RESEARCH REVIEW JOURNALS
Volume-04
Issue-04
April -2019

ISSN: 2455-3085 (Online)
RESEARCH REVIEW International Journal of Multidisciplinary
www.rrjournals.com [UGC Listed Journal]

Big Data Analytic Techniques Evaluation with Hadoop Framework for Health Care based Expert System

[1]Neha yadav, [2]P. Alagu Manoharan & [3]R. Radha Raman Chaudhary

[1]Research Scholar, Department of Computer Science & Engineering, Global Institute of technology & management (Affiliated by MDU University rohtak, HR) (India)
[2,3] Assistant professor, department of computer science & Engineering, Global Institute of technology and management (Affiliated by MDU University rohtak, HR)

ARTICLE DETAILS
Article History
Published Online: 15 April 2019

Keywords
Big Data, Hadoop, Map Reduce, Health care, Predicate analysis.

ABSTRACT
Research proposes Expert systems for Health care, to solve complex problems. To fulfill this objective Hadoop framework & big data analytic techniques has been considered capable. During medical treatment patterns such as insulin, serum, & plasma glucose concentration are required to test. The pattern discovery of predictive analysis consists of association rule mining, classification. Predictive analysis that is a method incorporating various techniques from data mining has been explained here. There are several research related to this field which are also discussed here. This research focuses on big data analytic techniques. Hadoop framework and health care based system are also discussed. The proposed work would be helpful in health care sector.

1. Introduction

Expert systems have been created first flourishing forms of artificial intelligence software. An expert system is a computer system. it emulate efficiency of decision making of a human expert. Expert systems have been formulated to face complex challenges. It has been done by reasoning through bodies of knowledge. Such knowledge has been represented mainly as if–then rules instead of through conventional procedural code. Expert system has been differentiating in inference engine & knowledge base. Knowledge base is representing facts & rules. On other hand, Inference engine is using rules to known facts to assume latest facts. Inference engines might have explanation & debugging abilities. Diabetes is a disease where body can't control level of sugars in blood. It talks about a group of metabolic diseases. Here individual has high blood sugar (blood sugar). It occurs because insulin production is insufficient. Other reason would be that sometime body's cells aren't responding correctly to insulin. You will find 3 major kinds of diabetes.

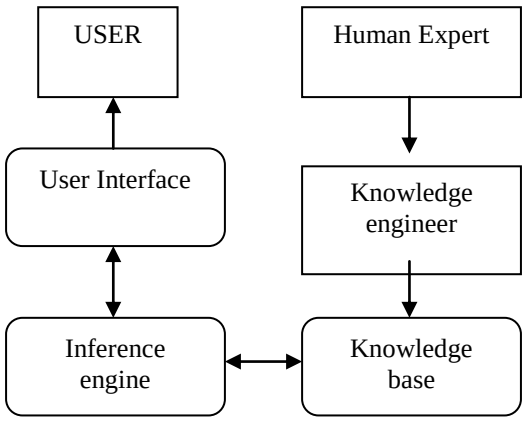


Fig 1 Expert System

2. Literature review

There are several researches in this field. In which some have been mentioned here:

In 2009, V. H. Bhat[1], et al. Introduced efficient Prediction Model for Diabetic Database Using Soft Computing Techniques,

In 2012, Abdullah A[2] ,et al. discussed applications of data mining for diabetes health care in young & old patients.

In 2012, K. Rajesh[3], et al. made research paper on Application of Data Mining Methods & Techniques for Diabetes Diagnosis.

In 2014, Sadhna[4], et al. did Analysis of Diabetic Data Set Using Hive & R.

In 2013, Sabibullah M[5], et al. proposed Diabetes Patient's Risk through Soft Computing Model.

In 2011, V. H. Bhat[6], et. al proposed Efficient Framework for Prediction in Healthcare

In 2012, Nishchol Mishra[7], did survey on predictive analytics, trends, applications, opportunities. They also considered different challenges in predictive analytics.

