Egyptian Informatics Journal (2011) **12**, 185–196

Cairo University

Egyptian Informatics Journal

www.elsevier.com/locate/eij
www.sciencedirect.com

ORIGINAL ARTICLE

Using incremental general regression neural network for learning mixture models from incomplete data

Ahmed R. Abas

Department of Computer Science, University College in Lieth for Male Students, Umm Al-Qura University, Makka Al-Mukarrama, Lieth, Kingdom of Saudi Arabia

Received 13 April 2011; revised 8 July 2011; accepted 25 July 2011
Available online 25 August 2011

KEYWORDS

Cluster analysis;
Expectation maximization;
Finite mixture models;
Incomplete data;
Incremental general regression neural network;
Local correlations

Abstract Finite mixture models (FMM) is a well-known pattern recognition method, in which parameters are commonly determined from complete data using the Expectation Maximization (EM) algorithm. In this paper, a new algorithm is proposed to determine FMM parameters from incomplete data. Compared with a modified EM algorithm that is proposed earlier the proposed algorithm has better performance than the modified EM algorithm when the dimensions containing missing values are at least moderately correlated with some of the complete dimensions.

1. Introduction

FMM is a partitional probabilistic algorithm that is commonly used in cluster analysis [1–17]. Parameters of the FMM are usually determined via the EM algorithm [18] from complete data. It is shown that when the mechanism that describes the occurrences of missing values in the data is missing at random (MAR), maximum likelihood inferences about parameters of the FMM that represents the data can be made from complete data [19]. In other words, the likelihood is simply the density of the complete data, which is a function of the FMM parameters [20]. Missing data are considered MAR if the probability of having missing values in a certain dimension of the data depends on values of the other complete dimensions but not on the true values of missing values [21]. However, determining FMM parameters using only complete data requires the data size to be large in order to obtain good fitting of the data distribution [22]. Due to practical problems in factorising the likelihood, it is commonly maximized iteratively via the EM algorithm for complete data. It is shown that the EM algorithm can be used in determining parameters of a multivariate normal distribution from incomplete data [21]. Missing values in the data are estimated using multivariate regression. The complete data with the residual covariances of the estimated values are then used in estimating the mean and the covariance matrix of the multivariate normal distribution. The regression

E-mail address: armohamed@uqu.edu.sa

Peer review under responsibility of Faculty of Computers and Information, Cairo University.
doi:10.1016/j.eij.2011.07.001

Production and hosting by Elsevier

coefficients and the residual covariances are estimated from the current model parameters such that the likelihood is maximized.

The EM algorithm is modified such that it can determine parameters of a mixture of multivariate normal distributions from incomplete data [23]. Missing values and some other statistics are estimated in the E-step from each model component as they are come from that component. These values with the observed ones are then used in the M-step to determine the parameters of that model. However, accuracy of the estimated values is limited due to the small number of estimators (kernels) used in the estimation of missing values. The number of these estimators equals to the number of components in the FMM. This in turn affects the accuracy of the leaned FMM parameters and hence its clustering results. This problem can be overcome if missing values in the data are estimated with high accuracy first, and then FMM parameters are learned via the EM algorithm. The computational complexity of the modified EM algorithm [23] is reduced during the EM iterations by incorporating two types of auxiliary binary indicator matrices corresponding to the observed and unobserved components of each datum [24]. The resulting EM algorithm is compared with a new proposed Data Augmentation (DA) computational algorithm for learning normal mixture models when the data are missing at random [24]. Experimental results show that DA imputation has considerable promising accuracy in the prediction of missing values when compared to the EM imputation, especially when the missing rate increases. However, both algorithms impute missing values using mixture model parameters and hence their imputations are sensitive to the prior information about density functions of mixture components and the size of data that are fully observed. A supervised classification method, called robust mixture discriminant analysis (RMDA), is proposed to handle label noised data [25]. The RMDA algorithm uses only fully observed data to learn mixture model parameters, and then uses the resulting mixture model to estimate labels and detect noisy ones. However, imputations made by the RMDA algorithm are sensitive to prior information about density functions of mixture components, the size of data that are fully observed and assumptions such as all the uncertain labels are in one feature.

