Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 4, April 2017, pg.1 – 4

Effective Dimensionality Reduction for Large Scale Datasets

Pragna Vedanthi, Pankaj Nalawade, Bronson Singh, Avishkar Watpade
Computer Department, Savitribai Phule Pune University, Nasik, India

Abstract— The recent explosion of data set size, in number of records as well as of attributes, has triggered the development of a number of big data platforms as well as parallel data analytics algorithms. At the same time though, it has pushed for the usage of data dimensionality reduction procedures. Therefore, before proceeding with any data analytics task, we needed to implement one or more dimensionality reduction techniques. Dimensionality reduction is not only useful to speed up algorithm execution, but actually might help with the final classification/clustering accuracy as well. Too much noisy or even faulty input data often lead to a less than desirable algorithm performance. Removing un-informative or dis-informative data columns might indeed help the algorithm find more general classification regions and rules and overall achieve better performances on new data.

Keywords— Distributed Data Mining, Data mining, PCA, Pearson Correlation Coefficient

I. INTRODUCTION

Data mining is one of the most promising domain in today's world and data storage and processing are some of the challenging task present in this domain. Dimensionality reduction helps in minimizing the anomalies in processing huge amount of data by reducing the number of variables used for storage. As everyday data is in heterogeneous form, so it is difficult to handle such data. To improvise the quality of the large scale data, PCA algorithm provides this opportunity.

II. LITERATURE SURVEY

The field of Data Mining has a different behaviour towards Big Data. It can deal with data-sets having size gigabytes or even tera bytes. The main concern over here is that the algorithms which are used in data mining operations work on small data sets and do not give better results on large data sets. To work efficiently with large data sets, the algorithms must have high scalability.

III. Proposed System

After a vast survey performed in literature [1], [2] and [3] we have studied the results and chosen few dimensional reduction techniques. But, still performing dimension reduction on Big Data is an issue. To study and differentiate data, is an issue as there are several

dimensions and which dimensions are necessary to select creates problem. And with these all reasons we got motivated to study the dimensionality reduction process and merged based algorithm to achieve the following:

To propose a system which performs merging of numerical dataset.

Performs dimensionality reduction on the data to reduce noise, dimensions for proper use of it.

IV. METHODOLIES OF PROBLEM SOLVING AND EFFICIENCY ISSUES

We started working with the "small" data set to evaluate a few classic dimensionality reduction methods. The relatively small number of data columns allows for faster evaluation and comparison of the different techniques. It was only after we had gained a clearer picture of the pros and cons of the evaluated dimensionality reduction methods that we approached the "large" data set for a more realistic analytics project. Here we used a cascade of the most promising techniques, as detected in the first phase of the project on the smaller data set. In this whitepaper, we concentrate on a few state-of-the-art methods to reduce input dimensionality and examine how they might affect the final classification accuracy. In particular, we implement and evaluate data columns reduction based on:

1. High number of missing values
2. Low variance
3. Pearson correlation coefficient
4. Principal Component Analysis

The most straightforward way to reduce data dimensionality is via the count of missing values. While there are many strategies to interpolate missing values in a data column, it is undeniable that if the majority of data values is missing, the data column itself cannot carry that much information. In most cases, for example, if a data column has only 5-10% of the possible values, it will likely not be useful for the classification of most records. The goal, then, becomes to remove those data columns with too many missing values, i.e. with more missing values in percent than a given threshold.

The Low Variance Filter node calculates each column variance and removes those columns with a variance value below a given threshold. Notice that the variance can only be calculated for numerical columns, i.e. this dimensionality reduction method applies only to numerical columns. Note, too, that the variance value depends on the column numerical range. Therefore data column ranges need to be normalized to make variance values independent from the column domain range.

Often input features are correlated, i.e. they depend on one another and carry similar 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. One of the two columns can be removed without decreasing the amount of information available for future tasks dramatically.