In 2014, Wullianallur Raghupathi[8], did Big data analytics in healthcare. They considered promise & potential of big data analytics.

3. Pattern discovery

It has been essential to test patterns for medical treatment. That may be like, plasma glucose concentration, serum insulin, diastolic blood pressure, diabetes pedigree, Body Mass Index BMI, age, time period of pregnancy. It is vital that pattern of discovery of predictive analysis must involve below given points [14]:

Association rule mining- Association between diabetic type & pages viewed for example laboratory results Clustering- clustering of same patterns of usage, etc.

Classification- categorization of health risk value by stage of patient health situation.

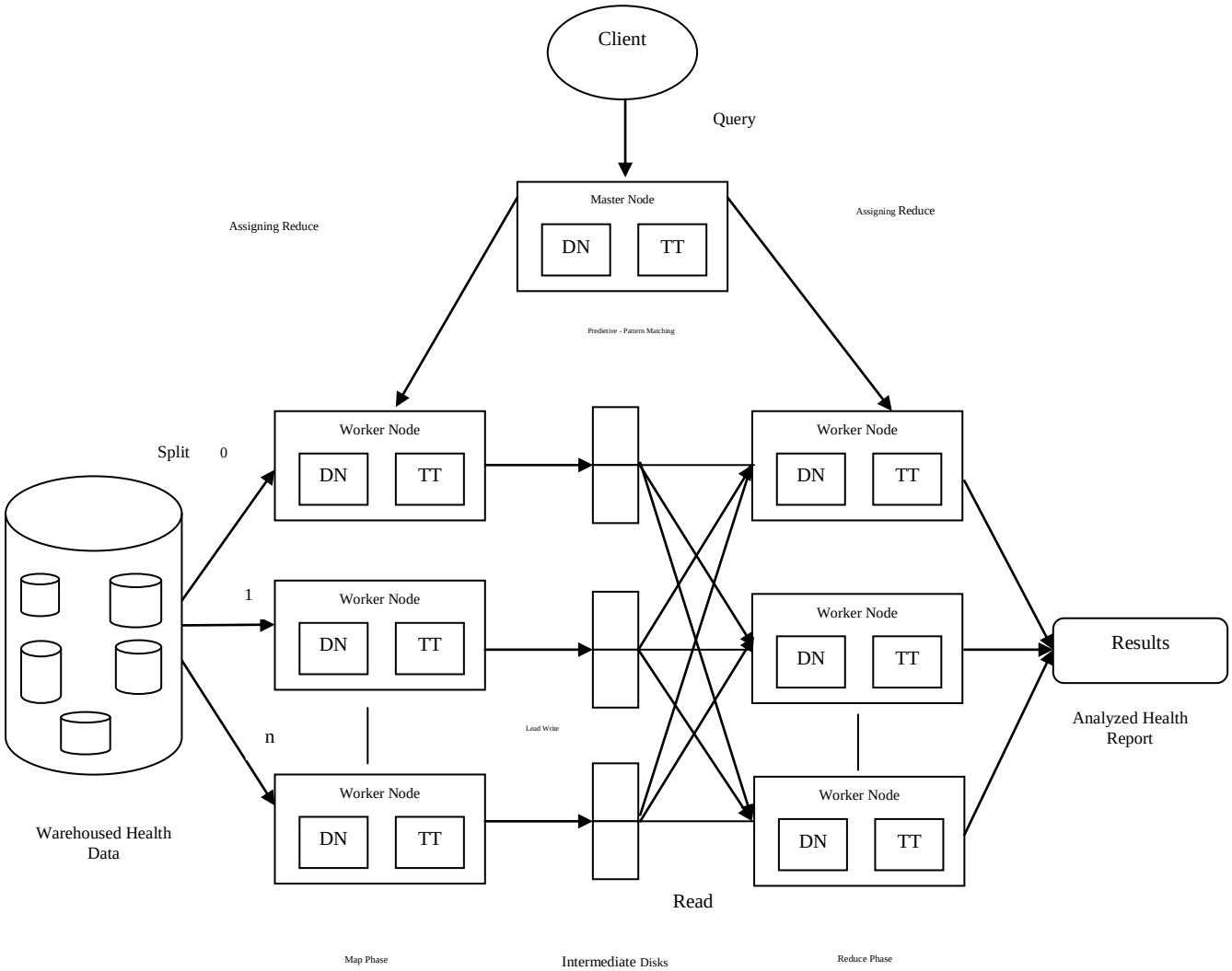


Fig 2 Health care model

When Hadoop system transfers dataset to warehoused, without delay map decrease task is executed. In mapping step, Master Node breaks huge data into smaller tasks for several Worker Nodes. Above figure deploys exact operation of predictive pattern matching system. The Master node is one consists of Name Node & Job Tracker, which always employs map & reduce task. The Worker Node or Slave Node receives order from Master Node, process pattern matching task for medical data with help of Data Node – Same Machine & Task Tracker. The predictive matching that is a procedure compares analyzed threshold value with obtained value. In condition of pattern matching procedure completion by all Worker Nodes that is depended on requirement, it was saved in midway disks. The whole procedure is called local write. It case of reduce task employment by Master Node, all other attached Worker Nodes will read processed data from midway disks. Based on query received from Client through Master Node, reduce task will be executed in Worker Node. The output got from reduce step will be distributed in several servers.

4. Motivation for research

To study existing techniques & framework & to find critical attributes which play major role in determining & predicting in advance possibility of diseases & its various stages, research work has been proposed. Map reduce in with predicates is applied in order to manage big data. There is requirement to use a better clustering technique to manage data. Along with this, there is need to implement & compare various clustering techniques.

5. Research gap

A literature review has revealed a few outputs on health completed different major data mechanism. The classical neural network design is utilized for prediction, on pre processed dataset. In predictive analysis of diabetic care using regression based data mining techniques to healthcare information, they realize patterns using SVM algorithm which identify best method of therapy for diabetes across various age. A comprehensive evaluation of Pima diabetic data set was performed effectively using of R and Hive. In this particular analysis we could possibly derive some fascinating facts, which may be employed to have prediction models. The gentle computing based prediction design was created for choosing risks built up by healthcare patients. They've experimented with real time medical data applying Genetic Algorithm. They obtained results pertaining to amount of danger that prone to both stroke and heart attack. A hybrid blend of Regression and Classification Trees (Genetic Algorithms and cart) to impute missing constant values & Self Organizing Feature Maps (SOFM) to impute categorical values was enhanced in.

Deploying a health info exchange (HIE) repository promote & integrate information within one point of strong information sharing.

6. Implementation

Map Reduction Implementation

MapReduce has been considered as a programming model. It is associated with implementation for processing & generating huge data sets. It works with a parallel, distributed algorithm on a cluster. A MapReduce program has been composition of a **map** procedure. It performs filtering as well as sorting just like sorting nutrient by first name into queues. Here each queue is for each name. A *reduce* method performs a summary operation like counting number of nutrient in each query presenting name frequencies. The MapReduce System is also called infrastructure. The code in c# has been written in order to get frequency of keywords in collected data. The input file for program would be keywords.txt & data.txt output file would be frequency.txt.

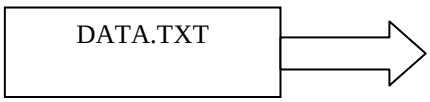
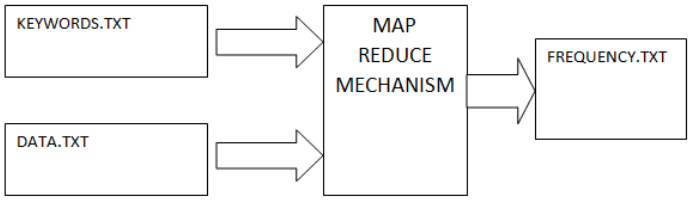


Fig 3 Process fl



And this csv file would be open in weka tool

Fig 4 Weka tool

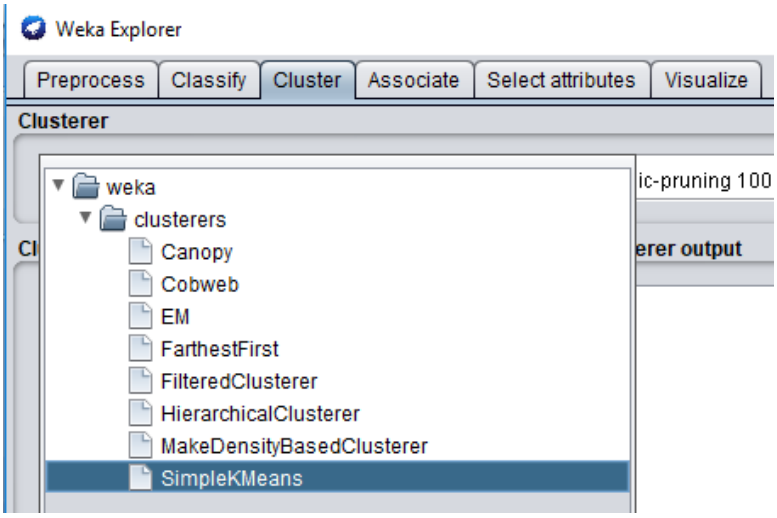
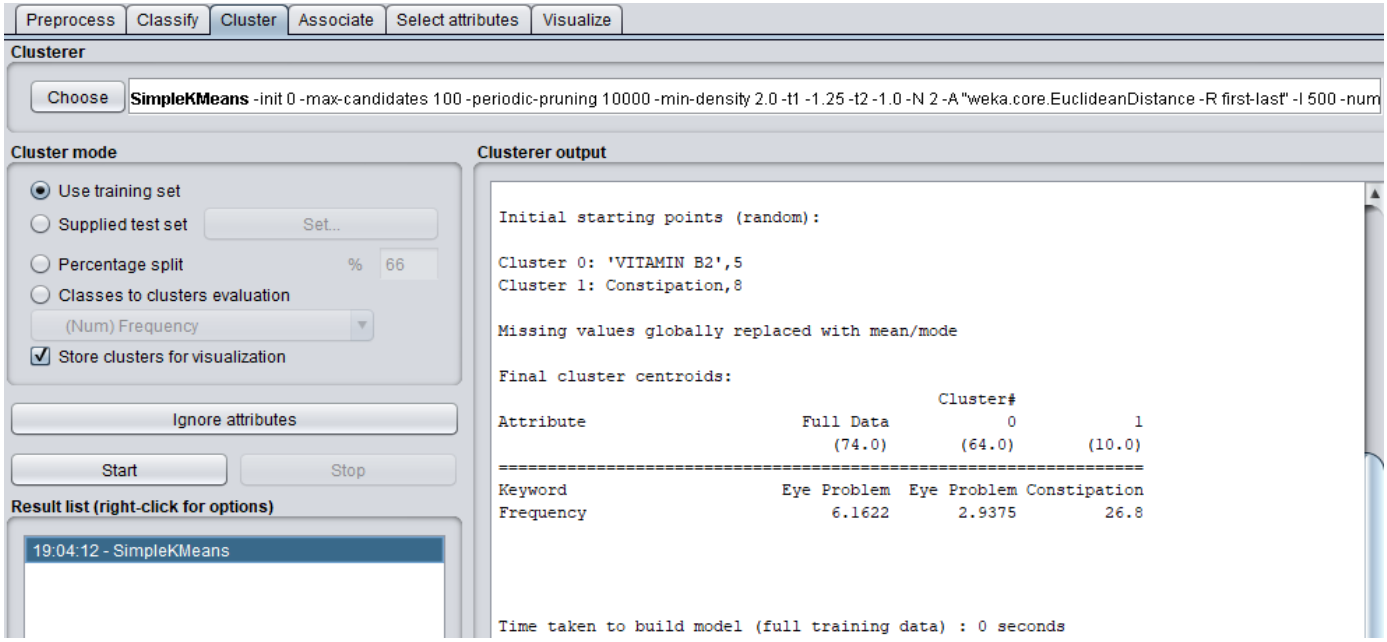


Fig 5 Kmean selection



K mean clustering using training set in weka tool.

Fig 6 Training set selection

OUTPUT:

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: frequency Instances: 74 Attributes: 2 Keyword Frequency

Test mode: evaluate on training data

=== Clustering model (full training set) === kMeans ======

Number of iterations: 4

Within cluster sum of squared errors: 71.19061909262763

Initial starting points (random):

Cluster 0: 'VITAMIN B2',5

Cluster 1: Constipation,8

Missing values globally replaced with mean/mode

Final cluster centroids:

		Cluster#	
Attribute	Full Data	0	1
	(74.0)	(64.0)	(10.0)
Keyword	Eye Problem	Eye Problem Constipation	
Frequency	6.1622	2.9375	26.8

Time taken to build model (full training data) : 0 seconds

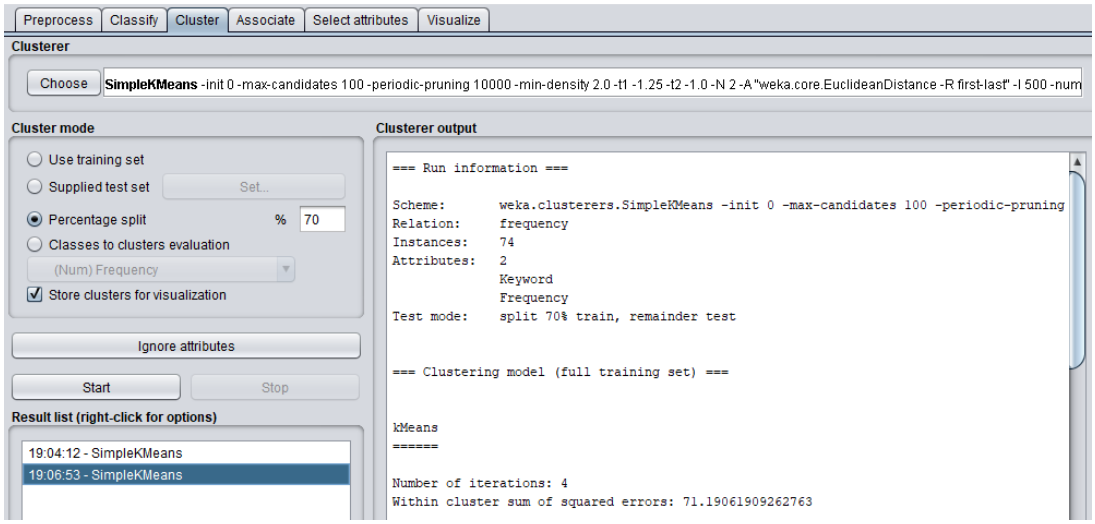
=== Model & evaluation on training set ===

Clustered Instances

0 64 (86%)

1 10 (14%)

Kmean clustering using 70% percentage split



Output:

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: frequency

Instances: 74

Attributes: 2

Keyword

Frequency

Test mode: split 70% train, remainder test

=== Clustering model (full training set) ===kMeans

Number of iterations: 4

Within cluster sum of squared errors: 71.19061909262763

Initial starting points (random):

Cluster 0: 'VITAMIN B2',5

Cluster 1: Constipation,8

Missing values globally replaced with mean/mode

Final cluster centroids:

		Cluster#	
Attribute	Full Data	0	1
	(74.0)	(64.0)	(10.0)
Keyword	Eye Problem	Eye Problem	Constipation
Frequency	6.1622	2.9375	26.8

