IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Volume 2, PP 81-84

Exploratory Data Analysis Using Dimension Reduction

Tejas Nanaware[1], Prashant Mahajan[2], Ravi Chandak[3], Pratik Deshpande[4], Prof. Mahendra Patil[5]

[1]*(Department of Computer Engineering, Atharva College of Engineering, Malad West, Mumbai, India)*
[2]*(Department of Computer Engineering, Atharva College of Engineering, Malad West, Mumbai, India)*
[3]*(Department of Computer Engineering, Atharva College of Engineering, Malad West, Mumbai, India)*
[4]*(Department of Computer Engineering, Atharva College of Engineering, Malad West, Mumbai, India)*
[5]*(HOD, Department of Computer Engineering, Atharva College of Engineering, Malad West, Mumbai, India)*

Abstract : *Exploratory Data Analysis (EDA) is a complex process having multiple steps. The motive of this paper is to explain how the method of dimensionality reduction is used to perform EDA. This paper outlines techniques used for EDA as well as approaches adopted to achieve dimensionality reduction. Particularly, two approaches are explained with the help visual representation of the data, namely Principal Component Analysis (PCA) and t- Distributed Stochastic Neighbor Embedding (t-SNE). We explain how high-dimensional data set can be visualized into the lower dimensions, retaining the implicit structure of the data. The visualizations produced by the two methods are compared to discuss the pros and cons of each method pertaining to our system.*

Keywords – *Data Analysis, Data Visualization, Dimensionality Reduction, Machine Learning, PCA, TSNE.*

I. INTRODUCTION

Exploratory Data Analysis (EDA) is an approach for data analysis that utilizes variety of techniques to summarize main characteristics of the data set, often with visual methods. EDA is useful for a range of purposes such as: Maximizing insights into a data set, mapping out underlying structure of the data, identifying useful variables, detecting outliers and anomalies and Testing a hypothesis. EDA is about getting to know and understanding your data before making any assumptions about it. Different techniques used for analysis of the data are outlined below:

1) Clustering and dimension reduction: Creates graphical displays of high-dimensional data with many variables.
2) Univariate visualization: Method of looking at a one variable of interest
3) Multivariate visualizations: Analysis of multiple variables at the same time
4) K-Means clustering
5) Predictive Models

A. DIMENSIONALITY REDUCTION

It is the process of reducing a high dimensional dataset to lower-dimensional representation that keeps most of its important structure [1]. If the data lies in a high dimensional space, then large amount of data is required to learn and teach a model. Also, some problems become unmanageable as the number of variables increases, giving rise to requirement of huge amount of training data and too many model parameters which in turn increases the complexity. For such scenarios dimensionality reduction comes handy as it aims at automatically finding lower-dimensional representation by reducing the effect of unimportant attributes. There are various methods used to perform dimensionality reduction. Some of them are explained here:

1. Low variance – In case of high number of dimensions, we should filter out the variables having low variance compared to others because these variables will not contribute to explaining variance in target variables.
2. Decision Trees – This can be used to tackle multiple problems like missing values, identifying significant variables and outliers.
3. Random Forest – It's inbuilt feature importance can be used to select a smaller subset of input features. Only drawback random forests is they favour numeric variables over binary variables.
4. High Correlation – Parameters with higher correlation can bring down the performance of the model. Selection of one of the variables in high correlational variables can be done using Variance Inflation Factor (VIF) where variables having VIF>5 can be dropped.
5. Factor Analysis – This method is extracts maximum common variance amongst all variables and put them into a common score. There are two ways to perform factor analysis – Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA).

6. Principal Component Analysis (PCA) – This technique performs a linear mapping of the data to lower dimensional space such that variance of the data is maximized.

Visualisation of high dimensional data is important problem across various domains which deals with the data of widely varying dimensionality. This paper highlights two methods to perform data analysis Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbour Embedding (t-SNE). t-SNE is a method of visualizing high-dimensional data by giving each data point a location in a two or three-dimensional space [6].

II. LITERATURE SURVEY

Dimension reduction for exploratory data analysis is being done is various modelling systems since several decades. Christopher J. C. Burges defines the process of dimension reduction where uninformative variance is discarded and compares various techniques. The main limitation of PCA is that it ignores correlations that are higher second order. He also provides a cautionary note stating that the behavior of higher dimensional data is different from lower dimensional data and also provides a simple approach that is find low dimensional projections that extract useful information from the data by maximizing a suitable objective function [4]. Matthew Brehmer et. al. focuses on characterization of five task sequences related to the visualization of dimensionally-reduced data: Synthesized dimensions, mapping a synthesized dimension to original dimensions, verifying clusters, naming clusters, matching clusters. Their abstract characterization of these task sequences fills a gap between the large body of technique-driven literature and analysts' domain problems [1]. Nandakishore Kambhatla and Todd K. Leen provided a local linear approach to dimension reduction which is fast to compute and provides accurate representations even though nonlinear dimension reduction techniques out-performs linear ones. They provided with a full implementation of transform coding, with comparisons between PCA, auto-associations, and VQPCA in terms of rate distortion curves [2].

Brehmer et. al. discusses best practices in exploratory factor analysis and provide recommendations for getting the most from your analysis. According to their research, while principal component analysis and Kaiser criterion are normally used, they are not optimal, especially when data do not meet assumptions. Factor analysis is preferable to principal component analysis to obtain optimal results [5]. Furthermore, the research on t-SNE is done by Maaten et. al. where they proposed that t-SNE is better than existing techniques at creating a simple map that reveals structure at many different scales. t-SNE is capable of capturing much of the structuring of high dimensional data, simultaneously revealing global structure such as the presence of clusters at several scales. They also discuss how cost function of t-SNE differs from the one used by SNE and potential weaknesses of this method [6]. Alexander Schulz et. al. states that non-parametric dimension reduction techniques like t-SNE leads to a powerful and flexible visualization of high dimensional data. A comprehensive comparison between PCA, MDS, SOM and t-SNE is given along with the cons and pros of all. The main issue of dimension reduction is that we have to take information loss into account and that there does not exist a direct method to map additional points. PCA assumes centered data and tries to find linear projections of the points by maximizing variance [3].

Based on the above research, we performed the data analysis using PCA and t-SNE. The graphical visualization of the data by using these two different method of analysis is also shown which can be further used for feature extraction process.

III. OUR SYSTEM

This system tackles the issue of data analysis by dimension reduction. When the system has many dimensions, the issue occurs when they are to be displayed graphically. In such a case, the dimensions are often averaged and that results in loss of accuracy. When a system has say two parameters, it can be represented with two-dimension graph, similarly, if a system has three parameters, it can be represented with three-dimension graph. This becomes a critical issue when there are more than 10 parameters as then the axes often have averaged the data to fit in a three-dimensional graph.

In our system, we have to analyse the student's data for predicting their campus placements and the package that they will receive. The data consists of the various skills that a student possesses as shown in Table 1. The data consists in the range between 0 to 100 as the percentage of the quality of the skills. The probability of being placed also exists as a range between 0 to 100 as probabilistic reasoning.

Table 1: Database Parameters

Coding Skills	Aptitude Skills	Technical Skills	Communication Skills	Core Knowledge	Presentation Skills	Academic Performance	Puzzle Solving skills
85	85	90	30	40	40	75	75
English Proficiency	Programming Skills	Management Skills	Projects	Internships	Training	Backlog	Placed
50	90	40	80	80	85	0	52

In order to express this database in a three-dimensional graph, we have to combine and average the student's skills as qualitative and quantitative parameters as X - axis and Y - axis and plot the probability of the student being placed on the Z - axis. This three-dimensional plot enables the user to read the dataset very clearly and is able to understand the situation of placement in the college. This plot is shown in Figure 1.

