International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 17 (2015) pp 37578-37583

Distributed Machine Learning Based Biocloud Prototype

Mansaf Alam
Department of Computer Science, Jamia Millia Islamia, New Delhi malam2@jmi. ac. in

Shuchi Sethi
Department of Computer Science, Jamia Millia Islamia, New Delhi shuchi. sethi@yahoo. com

Kashish A Shakil,
Department of Computer Science, Jamia Millia Islamia, New Delhi shakilkashish@yahoo. co. in

Abstract
Our proposed biocloud is a novel framework to analyze large biomedical data using latest cloud based distributed processing. Data capturing with scalable storage and distributed big data processing system for which authors have build a prototype allowing tremendous advantages to biomedical community in the form of accuracy, fast and efficient processing at a low cost with flexibility of scaling upto any degree. The framework we propose is based on map and reduce approach which is typical to cloud.
In this manuscript we code in Ruby for feature extraction and Hive to manage data using Hadoop distributed File System. Then all the data is analyzed using DBDP algorithm for distributed classification of extracted features. The major contributions include a framework specific to biomedical com- munity while being general enough to be applicable to any disease prediction, besides in DBDP algorithm unique mapping and reducing strategy so that balanced training is achieved and hypothesis is free from any bias.
Keywords: Wireless Sensor Network, Sensors, Named Data Networking, Decision coordination Priority, Wireless recharging, Energy replenishing.

Index Terms — Bio-Cloud, Cloudcomputing, Dengue, Distributed machine learning

Introduction

To design an architecture that efficiently supports diagnostics and prediction on cloud requires to overcome numerous challenges like cost, security and more. We leverage cloud computing scalability, storage and distributed processing to tackle the cost issues. As a case we apply distributed classification algorithm to obtain dengue predictions using patient's symptoms as parameter. This is an important step in reaching out to masses and making ubiquotious healthcare a reality.
The major challenges that exist in current analysis of biomedical data that are overcome by our proposed framework are mentioned:

Scalability:

To handle a large number of patients, a need for huge storage and reliable framework are prerequisites. Bio-Cloud provides horizontal scalability by being cloud enabled framework and ability to expand its resource pool as number of patients increase. This scalability is specifically useful in contagious diseases that have the tendency of becoming an epidemic where rise in number of patients may be exponential and traditional infrastructure will not be able to handle this burst requirement.
Missing values in biomedical data:
As data obtained by patients is usually incomplete and removing that data leads to creeping of inaccuracy and this being a typical feature of big data, Hadoop allows such data to be stored and processed which is a compelling reason to use Hive for preprocessing layer[3].

Economy

Biocloud as a service will offer low cost solution to enable this service availability in remote areas where diagnostics and predictive analysis are otherwise not affordable. In cloud network, storage, software service and monitoring costs are involved. Thus to make quality services available to users at low cost, it is important to maximize judicious use of resources at disposal and thats where middleware comes into picture by being a miser and distributing resources in most balanced way possible. Some amount of research has been done on ways of allocating resources [11], [12] and in future one might come up with new science of best ways of resource sharing.
Besides overcoming the challenges our framework supports green computing by minimizing the time complexity and providing meaningful results timely. The remainder of the paper is organized in the following manner: Related work section discusses all the efforts in biomedical predictive analysis of data. Next section is framework which explains architecture of our proposed framework in detail. Section 4 is a case study of Dengue disease which gives technical details in depth. Section 5 gives biocloud prototype in detail and the last section being the algorithm implementation for obtaining results and comparison is discussed. Dengue as a case study is chosen because it is life threatening disease whose outbreak can be controlled with the use of technology[7]. Dengue viruses have spread rapidly across regions in past few decades leading to an increased frequency of epidemics. It is regarded as the most rapidly spreading mosquito-borne viral disease in humans. Due to this reason an upsurge in research on dengue virology, immunology has been done in the past decade. Also there has been development of antivirals, vaccines and lot of research pertaining to this in biomedical fraternity.

