International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 6 (2015) pp. 15621-15630

http://www.ripublication.com

Mathematical Model With Multi Variate Data To Support Big Data Clustering Algorithms

Manishankar. S
Department of Computer Science,
Amrita Vishwa Vidyapeetham,Mysore Campus,India.
E-mail: manishankar1988@gmail.com.
Akshatha Prabhu
Department of Computer Science,
Amrita Vishwa Vidyapeetham,Mysore Campus,India.
E-mail: akshathaprabhu06@gmail.com.

Abstract

Growth of real data in servers is increasing day to day in proportion. Handling such a big data with the help of normal data warehouses and data centre is becoming highly inefficient. This has been handled to some extent with the help of Haddop. Then also a need arises to have a mathematical model to represent data efficiently so that processing becomes faster and clusterization as well as and storage is being carried out effectively. Technologies like Hadoop Mahout considers single valued or single varied distributions with matrix representations to process the data. Considering the varying attributes of data we propose a novel approach with a multivariate or multi valued distributed mathematical model which can handle diverse data.

Index Terms: Big data; medical data set ; multi valued attribute; clustering; categorial distance algorithm; correlation co-efficient.

Introduction

Big data has been termed as voluminous amount of data. There has always been a gargantuan challenge involved in processing these analogue data which have to be processed together with analysis, capture, visualization, storage and transfer that is different from traditional data ware house processing techniques [1]. Big data is characterized by volume, velocity and variety and mostly data set encompasses rich features, unstructured data withcomplex formation such as permutations, strings and graphs that lie in complex discrete space. Prediction and comparison of real data against existing trained data set is powered by big data [8]. Data is usually in terms of petabytes and terabytes. In totality, need arises to have a scalable and distributed

system which organizes and handles data. Usually machine learning tasks such as recommendation or classification or clustering is preferred to analyse data . Mining in big data has a major loop hole for the fact that it creates a confusion amongst the practitioners as to choose the techniques apt for the data sets [14]. Representation of data in terms of mathematical notations that can be easily fed as an input to the analytic system [4]. Training an effective machine learning platform with a mathematical model is a prerequisite for any big data handling domain. Analytics and visualization technique is as well important when we carry forward the processing of big data, big data users should have an ideal knowledge about the analytics taking place [2]. It depends purely on the type of data that we perform what kind of data mining technique we use in order to make a machine learning possible for the data sets [22]. Classifications performed is a much more difficult task as there might be an inadequate training for machine. As clusteriztion supports unsupervised learning , the training of the data set can be efficiently handled even though it is quite large [12]. There are many clustering techniques which help to analyse data set. Various clustering algorithms k-means, canopy, Fuzzy k-means, Streaming k-means, Spectral Clustering and many more. There has been some advancement in those techniques with regard to the mathematical model that has been used to represent the data. Data mining and business analytics can be performed easily if we have a data set which considers more number of attributes in order to analyse.

Existing System

Existing system is a scalable , machine learning library platform which supports analytics in larger data sets with various techniques. There has been an integration of recommendation , classification and clustering. The similarity between the features of various items are checked. Each item is represented as a vector. Between a pair of item, the distance is measured among these vectors using distance algorithms such as Manhattan , random walk , Euclidean. Items are grouped together which has minimum distance forming cluster of similar features [13].

The algorithms used to process data are supported by underlying mathematical model based on matrix algebra. Data and its features are represented in query vector. Since huge data sets are considered ,the matrix dimension is quite large. Considering a single valued decomposition with a dense or sparse matrix, sparse matrices may contain innumerable overlapping of zero entries. Dense matrices have huge collection of non zero entries. Dimensionality reduction is a better approach for dealing with sparse matrices where feature space of the data set is constructed as the column and the rows are mapped onto the feature space. This forms a reduced dimensional space. Distance between values is considerably large in case of sparse matrix. The loss of data during reduction in dimension can disrupt the feature space resulting in improper analysis.

In Eigen valued decomposition, singular vector of a given vector A is obtained in form of the eigenvectors of AT * A or $A * A^T$. This type of decompositions works in a stochastic manner [7]. Data sets with rich feature space mostly find a slower decomposition rate with singular value decomposition and contains a large fraction of

the variability. Large spatial correlations are often found in the larger data sets which exhibit spatial correlations with respect to this decomposition. Hence, it is tedious work to analyse a data with large correlations [19].

