their right hand sides. The degree of conflict can be written as:

\[
\text{degree of conflict}(RS) = \frac{|X_1 \cap X_2|}{|X_1 \cup X_2|}.
\]

4. Connections of Micro and Macro Evaluations

Micro and macro evaluation are closely related. In this section, we investigate their connections under a pre-defined condition by using the maximum-precision as user-defined criterion.

In general, macro evaluation of a rule set depends on the relationships between rules in the set. When rules conflict with each other, we need to find a conflict resolution based on the connections between micro and macro evaluations.

Suppose we have a classification rule set \( RS \), if two rules in \( RS \) involving the same subset \( X \) of objects and different class labels, that is,

\[ X \rightarrow D_i \text{ and } X \rightarrow D_j, \]

...
where $D_i \neq D_j$. They are called conflict rules. The maximum-precision criterion for each rule involving $X$ is denoted as:

$$X \rightarrow D_{\text{max}}(X),$$

where $D_{\text{max}}(X)$ is the class label with the maximum precision value in the rule set. We also assume that each subset $X \subseteq U$ is disjoint with each other and $\cup_{X \subseteq U} X = U$. Now we can examine the conflict resolutions of rule sets and their relationships with single rule measures.

The precision of the rule set can be interpreted as the ratio of number of correctly classified objects and the number of classified objects covered by all the rules in the set, written as:

$$\text{precision}(RS) = \frac{\# \text{ of correctly classified objects by } RS}{\# \text{ of classified objects by } RS} = \frac{|\cup_{X \subseteq U} (X \cap D_{\text{max}}(X))|}{|\cup_{X \subseteq U} X|} = \sum_{X \subseteq U} |X \cap D_{\text{max}}(X)| = \sum_{X \subseteq U} \frac{|X|}{|U|} \times \text{precision}(X \rightarrow D_{\text{max}}(X)),$$

where $\text{precision}(X \rightarrow D_{\text{max}}(X))$ is the precision for the rules involving $X$ with the maximum precision value. That is, the overall precision of a rule set is the weighted sum of the precision of individual rules in the set.

The coverage of the rule set can be interpreted as the ratio of number of correctly classified objects and the number of objects in $U$, written as:

$$\text{coverage}(RS) = \frac{\# \text{ of correctly classified objects by } RS}{\# \text{ of objects in } U} = \frac{|\cup_{X \subseteq U} (X \cap D_{\text{max}}(X))|}{|U|} = \sum_{X \subseteq U} |X \cap D_{\text{max}}(X)| = \sum_{X \subseteq U} \frac{|D_{\text{max}}(X)|}{|U|} \times \text{coverage}(X \rightarrow D_{\text{max}}(X)),$$

where $\text{coverage}(X \rightarrow D_{\text{max}}(X))$ is the coverage for the rules involving $X$ with the maximum precision value. That is, the overall coverage of a rule set is the weighted sum of the coverage of individual rules in the set.

The generality of the rule set can be interpreted as the ratio of number of objects covered by the rule set and the number of objects in $U$, written as:

$$\text{generality}(RS) = \frac{\# \text{ of objects covered by } RS}{\# \text{ of objects in } U} = \frac{|\cup_{X \subseteq U} X|}{|U|} = \sum_{X \subseteq U} \frac{|X|}{|U|} = \sum_{X \subseteq U} \text{generality}(X \rightarrow D_{\text{max}}(X)),$$

where $\text{generality}(X \rightarrow D_{\text{max}}(X))$ is the generality for the rules involving $X$ with the maximum precision value. That is, the overall generality of a rule set is the sum of the generality of individual rules in the set.

According to the relationships between micro and macro evaluation measures analyzed above, we can easily obtain the following outcomes.

For each individual rule in the set, let $\alpha \in (0, 1]$ denote the precision bound,

$$\text{precision}(X \rightarrow D_{\text{max}}(X)) \geq \alpha \Rightarrow \text{precision}(RS) \geq \alpha.$$

The outcome shows that the precision bound of individual rules is the same as the precision bound of the rule set. This implies that if the precision of each individual rule in the set is above a certain threshold, then the precision of the rule set is above the same threshold. However, the reverse is not necessarily true.

Similarly, we can get the following outcome for generality and complexity measures.

$$\text{generality}(X \rightarrow D_{\text{max}}(X)) \leq \text{generality}(RS).$$

and

$$\text{number of attributes}(X \rightarrow D_{\text{max}}(X)) \leq \text{number of attributes}(RS).$$

This outcome shows the generality and complexity of each individual rule is smaller than or equal to the generality and complexity of the rule set.

5. Conclusion and Future Works

In the past, researchers in the field of machine learning and data mining tended to learn rules from data sets based on single rule evaluation, while the overall rule
induction system performance depends on the performance of a set of rules. Although some rule set evaluation resolutions have been proposed, but single rule and rule set evaluation are being considered as two separate processes that eventually compared by some experimental results. In this paper, we look into the direct relationships between these two types of evaluations and try to build a theoretical connection between them by introducing a general framework for rule evaluation.

Micro and macro evaluation are introduced, the former is to measure the single rule performance and the latter is to measure the overall system performance of a set of rules. Selected micro evaluation measures are reviewed and the drawbacks are analyzed. For macro evaluation measures, we investigate the relationships between rules in the set, and the rule conflict resolutions are introduced. Under our pre-defined conditions, we show that macro evaluation can be explicitly expressed by micro evaluation.

The relationships between rules in a set are complicated. There are many conditions need to be considered. Our proposed conflict resolution is only based on one of these conditions, and we only analyze the relationships between micro and macro evaluation based on selected measures. In future works, we need to look into other relationships between rules and analyze more evaluation measures. The complete work will bring more insights into rule evaluation in terms of the overall rule induction system performance.

Acknowledgements

This study is partially supported by an NSERC Canada Discovery grant. The authors thank anonymous reviewers for their constructive comments.

References