Time taken to build model (full training data) : 0 seconds

=== Model & evaluation on test split ===kMeans

Number of iterations: 8

Within cluster sum of squared errors: 47.7993538560205

Initial starting points (random):

Cluster 0: Dither,0

Cluster 1: fainting,5

Missing values globally replaced with mean/mode

Final cluster centroids:

		Cluster#	
Attribute	Full Data	0	1
	(51.0)	(44.0)	(7.0)
Keyword	Eye Problem	Skin Disease	Eye Problem
Frequency	5.1569	3.0227	18.5714

Time taken to build model (percentage split) : 0 seconds

Clustered Instances

0 19 (83%)

1 4 (17%)

7. Scope of research

The research work would offer study existing techniques & framework. It also finds critical attributes which play major role in determining & predicting in advance possibility of diseases & its various stages. Map reduce in with predicates has been applied in order to manage big data. There would be requirement to use a better clustering technique to manage data. Along with this, there would be need to implement & compare various clustering techniques.. It would be helpful to develop data mining methods & techniques for healthcare. It would be useful to take significant predictions or decisions by employing big data analytics in diabetic field.

Reference

1. V. H. Bhat, P. G. Rao, & P. D. Shenoy, “An Efficient Prediction Model for Diabetic Database Using Soft Computing Techniques,” Architecture, Springer-Verlag Berlin Heidelberg, pp. 328-335, 2009..
2. Abdullah A. Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqui, “Application of data mining: Diabetes health care in young & old patients”, Journal of King Saud University – Computer & Information Sciences, vol. 25, pp. 127–136, 2012.
3. K. Rajesh, V. Sangeetha, “Application of Data Mining Methods & Techniques for Diabetes Diagnosis” in International Journal of Engineering & Innovative Technology (IJEIT) Vol 2(3), 2012.
4. Sadhana, Savitha Shetty, “Analysis of Diabetic Data Set Using Hive & R”, International Journal of Emerging Technology & Advanced Engineering, vol 4(7), 2014.
5. Sabibullah M, Shanmugasundaram V, Raja Priya K, “Diabetes Patient’s Risk through Soft Computing Model”, International Journal of Emerging Trends & Technology in Computer Science, vol 2(6), 2013.
6. V. H. Bhat, P. G. Rao, S. Krishna, & P. D. Shenoy, “An Efficient Framework for Prediction in Healthcare,” Most, Springer-Verlag Berlin Heidelberg , pp. 522-532, 2011.
7. Nishchol Mishra, Dr.Sanjay Silakari, “Predictive Analytics: A Survey, Trends, Applications, Oppurtunities & Challenges”, International Journal of Computer Science & Information Technologies, vol. 3(3), 4434- 4438 4434, 2012.
8. Wullianallur Raghupathi, & Viju Raghupathi, “Big data analytics in healthcare: promise & potential”, Health Information Science & Systems, vol. 2(3) pp. 2-10, 2014.
9. Mansoor Khan,” Managing Vulnerable Populations: How Predictive Analytics is a Key Component for Understanding User Behavior & Improving Care Quality”, Managed Care Outlook, vol 27(4), February 15, 2014.
10. Andrew Pearson, Qualex Asia, ”Predictive Analytics for Healthcare Industry”, Andrew Pearson, Qualex Asia Limited, 2012.
11. http://www.intel.com/content/www/us/en/healthcare-it/bigger-data-better-healthcare-idc-insights-whitepaper.html
12. D. Peter Augustine, “Leveraging Big Data analytics & Hadoop in Developing India’s Health Care Services”, International Journal of Computer Applications, vol 89(16), pp 44-50, 2014.
13. Muni kumar N, Manjula R,”Role of Big Data Analytics in Rural Health Care – A Step Towards Svasth Bharath”, International Journal of Computer Science & Information Technologies, vol 5(6), pp 7172-7178, 2014.
14. Andre W. Kushniruk, “Predictive Analytics & Forecasting in Health Care: Integrating Analytics with Electronic Health Records”, SAS Institute Inc, 2008.