Abbreviations used SVM: Support vector Machines
SV: Support vector U: Union

Related Work

Some research has been done in applications of cloud in biomedical field. Dr Buyya et al. [1]have designed a real time health monitoring system which has been discussed with ECG as a case study. Researchers have discussed especially those cases where scalable and economic monitoring can be provided to those who require these services very frequently. In this sensor data flows into the system where it is monitored for abnormality and if found it is reported.

In [4] researchers have designed bionimbus cloud management system for sharing and analyzing large gnomeics dataset. The purpose is to allow huge dataset processing using cloud. It is primarily based on OpenStack and has high performance cluster file system. It is useful to projects that require managing massive biomedical data.

In [6] researchers have discussed how the biomedical data processing faces issues like big data volume, highly intensive computation and high dimension. Cloud can solve the issue by its unique ways of resource allocation, data storage, computation and sharing. This paper first analyzes features of the cloud computing for processing biomedical data. Then some solutions of biomedical cloud are summarized. But it mainly gives an overview of how cloud is applicable to biomedical community in general and not meant for specific study.

In [8] researchers have focused on handing heterogeneous physiological data which requires cloud to handle multimodal, non stationary characteristics of physiological signals like in ECG, HBP, PPG. This healthcare system is advanced and sends out messages to patients if any abnormality in testing is found. But it requires high cost infrastructure whose implementation is a huge task in itself for developing economies. They have proposed a six layer architecture including a separate storage and service layer and cloud engine. Medical data mining algorithms help process raw signals from and to cloud storage.

Our work is related to above but our focus is use of distributed technology in helping early, accurate prediction of epidemic disease. Some other research in past have proposed an architecture for mobile management of chronic conditions and medical emergencies but those proposal have certain issues pertaining to time complexity. We have not discussed traditional approaches used for analysis here as authors believe that cloud takes care of most of the traditional issues except privacy and security which has been mentioned in conclusion section in brief.

Framework

The architecture of our proposed framework is composed of 3 layers.

At First Layer are the end users which includes entire biomedical community and researchers engaged in bioinformatics research. Our framework will be helpful to them in following manners:

1) Fast and Efficient processing of biomedical data.
2) Highly Accurate Diagnostics of diseases.
3) Early prediction of the same to prevent a disease from becoming an epidemic.
4) Distribution of results from anywhere in the world with ease.

Layer two is the Interface of our proposed Biocloud. This layer is discussed in prototype section. Purpose of this interface is to make the analysis and information capturing as user friendly as possible.

Layer three is preprocessing layer composed of three modules:

Module 1: Distributed processing using Hive based on HDFS
Module 2: Ruby is used for feature extraction in biomedical analysis.
Module 3: Classification of data using distributed approach using DBDP(Distributed BioMedical Data Prediction) algorithm.

To explore layer 3 in detail, its purpose is to be able to determine whether patients are infected with disease on large scale, all features are extracted from data using Hive which processes a query using Hadoop distributed file system. This parallel feature extraction in Bio cloud as service in figure [1] reduces time and space complexity of the algorithm as the matrix that will be obtained later for optimization will have much lower dimensions.

