Visualization of Mixed Attributed High-Dimensional Dataset Using Singular Value Decomposition

Bindiya M. Varghese(✉), A. Unnikrishnan, and K. Poulose Jacob

Rajagiri College of Social Sciences, Kalamassery, India
bindiya@rajagiri.edu

Abstract. The ability to present data or information in a pictorial format makes data visualization, one of the major requirement in all data mining efforts. A thorough study of techniques, which presents visualization, it was observed that many of the described techniques are dependent on data and the visualization needs support specific to domain. On contrary, the methods based on Eigen decomposition, for elements in a higher dimensional space give meaningful depiction. The illustration of the mixed attribute data and categorical data finally signifies the data set a point in higher dimensional space, the methods of singular value decomposition were applied for demonstration in reduced dimensions (2 and 3). The data set is then projected to lower dimensions, using the prominent singular values. The proposed methods are tested with datasets from UCI Repository and compared.

Keywords: Data visualization · Mixed attribute datasets · Dimensionality reduction · SVD

1 Introduction

Visualization implies presenting data in a pictorial form. Kim Bartke states that data visualization as the plotting of data into a Cartesian space. It helps the user to have a better insight into the data. Data visualization is graphical presentation of a dataset, which provides data analysts a quality understanding of the information contents in way that is more comprehensible. The spatial representation for high dimensional data will be very handy for envisaging the relationship between the attributes.

2 Various Approaches

Sándor Kromesch et al. as geometric methods, icon-based methods, pixel-oriented techniques, hierarchical techniques, graph-based methods, and hybrid class classify the most popular visualization techniques. Geometric projection visualization techniques map the attributes to a Cartesian plane like scatter plot, or to an arbitrary space such as parallel coordinates. A matrix of scatter plots is an array of scatter plots displaying all possible pair wise groupings of dimensions or coordinates. For n-dimensional data this

J.J. Jung and P. Kim (Eds.): BDTA 2016, LNICST 194, pp. 3–11, 2017.
DOI: 10.1007/978-3-319-58967-1_1

produces n(n − 1)/2 scatter plots with shared scales, although most often n^2 scatter plots are shown. A survey plot is a technique, which helps to extract the correlations between any two attributes of the dataset, mainly when the data is organized according to a particular dimension. Each horizontal splice in a plot relates to a particular data instance [1]. Parallel coordinate technique is a tool for envisaging multivariate data. This visualization technique maps the multi-dimensional element on to a number of axes, which are parallel. In pixel-oriented techniques the pixel representing an attribute value is colored based on the value in its domain. Recursive pattern methods orders the pixels in small clusters and arranges the clusters to form some global design [2]. The above-mentioned techniques are used to illustrate the Iris Data (UCI Repository [6]) in Figs. 1, 2, 3 and 4.

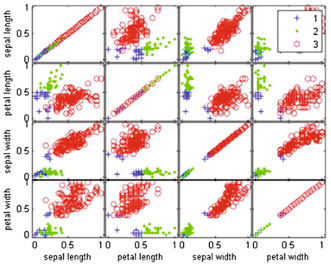

Fig. 1. Scatter matrix plot of Iris Data

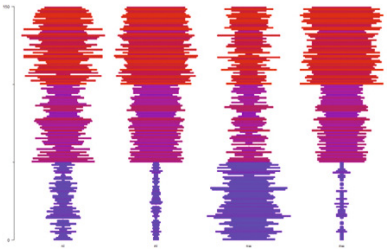

Fig. 2. Survey plot of Iris Data

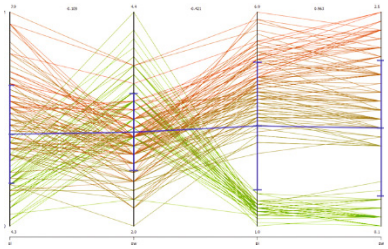

Fig. 3. Parallel coordinates of Iris Data.

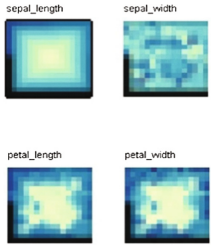

Fig. 4. Pixel oriented display of Iris Data.

