

Scoring Hypotheses from Threat Detection Technologies: Analogies to Machine Learning Evaluation

Robert C. Schrag* and Masami Takikawa†

*Global InfoTek, Inc., 1920 Association Dr, Suite 200, Reston, VA 20191 USA, rschrag@globalinfotek.com

†Cleverset, Inc., 673 NW Jackson Ave, Corvallis, OR 97330 USA, takikawa@cleverset.com

The authors were employed by Information and Transport, Inc. (IET) when this work was performed.

Abstract

We have developed efficient methods to score structured hypotheses from technologies that fuse evidence from massive data streams to detect threat phenomena. We have generalized metrics (precision, recall, F-value, and area under the precision-recall curve) traditionally used in the information retrieval and machine learning communities to realize object-oriented versions that accommodate inexact matching over structured hypotheses with weighted attributes. We also exploit the object-oriented precision and recall metrics in additional metrics that account for the costs of false-positive and false-negative threat reporting.

We have reported on our scoring methods more fully previously; the present brief presentation is offered to help make this work accessible to the machine learning community.

Introduction

Information fusion—collecting individual, disparate items of information into coherent, structured reports to provide a holistic situation assessment—is qualitatively similar to machine learning classification in some ways and different in others. Both tasks produce hypotheses whose veracity may be tested against an available standard for a given problem instance. In information fusion, the standard, known as “ground truth,” may be developed as part of a live or simulation-based experimental process. In supervised learning, the standard comes from class labels associated with training instances.

In binary classification (deciding which instances are and which are not members of a target class), the standard partitions hypotheses crisply into true positives, true negatives, false positives, and false negatives, from which we may compute precision, recall, accuracy, and other metrics of interest. This is schematized in Figure 1. (Imagine that the visual cues are available to us as readers but that the technology under test must rely on contextual cues, not shown.)

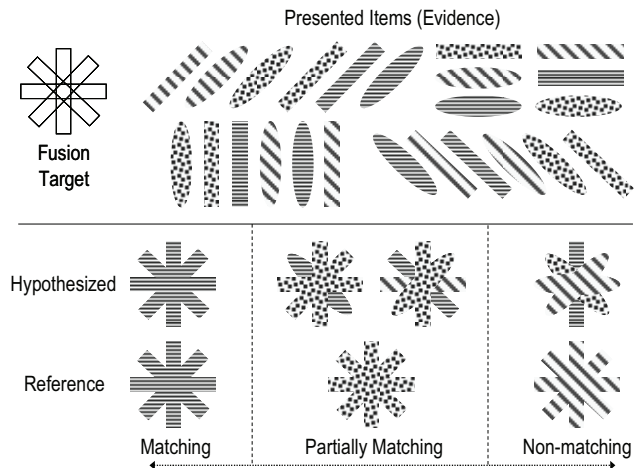


Figure 2: The information fusion task

In information fusion (even if there is only one target class), there are many possible combinations of existing information items into hypotheses that are structured as object instances with attribute values. We want to reward hypotheses that may not match a given ground truth instance exactly but are close. Approximate matching gives rise to a sort of continuum where the truth and falsity of being a match are not necessarily crisp as in binary classification. This is illustrated in Figure 2. (Imagine that the simple shapes falling along the four different axes correspond to outputs from four different sensors.) Along this continuum fall combinatorially many (possible) hypothesized instances that may defy the practical enumeration assumed by machine learning’s accuracy metric. In determining how closely two instances match each other, we sometimes need to accord different levels of importance or weight to their various attributes. For all these reasons, computing information fusion metrics demands a qualitatively different approach.

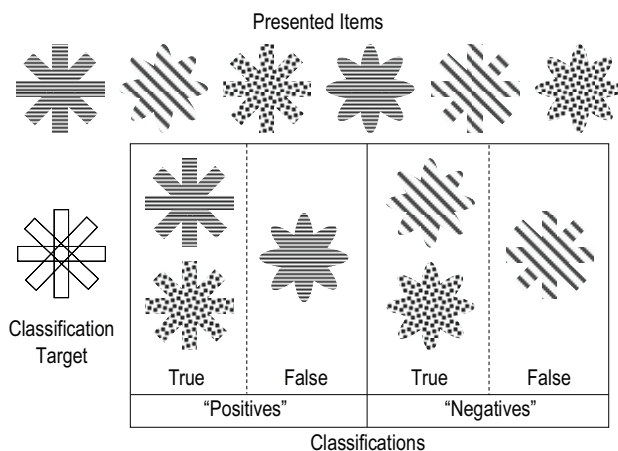


Figure 1: The binary classification task

