International Journal of Innovative Computing, Information and Control
Volume **5**, Number **5**, May **2009**

ICIC International ©2009 ISSN 1349-4198
pp. **1419–1428**

FINDING COVERAGE USING INCREMENTAL ATTRIBUTE COMBINATIONS

Jiyuan An and Yi-Ping Phoebe Chen

School of Engineering and Information Technology
Faculty of Science and Technology
Deakin University
Melbourne, VIC 3125 Australia
phoebe@deakin.edu.au

Received November 2007; revised April 2008

Abstract. *Coverage is the range that covers only positive samples in attribute (or feature) space. Finding coverage is the kernel problem in induction algorithms because of the fact that coverage can be used as rules to describe positive samples. To reflect the characteristic of training samples, it is desirable that the large coverage that cover more positive samples. However, it is difficult to find large coverage, because the attribute space is usually very high dimensionality. Many heuristic methods such as ID3, AQ and CN2 have been proposed to find large coverage. A robust algorithm also has been proposed to find the largest coverage, but the complexities of time and space are costly when the dimensionality becomes high. To overcome this drawback, this paper proposes an algorithm that adopts incremental feature combinations to effectively find the largest coverage. In this algorithm, the irrelevant coverage can be pruned away at early stages because potentially large coverage can be found earlier. Experiments show that the space and time needed to find the largest coverage has been significantly reduced.*
Keywords: Attribute combinations, Coverage, Concept learning

1. **Introduction.** Coverage is the range that covers only positive samples in attribute (or feature) space. Every example can be viewed as a point in the attribute space. Finding instances of coverage is the kernel problem in induction algorithms because coverage can describe positive samples as rules in concept learning. To reflect the characters of training samples, it is desirable that the coverage covers only positive samples and excludes negative samples. As the attribute space is usually very high dimensionality, finding large coverage, which covers more positive samples than any other coverage, is a very difficult task as a result of the so-called "curse of dimensionality" [2].

To determine the instances of coverage for all positive samples, many heuristic methods such as ID3 [12], AQ [7] and CN2 [3] have been proposed. These algorithms can be used effectively to solve practical problems [9,15-22]. They cannot, however, find the largest coverage. Sometimes we are asked to find the exact and largest coverage. The simplest manner is to enumerate all possible attribute combinations. However, it is very costly because of the exponential increase in number of attribute combinations.

The enormous improvements in computer performance have required the reconsidering of some algorithms that are too costly to be implemented. Now, even personal computers can possess 1G memory and 1GHz CPU. Based on development of hardware, we propose an algorithm by which to step-by-step find the largest coverage. Despite the required increase in memory space and CPU time, our algorithm can find the largest coverage to describe a specific class whereas heuristic methods, as mentioned earlier, cannot.