Available online at www.sciencedirect.com
SCIENCE DIRECT®

Pattern Recognition 39 (2006) 424–431

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Clustering noisy data in a reduced dimension space via multivariate regression trees

Christine Smyth*, Danny Coomans, Yvette Everingham

Statistics and Intelligent Data Analysis Group, School of Mathematical and Physical Sciences, James Cook University, Townsville QLD 4811, Australia

Received 20 July 2005

Abstract

Cluster analysis is sensitive to noise variables intrinsically contained within high dimensional data sets. As the size of data sets increases, clustering techniques robust to noise variables must be identified. This investigation gauges the capabilities of recent clustering algorithms applied to two real data sets increasingly perturbed by superfluous noise variables. The recent techniques include mixture models of factor analysers and auto-associative multivariate regression trees. Statistical techniques are integrated to create two approaches useful for clustering noisy data: multivariate regression trees with principal component scores and multivariate regression trees with factor scores. The tree techniques generate the superior clustering results.

© 2005 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.

Keywords: Cluster analysis; Noise variables; Multivariate regression trees; Dimension reduction

1. Introduction

A characteristic common to many data sets is “noise variables”. Noise variables contain no relevant information and mask the underlying structure of the data set. The prevalence of noise variables in data sets is increasing: an unavoidable consequence as the size of data sets increases.

It is known that care must be taken when clustering large noisy data sets because including superfluous variables may induce spurious clusters or blur existing cluster boundaries. The researcher must use a clustering algorithm suitable for noisy high dimensional data sets. These clustering algorithms will intrinsically incorporate dimension reduction.

Here we trial auto-associative multivariate regression trees [1] using noisy data. We further extend multivariate regression trees as a clustering technique by incorporating dimension reduction, and demonstrate their capabilities when clustering noisy data. We reduce the dimension of the data set globally using either factor or principal components analysis, and subsequently cluster in the reduced factor or principal components space via the regression tree. The capabilities of mixture models of factor analysers [2,3], a clustering technique featuring local dimensionality reduction, are also investigated.

We assess the potential of these algorithms when clustering noisy data sets, by perturbing two data sets via the introduction of superfluous variables. Clustering techniques are applied to the perturbed data set and their capabilities to recover the known clusters are gauged. This error perturbation experiment has been used previously by Milligan [4]. We extend Milligan’s experiment to include real-life data sets, whilst benchmarking to the classical *K*-means technique. Our results show the superiority of multivariate regression trees with principal component scores and/or factor scores when clustering noisy data.

2. Theory

2.1. Principal components analysis and factor analysis

Both principal components analysis and factor analysis are dimension reduction techniques. Principal components

* Corresponding author. Tel.: +617 4781 4237;
fax: +617 4781 5880.
E-mail address: christine.smyth@jcu.edu.au (C. Smyth).

0031-3203/$30.00 © 2005 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2005.09.003

analysis attempts to model the total variance of the original data set, via new uncorrelated variables called principal components [5]. The principal components are linear combinations of the original variables:

$$\boldsymbol{y} = A^{\mathrm{T}}\boldsymbol{x},$$

where $\boldsymbol{x}$ is a vector of the original variables; $\boldsymbol{y}$ is a p element vector of principal component scores; and A is obtained from the spectral decomposition of Σ.

There are p principal components. However, the first q principal components usually account for most of the variation within the data set. Dimension reduction is achieved by discarding the latter $p - q$ principal components. Then Σ is approximated by its first q eigenvectors. Observational units are represented in the q dimensional subspace via the first q principal component scores.

Factor analysis attempts to explain the variables by assuming that they can be generated as a linear combination of q unobservable common factors (usually $q \ll p$) plus a unique factor [5]. The factor analysis model is given by

$$\boldsymbol{x} = \boldsymbol{\mu} + F\boldsymbol{z} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\boldsymbol{\mu}$ is a mean vector; F is a $p \times q$ matrix of factor loadings; $\boldsymbol{z}$ is a q dimensional vector of hypothetical common factors; and $\boldsymbol{\varepsilon}$ is a unique factor. Because the $\boldsymbol{z}$ are hypothetical, imposing assumptions $\boldsymbol{z} \sim \mathrm{N}(0, I_q)$ and $\boldsymbol{\varepsilon} \sim \mathrm{N}(0, D)$ allows the estimation of F.

We see that unlike principal components analysis, factor analysis distinguishes between common and unique variance. The factor analysis model implies that $\Sigma = FF^{\mathrm{T}} + D$. The $p \times q$ matrix F contains the factor loadings. The factor loadings are the correlations of a variable with a common factor $\boldsymbol{z}$ [6]. D, a diagonal matrix, contains the specific variances of each variable: the unique variance of each variable that is not associated with the other variables. Therefore, the $p \times p$ covariance matrix Σ is modelled by a $p \times q$ matrix F and a diagonal matrix D, implying a substantial amount of dimension reduction if $q \ll p$.

Unlike principal components analysis, Eq. (1.1) shows that factor analysis does not provide a unique transformation from factors to variables. In fact, the solution can be rotated to make it more interpretable. Observational units can be represented in the q dimensional factor space by the estimated values of the hypothetical common factors, called “factor scores”.

As a data set increases in size, the factor analysis solution is likely to approach the principal components solution [5]. Despite this fact, we use both factor scores and principal component scores as response variables in the multivariate regression trees and investigate any differences.

2.2. Multivariate regression trees and auto-associative multivariate regression trees

Regression Trees begin with all the data contained within a single node. The root node is then split in two on the basis of the value of an explanatory variable so as to make the two new nodes more homogenous with respect to the response variables. The splitting process is continued until the terminal nodes (nodes not split in two) are sufficiently homogenous.

Mathematically, the binary decision function that spits a node is chosen such that it maximizes the decrease in $R(T)$ [7]. $R(T)$ is given by

$$R(T) = \frac{1}{n}\sum_{i \in \tilde{T}}\sum_{\boldsymbol{x}_i \in t}\left(\boldsymbol{y}_i - \bar{\boldsymbol{y}}(t)\right)^2, \tag{1.2}$$

where $\boldsymbol{x}_i$ is the vector of measurements of p explanatory variables for the ith observational unit; $\boldsymbol{y}_i$ is the vector of measurements of the response variables for the ith observational unit; $\tilde{T}$ is the set of all nodes and; $\bar{\boldsymbol{y}}(t)$ is the average response vector of node t.

Observational units within a terminal node are similar to each other with respect to the response variables. By replicating the explanatory variables as the response variables (that is, using identical response and explanatory variables) an auto-associative multivariate regression tree can be used as a divisive clustering technique.

We suggest a relaxed criterion for selecting the natural number of clusters found by an auto-associative multivariate regression tree: the “elbow” of the tree’s relative error curve. This tree will not attain optimal predictive performance, but any further splitting will result in only a small decrease in the heterogeneity of the terminal nodes. Therefore, at the number of nodes indicated by the location of the elbow, the clusters are sufficiently homogenous.

The location of the elbow is questionable. We deemed the elbow as the point, $\widehat{k}$, where the gradient of the relative error curve changed from being steep to gentle. Specifically, we chose $\widehat{k}$ as the k that minimized

$$abs\left(\frac{RE(k+1) - RE(k)}{RE(k) - RE(k-1)}\right),$$

where $RE(k)$ is the value of the relative error curve at k.

2.2.1. Multivariate regression trees with principal component scores and factor scores

Clustering via an auto-associative multivariate regression tree by replicating the explanatory variables as response variables is computationally intensive if the data set is large. Moreover, including redundant variables as response variables may induce suboptimal results. Reducing the dimension of the response variables may produce more stable results. We propose two techniques for reducing the

Download English Version:

https://daneshyari.com/en/article/531260

Download Persian Version:

https://daneshyari.com/article/531260

Daneshyari.com