Kümeleme ve Maksimum Olabilirlik Yaklaşımıyla Eksik Veri Tamamlama

A Missing Data Imputation Approach Using Clustering and Maximum Likelihood Estimation

Muammer ALBAYRAK[1], Kemal TURHAN[1], Burçin KURT[1]
[1]Department of Biostatistics and Medical Informatics, Karadeniz Technical University, Trabzon, Turkey
{m.albayrak, kemalturhan, burcinkurt}@ktu.edu.tr

***Özetçe*—Eksik veri sağlık alanındaki verilerde çeşitli nedenlerden dolayı sıklıkla karşılaşılan, veri analizi ve karar verme süreçlerini olumsuz etkileyen bir veri madenciliği problemidir. Eksik veri tamamlama probleminde yöntemin başarısı, verinin karakteristiği ve eksik verinin tipi gibi birçok faktörden etkilendiğinden halen önemli bir araştırma konusudur. Bu çalışmada, eksik veri problemine yönelik kümeleme ve maksimum olabilirlik (MO) tabanlı bir yaklaşım önerilmiştir. Önerilen yöntemde, UCI uluslararası açık kaynaklı veritabanındaki Dicle Üniversitesi Tıp Fakültesi tarafından hazırlanan "Mezotelyoma (Mesothelioma)" veri seti kullanılmış ve ilk olarak tamamen rasgele olarak kayıp (Missing completely at random;MCAR), rasgele olarak kayıp (Missing at random;MAR) ve rasgele olmayan kayıp (Missing not at random; MNAR) eksik veri örüntülerine uygun yeni veri setleri oluşturulmuştur. İkinci adımda, bu yeni veri setleri MO yönteminin hesaplama başarısını artırmak amacıyla eksik veri bulunan 3 özelliğin dâhil edilmediği bir k-means kümeleme işlemine tabi tutularak 10 kümeye ayrılmıştır. Son adımda, eksik verili özelliklerin tekrar eklendiği bu kümeler için MLE yöntemi ile eksik veriler tamamlanmış ve kümeler birleştirilerek tam veri seti elde edilmiştir. Üç adımda (veri eksiltme, kümeleme ve veri tamamlama) tamamlanan işlemler sonucunda elde edilen yeni veri setleri orijinal veri seti ile hata kareler ortalamasının karekökü (Root mean square Error;RMSE) kriterine göre karşılaştırılmış ve ortalama %96,5 başarı elde edilmiştir.**

Anahtar Kelimeler — Eksik veri; Kümeleme; MLE; KMO yaklaşımı.

***Abstract*—Missing data is a data mining problem that adversely affects data analysis and decision making processes that are frequently encountered in healthcare data for a variety of reasons. Missing data is still an important research topic because the success of the method is influenced by many factors such as the characteristics of the data and the type of the missing data. In this study, a clustering and maximum likelihood estimation (MLE) based approach to the missing data problem is proposed. In order to test the proposed method, the "Mesothelioma" (Mesothelioma) data set prepared by the Dicle University Medical School and uploaded to UCI international open source database was used. New data sets have been created that are compatible with missing data patterns such as Missing completely at random (MCAR), Missing at random (MAR), and Missing not at random (MNAR). In the second step, these new data sets are divided into clusters in order to increase the computation success of the MLE method by a k-means clustering process in which 3 features with missing data are not included. In the last step, the missing data are completed with the MLE method for these clusters in which the features with missing values are added again, and the clusters are merged to obtain the complete data set. The new data sets obtained as a result of the completed operations in three steps (data reduction, clustering and data completion) were compared with the original data set according to the root mean square error (RMSE) criterion, and an average of 96.5% success was achieved.**

Keywords — Missing data; Clustering; MLE; CMLE approach.

I. INTRODUCTION

Missing data is a data mining problem that adversely affects data analysis and decision making processes that are frequently encountered in healthcare data for a variety of reasons. It also leads to biased estimates of parameters, loss of information, decreased statistical power, increased standard errors, and weakened generalizability of findings [9]. Missing data is still an important research topic because the success of the method is influenced by many factors such as the characteristics of the data and the type of the missing data [1, 5, 9]. In this study, a clustering and maximum likelihood estimation (MLE) based approach is proposed to solve missing data problem. The main purpose of this approach is to increase the sensitivity and performance of estimating missing data points by MLE method by dividing the data into clusters before imputation.

978-1-5386-0633-9/17/$31.00 ©2017 IEEE

Missing data is a data mining problem that occurs in many different ways in the health field. Three different types of missing data are defined in the literature according to pattern type [1, 9].

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Missing Not at Random (MNAR)

In order to explain these mechanisms, suppose variable Y has some missing values. We will say that these values are MCAR if the probability of missing data on Y is unrelated to the value of Y itself or to the values of any other variable in the data set. We will say that these values are MAR if the probability of missing data on Y is unrelated to the value of Y after controlling for other variables in the analysis. And we will say that these values are MNAR if the missing values do depend on unobserved values [1].

