|  |
| --- |
| **A Project Based Seminar Report**  **on**  **‘Application based data collection and sequential DNA mining’**   Submitted to theSavitribai Phule Pune UniversityIn partial fulfillment for the award of the Degree ofBachelor of EngineeringinInformation TechnologybySatwik Suhas Ramchandre(T150028563 / 306074 & TE - IT)T.E. (Information Technology)Under the guidance of**Dr. Himangi Pande** **MIT_Logo**  **DEPARTMENT OF INFORMATION TECHNOLOGY**  **MAHARASHTRA INSTITUTE OF TECHNOLOGY**  **PUNE-411038**  **Academic Year 2018-2019**  **Semester II**  Affiliated to  uop_logo  Savitribai Phule Pune University |
| MIT_Logo  **MAEER’S**  **MAHARASHTRA INSTITUTE OF TECHNOLOGY, PUNE** DEPARTMENT OF INFORMATION TECHNOLOGY**CERTIFICATE** This is to certify that the project based seminar report entitled**”** **Application based data collection and sequential DNA mining”** being submitted by **Satwik Ramchandre (T150028563/306074 TE - IT)** is a record of bonafide work carried out by him/her under the supervision and guidance of Dr. Himangi Pande in partial fulfillment of the requirement for **TE (Information Technology Engineering) a 2015 course** of Savitribai Phule Pune University, Pune in the academic year 2018- 2019.  Date: 12/04/2019  Place: Pune Dr. Himagi Pande Prof. Sumedha Sirsikar **Seminar Guide Head of the Department, IT**  **Dr. L.K. Kshirsagar**  **Principal**  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  This Project Based Seminar report has been examined by us as per the Savitribai Phule Pune University, Pune requirements at Maharashtra Institute of Technology, Pune 411038 on . . . . . . . . . . .  Internal Examiner External Examiner |

###### *ACKNOWLEDGEMENT*

I take this opportunity to express my sincere appreciation for the cooperation given by Prof. Sumedha Sirsikar, HOD (Dept. of Information Technology) and need a special mention for all the motivation and support.

I am deeply indebted to my guide Dr. Himangi Pande for completion of this seminar for which she has guided and helped me going out of the way.

For all the efforts behind the seminar report, I would also like to express my sincere appreciation to the staff of the Dept. of Information Technology, MIT Pune, for their extended help and suggestions at every stage.

**Satwik Ramchandre**

Roll No: 306074

T.E. (Information Technology)

MIT, Pune 411038

## Abstract

Data mining provides an user oriented approach to extract and analyze hidden patterns from a massive database. Medical domain contains heterogeneous data in the form of text, numbers, images that can be mined properly to provide a variety of useful information to the healthcare experts.

The patterns obtained from such healthcare data can be useful for physicians to predict various diseases, their severity, etc. Gathering such sensitive data in an efficient and precise manner is a challenging task , especially provided how large the amount of such data is. In this report , effective means of collection of healthcare data are discussed . Application based data collection being the most common method for data gathering , is not usually flexible enough, which is also discussed in this report. We also present an algorithm to mine sequential patterns from huge DNA patterns.

**Keywords:**

*Data mining, Data collection, healthcare, Algorithm*

## INDEX

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr. Topic**  **No** | | | **Page No.** |
|  |  |  |  |
| 1. | INTRODUCTION | |  |
|  | 1.1 | Introduction to the project | 4 |
|  |  |  |  |
|  | 1.2 | Motivation for the project | 4 |
|  | 1.3 | Aim and Objective | 4 |
|  | 1.4 | Introduction to the seminar topic | 4 |
|  |  | |  |
| 2. | LITERATURE SURVEY | |  |
|  | 2.1 | Crowd-sourced Data Collection of Physical Activity and Health Status: An App Solution | 4 |
|  |  |  |  |
|  | 2.2  2.3  2.4  2.5 | Mining Sequential Patterns from Uncertain Big DNA in the Spark Framework  Towards Flexible Mobile Data Collection in Healthcare  Health-CPS: Healthcare Cyber-Physical System Assisted by Cloud and Big Data  Privacy-Protected Data Collection in Wireless Medical Sensor Networks | 4  4  5  5 |
|  |  |  |  |
|  |  |  |  |
| 3.  4. | METHODOLOGY  3.1  3.2  3.3  ALGORITHMS……    4.1 | Application based data collection  Flexibility of application based data collection  Privacy protection in application based data collection  Algorithm for mining sequential DNA patterns | 5  5  5  5 |
|  |  |  |  |
| 5.  6. | CONCLUSION  REFERENCES | | 6  6 |
|  |  | |  |
|  |  | |  |

