Big-Data Analytics, Machine Learning Algorithms and Scalable/Parallel/Distributed Algorithms

Anindita Desarkar and Ajanta Das

Abstract Smart data analysis has become a challenging task in today's environment where disparate data set is generated across the globe with enormous volume. So there is an absolute need of parallel and distributed framework along with appropriate algorithms which can handle these challenges. Various machine learning algorithms can be deployed effectively in this environment as they can work with minimal manual intervention. The objective of this chapter is first to present various issues faced in storing and processing big data and available tools, technologies and algorithms to deal with those problems along with one case study which describes an application in healthcare analytics. In the subsequent section it discusses few distributed algorithms which are widely used in the data mining domain. Finally it focuses on various machine learning algorithms and their roles in the big data analytics world.

Keywords Big data analytics · Machine learning · Distributed algorithms · Parallel algorithms

1 Introduction

Big data analytics is the methodology of processing and finding hidden patterns, unknown correlations, market trends and other useful business information from large volume of data sets consisting of heterogeneous data types, coming from various sources across the globe. Based on this analytical findings, more business

A. Desarkar · A. Das (✉)
Department of Computer Science & Engineering, Birla Institute of Technology, Mesra, Ranchi, 1582 Rajdanga Main Road, 4th Floor, Kolkata Campus, Kolkata 700107, India
e-mail: ajantadas@bitmesra.ac.in

A. Desarkar
e-mail: aninditadesarkar@gmail.com

C. Bhatt et al. (eds.), *Internet of Things and Big Data Technologies for Next Generation Healthcare*, Studies in Big Data 23,
DOI 10.1007/978-3-319-49736-5_8

growth can be achieved within a short period of time. Basic analytical methods and reporting tools which are working on calculating sums, counts, averages, execution of SQL queries depending on human intervention for the performed activities. This type of human dependency is a great challenge in the domain of big data where velocity, variety and volume are major concerns.

On the other hand, Internet of Things (IoT) which will be the next technological revolution is basically the other side of the coin where big data resides in one side. Basically IoT is the concept where every object or devices should have built in sensors to capture data across a network. So managing this huge amount of data, heterogeneous in nature is a huge challenge facing by all organizations. A proper analytics platform or framework is highly required for this data management and taking actions on it. These actions include event co-relation, metric calculation, and statistics preparation along with analytics and can vary depending on scenario.

Machine Learning comes into the picture which is perfect for exploiting the hidden knowledge within this large volume of distinct dataset with little reliance on human direction. It learns based on available data inputs and/or outputs, basically data driven and runs at machine scale, capable of handling huge variety of variables as well as data complexity which is quite essential in today's big data world. Machine learning consists various data analysis disciplines, starting from predictive analytics and data mining to pattern recognition and various algorithms are used for these purposes.

Big data analytics gained the momentum over traditional business intelligence program for its unique capability to deploy ideas into solutions, adapt with the changed environment along with its flexible nature. As a result, newer class of technologies that includes Hadoop and related tools such as YARN, MapReduce, Spark, Hive and Pig as well as NoSQL databases have been introduced. The current trend is using Hadoop as distributed data management system which has a flexible data storage mechanism for storing heterogeneous voluminous data coming across various sources. MapReduce is used as a parallel programming model in Hadoop for large scale data processing using commodity clusters.

Centralised healthcare monitoring and clinical analysis for caregiver or medical practitioner becoming a challengning issue. Healthcare monitoring system is based on lots of wearable body sensors, convenient handheld devices and broadband wireless services [1]. Electronic health records (EHR) includes various signal processing and time series data. So these different types of sources generate huge data, which turns into big data. Thus big data machine learning approaches need to be applied to provide clinical care. In this arena high performance computing or parallel/distributed algorithms ensures efficient storage and data retrieval for heterogeneous medical images or data [2].

The primary goal of this chapter is to discuss various machine learning algorithms along with their implementation roadmap in the analytics domain. As a first step, it discusses various challenges faced to handle big data and the mitigation techniques to overcome the same which includes common tools and technologies available in the market. It also includes various existing applications along with a proposed one in the healthcare domain which describes the importance of analytics

in our day to day life. There are various data mining algorithms which are basically machine learning algorithms, briefly described in the next section. The next section presents various types of machine learning methodologies includes supervised learning, unsupervised learning and reinforcement learning and few algorithms under each bucket. It also gives the brief description of the application where these algorithms are implemented successfully.

The structure of the chapter is organized as follows. Section 2 presents various challenges in big data Analytics and their mitigation techniques by applying various tools and technologies along with various applications in the healthcare domain which uses analytics for providing an effective and efficient solution. A case study on Healthcare Analytics is also included in this section. Detailed description of different Parallel/distributed algorithms and their role in big data analytics are described in Sect. 3. Section 4 focuses on various machine learning algorithms and their application in analytics domain. Section 5 concludes the chapter.

2 Big Data Analytics: Challenges and Mitigation Techniques

Big data is large amount of data, may be structured, semi-structured or unstructured in nature, generated from various sources across the globe. One major source is definitely Internet of Things (IoT) data which the IoT connected devices will produce.

There are three 'V's—Volume, Variety and Velocity which describes the characteristics of Big Data. It's produced in large volume from various sources across the world. Variety describes its heterogeneous sources and multiple data formats. Velocity is all about data streams in at an unprecedented speed from various sources. Another important concern about Big Data is, it's difficult to capture, process and manage by traditional software tools is a cost effective manner. In today's era of big data, it has become a real challenge to extract meaningful insights by applying traditional algorithms/methods from unstructured, imperfect and complex dataset in almost all the domains like Environmental study, biomedical science, Engineering etc. The challenges include understanding and prioritizing relevant data from the huge set, extracting data from master set where 90 % data reflects noise, security threat, costly tools and framework etc. So various innovative tools, technologies and frameworks have been developed to handle these challenges which includes Hadoop a distributed file system and framework for storing and processing huge amount of dataset using the MapReduce programming paradigm, different NoSQL data stores with flexible schema pattern, several big data analytics tool like Pentaho Business Analytics etc. This chapter describes these various tools and technologies in detailed fashion. It also gives the architectural advantage of these advanced technologies over the traditional ones.

2.1 Defining Big Data Analytics

Big data Analytics is the method of processing big data which is huge in volume and containing heterogeneous data types to discover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information with the help of set of innovative tools and technologies.

2.2 Challenges

There are lots of challenges which need to be taken care in a proper way for successful implementation in big data and Analytics. Figure 1 shows the percentage of various challenges in big data and analytics world which came as a survey result. Few major challenges are described below.

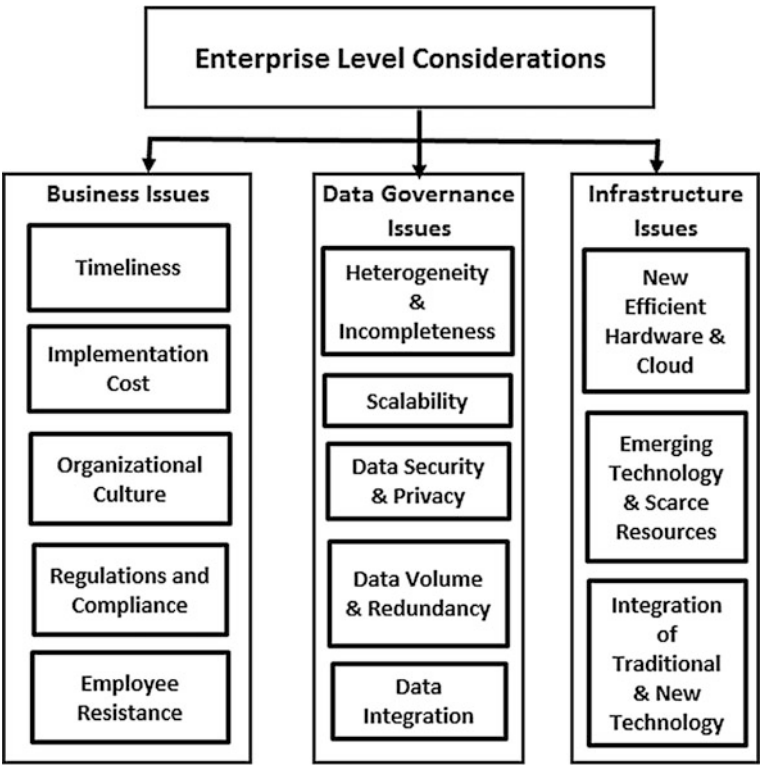

Fig. 1 Biggest challenges for success in big data and analytics

2.2.1 Heterogeneity and Incompleteness

One of the major features of big data is its heterogeneity because of its heterogeneous sources which is basically mixed data, collected based on various patterns or rules. Here the pattern and the rules and the properties of the pattern varies greatly depending on the variation of sources. And data can be both structured and unstructured in nature though 80% of the data is unstructured. So transforming these unstructured dataset into structured readable format for further processing is a major challenge for big data mining. To overcome this challenge, new technologies should be in place to deal with this kind of data. Sometimes data is also incomplete in nature. So integrating these heterogeneous and incomplete data within a specified time line and reasonable cost, is a great challenge. So data quality is a related challenge which comes with data variety automatically. Synchronization of data also another big concern as data is getting migrated from various sources, on different schedules and different rates.

2.2.2 Scalability

Large volume of data is another big concern for handling big data. Adequate processes and tools should be in place to process these huge volumes of data set at an acceptable speed so that important decisions can be made on time.

2.2.3 Timeliness

Time for analysing and deriving meaningful sights from big data is also an issue as it's comes in huge volume. Sometimes the analysis result is required immediately, like a suspected transaction in the credit card. Ideally it should be identified before the transaction completion by preventing the transaction from taking place at all. Full analysis of a user's purchase history may not be possible in real time but a partial analysis can be completed beforehand on base data so that computation on incremental data can be performed quickly to reach a decision on time.

2.2.4 Security and Privacy

The trustworthiness of big data needs to be verified as it's coming from heterogeneous sources. Appropriate techniques should be in place to find maliciously inserted data and to protect them from various security threats like accessing files and sniffing data packets by unauthorized user which are sent to the client, gaining access privilege by the unauthorized user which leads to unscheduled job submission, modification of job priority etc. Information security is a big concern where massive amount of data will be correlated, analyzed and mined for meaningful

patterns. Various security measures are available in the market to ensure this information security which should meet the following requirements.

- Basic functionality of the cluster should not be compromised
- Scaling should be done in the same manner as the cluster
- Mandatory big data characteristics should not be negotiated
- Security threats should be addressed in an appropriate manner to big data environments

Various authentication, authorization, encryption and audit trails can enhance the security of big data though the possibility of attack can be there. But that can be reduced by implementing the following techniques.

