*Journal of Computer Science*

Original Research Paper

# Missing Values Treatment and Feature Reduction Analysis to Enhance Classification

**Muralidharan, D., K. Renuka, Mulagala Jaswant, J. Karthikeyan and G.R. Brindha**

*SoC, SASTRA Deemed University, India*

*Article history:*
Received: 07-07-2019
Revised: 16-09-2019
Accepted: 19-02-2020

Corresponding Author:
Brindha, G.R.
SoC, SASTRA Deemed University, India
Email: brindha.gr@ict.sastra.edu

**Abstract:** Datasets may have large number of features which makes it hard and time consuming to classify. Additionally, they may have irrelevant and noise features too with missing values. The missing values should be treated in a proper way so that the classifier accuracy can be improved. There is also a need to reduce features and select only the features necessary to the classifier. Principal Component Analysis (PCA) is commonly considered for this process of reducing the number of features in a dataset. These reduced components can be applied as input to the classifiers. In this study, standard datasets are checked for missing values, classified using Support vector Machines (SVM) and Naive Bayes with and without reducing the features using PCA. Then, the proposed algorithm for missing value imputation is used on the datasets and the same analysis were carried out. The accuracy is evaluated using Confusion Matrix. The results are discussed with analysis based on the nature of features and missing values and how different datasets behave when used with machine learning algorithms.

**Keywords:** PCA, SVM, Naive Bayes, Missing Value Treatment

## Introduction

Databases have implicit information which are very important and are not explicitly known to everyone. Data mining techniques are used to extract this information by considering all records of the database. Each and every record of database has a lot of attributes to aid extracting any needed information from the database. However, the humongous volume of data base made the mining process very difficult and slow. When extracting one special information from the database, the task might be made difficult by the existence of other irrelevant attributes. Sometimes, some of the attributes act adversely and the obtained information may not be accurate. As these attributes might be important when obtaining some other specific information, they cannot be removed permanently from the database. Hence, it is important to consider only the relevant attributes and remove other attributes. This is called as 'feature selection'. Sometimes a new feature might be formed from the available multi-features, called 'feature extraction'. These methods have twofold advantage. First one is that the obtained information will be more accurate. The second advantage is the enhancement achieved in the processing speeds as only a subset of attributes are considered for extraction.

As mentioned above, this obtained information might be inaccurate and the obtained process will be made complex if the feature size is not reduced. But it is a very difficult task to consider all features to identify the relevant ones. Raymer *et al*. (2000) considered the task of choosing mainly relevant attributes. Qu *et al*. (2005) analyzed the correlation of features to minimize the dimension. Janecek *et al*. (2008) investigated the relationship between feature selection and classification accuracy. Intuitively it can be accepted that the accuracy will be reduced when wrong features are selected and/or relevant features are not selected by the feature selection algorithm. Burges and Christopher (2010) presented a guided tour in machine learning approach to reduce the dimensionality of records. PCA (Jolliffe, 1986; Sehgal *et al*., 2014) is one of the most accepted techniques for dimensionality reduction.

Spearman correlation method is used in improved imputation method to find the missing values and the performance of classification is given in RoC curve for the methods SVM, NB and KNN Elenita *et al*. (2019), Ghorbani and Desmarais (2017; Schmitt *et al*., 2015). Data set with missing values influence the algorithm by weaken it and reduce the accuracy (Sim *et al*. (2015; Kanchana and Thanamani, 2016). For classification problems, Naive Bayes classifier and SVM (Cristianini and Shawe-Taylor, 2000) are used in widespread. To improve the accuracy for SVM classifiers, Deisy *et al*. (2010) proposed a novel feature selection algorithm based on information theory.

Data mining techniques are classified into supervised and unsupervised learning. Unsupervised learning techniques include clustering which is used for applications like image processing. But here we concern with only the supervised learning which uses the class attribute and therefore called classification models. Support Vector Machine has attracted much attention recently and has been successfully used in various applications Cortes and Vapnik (1995; Kao *et al*., 2013; Muthu Rama Krishnan *et al*., 2010; Xie and Wang, 2011; Burbidge *et al*., 2001). As SVM produces accurate classification for both linear and non-linear relationships, it is preferred over other classifiers.