In order to remove highly correlated data columns, first we need to measure the correlation between pairs of columns using the Linear Correlation node, then we have to apply the Correlation Filter node to remove one of two highly correlated data columns

Analysis of some Dimensionality reduction methods can be visualized for understanding high dimensional space. The simplest approach is to use linear methods, but the complexity of the recent data sets makes these methods useless. Thus, the nonlinear versions of those linear methods were introduced to overcome their limitations. Nonlinear transformations can be used to project high-dimensional data-sets in a low-dimensional space. The system focuses on performing merging on low dimensional data obtained by applying Dimensionality Reduction techniques. The input data-set is numerical and is passed to PCA model for

Dimensionality Reduction. Depending upon the size of co-variance matrix in PCA, efficiency of the system will be affected. Divide and Conquer can be used to avoid high execution context.

PCA Algorithm:

PCA is a statistical method which uses orthogonal transformation (it is linear transformation which doesn't change even after performing rotation ad reflection operation upon the data) to convert set of observation of possibly correlated attributes into a set of values of unrelated data variables. It identifies patterns and finds patterns to reduce the dimensions of the data with minimal loss of information. The attributes are been converted to Principal attributes which are important and necessary for defining the data.

V. ALGORITHM

Input: Data

Output: Data with only Principal attributes

Steps:

1. Perform the orthogonal transformation
2. Select the Eigen vectors (Those attributes which are not affected by the above operation) and Eigen values
3. Sort the Eigen vectors and Eigen values according to decreasing order.
4. Select some subset of Eigen vectors as per their values as Principal Attributes of the data.

Complexity: O (p2 n+p3), where p is the features and n is data points.

Pearson Correlation Coefficient

The Pearson correlation coefficient is a very helpful statistical formula that measures the strength between variables and relationships. In the field of statistics, this formula is often referred to as the Pearson R test. When conducting a statistical test between two variables, it is a good idea to conduct a Pearson correlation coefficient value to determine just how strong that relationship is between those two variables.

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

- N = number of pairs of scores
- Σxy = sum of the products of paired scores
- Σx = sum of x scores
- Σy = sum of y scores
- Σx^2 = sum of squared x scores
- Σy^2 = sum of squared y scores

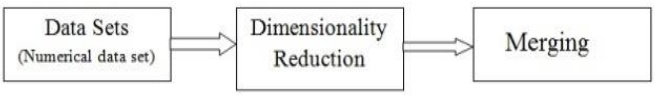

FIG: FLOW DIAGRAM

VI. CONCLUSIONS

The main purpose of this project is to address the issue of analysing and modelling both quantitative observations. Such a problem is widespread in medical fields, especially in occupational medicine. We have proposed a new methodology for processing such a type of information by connecting a set of well-known dimensional reduction techniques. This simple interconnection has created a system that can fill a significant number of unobserved data, before allowing to analyse and model on the same dimensionally reduced large scale data.

REFERENCES

[1] Vladimir Filkov and Steven Skiena, "Heterogeneous Data Integration with the Consensus Clustering Formalism"
[2] Rosaria Silipo , Iris Adae , Aron Hart , Michael Berthold, "Seven Techniques for Dimensionality Reduction"
[3] Éverton Santi, Daniel Aloise and Simon J. Blanchard " Heterogeneous Dissimilarities"
[4] Chuan Hu, Huiping Cao, Aspect-level Influence Discovery from Graphs
[5] Zhang, T.; Ramakrishnan, R.; Livny, M. BIRCH: "An efficient data clustering method for very large databases." In Proceedings of the 1996 ACM SIGMOD
[6] A. Ahmad and L. Dey, "A feature selection technique for classificatory analysis," Pattern Recog. Lett., pp. 43–56, 2005.
[7] C. Biernacki, G. Celeux, G. Govaert, and F. Langrognet, "Model-based cluster and discriminant analysis with the mixmod software," Comput. Stat. Data Anal., vol. 51, pp. 587–600, 2006.
[8] Everton Santi, Danial Aloise, Simon J. Blanchard , "A model for clustering data from heterogeneous dissimilarities," in European Journal of Operational Research
[9] H. A. L. Kiers, "Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables," Psychometrika, vol. 56, pp. 197–212, 1991
[10] L. Hunt and M. Jorgensen, "Clustering mixed data," Wiley Interdisciplinary Reviews: Data Mining Knowl. Discovery, vol. 1, pp. 352–361, 2011.

