

A Shrinking-Based Dimension Reduction Approach for Multi-Dimensional Data Analysis

Yong Shi and Aidong Zhang
Department of Computer Science and Engineering
State University of New York at Buffalo
Buffalo, NY 14260
{yongshi, azhang}@cse.buffalo.edu

Abstract

In this paper, we present continuous research on data analysis based on our previous work on the shrinking approach. Shrinking[2] is a novel data preprocessing technique which optimizes the inner structure of data inspired by the Newton's Universal Law of Gravitation[1] in the real world. It can be applied in many data mining fields. Following our previous work on the shrinking method for multi-dimensional data analysis in full data space, we propose a shrinking-based dimension reduction approach which tends to solve the dimension reduction problem from a new perspective. In this approach data are moved along the direction of the density gradient, thus making the inner structure of data more prominent. It is conducted on a sequence of grids with different cell sizes. Dimension reduction process is performed based on the difference of the data distribution projected on each dimension before and after the data-shrinking process. Those dimensions with dramatic variation of data distribution through the data-shrinking process are selected as good dimension candidates for further data analysis. This approach can assist to improve the performance of existing data analysis approaches. We demonstrate how this shrinking-based dimension reduction approach affects the clustering results of well known algorithms.

1. Introduction

With the advance of modern technology, the generation of multi-dimensional data has proceeded at an explosive rate in many disciplines. *Data preprocessing* procedures can greatly benefit the utilization and exploration of real data. *Shrinking*[2] is a novel data preprocessing technique which optimizes the inner structure of data inspired by the Newton's Universal Law of Gravitation[1] in the real world. It can be applied in many data mining fields.

In the previous work[2], we proposed a shrinking-based approach for multi-dimensional data analysis which consists of three steps: data shrinking, cluster detection, and cluster evaluation and selection. In the data-shrinking step, each data point moves along the direction of the density gradient and the data set shrinks toward the inside of the clusters. Data points are "attracted" by their neighbors and move to create denser clusters. The neighboring relationship of the points in the data set is grid-based. The data space is first subdivided into grid cells. Data points in sparse cells are considered to be noise or outliers and will be ignored in the data-shrinking process. Data-shrinking proceeds iteratively; in each iteration, we treat the points in each cell as a rigid body which is pulled as a unit toward the data centroid of those surrounding cells which have more points. Therefore, all points in a single cell participate in the same movement. The iterations terminate if the average movement of all points is less than a threshold or if the number of iterations exceeds a threshold.

2. Dimension Reduction

In this paper, we propose a shrinking-based dimension reduction approach for multi-dimensional data analysis to address the inadequacies of current clustering algorithms in handling multi-dimensional data. It tends to solve the dimension reduction problem from a new perspective. Dimension reduction process is performed based on the difference of the data distribution projected on each dimension through data-shrinking process. For those dimensions which make large contribution to the good results of data analysis (practically clustering process here), the alterations of the histogram variance of them through the data-shrinking process are significant. By evaluating the ratio of the histogram variances through data-shrinking process, good dimension candidates for further data analysis steps (e.g., clustering algorithms) can be picked out efficiently, and unqualified ones

are discarded. It can improve the clustering results of existing clustering algorithms.

2.1. Criteria for dimension reduction

Histogram variance of a dimension mentioned in the previous subsection can indicate the data distribution projected on that dimension to a certain degree. However, that is not necessarily the case. The alteration of the histogram variances through the data-shrinking process on each dimension reflects the characteristic aspects of the data distribution on the dimension much better than the histogram variance itself. For those dimensions which make large contribution to the good results of data analysis (practically clustering process here), the alterations of the histogram variance of them through the data-shrinking process are significant. By evaluating the ratio of the histogram variances through data-shrinking process on each dimension instead of the histogram variance itself, good dimension candidates for further data analysis steps (e.g., clustering algorithms) can be picked out efficiently, and unqualified ones are discarded.

2.2. Dimension reduction process

In this subsection we will present the details of the proposed approach. It is a grid-based approach.

Select reasonable grid scales. For grid-based approaches, it is crucial to properly select the grid size. However, proper grid size selection is problematical without prior knowledge of the structure of a data set. Instead of choosing a grid with a fixed cell size, we use a sequence of grids of different cell sizes. Details are in [2].

Perform data-shrinking and compute histogram variance alteration. After we get the set A of P reasonable grid scales, under each grid scale j, we perform the following steps:

a) For each dimension i, we first calculate the histogram variance $\sigma_{H_i}^2$.

b) The data-shrinking process is then performed. In the data-shrinking step, each data point moves along the direction of the density gradient and the data set shrinks toward the inside of the dense areas.

c) Let $\mu_{\tilde{H}_i}$ be the mean of the bin sizes of H_i , $|\tilde{H}_{ij}|$ be the size of bin H_{ij} after the data-shrinking process, we compute $\sigma_{\tilde{H}_i}^2$ as the histogram variance after the data-shrinking process:

$$\sigma_{\tilde{H}_i}^2 = \frac{\sum_j (|\tilde{H}_{ij}| - \mu_{\tilde{H}_i})^2}{\eta_i} \quad (1)$$

d) We evaluate the variance difference between the original histogram status and after-shrinking histogram status by

the ratio of the variances:

$$\gamma_{\sigma_{\tilde{H}_i}^2} = \tilde{\sigma}_{H_i}^2 / \sigma_{H_i}^2. \quad (2)$$

The sum of the histogram variance alterations on all the dimensions under a certain grid scale condition are calculated:

$$\Gamma = \sum_{i=1}^d \gamma_{\sigma_{\tilde{H}_i}^2} \quad (3)$$

Refine the set of grid scales. Suppose there are P reasonable grid scale candidates previously generated. We got P different histogram variance sums:

$$\mathbf{\Gamma} = \{\mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \dots, \mathbf{\Gamma}_p\}. \quad (4)$$

Under some grid scales, data can not be properly shrunk via the data-shrinking process. In other words, those grid scales can not help make the data distribution more piercing. In such cases, the variance differences through the data-shrinking process are not prominent. Those grid scales are discarded.

Select good dimension candidates. For those grid scales selected according to the sum of the histogram variance alterations, dimensions having significant histogram variance alteration through the data-shrinking process are selected as good candidates for clustering process, based on the integrated variance differences under these grid scales.

Under each selected grid scale λ , we sort dimensions D_1, D_2, \dots, D_d in descending order according to $\gamma_{\sigma_{\tilde{H}_i}^2}$. Suppose the dimension list after sorting is $\hat{D}_1, \hat{D}_2, \dots, \hat{D}_d$, we select the first several dimensions. The cut on the dimension list is performed as follows. To keep most valuable dimensions, the second half of the ordered dimension list is checked, and the cut spot is set on the first sharp descent dimension.

For each grid scale candidate, an ordered dimension list based on the histogram variance alteration is generated. The ultimate selection of valuable dimensions is based on the integration of the dimension selections on these qualified grid scales.

References

- [1] Rothman, Milton A. *The laws of physics*. New York, Basic Books, 1963.
- [2] Yong Shi, Yuqing Song and Aidong Zhang. A shrinking-based approach for multi-dimensional data analysis. In *the 29th VLDB conference*, pages 440–451, Berlin, Germany, September 2003.