EFFICIENT ALGORITHMS FOR MINING ARBITRARY SHAPED CLUSTERS

By

Vineet Chaoji

An Abstract of a Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major Subject: COMPUTER SCIENCE

The original of the complete thesis is on file
in the Rensselaer Polytechnic Institute Library

Examining Committee:

Dr. Mohammed J. Zaki, Thesis Adviser
Dr. Boleslaw Szymanski, Member
Dr. Mark Goldberg, Member
Dr. Malik Magdon-Ismail, Member
Dr. Taneli Mielikäinen, Member

Rensselaer Polytechnic Institute
Troy, New York

July 2009
(For Graduation August 2009)

ABSTRACT

Clustering is one of the fundamental data mining tasks. Many different clustering paradigms have been developed over the years, which include partitional, hierarchical, mixture model based, density-based, spectral, subspace, and so on. Traditional algorithms approach clustering as an optimization problem, wherein the objective is to minimize certain quality metrics such as the squared error. The resulting clusters are convex polytopes in d-dimensional metric space. For clusters that have arbitrary shapes, such a strategy does not work well. Clusters with arbitrary shapes are observed in many areas of science. For instance, spatial data gathered from Geographic Information Systems, data from weather satellites, data from studies on epidemiology and sensor data rarely possess regular shaped clusters. Image segmentation is an area of technology that deals extensively with arbitrary shaped regions and boundaries. In addition to the complex shapes some of the above applications generate large volumes of data. The set of clustering algorithms that identify irregular shaped clusters are referred to as *shape-based clustering algorithms*. These algorithms are the focus of this thesis.

Existing methods for identifying arbitrary shaped clusters include density-based, hierarchical and spectral algorithms. These methods suffer either in terms of the memory or time complexity, which can be quadratic or even cubic. This shortcoming has restricted these algorithms to datasets of moderate sizes. In this thesis we propose SPARCL, a simple and scalable algorithm for finding clusters with arbitrary shapes and sizes. SPARCL has a linear space and time complexity. SPARCL consists of two stages – the first stage runs a carefully initialized version of the Kmeans algorithm to generate many small seed clusters. The second stage iteratively merges the generated clusters to obtain the final shape-based clusters. The merging stage is guided by a similarity metric between the seed clusters. Experiments conducted on a variety of datasets highlight the effectiveness, efficiency, and scalability of our approach. On large datasets SPARCL is an order of magnitude faster than the best existing approaches. SPARCL can identify irregular shaped clusters that are

full-dimensional, i.e., the clusters span all the input dimensions.

We also propose an alternate algorithm for shape-based clustering. In prior clustering algorithms the objects remain static whereas the cluster representatives are modified iteratively. We propose an algorithm based on the movement of objects under a systematic process. On convergence, the *core structure* (or the *backbone*) of each cluster is identified. From the core, we can identify the shape-based clusters more easily. The algorithm operates in an iterative manner. During each iteration, a point can either be subsumed (the term "globbing" is used in this text) by another representative point and/or it moves towards a dense neighborhood. The stopping condition for this iterative process is formulated as a MDL model selection criterion. Experiments on large datasets indicate that the new approach can be an order of magnitude faster, while maintaining clustering quality comparable with SPARCL.

In the future, we plan to extend our work to identify *subspace clusters*. A subspace cluster spans a subset of the dimensions in the input space. The task of subspace clustering thus involves not only identifying the cluster members, but also the relevant dimensions for each cluster. Indexing spatial objects using the seed selection approach proposed in SPARCL is another line of work we intend to explore.