(a) ***Authentication Techniques***: Authentication is the process of identifying valid user or system before system access, like Kerberos. Access control privileges for user or system is provided by authorization process.
(b) ***Encryption and Key Management***: This process ensures confidentiality and security of user information which is also sensitive in nature. It protects data from malicious user access. Consistent protection is provided by file layer encryption across different platforms regardless of OS/platform type. But file layer encryption is not useful if unauthorized user can access the encryption keys. For these cases, key management service is used as a solution which is responsible for key distribution and certificates and manage different keys for each group, application, and user.
(c) ***Logging***: Managing log files is a solution to detect malicious users, failures etc. It provides a place to look if something fails or if something hacked. And periodical audit needs to be conducted in regular basis to find whether any unusual problem occurred.
(d) ***Watermarking Techniques***: Information hiding is another way of protecting sensitive information from unauthorized users which can be achieved by applying the techniques like Steganography, Cryptography and Watermarking. Among these few techniques, Watermarking plays a vital role in the healthcare domain which is used for protecting medical information as well as secured sharing and handling of medical images [3]. In the current electronic age, telemedicine, telediagnosis, teleconsultation are the new buzzwords where digital medical images need to be shared across the globe among various specialist doctors to obtain its benefit. So the primary concern here is to secure patient's information from the attack of any unauthorised user.

Digital watermarking is the solution to preserve the authenticity and integrity of the content of these medical images. A watermark which is basically a distinguishable mark created on paper at the time of production where as in case of digital watermarking, patterns of bits are inserted into a digital image, audio or video file which uniquely defines the file's copyright information without affecting its look

[4]. Another aspect of digital watermarking is its bit arrangement—here the bits are scattered in the whole file in such a way that it can't be found or manipulated [5].

Various important characteristics of this technique include invisibility, robustness, readability and security. Invisibility is the first visible thing because the watermark is not visible at all. It should be robust enough as embedded watermark should not be affected by any kind of attack or image manipulation. Readability is another major concern as it should convey good amount of information. Security is the primary concern which indicates that a watermark should be secret and must not be detectable by unauthorised user [6]. This requirement is normally achieved by cryptographic keys.

There are several applications of digital watermarking techniques which includes security verification (certification, authentication and conditional access), copyright protection, fingerprinting etc. [7]. For copyright protection, the owner's copyright information is inserted into the digital image which is invisible in nature. Here the authenticity can be proved by extracting that information in case of any dispute. In case of healcare related applications, it opens a new era as most of the stages are performed online. Here the electronic patient report and different medical images are sent to various hospitals for consultation. So by using various watermarking techniques, the confidentiality, security as well as the integrity of these online reports and images can be guaranteed. There are various watermarking techniques are available in the market, out of which two correlation-based (binary logo hiding) and two singular value decomposition (SVD)-based (gray logo hiding) watermarking algorithms are widely used for embedding ownership logo [8]. Reversible watermarking method, also known as Odd-Even method, works for watermark insertion and extraction in a biomedical image. This method has a huge data hiding capability, security and watermarked with great quality. Another remarkable feature is its correlation value which is 1 for both original and extracted watermark. The method is also quite robust as Peak Signal-to-Noise Ratio (PSNR) is high irrespective of the amount of embedded secret data [9].

2.2.5 Skills Availability

Processing and finding decisions from big data requires set of new tools and processes for which enough skilled resource is not available in the market. These new tools include various big data analytics tools along with various NoSQL databases. So project cost automatically increases to hire the available resources in a higher rate. If we consider the statistics regarding the talent gap, we can clearly understand the scenario. According to analyst firm McKinsey & Company, there may be an acute shortage of 140,000–190,000 people in the analytics domain by 2018 in the US itself. And in a report from Gartner analysts in 2012, 4.4 million IT jobs will be created globally in the big data domain by 2015 [10].

2.3 Mitigation Techniques

Various innovative tools, technologies and frameworks have been developed to handle big data efficiently which includes Hadoop a distributed file system and framework for storing and processing huge amount of dataset using the MapReduce programming paradigm, different NoSQL data stores with flexible schema pattern, several big data analytics tool like Pentaho Business Analytics etc. Following section gives brief description of these tools and technologies.

2.3.1 Introduction of Parallel Programming Approach: MapReduce

Parallel programming is a methodology where the processing of a particular job is divided into several modules and concurrent execution is performed on the modules. Here modules can run simultaneously on different CPUs where CPUs can be a part of a single machine or they are the part of a network. Improved performance and efficiency are two major motivators of parallel programming. Resource challenge also can be mitigated by using this methodology.

MapReduce, the parallel programming methodology was first developed within Google for huge amount of data processing. Due to its large volume, it was distributed into a set of CPUs for processing within reasonable amount of time. This distribution of data implies the implementation of parallel programming approach as same computation is performed on different dataset on different machine. MapReduce is an abstraction which enables us to perform computation where parallelization details, data distribution, load balancing and fault tolerance are hidden from the users. This programming model is currently adopted for processing large sets of data [11]. The fundamental tenets of this model are Map and Reduce functions. The function of Map is used for generation of set of intermediate key and value pairs, while Reduce function combines all intermediate values with same key. The model provides an abstract view of flow of data and control and the implementation of all data flow steps such as data partitioning, mapping, synchronization, communication and scheduling is made transparent to the users. User applications can use these two functions to manipulate the data flow. Data intensive programs that are based on this model can be executed PaaS. Hadoop which follows MapReduce model, has been used by many companies such as AOL, Amazon, Facebook, Yahoo and New York Times for running their business application. The basic functionality of MapReduce programming is described in the following Fig. 2.

Advantages of MapReduce Programming:

Advantages of MapReduce programming explained and discussed on the basis of scalability, cost effectiveness, flexibility, data processing speed, security management, parallelism, fault tolerance and simplicity in the following.

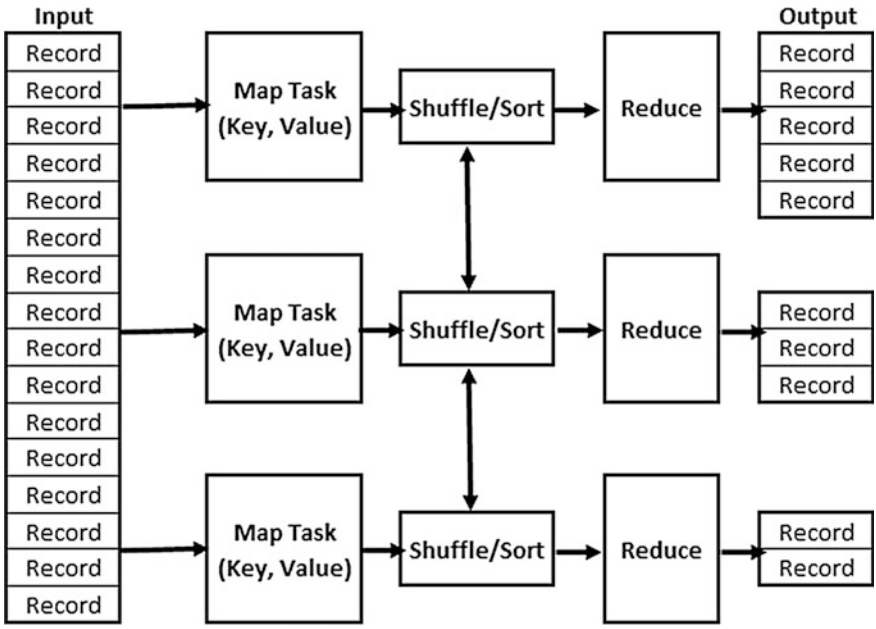

Fig. 2 MapReduce programming model [12]

(a) **Scalability**: MapReduce runs on Hadoop platform which is highly scalable in nature. These Hadoop servers can store and distribute huge dataset across various servers present globally. These servers are inexpensive in nature and parallel operation is acceptable. Depending on the business requirement, new server can be added easily to add more processing power. This advantage is not present in the traditional relational database system which cannot be scaled based on requirement. This MapReduce style enables business to run applications using a large number of nodes which involve huge amount of dataset.

(b) **Cost Effectiveness**: Cost reduction is another big advantage of MapReduce technique over the traditional solutions. Massive cost is associated in traditional system to process the huge dataset whereas implementing Hadoop architecture with MapReduce style reduces the cost to a great extent. It allows storage and huge data processing in an affordable manner.

(c) **Flexibility**: Flexibility is another keyword which made Hadoop MapReduce style so popular. It can be used to access any kind of new sources—structured or unstructured and process them to find the insight from these dataset. It is capable in processing the sources like social media related data, various emails, different log files, marketing related data to find strategic decisions.

(d) **Data Processing Speed**: Faster data processing which is the key requirement in today's e-era can be achievable by implementing Hadoop-MapReduce style. This technique takes minutes to process terabytes of data.

(e) **Security Management**: Managing security is a major aspect of any kind of business data. It is also assured in MapReduce as it works with Hadoop Distributed File System and HBase security which provide permission to the valid user only to access the data stored in the system.
(f) **Parallelism**: Parallel data processing is the primary characteristics of this MapReduce technique which segregates task in such a way that those can be executed in parallel mode. Multiple processors can work on a single task to complete it within shorter time period.
(g) **Fault Tolerance**: Fault tolerance which plays a vital role for business critical data as any kind of data loss leads to major security issues. Hadoop MapReduce methodology is capable of handling this data loss issues as the same data set is copied to various nodes in the network. So if a particular node fails, the data can be retrieved from other nodes which ensure data availability.
(h) **Simplicity**: Simple programming style is another characteristic of this discussed technique which allows programmers to handle task in a more efficient way. This MapReduce is written in java which is easy to learn and already widespread in the market. So mastering this programming style is not a big challenge for the developer community.

2.3.2 Distributed Approach in Cloud Environment—Hadoop Distributed File System for Data Storage and Processing

In reality, cloud computing is used extensively in large data processing applications where data is stored in distributed manner across the globe in various servers and also growing rapidly. Government institutions and large enterprises are leveraging cloud infrastructure to process data in efficiently and at faster rate. Major emphasis is on using parallel programming models to derive extreme capabilities of computing and storage hardware.

