

# Data Dimensionality Reduction Techniques : Review

Dr.K Bhargavi

TKR Engineering college, Hyderabad, Telangana

bhargavi.mtech@gmail.com

**Abstract**— Data science is the study of data. It involves developing methods of recording, storing, and analyzing data to effectively extract useful information. The goal of data science is to gain insights and knowledge from any type of data — both structured and unstructured.

Data science is related to computer science, but is a separate field. Computer science involves creating programs and algorithms to record and process data, while data science covers any type of data analysis, which may or may not use computers. Data science is more closely related to the mathematics field of Statistics, which includes the collection, organization, analysis, and presentation of data.

Because of the large amounts of data modern companies and organizations maintain, data science has become an integral part of IT. For example, a company that has petabytes of user data may use data science to develop effective ways to store, manage, and analyze the data. The company may use the scientific method to run tests and extract results that can provide meaningful insights about their users.

**Keywords**—data science, data, machine learning algorithms, reduction techniques, storage.

## I. INTRODUCTION

Data Science is a more forward-looking approach, an exploratory way with the focus on analyzing the past or current data and predicting the future outcomes with the aim of making informed decisions. It answers the open-ended questions as to “what” and “how” events occur.

Features	Data Science
Data Sources	Both Structured and Unstructured ( logs, cloud data, SQL, NoSQL, text)
Approach	Statistics, Machine Learning, Graph Analysis, Neuro- linguistic Programming (NLP)

Focus	Present and Future
Tools	RapidMiner, BigML, Weka, R

Table 1: Features of Data Science

A common mistake made in Data Science projects is rushing into data collection and analysis, without understanding the requirements or even framing the business problem properly. Therefore, it is very important for you to follow all the phases throughout the lifecycle of Data Science to ensure the smooth functioning of the project.

Here is a brief overview of the main phases of the Data Science Lifecycle:

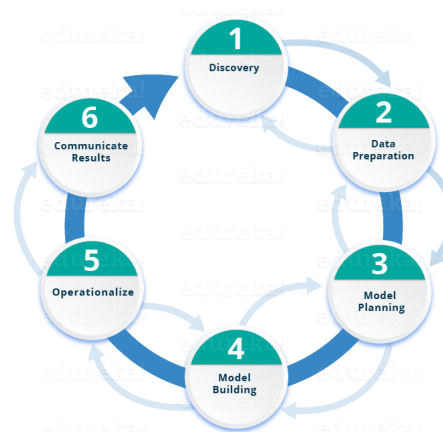


Fig 1: Life Cycle of Data Science

Figure 2 shows a comprehensive reference architecture consisting of an importer, an exporter, a data storage and access layer, a text mining engine, and a user interface. Based on the reference architecture we developed a collaborative web application with a Java back-end. As web application framework Play3 was used and

additionally the search engine Elasticsearch for efficient access to the textual data has been integrated.

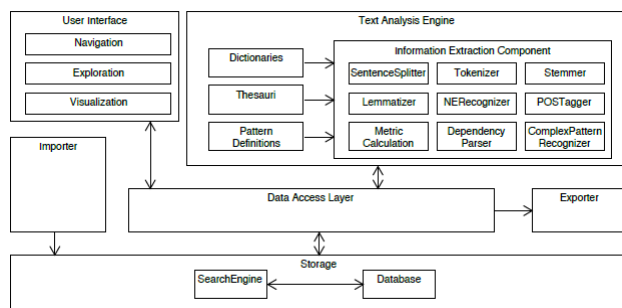


Fig 2: Legal data science reference architecture for collaborative environments

## II. DATA DIMENSIONALITY REDUCTION TECHNIQUES

There are seven different methods that have been applied in the data analytical space. They are illustrated in the Figure1. In most of the times the huge amount of data in data analytical process does not work well. So dimensionality reduction is applied to the large data. The methods of dimensionality reduction define about the reduction of data elements.

**A. Missing values:** The data column with enormous amount of missing values is calculated and then they are removed. The numbers of missing values are calculated using either statistical node or group node. The statistical nodes are used for the analysis of data and the group node is used in the techniques of DR. The missing values larger than the threshold are collected and then they are removed. If the threshold value found is larger than the original values then the reduction will be aggressive. The missing values are calculated using the formula which is shown in Figure.2. In KNIME tool the missing value node calculates the ratio of the missing values is calculated by number of missing values divided by the total number of rows.

**B. Low Variance Filter:** The variance of the data is calculated to find the number of information about the data column. In the limit case where the column cells assume a constant value, the variance would be 0 and the column would be of no help in the discrimination of different groups of data [5]. The data column lower than the threshold values of the data are filtered and then they are removed. The filter is applied to the data

to find the low variance. First normalization is done to all the columns to find the lower variance.

**C. High Correlation Filter:** The feature which are given as input are often correlated i.e. the features are dependent to one another and they too have same information. A data column with values highly correlated to those of another data column is not going to add very much new information to the existing pool of input features [5]. To reduce the correlated column, the correlation is measured using the linear correlation node between couple of columns. If any highly correlated column is present between the pair the particular data is removed. Correlation matrix is taken as input to the correlation filter. Filtering highly correlated data columns requires uniform data ranges again which can be obtained with a Normalize node [5].