Another simplest learning algorithm is Naive Bayes. It is based on the assumption of conditional independence. Though this conditional independence assumption is hardly ever seen in the real world, Naive Bayes classifier is classifying satisfactorily. This is because of the conditional independence assumption holds good if the dependences are distributed evenly among classes and/or if the total dependences may be ignored as they may cancel out each other.

The details of existing algorithms SVM, NB and the proposed algorithm are given in Materials and method section. The meta data about data set is also provided. The result and discussion section depicts the performance of the algorithm through comparative tables and graphs.

## Materials and Methods

Process I of Fig. 1 depicts that if the data contains missing values, then the proposed treatment algorithm updates the dataset and given as input to the classifier for training (70% of data) and the model is created and tested by 30% of test dataset. The same process is employed after applying PCA (II). The classified results are compared and analysed.

The data sets used for the proposed process (see references for links) is given in Table 1. Among 10 sets, last 5(in bold letters) have missing values and the features are both numerical and categorical.

### *Missing Values Treatment Algorithm*

The missing values are considered as the mean of the attribute column for numerical data base.

The missing values of a record are replaced by the attribute values of the least distance record from it. The algorithm is divided into 4 phases,

**Phase 1:** The vector which has missing values for an attribute and the records which have categorical attributes are considered as a set. Mathematically in the set for a record i, when $R(i, k) = ?$ then $R(j, k) \neq ?$ at least for one value of $k$ of the remaining records.

**Phase 2:** The Euclidean distances from the original vector to all other vectors are calculated using the following formula:

$$D(i,j)=\sum_{\substack{k=1;R(i,k)\neq NA\\ R(j,k)\neq NA}}^{M}(R(i,k)-R(j,k))^2$$

where, $R(i,k)$ is the k$^{th}$ attribute value of i$^{th}$ record.

Sinilarly, $R(j, k)$ is the k$^{th}$ attribute value of $j^{th}$ record. Missing value is mentioned as *NA*.

**Phase 3:** Find the least distance record '*j*' from '*i*' using the above equation and replace the missing value of $R(i, k)$ by the available value of $R(j, k)$.

**Phase 4:** Repeat Phase 1 to Phase 3 until all missing values of all records are replaced.

The following pseudo code explains the steps to replace the missing values in categorical databases

*Given: The Database has N Records with M Attributes*

```
while (the database contains missing values)
  i=0;              //First record
 while (i<N)          //Repeat till last record
  if (no missing value in the record)
   continue with next 'i' value
  else
   j=0        // the record to be compared
   while (j< N)
          // Compare Rec[i] others
    if ( i==j)
      continue with the next 'j' value
    else
                  //Compare the two records attribute wise
                  // kth attribute is compared.
      if ( Rec[i,k]==NA &  Rec[j,k]==NA))
            continue with the next 'j' value
          else if (attribute is numerical)
       Euclidean distance (Rec[i] and Rec[j]);
       Find the record 'j' which has the
       least Euclidian distance from record 'i';
             Replace the missing attributes of record
             'i' with the corresponding attributes of
             record 'j';
     else if (attribute is categorical)
       Cosine distance (Rec[i] and Rec[j]);
       Find the record 'j' which has the
       least distance from record 'i';
              Replace the missing attributes of record
              'i' with the corresponding attributes
              of record 'j';
```

### *Principal Component Analysis*

Principal Components Analysis (PCA) uses a set of linearly uncorrelated variables as principal components. To reduce the features, PCA converts a set of correlated variables in to a set of principal components. In this way, PCA reduces the number of features and selects or creates features which are vital for classification.

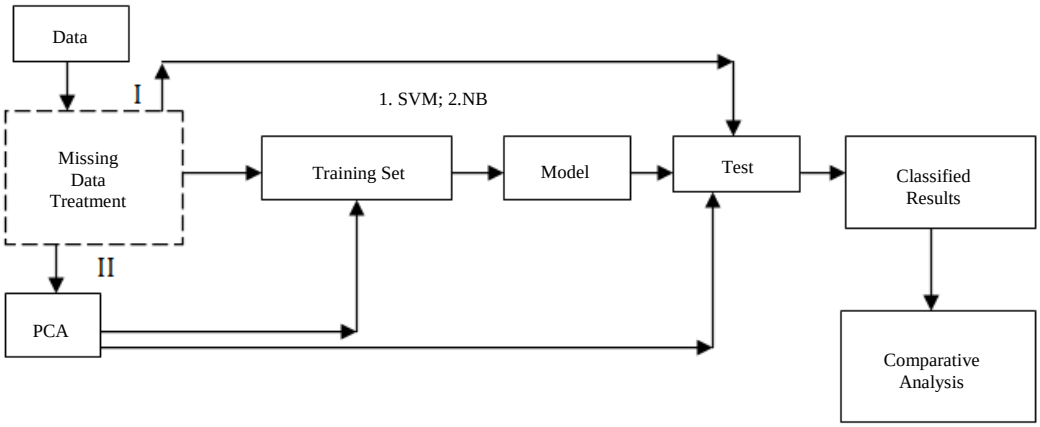


**Fig. 1:** Process flow of the proposed process

**Table 1:** Data set (without missing values) description

| Name | Size | Numerical | Categorical | Class |
|---|---|---|---|---|
| Frog-Species | 7196 | 23 | 3 | 3 |
| Ionosphere | 351 | 34 | 0 | 2 |
| Biodeg | 1055 | 41 | 0 | 2 |
| Sonar | 208 | 60 | 0 | 2 |
| Flight_delay | 2202 | 7 | 6 | 2 |
| Data set (with missing values) description | | | | |
| Credit | 690 | 6 | 9 | 2 |
| BreastCancer Wisconsin | 699 | 0 | 9 | 2 |
| Statlog | 270 | 5 | 8 | 2 |
| Autistic Spectrum | 104 | 3 | 18 | 2 |
| BreastCancer | 286 | 3 | 6 | 2 |

This mapping is done with the assumption of high variance gives high information. Eigen value decomposition on the centred Kernel matrix $K^c$ is used to do Principal Component Analysis in the kernel space (Pearson, 1901):

$$K^c(i,j) = \varphi(x_i).\varphi(x_j)$$
$$\varphi(x_j) = \varphi(x_i) - \frac{1}{N}\sum_{j=1}^{N}\varphi(x_j)$$

When, the variance in $k^{th}$ principal direction of the kernel space is $\sigma_k^2$ the Eigen values of $K^c$,$\lambda_k = N\sigma^2_k$

## *Support Vector Machine*

Support Vector Machine (SVM) comes under supervised learning algorithms, is naturally used for classifying problems (Vapnik, 1995). SVM does not only linear but also non-linear classifications accurately. It does classification in two major steps. First, the inner product of the data points in kernel, the feature space, is obtained. A hyper plane learning algorithm is applied in the feature space as the second step. The hyper plane consists of theoretically infinite number of planes to split the data sets for classification. Mathematically, SVM constructs linearly separating hyper planes in high dimensional vector spaces. Data points are indicated by a pair of tuples (feature values, classification). When such hyper planes provide maximal distance to the nearest data points, optimal classification will be occurred in training. This is obvious when the distance between nearest data points are less, no ambiguity occurs during classification.

## *Naive Bayes*

One of the most efficient and effective classification algorithms is Naive Bayes. It works based on the conditional probability theorem given by Bayes and got its name. The classification is made simple in Naive Bayes classifier when the features are independent for the given class variable. The classifier uses the following formula:

$$f_i(X) = \prod_{j=1}^{N} P(x_j \mid c_i)\, P(c_i)$$

where, $X = (x_1, x_2,..., x_N)$ denotes a feature vector and $c_j$, $j = 1,2,...,N$, denote possible class labels. Although independence is generally poor assumption NB generally competes well with more sophisticated classifiers. To avoid $f_i(X)$ becomes zero when $P(x)$ is zero, Laplace Estimator is used instead of Naive Bayes.

## Results and Discussion

The missing values treatment, PCA influence and performance of classifiers are discussed in this section. Table 2 depicts the accuracy results of the data sets with missing values. Initially the records with missing values are removed and given to the classifiers. Then the proposed missing value treatment algorithm is applied to the datasets and the missing values are updated. Because of the removal process, the data set size is reduced and so the classifiers have lack of training set which in turn leads to reduced accuracy. Whereas, the proposed missing values treatment process fills the records with perfect values and the accuracy is increased for both classifiers.

The analysis of accuracy reveals that based on the nature of the features, PCA can be applied (Fig. 2 and 3) Since every data set is different from each other as there are several feature processing, selection methods and threshold values. This study applies the popular PCA to convert the features into components. To get the number of components, the threshold is set as t = (Min(component value) + Max(Component value))/2; The component above this threshold values are the inputs to the classifiers. But there is no general rule that application of PCA will increase the classification performance. Especially when the future values are categorical the application of PCA converts the values into components (numerical) and the algorithm is unable to learn from the components.

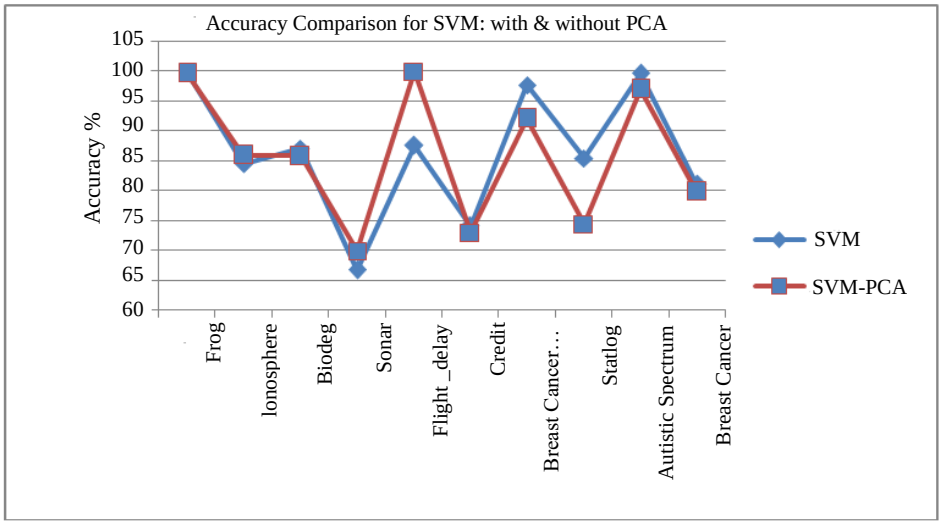


**Fig. 2:** PCA influence in SVM

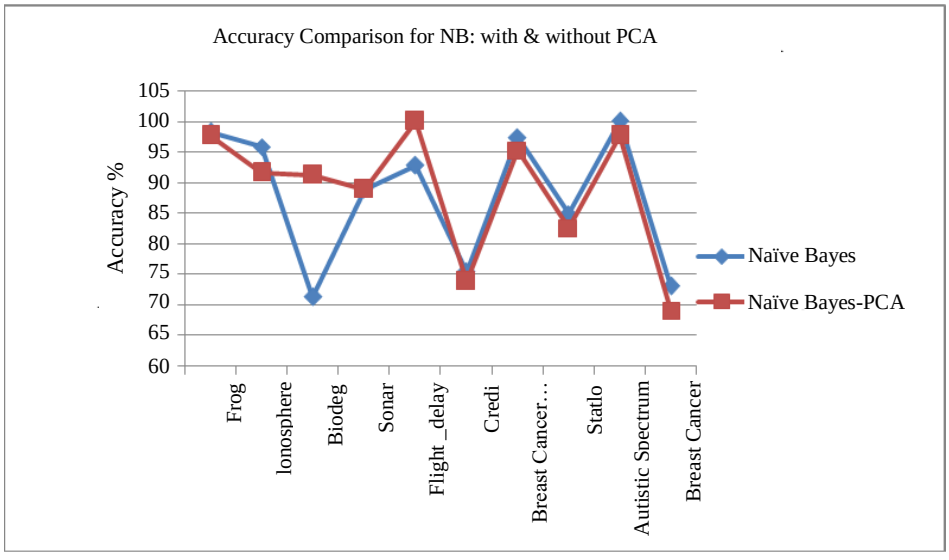


**Fig. 3:** PCA influence in NB

**Table 2:** Accuracy comparison between removal of missing values and treatment of missing values

| Dataset | Removal of rows with missing values + SVM | Treatment of missing values + SVM | Removal of rows with missing values + NB | Treatment of missing values + NB |
|---|---|---|---|---|
| Credit | 68.20% | 74.00% | 71.80% | 75.40% |
| BreastCancer Wisconsin | 92.30% | 97.74% | 93.60% | 97.30% |
| Statlog | 79.43% | 85.40% | 81.98% | 84.85% |
| Autistic Spectrum | 93.2% | 99.7% | 94.6% | 100% |
| BreastCancer | 76.7% | 81% | 67.4% | 73% |

**Table 3:** F1 Measure comparison: With and without PCA

| Dataset | SVM | SVM-PCA | Naïve Bayes | Naïve Bayes-PCA |
|---|---|---|---|---|
| Frog-Species | **99.53%** | 99.52% | 98.13% | 97.42% |
| Ionosphere | 84.37% | 85.76% | **95.54%** | 92.43% |
| Biodeg | 86.84% | 85.60% | 71.06% | **90.67%** |
| Sonar | 66.47% | 69.54% | 88.69% | **88.76%** |
| Flight_delay | 87.32% | 99.76% | 92.58% | **99.84%** |
| Credit | 73.67% | 72.56% | **75.06%** | 73.60% |
| BreastCancer Wisconsin | 97.44% | 91.25% | 97.05% | 95.03% |
| Statlog | 85.28% | 74.16% | 84.69% | 82.32% |
| Autistic Spectrum | 98.2% | 96.56% | 99.8% | 99.1% |
| BreastCancer | 80.1% | 78.4% | 71.9% | 68.7% |

This is obvious for the data sets Frog, Credit, Breast cancer Wisconsin, Statlog, Autistic Spectrum and Breast cancer. Whereas, the classifiers for Biodeg, Sonar and Flight_delay datasets in their PCA components format learn better and provided improved accuracy. But Ionosphere data is cannot be inferred in these cases. Though all the features of Ionosphere data are numerical values, most of them are negative values. So the classifiers unable learn from components and accuracy is reduced after PCA process. The weighted average of (Table 3) precision and recall measures also supports the discussion given based on accuracy results of classifiers compared with and without API); Fig. 3 says about the distribution.

## Conclusion

Infering from the graphs and the table we can see that the accuracy may decrease or increase if we use PCA before Naive Bayes and SVM after treating with the proposed method. Accuracy of naive bayes predominetly decreases while used with PCA. PCA combines the features and create a new set of features to classify. Probabilty of class features get affected while combining features.Since Naive Bayes uses probability clssification due to that PCA with Naive Bayes may not be efficient.So accuracy of Naive Bayes decreases when used with PCA. However accuracy of SVM may increases with PCA because both SVM uses vectors to classify and combining features may make it easier to use vectors for classification.

## Acknowledgement

This research was carried out at SASTRA University, School of Computing, Data science Research Lab and we would to extend our gratitude to the institution for providing us with the necessary computing resources and guidance.