Hadoop facilitates a distributed file system and framework which is able to store and process large volume of dataset using the MapReduce programming paradigm. A significant feature of Hadoop is data partitioning and computation across various hosts and the execution of application computations in parallel close to their data. A Hadoop cluster which is a collection of thousands of server can be scaled horizontally by adding more commodity servers based on computation need, storage need and I/O bandwidth. In Hadoop Distributed File System, file system metadata and application data are stored separately. Here metadata is stored in a dedicated server called NameNode and application data is stored in the server called DataNode. All the servers are completely connected and interact with one another by TCP based protocols. In summary, Hadoop framework is the perfect ground to develop applications capable of running on groups of machines, which can perform complete analysis for a large volume of data. So in the world of big data, it appears

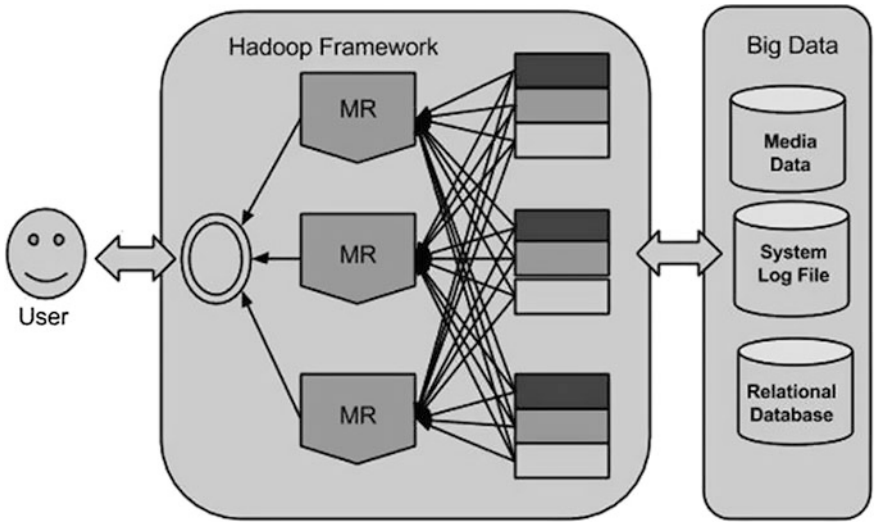

Fig. 3 Basic collaboration among MapReduce, Hadoop and Big data

as an effective solution and accepted across the globe. Figure 3 shows the basic collaboration between Hadoop Framework and Big Data technologies.

A. Hadoop Architecture:

Hadoop framework has mainly four modules which are depicted in the following Fig. 4.

i. **Common Utilities**: These contains the java libraries along with the utilities which is used by other modules of Hadoop.
ii. **YARN Framework**: YARN framework is used for job scheduling and cluster resource management.
iii. **Hadoop Distributed File System (HDFS™)**: It's distributed file system called as distributed storage also, supports master-slave architecture where master contains a single NameNode—responsible for managing file system metadata and one/more slave DataNodes for storing actual data. A HDFS file is the combination of several blocks and those blocks are stored in various DataNodes. The NameNode is responsible for mapping between blocks and DataNodes. On the other hand, DataNode looks after read and write operations with the file system.
iv. **MapReduce**: MapReduce is a software framework, capable of processing huge amount of dataset in parallel on large cluster of commodity servers reliably. It consists of two tasks: Map and Reduce. The functionality of these two tasks are described in the above section. Here the framework takes the responsibility of task scheduling, monitoring and re-execution in case of task failure.

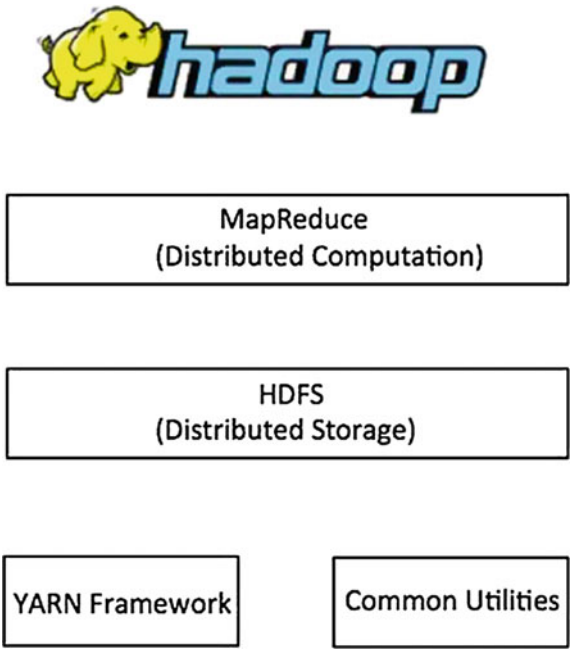

Fig. 4 Basic architecture of Hadoop [13]

B. Advantages of Hadoop Usage:

i. Support of horizontal scalability is one of the beneficial factors of using Hadoop in the big data environment as commodity servers can be added or deleted dynamically from the cluster depending on the business need without interrupting Hadoop's normal operation.
ii. Hadoop library is another unique feature which is enabled for identifying and handling failures at application layer instead of relying on hardware to be fault tolerant and on time availability.
iii. Hadoop framework gives the advantage of writing and testing on distributed environment. Automatic distribution of data and working across various machines are enabled here which utilizes the underlying parallelism of the CPU cores.

2.3.3 Schema Agnostic Data Model: NoSQL

NoSQL database provides a very relaxed approach in the world of data modelling because of its support for the elastic schema pattern and heterogeneous dataset. So management of this large volume of dataset becomes much easier compared to the relational database as data can be distributed automatically with the support of its

unique feature—flexible data model. NoSQL database also has the feature of integrated data caching by which data access latency can be reduced.

The main intention is to get rid from strict relational structure and to allow various models to be adapted to specific types of analyses. There are various models, out of which fours models are widely used which are discussed below.

- Key-value stores
- Document stores
- Column stores
- Graph database

A. Key Value Stores

The key value database which basically uses a hash table where a unique key and pointer to a particular item of data exists. The value is stored in the database in the form of a two valued tuple—one is the key which identifies the record and the other is actual data which is basically the value column. So by the key value, we can easily get all information about the object without traversing the whole database. The main features of key value store database are described below.

- The schema less format is ideal for storage of heterogeneous kind of data.
- Key can be any of the type: synthetic or auto-generated.
- A bucket consists logical group of keys—but it may happen that identical keys are present in different buckets. So here the real key is a hash—bucket + key
- Reading and writing operations are very fast as the key is indexed.
- In the light of CAP theorem, key value store structure supports Availability and Partition not Consistency.
- Read and write functionalities are supported by few functions like: Get (key)—returns the and, Put (key, value)—associated which is with the key, Multi-get (key1, key2, keyN), final Delete (key)—is used to remove the value for the key from the data store.
- It does not support the traditional relational database functionalities like atomicity or consistency simultaneously. It should be created as in built functionalities within the application itself.
- Maintaining unique keys may be difficult if data volume increases.
- However, it's not suitable in the scenario where queries are based on the value rather than on the key.
- It's not a good option for transactions or storing relational data.

Figure 5 presents how data is stored in key-value store database where all information about BIT Kolkata (by the key BIT_KOL) or BIT Mesra (by the key BIT_MESRA) can be retrieved in one shot.

B. Document Stores

In this database, data is also collection of key value pairs where value is compressed as a document and it embeds attribute metadata associated with stored contents. Here also data is identified by the key. It's used mainly to store large files such as Videos, music etc. These types of databases allow to fetch the data for an

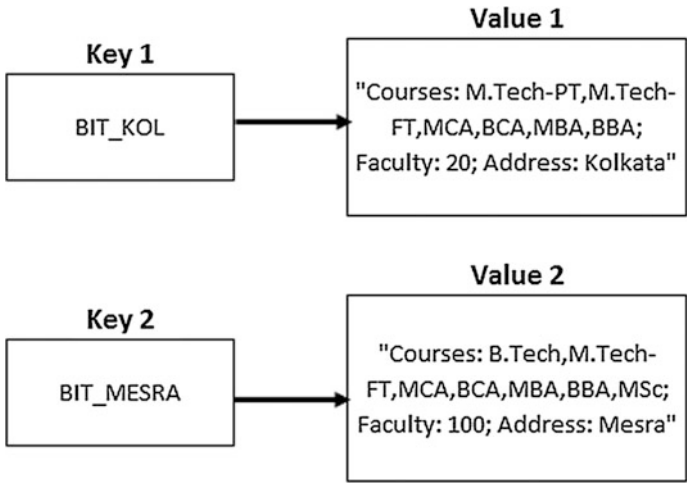

Fig. 5 Example of key—value store

entire page by a single query and most suitable for the applications like Facebook, Amazon.

- Tables do not store data and their relationships.
- Apache CouchDB and MongoDB are examples of document based databases. CouchDB uses JSON and JavaScript along with MapReduce and HTTP.
- It's schema less property makes addition of field in JSON documents a simple task as there is no need to define the changes first.
- Association of metadata with the data fastens the query the data based on the contents.

A single record in MongoDB is a document comprises key and value pairs. The value of these fields may be other documents, arrays and arrays of documents. Figure 6 depicts this kind of scenario.

C. Column Stores

In column store NoSQL database, data is stored in cells grouped together in columns of data instead of rows in relational databases. Here logical grouping of columns is called column families. The main advantage is that there is no restriction on the number of columns which a column family consists and again the columns can be created runtime.

- Here reading and writing operation are executed based on columns instead of rows.
- It stores all the cells corresponding to a particular column in the continuous disk entry instead of storing a single row in the continuous disk space which happens in relational database. For example, querying the conference paper names from the millions of rows is a time consuming task in relational database as it searches

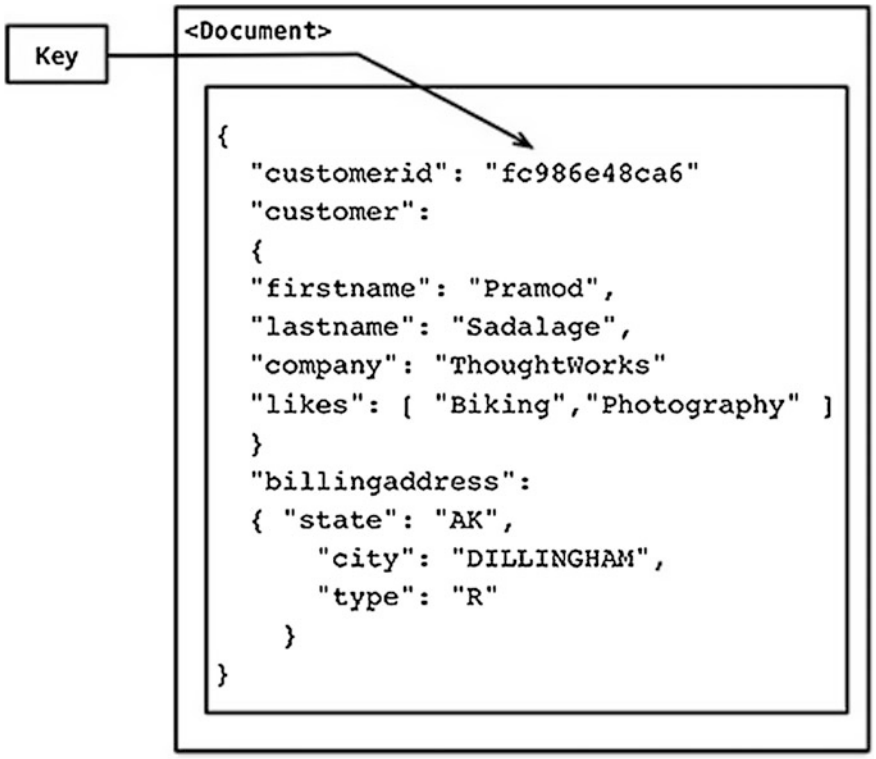

Fig. 6 Example of document store

every location to get the paper name whereas here it can be accessed by reaching one particular disk access as all the values for a particular column are stored in consecutive memory location.