**D. Ensemble Trees:** Ensemble trees are also known as random forest. For effective classification feature, selection is done. One approach to dimensionality reduction is to generate a large and carefully constructed set of trees against a target attribute and then use each attribute's usage statistics to find the most informative subset of features [5]. The score is calculated using level of the candidates and it defines about the relative attributes which are most predictive. A shallow tree has been generated for ensembling and here each tree is trained on the fraction of all attributes. If the particular attribute is selected as the best split one then the informative features are retained.

**E. Backward Feature Elimination:** The Backward Feature Elimination loop performs dimensionality reduction against a particular machine learning algorithm [5]. It is a simple iterative method and in each method the selected classification algorithm is performed for number of input features. Then one input feature is removed and the model is trained for (n-1) input features for number of times. The input feature with large number of error rate, after the removal of feature is recognized and they are eliminated. Then they are repeated for number of iterations like n-2, n-3, etc to find the larger error rate. The Backward Feature Elimination Filter finally visualizes the number of features that are kept at each iteration and the corresponding error rate [5]. This method can only be applied to the small number of dataset.



**F. Forward Feature:** Elimination Similarly to the Backward Feature Elimination approach, a Forward Feature Construction loop builds a number of pre-selected classifiers using an incremental number of input features [5]. The forward feature loop opens with one feature and other feature is added to it for every iterations. Both forward and backward are very expensive and computationally high. Running the optimization loop the best cutoffs in terms of lowest number of columns and best accuracy were determined for each one of the six dimensionality reduction methods and for the best performing model [6].

There are two different major techniques used in dimensionality reduction. They are feature selection and feature extraction .

**A. Feature Selection:** In machine learning and statistics, feature selection (FS), also known as variable selection or attribute selection or variable subset selection and it is the process of selecting a subset of relevant features (variables, predictors) for use in model construction [7]. This technique has three different reasons: They are Interpretation of model to make them efficient and simple for the users, Less training time and Reduction of variance for strengthened generalization. Feature selection is applied to the domains with larger features and chooses the feature according to the objective function. A feature selection algorithm can be seen as the combination of a search technique for proposing new feature subsets along with an evaluation measure which scores the different feature subsets [7]. Every subset of the feature is tested separately to minimize the error or noise rate. Feature selection is classified into three different classes: they are embedded, filter and wrapper method.

**B. Feature Reduction:** In feature reduction, the data of high dimensional space are transformed into lower dimensional space. Transforming the data is done in linear method or non linear method. The main linear technique for dimensionality reduction performs a linear mapping of the data to a lower-dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized. In order to reduce the features the correlation coefficient is calculated by finding the Eigen vectors and Eigen values. In this model

the relevant features for the class is selected by eliminating the redundant features.

### III. CONCLUSION

The main objective of this paper is to provide the overview of the dimensionality reduction. Before performing the reduction the estimation should be made for dimensions. Dimensionality reduction has been used in numerous fields for further techniques. It is used for providing the better result for the data analysis. The guidance should be given to the users handling dimensionality reduction in their data analytics. This paper discussed about the dimensionality reduction methods, techniques used, algorithms, fields where the reduction has been used. Further work is done by performing various dimensionality reduction algorithms and the techniques like clustering, classification, etc can be performed for comparative analysis.

### References

- The heading of the References section must not be numbered. All reference items must be in 8 pt font. Please use Regular and Italic styles to distinguish different fields as shown in the References section. Number the reference items consecutively in square brackets (e.g. [1]).
- [1] . Laurens van der Maaten "An Introduction to Dimensionality Reduction Using Matlab", MICC, Maastricht University.
  - [2]. G.N.Ramadevi and K.Usharani- "Study On Dimensionality Reduction Techniques And Applications" - 2- Vol 04, Special Issue01; 2013 -Publications Of Problems & Application In Engineering Research – Paper
  - [3]. Christopher J. C. Burges "Dimension Reduction: A Guided Tour- Foundations and Trends<sub>R</sub> in Machine Learning" Vol. 2, No. 4 (2009) 275–365 \_c 2010 DOI: 10.1561/2200000002.
  - [4] . C.O.S. Sorzan, J.Vargas and A. Pascual Montano "A survey of dimensionality reduction techniques"
  - [5]. Seven Techniques for Dimensionality Reduction-Open for innovative KNIME.
  - [6]. Thomas Navin Lal1, Olivier Chapelle, Jason Weston, and André Elisseeff - Learning with Local and Global Consistency-Max Planck Institute for Biological Cybernetics, Tübingen, Germany.
  - [7] . Mahesh, Bhasutkar, Maninti Venkateswarlu, and M. Raghavendra. "End-to-end congestion control techniques for Router." 2011 International Conference on Communication Systems and Network Technologies. IEEE, 2011.
  - [8]. Mahesh, B., and K. Shyam Sunder Reddy. "Router Aided Congestion Control Techniques." Second International Conference on Information Systems and Technology.
  - [9]. Mahesh, B. "Dynamic Update and Public Auditing with Dispute Arbitration for Cloud Data." Journal of Advanced Database Management & Systems 4.3 (2017): 14-19.



- [10]. Mahesh, B., et al. "A Review on Data Deduplication Techniques in Cloud." *Embedded Systems and Artificial Intelligence*. Springer, Singapore, 2020. 825-833.