I]List of Figures

**I)List of figures**

**Sr.No Title Page No.**

1. Data collection layer in Health-CPS application
2. Flexibility concerns in application based data collection
3. Secret sharing scheme

**CHAPTER 1**

1. **INTRODUCTION**

**1.1 Introduction to Project**

Machine learning is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Usually, machine learning uses data mining techniques and another learning algorithm to build models of what is happening behind some data so that it can predict future outcomes.

Data mining:

The purpose of data mining, whether it’s being used in healthcare or business, is to identify useful and understandable patterns by analyzing large sets of data. These data patterns help predict industry or information trends, and then determine what to do about them.

In the healthcare industry specifically, data mining can be used to decrease costs by increasing efficiencies, improve patient quality of life, and perhaps most importantly, save the lives of more patients.

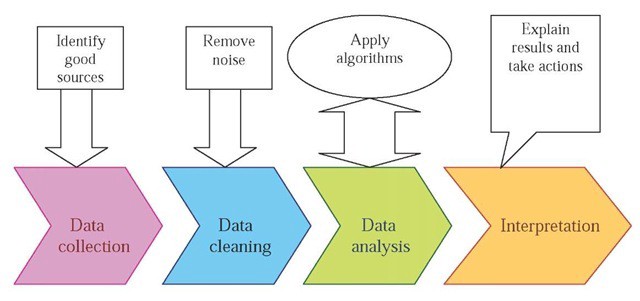
The data mining has 4 parts –

1) Data collection

2) Data Cleaning

3) Data Analysis

4) Data Interpretation and applications



**1.2 Motivation behind project topic**

The health industry collects huge amounts of health care data, which unfortunately are not “mined” to discover hidden information for effective decision making. Proper data mining is an extremely important step in increasing the reliability of raw data. The data needs to be collected from various reliable sources, then cleaned efficiently and analyzed properly to find out the true meaning hidden in the data to be used and applied effectively in the real world. Proper merging of data collection, data cleaning, data analyzing and data interpretation is needed so that the data mining process on healthcare is a success.

**1.3 Aim and Objective(s) of the work**

To design a system/framework that refines the raw data collected/generated in healthcare domain into valuable information which can be used for different applications

**1.4 Introduction to Seminar Topic**

The data which is collected from various homogeneous or heterogeneous cancer research data sources is dirty in nature, and needs to be cleaned in order to be analyzed and interpreted effectively. There are various data cleaning algorithms, from which a few have been studied in this seminar.

**1.4 Introduction to Seminar Topic**

Healthcare sector has tremendous amounts of data which is very useful in multiple number of ways. The data can be processed effectively using different techniques and algorithms to produce results and predict certain diseases. However , gathering of such data is a hurdle since the amount of data is large , and the precision and other factors like privacy, quality, flexibility need to be handled as well. The data is sensitive and must be as precise as possible, and also should be easily collected. In this seminar, we discuss about application based healthcare data collection , and an algorithm for mining of sequential DNA patterns from a big data DNA.

**CHAPTER 2**

**2. LITERATURE SURVEY**

**2.1 Crowdsourced Data Collection of Physical Activity and Health Status: An App Solution :** **Daniel Kelly, Brian Caulfield, and Kevin Curran**

**Publication:** ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2017 P. Perego et al. (Eds.): MobiHealth 2016, LNICST 192, pp. 151–159, 2017

Health status and related data and metrics are vital to understand a patient’s health. However, the precision and the means of measuring such data are very limited. In this paper, an app system for mobile phones is implemented to take into account the problem of lack of such health related data, and is gathered using sensors and a SF-36 questionnaire. Preliminary analysis of the data also shows a statistically significant correlation between the amount of time a participant is active and the health status of the participant.

**2.2 Mining Sequential Patterns from Uncertain Big DNA in the Spark Framework**

**Fan Jiang , Carson K. Leung , Oluwafemi A. Sarumi, and Christine Y. Zhang**

**Publication:** 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)

In bioinformatics, there is terabytes and more of data that can be easily generated using next generation sequencing technologies (NGS).However such data is usually filled with and some errors. This uncertain data is filled with many patterns which is the useful data needed to process further. This paper goes over an algorithm which is used to mine such sequential patterns, referred to as ‘motifs’ from huge amounts of DNA sequences. The algorithm uses collection of resilient distributed datasets (RDDs) to mine the motifs.

**2.3 Towards Flexible Mobile Data Collection in Healthcare**

**Johannes Schobel, Rudiger Pryss, Marc Schickler, Manfred Reichert**

**Publication:** 2016 IEEE 29th International Symposium on Computer-Based Medical Systems

Usually, the implementation of mobile healthcare applications for collecting patient data is cumbersome and time-consuming due to scenario-specific requirements as well as continuous adaptations to already existing mobile applications. This paper discusses flexibility issues considered by a generic and sophisticated framework for realizing mobile data collection applications. Thereby, flexibility is discussed along different phases of data collection scenarios. Altogether, the realized flexibility significantly increases the practical benefit of smart mobile devices in healthcare data collection scenarios.

**2.4 Health-CPS: Healthcare Cyber-Physical System Assisted by Cloud and Big Data**

**Zhang, Yin; Qiu, Meikang; Tsai, Chun-Wei; Hassan, Mohammad**

**Publication:** IEEE SYSTEMS JOURNAL

To provide a more convenient service and environment of healthcare, this paper proposes a cyber-physical system for patient-centric healthcare applications and services, called Health-CPS, built on cloud and big data analytics technologies. This system consists of a data collection layer with a unified standard, a data management layer for distributed storage and parallel computing, and a data-oriented service layer. The results of this study show that the technologies of cloud and big data can be used to enhance the performance of the healthcare system so that humans can then enjoy various smart healthcare applications and services.

**2.5 Privacy-Protected Data Collection in Wireless Medical Sensor Networks**

**Md Zakirul Alam Bhuiyan, Mdaliuz Zaman, Guojun Wang, Tian Wang and Jie Wu**

**Publication**: 2017 International Conference on Networking, Architecture, and Storage (NAS)

Medical data collection in healthcare monitoring applications through traditional frameworks raise serious concerns of patient data privacy and security, due to numerous security threats and attacks. This paper investigates the issues and concerns with privacy protected data collection and proposes a privacy protected data collection framework. A distributed database is considered consisting of multiple edge servers , where each server receives a part of the patient data.

**CHAPTER 3**

**3. METHODOLOGY**

This section discusses about the methodologies used for data collection, and the issues and possible improvements in data collection systems which are discussed about in the report.

**3.1 Application based data collection**

Health Status measurements, such as Health Related Quality of Life (HRQOL), are used as a means of quantifying the impact of chronic disease on a patients’ daily life .These measures are vital in understanding a patients’ health and their response to particular treatments and have become a central feature in many chronic disease studies Questionnaires are used to evaluate health status. However, evidence suggests questionnaire results are only useful in large groups and should not be relied upon on an individual basis.

Modern smartphones are equipped with multiple sensors. Smartphones can therefore enable a new type of data collection by harnessing the power of the crowd. Crowd-sourced data collection, using smart-phones, presents a major opportunity to collect sensor data from a large, and varied, set of participants.

Using motion sensors to gather data-   
The Accelerometer, built into a participants’ Smartphone, is used by the App to measure physical activity. Sensor data capture and recording is performed in the background, and data is recorded constantly while the App is enabled.

Two types of readings were proposed:

(1) Total Movement Duration (TMD) and

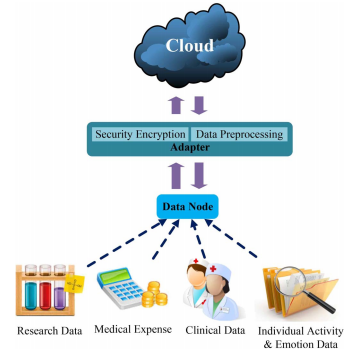
(2) Average Stationary Period (ASP).

TMD specifies the total amount of time in which the phone was detected as moving during a given day. The phone was deemed to be moving if the variance of the accelerometer magnitude was greater than a predefined threshold. For each 2 s window where the phone was deemed to be moving, 2 s were added to the overall TMD measure for that day.

ASP was calculated as the average of a set of stationary period durations for a given day. The set of stationary period durations store the set of times between when the phone stopped moving and when the phone started to move again .

Another application called as Health-CPS (Cyber Physical System) was also considered for this report. The application consists of various layers , however we shall only consider the Data Collection layer for our scope.

In the data collection layer, various healthcare data are collected by the data nodes and are transmitted to the cloud through the configurable adapters that provide the functionality to preprocess and encrypt the data.



The above figure shows an overview of the data collection layer used in the Health-CPS application.

Data node collects the data of various types, and thus the data node can be divided into four types:

1. **Research data** –

Drug research and development institutions and other scientific research institutions have accumulated a large amount of research data, such as clinical trial data and high-throughput screening data. These digital data, including individual or clinical gene or protein data, can help identify the drug side effect and the new effect

1. **Medical expense data-**

Medical behaviors generate massive expense data, such as medical bill and medical insurance reimbursement, which are not the traditional healthcare data, but it can be used to analyze and estimate the medical cost.

1. **Clinical Data -**

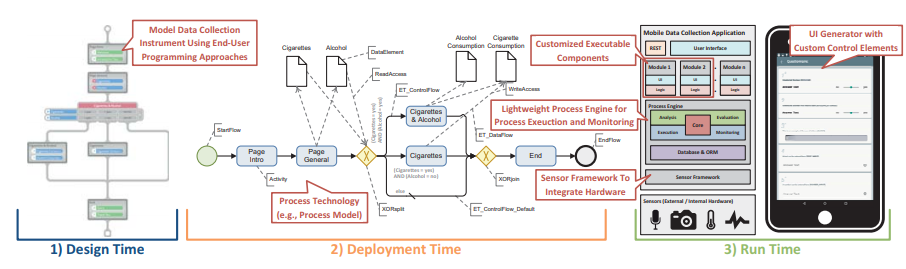
This is the typical medical data, usually collected by medical service providers for clinical diagnosis such as EMR, and medical image. These data can be unified, managed, and opened to researchers with a necessary precondition for ensuring the privacy of the patient.

1. **Individual activity and emotion data -**

This kind of data is not generated as a part of the healthcare sector, but is relevant to a user’s personal health. For example, retail consumption habits of a user, data collected from wearable devices, sensors, etc. Particularly for a recovering patient, doctors can monitor the emotions and the personal health of that patient.

**3.2 Flexibility of application based data collection**

With most of the data collection processes being done using smartphone based applications, a number of crucial requirements are identified on a large scale. The applications need to flexible right from the user interface, language and also adapt to changing requirements.



As seen in the above figure, flexibility issues can be considered in the three given phases as-

1. **Design time-**

Through a model-based approach domain experts are enabled to easily define the logic and structure of the data collection instrument themselves. Advanced wizards guide domain experts through the process of defining navigation paths for instruments. Finally, sensors for collecting data during run time are modeled on an abstract level.

1. **Deployment time-**

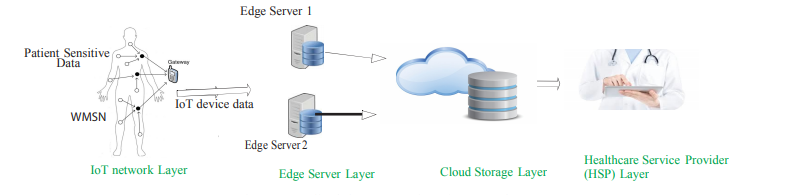
The designed data collection instrument is mapped to a process model and is installed on respective smart mobile devices. Process management technology, in turn, is utilized to ensure a flexible and robust execution of the process instance during data collection.

1. **Run time-**

Run time flexibility needs to address multiple issues. The first one being a lightweight framework for smooth execution of the process model which is run on the smartphone. Second, an advanced UI generator that combines all the executable components and UI fragments and displays them to the user. Third, a sophisticated sensor framework that allows both connecting internal and external sensors with the application. The combined use of the various technologies enables flexibility for all phases of mobile data collection applications.

* 1. **Privacy protection in application based data collection**

Healthcare monitoring applications are considered to be very promising, where patients can be monitored in hospitals or even at homes, using wireless sensor networks. These are connected to cloud computing systems which allows healthcare service providers to access and share the data.



As the above figure illustrates, a secret sharing scheme is introduced for the privacy protection. The general idea of a secret sharing scheme is that, a secret is encoded into a number shares. Each edge server receives one share. This is known as the (n, m) - threshold secret sharing in which any m or more shares can be used to reconstruct the secret and in which the size of a share is the same as the size of the secret. However, these secret sharing methods have heavy computation costs. In this framework, the authors have used Slepian-Wolf Secret Sharing Scheme (SWC-SS), which is more efficient in terms of computation, communication and storage costs.

**CHAPTER 4**

**4.** **ALGORITHMS**

* 1. **Mining sequential DNA patterns**

Given that ∑ = {A, C, G, T} is a set of finite DNA alphabets, where A (for Adenine), C (for Cytosine), G (for Guanine) and T (for Thiamine) are set of DNA characters or bases. A DNA sequence—denoted by S—over an alphabet ∑ is an ordered and finite list of alphabets from ∑.

For example, S = {ATTCGTATGTCTATAGTTGATTTG} is a DNA sequence from ∑, and S is denoted by 〈s1, s2, s3..., sn〉 where si ∈ ∑. And, |S| represents the length of sequence S. A motif—denoted by m—is a contiguous sub-sequence of sequence S that satisfies the following user-specified parameters:

1. A sub-sequence length (L);

2. An uncertainty factor (U), which is a measure of hamming distance among the sub-sequences; and

3. A minimum support threshold (δ).

**Overview of the algorithm-**

1. First, the algorithm reads DNA sequence from a text file, slits and caches RDD across all the memories of the workers.

2. Then, the algorithm generates and optimizes pairs of frequent subsequence.

3. Afterwards, the algorithm generates frequent patterns satisfying the user-specified uncertainty distance factor U and minimum support threshold δ=16.

4. Finally, the algorithm groups subsequences into different length and outputs the result to the driver**.**

**Details of the algorithm-**

STAGE I:

In Stage I, the algorithm reads the DNA sequence from a text file on the local disk, converts the sequence to RDD, and then partitions RDD across all the memories of the worker nodes. The algorithm also defines a distance function to estimate the similarities among the subsequences.

Pseudo Code:

*// Read DNA sequence from a text file; slit and cache RDD across all the*

*// memories of the workers.*

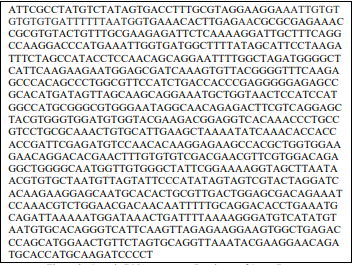
*val data ="file:///tmp/sequence.txt"*

*val seq = sc.textFile(data, 2).cache()*

*flatMap[U] (T->SeqU): RDD[U]*

*sc.broadcast(RDD): RDD[U]*

*Sample DNA sequence:*



STAGE II:

In Stage II, the algorithm uses a “for” loop to process the RDD using the minimum and maximum sub-sequence lengths defined in Stage I of the algorithm. It then applies the map operator to generate pairs of all subsequences of the input sequence on each partition of RDD on the worker nodes. It also uses the distinct operator to optimize pairs of the substrings and retain only unique pairs on each node.

