33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Missing Data Imputation for Remote CHF Patient Monitoring Systems

Myung-kyung Suh, Jonathan Woodbridge, Mars Lan, Alex Bui, Lorraine S. Evangelista, Majid Sarrafzadeh

Abstract— **Congestive heart failure (CHF) is a leading cause of death in the United States. WANDA is a wireless health project that leverages sensor technology and wireless communication to monitor the health status of patients with CHF. The first pilot study of WANDA showed the system's effectiveness for patients with CHF. However, WANDA experienced a considerable amount of missing data due to system misuse, nonuse, and failure. Missing data is highly undesirable as automated alarms may fail to notify healthcare professionals of potentially dangerous patient conditions. In this study, we exploit machine learning techniques including projection adjustment by contribution estimation regression (PACE), Bayesian methods, and voting feature interval (VFI) algorithms to predict both non-binomial and binomial data. The experimental results show that the aforementioned algorithms are superior to other methods with high accuracy and recall. This approach also shows an improved ability to predict missing data when training on entire populations, as opposed to training unique classifiers for each individual.**

I. Introduction

CONGESTIVE heart failure is a leading cause of death in the United States with approximately 670,000 individuals diagnosed every year [1]. The sequelae of CHF are well known, with frequent decompensation of the chronic state resulting in recurrent hospitalizations. Experts believe that constant monitoring of patients with CHF is essential to a patient's health.

Remote patient monitoring is a promising solution for an expanding population of CHF patients who are unable to access clinics due to the lack of resources, location, or infirmity. Medical care facilitated by remote technology has the potential to enable early detection of key clinical symptoms indicative of CHF-related decompensation. Such remote technologies can also enable health professionals to offer surveillance, advice, and continuity of care to trigger early implementation of strategies that enhance adherence behaviors. WANDA [2][3] is a wireless health project that leverages sensor technologies and communication to monitor the health status of patients with CHF. WANDA monitors vital signs and other information deemed critical to CHF assessment, including weight, blood pressure, heart rate, activity, and daily somatic awareness scale questionnaires [4][5][6]. The effectiveness of WANDA for CHF patients was shown in [3].

However, the first randomized trial of WANDA experienced a considerable amount of missing data: only 33% of the somatic questionnaires were completed; and 55.7% of data had missing values for weight, systolic and diastolic blood pressure, and heart rate data. Moreover, 22.2% of patients experienced system misuse and requested help to accustom themselves the WANDA's technologies. Missing data was further caused by system nonuse and service disorder (such as a network failure, resulting in as much as 6.3% of all of the missing data). Analysis found that system non-use was often due to patients' lack of time or interest to participate in the study. Notably, other studies have experienced similar data loss [5][7]. It is critical for a remote monitoring system such as WANDA to collect and store all monitored vital signs. Any unhealthy changes in a patient's vital signs must be addressed promptly in order to prevent further degradation of a patient's health.

Missing data is especially common in randomized controlled trials. Wood's study [8] showed that 89% of 71 trials published in 2001 in well-known journals (British Medical Journal, BMJ; Journal of the American Medical Association, JAMA; Lancet; and New England Journal of Medicine, NEJM) reported having partly missing outcome values. Many studies applied last observation carried forward, worst case imputation, and complete case analysis techniques that can lead to biased results. To date, there has been no study on missing data imputation in CHF randomized trials.

The objective of this study is to enhance the accuracy of CHF missing data imputation using different data mining techniques. Data imputation allows a patient monitoring system to detect an unhealthy change in a vital sign even when that data is not collected by the system. In this work, we exploit the projection adjustment by contribution estimation (PACE) regression method for predicting and imputing non-binomial data such questionnaire responses. Bayesian methods and voting feature interval (VFI) are used to impute binomial data. The results of these methods are compared using accuracy and correlation efficient values for non-binomial cases, and recall values for binomial cases. The previous methods are compared with several other popular data mining methods. The experimental results show that the aforementioned methods are superior to other methods for CHF patient data imputation.

This study is supported by NIH/National Library of Medicine Medical Informatics Training Program Grant T15 LM07356 and the University of California, Los Angeles, Resource Centers for Minority Aging Research/ Center for Health Improvement of Minority Elderly (RCMAR/CHIME) under NIH/NIA Grant P30-AG02-1684.

M. Suh (dmksuh@ucla.edu), J. Woodbridge (jwoodbridge@ucla.edu), M. Lan (marslan@cs.ucla.edu) and M. Sarrafzadeh (majid@cs.ucla.edu) are with the Computer Science Department, University of California, Los Angeles, CA, 90095, USA.

A. Bui (buia@mii.ucla.edu) is with the Medical Imaging Informatics Group, Department of Radiological Sciences, University of California, Los Angeles, CA, 90024, USA.

L. Evangelista (l.evangelista@uci.edu) is with the Program of Nursing Science, University of California, Irvine, CA, 92697-3959, USA.

TABLE 1. CORRELATION COEFFICIENT VALUES OF EACH TECHNIQUE FOR NON-BINOMIAL DATA

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}
Linear Regression	0.61	0.79	0.41	0.28	0.3	0.78	0.44	0.88	0.29	0.9	0.42	0.84
Simple Linear Regression	0.5	0.75	0.34	0	0.2	0.85	0.32	0.9	0.24	0.92	0.4	0.85
Pace Regression	**0.62**	**0.79**	**0.42**	**0.3**	**0.32**	**0.82**	**0.52**	**0.88**	**0.29**	**0.9**	**0.42**	**0.85**
Isotonic Regression	0.53	0.76	0.23	0.12	0.18	0.86	0.38	0.89	0.1	0.92	0.29	0.85

II. METHODOLOGY

A. Subjects and Datasets

WANDA was approved by the UCLA institutional review board (IRB). Since November 2009, the WANDA system has been used for health data collection on 26 different subjects. The population of the participants is approximately 68% male; 40% White, 13% Black, 32% Latino, and 15% Asian/Pacific Islander, with a mean age of approximately 68.7 ± 12.1. Study participants were all provided with Bluetooth weight scales, blood pressure monitors, land line gateways, and personal activity monitor devices [2]. Each captured data instance for the study comprises 37 different attributes, including: time-stamps; weight; diastolic/systolic blood pressure; heart rate; metabolic equivalents (METs); calorie expenditure; and numeric responses to twelve somatic awareness questions. Each data instance is gathered from each subject once a day. The total number of instances used in this study is 1090.

This study uses the missing at random (MAR) hypothesis [9]. MAR assumes that missing data is dependent on observed data. Hence, missing data can be predicted by resident data.

B. Non-binomial Case Imputation

WANDA employs the Heart Failure Somatic Awareness Scale (HFSAS, [10]) which is a 12-item Likert-type scale to measure awareness of signs and symptoms specific to CHF. A 4-point Likert-type scale is used to ascertain how much a patient is bothered by a symptom (0: not at all, 1: a little, 2: a great deal, 3: extremely).

In order to predict missing answers, we exploit the projection adjustment by contribution estimation regression algorithm (PACE) [11] (rounding any non-integer value returned by PACE). This method is based on maximum likelihood estimation (MLE) and an empirical Bayes framework to minimize the Kullback-Leibler (KL) distance between the original and the estimation function. First, the PACE algorithm transforms parameters using MLE's asymptotic normality property [12] to convert the original parameters. The algorithm utilizes the empirical Bayes estimator in (1):

$$\widetilde{\theta_i^{EB}} = \frac{\int \theta f(x_i|\theta) dG_k(\theta)}{\int f(x_i|\theta) dG_k(\theta)} \quad (1)$$

where $\tilde{\theta}(x)$ is the estimator, $f(x_i|\theta_i)$ is a probability density function (PDF) and G_k is a consistent estimator of G which is the mixing distribution of the mixture $f_G(x) = \int f(x|\theta) dG$. Using (2), the developed algorithm minimizes the KL distance between f and $\tilde{f}$ in (2):

$$\Delta_{KL}(f, \tilde{f}) = E_f \log\left(\frac{f}{\tilde{f}}\right) = \int log\left(\frac{f}{\tilde{f}}\right) f \, (2)$$

This method especially shows better results in high dimensional data spaces and is applied to complete cases that have all 12 answered questions to evaluate the accuracy.

