Study of Influence of Dimension Reductions of High Dimensional Datasets in Classification Problem

Mohd. Salman Hossain Bhuiyan
ID: 2015–2–60–085

Nabil Al Raian
ID: 2015–2–60–081

Shahad Iqbal Leon
ID: 2015–2–60–110

A thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering

Department of Computer Science and Engineering
East West University
Dhaka-1212, Bangladesh

December, 2019

ABSTRACT

In our day to day life we develop many applications based on datasets. In case of high dimensional dataset, we often face some problems while building any data mining model. When there are too many attributes in the dataset, then there may be dependency between attributes. There may be some irrelevant attributes too. So, we get less accuracy in the data mining because of the influence of dependent and irrelevant attributes. So, in order to solve this problem, we need to reduce the dimensions of the dataset. In this work, we experimentally tested influence of dimension reduction on classification problems. For this purpose, we used 4 different datasets. We used backward elimination method to reduce the dimension of the dataset down to seven dimensions. We have experimented with Multi-Layer Perceptron, Naïve Bayes, Decision Tree, K-Nearest Neighbor, and Support Vector Machine classification methods. We used 10-fold validation to train and test our dataset. Experimental results show that when the dimension is reduced, then the accuracy is improved for some classification algorithm like Multi-Layer Perceptron, Naïve Bayes and Random Forest. We come up with a conclusion that if we exclude the less significant attributes, then the classification model gives better accuracy than it does without dimension reduction.

DECLARATION

We, hereby, declare that the work presented in this thesis is the outcome of the investi- gation performed by us under the supervision of **Musharrat Khan**, Department of Computer Science and Engineering, East West University. We also declare that no part of this thesis has been or is being submitted elsewhere for the award of any degree or diploma.

Countersigned

.....

(**Musharrat Khan)**
Supervisor

Signature

.....

(Shahad Iqbal Leon)
ID: 2015–2–60–110

Signature

.....

(Mohd. Salman Hossain Bhuiyan)
ID: 2015–2–60–085

Signature

.....

(Nabil Al Raian)
ID: 2015–2–60–081

LETTER OF ACCEPTANCE

We, hereby, declare the thesis is from the student's own work and best effort of us and all other sources of information used in this paper have been acknowledged. This thesis has been submitted with our approval.

Supervisor

Signature

.....

(Musharrat Khan)

Department of Computer Science and Engineering

East West University, Dhaka

Chairperson

Signature

.....

(Dr. Taskeed Jabid)

Associate Professor

Department of Computer Science and Engineering

East West University

ACKNOWLEDGMENTS

As it is true for everyone, we have also arrived at this point of achieving a goal in our life through various interactions with and help from other people. However, written words are often elusive and harbor diverse interpretations even in one's mother language. Therefore, we would not like to make efforts to find best words to express our thankfulness other than simply listing those people who have contributed to this thesis itself in an essential way. This work was carried out in the Department of Computer Science and Engineering at East West University, Bangladesh.

First of all, we would like to express our deepest gratitude to the almighty Allah for the blessings on us. Next, our special thanks go to our supervisor, **Musharrat Khan**, who gave us this opportunity, initiated us into the field of "**Study of Influence of Dimension Reductions of High Dimensional Datasets in Classification Problem",** and without whom this work would not have been possible. Her encouragements, visionaries and thoughtful comments and suggestions, unforgettable support at every stage of our B.Sc. study were simply appreciating and essential. Her ability to help us enough to finally answer our own question correctly is something valuable what we have learned and we would try to emulate, if ever we get the opportunity.

There are numerous other people too who have shown me their constant support and friendship in various ways, directly or indirectly related to our academic life. We will remember them in our heart and hope to find a more appropriate place to acknowledge them in the future.

Table of Contents

Table of Figures

CHAPTER 1 INTRODUCTION

1.1 Classification Problem

A classification problem is a problem where we try to predict the class or category for a given observation. It is like a mapping function (f) from input variable(x) to output variable(y). The output variables are called labels. Classification problems predict a continuous value as the probability for each output class. The probabilities can be the likelihood or confidence belonging to each class. A predicted value can be converted into class label by selecting the class label that has the highest probability. An algorithm that is capable of learning a predictive classification model is called classification algorithm. The accuracy of a classification algorithm is the percentage of correctly classified examples out of all predictions made.

1.2 Common Classification Techniques

There are some commonly used classification techniques which we use very often. Such as K- nearest neighbors (KNN), Neural Network (NN), Decision tree (DT), Support vector machine (SVM), Naïve Bayes, Random forest (RF). We will briefly discuss about them one by one.

1.2.1 K-nearest neighbors

K-nearest neighbors is a classification model that classifies data points based on the points that are similar to it. It uses the test data to make a guess on where an unclassified data should be classified. The first step of implementing k-nearest neighbors is to transform the data points into mathematical value. Then the algorithm finds the distance between these mathematical values of these points. The most common way of finding the distance is the Euclidian distance[1]. If the point is (x1, y1) and (x2, y2) in 2-dimensional space, then the Euclidean distance between them is,

$$\sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

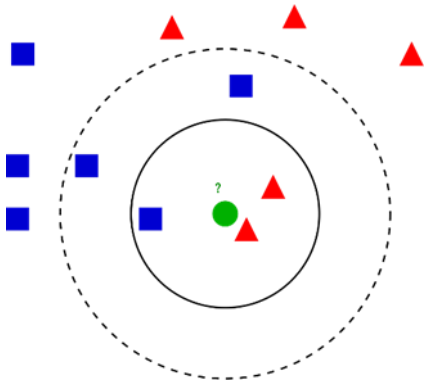

Figure 1. 1 K-nearest neighbors

The test sample (green dot) should be classified either to blue squares or to red triangles. If *k = 3* (solid line circle) it is assigned to the red triangles because there are 2 triangles and only 1 square inside the inner circle. If *k = 5* (dashed line circle) it is assigned to the blue squares (3 squares vs. 2 triangles inside the outer circle)[2].

1.2.2 Random forest

The random forest algorithm is composed of different decision trees each of the trees have same nodes but different data that leads to different leaves. Then it merges the decisions of all decision tree that were previously generated and find an answer which is the average of all the decision trees. The random forest algorithm is a supervised learning model. Before applying random forest, we need to perform something called one-hot encoding which means assigning a number to a categorical value. While performing random forest in a particular dataset we often use Gini index. The Gini index formula uses the class and probability to determine the Gini of each branch on a node. Entropy can also be used to determine how nodes branches in a decision tree. Entropy uses the probability of a certain outcome in order to make a decision how the node should branch[1].

1.2.3 Naïve Bayes

Naïve Bayes classifier is a classification algorithm that classifies data points based on Bayes theorem. It is called naïve because each assume feature of data points is completely independent to one another. The algorithm uses the probabilities of certain events being true given other events are true in order to make predictions on new data point. This is the fact that makes this algorithm so unique compared to other classification algorithms[1].

1.2.4 Decision tree

Decision tree is a supervised machine learning which delegate simple classifying example. By using decision tree, a classification model can be made from some input data set. Decision tree build a structure like tree where some conditions and questions organized a tree structure. After decision tree formed, some test data has been applied on that tree and appropriate branch based on outcome has been found. Here, each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. Decision tree starts from the tree root and data are split into some features[1].

1.2.5 Multi-layer Feed-forward neural network

The back-propagation algorithm executes learning on a multilayer feed-forward neural network. Feed-forward neural network is an inner connection of perceptron. It iteratively learns a set of weights for prediction of the class label of tuples. A multilayer feed-forward neural network consists of an input layer, each layer indicates number of perceptron's[1].

1.2.6 Support vector machine

For linear and nonlinear data Support Vector Machine is a classification approach which ascertain by a perceive hyperplane. Nonlinear mapping is used to changing the actual training data into a higher dimension. Within this new dimension, it searches for the linear optimal separating hyperplane. With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane. The SVM finds this hyperplane using support vectors and margins[1].

1.2.7 Logistic Regression

In this process to evaluate the output of the algorithm there can be many independent values. So, to implement this algorithm multiple number of outcomes will be needed. In order to search the most applicable model this algorithm can be very helpful. This algorithm mainly focuses on the quantitively process[1].

1.2.8 Stochastic Gradient Descent

The models which are linear for them the best choice is to apply the stochastic gradient descent algorithm. In case when the size of the dataset is very large this algorithm can come up with a

very efficient result. Different types of functions are used in this algorithm such as loss function. The algorithm also provides penalties. The algorithm may have some important parameters that needed to be pass through the function call. In order to scale the attributes this algorithm is one of the best choices[1].

1.3 Dependent and irrelevant attributes in high dimensional datasets

The attributes which have maximum similarity or dependency to other attributes are called dependent attributes. Irrelevant attributes are those which have less dependency on other attributes. Before applying any classification process, we should remove the irrelevant attributes as it degrades the performance of the model.

1.4 Dimension reduction algorithms

1.4.1 Missing value ratio

When we have too many missing values in the dataset, we need to set a threshold. The threshold refers to the termination of a variable. Which means if a particular variable crosses the threshold it will be automatically drop from the dataset[1].

1.4.2 Low variance filter

In a given dataset where all the observations of a variable have same value then the variable will have zero variance. Then we calculate the variance of other variables and then drop those variables which have less variance compared to others. The variables with low variance does nor effect the target value so it will not affect the model[1].

1.4.3 High correlation filter

High correlation between two variables means they have similar trends and likely to hold similar information. This can damage the performance of a model drastically .So, we calculate the correlation between independent numerical values and if the correlation crosses a certain threshold value it will be dropped from the dataset[1].

1.4.4 Backward feature elimination

In this process we first take n variables present in the dataset and train the model. Then we calculate the performance of the model. Then we compute the performance of the model after eliminating each variable. Then we try to find exclusion of which variables have less impact on the dataset. Then we drop those variables[1].

1.4.5 Forward feature elimination

In this process we train the model for each variable separately. The variable with the best performance is selected as starting variable. Then repeat the process and add one variable at a time until the variable that produces the highest performance is retained[1].

1.4.6 Principal component analysis

Principal component analysis is a technique which helps us to extract new set of variables from existing set of variables. These newly extracted components are called principal components. It is extracted in such a way that first principal component explains maximum variance in the dataset. This way second principle component tries to explain the remaining variance in the dataset which are uncorrelated to the first principle component[1].

1.4.7 Independent component analysis

To perform the analysis, we need some factors which are independent. It is assumed that the variables can be latent variable. Each of the variable will be different to each other [1].

1.4.8 Uniform Manifold approximation and manipulation

This algorithm can handle high dimensional dataset easily. It combines the power of visualization with the ability to reduce the dimensions of data. It preserves both local and global structures of data. UMAP maps nearby points on the manifold to nearby points in the low dimensional representations[1].

1.4.9 Methods based on projections

In projection technique multi-dimensional data is represented by projecting its point in low dimensional space. First we take a manifold that is closed to the data then project the data onto the manifold finally for representation we unfold the manifold[1].

1.5 Objective

Our objective is to study the effect of dimension reduction in classification of some selected high dimensional datasets.

1.6 Methodology

At first, we have loaded the dataset in the Jupiter notebook. Then we used backward elimination technique to apply dimension reduction in our dataset. We used the ordinary least square (OLS) model. First, we calculate the accuracy of the model without applying dimension reduction. Then in a loop we calculate the pvalue iteratively and the feature with the larger pvalue is dropped from the dataset then we put the features in the classification model using 10-fold validation. Pvalue is the performance metric which is used to measure the feature performance. If the Pvalue is greater than 0.5 then the particular feature will be extracted from the model. So, this way each time the pvalue of larger feature is dropped from the dataset. In this iterative approach there will be at least seven attributes after reducing dimensions. After reducing the dimension, the overall accuracy of the model has been improved. That means there is dependency in the dimensions of the dataset.

K-fold validation is a resampling procedure which is used to evaluate machine learning models on limited sample data. In train_test_split we train and test the data as of our choice. But here the case is not that. Here in this procedure a single parameter called k refers to the number of groups that a given data sample is split into. For example, when the value of k=10, it means the whole dataset will be divided into 10 experiments. So, the first experiment will be taken for the testing part and rest of the experiments will be taken as training part. Similarly, after working with the first experiment it will take the second experiment as testing part and remaining 9 experiments as training part. This is an iterative approach.

1.7 Contribution of the work

In order to enhance the accuracy score of a classification model we apply several techniques. Dimension reduction is one of the approaches we can apply in the classification model. In our paper we have used four datasets which are Diabetic retinopathy Debrecen dataset, SPECTF Heart Data Set, Online shoppers purchasing intention dataset And Cardiotocography dataset. We have applied six different classification algorithms to evaluate the model. As the datasets are high dimensional dataset so there are many irrelevant dimensions. Also, there are dependency between the dimensions. So, we applied dimension reduction by implementing backward feature elimination using ordinary least square (OLS) model. In Diabetic retinopathy Debrecen dataset multi-layer perceptron and Random forest give better accuracy. In SPECTF Heart Data Set multi-layer perceptron and Random forest give better accuracy. In Online shoppers purchasing intention dataset Random Forest gives better result than Multi-Layer Perceptron, Naïve Bayes, Decision Tree, K-Nearest Neighbor, and Support Vector Machine. In Cardiotocography dataset Random Forest gives better result than Multi-Layer Perceptron, Naïve Bayes, Decision Tree, K-Nearest Neighbor, and Support Vector Machine.

CHAPTER 2 LITERATURE REVIEW

In[3], a new dimension reduction algorithm was introduced named PSO-PCNN. This algorithm is based on feature selection. The input is passed to the pulse coupled neural network from the Particle Swarm Optimization. The name of the dataset is health care. The accuracy is 95% when it was compared to another optimization algorithm.

In[4], they discussed about the improvement of variational autoencoder (VAE). They also applied dimension reduction to the dataset. The name of the dataset is HDLSS dataset. In the experiment they used fourteen high-dimensional datasets. The dataset was taken from the Arizona state University repository1.

In[5], they came up with a new idea of semi-supervised dimension reduction. This is actually a hyperplane-based process. To make a good separating hyperplane features are extracted from the data. This is actually an optimization problem. They introduced an efficient algorithm to solve it. From the experimental result it can be said that the DRSVM algorithm gives better accuracy than the other algorithms. The accuracy difference between the L1-DRSVM and L2-DRSM is comparatively similar. While experimenting they also saw that the accuracy of DRSVM and SVD+SVM is kind of same.

In[6], they applied a new framework in order to reduce the dimension in the dataset. They worked with three different kinds of strategies. One is reducing the dimension based on selecting the features. Another one is liner reduction. The last strategy is non-linear reduction. From the experimental result it can be said that all of these strategies are very efficient. To Categorize the features the mutual information is used. In order to extract the variables from the dataset kernel linear method was used. The accuracy of the of the two algorithms have very less amount of difference to create the image. The total time needed to compute the accuracy of LKPCA is much less compared to PCA. The lengthy of 6the matrix y'*y is much shorter than y*y' in the data.

In[7], they defined a new combining method which includes dimension reduction and hashing both in order to retrieve the image properly. By applying the CNN algorithm, the features of the dimension are extracted. In order to remove the deep attributes of the dataset they used Neighbor embedding. In order to perform the process Sparse Projection was implemented. To apply the clustering in the dataset the reduce in the attributes was done. To visualize the process Two-dimensional scatterplot was used. Then they performed an empirical approach. Two human programmers performed the task. They pictured 816 scatter plots from 75 datasets and four-dimension extraction techniques.

In[8], their research topic is mainly based on bioinformatics. They used various kinds of datasets like the dataset of agriculture, the dataset of various genes and proteins. Their paper is actually based on comparison. In order to enhance the accuracy of the dataset the different type of dataset is used for the purpose.

In[9], they discussed about the effects of dimension reduction in the real life example. In order to reduce the attribute in each iteration they applied the most common dimension reduction technique which is Principle Component Analysis. The analysis is done by the main component of the dataset. In this process there is no chance of missing the data even though the analysis gives very good accuracy.

In[10], they used the common student datasets in order to build the model. Many common classification algorithms are used also some algorithms which are mainly based on filtering.

In[11], This process mainly focuses on the feature selection. They also analyzed the process of reducing every dimension. Each dimension has been categorized based on the attributes. The algorithm is implemented in such a way that the user does not have any idea about the information.

In[12], they actually defined an application for the user. They also analyzed the drawbacks of the algorithm. They applied The FeatureSelect based on some very basic common filtering algorithms. There are three types of users in the system which are mainly focuses on some optimized algorithms. To apply any kind of analysis into the system the system provides very simple interface. The system can handle all types of data. To calculate the accuracy score and F1 score the user gives overall better accuracy using FeatureSelect.

CHAPTER 3 EXPERIMENTAL RESULTS AND OBSERVATION

We have applied six different classification algorithms in this study. They are K-nearest neighbors, Decision tree, Naïve Bayes, Multi-layer perceptron, Random forest, Support vector machine. After reading the dataset we applied dimension reduction using ordinary least square (OLS) model. We have used four datasets which are Diabetic retinopathy Debrecen dataset, SPECTF Heart Data Set, Online shoppers purchasing intention dataset And Cardiotocography dataset[13].

3.1 Diabetic retinopathy Debrecen dataset

Description of the Dataset

This dataset contains features extracted from the Messidor image set to predict whether an image contains signs of diabetic retinopathy or not. All features represent either a detected lesion, a descriptive feature of an anatomical part or an image level descriptor. . The dataset has 20 attributes where 19 are input features and the remaining one is the output feature.

The experimental result is shown in the figure 3.1

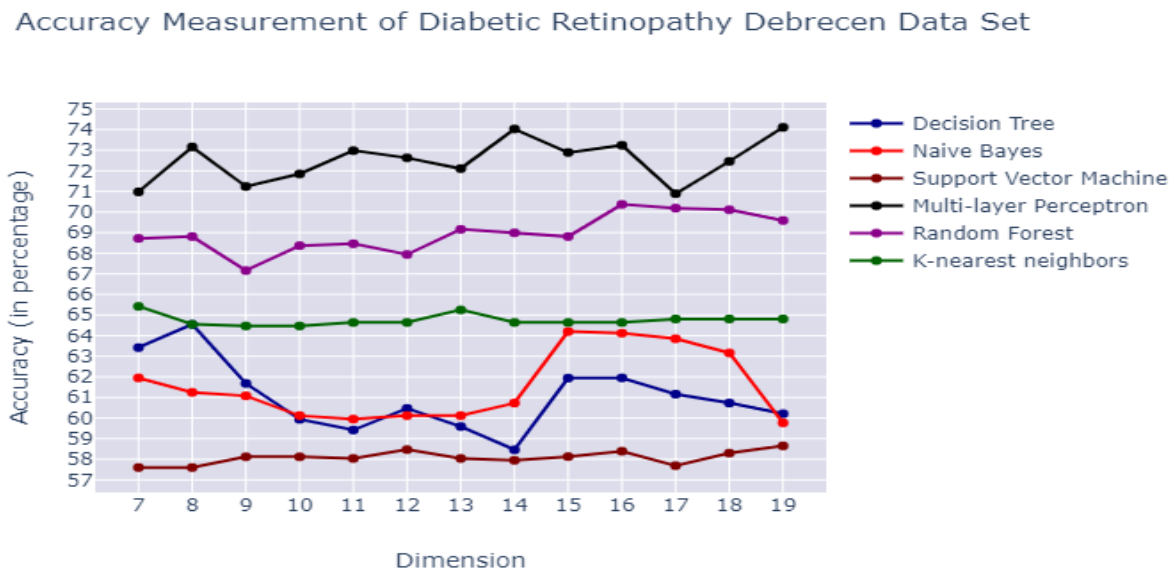

Figure 3. 1 Accuracy measurement of Diabetic Retinopathy Debrecen Data Set

From the experimental result we have found following observation.

1. Multi-layer perceptron and Random forest give overall better accuracy than Decision tree, Naïve Bayes, K-nearest neighbors and Support vector machines.
2. Multi-layer perceptron, Random forest, Decision tree, K-nearest neighbors, Naïve Bayes produce better result when dimension is reduced to 8 or from original dimension of 19. This implies that the dataset has many dependent attributes.

3.2 SPECTF Heart Dataset

Description of the Dataset

Data on cardiac single proton emission computed Tomography (SPECTF) images. Each patient classified into two categories normal and abnormal. The dataset has 45 attributes where 44 are input features and the remaining one is the output feature.

The experimental result is shown in the figure 3.2

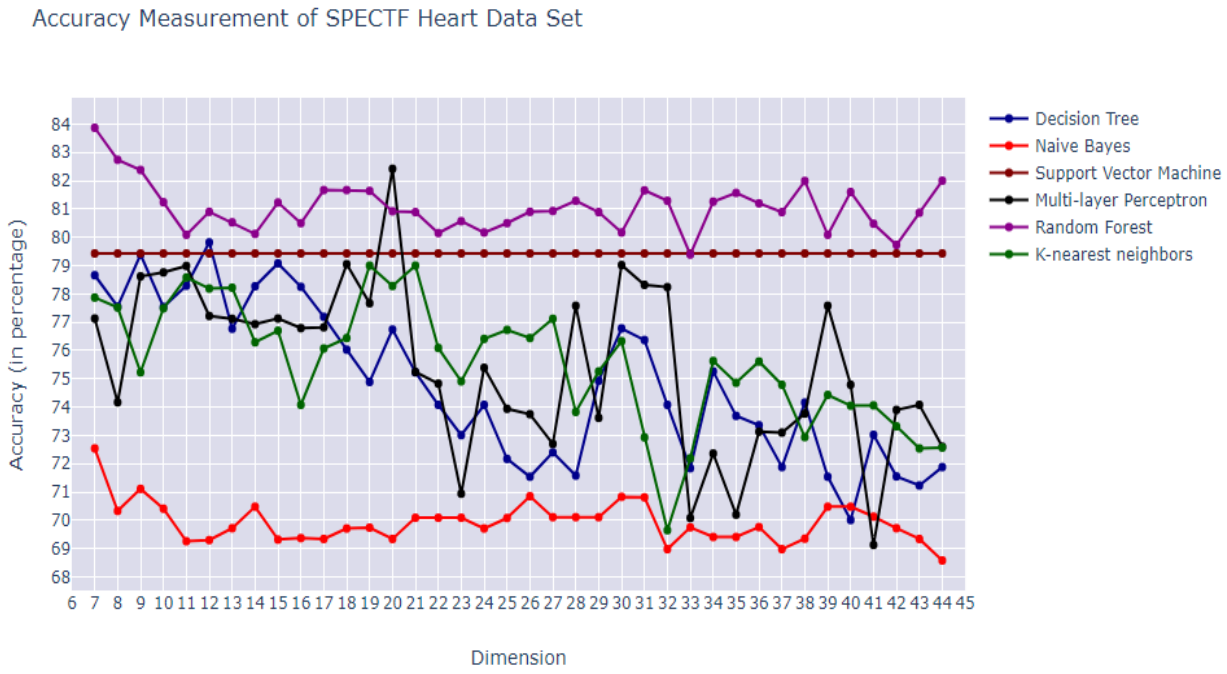

Figure 3. 2 Accuracy measurement of SPECTF Heart Data Set

From the experimental result we have found following observation.

1. Random Forest and Multi-Layer Perceptron give overall better accuracy than Naïve Bayes, Decision Tree, K-nearest neighbors, Support Vector Machine.
2. Random Forest gives best result for dimension 7.
3. Multi-Layer Perceptron gives best result for dimension 20.
4. Decision Tree gives best result for dimension 12.
   Above observations show that dimensionality reduction to 7 to 20 from original dimension of 44 gives better result. Therefore, the dataset has many dependent attributes.
5. Except Support Vector Machine, accuracies of all other classification techniques widely vary with the reduction of dimensions. The reason of this behavior is that this dataset has 44 attributes and many of them are dependent and irrelevant.

3.3 Online shoppers purchasing intention dataset

Description of the Dataset

The dataset consists of feature vectors belonging to 12,330 sessions. The dataset was formed so that each session would belong to different user in a 1-year to period to avoid any tendency to a specific campaign, special day, user profile or period. . The dataset has 18 attributes where 17 are input features and the remaining one is the output feature.

The experimental result is shown in the figure 3.3

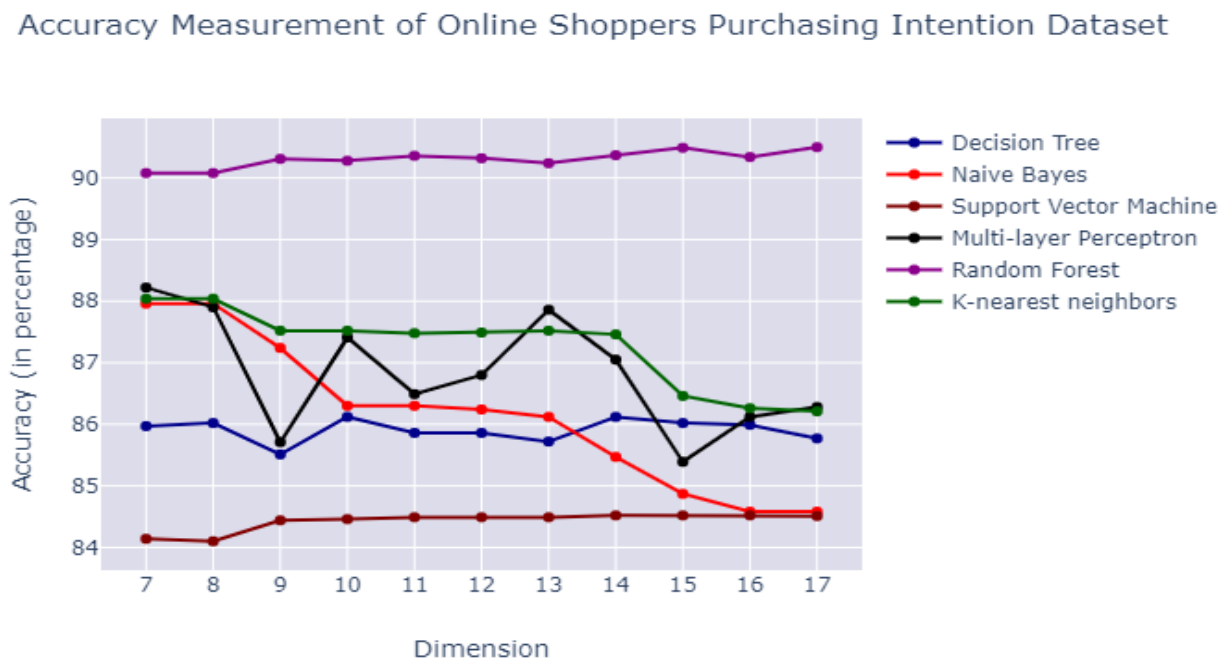

Figure 3. 3 Accuracy measurement of Online Shoppers Purchasing Intention Data Set

From the experimental result we have found following observation.

1. Random forest gives better result than Multi-layer perceptron, Naïve Bayes, Support Vector Machine, K-nearest neighbors, Decision Tree. But this is invariant to dimension reduction.
2. The other methods have some performance improvement for dimension reduction. But their overall accuracy is much smaller than Random Forest.

3. From this observation it can be said that the dataset does not have significant number of dependent attributes.

3.4 Cardiotocography dataset

Description of the dataset

2126 fetal cardiotocograms (CTGs) were automatically processed and the respective diagnostic features measured. The (CTGs) were also classified by three expert obstetricians and a consensuses classification label assigned to each of them. The dataset has 22 attributes where 21 are input features and the remaining one is the output feature.

The experimental result is shown in the figure 3.4

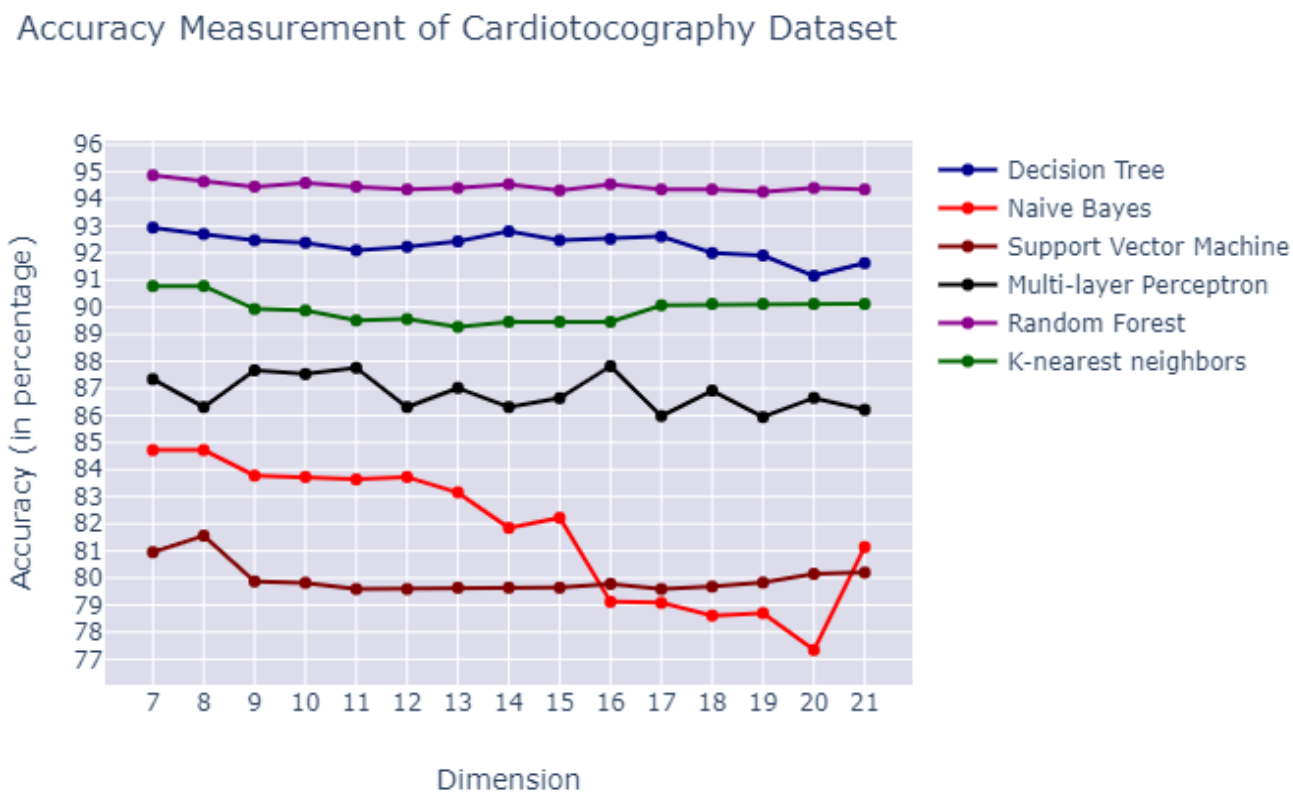

Figure 3. 4 Accuracy measurement of Cardiotocography Data Set

From the experimental result we have found following observation.