Proposed System

The system aims at processing bigger organizational data and analysis with the machine learning based platform and a new mathematical model. In order to have an enhanced solution for segregation of multivariate data. Data in most of the organisation is composed of varied , unstructured and speculative types [11]. Considering a model which support single valued attributes cannot process these data efficiently where storage and retrieval is a herculean task. Always there is a need for multiple parameters with multiple values. So a desirable solution is one with multi valued matrix model. This solution works under the domain of matrix model of covariance and mean vector. The system aims at processing bigger organizational data and analysis with the machine learning based platform and a new mathematical model. In order to have an enhanced solution for segregation of multivariate data. Data in most of the organisation is composed of varied , unstructured and speculative types [11]. Considering a model which support single valued attributes cannot process these data efficiently where storage and retrieval is a herculean task. Always there is a need for multiple parameters with multiple values. So a desirable solution is one with multi valued matrix model. This solution works under the domain of matrix model of covariance and mean vector.

Methodology

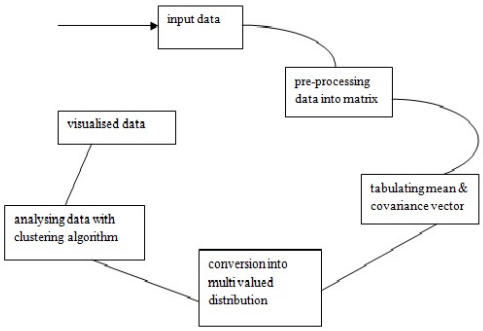

Figure: Architecture Diagram

System is a integrated approach of various components which process data and clusters it to a precise format which suits high scalability processing in comparison with big data. Input data is huge data that arises from any service oriented data collection platform [5]. Data is converted into a desired format through pre-processing where some unique steps are followed in congruence with the algorithm. As a

preliminary step input data is converted into its equivalent matrix form. Mean and covariance vector of each individual matrix which is formed out of input data is converted and pre-processed [20]. It involves mainly the formulation of correlation factor which gives stronger relations between multi valued attributes. Multi valued distribution is applied to the covariance matrix and correlation co-efficient in order to enhance the clustering algorithm used to segregate data for efficient storage[16]. The clustering algorithm working with a concept of calculating equivalent centroids in order to supplement the multi valued data.

The data has to be pre-processed into a random vector A and composed of k different components A1,A2,A3.....Ak and joint cumulative distribution function F(X) tabulated as FX(x) = P(X<=x) where X is the random variable and x is the range value. Probability that the random variable X lies in the interval [a,b] is P(a<X<b) = FX(b) - FX(a).The joint probability density function is the integration of random variables in the given range which is given [21] by

$$F_X(x) = \int_{-\infty}^{x} fx(t)dt$$

The density function gives a better idea for proposing a matrix model. Multivariate Gaussians are parameterised by mean vector V and co-variance vector Z. The parameters of ith and jth component of the covariance matrix is the covariance between diagonal elements of the matrix[17].

$$p(x)=(2\Pi \det \Sigma)\text{-}p/2\exp(-1/2(x\text{-}u), \Sigma\text{-}1(x\text{-}\mu)\}$$

Normally the vector space is a Gaussian elliptical space which has equal geometric distance from the mean vector. Correlation co-efficient is found for multi values in order to interpret better range of data. The linear algebra models Gaussian as symmetric and positive definite values represented as Eigen values and Eigen vectors given by [6].

$$E= V^{T}dv$$

where dv is the matrix of diagonal elements of all possible combination of Eigen values computed from random vector A and E is the Eigen vector which represents the Gaussian vector space [19].

The representation of data in matrix for the vector space helps the clustering algorithm such as spectral clustering k means to segregate data. Here arises a need for a clustering algorithm that exactly takes the consideration of multi valued Guassian vector space.

