Chapter VIII
QROC: A Variation of ROC Space to Analyze Item Set Costs/Benefits in Association Rules

Ronaldo Cristiano Prati
UFABC Universidade Federal do ABC, Brazil

ABSTRACT

Receiver Operating Characteristics (ROC) graph is a popular way of assessing the performance of classification rules. However, as such graphs are based on class conditional probabilities, they are inappropriate to evaluate the quality of association rules. This follows from the fact that there is no class in association rule mining, and the consequent part of two different association rules might not have any correlation at all. This chapter presents an extension of ROC graphs, named QROC (for Quality ROC), which can be used in association rule context. Furthermore, QROC can be used to help analysts to evaluate the relative interestingness among different association rules in different cost scenarios.

INTRODUCTION

In numerous data mining applications, a key issue to discover useful and actionable knowledge from association rules is to properly select the (probably) most profitable rules out of a large set of generated rules. To this end, it would be desirable to properly take into account expected costs/benefits of different item sets of different rules.

Receiver Operating Characteristic (ROC) analysis was developed in the context of signal-detection theory in engineering applications to examine the association between the presence and absence of a signal and the ability of a detector to sense the difference between them (Egan, 1975). These methods have been widely adopted for use in machine learning and data mining research thanks to its ability in visualizing and organizing
classification performance without regard to class distributions or misclassification costs, therefore providing a direct and natural way to incorporate cost/benefit analysis for different classification thresholds (Provost, Fawcett, & Kohavi, 1998). Furthermore, the ROC convex hull (ROCCH) provides a straightforward tool to select possibly optimal models and to discard sub-optimal ones, independently from (and prior to specifying) the cost context or the class distribution (Provost & Fawcett, 2001).

Nevertheless, the use of ROC analysis in machine learning and data mining concentrates mostly in classification tasks, where the area under the ROC curve (AUC) is used as a performance measure (Bradley, 1997). Although it is possible to plot association rules in the ROC space, the properties that make ROC analysis suitable for cost/benefit analysis in classification context do not hold in the association rule context. These properties came from the assumption that, even though cost context or class distribution are free to vary, in classification context they vary in the same way for all rules. This is because in classification, all classification rules have the same attribute in the consequent (the “target class” attribute). However, in the association rule context, the consequent is an item set that may contain any item (attribute) that does not appear in the rule antecedent. Therefore, ROC analysis cannot be used in the same direct and natural way to incorporate costs/benefits in association rule context.

This chapter proposes the use of a variation of ROC graphs, named QROC, for association rule evaluation. The benefit of using a ROC-like graph for analyzing association rules is twofold: first, this framework can be used to prune uninteresting rules from a large set of association rules. Second, it can help the expert in analyzing the performance of the association rules in different cost scenarios.

This chapter is organized as follows: Section “ROC space” presents some concepts of ROC analysis in the classification rule evaluation context, as well as it describes some pitfalls of using ROC analysis to association rules evaluation. Section “Quality ROC – QROC” presents an extension to ROC analysis, named QROC, which can be used in the association rule context. This section also presents a geometric interpretation of the ROC re-scaling process from which the QROC space is derived, as well as how cost information can be used to derive cost-sensitive metrics for rule evaluation and geometric interpretations of some well-known rule measures used for association rules evaluation. Section “An illustrative example” presents example on how QROC can be used to analyze a set of association rules generated for basket analysis. Section “New trends and Related Work” present related research. Finally, Section “Concluding remarks” concludes.

**BACKGROUND**

Rule learning has been mainly addressed from two different perspectives: predictive and descriptive tasks. Rule learning in predictive tasks is mainly concerned in generating classification rules that form a classifier. On the other hand, rule generation in descriptive tasks focus in finding all rules over a certain confidence that summarizes the data. However, in a broader sense, rules (either predictive or descriptive) can be considered as an association of two binary variables: the rule antecedent and the rule consequent. Rules of the form:

\[
antecedent \rightarrow \text{consequent},
\]

where both antecedent and consequent are conjunctions of features (for classification rules, the consequent always refers to a single feature), whereas they do not have features in common. The antecedent is also called left-hand side, premise, condition, tail or body and the consequent is called right-hand side, conclusion or head. Throughout this chapter, we will also use the general notation...