1. Random Forest gives better result than Multi-Layer Perceptron, Naïve Bayes, Decision Tree, K-Nearest Neighbor, and Support Vector Machine. But this is invariant to dimension reduction.

2. Naïve Bayes, produces better result when number of dimensions are reduced. It produces the best result for attribute reduction to 7 from 21. However, it performs worst with 20 attributes. For original data, it is better than Support Vector Machine.

CHAPTER 4 CONCLUSION AND FUTURE WORKS

Among the experimented datasets, the dataset with highest number of attributes shows higher accuracy for dimension reduction. It can be interpreted as that high dimensional dataset has many dependent attributes. This sort of high dimensional dataset practically has many irrelevant attributes. We have used backward elimination to reduce the dimension using Ordinary least square model. We have reduced the data dimension from a particular number n (n= number of dimensions) to 7. In each iteration we have excluded the less significant dimension based on the pvalue. Finally, we have come up with an idea that for high dimensional datasets dimension reduction is a very useful technique. After applying dimension reduction, we have found significant improvement in the accuracy score of some classification algorithm like Multi-layer Perceptron, Naïve Bayes and Random forest. Dimension reduction of this sort of datasets produce better performance in classification problem.

Our future work is the study of statistical nature of datasets and their influences on the dimension reduction. The influence of dimension reduction on the performance of data classification may also be investigated.

REFERENCES

[1] S. Agarwal, *Data mining: Data mining concepts and techniques*. 2014.

[2] "K nearest neighbor." [Online]. Available: https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#/media/File:KnnClassification.svg.

[3] S. Rajeswari, M. S. Josephine, and V. Jeyabalaraja, "Dimension reduction: A PSO-PCNN optimization approach for attribute selection in high-dimensional medical database," *IEEE Int. Conf. Power, Control. Signals Instrum. Eng. ICPCSI 2017*, pp. 2306–2309, 2018.

[4] M. S. Mahmud and X. Fu, "Unsupervised classification of high-dimension and low-sample data with variational autoencoder based dimensionality reduction," *2019 4th IEEE Int. Conf. Adv. Robot. Mechatronics, ICARM 2019*, pp. 498–503, 2019.

[5] H. Fang, M. Cheng, and C. J. Hsieh, "A hyperplane-based algorithm for semi-supervised dimension reduction," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, vol. 2017-Novem, pp. 101–110, 2017.

[6] A. Anaissi, P. J. Kennedy, and M. Goyal, "A framework for high dimensional data reduction in the microarray domain," *Proc. 2010 IEEE 5th Int. Conf. Bio-Inspired Comput. Theor. Appl. BIC-TA 2010*, pp. 903–907, 2010.

[7] Y. Li, Z. Miao, Y. Xu, H. Li, and Y. Zhang, "Combining Nonlinear Dimension Reduction and Hashing Method for Efficient Image Retrieval," *Proc. - 2016 12th Int. Conf. Semant. Knowl. Grids, SKG 2016*, pp. 126–130, 2017.

[8] M. Sedlmair, T. Munzner, and M. Tory, "Empirical guidance on scatterplot and dimension reduction technique choices," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, pp. 2634–2643, 2013.

[9] G. . Ramadevi and K. Usharani, "Study on Dimensionality Reduction Techniques and Applications," *Publ. Probl. Appl. Eng. Res.*, vol. 04, no. 1, pp. 134–140, 2013.

[10] M. Zaffar, M. A. Hashmani, K. S. Savita, and S. S. H. Rizvi, "A study of feature selection algorithms for predicting students academic performance," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 5, pp. 541–549, 2018.

[11] V. Kumar, “Feature Selection: A literature Review,” *Smart Comput. Rev.*, vol. 4, no. 3, 2014.

[12] Y. Masoudi-Sobhanzadeh, H. Motieghader, and A. Masoudi-Nejad, “FeatureSelect: A software for feature selection based on machine learning approaches,” *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–17, 2019.

[13] “UCI Machine Learning.” [Online]. Available: https://archive.ics.uci.edu/ml/index.php.