Clustering Algorithm For Multi Valued Attributes

Data that is processed in the system contains categorial multivalued attrribute. In order to segregate the data for better analysis and retrieval there arises a need for an efficient clustering algorithm , many of the non-clustering algorithms which works on normal data sets fails to process such a huge data. For enhancing our approach we need a tabulation of distance between various attributes of the categorical data. The

data being represented in the matrix as explained in the mathematical model. Here there is a requirement to compute distance between each pair of attributes so that we have a better clustering of data.

Considering a data set D={d1, d2, d3.....dn} and set E={X1,X2,X3.....Xn} , n values of instances for the data set, assuming m multi valued attributes of the categorial data set. Initially considering X1 to be a categorial attribute , we need to compute the distance between any pair of its instances. Selection of attribute based on a threshold values of their instances is critical for the algorithm to give accurate result [18].

The computation begins with the pre-processing of data in matrix form as explained in the mathematical model to calculate mean and covariance vector. The pre-processed matrix is taken as input to the algorithm to calculate the distance between parameters.

To begin with we have to identify those parameters which are critical to the data using formulation of threshold values of parameters. Threshold values are dependent on the association and correlation between values of parameters , so knowledge of the range value of parameters is essential. We compute range as function of cumulative distribution explained in methodology.

Threshold value

$$TX = \frac{\sum(X1,X2,X3.....Xn) \times \Delta (Fx(b) - Fx(a))}{NX}$$

where Xi - instance of data valaue di

Δ FX - difference between min and max range values

N - total number of instances in the sample space

Based on the threshold , the choice of evaluation parameters for the clustering algorithm.

The process continues with finding categorial distance between evaluation parameters,

Distances $S(i,j) = \sqrt{\sum} Txi \sum(Pyj(T(Xi/y) - T(Xj/y))2)$

where Tx - threshold

Pyj - Probability of evaluation parameter

T(Xi/y) - Occurrence of threshold with varying instance of X in correlation with y.

Algorithm

Categorical Parameter Distance(S,T,P,N)

1. Vector Space A = PRE[A]
2. Compute T with Mean and Co-variance as n

$$T= \frac{\sum Xi+\sum Zi - (Vi*(Fxmin- Fxmax))}{N1}$$

$i=0$

3. Tabulate the standard deviation

$$\sigma = \sqrt{\frac{\sum((X) - \bar{X})2}{N-1}}$$

4. Find the correlation co-efficient [15]
   $C = \frac{Sxy}{\sigma}$
5. If the instance X satisfies threshold then probability P is computed by
   $P = (T(Xi/y) - T(Xj/y))^2$
6. For all instances Xi Xj €X and y, the varying instance Calculate categorical distance
   $Di,j = \sqrt{\sum} Txi \sum(Pyj(T(Xi/y) - T(Xj/y))^2)$
7. Calculate the centroid value as the mean of distance of various matrix with different Dij.
   $G = \frac{\sum \Delta Dij}{N}$
8. Divide the cluster according to varying values of G
9. End

The clusterization is much efficient since the distance calculated from multiple attributes provides an exact estimation of throughput of data processed and stored. The algorithm proposed suits the platform as such of big data.

A data set with mutivalued heterogeneous data process from a huge data warehouse.

Implementation Details

The system is made up of a huge data server and an application program that supports collection of data from various medical databases [3]. The data is interglued with multi valued attributes essential for faster segregation and processing of information. The data in comma separated format is taken as the input. The specific attributes are considered during pre-processing the input data. The distance and the centroid for the attributes are formulate through pre-processing of data. Data is supplemented with multi valued parameters representing various features essential for clusterization. Preprocessing involves conversion of data from csv format to matrix form using heat map reshaping the parameter values to get the data such as one row per column /row combination using Matlab function which reads an entire csv file starting at a specific row and column offset , read specific range from csv value. Functions such as csv read and csv write helps for easy conversion to matrix form. To find the evaluation parameters we perform the Categorical Parameter Distance algorithm. Algorithm performs mapping of features of data to a Gaussian vector space. The threshold is calculated with mean co-variance vector and correlation co-efficient [16]. The probability for each instance is tabulated with the help of range of variation of multiple values. The categorial distance for all instances and the varying instance is found. The algorithm further continues with the finding of centroid as the mean of

distance of various matrix with different categorial distance which is further divided with various centroid values. Centroid gives the proper area of the cluster, with multi values present, cluster will be with more outlier. A file system has been developed in resemblance with hadoop [10]. There is a need to develop a file system similar to that of hadoop wherein data processing is supported with multi valued attributes.