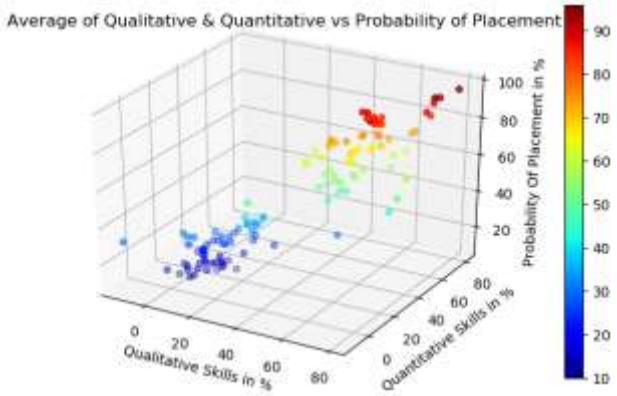

Figure 1: Combine and average the student's skills

On the other hand, the PCA and t-SNE algorithms performs better as compared to the averaging method as the data loss is significant. PCA provides a fixed plot whereas t-SNE gives a different plot every time. PCA and t-SNE preserves the Euclidean distance between the points which enables the prediction algorithm to perform more efficiently. On performing PCA and t-SNE plot, we achieve preserved points and moreover, we are able to plot the data in a simple two-dimensional graph. Figure 2 and Figure 3 shows the plot of PCA and t-SNE respectively.

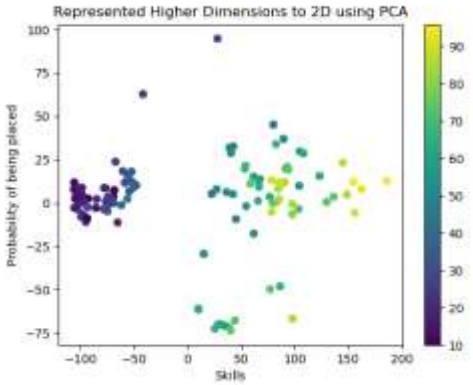

Figure 2: Dimensions reduced using PCA

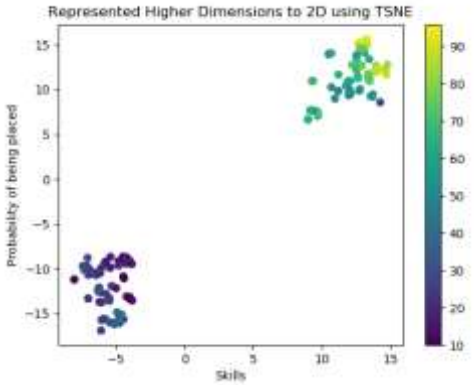

Figure 3: Dimensions reduced using t-SNE

It has been observed that the plot of t-SNE occurs different at each run and thus you get a different plot at each run. Moreover, PCA works better than t-SNE if the number of dimensions to reduce are less. Hence, while reducing four dimensions into a three-dimensional space, t-SNE will not infer the mappings properly thus in such a case PCA works better by simply keeping the top two dimensions.

IV. Conclusion

Exploratory Data Analysis (EDA) is a complex, multi-step process. We learnt about various methods of dimensionality reduction and its two particular algorithms: Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE). Visualization of the graphs help us to know about how higher dimension data can be visualized to lower dimensions. It has been observed that t-SNE works better than PCA if the number of dimensions are very high. Furthermore, the graph of t-SNE will change during every run as t-SNE transforms the samples into different spaces that preserves distances between them and not the value of the data sample.

Acknowledgements

The authors thank Prof. Mahendra Patil for his support and guidance that has deeply influenced the quality of this manuscript.

REFERENCES

[1]. Brehmer, M., Sedlmair, M., Ingram, S., & Munzner, T. (2014, November). Visualizing dimensionally-reduced data: Interviews with analysts and a characterization of task sequences. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*(pp. 1-8). ACM.

[2]. Kambhatla, N., & Leen, T. K. (1997). Dimension reduction by local principal component analysis. *Neural computation*, *9*(7), 1493-1516.

[3]. Gisbrecht, A., Schulz, A., & Hammer, B. (2015). Parametric nonlinear dimensionality reduction using kernel t-SNE. *Neurocomputing*, *147*, 71-82.

[4]. Burges, C. J. (2010). Dimension reduction: A guided tour. *Foundations and Trends® in Machine Learning*, *2*(4), 275-365.

[5]. Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research & evaluation*, *10*(7), 1-9.

[6]. Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*(Nov), 2579-2605.