International Journal of Engineering Research and Development
e-ISSN: 2278-067X, **p**-ISSN : 2278-800X, www.ijerd.com
Volume 5, Issue 1 (November 2012), PP.05-07

K-Nearest Neighbor in Missing Data Imputation

Ms.R.Malarvizhi[1], Dr.Antony Selvadoss Thanamani[2],
[1]Research Scholar ,Research Department of Computer Science, NGM College, 90 Palghat Road,Pollachi, ,Bharathiyar University,Coimbatore.
[2]Professor and HOD, Research Department of Computer Science, NGM College 90 Palghat Road, Pollachi Bharathiyar University, Coimbatore.

Abstract:- We propose a comparative study on single imputation techniques such as Mean, Median, and Standard Deviation combined with k-NN algorithm. Training set with their corresponding class groups the data of different sizes. The above techniques are applied in each group and the results are compared. Median/ Standard Deviation shown better result than Mean Substitution.

Keywords:- Mean Substitution, Median Substitution, Standard deviation, kNN algorithm, Training Set

I. INTRODUCTION

Missing data is one of the problems which are to be solved for real-time application. Traditional and Modern Methods are there for solving this problem. The variables may be of Missing Completely at Random, Missing at Random, Missing not at Random. Each variable should be treated separately. k-NN algorithm is used to group the data set into different groups. The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples. In the classification phase, *k* is a user-defined constant, and an unlabeled vector is classified by assigning the label which is most frequent among the *k* training samples nearest to that query point. Usually Euclidean distance is used as the distance metric. After grouping of data, missing data in each group is imputed by Mean/ Median/Standard Deviation. The results are compared in different percentage of accuracy.

II. LITERATURE SURVEY

Rubin and Little have defined three classes of them: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR).In the case of statistical modelling, when external knowledge is available about the dependencies between the missing values as well as about the dependencies between missing and observed data, one might use the Markov or Gibbs processes for imputation. Likelihood based tests have been proposed by Fuchs (1982) for contingency tables, and by Little (1988) for multivariate normal data. A nonparametric test has been proposed by Diggle (1989 for preliminary screening. Rideout and Diggle (1991) have proposed a parametric test which requires the modelling of the missing-data mechanism. Chen and Little (1999) have generalized Little's (1988) basic idea of constructing test statistics. They avoid distributional assumptions, whereas Little (1988) assumed normal data. Some specific tests for linear regression models have been proposed too. Simon and Simonoff (1986) have written an article in which they describe tools for MAR diagnostic and for other purposes. They make no assumptions about the nature of the missing value process. Simonoff has introduced (1998) a test to detect non-MCAR mechanisms. His diagnostics are based on standard outlier and leverage-point regression diagnostics. Recently Toutenburg and Fieger (2001) introduced methods to analyse and detect non-MCAR processes for missing covariates. They use an outlier detection to identify non-MCAR cases.

III. SINGLE IMPUTATION TECHNIQUES

A. Mean Substitution

The most commonly practiced approach is mean substitution— single imputation techniques. Mean substitution replaces missing values on a variable with the mean value of the observed values. The imputed missing values are contingent upon one and only one variable – the between subjects mean for that variable based on the available data. Mean substitution preserves the mean of a variables distribution; however, mean substitution typically distorts other characteristics of a variables distribution.

B. Median Substitution

Mean or median substitution of covariates and outcome variables is still frequently used. This method is slightly improved by first stratifying the data into subgroups and using the subgroup average. Median imputation results in the median of the entire data set being the same as it would be with case deletion, but the variability between individuals' responses is decreased, biasing variances and covariances toward zero.

C. Standard Deviation

The standard deviation measures the spread of the data about the mean value. It is useful in comparing sets of data which may have the same mean but a different range. The Standard Deviation is given by the formula

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

σ = lower case sigma
$\sum$ = capital sigma
$\bar{x}$ = x bar

D. k-Nearest Neighbor Algorithm for Classification

If each sample in our data set has n attributes which we combine to form an n-dimensional vector: x = (x1, x2, . .xn). These n attributes are considered to be the independent variables. Each sample also has another attribute, denoted by y (the dependent variable), whose value depends on the other n attributes x. We assume that y is a categoric variable, and there is a scalar function, f, which assigns a class, y = f(x) to every such vectors. We suppose that a set of T such vectors are given together with their corresponding classes: x(i), y(i) for i = 1, 2, . . . , T. This set is referred to as the training set.

The idea in k-Nearest Neighbor methods is to identify k samples in the training set whose independent variables x are similar to u, and to use these k samples to classify this new sample into a class, v. f is a smooth function, a reasonable idea is to look for samples in our training data that are near it (in terms of the independent variables) and then to compute v from the values of y for these samples. The distance or dissimilarity measure can be computed between samples by measuring distance using Euclidean distance.

The Euclidean distance between the points is

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

The simplest case is k = 1 where we find the sample in the training set that is closest (the nearest neighbor) to u and set v = y where y is the class of the nearest neighboring sample. For k-NN , find the nearest k neighbors of u and then use a majority decision rule to classify the new sample. The advantage is that higher values of k provide smoothing that reduces the risk of over-fitting due to noise in the training data. In typical applications k is in units or tens rather than in hundreds or thousands. Notice that if k = n, the number of samples in the training data set, we are merely predicting the class that has the majority in the training data for all samples irrespective of u.

IV. EXPERIMENTAL ANALYSIS AND RESULT

A dataset consisting of 5000 records with 5 variables has been taken for analysis. The test dataset is prepared with some data's missing. The missing percentage varies as 2, 5, 10, 15 and 20. k-NN algorithm is implemented with different training data and its corresponding class. The group size also differs as 3, 6 and 9. Each group is separately assigned with mean, median and standard deviation. The results are shown in the below table.

Table 1: Classification of Data into Various Groups

Percentage of Missing	**3 GROUPS**			**6 GROUPS**			**9 GROUPS**		
	Mean Sub	**Med Sub**	**Std Dev**	**Mean Sub**	**Med Sub**	**Std Dev**	**Mean Sub**	**Med Sub**	**Std Dev**
2	67	70	70	69	71	71	71	73	72
5	69	72	72	75	76	76	74	77	77
10	66	71	71	72	78	78	71	80	80
15	64	72	72	66	77	77	66	79	79
20	60	72	72	55	78	77	50	80	80

The below table shows the average percentage of the above table. The result shows that median and standard deviation has some improvement over mean substitution. There is also a gradual improvement in the percentage of accuracy in case of different sizes of groups.

Table 2: Average of the above Method

Percentage of Missing	Mean Substitution	Median Substitution	Standard Deviation
2	69	71	71
5	73	75	75
10	70	76	76
15	65	76	76
20	55	77	76

V. CONCLUSION AND FUTURE WORK

k-NN algorithm is one of the famous classifier for grouping up of data .Traditional Methods such us Mean/ Median and Standard Deviation is used to improve the performance of accuracy in missing data imputation. This can be further enhanced by comparing with some other machine learning techniques like SOM, MLP.

REFERENCE

[1]. Allison, P.D-"Missing Data", Thousand Oaks, CA: Sage -2001.

[2]. Bennett, D.A. "How can I deal with missing data in my study? Australian and New Zealand Journal of Public Health", 25, pp.464 – 469, 2001.

[3]. Graham, J.W. "Adding missing-data-relevant variables to FIML- based structural equation models. Structural Equation Modeling", 10, pp.80 – 100, 2003.

[4]. Graham, J.W, "Missing Data Analysis: Making it work in the real world. Annual Review of Psychology", 60, 549 – 576 , 2009.

[5]. Gabriel L.Schlomer, Sheri Bauman, and Noel A. Card : " Best Practices for Missing Data Management in Counseling Psychology" , Journal of Counseling Psychology 2010, Vol.57.No 1,1 – 10.

[6]. Jeffrey C.Wayman , "Multiple Imputation For Missing Data : What Is It And How Can I Use It?" , Paper presented at the 2003 Annual Meeting of the American Educational Research Association, Chicago, IL ,pp . 2 -16, 2003.

[7]. A.Rogier T.Donders, Geert J.M.G Vander Heljden, Theo Stijnen, Kernel G.M Moons, "Review: A gentle introduction to imputation of missing values" , Journal of Clinical Epidemiology 59 , pp.1087 – 1091, 2006.

[8]. Kin Wagstaff ,"Clustering with Missing Values : No Imputation Required" -NSF grant IIS-0325329,pp.1-10.

[9]. S.Hichao Zhang , Jilian Zhang, Xiaofeng Zhu, Yongsong Qin,chengqi Zhang , "Missing Value Imputation Based on Data Clustering", Springer-Verlag Berlin, Heidelberg ,2008.

[10]. Richard J.Hathuway , James C.Bezex, Jacalyn M.Huband , "Scalable Visual Assessment of Cluster Tendency for Large Data Sets", Pattern Recognition ,Volume 39, Issue 7,pp,1315-1324- Feb 2006.

[11]. Qinbao Song, Martin Shepperd ,"A New Imputation Method for Small Software Project Data set", The Journal of Systems and Software 80 ,pp,51–62, 2007.

[12]. Gabriel L.Scholmer, Sheri Bauman and Noel A.card "Best practices for Missing Data Management in Counseling Psychology", Journal of Counseling Psychology, Vol. 57, No. 1,pp. 1–10,2010.

[13]. R.Kavitha Kumar, Dr.R.M Chandrasekar,"Missing Data Imputation in Cardiac Data Set" ,International Journal on Computer Science and Engineering , Vol.02 , No.05,pp-1836 – 1840 , 2010.

[14]. Jinhai Ma, Noori Aichar –Danesh , Lisa Dolovich, Lahana Thabane , "Imputation Strategies for Missing Binary Outcomes in Cluster Randomized Trials"- BMC Med Res Methodol. 2011; pp- 11: 18. – 2011.

[15]. R.S.Somasundaram , R.Nedunchezhian , "Evaluation of Three Simple Imputation Methods for Enhancing Preprocessing of Data with Missing Values", International Journal of Computer Applications (0975 – 8887) Volume 21 – No.10 ,pp.14-19 ,May 2011.

[16]. K.Raja , G.Tholkappia Arasu , Chitra. S.Nair , "Imputation Framework for Missing Value" , International Journal of Computer Trends and Technology – volume3Issue2 – 2012.

[17]. BOB L.Wall , Jeff K.Elser – "Imputation of Missing Data for Input to Support Vector Machines" ,