Pseudo Code:

*// Generate and optimize pairs of frequent subsequence*

*sc.parallelize(*

*for{*

*i -> min\_len to max\_len*

*j-> 0 to (seq\_len – i)*

*} yield (j, i+j))*

*.map{case(j,i) ≥ S\_broadcast.value.substring(j, i)}*

*.distinct*

*Distinct sub sequences-*



STAGE III:

In Stage III, the algorithm generates frequent patterns satisfying the uncertainty distance function U and the minimum support threshold δ on each partition of RDD. The uncertainty distance function compares the pairs of the distinct sub-sequences generated from Stage II of the algorithm by calling the Hamming distance function declared in Stage I of the algorithm.

Pseudo Code:

*// Generate frequent patterns with minimum support and distance factor*

*flatMap(x=>*

*for{dist(*

*v <- δ to (seq\_len ÷ min\_len)*

*)} yield (x -> v) ≤ U)*

*.filter{case(x,v)=>S\_broadcast.value.indexOf(x\*v)≥0}*

*Sub sequences satisfying the minimum support threshold*



STAGE IV:

In Stage IV, the algorithm groups the frequent subsequences on each partition of RDD (on all the worker nodes) by their length, and uses the collect action to send the output from all partitions to the driver.

Pseudo Code :

*// Group subsequences into different length and output result to the driver.*

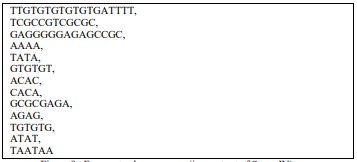
*.groupByKey*

*.map{case(k, v)=>k->v.max}*

*.collect*

*.map{case(k,v)=>k\*v}.foreach(println)*

*Frequent sub sequences-*



**CHAPTER 5**

**5. CONCLUSION**

In this report , we went through application based healthcare data collection , which is the most widely used technique to collect healthcare related data, and how the applications face issues of flexibility , and privacy protection of patient’s data. With ever changing needs in the healthcare sector, applications need to be flexible enough to be able to adapt to the newer requirements and also be secure enough to protect the data privacy from any kind of intrusions and threats. An algorithm to find out sequential patterns from a large DNA pattern was also discussed in the report. The extracted DNA sequences can be further used as sensitive healthcare data.

**CHAPTER 6**

**6. REFERENCES**

List of all the material that are used from various sources for making this project proposals:

*[1] Daniel Kelly, Brian Caulfield, and Kevin Curran*

*Crowdsourced Data Collection of Physical Activity and Health Status: An App Solution*

*ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2017 P. Perego et al. (Eds.): MobiHealth 2016, LNICST 192, pp. 151–159, 2017*

*[2] Fan Jiang , Carson K. Leung , Oluwafemi A. Sarumi, and Christine Y. Zhang*

*Mining Sequential Patterns from Uncertain Big DNA in the Spark Framework*

*2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*

*\*

*[3] Johannes Schobel, Rudiger Pryss, Marc Schickler, Manfred Reichert*

*Towards Flexible Mobile Data Collection in Healthcare*

*2016 IEEE 29th International Symposium on Computer-Based Medical Systems*

*[4] Zhang, Yin; Qiu, Meikang; Tsai, Chun-Wei; Hassan, Mohammad*

*“Health-CPS: Healthcare Cyber-Physical System Assisted by Cloud and Big Data*

*IEEE SYSTEMS JOURNAL*

*[5] Md Zakirul Alam Bhuiyan, Mdaliuz Zaman, Guojun Wang, Tian Wang and Jie Wu.*

*Privacy-Protected Data Collection in Wireless Medical Sensor Networks*

*2017 International Conference on Networking, Architecture, and Storage (NAS)*