Data Table

The implementation has been done with considering data from a medical database and record has been verified to contain multi valued attributes. Data first converted in a csv format is pre-processed to matrix and calculation of mean , co-variance and evaluation parameter is done. After pre-processing data header, the file is stored into file system that has been developed.

I0 Impedance range value with reference to 0

Table 1: Table of evaluation parameters

PA500	Phase angle at 500 KHz
HFS	Max frequency shift value.
DA	Impedance distance between spectral ends
AREA	area under spectrum
A/DA	area normalized by DA
MAX IP	maximum of the spectrum
DR	distance between I0 and real part of the maximum frequency point
P	length of the spectral curve

Here PA500 is a multi valued attribute at various phase angles, the value of PA500 of 500 KHz varies. similarly other multi varied attributes like HFS,DA,DR and P are also taken in to consideration for faster pre processing. Each of the data values are mapped into cluster 0 and cluster 1 according to the threshold value present for each instance. Here the total sample is split into 30% in the cluster 0 and 70% in the cluster 1.The centroid distance always varies according to the group of instances present. The file system is processed into sectors on the basis of these clusters.

Table 2: Table of clusters

Cluster centroids:

		Cluster#	
Attribute	Full Data	0	1
	(106)	(31)	(75)
Case #	53.5	90.5806	38.1733
Class	adi	adi	car
I0	784.2516	1879.9877	331.3474
PA500	0.1201	0.0746	0.139
HFS	0.1147	0.1108	0.1163
DA	190.5686	416.2653	97.2807
Area	7335.1552	19830.9045	2170.2454
A/DA	23.4738	41.1624	16.1625
Max IP	75.3813	164.596	38.5058
DR	166.7106	361.2298	86.3093
P	810.6381	1886.9889	365.7465

To enhance faster processing of the very large data sets some map reduce functions can be used together so that Peta bytes of unstructured multi valued data is taken care while storage [9]. The simulation is done considering only data sets of few thousand records.

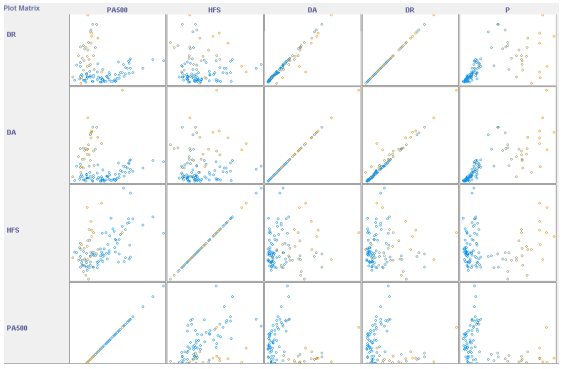

Figure 1: Visualisation of clusters

The above figure depicts the visualisation of clusters with various multi valued attributes. Each cluster column represents the value of the pair in the cluster 0 and cluster 1 represented with different colours. The intersecting pair of value represents the data characterised by both cluster. There are sectors with large near zero variance and large positive covariance.

Conclusion and Future Scope

The system is modelled and characterised by multi valued attributes based on mathematical model, calculating the essential parameters thus forming an efficient data set to perform clusterization. Cluster values is calculated and plotted to the sectors inside the file system. The system is found to give efficient performance and

segregation in case of test data. Considering the various performance issue , it can be scaled to larger big data. Processing capabilities can be subjected to test and factors can be improved. Consideration of integrated inter domain data feature analysis can be an enhancement of the model. A priority based model data can be included in the calculation of a rank based correlation to improve the scope of the model.

Acknowledgements

With divine intervention , we express our deepest gratitude to our parents and family members in supporting and motivating us during the research that has lead to publication of this paper.Thanking the Institution for full fledged support , we also remember the responsiveness of our research guides in fine tuning this paper and the co -operation of colleagues.

