Chapter III

Exploring Similarities Across High-Dimensional Datasets

Karlton Sequeira, Rensselaer Polytechnic Institute, USA

Mohammed Zaki, Rensselaer Polytechnic Institute, USA

Abstract

Very often, related data may be collected by a number of sources, which may be unable to share their entire datasets for reasons like confidentiality agreements, dataset size, and so forth. However, these sources may be willing to share a condensed model of their datasets. If some substructures of the condensed models of such datasets, from different sources, are found to be unusually similar, policies successfully applied to one may be successfully applied to the others. In this chapter, we propose a framework for constructing condensed models of datasets and algorithms to find similar substructure in pairs of such models. The algorithms are based on the tensor product. We test our framework on pairs of synthetic datasets and compare our algorithms with an existing one. Finally, we apply it to basketball player statistics for two National Basketball Association (NBA) seasons, and to breast cancer datasets. The results are statistically more interesting than results obtained from independent analysis of the datasets.

Copyright © 2007, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.
Introduction

Often, data may be collected by a number of sources. These sources may be geographically far apart. There are a number of disadvantages in transferring the datasets from their source to a central location for processing. These include less reliability, security, higher computational and storage requirements, and so forth. It may be preferable to share condensed models of the datasets. Similarly, for reasons like confidentiality agreements, it may be required to use condensed models of datasets, which obfuscate individual details while conveying structural information about the datasets. Lastly, the datasets may have slightly different dimensionality or transformations like rotations, with respect to each other. This may preclude simply appending the datasets to each other and processing them.

If unusually similar substructure can be detected from the condensed models of some of the datasets, then policies successfully applied to one may be successfully applied to the others. For example, two consumer markets \(A\) and \(B\) differing in geography, economy, political orientation, or some other way may have some unusually similar consumer profiles. This may prompt sales managers in \(B\) to use successful sales strategies employed by sales managers in \(A\) for consumer profiles in which they are unusually similar. Also, profiles which are unusually dissimilar to any of those in the other graph are particularly interesting. The latter is analogous to the problem of finding contrast sets (Bay & Pazzani, 2001). Additionally, determining similarities and dissimilarities between snapshots of a dataset taken over multiple time intervals can help in identifying how the dataset characteristics evolve over time (Ganti, Gehrke, Ramakrishnan, & Loh, 1999).

A dataset may be a set of points drawn in possibly different proportions, from a mixture of unknown, multivariate, and perhaps non-parametric distributions. A significant number of the points may be noisy. There may be missing values as well. We currently assume that the dataset may belong to non-identical attribute spaces, which are mixtures of nominal and continuous variables. The datasets may be subject to translational, rotational, and scaling transformations as well. High-dimensional datasets are inherently sparse. It has been shown that under certain reasonable assumptions on the data distribution, the ratio of the distances of the nearest and farthest neighbors to a given target is almost 1 for a variety of distance functions and data distributions (Beyer, Goldstein, Ramakrishnan, & Shaft, 1999). Hence, traditional distance metrics which treat every dimension with equal importance have little meaning. Algorithms using such dissimilarity measures as a building block for application to high-dimensional datasets may produce meaningless results due to this lack of contrast.
Combining kNN Imputation and Bootstrap Calibrated: Empirical Likelihood for Incomplete Data Analysis
Yongsong Qin, Shichao Zhang and Chengqi Zhang (2010). International Journal of Data Warehousing and Mining (pp. 61-73).
www.igi-global.com/article/combining-knn-imputation-bootstrap-calibrated/46943?camid=4v1a