C. Binomial Case Imputation

A binomial approach is used to predict alarms normally triggered by abnormal data values (e.g., drastic weight changes, unhealthy blood pressure, etc.) given missing data. For example, the system should trigger an alarm if a patient has an extreme change in weight – even when the weight value was not collected by WANDA. We use naïve Bayes, a Bayesian network, and VFI to detect such changes in order to alert caregivers.

Naïve Bayes and Bayesian network classifiers are algorithms that approach the classification problem using the conditional probabilities of the features [13]. A Bayesian network is a directed acyclic graph (DAG) over a set of variables X, where the outgoing edges of a variable x_i specifies all variables that depend on x_i. The probability of an outcome is determined as:

$$P(X) = \prod_{x \in X} p(x|par(x)) \quad (3)$$

where $X = \{x_1, x_2, \ldots, x_k\}$ is a set of variables, and par(x) is the set of parents of x in a Bayesian network. The probability of the instance belonging to a single class is calculated by using the prior probabilities of classes and the feature values for an instance. Naive Bayesian method assumes that features are independent and there are no hidden or latent attributes in the prediction process. As such, the experimental results for naïve Bayes and Bayesian network can be slightly different as $p(class) = \frac{1+N(class)}{N(class)+N(instances)}$ for naïve Bayes and p(class) =

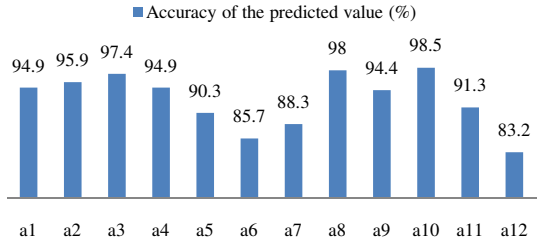

Figure 1. Accuracies of the predicted data for non-binomial cases

TABLE 2. RECALL VALUES OF WEIGHT, SYSTOLIC, DIASTOLIC BLOOD PRESSURE AND HEART RATE VALUES

	Weight	Systolic	Diastolic	Heart Rate
C 4.5	0	0	0.17	0.16
Random Tree	0.05	0	0.11	0.19
Naïve Bayes	**0.7**	**0.71**	**0.89**	**0.91**
Bayesian Network	**0.65**	**0.71**	**0.78**	**0.78**
VFI	**0.65**	**0.71**	**0.67**	**0.78**
Nnge	0.05	0	0.11	0.28
PART	0.08	0	0.28	0.25
DTNB	0.03	0	0.44	0.13
Decision Table	0	0	0.22	0.06
Rotation Forest	0.05	0	0.17	0.03

$\frac{\frac{1}{2}+N(class)}{N(class)\times\frac{1}{2}+N(instances)}$ for Bayesian network where N(x) is the number of sets or instances.

VFI is a categorical classification algorithm and considers each feature independently as Bayes methods [14]. The classification of a new instance is based on a vote among the classifications built by the value of each feature. While training, the VFI algorithm constructs intervals for each feature. For the classification, a single value and the votes of each class in that interval are calculated for each interval. For each class c, feature f gives a vote value:

$$feature\_vote[f,c] = \frac{interval\_class\_count[f,i,c]}{class\_count[c]} \quad (4)$$

where interval_class_count[f, i, c] is the number of instances of class c which is a member of interval i of feature f. The class with the highest total vote is predicted to be the class of the test instance.

In the Bayes methods, each feature participates in the classification by assigning probability for each class and the final probability of a class is the product of each probability measured on each feature. In VFI, each feature distributes its vote among classes and the final vote of a class is the sum of each vote given the features.

All 1090 instances of data are complete (i.e., contain all 37 data values). Instances were divided into to two groups: training and testing. Values from the testing set predicted by the data imputation techniques were compared to their actual values to evaluate the effectiveness of each system.

