9 July 2002

Classification by Aggregating Emerging Patterns

ABSTRACT

Emerging patterns (EPs) are itemsets whose supports change significantly from one dataset to another; they capture multi-attribute contrasts between data classes, or trends over time. we propose a new classifier, CAEP, using the following main ideas based on EPs: (i) Each EP can sharply differentiate the class membership of a (possibly small) fraction of instances containing the EP, due to the big difference between its supports in the opposing classes; we define the differentiating power of the EP in terms of the supports and their ratio, on instances containing the EP. (ii) For each instance t, by aggregating the differentiating power of a fixed, automatically selected set of EPs, a score is obtained for each class. The scores for all classes are normalized and the largest score determines t's class. CAEP is suitable for many applications, even those with large volumes of high (e.g. 45) dimensional data; it does not depend on dimension reduction on data; and it is usually equally accurate on all classes even if their populations are unbalanced. Experiments show that CAEP has consistent good predictive accuracy, and it almost always outperforms C4.5 and CBA. By using efficient, border-based algorithms (developed elsewhere) to discover EPs, CAEP scales up on data volume and dimensionality.



G00002



162162