Pattern Recognition Letters 77 (2016) 28–34

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Improving clustering performance by incorporating uncertainty☆

Maha Bakoben[a,*], Anthony Bellotti[a], Niall Adams[a,b]

[a] Department of Mathematics, Imperial College London, London SW7 2AZ, United Kingdom
[b] Heilbronn Institute for Mathematical Research, University of Bristol, Bristol BS8 9AG, United Kingdom

ARTICLE INFO

Article history:
Received 25 July 2015
Available online 11 March 2016

Keywords:
Clustering with uncertainty
Ellipsoid dissimilarity measures
Confidence ellipsoids
Time series clustering

ABSTRACT

In more challenging problems the input to a clustering problem is not raw data objects, but rather parametric statistical summaries of the data objects. For example, time series of different lengths may be clustered on the basis of estimated parameters from autoregression models. Such summary procedures usually provide estimates of uncertainty for parameters, and ignoring this source of uncertainty affects the recovery of the true clusters. This paper is concerned with the incorporation of this source of uncertainty in the clustering procedure. A new dissimilarity measure is developed based on geometric overlap of confidence ellipsoids implied by the uncertainty estimates. In extensive simulation studies and a synthetic time series benchmark dataset, this new measure is shown to yield improved performance over standard approaches.

1. Introduction

Cluster analysis is a well-known unsupervised technique, which aims at assigning data objects into homogeneous groups [4]. Groups are formed based on a predefined dissimilarity measure between objects. Hence, the performance of clustering algorithms is highly affected by the dissimilarity measure used in the clustering process.

Two important properties of any data set are often ignored in the conventional computation of dissimilarities between objects using standard measures such as Euclidean distance, Manhattan distance and many others: the first property is error information associated with each variable, and the second is the dependence between these variables. Discarding such properties is more likely to result in imprecise cluster assignments.

We present a dissimilarity measure that considers error information associated with n data objects in a p-dimensional space $\{Y_i\}_{i=1}^n = \{(\mathbf{y}_1, \ldots, \mathbf{y}_p)_i^T\}$, where each element $\{\mathbf{y}_j\}_{j=1}^p$ is a vector-valued quantity which could be one value or a time series. Note that the proposed measure can thus handle different length time series objects. In this paper, data objects $\{Y_i\}_{i=1}^n$ are represented by vectors of model coefficients $\hat{\boldsymbol{\beta}}_i$ for $i = 1, \ldots, n$, which are estimated from some statistical models. Such statistical models estimate the covariance matrix $\Psi(\hat{\boldsymbol{\beta}}_i)$ for $i = 1, \ldots, n$, which defines error information or the uncertainties of model coefficients; these terms will be used interchangeably throughout this paper. We consider $\Psi(\hat{\boldsymbol{\beta}}_i)$ in the computation of dissimilarities between model coefficients to obtain consistent cluster results.

The importance of incorporating the uncertainty associated with point estimates in dissimilarity measures is illustrated by the example in Fig. 1. Coefficient estimates in Fig. 1 are sampled from two-dimensional Gaussian distributions with small variances. These coefficients are clustered using the well-known k-medoid clustering methods [4,7] with standard Euclidean distance. As can be seen, perfect cluster assignments are obtained using Euclidean distance (left plot in Fig. 1). However, this standard dissimilarity measure frequently fails to identify the correct clusters when data points are impaired by high amount of variability as shown in the right plot of Fig. 1. Standardising variables often handles problems of location and scale, however the correlations between variables are discarded. Cluster analysis with Euclidean distance might be superior only in the absence of correlation between variables.

Although the consideration of the implicit uncertainty in cluster analysis is an active research topic, only a few studies have incorporated the uncertainty explicitly. An early approach to this problem was developed by Chaudhuri and Bhowmik [3], however its applicability is limited to uniformly distributed error. Another paper by Kumar and Patel [8] in which they consider clustering model coefficients with uncertainty. The dissimilarity measure suggested in [8] is the Mahalanobis distance, which incorporates uncertainty explained by the covariance matrix between coefficient estimates. The covariance matrix in the Mahalanobis

☆ This paper has been recommended for acceptance by Dr. G. Moser.

* Corresponding author. Tel.: +44 (0) 784 108 0583.
E-mail address: m.bakoben11@ic.ac.uk, maha_bakoben@yahoo.com (M. Bakoben).

http://dx.doi.org/10.1016/j.patrec.2016.03.004
0167-8655/

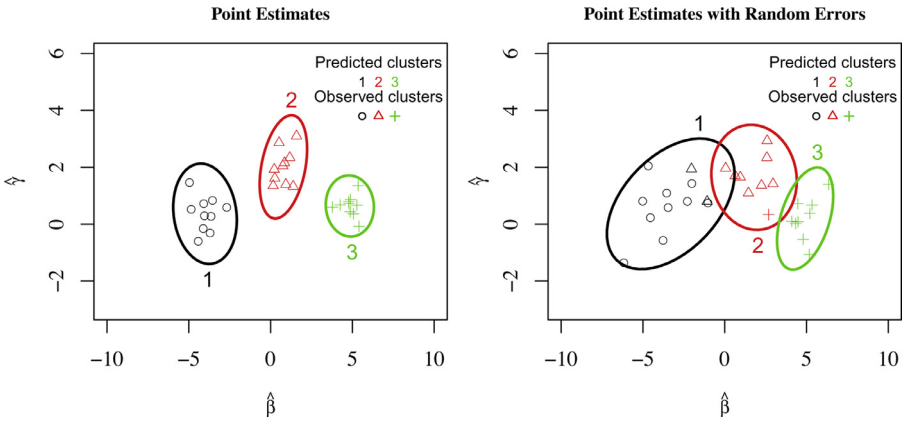

Fig. 1. Data points in both plots are random sample drawn from bivariate Gaussian distributions with means $(\beta_1, \gamma_1)^T = (-4, 0.5)^T$, $(\beta_2, \gamma_2)^T = (1, 2)^T$ and $(\beta_3, \gamma_3)^T = (5, 0.5)^T$ for clusters $\mathcal{C}_1, \mathcal{C}_2$ and $\mathcal{C}_3$, whereas the variances are assumed to be zero in the left plot, and have some positive values in the right plot. Each boundary indicates an iso-density contour of the generating distribution of the predicted cluster. Euclidean distance was used in the clustering process.

distance assigns higher weights to distance between estimates if the orientations of their distributions are different [11], although the degree of similarity between estimates with different orientations of distributions might be higher than estimates in the same orientations. In response to these shortcomings, we propose a dissimilarity measure that is defined based on geometric overlaps between the distributions of coefficient estimates regardless of the orientations. Another advantage of the proposed dissimilarity measure is the ability to achieve good clustering performance without making any assumption on the error distribution as the measure is defined on confidence regions however they are constructed, whereas the Mahalanobis distance is only optimal for clustering data from Gaussian distributions.

Our dissimilarity measure considers the geometrical overlap of uncertainties associated with coefficient estimates which define a $(1-\alpha)\%$ confidence region. This confidence region is represented by an ellipsoid with parameters $\boldsymbol{\mu}$ and Ψ; they represent the centre of the ellipsoid defined by the point estimates, and its orientation and size which are defined by the covariance matrix between estimates. In the computation of the proposed dissimilarity measure, we consider part of the joint distribution that is determined by the significance level α. Optimal clustering performance is expected before the confidence region reaches the 100% confidence level. In principle, the free parameter α could be tuned by computation procedures with a proxy measure for clustering performance.

The advantage of the ellipsoid dissimilarity measure is the ability to identify clusters from all possible values of an estimate rather than producing results that are restricted to single values obtained from point estimates. Therefore, outcomes of clustering using the ellipsoid confidence regions are expected to be more stable than the clustering using conventional dissimilarity measures. We test the performance of the proposed ellipsoid dissimilarity measure on simulated data from bivariate Gaussian distributions and on coefficient estimates from vector autoregression models.

This paper is structured as follows: Sections 2 and 3 describe in detail the proposed ellipsoid based dissimilarity measure. Section 4 includes the outcomes of the simulation studies. In Section 5, we test the performance of the proposed measure on a benchmark dataset of control charts. Finally we present a summary including directions for future work, in Section 6.

2. Geometrical representation for uncertainty

Assume a vector of coefficients $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ is estimated by fitting a statistical model to some data. The uncertainty associated with the estimate $\hat{\boldsymbol{\beta}}$ is represented by the estimated covariance matrix $\widehat{\Psi} \in \mathbb{R}^{p\times p}$, where $\widehat{\Psi}$ is a positive definite matrix.

2.1. Confidence interval ellipsoids

In the proposed approach of incorporating uncertainty in the dissimilarity measure, the uncertainty of coefficient estimates is geometrically represented by an ellipsoid $\mathcal{E}(\boldsymbol{\mu}, \widehat{\Psi})$ defined by,

$$\mathcal{E}(\boldsymbol{\mu}, \widehat{\Psi}) : \{(\mathbf{x}-\boldsymbol{\mu})^T (c\widehat{\Psi})^{-1} (\mathbf{x}-\boldsymbol{\mu}) \leq 1\}, \tag{1}$$

where the centre of the ellipsoid is the point estimate $\boldsymbol{\mu} = \hat{\boldsymbol{\beta}}$ under the model and the size and orientation are defined by the estimated covariance matrix $\widehat{\Psi}$. Computations of principal axes and radii of ellipsoids are involved in the computation of the dissimilarity measure, in which principal axes are defined by the eigenvectors of $c\widehat{\Psi}$ denoted by $\{\mathbf{v}_i\}_{i=1}^p$ and radii are obtained from the corresponding eigenvalues $\{\lambda_i\}_{i=1}^p$ by $\{\frac{1}{\sqrt{\lambda_i}}\}_{i=1}^p$ [6].

That is, ellipsoids are constructed to represent confidence regions for simultaneous inference about the coefficients $\hat{\boldsymbol{\beta}} = (b_1, \ldots, b_p)$. These confidence regions are controlled by the scalar multiplier $c = t_{\tau,1-\alpha/2}$, that is the quantile of t-distribution where $\tau = n - p - 1$. For example, when $n > 30$, $p = 2$ and $c = 1$, the representation of an ellipsoid in two-dimension defines an ellipse which corresponds to a 0.84 confidence region for the inference on (b_1, b_2), as shown in Fig. 2. Note that the true values of (b_1, b_2) can be anywhere inside the ellipse.

Typically, the projections of ellipsoids on the coordinate axes define the confidence intervals for the individual coefficient by,

$$b_j \pm t_{\tau,1-\alpha/2}\ S_{b_j}, \quad j = 1, \ldots, p. \tag{2}$$

The confidence interval level in Eq. (2) corresponds to $(1-\alpha)\%$ quantile of t-distribution, where S_{b_j} is an estimate of σ_{b_j} the square root of the uncertainty associated with b_j in the diagonal of $\widehat{\Psi}$.

However, the previous multiplier c defines the individual $(1-\alpha)\%$ confidence intervals for $\hat{\boldsymbol{\beta}}$ but the joint confidence level is reduced to $(1-p\alpha)\%$. In order to increase the confidence level for joint inference, Bonferroni confidence intervals [12] can be constructed. Bonferroni's adjustment to the significance level for the inference about the individual parameter is $1-\alpha/(2p)$ to obtain a $(1-\alpha)\%$ conservative confidence region for the joint inference on $\hat{\boldsymbol{\beta}} = (b_1, \ldots, b_p)$. Thus, we only need to replace the multiplier $t_{\tau,1-\alpha/2}$ by $t_{\tau,1-\alpha/(2p)}$ in Eq. (1) to achieve the required level of confidence. Notice that the volume of the ellipsoid with