Below is the example of column store database where Course_Information and Address are two column families, represented in Fig. 7. Here if we want to know what the distinct locations of BIT are, we can get it from the column family: "Address", we don't need to traverse the full table but in case of RDBMS, the full table scan is required to get the information.

D. Graph Database

Graph database defines each entry as to how it relates to another item including a pointer to the next item. It stores the relationships directly so that the optimal route between two nodes can be found easily. Here each node consists data about one particular item along with the link of next item.

- Most of the graph databases store value in key-value or document store fashion. In addition to that concept, they store the relationship also which makes the performance faster where data is highly interrelated in nature. This additional

Key	Column Family: Course_Information		Column Family: Address	
	Course Name	Number_Faculty Members	City	PIN
BIT_KOL	M.Tech-FT	8	Kolkata	700107
BIT_KOL	M.Tech-PT	8	Kolkata	700107
BIT_KOL	BBA	5	Kolkata	700107
BIT_KOL	MBA	6	Kolkata	700107
BIT_KOL	BCA	8	Kolkata	700107
BIT_KOL	MCA	7	Kolkata	700107
BIT_MESRA	B.Tech	50	Mesra	835215
BIT_MESRA	M.Tech-FT	40	Mesra	835215
BIT_MESRA	MSc	30	Mesra	835215
BIT_MESRA	BBA	35	Mesra	835215
BIT_MESRA	BCA	35	Mesra	835215
BIT_MESRA	MCA	30	Mesra	835215
BIT_MESRA	MBA	32	Mesra	835215

Fig. 7 Example of column store

feature of storing relationship allows complex hierarchies to be traversed quickly.

- Graph database consists mainly three elements: Nodes, properties and edges. Nodes represent the entities for which we want to keep information like people, business etc., properties holds information related to that node, edges are the lines which creates connection between two nodes, or between nodes and properties and they represent the relationship between the two.
- Graph database is most suitable in the circumstances where data is highly linked to other data in the database and finding relationship among data is a primary requirement including the shortest path between two objects.
- Also very effective for items which vary frequently with time. In those cases change in a particular node will be very easy in this kind of structure.

Figure 8 presents a real life example of graph database where user, page, tag and invitations are the nodes. Login and name are the properties of the node “user”, html and create_ts are the properties of node “page” and the arrows indicate the properties. Here the nodes “user”, “page”, “tag” and “invitation” are interconnected in many ways. So handling these kind of cases by traditional RDBMS is quite complex as every relationship needs to be stored separately which can be easily achieved by the node “property” here. Here every node contains a list of relationship records which indicates the relationship with other connecting nodes. So the database just follows the list and access directly the connected nodes. As a result extensive search or match operation can be avoided compared to traditional RDBMS.

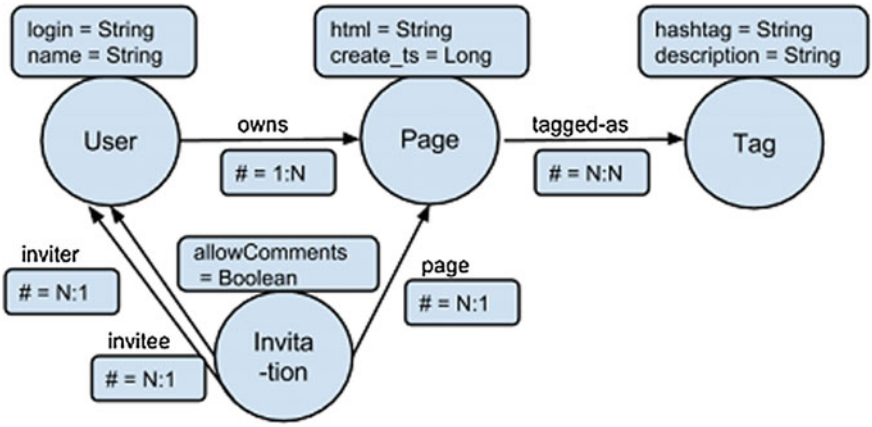

Fig. 8 Example of graph database [14]

2.3.4 Machine Learning Algorithms for Big Data Mining

Machine Learning is a branch which enables computers to learn by their own. No explicit programming is required for this purpose. Here dataset is the source of learning of the algorithms. As a next step, it identifies patterns like trend classification and then automates the output. It may be sorting data into categories or making predictions on future outputs. On the other hand, data mining is the methodology of identifying hidden patterns and features in the data set. The machine learning techniques are used in the data mining domain very often and unsupervised machine learning follows the same principle as data mining. There are various machine learning algorithms which are used for mining large data sets, some of them are Support Vector Machines, Decision Trees, Dimensionality reduction, Neural Networks etc. The brief descriptions of these algorithms are given in the subsequent section.

2.4 Healthcare Analytics—Application for Handling Big Data

In the near future, big data will affect every aspect of our life. There are few areas where already it started to create differences by adding insights from this huge dataset. According to the McKinsey Global Institute, there are four broad areas which deal with big data and can be benefited by analyzing this huge dataset to predict the future. These include applications in public sector, healthcare domain, manufacturing industry and retail domain. Following section briefly describes how big data can be used to add real values in healthcare domain.

2.4.1 Existing Healthcare Applications

Optimizing Treatment and Predict the Risk of Disease: Existing treatments for various diseases can be optimized by analyzing the existing cases which is undoubtedly big data and also risk prediction before appearing of various vulnerable diseases, can be performed by analyzing these huge dataset. The huge computing power of big data analytics help us to decode the complete DNA strings in minute, which also allows us to predict various disease patterns. This will definitely be a great value addition in the field of medical science.

Predicting Early Steps for Premature Newborns: Various big data tools and techniques are already in place to monitor premature and sick babies. These algorithms can predict infections 24 h before they appear by recording and analyzing every heartbeat, breathing patterns etc. In this way, the medical team can intervene early and able to help these fragile babies where time is the most crucial thing.

Prediction of Epidemics and Disease Outbreaks: Big data analytics enable us to predict various epidemics and disease outbreaks on time so that effective measures can be taken to handle the emergency situation if arises. Social media analytics also plays a vital role here for getting various inputs from different people across the country and integrating them to develop the insights.

2.4.2 Healthcare Analytics—Case Study

Big data Analytics is solving various problems and used to meet the business goal across the world in almost every domains like retail, banking, manufacturing, telecommunication, financial sectors. Healthcare Analytics is one such domain where analytics is widely used to obtain any important decision. In the Healthcare domain, huge amount of data has generated from record keeping, compliance and regulatory requirements, patient care in the last few decades which are in paper or hard copy form. Digitization of this large amount of data is the current plan as discovering associations, pattern understanding and trend analysis on this dataset opens a new era in the Healthcare domain which is a crucial part of big data analytics. So applying big data analytics in the healthcare and medical sector is a big step in achieving the facility in lower cost. In this way, it extracts insights from this large dataset which enables to perform various life saving predictions and better decisions on time.

Nowadays, inadequate Medical Facility is one of the primary and vulnerable issues in our society, especially in the developing countries. Our traditional hospital facility is juggling the problems among deficient infrastructure, deficient manpower, unmanageable patient load, equivocal quality of services, high expenditure etc. We are proposing one application, Med-App which intelligently provides the solution for most of the problems occurred due to inadequate medical facilities.

A. Motivation

The primary motivators behind the application includes the following along with some others.

- **Lack of Medical Infrastructure**: Public and private hospitals are incapable to handle the huge population pressure which are increasing exponentially. The main reason of their incapability is inadequate infrastructure and unavailability of medical practitioners.
- **Timeliness**: Another very crucial factor for a sick patient is time, starting the treatment as soon as possible which can save life many times. But too often, access to doctors—and particularly to specialists—can be a difficult challenge. Sometimes, the waiting period in the public hospital is too high that it becomes infeasible for the patients to wait for such a long time as time is the most crucial factor for the vulnerable diseases.
- **Sudden Need of Medical Facility**: Need of Medical facility when user is on the move, it may be suggestion/advice from medical practitioner or availing the medicines or support for immediate hospitalization. So there is an absolute need of an effective and easily accessible health care system which would satisfy the needs of diverse groups within their population.

B. Solution Framework

The proposed medical application—MedApp will address the above issues and can be an effective solution for most of the cases. This is a data analytics based layered framework which has been depicted below by the following Fig. 9. It consists of mainly three layers:

(a) *Source of data*: *Symptoms received from patients*
(b) *Analytics and Knowledge Discovery*
(c) *Visualization and Interpretation*

The layered wise working methodology is described in the following:

(a) *Source of data*: includes various types of symptoms from various patients
(b) *Analytics and Knowledge Discovery*

i. Preliminary analysis will be carried out from the Medical Solution master database
ii. More Predictive Analysis or Analytics will be performed to provide the suggestive medicines or necessary measures which should be taken by the patient during the crisis period or before reaching the crisis point. Historical Dataset and Symptom Database both will be treated as two main sources of input generation for better prediction. Registered Practitioner Database, Registered Hospital Database and Registered Emergency Services will provide the list of registered doctors, registered hospitals and registered emergency services like ambulance accordingly who can extend their support during this emergency.

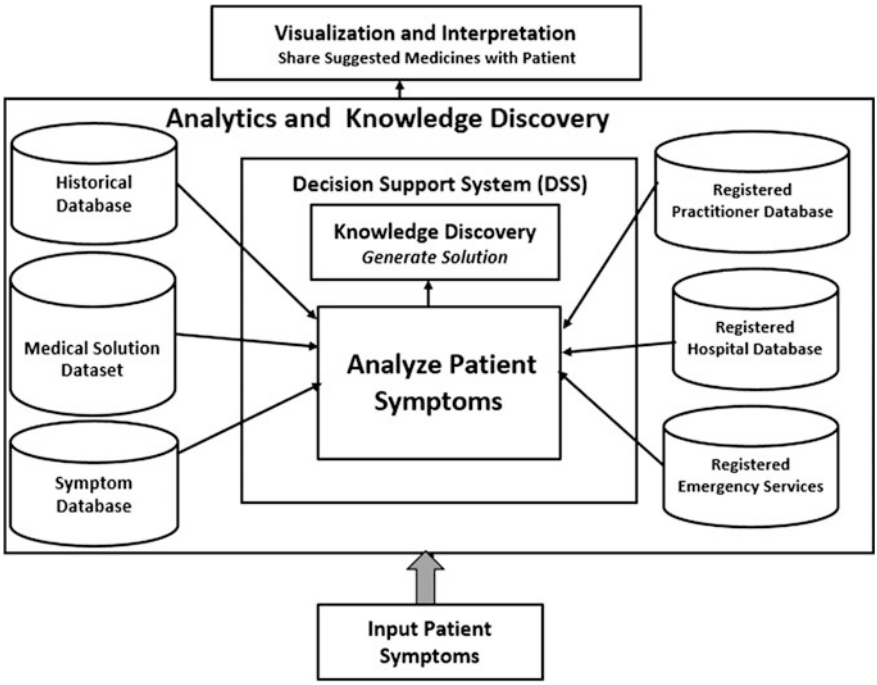

Fig. 9 Solution framework for MedApp application

iii. The next part is Knowledge Discovery which would be performed based on the prediction generated in the previous level.

(c) *Visualization and Interpretation*

The last layer will represent the decision which should be communicated to the user

C. Application Functionality

The brief solution approach of MedApp application is depicted below with the help of flow chart diagram represented by Fig. 10.

The high level details of the above steps are described below.

- **Handling Emergency Situation (Module A)**:

Few databases need to be created which will act as providing necessary inputs to the application. One is Symptom database which will contain various symptoms and corresponding root cause and the other is Medical Solution database which will contain the root cause of the problem and its corresponding solution in terms of medicine and other measures. The next action is populating those databases collecting inputs from patient. The symptom recognizer contains the parameters like Age, Sex, Chief Symptom: (headache/fever/stomach upset/Pain), Onset: (Gradual/Abrupt),

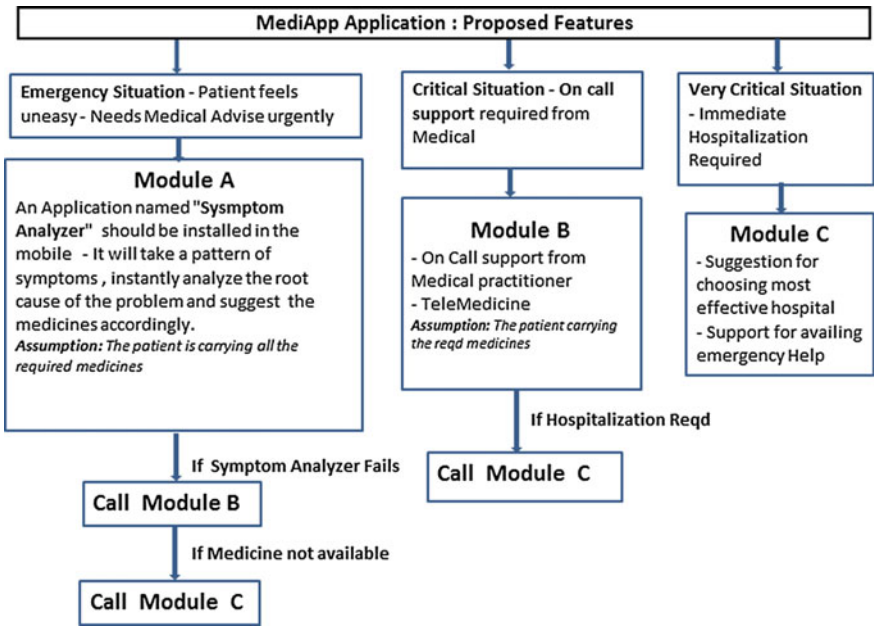

Fig. 10 Basic functionalities of MedApp application

Location of Symptom: (Head/hand/Leg/Abdomen/Neck), Intensity: (Intense/Moderate), Exacerbating Factors (what makes it worse?): (Taking water/taking Food), Ameliorating Factors (what makes it better?), Associated Symptoms etc. Then proper analysis should be done based on the inputs received from the patient and the dataset present in the above mentioned databases. If appropriate solution is not achieved, Module B is called for doctors on call consultation.

- **Handling Critical Situation (Module B)**:

This module describes the functionalities required for on call support of medical practitioner during emergency situation. As a first step of the process, registration needs to be done with set of hospitals, nursing homes and doctors. The application will contain the list of doctors/hospitals along with their contact details who have successfully registered. In the emergency situation, they will be contacted for on call and other support. The patient will receive the list of available doctors/facilitators so that they can be reached for providing help at crisis period. The consultation fees will be paid online to the facilitator.

- **Handling Very Critical Situation (Module C)**:

This module consists the functionalities which are required at the very critical situation like immediate hospitalization. During emergency period, this application finds the closest/closer/close by hospitals/clinics, contact with them for emergency help via messages/E mail and send few important queries which are very crucial at

that stage like Ambulance facility, Hospital Bed Availability, Hospital Bed charges, Availability of specialist Doctors, Availability of cashless facility, Advanced Ambulance facility availability etc. The patient will choose the most appropriate hospital based on the query result received through the application. The hospital will arrange a bed and other facilities like ambulance when get the confirmation from the patient. The patient will reach the hospital availing the facility so that time can be saved which is most valuable at that situation.

3 Scalable/Distributed Algorithms and Their Usage in Big Data Analytics

In the era of big data, innovative tools and techniques are required to extract, process and mine enormous volume of data and finding meaningful insight from the dataset. These techniques should be scalable enough to support on demand requirement and also distributed in nature to enable parallel processing. The following section describes few such algorithms briefly.

3.1 Frequent Item Set Mining—FP-Growth Algorithm

Frequent pattern searching in large database is a very important as well as expensive task in the data mining world in the last few years. The FP Growth algorithm is an efficient and effective way to mine huge dataset which is also scalable in nature. It was found that this algorithm worked better compared to other similar algorithms like Apriori algorithm and Tree Projection. The main reason of its improved performance lies within its methodology of finding frequent item set without using candidate generation. This is parallel in nature and based on divide and conquer strategy. The main feature of this method is its usage of special kind of data structure which is Frequent Pattern Tree (FP Tree), responsible for keeping item set association information. The brief description of the methodology is described below [15, 16].

- A compressed data structure called FP tree is built using 2 passes by compressing the input database to represent the frequent items.
- As a next step, the compressed database is divided into set of conditional databases where each represents one frequent pattern.
- Finally each conditional database is mined separately. As a result frequent item set can be directly extracted from the FP tree.

(a) *Frequent Pattern (FP) Tree Construction:*

- Scan the database of Transaction T once to find out the frequent 1-itemset L
- Arrange the support count in descending order to get the L1
- Take the "Null" as the root node when scan the transaction item set for the second time
- construct the FP-tree base on L1

(b) *Frequent Itemset Generation from the FPtree after constructing the original FP-tree*

- Produce conditional pattern base for every node in the FPtree
- Build the corresponding conditional FP-tree from the conditional pattern base
- Mine the conditional FP-tree recursively and increase the frequent item set belong to it at the same time by producing the involved frequent item set immediately if the conditional FP-tree only contains one path. Otherwise, increase the minimum item of support count if the suffix pattern.
- Then construct the conditional pattern base and conditional FP-tree (The conditional pattern base is the entire branch which takes (E) as their leaf node in the FP-tree. The conditional FP-tree of is a new FP-subtree taking the conditional pattern base as its transaction and constructed in the same way of the original FP-tree).

Fig. 11 Major steps of frequent item set mining algorithm [17]

Following are the major deciding factors to choose this algorithm over the other frequent pattern searching techniques.

- Taking advantages of Divide and Conquer strategy
- No mandate on candidate generation
- Repetitive scan of full database is not required

Major Steps of the Algorithm:

The core methodology of this algorithm is divided into two steps: Frequent Pattern (FP) Tree Construction and Frequent Itemset Generation. The algorithm is depicted below with the help of Fig. 11.

3.2 Deep Learning

Deep learning is a new branch of machine learning whose main objective is moving machine learning nearer to one of its basic goals: artificial intelligence. This learning methodology is responsible for extracting high-level, complex abstractions

as data representations with the help of a hierarchical learning process. Complex abstractions are cultured at a specific level depending on comparatively easier abstractions created in the preceding level in the hierarchy. One important feature of deep learning is its capability to analyze huge amount of unsupervised dataset which is extremely required in the big data analytics as the data here is unlabeled and uncategorized. Few well known applications include extraction of complicated patterns from huge dataset, quicker information retrieval, tagging of data etc. The basic advantages are its robustness, generalizability and scalability. In this learning methodology, designing the features in advance is not required, features are automatically learned to be optimal for the specific task. This is also robust to the natural variation of data as learning is automatic. It's also generic in nature as it can be used for several applications and data. Scalability is another very important characteristics of this methodology as here performance improves with the increase of data and it's massively parallelizable. Another big advantage is it can extract representations from unsupervised data without the manual intervention which is extremely effective in the domain of big data as the data volume is huge.

Following are some common applications of deep learning in the big data analytics [18].

Semantic Indexing: Information retrieval is one of the key tasks in big data analytics which is hugely depends on efficient storage and retrieval process. Here the challenge is increased as data volume is huge and also heterogeneous in nature includes text, image, video and audio. In this situation, the semantic indexing is extremely helpful which presents the data in more efficient manner, automatically helps in the process of knowledge discovery and comprehension. Deep learning comes into the picture as it is able to generate high level abstract data representations which can be utilized for semantic indexing. Complex association and factors are revealed by these representations which leads to semantic knowledge and understanding. As data representation plays an important role in data indexing, deep learning is used to provide a semantic and relational understanding of complex data along with a vector representation of data instances which leads to faster searching and information retrieval.

Discriminative tasks and Semantic tagging: Finding nonlinear features from raw data is a challenge in big data analytics to perform discriminative tasks. Deep learning algorithms can be used in this scenario to extract complicated nonlinear features from the huge dataset and then simple linear models are used to perform discriminative tasks by taking the extracted features as input. The main advantages of this approach includes adding nonlinearity to the data analysis and applying comparatively easier linear models on the extracted features which is computationally efficient, automatically a great advantage in the big data analytics. Hence, huge amount of input data is used to develop nonlinear features which is a great advantage for the data analysts as the knowledge present in the data can be utilised effectively. In this way, data analytics can be benefited to a great extent by implementing deep learning techniques.

4 Machine Learning Algorithms and Their Applications in Big Data Analytics

Machine Learning is a methodology in data analytics domain which automates analytical model building. It allows finding hidden insights from large dataset by using appropriate methods which iteratively learns from data without being explicitly programmed.

In the world of big data, large and heterogeneous datasets are two major challenges for the traditional approaches like trial and error to extract meaningful information from this dataset. Also very few tools allow to process this huge complex dataset within reasonable amount of time. Machine Learning, a novel and rapidly expanding research domain provides effective and efficient solution of the above issues by implementing appropriate machine learning techniques which differs from the traditional approaches [19].

4.1 Classification of Machine Learning Algorithms

An algorithm can model a problem in various ways depending on its interaction with input data. So choosing the appropriate Learning Style is the first thing which needs to be considered by these machine learning algorithms. There are few learning styles or learning models which a machine leaning algorithm can adopt to get the desired output. The common learning style includes Supervised Learning, Unsupervised Learning, Reinforcement Learning, Transduction, Learning to learn. Brief description of these learning styles are depicted below [20, 21].