References

[1] Big Data Clustering via Sketching and Validation Traganitis, P.A. ; Slavakis, K. ; Giannakis, G.B. Selected Topics in Signal Processing, IEEE Journal of Volume: PP Issue: 99

[2] Big data machine learning and graph analytics: Current state and future challenges, Huang, H.H. ; Hang Liu ,Big Data (Big Data), 2014 IEEE International Conference

[3] Big Data in Health Informatics Architecture ,Onyejekwe, E.R. , Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference

[4] Big Data analytics, Singh, S. ; Singh, N, Communication, Information & Computing Technology (ICCICT), 2012 International Conference

[5] Real-time big data analytics: Applications and challenges , Mohamed, N. ; Al-Jaroodi, J. , High Performance Computing & Simulation (HPCS), 2014 International Conference

[6] Multivariate Generalized Gaussian Distribution: Convexity and Graphical Models, Teng Zhang ; Wiesel, A. ; Greco, M.S , Signal Processing, IEEE Transactions,Volume:61, Issue: 16

[7] On the Multivariate \alpha – \mu Distribution: New Exact Analytical Formulations , Rabelo, G.S. ; Yacoub, M.D.; de Souza, R.A.A., Vehicular Technology, IEEE Transactions Volume:60, Issue: 8

[8] Big data challenges in information engineering curriculum, Okur, M.C.; Buyukkececi, M , EAEEIE (EAEEIE), 2014 25th Annual Conference

[9] 5Ws Model for Big Data Analysis and Visu alization , Jinson Zhang ; Mao Lin Huang, Com putational Science and Engineering (CSE), 2013 IEEE 16th International Conference

[10] Big Data Analysis with Signal Processing on Graphs: Representation and processing of massive data sets with irregular structure,Sandryhaila, A. ; Moura, J.M.F ,Signal Processing Magazine, IEEE , Volume:31, Issue: 5

[11] Towards a User-Friendly Loading System for the Analysis of Big Data in the Internet of Things, Mesiti, M.; Valtolina, S. ,Computer Software and Applications Conference Workshops (COMPSACW), 2014 IEEE 38th International Conference

[12] Text clustering using statistical and semantic data, Benghabrit, A. ; Ouhbi, B.; Behja, H.; Frikh, B., Computer and Information Technology (WCCIT), 2013 World Congress

[13] Similarity detection among data files-a machine learning approach,Dash, M.; Liu, H,Knowledge and Data Engineering Exchange Workshop, 1997.

[14] Addressing big data problem using Hadoop and Map Reduce,Patel, A.B.; Birla, M.; Nair, U, Engineering (NUiCONE), 2012 Nirma University International Confence.

[15] Asymptotic Properties of Order Statistics Corre lation Coefficient in the Normal Cases, Weichao Xu; Chunqi Chang; Hung, Y.S.; Chin Wan Fung, P. Signal Processing, IEEE Transactions, Volume: 56, Issue: 6

[16] Correlation coefficient estimation for stochastic FDTD method, Bisheh, K.M.; Zakeri, B.; Andargoli, S.M.H, Telecommunications (IST), 2014 7th International Symposium.

[17] Early Termination Algorithms for Correlation Coefficient Based Block Matching, Mahmood, A.; Khan, S., Image Processing, 2007. ICIP 2007. IEEE International Conference on Volume:2

[18] An enhanced entity-attribute-value data model for representing high dimensional and sparse healthcare data, Kamau, A. ; Mwangi, W. IST-Africa Conference and Exhibition (IST- Africa), 2013

[19] Regularization networks for approximating multi-valued functions: learning ambiguous input-output mappings from examples. Shizawa, M. Neural Networks, 1994. IEEE World Congress on Computation al Intelligence., 1994 IEEE International Conference, Volume:1

[20] Multi-valued functional decomposition as a machine learning method Files, C.M. ; Perkowski, M.A. Multiple-Valued Logic, 1998. Proceedings. 1998 28th IEEE International Symposium

[21] Evaluation of Trend Localization with Multi-Variate Visualizations , Livingston, M.A. ; Decker, J.W. Visualization and Computer Graphics, IEEE Transactions on Volume:17 , Issue: 12

[22] Application on web mining for web usability analysis Jian-Li Duan ; Shu-Xia Liu Machine Learning and Cybernetics (ICMLC), 2012 International Conference on Volume:5