III. RESULTS

For non-binomial data, PACE [11], linear [16], simple linear [17] and isotonic regression [18] methods were applied. Table 1**Error! Reference source not found.** shows the correlation coefficient values of each method to predict each answer a_i. Correlation coefficient is a measure of the least square fitting values between the predicted and original data. For a given N data points (X,Y), the correlation coefficient $\rho_{X,Y}$ is given as equation (5) where COV(X,Y) is a covariance between X and Y and σ_X , σ_Y are standard deviation values of X and Y. The experimental results show that PACE regression method works better on average than other given regression methods.

$$\rho_{X,Y} = \frac{COV(X,Y)}{\sigma_X \times \sigma_Y} \quad (5)$$

After calculating the coefficient and constant variables, the developed algorithm determines missing values using PACE regression (rounding any non-integer value returned by PACE). The accuracies of the obtained values range between 85.7% and 98.5% (Figure 1).

The binomial case predicts a potential abnormal vital sign when no data value exists within WANDA's database using other existing attributes. C4.5 [19], random tree [20], naïve Bayes [21], Bayesian network [22], VFI [23], nearest neighbor [24], PART [25], DTNB [26], decision table [27], and rotation table [28] algorithms were applied and their recall values were compared. For each method, ten-fold cross validation was applied. In ten-fold validation, the original sample is randomly partitioned into ten subsets and a single subset is held as a testing model, with the remaining nine subsets are used as training data. This cross-validation process is then repeated ten times, using a new subset as a testing model for each repetition. Recall values are given as:

$$recall = \frac{Tp}{Tp+Fn} \quad (6)$$

where T_p is true positive and F_n is false negative. The experimental result (Table 2) shows that naive Bayes, Bayesian network, and VFI have recall values of up to 0.7 for weight, 0.714 for systolic blood pressure, 0.889 for diastolic blood pressure and 0.906 for heart rate values.

Classifiers were trained in two ways. First, unique classifiers were created for each individual where only data collected from an individual was used to predict values from the same individual. Second, a grouped classifier was created using data from the entire population. Both the individual and grouped classifiers were compared using ten-fold validation to test data from 16 patients. The recall values of weight, blood pressure, and heart rate are improved when training on the entire group's data as compared with training each individual's data separately (TABLE 3). For questionnaire data, the accuracies of results were also better when training on all patients' data. When training individually, 75% of patients' data showed 0% accuracy. This is because the entire group has bigger number of data and many individual share similarities in monitored attributes, such as age, symptoms of CHF, etc.

TABLE 3. RECALL VALUES OF DATA FOR INDIVIDUAL AND GROUP

		P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	Group
weight	Naïve Bayes	0	0	0	0	0	.33	0	0	0	0	0	1	0	.65	0	0	**.7**
	Bayes Net	0	0	0	0	0	.33	0	0	0	0	0	0	0	.88	0	0	**.65**
	VFI	0	0	1	0	0	.33	0	.33	0	0	0	1	0	.82	0	0	**.65**
systolic	Naïve Bayes	0	.33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**.71**
	Bayes Net	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**.71**
	VFI	0	.33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**.71**
diastolic	Naïve Bayes	0	.85	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**.89**
	Bayes Net	0	.39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**.78**
	VFI	0	.62	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**.67**
heart rate	Naïve Bayes	0	.54	0	0	0	0	0	0	0	0	0	0	0	0	0	.92	**.91**
	Bayes Net	0	.31	0	0	0	0	0	0	0	0	0	0	0	0	0	.75	**.78**
	VFI	0	.69	0	0	0	0	0	0	0	0	0	0	0	0	0	.67	**.78**

IV. CONCLUSION

Heart failure is a leading cause of death in the United States with approximately 4,600,000 Americans suffering from heart failure. The wireless health monitoring system WANDA exploits sensors and wireless communication techniques to monitor and provide guidance and feedback to patients with CHF. WANDA was shown to be highly effective for patients with CHF [3]. However, the first pilot study of WANDA experienced a considerable amount of missing data.

This study enhanced the accuracy of the CHF missing data using the PACE regression method for predicting and imputing non-binomial data; and Bayesian methods and voting feature interval for binomial data. The experimental results show that PACE regression works better than linear regression, simple linear regression, and isotonic regression methods with accuracy values of more than 85.7%. The experiment comparing Bayes and VFI methods with other algorithms proves that Bayes and VFI algorithms work better (Table 2) with recall values of up to 0.7 for weight, 0.714 for systolic blood pressure, 0.889 for diastolic blood pressure and 0.906 for heart rate values. This study also showed that increased accuracy is obtained by training on a large population as opposed to training the classifiers for each individual independently.

REFERENCES

[1] Lloyd-Jones, D. (2009). Heart disease and stroke statistics--2009 update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Circulation, 119(3), e21-.

[2] Suh, M. (2010). WANDA B.: Weight and activity with blood pressure monitoring system for heart failure patients. In 2010 IEEE International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM). (pp. 1-).

[3] Suh, M. (2010). An automated vital sign monitoring system for congestive heart failure patients. In Proceedings of the ACM international conference on Health informatics - IHI '10. (pp. 108-).

[4] Chaudhry, S I. (2007). Randomized trial of telemonitoring to improve heart failure outcomes (Tele-HF): Study design. Journal of cardiac failure, 13(9), 709-.

[5] Chaudhry, S I. (2010). Telemonitoring in patients with heart failure. The New England journal of medicine, 363(24), 2301-.

[6] Stone, R A. (2010). Active Care Management Supported by Home Telemonitoring in Veterans With Type 2 Diabetes. Diabetes care, 33(3), 478-.

[7] Desai, A S. (2010). Connecting the circle from home to heart-failure disease management. The New England journal of medicine, 363(24), 2364-.

[8] Wood, A M. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. Clinical trials, 1(4), 368-.

[9] Schafer, J L. (2002). Missing data: Our view of the state of the art. Psychological methods, 7(2), 147-.

[10] Jurgens, C Y. (2006). Psychometric testing of the heart failure somatic awareness scale. Journal of cardiovascular nursing, 21(2), 95-.

[11] Yang, Y.(2002). Modeling for optimal probability prediction. In Proceedings of ICML.

[12] Fahrmeir, L. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. Annals of statistics, 13(1), 342-.

[13] Duda, R. O. (1973). Pattern classification and scene analysis,New York: Wiley.

[14] G. Demiroz.. (1997). Classification by voting feature intervals. In: 9th European Conference on Machine Learning, 85-92.

[15] Gu¨venir, H. A., & Sirin, I. (1996). Classification by feature partitioning.Machine Learning, 23, 47–67.

[16] Draper, N.R. (1998). Applied Regression Analysis. Wiley Series in Probability and Statistics.

[17] Heiberger, R M. (2009). Simple Linear Regression. In R Through Excel. (pp. 193-).

[18] Best, M J. (1990). Active set algorithms for isotonic regression; a unifying framework. Mathematical Programming, 47(1-3), 425-.

[19] Quinlan, J R. (1993). C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning).

[20] Aldous, D. (1991). The continuum random tree II: an overview. Stochastic Analysis, 167, 23-.

[21] George H. (1995). Estimating Continuous Distributions in Bayesian Classifiers. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. pp. 338-345.

[22] Friedman, N. (1997). Bayesian network classifiers. Machine learning, 29(2/3), 131-.

[23] Demiroz, G. (1997).Classification by voting feature intervals. *ECML-97*.

[24] Martin, B. (1995). Instance-Based learning : Nearest Neighbor With Generalization, Master Thesis, University of Waikato, Hamilton, New Zealand

[25] Frank, E. (1998). Generating Accurate Rule Sets Without Global Optimization. Machine Learning: Proceedings of the Fifteenth International Conference.

[26] Hall,M . (2008). Combining Naive Bayes and Decision Tables. In Proceedings of the 21st Florida Artificial Intelligence Society Conference (FLAIRS).

[27] Kohavi, R. (1995). The Power of Decision Tables. In Proc European Conference on Machine Learning.

[28] Rodriguez, J. (2006). Rotation Forest: A new classifier ensemble method. IEEE Transactions on Pattern Analysis and Machine Intelligence. 28(10):1619-1630.