4.1.1 Supervised Learning

Supervised learning is a type of learning which is appropriate when correct results are assigned to the training instances that can predict the progress of learning. This is a very common method in classification problems where the goal is mostly to get the computer to learn a classification system that is already created. A very common example in classification learning is Digit recognition. In general, it's most appropriate where classification generation is the main objective and also can be done easily. The most common areas of implementing this methodology is training neural networks and decision trees. For neural networks, it's used to find the error of the network and also for doing necessary adjustments in the network to minimize it. In case of decision trees, classifications are providing information about the attributes which can be utilized to provide the solution of classification puzzle. Figure 12 depicts the basic methodology in supervised learning. The main objective in supervised learning is to build a model which is able to do prediction based on

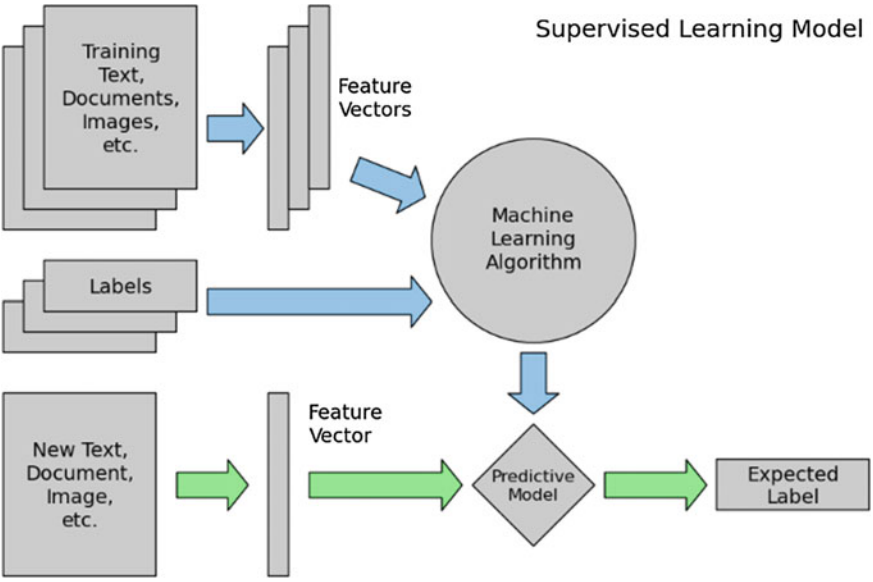

Fig. 12 Supervised learning model [22]

evidence in the presence of uncertainty. For doing this, first it takes known input data along with known responses and based on that it trains a model so that it is able to generate meaning prediction with new set of inputs [23, 24].

Steps in Supervised Learning: The basic steps in supervised learning includes the following [25].

- **Data Preparation**: This is the beginning step of it which starts with input data preparation in some specified format.
- **Choosing appropriate Algorithm**: There are several characteristics of the algorithms which includes training speed, memory usage, prediction accuracy, transparency etc. So appropriate algorithm needs to be chosen based on the requirement.
- **Fit appropriate Model**: Various algorithms have various fitting functions, so fitting function plays very important role in the selection process of appropriate algorithm.
- **Selection of appropriate validation methodology**: There are various validation methods are available which are used to check the accuracy of the resulting fitted model like examine the resubstitution error, cross validation error, out of bag error for bagged decision trees.
- **Test and Update**: After model validation, further tuning can be required to achieve better accuracy, better speed, less memory usage etc.
- **Using final model for prediction**: Final model can be used for prediction of new dataset.

Importance of Supervised Learning Algorithms in Big data Analytics: Machine learning is a very ideal solution for exploiting new opportunities from this huge volume of data as it requires minimum human interaction. Also this methodology is data driven and runs at machine scale. It is also capable of handling huge variety of variables coming from heterogeneous sources. Another big advantage of this method is its improved performance in the presence of large dataset. The machine learning system will provide better prediction if it learns more which is possible if it feeds more data.

Common Algorithms in Supervised learning: The following section describes the commonly used algorithms in supervised learning.

A. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used both for classification and regression problems though mostly used in classification challenges. Here we plot every data item as a point in n-dimensional space where n denotes the feature number. The value of the particular co-ordinate is basically value of the feature. The following example explains it in a better way.

Suppose there is a sample set of population containing 50 % male and 50 % female. We need to create few rules by observing the features of this sample set so that the gender of a new person can be identified correctly based on this ruleset. This is basically a problem of classification domain which can be solved efficiently by Support Vector Machine (SVM). Here the sample features for observation are height and hair length. First we plot the data based on these two features which clearly classify the set into two segments in the following Fig. 13.

Here the circles represent the female population and squares represent the male population. The two rules can be created based on the observation of the above set.

- Male population has the higher average height
- Female population has the longer hair.

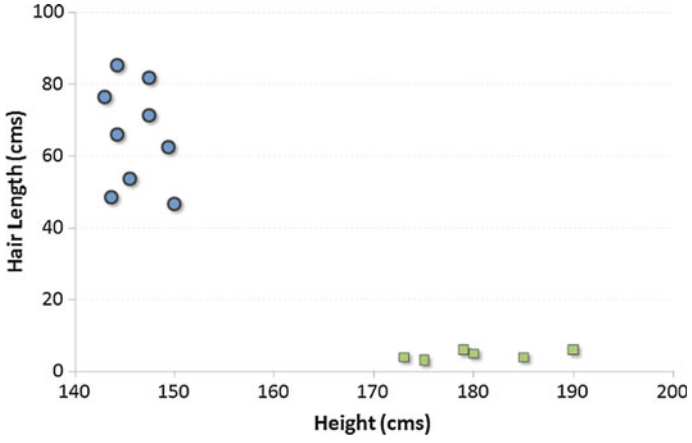

Fig. 13 Mapping of features: hair length and height [26]

Support Vector is the coordinate of individual observation. Here (45, 150) is a support vector which represents a female. Support Vector Machine is a frontier which can segregate the two groups in an optimum way. Various strategies exist in the market which are quiet effective to find the optimum frontier as multiple frontiers can exist in a problem. The simplest way to understand the objective function in a SVM is to discover the smallest minimum distance of the frontier from closest support vector which can belong to any class. After finding all the distances for all the frontiers, we just select the frontier with maximum distance from closest support vector.

B. Naive Bayes Algorithm

Naive Bayes algorithm is one of the fastest classification algorithm which works based on Bayes theorem of probability for prediction of the class of unknown dataset. The basic assumption of Naïve Bayes classifier is that there should not be any relation between the presences of specific feature in a class with the presence of any other feature. This model is very suitable and useful for large data sets.

Posterior probability P(c|x) can be calculated with the help of Bayes theorem based on P(c), P(x) and P(x|c). The following Fig. 14 depicts the formula and meaning of the variables.

There are various advantages of this methodology over the other ones like this is the simplest and extremely fast in prediction of test data set and also capable of multi class prediction. The most common usages include real time prediction, multi class prediction, sentiment analysis etc.

C. Decision Tree Classifiers

Decision tree is a well-known supervised learning method used for classification and regression. A decision tree of a pair (x; y) denotes a function which takes the input attribute x (Boolean, discrete, continuous) and outputs a simple Boolean y. This is basically a predictive model which is used to map the observations regarding an item to conclusion about the item's target value. This can be used to visually and explicitly represent decisions. The ultimate goal of this method is predicting the value of a target variable based on simple decision rules concluded from the data features.

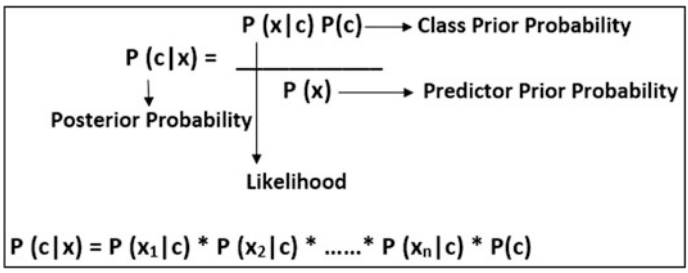

Fig. 14 Bayes theorem

```
buildtree (examples, questions, default)
/* examples: a list of training examples
questions: a set of candidate questions, e.g., "what's the value of feature xi?"
default: default label prediction, e.g., over-all majority vote */
IF empty(examples) THEN return (default)
IF (examples have same label y) THEN return(y)
IF empty(questions) THEN return(majority vote in examples)
q = best_question(examples, questions)
Let there be n answers to q
        – Create and return an internal node with n children
        – The ith child is built by calling
                Buildtree ({example|q=ith answer}, questions\{q}, default)
```

Fig. 15 Outline of decision tree algorithm [27]

A decision tree is also called as classification tree where each non leaf node is denoted with an input feature and the arcs which are joined with the nodes (labelled with feature) are labelled with each of the possible values of the feature. And each leaf of the tree represents a class or a probability distribution over the classes. The outline of the decision tree algorithm is depicted in the following Fig. 15.

4.1.2 Unsupervised Learning

Unsupervised learning is more complicated approach than supervised learning. Here the objective is to learn something by the computer by its own. There are primarily two approaches available in this type of learning. The first approach is teaching the agent with the help of reward system which is an indicator of success. This approach is most suitable into the decision problem framework where the goal is making decisions for maximizing rewards instead of producing a classification. The second type of approach is clustering where the goal is finding similar patterns in the training dataset instead of maximizing a utility function. There are various techniques used in unsupervised learning includes K-means clustering algorithm, dimensionality reduction techniques etc. Few common areas where this type of learning methodology is most suitable are determine the most important feature for distinguishing between galaxies where detailed observation of detailed galaxies are present, for the blind source separation problems etc. [24, 28].

Steps in Unsupervised Learning: The following Fig. 16 depicts the different steps involved in unsupervised learning.

Importance of Unsupervised Learning Algorithms in Big data Analytics: Unsupervised learning is one of the most effective way for analyzing big data as no

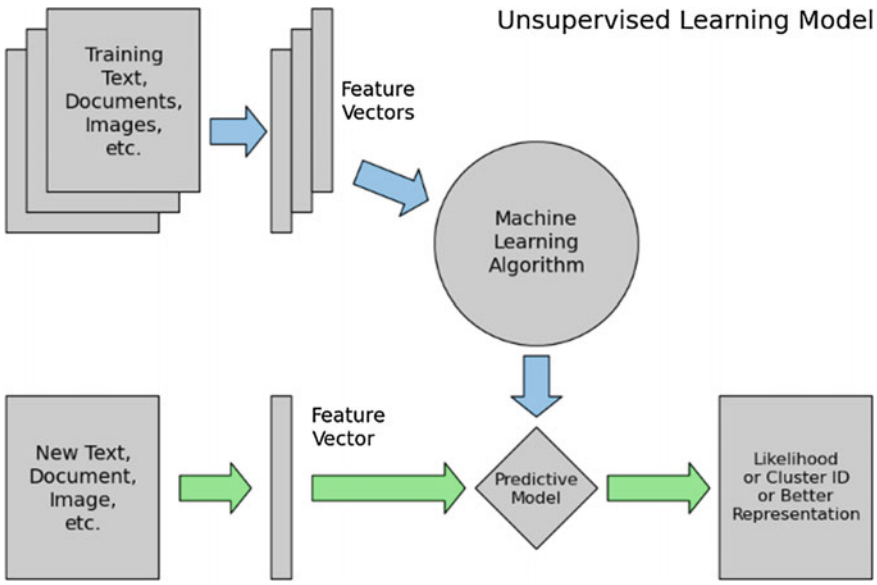

Fig. 16 Unsupervised learning model [22]

training set data is required here. In the big data domain, analysis is normally required on the dataset under exploration where predefined rule set is not available. So in this situation, unsupervised learning is quite effective to find useful patterns above and beyond noise.

Common Algorithms in Unsupervised learning: The following section describes the commonly used algorithms in unsupervised learning.

A. Clustering Algorithms:

Clustering is a popular concept which groups organization of unlabeled data based on similarity. So as a result, similar kind of data belongs to one group and other reside in another group. There are mainly three types of clustering algorithms are available, out of which K-means is the most widely used technique [29].

- **Bayesian Algorithms**: The major goal of this kind of algorithm is to generate a posteriori distribution over the collection of all partitions of the data.
- **Hierarchical Algorithms**: These type of algorithms find successive clusters using the clusters used previously. These algorithms can have two approaches again. First one is Agglomerative algorithms which starts with each element as a separate cluster and combine them into successively huge clusters. Second one is Divisive algorithms which starts with the complete set and continue to split it into successively reduced clusters.
- **Partition Algorithms**: These type of algorithms find all clusters at the same time but can also be used as divisive algorithms in the hierarchical clustering. K-means clustering algorithm resides in this group.

K-Means Clustering: K-means is the easiest unsupervised learning algorithm which is used as a well-known solution of clustering problems. The following figure describes the algorithm briefly. The main idea is to classify a defined data set through a specific number of clusters (let k clusters) fixed apriori. The first thing is defining k centers, one for each cluster. These centers needs to be placed in a cunning way as it various locations lead to various results. So the better choice is to place them in such a way that maximum gap is maintained among them. As a next step, we need to consider each point belongs to a given data set and associate it to the closest center. The first step will be completed when no point is left and an early group age is completed. Now we need to recalculate k new centroids as barycenter of the clusters resulting from the past step. After finding the k new centroids, a new binding needs to be done between the same data set points and the nearest new center. A loop has to be formed. As a result of this loop, it may be observed that k centers change their locations step by step until no more changes are required. Finally, the objective is minimizing an objective function commonly referred as squared error function given by the following Fig. 17.

The brief description of the algorithm is as follows in following Fig. 18.

B. Dimensionality Reduction Techniques:

In the world of big data, the volume of dataset increased tremendously and it leads to lots of redundancy. So it needs a treatment of dimensionality reduction to remove unwanted dimension. These techniques refer to the process of converting data set with higher set of dimensions into lower set of dimensions but ensuring to convey same information. These techniques are very common for achieving better features in classification or regression task in machine learning domain. One very common area of implementing this technique is image processing. There are various methods available in dimensionality reduction, some of them are depicted below in the following.

- **Missing Values**: In big data analytics, we face the problem of missing values very often. It's better to drop the variables if the rate of missing values for those variables are high with the help of appropriate methods.

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\| x_i - v_j \right\| \right)^2$$

where,

'$||x_i - v_j||$' is the Euclidean distance between x_i and v_j.

'c_i' is the number of data points in i^{th} cluster.

'c' is the number of cluster centers.

Fig. 17 Squared error function

Input: Data points D, Number of Clusters k

Step 1: Initialize k centroids randomly.

Step 2: Associate each data point in D with the nearest centroid. This will divide the data points into k clusters.

Step 3: Recalculate the position of centroids.

Repeat Steps 2 and 3 until there are no more changes in the membership of the data points.

Output: Data points with cluster memberships

Fig. 18 Pseudo code of K-means clustering algorithm [17]

- **Low Variance**: We can encounter the constant variable in our data set which has little power to improve the model. In such cases, it's better to drop such variables from the data as it will not explain the variation in target variables.
- **Random Forest**: This is almost similar to the previous technique, decision trees. It's always recommended using the inbuilt feature importance given by random forests to select a smaller subset of input features.
- **Principal Component Analysis**: This technique is used very often in real world. Here variables are converted into a new set of variables which are linear combination of original variables. This new set of variables are referred as principal components.

4.1.3 Reinforcement Learning

Reinforcement learning is a kind of machine learning which is a branch of artificial intelligence. It enables machines and software agents to determine the ideal behaviour in a specific scenario automatically to maximize the performance. Basically this is the learning from interaction with environment. The agent learns here from the consequences of its actions instead of explicitly programmed and determine the new course of actions based on exploitation and exploration. So it can be called as "trial and error" learning also. Based on this learning, the algorithm modifies its strategy to achieve the optimum performance. Various algorithms are available in reinforcement learning to handle the issues. The main advantages include lesser time requirement in designing a solution with slight manual intervention [30].

The following Fig. 19 describes the abstract view of reinforcement learning agent in its environment. At a specific time instant, a state, an event and a reward are observed by an agent from its operating environment, the agent performs learning, takes necessary decision and actions accordingly.

At any time variant t, the agent performs a suitable action so that maximum reward is achievable in the next time instant t + 1. Learning engine which is the most important component, offers knowledge of the operating environment based

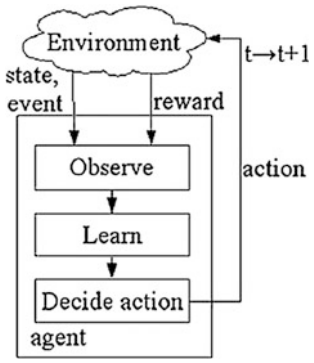

Fig. 19 Abstract view of reinforcement learning agent

on the observation of the consequences of its past actions which includes the state, event and reward.

The following Fig. 20 describes the flowchart of reinforcement learning approach. At time t, an agent chooses a subset of actions to adhere a set of rules. As a next step, it chooses either exploration which is a random action selected to increase the knowledge of the environment or exploitation action which is the best known action derived from Q table. At the next time instant t + 1, it watches the consequences of the past actions including state, event and reward and updates Q tables and rules accordingly.

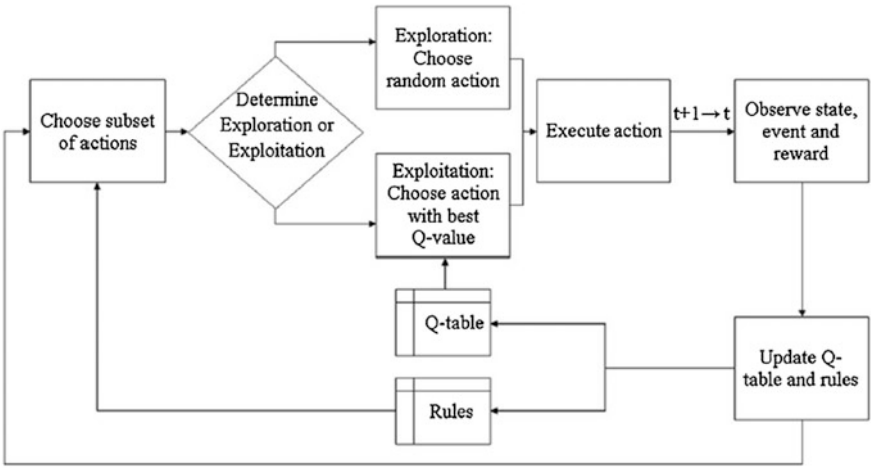

Fig. 20 Flowchart of reinforcement learning approach

Reinforcement learning techniques are extremely useful in big data analytics as it's capable of handling huge amount of data compared to other methods. These techniques automatically learn from past experiences (which is huge in nature) without much manual intervention. It ensures more accuracy in prediction as more examples can be integrated within the predictive model. The following are various real life applications of reinforcement learning.

A. Predictive Voice Analytics in Call Center:

Predictive voice analytics program in call center observes human speech pattern to obtain emotional tone and behaviour and predict future behaviours based on that. Here the reward can be defined as a positive call outcome when the customer agrees to pay for some product. The prediction algorithm explores how the customer speaks based on the voice analysis. It ultimately do the predictive analysis of the customer's future behaviour based on the current one. The reinforcement learning technique will collect data from thousands of calls which includes both positive and negative results and connect individual feature vectors from each call to the end result. After getting the aggregated results data, the efficiency and accuracy of the algorithm is increased. So in summary, reinforcement learning tunes the analytics program in a better way so that it can predict the customer behaviour more accurately which saves human effort to a great extent.

The following Table 1 describes various features and characteristics of the problems based on which we can select the appropriate machine learning algorithm.

4.2 Applications of Machine Learning Algorithms

The primary goal of machine learning research is to use the learning algorithms as a solution of real life challenges which includes fraud detection, result of web searching, sentiment analysis, credit scoring, various prediction in the automobile industry, new pricing models, image and its pattern recognition, filtering spam email.

Support vector machines is widely used in mainly in the pattern classification problems and nonlinear regressions. Two such pattern classification problems are cancer diagnosis based on microarray gene expression data and other is protein secondary structure prediction. Here prediction means supervised classification which has two steps. In the first step, a support vector machine is trained as a classifier with a part of the data in a particular protein sequence data set. After that the classifier is used to classify the residual data in the set as a second step.

Nowadays, few innovative techniques are used to determine cancer type instead of the traditional approaches. Traditional approaches are based on the morphological appearances of cancers. Sometimes it becomes really difficult to find clear distinction among the cancer types as it's only based on their appearance. Gene expression based cancer classifiers have achieved satisfactory results for specific type of cancers like lymphoma, leukemia, breast cancer, liver cancer etc. [31].

Table 1 Various parameters for choosing appropriate algorithm

Machine learning types	Algorithm specific comparison parameters
Supervised learning	1. Naïve Bayes, a supervised learning method has the advantage of its *quicker convergence*. If the conditional independence assumption is true, it will converge quicker than the discriminative models like logistic regression and if the assumption is not true, it's still also effective in practical scenario
	2. Support Vector Machine, another supervised technique has the high level of *accuracy*. It's very suitable in the text classification problems where high dimensional spaces are the common practise
	3. Decision Tree is another technique which can handle *feature interaction* quite easily and *non-parametric* in nature. So no need to worry about whether data is linearly separable
Unsupervised learning	1. Cluster Analysis which is used as the most common type of unsupervised learning, works for *explanatory data analysis* to find hidden patterns or data grouping. There are various types of clustering algorithms out of which K-means clustering is the most popular one. The main advantages of K-means clustering is its usage of simple principles which can be explained in non-statistical terms. It's also *highly flexible* in nature and also *adaptable* with simple adjustments. The main advantage lies in its performance for real world applications
	2. Dimensionality Reduction is another popular technique which is used to reduce time and required storage space. The machine learning model is automatically improved with the *removal of multi-collinearity*. Data visualization becomes much easier when dimensions are reduced to 2D or 3D
Reinforcement learning	1. Reinforcement learning can choose an action is response to a data point. This technique is capable of balancing *exploration and exploitation* whereas the supervised techniques are purely explorative in nature
	2. *Minimal manual intervention* is another major aspect of this technique
	3. Capable of implementing *artificial intelligence* as required

Naive Bayes classifier is a type of supervised learning which is used to classify spam and non-spam emails, classify among articles among technology, politics and sports, capable to check a piece of text whether expressing positive and negative emotions etc.