3 Framework for Visualizing Mixed Attributed High-Dimensional Data

This research paper proposes a general framework to visualize a high-dimensional mixed dataset in two phases. To make it equipped for any general clustering algorithm, the mixed dataset is transformed into a uniform format. The phase includes an extension to completely categorical dataset where a frequency of the occurrence of data is made use of. A gridded representation of lower dimensional data, which was, normalized in a uniform format, is generated in the second phase. An intermediate phase requires the dimensions to be reduced.

3.1 Dataset Preprocessing

A machine-learning scheme receives an input or a set of instances, which are to be mined based on any classification, association rules or clustering techniques. The input dataset are described by the values of a set of fixed attributes after a detailed elimination of unwanted variables. To prepare this dataset adequate for data mining approaches, these data are examined for its basic data types and features. The source data is normally categorized as structured data, semi-structured or unstructured data. The structured data can be broadly classified into following types; Numeric, categorical, ordinal, nominal, ratio and interval. If dataset comprises of mixed attributes, i.e. a combination of numerical and categorical variables, then the usual approach for pre-processing is to distinctly cater different data types. Many of the clustering algorithms acts well with numeric data. The framework researched in this study starts with the various techniques to convert the categorical attributes to a numerical equivalent.

The type or magnitude of variables are not distinguished by many of the mining algorithms. The results will be affected the dominance on one particular variable over others, the algorithm treats all variables equally. Normalizing the data will eliminate this preference, thereby bringing the parameters of different units and scales into a similar plane. One of the most commonly used normalization techniques is Min-Max normalization, which performs a linear transformation on the raw data. Min-Max normalization maps a value v to v' in the range [new_min$_A$, new_max$_A$] by computing

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A \tag{1}$$

where min_A and max_A are the minimum and maximum value of the attribute respectively. In z-score normalization, the attribute values are normalized based on the mean and the standard deviation of values. A value v is normalized to v' by calculating

$$v' = \frac{v - \mu}{\sigma}, \tag{2}$$

where μ and σ are the mean and the standard deviation of the values respectively.

3.2 Conversion of a Mixed Dataset into a Uniform Format

Pure numeric data or on pure categorical data is apt for many of the mining algorithms. The mixed dataset in its raw form is not suitable for applying the algorithms as such. However, the real world data contains multiple types of attributes. In order to apply a data-mining algorithm, either it is required to convert the complete database to categorical or numerical in type. The proposed work is based on computing the dissimilarity computed with the co-occurrence matrices [3].

The dataset containing categorical variables is inputted as a n * p matrix, where n is the number of instances and p the number of categorical attributes. All combinations generated by p variables can be a part of the sample space. Traditional practice is to convert categorical values into a named set, for example, like low, medium and high to

numeric scale. Nevertheless, certain categorical variables cannot be ordered logically like geographical data or a weapon data in crime dataset. Exploration of the relationship among the attributes may eventually help in converting the mixed set into a numerical dataset. The idea of co-occurrence is considered as a foundation in this study to find the similarities between the categorical variables.

Co-occurrence indicates that when two objects shows up frequently together, and then there is always a possibility of strong similarity between them. Closer numeric values can be assigned to those values, which is co-occurred in the dataset.

The process of pre-processing starts with the normalization of the data using any of the normalization techniques explained in the previous section. Normalization is required to eliminate the dominance of one attribute over other because of its domain size. A base attribute is selected from all categorical attribute. The selection is based on the criteria that the base item must have maximum variation of elements in the domain. The elements present in the observations of base attribute can be termed as base items. Construction of a co-occurrence matrix M of size $n * n$, where n is the number of total categorical items; m_{ij} represents the co-occurrence between item i and item j; m_{ii} represents the occurrence of item i. The similarity matrix D is given by

$$D_{xy} = \frac{|m(X,Y)|}{|m(X)| + |m(Y)| - |m(X,Y)|} \tag{3}$$

where

X the occurrence of item x;
Y the occurrence of item y;
$m(X)$ is the set of objects having the item x;
$m(X, Y)$ is the set of objects comprising both x and y.

Finally, the matrix D describes the similarity between the categorical items; higher the value, higher the similarity.

The second stage of the algorithm continues with finding a numeric feature in the same instance, which minimizes the within group variance to base attribute. The group variance can be found out by applying the following formula

$$SS_W = \sum\nolimits_j \sum\nolimits_i \left(X_{ij} - \bar{X}_j\right)^2 \tag{4}$$

where $\bar{X}_j$ is the mean of mapping numeric attribute of j^{th} base item and X_{ij} is the i^{th} value in mapping numeric attribute of j^{th} base item. All non-base items of the categorical type can be computed by applying the following formula

$$F(x) = \sum\nolimits_{j=1}^{d} a_i * v_i \tag{5}$$

where d is the number of base item; a_i is the similarity between item x and i^{th} base item taken from D_{xy}; v_i is the measured value of i^{th} base item. Thus all the attributes in the dataset is given a numeric value.

3.3 Extension of Algorithm to Entirely Categorical Dataset

There are real world datasets, which has no or least number of numerical attributes. The given technique of finding co-occurrence and computing the group variance becomes impossible with such kind of data. To deal with the whole categorical datasets, an extension is proposed here. The first step is to add a temporary attribute FREQ to the original dataset. Each value in this column will indicate the frequency of the base item computed as per earlier methods. Specifically, each base item have its frequency in its corresponding FREQ column. The frequency is a direct indication of the strength of occurrence and is perfect to proceed with further steps.

Here a sample dataset is presented with four attributes. Attribute 2 has maximum variation of data items and hence chosen as the base item explained in the section above. The new attribute column FREQ is provisionally added to the original dataset with corresponding frequency of each base item in Attribute 2. The mean of the base items can be computed using the new FREQ column. The group variance within the base attribute can be computed as given in Eq. (4). Finally all non-base items can be given a numerical equivalent as given in Eq. (5) (Tables 1 and 2).

Table 1. Categorical dataset with four attributes

Attribute 1	Attribute 2	Attribute 3	Attribute 4
A1	B1	C1	D1
A1	B1	C2	D2
A1	B2	C1	D2
A1	B2	C2	D2
A2	B3	C3	D1
A2	B4	C3	D2
A2	B5	C3	D1

Table 2. Modified dataset with FREQ column

Attribute 1	Attribute 2	Attribute 3	Attribute 4	FREQ
A1	B1	C1	D1	2
A1	B1	C2	D2	2
A1	B2	C1	D2	2
A1	B2	C2	D2	2
A2	B3	C3	D1	1
A2	B4	C3	D2	1
A2	B5	C3	D1	1

3.4 Gridded Representation of High Dimensional Dataset

Dimensionality reduction is one of the basic functions in the steps of knowledge discovery in database is required to reduce the computational load as well as for exploratory data analysis. Each point in the observation can be characterized by n points

in a n-dimensional space. There is a high possibility that this n-dimensional representation consists of sparse data. In practice, it is easier to deal with a lower dimensional dataset rather with a high dimensional dataset. However, it is mandatory to do nonlossy transformation from a higher space to a lower plane.

3.5 Dimensionality Reduction Methods

Given a set of data points $\{x_1, x_2, \ldots x_n\}$, the low-dimensional representation is

$$xi \in Rd \rightarrow yi \in Dp\,(p << d) \tag{6}$$

which preserves the information in the original dataset can be considered as a good dimensionality reduction technique Basically, there are linear or non-linear techniques for dimensionality reduction [5].

Most common Linear Dimensionality Reduction techniques are Principal Component Analysis (PCA) and Singular valued decomposition (SVD). Principal Component Analysis (PCA) replaces the original features of a data set with a smaller number of uncorrelated attributes called the principle components. If the original data set of dimension D contains highly correlated variables, then there is an effective dimensionality, d < D, explains most of the data.

Principal Component Analysis is based on Eigen value decomposition of the covariance matrix C into

$$C = PDP^T \tag{7}$$

where P is orthogonal and D is a diagonal matrix given by

$$D = Diag(\lambda_1, \lambda_2, \lambda_3 \ldots \lambda_n). \tag{8}$$

The columns of P are eigenvectors $Cx_i = \lambda x_i$ for the eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3 \ldots \geq \lambda_n$.

Given a complex matrix A having m rows and n columns, the matrix product U $\sum$ V is a singular value decomposition for a given matrix A if U and V, respectively, have orthonormal columns and has nonnegative elements on its principal diagonal and zeros elsewhere. i.e.

$$A_{m*n} = U_{m*r}\Sigma_{r*n}V^T_{n*n} \tag{9}$$

Where U is an orthogonal matrix, a diagonal matrix, and the transpose of an orthogonal matrix where U and V are orthogonal coordinate transformations and Σ is a rectangular-diagonal matrix of singular values. The diagonal values of Σ viz. (σ_1, $\sigma_{2,} \ldots$, σ_n) are called the singular values. The i^{th} singular value shows the amount of variation along the i^{th} dimension.

SVD can be used as a numerically reliable assessment of the effective rank of the matrix. The computational complexity of finding U, Σ and V in SVD when applied to a dataset of size $m * n$ (usually $m >> n$) is $4m^2n + 8mn^2 + 9n^3$.

3.6 Visualization of Dataset Using Singular Value Decomposition

The reduction of dimensions without losing the information is a greater challenge in data mining steps. While mapping a non-spatial data to 2 D plane or 3 D space, the spatial information is being added to the non-spatial component. After the initial phase of normalizing data to a range of [0 1] the dataset is applied with Singular value decomposition for dimensionality reduction. SVD when compared to PCA, it acts on the direct data matrix and the prominent singular values are taken as the principal components. It can be observed that the vectors other than the k singular values are negligible and approximate to 0. The dot product of the original data matrix to the reduced Matrix of size (n * k) is computed. When k equals 2, the matrix can be plotted to an x-y plane and when k equals 3, matrix can be plotted to space.

The concise algorithm is given below.

Input: X: of M x N size
k: dimension ; k<<N
Output: X_{reduce} *with k dimensions*
Steps:
1. Apply min-max normalization to normalize $X_{(M \times N)}$ *into* $X\text{-}Norm_{(M \times N)}$ *a range of [0 1] ;*

$$v' = \frac{v - min_A}{max_A - min_A}\left(new\_max_A - new\_min_A\right) + new\_min_A$$

2. $[U \; \Sigma \; V] = SVD \; (X - Norm_{(M \times N)}{}^T)$

3. Set principal components = first k columns of U

$$U_{reduce} = \begin{bmatrix} e_{11} & e_{12} & \ldots e_{1k} \\ e_{21} & e_{22} & \ldots e_{2k} \\ e_{n1} & e_{n2} & \ldots e_{nk} \end{bmatrix}_{n * k}$$

4. $X_{reduce_{m*k}} = X^T{}_{m*n} * \; U_{reduce_{n*k}}$

4 Research with Various Datasets

The proposed research is tested with five major multi variate datasets from UCI repository and are discussed below. The multivariate IRIS dataset consists of 150 observations from three species of Iris: Iris setosa, Iris virginica and Iris versicolor. The length and width of petals and sepals of all three species are recorded in centimeters. Yeast is yet another multivariate set containing 1484 observations used for cellular localization sites of proteins [7] with 8 attributes. The Thyroid dataset contains 9172 instances of thyroid disease records and is provided by UCI Repository. The most frequent value technique is used to fill in the missing values. Wine dataset extracts data of 13 elements found in three types of Wine grown in the alike region of Italy. It contains 178 instances. To experiment the whole categorical case explained in the preprocessing Sect. 3.3 of this study, breast dataset with 9 categorical variables are used.

The following figures are ordered as (a) singular plot of datasets, (b) 2D plot and (c) 3D plot. Iris, Yeast, Wine, Thyroid and Breast Cancer datasets are explored in the given order from Figs. 5, 6, 7, 8 and 9.

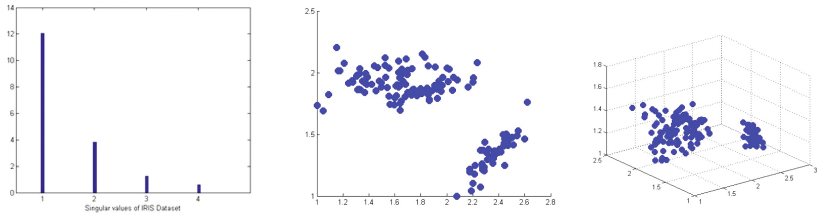

Fig. 5. a) singular plot of dataset of Iris Dataset, b) 2D plot of dataset of Iris Dataset and c) 3D plot of dataset of Iris Dataset

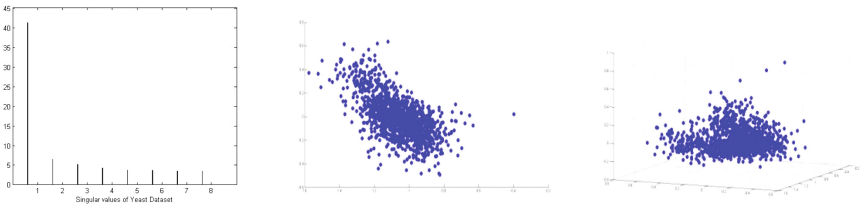

Fig. 6. a) singular plot of dataset of Yeast Dataset, b) 2D plot of dataset of Yeast Dataset and c) 3D plot of dataset of Yeast Dataset

