Analyzing Visualization of Big Data for Large Unstructured Data Sets using Self Organizing Maps

G. Reshma1 K. Swarupa Rani2 D. Leela Dharami3
1,2,3Assistant Professor
1,2,3Department of Information Technology
1,2,3PVPSIT, Vijayawada

Abstract— Big data brings new challenges to visualization that must be taken into account. One of the most common definitions of big data is data that is of such volume, variety and velocity. SOM combines different techniques which helpful in visualization so that the organization can move beyond its comfort zone technologically to derive intelligence for effective decisions.

Key words: Visualization, SOM

I. INTRODUCTION

Big data brings new challenges to visualization because of the speed, size and diversity of data that must be taken into account. The cardinality of the columns you are trying to visualize should also be considered. One of the most common definitions of big data is data that is of such volume, variety and velocity that an organization must move beyond its comfort zone technologically to derive intelligence for effective decisions.

- Volume refers to the size of the data.
- Variety describes whether the data is structured, semi structured or unstructured.
- Velocity is the speed at which data pours in and how frequently it changes

II. LARGE DATA VOLUMES

One challenge when working with big data is how to display results of data exploration and analysis in a way that is meaningful and not overwhelming. You may need a new way to look at the data that collapses and condenses the results in an intuitive fashion but still displays graphs and charts that decision makers are accustomed to seeing, and provide users with the ability to easily explore data on their own in real time.

When working with massive amounts of data, it can be difficult to immediately grasp what visual might be the best to use. To analyze the data there are some methods to reduce the dimensionality of the data. one of such methods is to project the data into low dimensional space which can be either linear or non linear. After reducing to low dimensional space, the relations among the data items to be preserved. projection to a plane is often used for data visualization. there are methods existed for projection of data into a plane like sammon mapping, multi dimensional scaling, curvilinear component. in all of these methods relations will be preserved. by inspection of the projected data, user can identify clusters which are of similar data and different from others.

III. SOM

Using Self-Organizing Map (SOM)[13],the amount of data can be reduced by clustering and projection of the data into a low dimensional space. SOM is a grid of units, each is similar to the data vectors under analysis. units in SOM are called cluster centers .thus each represents a number of original data items. there are som many clustering methods are came into existence where as in SOM, the units became organized in such a way that nearby grids are similar to another. in that way a continuous ordered display of data items helpful in understanding the structures in the data sets The topology of the grid can be almost anything, but in practice rectangular two-dimensional grids are preferred since they are easy to display. All these projection methods, including the SOM, require the data to be numeric.

IV. SOM TRAINING

For SOM training, the weight vector associated with each neuron moves to become the center of a cluster of input vectors. In addition, neurons that are adjacent to each other in the topology should also move close to each other in the input space, therefore it is possible to visualize a high-dimensional inputs space in the two dimensions of the network topology.

Fig 1: Som Topology

The default topology of the SOM is hexagonal. This figure shows the neuron locations in the topology, and indicates how many of the training data are associated with each of the neurons (cluster centers). The topology is a 10-by-10 grid, so there are 100 neurons. The maximum number of hits associated with any neuron is 22. Thus, there are 22 input vectors in that cluster.

You can also visualize the SOM by displaying weight planes (also referred to as component planes).
This figure shows a weight plane for each element of the input vector (two, in this case). They are visualizations of the weights that connect each input to each of the neurons. (Darker colors represent larger weights.) If the connection patterns of two inputs were very similar, you can assume that the inputs are highly correlated. In this case, input 1 has connections that are very different than those of input 2.

However, usually without exception, in manufacturing applications the data is of mixed type consisting of both numerical and categorical variables.

V. DIFFERENT VARIETIES OF DATA (SEMI STRUCTURED AND UNSTRUCTURED)

Data variety brings challenges because semi structured and unstructured data require new visualization techniques. A word cloud visual (where the size of the word represents its frequency within a body of text) can be used on unstructured data as a way to display high- or low-frequency words. One example to represents different varieties of data by using the self organizing semantic map is word cloud.