Decision tree classifier is widely used in the domain of astronomy. Noise is filtered from Hubble space telescope images with the help of decision trees. It also helps in star galaxy classification, determining galaxy counts etc. In the domain of biomedical engineering, it's used to identify features to be used in implantable devices. Automatic induction of decision trees is used for controlling of nonlinear dynamic systems in the control system domain. This technique is also widely used in the domain of medicine and molecular biology for diagnosing various disorders. Human Genome project, a great initiate from molecular biology has deployed this technique for analysing amino acid sequences.

K-means clustering is useful for undirected knowledge discovery. It is widely used in the areas ranging from unsupervised learning of neural network, pattern recognition, classification analysis, image processing etc.

It's applied extensively in various problems of data mining domain. In the field of image processing, it's used for choosing colour palettes on old fashioned graphical display devices and image quantization.

Dimensionality Reduction is an important technique in unsupervised learning to achieve better visualization, data in compressed format for efficient storage and retrieval and noise reduction used to gain positive effect on query accuracy. Document classification in a real life problem where this methodology is used widely. Here the objective is to classify unlabelled documents into categories which has thousands of terms. Another area is gene expression microarray analysis where the goal is to classify unlabelled samples into known disease types. Here the main challenge is the presence of thousands of genes along with few samples.

Reinforcement Learning which a very popular technique in today's machine learning world where the agent learns to perform a task based on the past learning experiences from the environment. The basic idea is the reinforcement outcome becomes positive if the goal is achieved and it's treated as negative if obstruction is faced. Video games and Robotics are such fields where there is a wide application of this learning methodology. In general an agent which is a game character or robot is present here which moves within the environment. The agent is allowed to perform task while moving. If it faces obstacles while moving, the outcome is negative and if goal is achieved, the outcome is positive.

This learning technique is also used in optimization of anaemia management among the patients who are undergoing hemodialysis. This is a very well-known problem in Nephrology where optimal Erythropoietin (EPO) dosages can be obtained by proper administration for an adequate long tern anemia management. The suitability of this methodology here its way of tacking the problem for obtaining long term stability in patients' haemoglobin level. If the patient is in a certain state, this technique suggests the sequence of actions which guides the patient to the best possible state [32].

Another real life application of this learning is the optimization of a marketing campaign. The basic approach is using data from marketing campaign to provide suggestion to the company policy for achieving long time organization goals which is achieved by implementing this type of learning [32].

Discussion and Future Scope:

The fundamental aspect of machine learning is to provide analytical solutions which can be created based on studying past data models. Data analysis is supported to a great extent where past data models, various trends and patterns work as the learning inputs whereas automated algorithmic systems is the final outcome. In today's world, data analytics and prediction are the keywords, without which it will be difficult for us to sustain in the future. As per the future prediction across the globe, machine learning will remove human intervention from the world of analytics which is completely dependent on prebuilt algorithms for doing various

predictions and analysis. There are various learning procedures which has the capability to study and learn from past experiences and based on that it simulates the human decision making process. It acts as an effective solution making tool in the domain of demand forecasting also. It removes the human intervention as well as biasness in demand planning activities. As it has the inherent capability of learning from past and current data, it's capable of handling challenges arise due to demand variation.

The Internet of Things has given a new leash to the traditional machine learning techniques. Some common machine learning applications include customer feedback in Twitter, self-driven Google car, various fraud detection systems which are capable of handling huge amount of heterogeneous data. According to future prediction across the globe, the global community will be witness a remarkable growth in the near future in smart applications, digital assistants and various usage of artificial intelligence. Machine learning will take the lead in these emerging technologies. Vendors will be pushed to provide new machine learning tools to cope up with the increased demand. Though these ready products will be available in the market, there will be a huge requirement to customize them and create more advance model according to the specific need. Machine. According to McKinsey, the implementation of this emerging technology will enable the business to work with reduced manpower which will definitely help them in reducing operational cost. Global investment banks are welcoming automated trading which increases the probability of making profits by at least 30 %. As a result, more data scientists and big data experts will be required for making the business successful. In Germany, an algorithm for reading street sign has achieved 99.4 % success rate where for human it's 5 % only. Google and Amazon are some big names whose reliability increases on machine learning instead of domain experts to make more profit in the business. In summary, machine learning will work a major differentiator in the all kind of industry.

5 Conclusion

In today's era of big data, it has become a real challenge to extract meaningful insights by applying traditional algorithms/methods from unstructured, imperfect and complex dataset in almost all the domains like Environmental study, biomedical science, Engineering etc. The challenges include understanding and prioritizing relevant data from the huge set, extracting data from master set where 90 % data reflects noise, security threat, costly tools and framework etc. So various innovative tools, technologies and frameworks have been developed to handle these challenges which includes Hadoop a distributed file system and framework for storing and processing huge amount of dataset using the MapReduce programming paradigm, different NoSQL data stores with flexible schema pattern, several machine learning algorithms includes supervised, unsupervised and reinforcement

learning etc. This chapter describes these various tools, technologies, machine learning algorithms along with their application in the analytics domain in detailed fashion. These applications help to gain clearer picture on the usages of these machine learning algorithms in the world of big data.

References

1. Clifton, D.A., Niehaus, K.E., Charlton, P., Colopy, G.W.: Health informatics via machine learning for the clinical management of patients. Yearbook Med. Inform. **10**(1), 38 (2015)
2. Moazeni, M.: Parallel Algorithms for Medical Informatics on Data-Parallel Many-Core Processors (2013)
3. Acharjee, S., Ray, R., Chakraborty, S., Nath, S., Dey, N.: Watermarking in motion vector for security enhancement of medical videos. In: 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), pp. 532–537. IEEE (2014, July)
4. Bose, S., Acharjee, S., Chowdhury, S. R., Chakraborty, S., Dey, N.: Effect of watermarking in vector quantization based image compression. In: 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), pp. 503–508. IEEE (2014, July)
5. Rathi, S.C., Inamdar, V.S.: Analysis of watermarking techniques for medical images preserving ROI. In: Computer Science & Information Technology (CS & IT 05)-open access-Computer Science Conference Proceedings (CSCP), pp. 297–308 (2012)
6. Coatrieux, G., Lecornu, L., Sankur, B., Roux, C.: A review of image watermarking applications in healthcare. In: Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE, pp. 4691–4694. IEEE (2006, August)
7. Abd-Eldayem, M.M.: A proposed security technique based on watermarking and encryption for digital imaging and communications in medicine. Egypt. Inform. J. **14**(1), 1–13 (2013)
8. Suri, J., Dey, N., Bose, S., Das, A., Chaudhuri, S.S., Saba, L., Nicolaides, A.: 2084743 diagnostic preservation of atherosclerotic ultrasound video for stroke telemedicine in watermarking framework. Ultrasound Med. Biol. **41**(4), S133 (2015)
9. Pal, A.K., Dey, N., Samanta, S., Das, A., Chaudhuri, S.S.: A hybrid reversible watermarking technique for color biomedical images. In: 2013 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1–6. IEEE (2013, December)
10. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A. H.:. Big Data: The Next Frontier for Innovation, Competition, and Productivity (2011)
11. Kamal, S., Ripon, S.H., Dey, N., Ashour, A.S., Santhi, V.: A MapReduce approach to diminish imbalance parameters for big deoxyribonucleic acid dataset. Comput. Methods Programs Biomed. **131**, 191–206 (2016)
12. A presentation on MapReduce. http://www.slideshare.net/nishantgandhi99/map-reduce-programming-model-to-solve-graph-problems
13. A tutorial on "Introduction to Hadoop". http://www.tutorialspoint.com/hadoop/hadoop_introduction.htm
14. A whitepaper on "Graph Database". http://lambdazen.blogspot.com/2014/01/from-entity-relationship-to-property.html
15. Sidhu, S., Meena, U.K., Nawani, A., Gupta, H., Thakur, N.: FP Growth algorithm implementation. Int. J. Comput. Appl. **93**(8) (2014)
16. A whitepaper on "Data Mining Algorithms In R/Frequent Pattern Mining/The FP-Growth Algorithm". https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Frequent_Pattern_Mining/The_FP-Growth_Algorithm

17. Verhein, F.: Frequent Pattern Growth (FP-Growth) Algorithm. School of Information Studies, The University of Sydney, Australia (2008)
18. Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R., Muharemagic, E.: Deep learning applications and challenges in big data analytics. J. Big Data **2**(1), 1 (2015)
19. Brownlee, J.: A Tour of Machine Learning Algorithms. A post available at http://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/
20. Qiu, J., Wu, Q., Ding, G., Xu, Y., Feng, S.: A survey of machine learning for big data processing. EURASIP J. Adv. Signal Process. **2016**(1), 1–16 (2016)
21. Oberlin, S.: Machine learning, cognition, and big data. CA Technology Exchange, 44 (2012)
22. A learning material on "Machine Learning 101: General Concepts". http://www.astroml.org/sklearn_tutorial/general_concepts.html
23. Machine Learning—What it is & Why it Matters. http://www.sas.com/en_id/insights/analytics/machine-learning.html
24. Machine Learning, Part I: Supervised and Unsupervised Learning. http://www.aihorizon.com/essays/generalai/supervised_unsupervised_machine_learning.htm
25. Supervised Learning Workflow and Algorithms. http://in.mathworks.com/help/stats/supervised-learning-machine-learning-workflow-and-algorithms.html?requestedDomain=www.mathworks.com
26. A blog on "Understanding Support Vector Machine Algorithm from Examples". http://www.analyticsvidhya.com/blog/2014/10/support-vector-machine-simplified/
27. A lecture note on "Machine Learning: Decision Trees". http://pages.cs.wisc.edu/∼jerryzhu/cs540/handouts/dt.pdf
28. Ray, S.: Essentials of Machine Learning Algorithms (with Python and R Codes). A post at AnalyticsVidhya available at http://www.analyticsvidhya.com/blog/2015/08/common-machine-learning-algorithms/
29. Cios, K.J., Swiniarski, R.W., Pedrycz, W., Kurgan, L.A.: Unsupervised learning: clustering. In: Data Mining, pp. 257–288. Springer US (2007)
30. Yau, K.L.A., Komisarczuk, P., Teal, P.D.: Reinforcement learning for context awareness and intelligence in wireless networks: review, new features and open issues. J. Netw. Comput. Appl. **35**(1), 253–267 (2012)
31. Wang, L. (ed.): Support Vector Machines: Theory and Applications, vol. 177. Springer Science & Business Media (2005)
32. Martín-Guerrero, J.D., Soria-Olivas, E., Martínez-Sober, M., Serrrano-López, A.J., Magdalena-Benedito, R., Gómez-Sanchis, J.: Use of reinforcement learning in two real applications. In: European Workshop on Reinforcement Learning, pp. 191–204. Springer Berlin Heidelberg (2008)