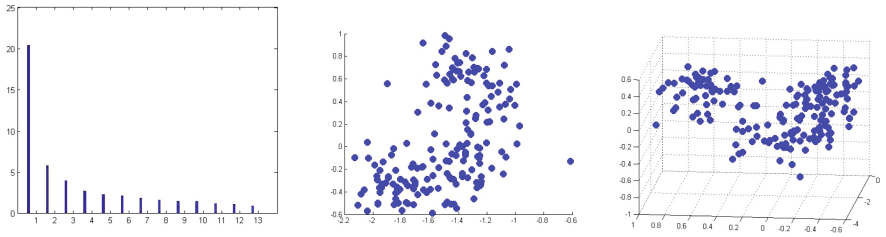

Fig. 7. a) singular plot of dataset of Wine Dataset, b) 2D plot of dataset of Wine Dataset and c) 3D plot of dataset of Wine Dataset

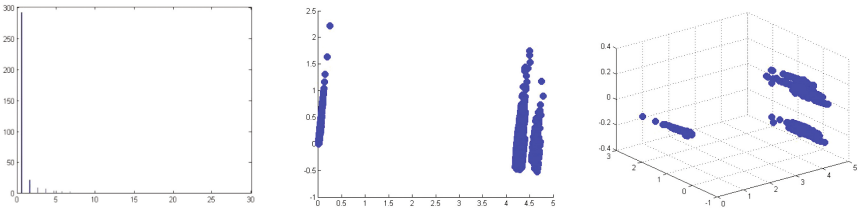

Fig. 8. a) singular plot of dataset of Thyroid Dataset, b) 2D plot of dataset of Thyroid Dataset and c) 3D plot of dataset of Thyroid Dataset

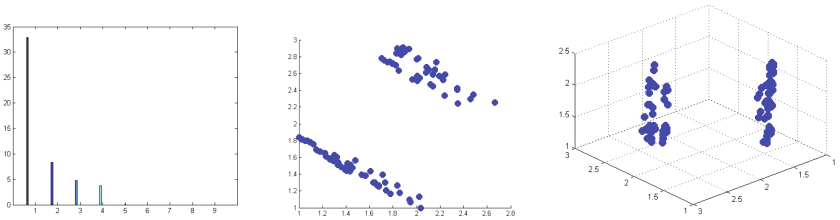

Fig. 9. a) singular plot of dataset of Breast Cancer Dataset, b) 2D plot of dataset of Breast Cancer Dataset and c) 3D plot of Breast Cancer Dataset

5 Conclusion

Existing visualization techniques are dependent on data, but an effective visualization demands the data to be independent on its type and scale. After a thorough exploration of the framework explored in this study using various multivariate datasets, Eigen decomposition or a singular value decomposition helps in representing data visually. SVD not only eliminates the curse of dimensionality, but also the pictorial mapping of a higher dimensional plane to lower dimensions, helps the user to get a better grasp on the data.

Acknowledgements. This study was conducted as a part of doctoral studies of the main author under the guidance of the co-authors and is approved by the concerned university.

References

1. Rao, R., Card, S.K.: The table lens: merging graphical and symbolic representations in an interactive focus+context visualization for tabular information. In: Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (1994)
2. Keim, D.A., Kriegel, H.-P., Ankerst, M.: Recursive pattern: a technique for visualizing very large amounts of data. In: Proceedings of the Visualization 1995, Atlanda (1995)
3. Shih, M.-Y., Jheng, J.-W., Lai, L.-F.: A two step method for clustering mixed categorical and numeric data. Tamkang J. Sci. Eng. **13**(1), 11–19 (2010)
4. Spence, R.: Information Visualization. Addison Wesley/ACM Press, New York (2000)
5. van der Maaten, L.J.P., Postma, E.O., van den Herik, H.J.: Dimensionality reduction: a comparative review (2008)
6. Fisher, R.A.: UCI Machine Learning Repository: Iris Data Set, January 2011
7. https://archive.ics.uci.edu/ml/index.html