In this paper, a new algorithm for determining FMM parameters from incomplete data is proposed. The proposed algorithm is a combination of the Incremental General Regression Neural Network (IGRNN) [26] and the EM algorithm [18]. It estimates missing values in the data using the IGRNN and then uses the EM algorithm to determine FMM parameters. Performance of the proposed algorithm is investigated against the use of the modified EM algorithm [23] in learning FMM from incomplete data. The motivation of this investigation is to test the effect of the accuracy of the estimated values of missing data from both algorithms on the clustering behaviour of the resulting mixture models. This paper is organised as follows. Section 2 describes the modified EM algorithm [23]. Section 3 describes the proposed algorithm. Section 4 describes experiments that are carried out to compare both of the described algorithms. Section 5 discusses the results obtained from experiments. Section 6 concludes the paper and summarises its findings.

2. The modified EM algorithm [23]

The modified EM algorithm [23] is proposed for determining parameters of a mixture of multivariate normal distribution from incomplete data provided that missing values are Missing At Random (MAR) [21]. We refer to this algorithm as the EMH algorithm in the rest of this paper. The EMH algorithm is described as follows.

Suppose that $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is a data set that is composed of n patterns and d dimensions such that each pattern is represented as $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{id}]^T$. This data set is assumed to be generated randomly from a mixture of K multivariate distributions with unknown mixing coefficients $p(c)$, where c = 1, 2,…,K. Let the probability density component of $\mathbf{x}_i$ from the kth multivariate distribution be $p(\mathbf{x}_i| \, k)$. The commonly used distribution is the Gaussian $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$, where $\boldsymbol{\mu},\boldsymbol{\Sigma}$ are the mean and the covariance matrix, respectively [23]. This distribution is preferred to other distributions for the EM algorithm because it has a small number of parameters that need to be estimated and also computing its derivative is simple. The density of $\mathbf{x}_i$ can be written as $p(\mathbf{x}_i) = \sum_{c=1}^{K} p(c)p(\mathbf{x}_i|c)$, where $\sum_{c=1}^{\mathrm{K}} p(c) = 1, 0 \leqslant p(c) \leqslant 1$, for $c = 1, 2, \ldots, \mathrm{K}$.

When $\mathbf{X}$ contains MAR missing values, the pattern $\mathbf{x}_i$ can be denoted as $\mathbf{x}_i = (\mathbf{x}_{i,obs}, \mathbf{x}_{i,mis})$, where $\mathbf{x}_{i,obs}$ stands for the observed values, and $\mathbf{x}_{i,mis}$ stands for the missing values for pattern $\mathbf{x}_i$. In fitting the FMM, there are two types of missing values that have to be considered; one is the values of the cluster membership dimensions for each pattern $\mathbf{z}_i = [z_{i1}, z_{i2}, \ldots, z_{iK}]^T$, where $i = 1, 2, \ldots, n$ and the other is the missing values in the data matrix $\mathbf{X}$. Each value in the cluster membership vector z_{ij} represents the probability by which a certain pattern $\mathbf{x}_i$ in the data matrix $\mathbf{X}$ is generated from the jth component in the FMM. In the EM algorithm, $\mathbf{z}_i$ is approximated by the posterior probabilities when $\mathbf{x}_i$ is fed to the model. In the E-step, all $\hat{\mathbf{z}}_i$'s are determined besides some statistical moments necessary for the M-step using observed values for each pattern. While in the M-step, the new estimates of the FMM parameters are determined using the observed data and the statistical moments determined in the E step. Both the E and the M steps are alternated until convergence. For more details and description of this algorithm see [23].

3. The proposed algorithm for determining FMM parameters from incomplete data

The Incremental General Regression Neural Network (IGRNN) [26] is proposed for estimating missing values in numeric data sets. It is shown that the IGRNN produces highly accurate estimations for missing values in the case of a data set that has strong correlations among its dimensions [26]. In this Section, it is proposed to combine this algorithm with the EM algorithm [18] for determining FMM parameters from incomplete data. First, the proposed algorithm estimates missing values in the data set using the IGRNN. Second, it estimates parameters of the FMM that can be used in clustering the data using the resulting complete data and the EM algorithm. The proposed algorithm is referred to as the IGRNNEM algorithm in the rest of this paper. The IGRNNEM algorithm is described as follows: