Estimation of Missing Values in Clinical Laboratory Measurements of ICU Patients Using a Weighted K-Nearest Neighbors Algorithm

OT Abdala, M Saeed

Massachusetts Institute of Technology, Cambridge, MA, USA

Abstract

In the modern intensive care unit (ICU), the physiologic state of critically-ill patients is monitored through a diverse array of biosensors and laboratory measurements. The sheer volume of data that is collected has overwhelmed clinicians charged with assimilating and transforming the data into clinical hypotheses. The development of automated algorithms with vigilant monitoring and clinical decision-support capabilities would help to alleviate this "information-overload" challenge. The inherent noise and measurement error is an added level of complication to the real-time analysis and interpretation of medical data. One class of "noise" in medical data can be characterized by the absence or unavailability of a desired measurement. We have analyzed a large collection of clinical laboratory data (blood chemistry, blood gasses, complete blood counts) from over 600 ICU/CCU patients in the MIMIC II database. An analysis of the frequency of missing data values across patient records for each measurement was completed. Furthermore, we have developed a novel method to estimate the values of missing data by the use of a weighted K-nearest neighbors algorithm. We propose a weighting scheme that exploits the correlation between a "missing" dimension and available data values from other fields. We compare our technique with several popular missing value estimation techniques: principal components analysis, least squares estimation, mean imputation, and classical k-nearest neighbors. The mean standardized imputation error ranges from a minimum of 0.31 to a maximum of 0.75 depending on the imputed dimension. The mean standardized imputation error over all dimensions is 0.45.

1. Introduction

The modern intensive care unit (ICU) has an impressive array of biomedical sensors and monitoring systems to help clinicians assess the physiologic state of acutely-ill patients. These measurements vary in terms of their invasiveness, frequency by which they are measured (seconds to days), costs, measurement protocol simplicity, and measurement reliability. The results of these measurements aid in the diagnosis of disease and direction of therapy.

Often, a measurement is most informative when interpreted with other complimentary diagnostic data. For example, scoring systems that utilize an array of diagnostic measurements have become a standard with which to risk-stratify critical care patients.

One of the most widely accepted scoring system is the Acute Physiology, Age and Chronic Health Evaluation (APACHE) system. APACHE and others [2] derive an acuity score based upon a set of routinely measured variables obtained from vital signs, standardized automated blood tests and arterial blood gas tests.

However, scoring systems usually cannot accommodate missing data and ad-hoc techniques are usually resorted to in order to overcome such challenges. This problem is not only encountered in real-time ICU monitoring, but also in retrospective analyses of clinical data. There are many possible explanations for why a data value may be unavailable: the measurements were simply not made, human or machine error in processing a blood sample, and error in transmitting or storing data values into their respective patient records.

When missing data are encountered, clinicians or researchers may often simply not derive any measurement or score for that patient. Alternatively, they may impute the missing value by assuming the value is "normal" and utilize the statistical mean value of that respective component across the total patient population [2]. However, mean imputation methods ignore the statistical correlation that may be present between different components of a physiologic vector. Several methods have been proposed to exploit the statistical relationships between clinical data components.

In this paper, we present a comparative analysis of the performance of several imputation techniques on ICU clinical data. We also present a novel algorithm that is capable of simultaneously estimating several missing components using a weighted K-nearest neighbors algorithm. In the next section, we describe the methodology we utilized for this study. We briefly describe the algorithms that we evaluated. We provide a detailed description of our novel algorithm and its performance on a rich clinical ICU data set. Then, we

0276-6547/04 $20.00 © 2004 IEEE

Computers in Cardiology 2004;31:693-696.

discuss the major results of this study and suggest future extensions of this work.

2. Methods

Database Used

The MIMIC II [3] database was used as a source of data for this study. MIMIC II includes recordings of waveforms, trend plots, ventilator settings, lab values, and text notes. The lab values are of critical significance in assessing the severity of a patient's condition. This, in addition to the fact that these data are frequently unavailable, makes imputing them of high importance in robust and accurate acuity scoring. The following Lab Values are used in the trial:

- pH
- PaCO2
- PaO2
- BUN
- WBC
- RBC
- Calcium
- Sodium
- Glucose
- Creatinine
- Hemoglobin
- Hematocrit
- Chloride
- Magnesium
- Potassium
- Platelets

Creation of the Test and Training Sets

As was mentioned, lab values are often missing. So, in order to create a test and training set, we took vectors only at times where a full complement of lab values was present. In MIMIC II, this left us with 906 patients. We then split the patients up into 80% to be used as a training set, and 20% to be used as a test set. We randomly performed this split 10 times to ensure that the results reflected the database well and did not amplify the results for anomalous situations from a small number of patients.

Algorithms

In the dropout model that we selected, we remove from 1 to 8 dimensions from the test set data. In a Monte Carlo fashion, we randomly select which of the 16 dimensions to remove 100 times for each number of missing dimensions. This ensures that the results are not dependent on random selections of data points and are reflective of the performance on the entire MIMIC II database.

Mean-Imputation

In mean imputation, we simply calculate the dimension mean from the training set and replace all the missing values in the test set with their corresponding training set mean.

SVD

In SVD imputation [5], we calculate 16 eigenvectors from the training set. We then take the present dimensions from the test set and project them onto the eigenvectors in the space spanned by the present dimensions. This gives us the combination of the eigenvectors that produce the present dimensions of the vector to be imputed. We then take this combination and apply it to the eigenvectors in the missing dimensions and use this as the imputed values for the test set.

KNN

In KNN imputation [1], we search the training set for the closest K neighbors in a Euclidean sense and in the present dimensions, to the vector we wish to impute. We then take the mean of these closest K vectors and replace the missing values with these means.

Novel Approach – Weighted KNN

Our novel approach seeks to add to the standard KNN method by proposing that the calculation for how close a vector is to this vector should not be equally based on all of the present dimensions. For example, if pH is highly correlated with a missing dimension, such as PaCO2, we weigh the pH dimension higher in the distance between two vectors, v_1 and v_2. Therefore, in our distance metric, when imputing a missing dimension, m, we weight each dimension i by the respective correlation coefficient, p_{im}. Thus, our distance metric $D(v_1,v_2)$ for N-dimensional vectors is given by the following equation:

$$D(v_1, v_2) = \sum_{i=1}^{N} \alpha_{i,m} \left| d_{1i} - d_{2i} \right|$$

$$\alpha_{i,m} = f(\sigma_i^2, \rho_{i,m})$$

where α_{im} is proportional to p_{im} and inversely proportional to the variance of the dimension i, σ_i^2. A second addition to the standard KNN algorithm is the use of a weighting on how much a vector can contribute to the imputation based on how close it is to the vector we are trying to impute. The weighting that we are using for this is $(1/D(v_1,v_2))^2$. The main advantage to using this weighting is that it reduces the dependence on the selection of K by making very dissimilar vectors contribute much less to the imputation.

3. Results

Evaluation Criteria

In order to evaluate our results, we are using normalized mean absolute error. For the trials with multiple missing dimensions, we average this over all these missing dimensions.

Average Performance of Techniques

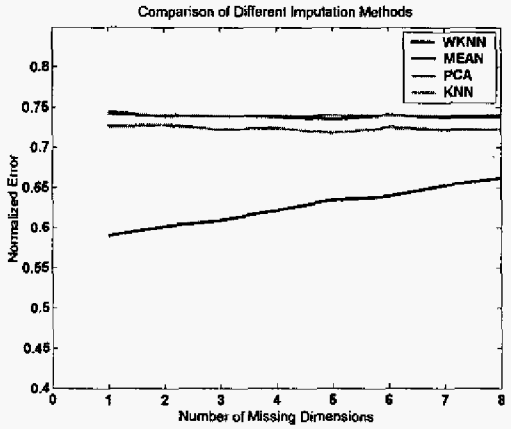

Figure 1: Estimation error as a function of the number of missing dimensions;

Figure 1 demonstrates how each of the imputation methods performs as we increase the number of missing dimensions. We see that the average normalized error using mean imputation is .75. On this data, PCA imputation does only marginally better than imputing mean values. KNN also consistently does slightly better than mean imputation, but the gain is still minimal. Rather than a mean normalized error of .75, the error is around .72. The most significant feature in this comparison of imputation methods is how the performance of all methods other than Weighted-KNN is relatively constant over the number of imputed dimensions. The fact that the performance is less related to the number of missing dimensions indicates that these methods are not using the information contained in the present dimensions to impute the missing dimensions. This is the main motivation behind the weighting we have chosen for our algorithm.

Differences between Imputed Dimensions

Since we are exploiting correlations between different dimensions in our Weighted-KNN distance metric, we expect that we will achieve better performance on the dimensions that exhibit high correlations with each other. This difference will be more peaked in the case where we drop out one dimension at a time and see how well the algorithm can estimate the missing dimension from the remaining dimensions. For example, note in figure 2 that red blood cells (RBC), hemoglobin, and hematocrit (dimensions 6,11,and 12) jointly have high correlation coefficients (all above .85). This contributes to the algorithm's ability to effectively estimate these values. Figure 3 details the one-dimensional performance across all dimensions. For RBC, hemoglobin, and hematocrit, we observe a relatively low mean normalized error of between .3 and .37. As a second example, Sodium and Chloride have a high correlation coefficient of around .7. The algorithm imputes these dimensions with a mean normalized error between .5 and .55. Therefore, as per our expectation, dimensions that have higher correlations are easier to impute, and the Weighted-KNN algorithm exploits this more so than alternate algorithms.

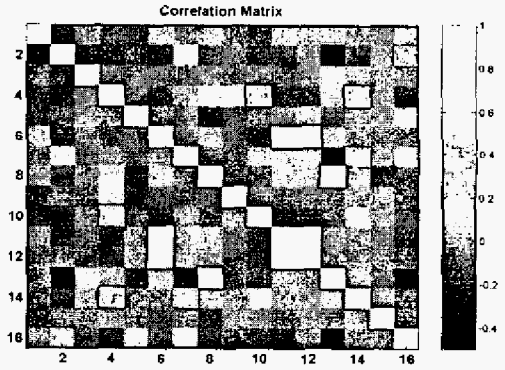

Figure 2: Correlation coefficient matrix to illustrate degree of correlation between different dimensions.

4. Discussion and conclusions

The W-KNN algorithm was demonstrated to have superior results over several established imputation techniques. Further research is possible for improving the W-KNN algorithm. One might be able to factor in therapeutic knowledge into a distance metric. For example, if a particular patient is receiving a potassium-wasting diuretic, the relationship between potassium and other lab values may differ in comparison to the overall patient population. Thus, a framework that factored such information into imputation decision rules may have improved performance over techniques that ignore such knowledge.

Also, one may model the temporal information between lab values from the same patient to predict future lab values as well as missing values using the W-KNN algorithm. As more clinical data in the MIMIC-II database is readily available, such techniques will be possible to develop and evaluate.

Address for correspondence

Mohammed Saeed
45 Carleton St. E25-505
Cambridge, MA 02142
msaeed@mit.edu

References

[1] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman, R. Missing value estimation methods for DNA microarrays. Bioinformatics 2001,17:6 520-525.

[2] Perez, Adriana, et al. Use of the mean, hot deck and multiple imputation techniques to predict outcome in intensive care unit patients in Columbia. Statistics in Medicine 2002; 21:3885-3896.

[3] Saeed, Mohammed, Lieu C, Raber G, and Mark, RG. MIMIC-II: A Massive Temporal ICU Patient Database to Support Research in Intelligent Patient Monitoring. Comput Cardiol. 2002;29:641-4

[4] Hastie T, Tibshirani R, Sherlock G, Eisen M, Brown P, Botstien D. Imputing Missing Data for Gene Expression Arrays. Technical Report, Diision of Biostatistics, Stanford University 1999.

[5] Fodor IK, A survey of dimension reduction techniques. Lawrence Livermore National Laboratory technical report, June 2002.

[6] Ennett C, Frize M. Validation of a Hybrid Approach for Imputing Missing Data. Proc IEEE EMBS/BMES 2003

[7] Bernaards C, Farmer M, Qi K, Dulai G, Ganz P, Kahn K. Comparison of Two Multiple Imputation Procedures in a Cancer Screening Survey. Journal of Data Science 2003.

Figure 3: