Using bagging k nn imputation english language essay

Linguistics, English

Abstract— Researchers in the database community have aroused great interest in handling high dimensional data sets for the past decades. Today's business captures inundate sets of data which includes digital documents, web pages-customer databases, hyper-spectral imagery, social networks, gene arrays, proteomics data, neurobiological signals, high dimensional dynamical systems, sensor networks, financial transactions and traffic statistics thereby generating massive high dimensional datasets. DNA microarray paves methods in identifying different expression levels of thousands of genes during biological process. The problem with microarrays is to measure gene expression from thousands of genes (features) from only tens of hundreds of samples. Microarray data often contain several missing values that may affect subsequent analysis. In this paper, a novel approach on imputation using k-NN with bagging method is proposed to handle missing value. The experimental result shows that the proposed method outperforms other methods in terms of distance and density of clusters. The proposed approach has enhanced the performance of traditional k-NN impute using bagging method. Keywords- clustering, microarray, missing value, bagging

Introduction

The various developments in the current century invest in collecting and storing unimaginable size of data. Microarrays, micro RNA expression, methalaytion, proteomics, functional MRI's, finance – time series data, climate data – spatial, spatio-temporal, neuro imaging, netflix moving rating data are a few high dimensional sources flooding data in scientific and business world. Microarray technology monitors thousands of gene expression levels simultaneously [1]. Matrices with rows(expression level of genes) and columns(experimental conditions) were produced at different conditions of microarray experiments. Most often, expression time series under different conditions, enables obtaining significant implications from gene expression data. The derived data sets may be evaluated by methods such as cluster analysis, correlation analysis and determination of mutual information content. The resulting data sets obtained from the experiments often consist of some missing values which may be due to scratches or spots on the slide, dust, insufficient resolution, image corruption and hybridization failures [2]. Hence microarray data exploration needs to preprocess in handling or predicting the missing value. The first method to handle the missing value includes: eliminating the records which has even a single value missing, that loses the entire entry which could provide useful information thereby produce invalid results [3]. The second approach involves a numerical substitution by a constant 0[3] or mean substitution [4], which may misinterpret the association among variables. Imputation is the third approach that selects the gene with missing value first and predicts them using the observed values of selected gene. The weighted K-nearest neighbor imputation (k-NN impute) is most commonly used method which reforms the missing values using a weighted average of k most similar genes [5]. It has better performance than simple methods except that it requires the input parameter k from the user. There are cluster-based algorithms [2, 6] that deal with missing values with no user-specific parameters [7]. These methods make no use on external information except on the expression data. In this paper, a novel approach is proposed for estimating (predicting) missing values using k-NN bagging based imputation method to handle incomplete data set. The evaluation results on gene annotations have shown that the proposed method is an effective approach for clustering incomplete gene expression data. The structure of this paper is as follows: Section 2 reviews the related work. In Section 3, the proposed missing value handling technique is discussed. Section 4 presents the performance evaluation, and finally conclusion on the proposed work is discussed in Section 5.

RELATED WORK

DNA microarray technology measures the mRNA levels of thousands of genes under certain experiments simultaneously. It gives a global overview of gene expression profiles in particular cells or tissues, so it has become one of the most prominent tools In functional genomics gene expression profile in tissues has become a vital tool in research. The objective of this research is to discover various biological classifications, identification of relevant genes and cancer predictions[8-14]. The existing multivariate analysis methods generate undesirable results in spite of the high percentage of missing values, e. g. hierarchical clustering and the support vector machine classifier [15, 16]. Moreover, many analysis methods such as principal component analysis (PCA), singular value decomposition (SVD) and generalized SVD (GSVD) cannot be applied to datasets with missing values [17-19]. An estimate on microarray datasets says more than 5% missing values and up to 90% of genes are affected [20, 21]. Even though unnoticed, missing value imputation minimizes the effect of missing values on the microarray data analysis. It is necessary, to confirm the exploration method of the microarray data with missing values by repeating the experiments, that is very costly and time consuming [22, 23]. Deleting the genes with missing values from promoting analysis, swapping missing values by zeros, or filling with row or column averages are the other methods to handle missing values [24, 25]. Correlation of data need to considered for presenting the best method which utilizes the relationships of data found in the dataset, which is missing in the above methods [26]. Hence, several imputation methods have been proposed since 2001, such as k-nearest neighbor (KNN), singular value decomposition(SVD), local least square(LLS), Bayesian principal component analysis(BPCA), Gaussian mixture clustering (GMC), collateral missing value estimation(CMVE) and weighted nearest neighbors method(WeNNI)[21, 23, 27-30]. The above said methods are completely based on the gene expression datasets themselves with no usage of external microarray datasets or biological related information.

MISSING VALUE HANDLING TECHNIQUES

The three most familiar, approaches on missing value handling techniques are studied in this section.

3. 1 Listwise Deletion (LD)

The foremost and commonly used method for dealing missing data was Listwise deletion method. The method considers only the row with complete data, whereas disregards the row containing a single lost value [31]. Most of the studies issued during 1998 and 2002 had used this method. Most statistical packages like SPSS, SYSTAT and SAS had the method as default setting for multivariate and univariate statistical procedures. It can be used when correlations among the variables are less and when the missing values are very small [35]. Unfairness occurs only when the data are missing completely at random, otherwise it discard only a few of the missing entries of data [31]. In linear regression method, the probability of missingness depends on the value of that variable, where listwise deletion is tough on an independent variable, whereas in a logistic regression model, listwise deletion can accept either nonrandom missingness on the dependent variable or nonrandom missingness on the independent variables.

3. 2. Pairwise Deletion (PD)

PD preserves all available data provided by a subject. It undergoes calculation of statistical inferences such as t-, z-. chi-square and descriptive statistics from non-filled data on each variable. While comparing PD with LD, it's not widely used as LD, only 7. 6% of the studies used PD. The statistical packages like SPSS, SYSTAT, and SAS has the default setting for descriptive, correlation, and regression analysis using either correlation or covariance matrices. PD may be used when there is a small number of missing cases on each variable relative to the total sample size, and a large number of variables are involved[35]. The PD method when related with LD method, takes interpretations from incomplete data. One shortcoming is that the sample data change from variable to variable, which paves difficulty in determining sample size and degrees of freedom. Since the entire data matrix is considered for multivariate statistical analyses the result may become undesired. Algorithm: List/PairWise DeletionProducing complete sample set from the given experimental samples, S by deleting the records whose attribute contains missing value. Input: Sample set, S with missing gene values. Output: Sample set, S, contains instances with no missing gene values.

Method:

for each sample smp in Sfor each gene g of smpif value of g is nulldelete smp{end if}{end for}{end for}

3. 3. Mean Substitution (MS)

The MS method uses mean computation to fill the missing values in the data [35]. Mean of the variable is considered to be one optimal way of filling the missing value of the variable. MS does not skip any information that is available rather than LD and PD methods. Few of the reviews have used this method to handle missing data. Missing value imputation using sub-group mean is another deviation of Mean Substitution. For example, if the sample with a missing value is a Democrat, the mean for all democrat is computed and inserted in place of the missing value. Although this method is not a conventional method of replacing the overall mean of the variable [35]. In spite of the variations Mean Substitution has some drawbacks such as overrated sample size, undervalued variance, correlation being undesirably biased and incorrect illustration of the computed values being distributed since the addition of values equal to the mean. Such partiality presented into the correlation, population variance, and variable distribution highly subject to the extent of missing data on the actual values that are absent [35]. Algorithm: Replacing missing gene by substituting the mean value of other gene in the sample setInput: Sample set, S with missing gene valuesOutput: Sample set, S, contains instances with no missing gene values.

Method:

for each selected gene g in SCalculate the Arithmetic mean Value (AMV) of gene{end for}for each selected gene g in Sfor each sample smp of gif the value of g is nullfill the value of g as AMV{end if }{end for}{end for}

PROPOSED METHOD

The proposed work undergoes data collection, data transformation, missing value handling and prediction models using clustering techniques. Figure 1 depicts proposed framework for handling missing value in gene expression data. First, the data set is collected from the public data set of Gene Expression Pattern Analysis Suite –Yeast cell cycle[36]. The collected raw dataset is preprocessed using min-max normalization and further discretized. Then the missing value is treated with the existing and proposed methods. Finally, clusters are formed and the results are analyzed.

4. 1 Data Preprocessing

Data transformation undergoes normalization and discretization. Normalization [32] is a data preprocessing tool used in data mining system. The normalization process undergoes transforming the values of the variable in the dataset such that they lie within the specified range such as 0 to 1. The techniques such as classification algorithms involving neural networks, distance measurements such as nearest neighbor classification and clustering widely use normalization process. Min-max normalization, z-score normalization and normalization by decimal scaling are the commonly used data normalization methods.

4. 1. 1 Min Max Normalization:

This method does a linear transformation on the original data. Min-max normalization maps a value dx of PV to dx′ in the range [new_min(p), new_max(p)]. The min-max normalization is calculated by the following formula:(1)where min(v) = minimum value of variable, max(v) = maximum value of variable. In this case min-max normalization maps a value x of V to x′ in the range [0, 1], so put new_min(v)= 0 and new_max(v)= 1 in the above equation. The simplified formula of min-max normalization is given below:(2)Min max normalization conserves the association among the original data values.

4. 2 Discretization:

The objective of discretization process is to induce a list of intervals that split the numerical domain of a continuous explanatory attribute. EFD [33]Equal Frequency Discretization divides the sorted values into k intervals so that each interval contains approximately the same number of training instances. Thus each interval contains n= k (possibly duplicated) adjacent values. K is a user predefined parameter.

4. 3 Proposed Approach – k-NN Imputation with Bagging

This paper focuses on enhancing the performance of traditional k-NN approach with the help of bagging technique.

4. 3. 1 k-NN impute algorithm

The k-NN-based method selects genes to impute missing values with expression profiles similar to the gene of interest. While considering gene A that has one missing value in experiment 1, this method would find K other genes, which have a value present in experiment 1, with expression most similar to A in experiments 2–N (where N is the total number of experiments). A weighted average of values in experiment 1 from the K closest genes is then used as an estimate for the missing value in gene A. The similarity of each gene with weight is considered under weighted average for each gene. After examining a number of metrics for gene similarity (Pearson correlation, Euclidean distance, variance minimization), it shows that Euclidean distance was a sufficiently accurate norm. This finding is somewhat surprising, given that the Euclidean distance measure is often sensitive to outliers, which could be present in microarray data. However, it is found that log-transforming the data seems to sufficiently reduce the effect of outliers on gene similarity determination.

Algorithm: k-NN impute

Input: Sample set, S with missing gene values. Output: Sample set, S, contains instances with no missing gene valuesSelect the gene A with expression which contains missing valueInput the value k for finding neighbour gene with complete expression to find k nearest neighbours. Compute the Euclidean distance between the gene A with missing value and all the training expression with absolute gene value. Sort the distance and determine k no of nearest neighbors based on the kth minimum distance.

Group the categories of those neighbors. Substitute the missing value by corresponding gene value based on weighted average of the most similar complete gene expressionBagging merges a huge number of learners, where each learner uses a bootstrap sample of the original training set. In this paper the k-nn classifier is used as a learner. A Bootstrap sample is generated by uniformly sampling m gene expressions from the training set with replacementT bootstrap samples BS1; BS2… BST is generated and a classifier Ci is built from each bootstrap sample Bi. A final classifier C* is built from C1; C2… CT whose output is the class predicted most often by its sub-classifiers, with ties broken arbitrarily.

4. 3. 2 Bagging Procedure

Classifier generation

Step 1. Create t data sets from a database applying the sampling with replacement scheme. Step 2. Apply a learning algorithm to each sample training data set. Step 3. For an object with unknown decision, make predictions with each of the t classifiers. Step 4. Select the most frequently predicted decision

Algorithm

Build the model: for m = 1 to M:(a) Bootstrap sample BSm of size N with replacement from the original training set S with equal weight.(b) Train a k-nn Gm(x) to the bootstrap sample BSm.

Predicting:

For m = 1 to M: Apply Gm to the testing set BTT . Classifier using

5. Prediction models using clustering technique

Data clustering aims at grouping of objects such that data points that belong to one cluster are similar whereas objects in different clusters are distinct. Clustering can define dense and sparse regions further identify overall distribution patterns and interesting correlations among data attributes. The selection of clustering algorithm depends on the type of the data and the purpose and application. Partitioning, hierarchical, density-based methods, grid-based methods and model-based methods are the various clustering techniques. 5. 1 DBSCAN: The density-based clustering methods aim at discovering arbitrary shape clusters. The clusters are represented as dense region of objects in the object space separated from low density regions. The DBSCAN(Density Based Spatial Clustering of Applications with Noise) [37] is a density-based clustering algorithm which defines clusters as density-connected points. Every point not contained in any cluster is considered to be noise. This method requires two input parameters: minimum objects(μ) and radius(ε). The method has the following definitions: Definition 1: (ε – neighborhood of a point) The neighborhood within a radius ε of a given object is the ε-neighborhood of the object, defined as NEps(x) = { x D | dist(x, q) ≤Eps }Core object and border point: If the ε-neighborhood of a point contains atleast a minimum number of points then the point is said to be core object and the points on the border of the cluster is border point. Definition 2: (Directly density-reachable) A point p is directly density-reachable from a point q wrt. Eps, MinPts ifx NEps(y) and| NEps(y)| ≥ MinPtsDefinition 3: (Density-reachable) A point x is density-reachable from a point y wrt. Eps and MinPts if there is a chain of points x1,..., xn, x1= y, xn=

x such that xi+1 is directly density-reachable from xi. Definition 4 : (Density-connected) A point x is density-connected to a point y wrt. Eps and MinPts if there is a point o such that both, x and y are density-reachable from o wrt. Eps and MinPts. Definition 5: (Cluster) Let D be a database of points. A cluster C wrt. Eps and MinPts is a non-empty subset of D satisfying the following conditions: i) x, y: if x C and y is density-reachable from x wrt. Eps and MinPts, then y C. ii) x, y C : x is density-connected to y wrt. Eps and MinPts. Definition 6: (noise) Let C1,…Ck be the clusters of the database D wrt. Parameters Epsi and MinPtsi, i= 1,…., k. The noise is defined as the set of points in the database D not belonging to any cluster Ci. i. e. noise= { x D | i: x Ci }

Description of the Algorithm

Select an arbitrary point xRetrieve all points density-reachable from x wrt. Eps and MinPtsIf x is a core point, a cluster is formedIf x is a border point, no points are density reachable from x, the method selects the next point of the database. The process is continued until all the points have been visited.

5. 2 Support vector clustering

Support Vector clustering does not take user specific parameters such as number or shape of clusters, which is more suitable for low-dimensional data. Kernel function is used to map data space to high dimensional feature space.

Farthest first

Farthest first[34] is a variant of K Means which points each cluster centre in turn at the point furthermost from the existing cluster centre. This point must lie within the data area. This greatly speeds up the clustering in most of the cases since less reassignment.

5. 4 K-mean

Simple K-means algorithm is a type of unsupervised algorithm in which items are moved among the set of cluster until required set is reached. This algorithm is used to classify the data set, provided the number of cluster is given in prior. This algorithm is iterative in nature. The algorithm for Simple K-means algorithm is given as:

Algorithm: Simple K-means clustering algorithm

Input: Set of Elements or Database of transactionD= {t1, t2, t3, …., tn}Number of required Cluster kOutput: Set of Cluster K

Method:

Make initial guesses for the means m1, m2... mk; RepeatAssign each element ti to the cluster having the closest mean. Calculate the new mean for each cluster. Until there are no changes in any mean.

5. 5 X-mean

In the k-means algorithm, the number of clusters k is an input parameter specified by the user. In the x-means algorithm, the BIC or Schwarz criterion is used globally and locally in order to find the best number of clusters k. Given a data set D = {x1, x2, . . . , xn} containing n objects in a d-dimensional space and a family of alternative models Mj= {C1, C2, . . . , Ck}, (e. g., different models correspond to solutions with different values of k), the posterior probabilities P(Mj| D) are used to score the models. The Schwarz criterion can be used to approximate the posteriors.

EXPERIMENTAL RESULTS

The experiment was conducted on the publicly available yeast cell cycle dataset which contains 6178 genes with 77 conditions each. There are 28127 missing values (5. 9 % of the total)[36]. The performance metrics used in this experiment are cluster distance and performance of cluster density. Table 1 & 2 provides the results of the performance of elimination method using average within cluster distance and within centroid distance. Table 3 & 4 provides the results of the performance of replacing by average mean method using average within cluster distance and within centroid distance. Table 5 & 6 provides the results of the performance of k-NN imputation method using average within cluster distance and within centroid distance. Table 7 provides the results of the performance of enhanced k-NN bagging method using average within cluster distance. The performance of k-NN impute with bagging shows better results rather than k-NN imputation. In this work the implementations of algorithms were done by a machine learning algorithm tool Rapid Miner version 5. The performance results of nominal values are to be assumed as positive values. Figure 2 shows the average within cluster distance for elimination method. Figure 3 shows the average within cluster distance using average mean. Figure 4 shows the average within cluster distance of k-NN imputation. Figure 4 shows the average within cluster distance of k-NN imputation with bagging method.

Vi. CONCLUSION

In this paper, we explore missing value handling using the different techniques. Gene expression data suffer from missing values which may affect subsequent analysis. The main motivation of the proposed work is that k-NN is an unstable learning algorithm and Bagging works well for " unstable" learning algorithms. So instead of using traditional k-nn imputation the performance of predicting missing value was greatly improved by bagging k-nn. The performance evaluation shows that among the other missing value handling techniques the proposed method performs best.