## Author's Contributions

**Mulagala Jaswant and J. Karthikeyan:** Coding of CNN and formatting

**D. Muralidharan:** Content writing and Dataset selection.

**G.R. Brindha:** Structural design of paper and Analysis.

**K. Renuka:** Content writing and Content revision.

## Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

## References

Burbidge, R., M. Trotter, B. Buxton and S. Holden, 2001. Drug design by machine learning: Support vector machines for pharmaceutical data analysis. Comput. Chem., 26: 5-14.

Burges, C.J.C., 2010. Dimension reduction: A guided tour. Foundations Trends Machine Learning, 2: 275-365. DOI: 10.1561/2200000002

Cortes, C. and V. Vapnik, 1995. Support-vector networks. Machine Learning, 20: 273-297.

Cristianini, N. and J. Shawe-Taylor, 2000. An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge, England

Deisy, C., S. Baskar, N. Ramraj, J.S. Koori and P. Jeevanandam, 2010. A novel information theoretic-interact algorithm (IT-IN) for feature selection using three machine learning algorithms. Expert Syst. Applic., 37: 7589-7597.

Elenita, T.C., A.M. Sison and R.P. Medina, 2019. Application of the modified imputation method to missing data to increase classification performance. Proceedings of the IEEE 4th International Conference on Computer and Communication Systems, Feb. 23-25, IEEE Xplore press, Singapore. DOI: 10.1109/CCOMS.2019.8821632

Ghorbani, S. and M.C. Desmarais, 2017. Performance comparison of recent imputation methods for classification tasks over binary data. Appl. Artif. Intell., 31: 1-22.

Janecek, A., N. Wilfried M.D. Gansterer and E. Gerhard, 2008.On the relationship between feature selection and classification accuracy. J. Mach. Learning Res. Proc. 4: 90-105.

Jolliffe, I.T., 1986. Principal component analuysi. Springer Verlag, New York.

Kanchana, S. and A.S. Thanamani, 2016. Elevating the accuracy of missing data imputation using bolzano classifier. Int. J. Eng. Technol., 8: 138-145.

Kao, H.Y., T.K. Chang and Y.C. Chang, 2013. Classification of hospitalweb security efficiency using data envelopment analysis and support vector machine. Math. Problems Eng.

Muthu Rama Krishnan, M., S. Banerjee, C. Chakraborty, C. Chakraborty and A.K. Ray, 2010. Statistical analysis of mammographic features and its classification using support vector machine. Expert Syst. Applic., 37: 470-478.

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. Philosophical Magazine, 2: 559-572.

Qu, G., S. Hariri and M. Yousif, 2005. A new dependency and correlation analysis for features. IEEE Trans. Knowle. Data Eng., 17: 1199-1207. DOI: 10.1109/TKDE.2005.136

Raymer, M.L., F. William Punch, D. Erik Goodman, K. Leslie and A.K. Jain, 2000. Dimensionality reduction using genetic algorithms. IEEE Trans. Evolutionary Comput., 4: 164-171.

Schmitt, P., J. Mandel and M. Guedj, 2015. A comparison of six methods for missing data imputation. J. Biomet. Biostat., 6: 224. DOI: 10.4172/2155-6180.1000224

Sehgal, S., H. Singh, M. Agarwal, V. Bhasker and Shantanu, 2014. Data analysis using principal component analysis. Proceedings of the International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom), Nov 7-8, IEEE Xplore press, Greater Noida, India, Greater Noida, pp: 45-48. DOI: 10.1109/MedCom.2014.7005973

Sim, J., J.S. Lee and O. Kwon, 2015. Missing values and optimal selection of an imputation method and classification algorithm to improve the accuracy of ubiquitous computing applications.

Vapnik, V., 1995. The nature of statistical learning theory. Springer-Verlag, New York, NY.

Xie, J.Y. and C.X. Wang, 2011. Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases. Expert Syst. Applic., 38: 5809-5815.