TITLE: SOFT COMPUTING TECHNIQUES FOR BIG DATA ANALYSIS: A REVIEW

Vedant Saini	Ishita Kohli	Karan Mahajan	Sukriti Bahri
Student DAVIET,	Student DAVIET,	Student DAVIET,	Student DAVIET,
Jalandhar	Jalandhar	Jalandhar	Jalandhar

ABSTRACT:

Analysis on large scale data .i,e. Big Data has been proposed in four techniques in this research paper, namely Neural network , fuzzy Logic , Ant Colony Optimization and Genetic Algorithm. Sorting of large amount of data is a difficult and time consuming task but using these techniques Big Data analysis becomes easy and more efficient.

KEYWORDS

Neural Network, Fuzzy Logic, Ant Colony Optimization, Genetic Algorithm.

INTRODUCTION

Big data can be defined as a large set of data that we are not able to solve through traditional approach. The data can be unstructured as well as multi - structured. Big data, today has its application in every field ranging from social networking sites, Artificial Intelligence, Cloud computing, IoT(Internet of things), Machine learning etc.

In general, everything generated around us is big data. It has changed the way people work. Organiation has realized the importance of it and many business and IT companies have come up together and helped in improvement of their decision making capacity.

Big data has defined in form of **three V's**:-

Volume - Volume is the amount of data that organization collect from resources and maintain for the future use.

Velocity -Velocity is the amount of time required to fetch the meaningful data from the records

Variety – Variety basically refers to the type or variety of data that we come across. Data from records can be in form of text, video, image or any other format.

Recently another **V** has been added.

Variability - Data flow from the records can be very inconsistent.At times when one or the other thing is in trend this flow can be very inconsistent.

1. Neural Network

Neural Network theory is divided in various aspects such as simulation modeling , network designing and to provide efficient performance.

This section of the paper describes various neural network that can be merged with Big data to provide a better, fast and effective results.

In the previous work it has been concluded that the number of neurons and number of hidden layers can affect performance as the small number of layers can be processed easily. Number of hidden layer increase the accuracy of learning, but it may increase the learning time and a large data set is difficult to interpret at one time. In this paper a large data set would be divided into multiple subsets each of which is trained using neural networks, then overall accuracy will be improved using techniques that replaces the weight from the smallest error to each node of neuron. The weights of the best small data sets are used to create a new network. Using this neural network technique the accuracy is increased and training time is drastically decreased.

1.1. Back propagation Neural Network

Back propagation is a technique in which neurons connect each other with weighted connections. The signal transmitted to the next neuron is weighted by link connecting both the nodes. Threshold and weight will be determined by transfer function. It will be transmitted by the neuron in the hidden layer.

$A_j(b) = \text{sigmoid}(\sum_{i-1}^{n} \quad x_i \cdot (b) * w_{ij}(b) - \theta_j)$

J is learning example, n is number of output of j neurons in hidden layer, Xi is input I that transmit to neuron and $A_j$is output. W is weight of each neuron,$\theta_j$is threshold.

The sigmoid function is :

$$\text{Sigmoid}(x)=\frac{1}{1+e^{-x}}$$

The values of above will be put in :

A_k(b)=sigmoid($\sum_{j=1}^{m} \quad X_{jk}(b) * W_{ij}$(b)- θ_k)

K is learning example in data, m denotes number of neurons in output layer ,k denotes output layer,X_{jk} is input from hidden layer,A is output .W is weight of each neuron and θ_k is threshold of output.

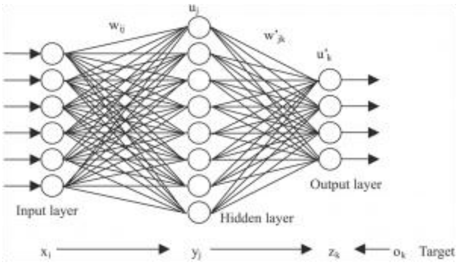


Fig 1.1 Back propagation

1.2 TECHNIQUES IN NEURAL NETWORK

1.2.1 DEEP NEURAL NETWORK

This method aims to improve training time by dividing the data into multiple processing layer or subsets with the help of an algorithm. Hence, processing the smaller subsets or layers makes the computation easy and efficient. Each subset will give a weight in result of its computation. After computation of each subset, a best weight will be chosen from the best trained subset and it will be the starting weight in the weight integration process.

1.2.2 WEIGHT INTEGRATION

As soon as the best weight is selected, if two weights are close to each other from different nodes, their average will be calculated.

Wavg = $\frac{(w1+w2)}{2}$

W1 and w2 are weights from two different nodes that are close to each other.

A net output function is applied to all the best weights and the result of that will be applied an activated function to give output as shown in diagram.

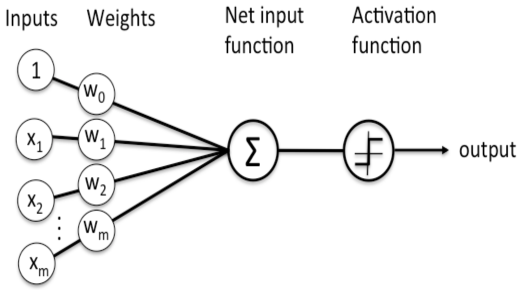


Fig 1.2 Network Integration in deep neural

2. FUZZY LOGIC

Till now it has been made clear that big data is huge and it is merely impossible to look into the data and derive to conclusion .Therefore, techniques are being developed to find the related data and their likeness and relationship between them from raw data that we have collected. Fuzzy technique basically refers to clustering the data. Clustering is itself one of the biggest unsupervised problem today. It refers to the finding of structure in the unlabeled section data. Structure is the finding of definite pattern from raw data and then organizing it into groups of similarities where similar objects are kept in same group.

2.1 C Means Clustering

Clustering is the first and foremost step of dividing data into groups that are being used as knowledge discovery tool .This further can be used for indexing, labeling and compression. In Fuzzy C Means clustering the same data item can participate in two or more clusters, and each data element is assigned with the membership values that indicates the strength of that data element with the particular cluster . The fuzzy C Means can be accomplished through the FCM algorithm which can be seen as followed :-

1. *Initialize U=[u_{ij}] matrix, $U^{(0)}$*
2. *At k-step: calculate the centers vectors $C^{(k)}$=[c_j] with $U^{(k)}$*

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

3. *Update $U^{(k)}$, $U^{(k+1)}$*

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

4. *If $\|| U^{(k+1)} - U^{(k)} \|| < \varepsilon$ then STOP; otherwise return to step 2.*

It partitions the finite collection for n number of elements into the collection of c fuzzy clusters.

2.2 K Means Clustering

K-means is one of the simplest unsupervised technique where data is being classified into clusters on the basis of centroid where we give each cluster one centroid, i.e. defining k centroids one for each cluster. Placement of centroids is an important issue, different location causes different results so it is better to keep these centroids as far as possible. Then data elements are considered and associated to the nearest centroid. The k-means algorithm is as followed:-

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

3. GENETIC ALGORITHM

Genetic algorithms are used to control evolution which is based on some heuristic techniques. They are used to solve a problem when little information is known about a problem. Let us consider u be the set of sample points, u = {u1,u2,u3,….,un}. Then we can obtain optimal set of recombination and selection on the basis of analysis from genetic algorithms. Thus from above sample points we can get one set {u1,u2,….,ui} which have same properties as another set {uj,uk,……un}.

Thus, GA is based on five basic principles:

1. INITIALIZATION:

The initialization of GA is marked by the set containing the whole sample points i.e. the whole population. It may contain tables which are input to a database or some inputs of real life scenarios.

2. SELECTON :

Selection is based on Darwin's theory of "survival of the fittest". A subset is created from the set found from previous step. This subset helps in categorization of data. They contain data that is logically related to each other at some point of time. They may contain multiple sets with information such as domain data.

3. CROSS OVER/RECOMBINATION:

This method allows the cross over of a gene, an individual member of any set with a gene of another set. This recombination results in exchange of behavior and trends creating logical relation between set and thus helps to reduce the degree of randomness among sets.

4. MUTATION:

This technique allows us to generate genetic diversity among different genes. Cross over may cause inhibition of one's properties but mutation helps to maintain individuality among the genes.

5. ACCEPTANCE:

This technique allows the generation of new offspring, i.e. candidate for next step. But all candidates are not accurate for future steps ,Thus elimination occurs.

3.1 Big Data Analysis using the Genetic Algorithm:

Big data is defined as the datasets whose rate of increase is exponentially high within less time thus it is very difficult to analyze them using typical data mining tools. Thus the typical data mining tools needs to be replaced by some efficient and adaptive techniques in order to increase the degree of efficiency. When we use GA over data mining we obtain a great robust, computationally efficient and adaptive systems.

3.1.1 Calculation of Expectancy:

When the summation is finite and convergent the value of any random variable becomes consistent. Tensor is used in order to present the idea of convergence of any variable which is defined as a point in sample space that has a fixed value and direction but its direction is dependent on other point's direction. Thus, there exists a degree of interrelation between the random variables, which helps to the degree of randomness in some direction.

The inverse sigmoid equation is given as follows:

$$f(x) = {}^{1}/_{1+e^{x}} + {}^{1}/_{1+e^{-x}}$$

Here we find out two mapper functions:

1) GA Mapper 1: (sin $x*$ cos$x$)

2) GA Mapper 2: $\sum$ $(-1)(2n+1)$ $(2n)!$ $+$ $xlnx$ $+$ $ceil(sinx)$

3.1.2 Validation of Equation:

Initialization:

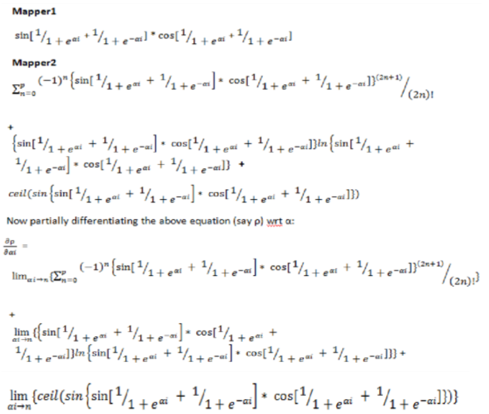
A base for validation is shown here before applying GA. The initialization is done using population gathering. Let us consider first set containing elements {x1,x2,……,xxn}. In worst case complexity each element is considered for final population, thus the final result become the superset of the first set. This shows that loop invariant concept holds just before GA starts.

Maintenance: There is a need to show consistency between ith and (i+1)th step in order to validate loop variant concept. At the i th step, then result is obtained from (i-1)th step.

The sigmoid: ${}^{1}/_{1+e^{\alpha i}} + {}^{1}/_{1+e^{-\alpha i}}$

After applying the sigmoid, a probability density function (pdf) is found with respect to all members of other respective set. Now we apply the respective map.

Termination: The termination is formed during last iteration. Consider this last iteration be nth, so we plot the ωn-2,n-1 and ωn-1,n values and check their continuity. Thus, we can form the validation of the equation easily. Thus this way helps to create a great degree of accuracy by analyzing big data and its concepts.



Mapper1

$$\sin[{}^{1}/_{1+e^{\alpha i}} + {}^{1}/_{1+e^{-\alpha i}}] * \cos[{}^{1}/_{1+e^{\alpha i}} + {}^{1}/_{1+e^{-\alpha i}}]$$

Mapper2

$$\Sigma_{n=0}^{p} {}^{(-1)^n \left\{\sin[{}^{1}/_{1+e^{\alpha i}} + {}^{1}/_{1+e^{-\alpha i}}\right] * \cos[{}^{1}/_{1+e^{\alpha i}} + {}^{1}/_{1+e^{-\alpha i}}]\}^{(2n+1)}}/_{(2n)!}$$

$$+$$

$$\left\{\sin[{}^{1}/_{1+e^{\alpha i}} + {}^{1}/_{1+e^{-\alpha i}}\right] * \cos[{}^{1}/_{1+e^{\alpha i}} + {}^{1}/_{1+e^{-\alpha i}}]\} ln \left\{\sin[{}^{1}/_{1+e^{\alpha i}} + {}^{1}/_{1+e^{-\alpha i}}\right] * \cos[{}^{1}/_{1+e^{\alpha i}} + {}^{1}/_{1+e^{-\alpha i}}]\} +$$

$$ceil(sin\left\{\sin[{}^{1}/_{1+e^{\alpha i}} + {}^{1}/_{1+e^{-\alpha i}}\right] * \cos[{}^{1}/_{1+e^{\alpha i}} + {}^{1}/_{1+e^{-\alpha i}}]\})$$

Now partially differentiating the above equation (say ρ) wrt α:

$$\frac{\partial \rho}{\partial \alpha i} =$$

$$\lim_{\alpha i \to n}\{\Sigma_{n=0}^{p} {}^{(-1)^n \left\{\sin[{}^{1}/_{1+e^{\alpha i}} + {}^{1}/_{1+e^{-\alpha i}}\right] * \cos[{}^{1}/_{1+e^{\alpha i}} + {}^{1}/_{1+e^{-\alpha i}}]\}^{(2n+1)}}/_{(2n)!}\}$$

$$+$$

$$\lim_{\alpha i \to n}(\left\{\sin[{}^{1}/_{1+e^{\alpha i}} + {}^{1}/_{1+e^{-\alpha i}}\right] * \cos[{}^{1}/_{1+e^{\alpha i}} + {}^{1}/_{1+e^{-\alpha i}}]\} ln \left\{\sin[{}^{1}/_{1+e^{\alpha i}} + {}^{1}/_{1+e^{-\alpha i}}\right] * \cos[{}^{1}/_{1+e^{\alpha i}} + {}^{1}/_{1+e^{-\alpha i}}]\}\} +$$

$$\lim_{\alpha i \to n}\{ceil(sin\left\{\sin[{}^{1}/_{1+e^{\alpha i}} + {}^{1}/_{1+e^{-\alpha i}}\right] * \cos[{}^{1}/_{1+e^{\alpha i}} + {}^{1}/_{1+e^{-\alpha i}}]\})\}$$

Fig 3.1 Mapper functions

4. ANT COLONY OPTIMIZATION ALGORITHM(ACO)

Data retrieval is the important task in the database. Cloud data retrieval is one of the major issue in the large databases available. Retrieving the data through the query language from the database is very difficult. There are many searching techniques used by the cloud servers for the retrieval of data. The optimization technique can be used for the retrieval of data. There are many data retrieval techniques like Boolean Symmetric Searchable Encryption, Secure Ranked Keyword Search over Encrypted Data and Privacy-Preserving Multi-keyword Ranked Search over Encrypted Cloud Data, Over Encrypted Data in Cloud Computing, etc. The ant colony optimization technique is used for solving the computational problems. The motive of the first algorithm was to find the optimized path through the graph the way the ants seek their path in the colony in the search of the source of food. In this algorithm there were several problems that occurred depending upon the behavior of the ants and the movement of the ants.

In the environmental aspect of the world when the ants find the way to the source of the food and there come the further more ants in search of the food they don't tend to follow the different path to reach the source the ants tend to follow the same path followed by others. But over time the pheromone trail starts to evaporate to reduce the length of the path then their comes the further more paths of the reduced length. And the ants tend to follow the path that is of the shorter length. the evaporation of the longer paths is also of advantage to the users because it helps in the easy calculation of an optimal shortest path. if there was no evaporation of the paths the ants would be attracted to follow the same path followed by the previous ones. The overall result of this is that the ant finds the good short path from the colony to find the source for food.

4.1 EXTENSIONS TO ANT COLONY OPTIMIZATION ALGORITHM:-

4.1.1 ELITIST ANT SYSTEM:

The global best solution deposits pheromone on every iteration along with all the other ants.

Max-min ant system (MMAS):

Added maximum and minimum pheromone amounts [tmax ,tmin]. Only global best or iteration best tour deposited pheromone <MAZ>. All edges are initialized to tmin and reinitialized to tmax when nearing stagnation.

Rank-based ant system :

In this the solutions are ranked according to the lengths of the path. The solutions with the shortest length are more followed then the paths followed of the longer paths.

Continuous orthogonal ant colony (COAC):

This is to enable ants to search for the solution collaboratively and effectively. By using an orthogonal design method, ants can explore the chosen regions rapidly and efficiently.

PSUEDO CODE AND FORMULA:-

```
procedure ACO_MetaHeuristic
while(not_termination)
generateSolutions()
daemonActions()
pheromoneUpdate()
end while
end procedure
```

EDGE SELECTION:-

An ant is an important agent in the ant colony optimization algorithm. The intermediate solutions are termed as the solution states. At every iteration of an algorithm, the ant moves from state x to state y.

APPLICATIONS:-

Ant Colony Optimization is the technique used to find the shortest path for routing vehicles. It is also used to design various algorithms to find the shortest path algorithms such as travelling salesman problem, The ant colony optimization continuously and adapt to changes in real time. This also helps in network routing and urban transportation system. The other applications are scheduling problem, assignment problem, set problem, image processing.

The first algorithm based on ant colony optimization is called as ant system and was used to solve the travelling salesman problem and the goal is to find the shortest path to make a round to the cities. At each stage the ant reaches every city following some rules.

1. It must visit each city exactly once;

2. A distant city has less chance of being chosen (the visibility);

3. The more intense the pheromone trail laid out on an edge between two cities, the greater the probability that that edge will be chosen;

4. Having completed its journey, the ant deposits more pheromones on all edges it traversed, if the journey is short;

5. After each iteration, trails of pheromones evaporate.

CONCLUSION:

In this paper we have discussed various techniques inspired from natural computing such as neural networks, fuzzy technique, genetic algorithm and Ant colony optimization technique for big analytics of various authors in which they have conducted experiments using these techniques for large data sets. Neural network is fast in computing small data sets/subsets as compared to computing the original datasets and by the use of network integration technique the results are found to be accurate. Fuzzy technique is helpful where large set of data need to be reduced and can further be improved by adding indexing for easy retrieval of information. Genetic algorithms are appropriate for such problems where outcome is very unpredictable and problem specification is difficult to formulate. The Ant colony algorithm is helpful in finding the shortest path to reach the destination. Thus out of these 4 soft computing techniques, neural network and fuzzy logic are hard to use as they don't provide any guidelines. Genetic algorithm and ACO can offer some significant improvements in accuracy. But the best technique so far used is of Genetic Algorithm as it provides evolutionary methods for effort estimation.

REFERENCES :

1. Munawar Hasan," Genetic Algorithm and its application to big data analysis" , International

Journal of Scientific & Engineering Research, January 2014, p1991-1996.

2. Kritsanatt Boonkiatpong and Sukree Sinthupinyo, "Applying neural networks on Large Scale Data", International Conference on Information and Electronics Engineering, 2011 ,p189-193.

3. https://deeplearning4j.org/neuralnet-overview#introduction-to-deep-neural-networks

4. http://www.aco-metaheuristic.org/

5. http://geneticalgorithms.ai-depot.com/Tutorial/Overview.html

6. K. Shibata and Y. Ikeda," Effect of number of hidden neurons on learning in large-scale layered neural networks", ICROS-SICE International Joint Conference, 2009, p5008-5013.

7. http://tranquy.info/what-are-machine-learning-deep-learning-and-how-they-are-difference/

8.S.Sangeetha, S.Kannimuthu and P.D. Mahendhiran, “Survey on Big Data Analytics and its Applications”, International Journal of Computer Applications(0975 - 8887) Volume 153-No 12,November 2016,p9-12.

9. http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html

10. http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html

11. https://www.mssqltips.com/sqlservertip/3132/big-data-basics--part-1--introduction-to-big-data/

12. S.Prabha and P.Kola Sujatha, “Reduction of Big Data Sets using Fuzzy Clustering”, International Journal of Advanced Research In Computer Engineering & Technology(IJARCET) Volume 3 Issue 6,June 2014,p2235- 2238.