Figure 1 shows the three different types of missing data patterns.

	Job performance ratings			
IQ	Complete	MCAR	MAR	MNAR
78	9	—	—	9
84	13	13	—	13
84	10	—	—	10
85	8	8	—	—
87	7	7	—	—
91	7	7	7	—
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	7	—	7	—
99	7	7	7	—
105	10	10	10	10
105	11	11	11	11
106	15	15	15	15
108	10	10	10	10
112	10	—	10	10
113	12	12	12	12
115	14	14	14	14
118	16	16	16	16
134	12	—	12	12

Figure 1. Missing data patterns [11]

The proportion of missing data is directly related to the quality of statistical inferences. Yet, there is no established cut off from the literature regarding an acceptable percentage of missing data in a data set for valid statistical inferences [9].

The most preferred modern methods for completing missing data are Multiple Imputation and MLE based methods [3, 8, 10].

In ECRI Patient Safety Organizations' report published in 2015, missing data is shown in the top ten of the problems arising from medical technologies in terms of patient safety. Situations that cause missing data in health field can be expressed as following.

- Incompatible patient records
- Network limitations, data entry delays or configuration errors
- Clock synchronization errors between different medical devices and systems
- Default values are inadvertently used or incorrectly given pre-filled fields
- Inconsistencies in patient information when using both paper and electronic records
- Copy old information and paste it in a new report

II. DATASET

In this study, a dataset named "Mesothelioma disease data set" which was prepared at the Faculty of Medicine of Dicle University and loaded on UCI (University of California, Irvine) Machine Learning Repository database was used. This dataset contains 324 patient records. Each record has 34 features.

The original data set does not contain missing data. Three of the 34 features (duration of asbestos exposure, pleural protein, pleural albumin) were selected and new features were generated from these features in accordance with 3 different missing data patterns (MCAR, MAR, MNAR). Selected features have ordinal and continuous data types. Table 1 shows the descriptive statistics of the selected features in the original dataset.

Feature Name	**Mean**	**Standart Error**	Variance
Duration of Asbestos Exposure	30,18	16,41	269,6
Pleural Protein	3,93	1,57	2,48
Pleural Albumin	2,07	0,91	0,83

Table 1. Descriptive statistics of the selected features

III. CLUSTERING AND MAXIMUM LIKELIHOOD ESTIMATION APPROACH

The data is divided into clusters and the estimation is carried out within each clusters themselves before the estimation of missing data points is done with the MLE method.

The k-means clustering method is preferred for clustering. The K-means clustering method divides a set of N data objects into K sets of input parameters. K-means is one of the most commonly used clustering algorithms [2, 6]. The k-means algorithm is often used in clustering applications but its usage requires a complete data matrix. Missing data, however, are common in many applications [4].

Many modern statistical methods that are widely used today work on MLE. MLE also plays a central role in missing data analysis and is one of the two approaches (Multiple Imputation - Maximum Likelihood) that researchers have seen as the latest technology [3, 7].

IV. METHOD

A. Data Reduction

The reason for using a dataset that does not contain missing data in the original case is the desire to have a complete dataset that we can use as a reference to test the success of the proposed method.

Three new features have been derived from three selected features (duration of asbestos exposure, pleural protein, pleural albumin) in accordance with three different missing data patterns (MCAR, MAR, MNAR). The missing data rate in the derived properties is 16% (52/324). The missing data rate is fixed for all missing data types and for all variables.

MCAR			
Feature Name	**Mean**	**Standart Error**	**Variance**
Duration of Asbestos Exposure	30,68	16,44	270,5
Pleural Protein	3,93	1,58	2,51
Pleural Albumin	2,08	0,89	0,80
MAR			
Duration of Asbestos Exposure	30,62	16,30	265,7
Pleural Protein	3,90	1,58	2,50
Pleural Albumin	2,07	0,90	0,81
MNAR			
Duration of Asbestos Exposure	30,03	16,46	270,9
Pleural Protein	3,95	1,56	2,44
Pleural Albumin	2,03	0,93	0,86

Table 2. Descriptive statistics of the selected features in the new datasets

B. Clustering

The new data sets that are created with missing data are divided into 10 clusters by the k-means clustering method in the MATLAB environment.

Since the K-means clustering method cannot work with missing datasets, features with missing values are removed from the dataset by preserving dataset indices before clustering. After the clustering process, the removed features are added to the rows of the clusters and 10 clusters with missing data are obtained. Figure 2 shows the block diagram of the proposed clustering approach.

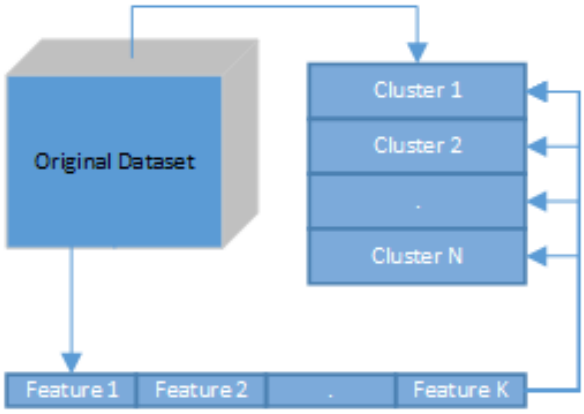

Figure 2. Block diagram of the proposed clustering approach

C. Data Completion

The 10 sets of incomplete data that are generated are completed separately in the MATLAB environment with the MLE method. Completed clusters are merged to obtain the completed dataset. Figure 3 shows the block diagram of the proposed data completion method.

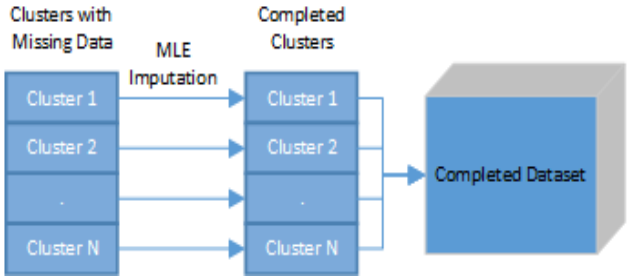

Figure 3. Block diagram of the proposed data completion method.

V. CONCLUSIONS AND DISCUSSION

After the data completion, the three feature vectors are compared with the feature vectors in the original data set by the root mean square error (RMSE) criterion. Table 3 shows the RMSE values and success rates of the proposed method for given dataset.

	Duration of Asbestos Exposure	**Pleural Protein**	**Pleural Albumin**
MCAR	2,7016 **(%96,1)**	0,2155 **(%96,7)**	0,1512 **(%96,5)**
MAR	2,8508 **(%95,9)**	0,2182 **(%96,7)**	0,1453 **(%96,6)**
MNAR	2,4704 **(%96,4)**	0,2677 **(%96,0)**	0,1006 **(%97,7)**

Table 3. RMSE values and success rates of the proposed method

Descriptive statistics of the features in the completed dataset are given in Table 4.

MCAR			
Feature Name	**Mean**	**Standart Error**	**Variance**
Duration of Asbestos Exposure	30,75	15,09	227,9
Pleural Protein	3,95	1,48	2,21
Pleural Albumin	2,08	0,83	0,70
MAR			
Duration of Asbestos Exposure	30,77	14,97	224,3
Pleural Protein	3,91	1,48	2,19
Pleural Albumin	2,08	0,84	0,71
MNAR			
Duration of Asbestos Exposure	30,03	15,23	232,0
Pleural Protein	3,97	1,46	2,14
Pleural Albumin	2,06	0,87	0,76

Table 4. Descriptive statistics of the features in the completed dataset

The comparison results show that the feature vectors obtained as a result of data reduction, clustering and data completion operations are approximately 96.5% similar to the feature vectors in the original data set.

VI. FUTURE WORK

In order to test the acceptability of the applied clustering-maximum likelihood approach, it would be beneficial to repeat this study;

- With different number of clusters
- With different sized data sets
- With different types of feature vectors

At the same time, the applied approach needs to be compared with other popular missing data completion approaches.

REFERENCES

[1] Soley-Bori, M., `` Dealing with missing data: Key assumptions and methods for applied analysis", Boston University *Technical Report No. 4*, 2013.

[2] Hathaway, R. J., *"Fuzzy c-Means Clustering of Incomplete Data"*, IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, Vol. 31, No. 5, 2001.

[3] Allison, P. D., " Handling Missing Data by Maximum Likelihood ", SAS Global Forum, Statistics and Data Analysis, 2012.

[4] Jocelyn T. C., Eric C. C., Richard G. B., "k-POD: A Method for k-Means Clustering of Missing Data", The American Statistician, 70:1, 91-99, 2016.

[5] Kiri L. W., Victoria G. L., "Making the Most of Missing Values: Object Clustering with Partial Data in Astronomy", Astronomical Data Analysis Software and Systems, 2015.

[6] Stefan E. W., "Methods for Clustering Data with Missing Values", University of Leiden, Mathematical Institute, Statistical Science for the Life and Behavioural Sciences, Master Thesis, 2015.

[7] Geaur R., Zahidul I., "Missing value imputation using a fuzzy clustering-based EM approach", Knowl Inf Syst 46:389–422, 2015.

[8] Donald B. R., "Multiple Imputation for Nonresponse in Surveys", Wiley series in probability and mathematical statistics. Applied probability and statistics, ISSN 0271-6232.

[9] Yiran D., Chao-Ying J. P., "Principled missing data methods for researchers", SpringerPlus, 2:222, 2013.

[10] Katsuhiro H., Nobukazu S., Hidetomo I., "Simultaneous Approach to Principal Component Analysis and Fuzzy Clusteringwith Missing Values", IFSA World Congress and 20th NAFIPS International Conference, 2001.

[11] Enders, C., K., "Applied Missing Data Analysis", The Guilford Press, New York, 2010