A. Self organizing Semantic Map

SOM takes the concept of word clouds to make associations among words and then organize these words into topics based on how the words are being used. self organizing semantic maps also can be referred as word category maps are SOMs that have been organized according to similarities among words, and those are measured by the similarity of the short contents of the words.SOM algorithm is based on competitive learning which can be described as the artificial neurons of the netwok become sensitive to different input categories. the neurons are from usually 2-D array. we can say it as a map.

When an input vector is processed, the best matching neuron on the map can be obtained by the closest Euclidean distance wins the competition and it is updated as well as its neighborhood according to the adaptation rule. A model vector is associated with each neuron, the vector specifies the coordinates of a neuron in the original space. the amount of the winning neuron and its neighborhood adapt is governed by neighborhood kernel. The kernel which is chosen to be large in the beginning of the learning process which helps in global ordering of the map. the width and height of kernel is decreasing slowly during learning.

In the unsupervised formation of the word category map, each input x consists of an encoded word and its averaged context. Each word in the vocabulary is encoded as a n-dimensional random vector. Word category map can be formed by creating a unique random vector for each word in the vocabulary. Each word to be mapped in the input text collection and find all the instances of each word. those words are keyword. After identifying key words calculate the average over the contents of each keyword.that random codes are which are generated above are used in the calculation. The context may consist of, e.g., the preceding and the succeeding word, or some other window over the context. As a result each key word is associated with a contextual fingerprint. If the original random encoding for a single word is, for instance, 90-dimensional, the resulting vector of key word with context is 270-dimensional if one neighboring word from both sides is included. The classifying information is in the context. Therefore the key word part of the vector is multiplied by a small scalar, ε. Each vector formed above is input to the SOM. The resulting map is labeled after the training process by finding best-matching neuron for each vector and labeling the neuron with the key part of the vector. The overall organization of a word category map reflects the syntactic categorization of the words. The SOM performs a non-linear mapping from the original multi-dimensional space into the two-dimensional lattice. To visualize the nonlinearities, one may use U-matrices to show the distances between neighboring map nodes in the original input space.

The semantic representation of the target words can be further processed by a self-organizing map (SOM; Kohonen, Self-organizing maps, 2001), an unsupervised neural network model that provides efficient data extraction and representation. Due to its topography-preserving features, the SOM projects the statistical structure of the context onto a 2-D space, such that words with similar meanings cluster together, forming groups that correspond to lexically meaningful categories. Such a representation system has its applications in a variety of contexts, including computational modeling of language acquisition and processing. There are several very valuable things that a SOM can do for you. One of those is to show correlation of data. The SOM's show text that is correlated to other text. In the medical field, working with medical records, this ability to correlate is very attractive. Another thing SOM's can do is to enable textual drill down processing. In textual drill down processing the analyst goes from one level of analysis to a lower level of analysis until the specific detail is found. Depending on how the data is arranged and integrated, SOM's support qualified analysis of data. For example in an organization the qualified analysis can be done by the following process. First the analyst looks at records for employees. Then the analyst looks for records for women employees. Then the analyst looks for records for women college graduates. Then the analyst looks for records for women college graduates who are older than 50, and so forth. If the unstructured data is properly conditioned and edited, in this similar ways SOM's can also yield insightful analysis for the selection of qualified data.

B. Different Velocities of Data

A correlation matrix combines big data and fast response times to quickly identify which variables among the millions or billions are related. It also shows how strong the relationship is between the variables. Feature Clustering...
Method based on Self-organizing Maps. We explore the visual properties of the SOM (namely a particular visualization technique called Component Planes), which can be seen as a "sliced" version of the SOM, to obtain correlations between features. Each feature in the SOM has its own component plane. If two component planes are similar to each other, this means that the two features associated with them are correlated.

VI. CONCLUSION

Visualizing your data is challenging. It is much easier to understand information in a visual compared to a large table with lots of rows and columns. In this paper we are presenting SOM visualizing techniques like topology, word clouds, component planes which are helpful in many experiments to visualize the big data which comes in three varieties i.e., volume, velocity and variety. Hence we can represent the unstructured big data using SOM visualization techniques.

REFERENCES