ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

Survey on Medical Cases Creation using Medical Events

Radhika Modani[1], Dr Parag Kulkarni[2] , Prof. Mukta Takalikar[3]

M.E. Student, Computer Engineering, Pune Institute of Computer Technology, Pune, India[1]

CEO, Chief Scientist, iknowlation Research Labs, Pune, India[2]

Asst. Professor, Computer Engineering, Pune Institute of Computer Technology, Pune, India[3]

ABSTRACT: The patient records exist in many forms and in a different location. The record fails to gain an overall picture of the patient's health condition. The medical cases creation system delivers accessible and accurate information of the patient to health-care providers. The fundamental function of medical cases is to record, monitor, retrieve, analyze and predict all event and cases which are an encounter between patient and the health-care system. The cases are created by imputing missing value in raw medical data-set. Researchers are working on raw medical data issue to present more modern techniques. Despite the fact that numbers of strategies are available, researchers are solving troubles in seeking an appropriate technique. This survey paper present a review of the imputing missing data strategy (IMDS) and talks about the strategies that are analyzed in the works.

KEYWORDS: IMDS, Imputing missing value; Medical cases creation system

I. INTRODUCTION

With the advent of electronic health records, more data is continuously collected for individual patients, and more data is available for review from past patients. Despite this, it has not yet been possible to successfully use this data to systematically build computer-based decision support systems that can produce clinical recommendations to assist clinicians in providing individualized health-care. Medical Decision making should use relevant data from many distributed systems instead of a single data source to maximize its applicability but real-world medical data are often based on missing information. This is referred as the medical information challenge.

In past, the specialist applies their insight in the therapeutic choice and finding framework. After applying their insight they make a watchful treatment on the premise of patients clinical exam result in a blend of their history. There is the need to give precise determination and treatment to offer assistance in patient recuperation. Various variables which can impact customary restorative determination process are introduced. The data mining is broadly utilized as a part of PC based therapeutic analysis, which utilizes the medicinal cases to get the conclusion run the show.

Now a days, large volume of data available in the medical system which gives the opportunity to construct computer based patient medical cases. Two issues are important in the construction of medical diagnosis decision system: the problem of constructing medical cases directly from raw data by imputing missing value and creating medical system with respect to user.

Data mining is the process which provides a concept to attract attention of users due to high availability of huge amount of data and need to convert such data into useful information. Data preparation is a principal phase of data investigation [16][17]. Three types of mean imputation methods presented on missing data [13]. Rubin investigated about inference and missing data and multiple imputations for non-reaction in the overview [14]. Allison explored estimates of linear models with incomplete data and on missing data [15]. Myrtveit et al. connected missing data strategies to a software project data set, and assessed four missing data procedure are list wise deletion (LD), mean imputation (MI), similar response pattern imputation (SRPI) and full information maximum likelihood (FIML)[18]. Junninen et al. evaluated and compared univariate and multivariate methods for missing data imputation in air quality data sets [19].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

The scope of medical cases creation is firstly it is suitable for a much medical application because it helps to impute missing data in partial information. Secondly, it is suitable for further validation which is useful for many users including doctors, nurses, other medical staff, and patient. It is also suitable for predicting the missing values from raw data, which helps in medical diagnosis. It provides a solution to deal with the complexity of relationship and inter-dependencies in medical records.

Types of Incomplete Data are little and David C. Howell characterizes a list of missing mechanisms, which are generally acknowledged by the community. There are three mechanisms under which missing data can happen [21]:

1) Missing completely at random (MCAR): the fact that any observation is missing is completely unrelated to the values of the data for the other variables or to the non-missing data elements in the variable for the missing data. This circumstance is uncommon in real world and is generally talked about in statistical theory.

2) Missing at random (MAR): Data is considered to be missing at random if the data meet the requirement that missingness does not depend on the value after controlling for all the other variables. The key aspect about MAR is that the values of the missing data can somehow be predicted from some of the other variables being studied.

3) Not missing at random (NMAR). If the probability that an observation is missing depends on information that is not observed, this kind of missing data is called not missing at random. This circumstance is generally confused and there is no universal solution.

II. RELATED WORK

A very few research papers that tangentially address these issues are available and we are going to discuss about the missing value imputation techniques:

In [1] authors had proposed a method based on density clustering and gray relational analysis. In the given missing data set, DBSCAN can accurately cluster samples, although there are noise instances and it can find clusters of different shape in spatial database. Grey relational analysis is introduced to compensate for the regret caused by using the mathematical statistics method. There isn't a high requirement for the number of data and whether have rules. Grey relational grade as similarity metrics between missing values data and complete data can accurately impute the missing attribute values and effectively improve the accuracy of aided medical diagnosis. Using gray relational analysis in the clusters can reduce complexity for missing data imputation relative to the whole data set and better mine the useful information in the existing data. But this technique has problems, such as concept drift. It may cause the lack of auxiliary information and bring the problems for imputation.

In [2] authors had proposed a hybrid missing data completion method named multiple imputation using gray-system-theory and entropy based on clustering (MIGEC). Firstly, the specific record with the least missing parametric values is firstly allocated to the closest group quantified by the GST based distance metric. Next, each missing attributive value of the record is estimated by the proposed multiple imputation. Then the imputed item is included into the complete set along with excluding the original copy from the incomplete set. And the next element in the rearranged incomplete data set repeats the similar solution until no more elements exist in that set.

In [5] a novel framework for a medical decision supports system that integrates knowledge-based and learning-based systems. The system leverages the benefits of machine learning, structured knowledge representation and logic-based inference to facilitate robust, intelligent decision support. It provides a solution to deal with the complexity of relationships and inter-dependencies in medical decisions. Imputation models are generated in a preprocessing stage and then integrated with the ontological system, allowing the system to maintain real time performance capability. It provides explainable responses to queries and more critically is also robust to missing data. This results in a patient-centric evidence-based decision support system. It depends on three real world information sources: a large dataset of patient histories, a drug interaction registry, and a collection of medication prescription protocols.

In [6] a novel method to impute missing data, named feature weighted gray KNN (FWGKNN) imputation algorithm had proposed. In this algorithm, mutual information (MI) is applied in the relevance measure between random features. Compared to other imputation methods based on gray theory, proposed gray distance metric takes into consideration the relevance between features based on the concept of MI. Moreover, this method imputes missing data iteratively in each class. To improve imputation performance, the FWGKNN method imputes instances with missing data according to the amount of missing data in ascending order.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

In [7] authors had proposed different substitution methods for the replacement of missing data values for the use of these cases in a neural network based decision support system for acute appendicitis. In the substituting means-method, the missing data values are replaced with the mean values computed from the complete data cases. In the random values method, the replacement is carried out with generated random values, where the mean and standard deviation,

the values in the complete data cases are taken into account in generating the numbers. In the nearest neighbor method, the most similar patient cases are searched from the complete data cases by computing Euclidean distance measures between all cases and selecting the closest one. The missing values are then replaced by the values of this closest case. In the neural network-method, all possible combinations of missing data value patterns are searched. For each of these patterns a separate Multilayer Perceptron (MLP) neural network is trained, with the back-propagation algorithm and with the complete cases of the database, to estimate the values of the variables in the pattern.

In [10] filter imputation method, for ordinal, correlated data and nearest neighbor based retrieval mechanism are introduce for dealing with missing values in a case-based reasoning system. The filter imputation method is similar to the weighted k-nearest neighbour (wKNN) imputation method but has several advantages. One of the shortcomings of the wKNN method is that finding appropriate weights can be a difficult and time-consuming task, especially if the correlation between attributes is not clearly defined. In the filter method, the exact value of the correlation coefficients between attributes is not required; knowing which attributes are correlated to each other suffices. In addition, in the wKNN method the number of similar cases retrieved has to be carefully chosen. If it is too large it might include cases that are quite dissimilar and therefore irrelevant for the imputation, while if it is too small the imputed value will be vulnerable to outliers or extreme values and biasing. In the filter method we utilize all cases that have the same value as the attribute used for filtering. The concept of the filter method is simple, allowing quick and easy implementation.

In [11] had proposed personalized clinical decision support system using discovery engine (DE) which Discover different relevant features and use them to recommend personalized clinical decisions. The DE discovers/learns which features/characteristics of a patient are most informative in predicting the success of a clinical decision. For instance, the tumor grade may be found to be relevant for predicting the success of a certain type of chemotherapy in a patient, but not the success of another type of chemotherapy. Thus, different features may be discovered to be relevant for different decisions. Then, when a clinician requests the recommendation from DE for a specific patient, DE decides the best clinical recommendation for the patient which has the best-estimated outcomes. The outcome of a decision is estimated based on the values of the relevant features (i.e. the features found to be relevant for that decision) of the patient. For instance, if the tumor grade was found to be relevant for a certain chemotherapy that chemotherapy will or will not be recommended to that patient depending on that patient's tumor grade.

In [14] Mean imputation method is explained which is frequently used methods. It comprises of replacing the missing data for a given component or attribute by the mean of all known values of that attribute in the class where the instance with missing attribute belongs.

In [22] explain hot deck imputation method. Given an incomplete pattern, HD replaces the missing data with values form input data vector that is nearest in terms of the attributes that are known in both patterns. HD attempts to protect the distribution by substituting different observed values for each missing .The similar method of HD is Cold deck imputation method which takes other data source than current dataset.

K-Means is used to classify or to group the objects based on attributes/features into k number of group [23]. The grouping is finished by minimizing the sum of squares of distances between data and the corresponding cluster centroid. It provides quick and precise way of estimating missing values.

In FKMI, membership function plays an important role. Membership function is allocated with every data object that depicts in what degree the data object is belonging to the particular cluster. Data objects would not get allotted to concrete cluster which is indicated by centroid of cluster (as in the case of K means), this is because of the various membership degrees of every data with entire K clusters [1].

REGRESSION IMPUTATION: Using regression method for imputation, the values from the features are observed and then predicted values are used for filling Missing values [24].

MULTIPLE IMPUTATIONS: The imputed values are draws from a distribution, so they inherently contain some variation. Thus, multiple imputations (MI) illuminates the limitations of single imputation by presenting an additional

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

form of error based on variation in the parameter estimates across the imputation, which is called between imputation errors [2].

Methods	Overview	Remarks
Mean Value imputation Method	Filling the missing values with arithmetic mean of the available cases.	Advantages • It is applicable for all type of missingness Disadvantages • Reduces the variability of the data. •Affect the measures of association.
Hotdeck (HD) Imputation	Replaces each missing value with a random draw from a subsample of respondents that scored similarly on a data set of matching variables.	Advantages • It generates a complete data set. Disadvantages • Not well suited for estimating measures of association. • Produce substantially biased estimates of correlation and regression coefficients.
K-Nearest Neighbor Imputation (KNNI)	This method uses k-nearest neighbour algorithms to estimate and replace missing data. It can estimate both qualitative attributes and quantitative attributes.	Advantages • Robust to noisy training data. • Effective if the training data is large Disadvantages • Need to determine value of parameter K • Distance based learning is not clear which type of distance to use.
K-Means clustering method	uses algorithm called nearest neighbour to impute the values in the same way as KNNI	Advantages • If variables are huge, then K-Means most of the times computationally faster than hierarchical clustering, if we keep k smalls Disadvantages • Difficult to predict K-Value. •It does not work well with data of Different size and Different density
Fuzzy K-Means clustering Imputation (FKMI)	Unreferenced attributes for every uncompleted data are substituted by FKMI on the basis of membership degrees and cluster centroid values.	Advantages • Gives best result for overlapped data set and comparatively better then k-means algorithm. • data point is assigned membership to each cluster center as a result of which data point may belong to more than one cluster center. Disadvantages • Apriori specification of the number of clusters. • Euclidean distance measures can unequally weight underlying factors.
Regression Imputation	Replaces missing values with predicted scores from a regression equation by using information from the complete variables.	Advantages • It generates a complete data set. • Variables tend to be correlated Disadvantages • Over estimate correlation • bias
Multiple Imputation (MI)	Creates several copies of the data and imputes each copy with different plausible estimates of missing values.	MI illuminates the limitations of single imputation. It replaces each missing item with two or more acceptable values.

TABLE 1: IMPUTATION METHOD

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

III. RESULTS

Ignoring missing values will create biased estimate in the analysis of data. With observational studies there are many methods that have been identified for dealing with missing observations. Some of the earlier solutions, such as hot deck imputation, mean substitution and pair wise deletion are slowly tending to fall by the wayside because they lead to bias in parameter estimation. The most important techniques is the multiple imputation (MI). It rely on iterative solutions in which the parameter estimates lead to imputed values, which in turn change the parameter estimates, and so on. MI is an interesting approach because it uses randomized techniques to do its imputation, and then relies on multiple imputed datasets for the analysis.

IV. CONCLUSION AND FUTURE WORK

This survey mainly focuses on the study of missing data handling method in data mining, imputation methods are broadly used to fill the missing values of various kinds of datasets. In this survey, the overall views on the handling missing methods are discussed. Thus it clearly seen that numerous strategies are proposed for handling missing values present in the dataset. Further, these imputation methods are compared along with their reviews.

REFERENCES

1. Li Peng, Zhang Ting-ting, LiangTian-ge and Zhang Kai-hui, Missing Value Imputation Method Based on Density Clustering and Grey Relational Analysis, International Journal of Multimedia and Ubiquitous Engineering 2015.
2. Ting Tian, Bing Yu, Dan Yu and Shilong Ma, Missing data analyses: a hybrid multiple imputation algorithm using Gray System Theory and entropy based on clustering, Springer Science+Business Media New York 2013.
3. Ritu Chauhan and Harleen Kaur, A Knowledge Driven Model: Extract Knowledge from High Dimensional Medical Databases, Interna-tional Conference on Machine Intelligence and Research and Advancement,2013.
4. D. L. Hudson, Development of Health Diagnostics Based on Personalized Medical Models, International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015.
5. Taranath N.L., Shanthakumar B Patil, Premajyothi Patil and C.K.Subbaraya,Medical Decision Support System for the Missing Data using Data Mining, IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2015.
6. Ruilin Pan, Tingsheng Yang, Jianhua Cao, Ke Lu and Zhanchao Zhang, Missing data imputation by K nearest neighbours based on grey relational structure and mutual information, Springer Science and Business Media New York, 2015.
7. Erkki Pesonen, Matti Eskelinen, Martti Juhola, Treatment of missing data values in a neural network based decision support system for acute abdominal pain, Elsevier Science, Artificial Intelligence in Medicine 13, 1998.
8. Ponrudee Netisopakul and Waranyu Saapajit, Prediagnosis Doctor Simulation Using Case-Based Techniques, IEEE World Congress on Computer Science and Information Engineering, 2009.
9. Kamran Farooq, Peipei Yang, Amir Hussain, Kaizhu Huang, Calum MacRae, Chris Eckl and Warner Slack, Efficient clinical decision making by learning from missing clinical data, IEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE), 2013.
10. Rupa Jagannathan and Sanja Petrovic, Dealing with Missing Values in a Clinical Case-Based Reasoning System, 2nd IEEE International Conference on Computer Science and Information Technology, 2013.
11. Jinsung Yoon, Camelia Davtyan, MD, FACP, and Mihaela van der Schaar, Discovery and Clinical Decision Support for Personalized Health-care" IEEE Journal of Biomedical and Health Informatics, 2015.
12. Hisao Ishbuch, Aluhu-0 Miyazalu and Hideo Tanaka, Neural-Network Based Diagnosis Systems for Incomplete Data with Missing Inputs, IEEE World Congress on Computational Intelligence, 1994.
13. Noor, M. N., Yahaya, A. S., Ramli, N. A., & Al Bakri, A. M. M., Mean imputation techniques for filling the missing observations in air pollution dataset, Key Engineering Materials 594-599:902-908 Trans Tech Publications 2014.
14. Rubin, D. B., Inference and missing data. Biometrika, 63(3):581-592. 1976.
15. Allison, P. D., Estimation of linear models with incomplete data. Sociological methodology, 71-103 1987.
16. Smyth, P., Data mining at the interface of computer science and statistics. In Data mining for scientific and engineering applications 35-61. Springer US 2001.
17. Zhang, S., Zhang, C., & Yang, Q. , Data preparation for data mining.Applied Artificial Intelligence, 17(5-6):375-381, 2013.
18. Myrtveit, I., Stensrud, E., & Olsson, U. H., Analyzing data sets with missing fata: An empirical evaluation of imputation methods and likelihood-based methods. IEEE Transactions on Software Engineering, 27:999–1013 (2001).
19. Heikki Junninen, Harri Niska, Kari Tuppurainen, Juhani Ruuskanen, Mikko Kolehmainen, Methods for imputation of missing values in air quality data sets, Atmospheric Environment 38:2895–2907, 2014.
20. Swati Jain, Dr Naveen Chodhary, Mrs Kalpana Jain, A survey paper on missing data in data mining, IJIERT, VOLUME 3, ISSUE 12, Dec.-2016
21. David C. Howell, The Treatment of Missing Data, Howell, D.C. (2008) The analysis of missing data. In Outhwaite, W. & Turner, S. Handbook of Social Science Methodology. London: Sage.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

22. Dr. A.Sumathi, Missing Value Imputation Techniques Depth Survey And an Imputation Algorithm To Improve The Efficiency Of Imputation. IEEE- Fourth International Conference on Advanced Computing, IcoAC (2012).
23. Kin Wagstaff, Clustering with Missing Values: No Imputation Required,NSF grant IIS-0325329:1-10.
24. Y. Kou, C.-T. Lu, and D. Chen, Spatial weighted outlier detection. In Proceedings of the Sixth SIAM International Conference on Data Mining, 614–618,Bethesda, Maryland, USA, 2016.

BIOGRAPHY

Radhika Modani received the B.E. degree in Computer Science and Engineering from G. H. Raisoni College of Engineering, Nagpur in 2014. Now doing PG in Computer Engineering, from Pune Institute of Computer Technology, Pune, India and intenship at iKnowlation Research Lab.

Dr Parag Kulkarni is Chief Scientist and CEO of the iKnowlation Research Labs Pvt Ltd**,** an innovation, strategy and business consulting and product development organization. He has been visiting professor/researcher at technical and B-schools of repute including IIM, Masaryk University – Brno, COEP Pune.

Prof. Mukta Takalikar is working as Professor at Pune Institute of Computer Technology, Pune, Maharashtra, India. She has completed her B.E. and M.E. in computer engineering. Her area of interest is NLP, Machine Learning and HPC.