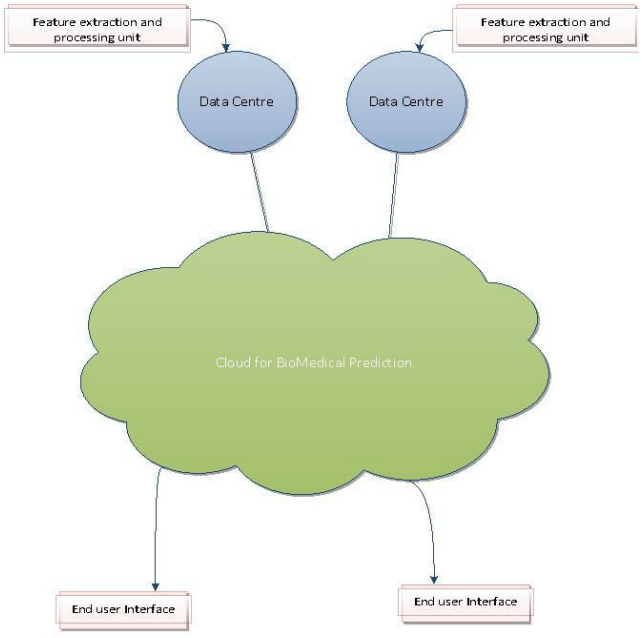

Fig. 1. Modules of Bio-Cloud as a Service

The query results are then pulled by Ruby for further processing. Ruby will extract relevant information for preprocessing the data. This is important layer (Preprocessing Layer) which if separate from processing will act as a controlling mechanism for Big data in Bio cloud and

different types of data can be just standardized with minor changes in this layer. This will introduce flexibility in our model. We used ruby on cloud based VM where certain features were reduced that were colinear and showed high dependency on each other by replacing then by a single feature and ignoring the symptom that is not related to the disease prediction. The code was written in ruby after obtaining such analytics about features. This works as cushion for different predictions by just changing the features in this small module, we have created another abstraction layer in the framework. The debian system based VM allowed vi editor to write code using Ruby and interpreter compiled and displayed using simple string manipulation functions. Use of Ruby made the code very precise and easy besides its unique quality of making the code DRY.

The preprocessed data then flows into the next module which is Distributed Parallel training algorithm as shown in figure [2]. The results are thus obtained and well formed hypothesis can be used for future predictions of disease by entering information about patient just by using a terminal sitting at home which has been implemented by us and in next sections we will be explaining our prototype.

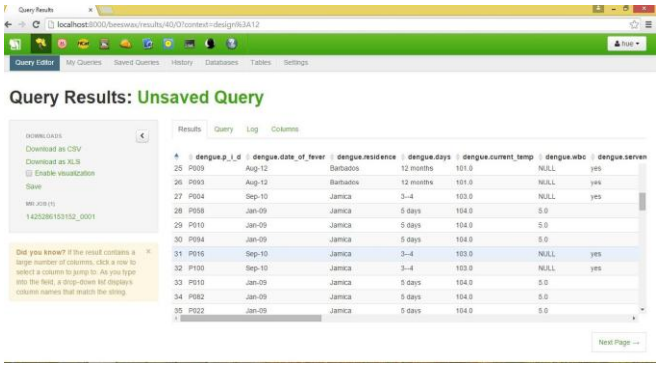

Fig. 3. HDFS based SQL: HIVE

A. HDFS

At the preprocessing layer Hive works to capture and create structure for cloud. When data is stored in cloud based distributed hadoop based file system, certain advantages are offered making Hive a good choice. The Apache Hive data warehouse software helps in querying and managing large datasets stored in a distributed manner. Hive provides a mechanism to project structure onto data and query the data using a SQL-like language. It is built on top of hadoop and helps in data summarization, query, and analysis.

B. DBDP algorithm

The SVM training is used parallely by distributing the big data in row order and separate clusters on cloud will process small datasets saving on time and storage. DBDP algorithm is specifically designed for Offering Bio-Cloud as a service. This uses the available resources to train fast and especially useful when a disease spreads fast and huge amount of patient data pours in on cloud which couldnt have been handled by a traditional algorithm. The algorithm is given below.

The algorithm first partitions the set into parts such that positive and negative training examples have equal number. This is done to neutralize the effect of imbalanced training leading to higher accuracy in training. SVM is a natural choice for biomedical framework as it is highly accurate and low cost algorithm. Though computation time is an issue when data exceeds a threshold but this will be taken care by our distributed algorithm. Our algorithm makes training feasible in cases where SVM traditionally is known not to be able to converge by optimization. Then training is done using SVM for 3 parts of the dataset while 2 are reserved for testing. This may be repeated for cross validation. After obtaining the hypothesis we take union of alphas and make sure that only positive alphas are included. Based on that hypothesis we test our 2 test parts of dataset.

Algorithm 1 DBDP Algorithm
Procedure: DISTRIBUTED PROCESSING OF DENGUE DATASET

[1] var i=5
[2] group i partition dataset s. t. n +ve and n -ve
[3] Loop i = i+1 till i =5
[4] For i=3 repeat next step
[5] distributed SVM training: mapping
[6] End Loop
[7] SVU=SV1 U SV2 U... SV5
[8] alpha g. t. 0
[9] Classify group4 and group5 with SVU
[10] close;

CASE STUDY: DENGUE

Dengue virus infections can range from being asymptomatic infection to dengue fever (DF). DF is characterized by high fever, severe headache, myalgia, arthralgia, pain and rash. There is a rising need for inexpensive dengue diagnostic tests that can be used for solitary diagnostics case to outbreak investigations. This will help in controlling epidemics. Use of Biomarkers to indicate the probable presence of this virus has to some extent contained the disease. But still it is far from being totally controlled. The issues in detection are posed by fact that optimal window for diagnosing a dengue infection is from the onset of fever to 10 days post-infection, one issue is if number of patients are very high and its going epidemic it would be hard to diagnose in given time frame and secondly if DF is asymptomatic.

The framework when applied to the dengue prediction gave excellent results. With ease we were able to validate the advantages that our Layer3 offers over the existing

Techniques used in diagnostics and prediction. Here we discuss in detail how our experiments were carried out on dengue dataset. The figure shows some of the features that we originally had for dengue dataset. There were more than 15 dimensions in dengue dataset that included date of fever, residence, days of fever, current temperature, wbc, severe headache, pain behind eyes, joint muscle ache, metallic taste in mouth, appetite loss, nausea and vomiting, diarrhea, hemoglobin, hematocri, platelet. After analysis of the

features we implemented our code in Ruby language to extract the most relevant two features which did not have any colinear dependence and had direct relation with dengue one being platelet. Then after prerocessing was complete the system is trained using DBDP algorithm. This trained algorithm can further be used for predictive analysis of dengue especially in areas where history of dengue outbreak exists.

The major issue arises when dengue is absolutely asymptomatic. For such cases we have introduced flexibility of additional module in our layer where instead of symptoms Biomedical researchers can use Biomarkers to predict the outbreak or presence of dengue in advance. This will provide low cost solution to prevent dengue becoming an epidemic in remote areas.

BIOCLOUD PROTOTYPE

Technologies used in obtaining the prototype are java and oracle. This has been tested to run on different types of machine like windows server, windows i3 dual core. Also it was test run at VMs with small configurations to test its suitability on cloud and ability to capture data at remote places by requiring low bandwidth, it becomes suitable for countries with network issues. In figure 4 we show the welcome screen. it shows 2 symptoms on basis of which one can decide the approach to follow. If none of 2 parameters exist, course choice needs to be made whether checking the patient for biomarkers should be done else if the two symptoms appear to exist, then whether other symptoms may be captured before applying the hypothesis. the figure shows the interface where the two symptoms are entered. This will pass the flow to layer 3 for further processing after making decision of choice. If choice 1 is made then a separate set of parameters can be extracted using module 1 and 2, which are part of preprocessing, of layer3 else the same parameters will be chosen as we have discussed in our case study in detail. Module 3 of layer 3 will then do processing.

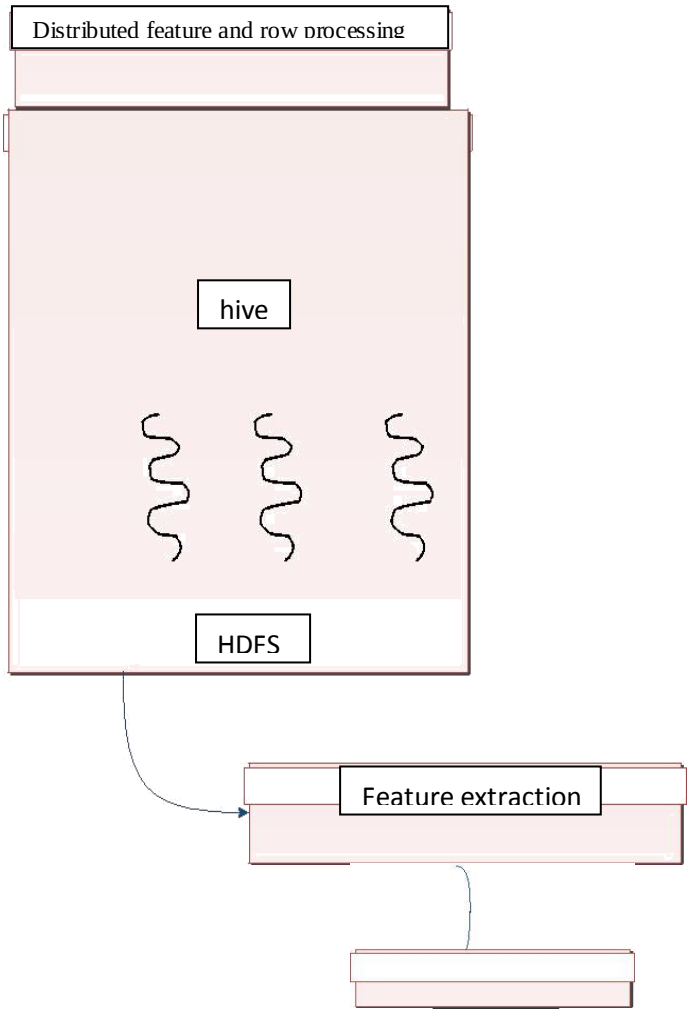

Fig. 2. Distributed Processing of Features and Rows in Bio-Cloud

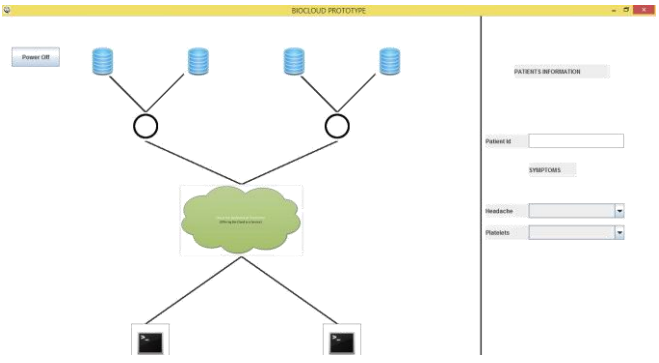

Fig. 4. A SNAPSHOT OF OUR PROTOTYPE

RESULTS AND DISCUSSION

Extensive use of tools was done to validate our framework. Some of the tools used in experiment are given in table below. Primary data collection was done using the prototype discussed in the section above. After preprocessing, data is used in training. Scripting of DBDP algo was done in matlab 2013a by distributing the data on different VMs and results were obtained. Also for comparison purpose, all dataset was

trained using linear SVM with soft margin to achieve accuracy. Besides we applied 10-fold cross validation in single machine training. The algorithm discussed above is based on map reduce approach. The results obtained are shown in figures below.

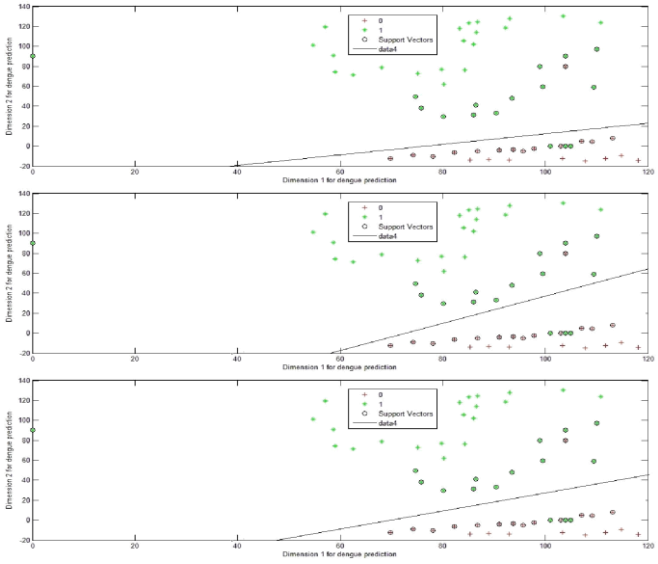

Fig. 5. Figure of distributed algo results

Two control classes are used. After comparing the results of DBDP and vanilla SVM classification we observed that error rate and correct rate were very similar. But major advantages were achieved when big dataset wasnt pro-cessed by vanilla. Variable cp captured class performance and cp. ErrorRate and cp. CorrectRate are displayed in table below for different dataset sizes. Not only the accuracy but DBDP has scalability while regular algorithms cannot handle the 3Vs of Big data, those are volume, velocity and variety while DBDP has proved its ability to handle volume of data as well as we tested it on all the features we had mentioned above. The algorithm was able to converge very fast.

Table I: Characteristics of computing resources used in the experiment

HardWare	SoftWare
Windows Server 2008	5 VMs with Windows OS with expandable RAM
4 CPU cores	Ruby language for feature extraction
32 GB RAM	DBDP algorithm coded in Matlab 2013a

Table II: Comparison Table(Rounded Values)

Dataset	DBDP CorrectRate	DBDP ErrorRate	SVM CorrectRate	SVM ErrorRate
100	1	0	1	0
500	1	0	. 99	. 01
7000	. 90	. 1	-	-

Conclusion

Cloud is a technology meant for biomedical applications to make services available to all at low cost and in remote ares. though certain issues do remain that need to be addressed, most important of which is security and privacy of patient data. Number of researchers[2], [5], [9], [10] are working in this area to provide technical security and privacy but besides that authors believe that SLAs have to be made strong. Service level agreements have to be bound on the provider so that privacy of medical data is not compromised. This will be future direction of our work. Also to improve upon our work in prototype, we plan to implement in rails as it is more scalable and light on web and appropriate for a Saas app than our existing prototype and some work is already in place in this direction.

References

[1] Pandey, Suraj, et al. ”An autonomic cloud environment for hosting ECG data analysis services. ” Future Generation Computer Systems 28. 1 (2012): 147-154.

[2] Beck, Martin, et al. ”GeneCloud: Secure Cloud Computing for Biomedical Research. ” Trusted Cloud Computing. Springer Interna-tional Publishing, 2014. 3-14.

[3] Feng, Mengling, et al. ”Management and analytic of biomedical big data with cloud-based in-memory database and dynamic querying: a hands-on experience with real-world data. ” Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014.

[4] Heath, Allison P., et al. ”Bionimbus: a cloud for managing, analyzing and sharing large genomics datasets. ” Journal of the American Medical Informatics Association 21. 6 (2014): 969-975.

[5] Dove, Edward S., et al. ”Genomic cloud computing: legal and ethical points to consider. ” European Journal of Human Genetics (2014).

[6] Lei, Sun, et al. ”Cloud computing solutions for processing biomedical data. ” Journal of Electronic Measurement and Instrumentation 11 (2014): 004.

[7] Guzman, Maria G., and Eva Harris. ”Dengue. ” The Lancet (2014).

[8] Li, Ye, et al. ”HCloud, a Healthcare-Oriented Cloud System with Im-proved Efficiency in Biomedical Data Processing. ” Cloud Computing with e-Science Applications (2015): 163.

[9] Alam, Mansaf, and Shuchi Sethi. ”Covert Channel Detection Tech-niques in Cloud. ” (2013): 3-02.

[10] Sethi, Shuchi, Kashish Ara Shakil, and Mansaf Alam. ”Seeking Black Lining In Cloud. ” arXiv preprint arXiv: 1501. 04473 (2015).

[11] Alam, Mansaf, and Kashish Ara Shakil. ”An NBDMMM Algorithm Based Framework for Allocation of Resources in Cloud. ” arXiv preprint arXiv: 1412. 8028 (2014).

[12] Alam, Mansaf, and Kashish Ara Shakil. ”Recent Developments in Cloud Based Systems: State of Art. ” arXiv preprint arXiv: 1501. 01323 (2015)