International Journal of Science and Engineering Research (IJ0SER),

Vol 6 Issue 4 April -2018
3221 5687, (P) 3221 568X

Disease Prediction by Mining Unstructured data and dead data series of patient's historical record

[1] Harshini.K , [2]Israth Sulthana Begam.I [3]Harinitha.S, [4] S.Suresh Kumar
[1] [2] [3] Students SREC, [4]Assistant professor
Sri Ramakrishna Engineering College,
Tamilnadu,India

Abstract:-Disease Prediction by Machine Learning over Big Data from Healthcare Communities .Predicting patient at risk is challenging especially due to its heterogeneity, intrinsic noise, and particularly the large volume of unstructured data. In this paper, we introduced an effective and efficient graph-based semi-supervised algorithm namely SHG-Health to meet these challenges. We propose a new optimized SHG (Semi-supervised Heterogeneous Graph on Health) based multimodal disease risk prediction (MDRP) algorithm using structured and unstructured data from hospital data. We additionally implement dead data set to efficiently characterize the region based health data for incrementing accuracy of risk prediction. None of the existing work focused on both data types in the area of medical big data analytics. Compared to several typical prediction algorithms, the prediction accuracy of our proposed algorithm reaches 94.8% with a convergence speed which is faster than that of the existing risk prediction algorithm.

OBJECTIVE:-

- **To handle decease prediction problem in the presence of large structured as well as unstructured dataset.**
- **To perform flexible operations even if heterogeneity(in the presence of multiple symptoms for a single disease) appears in the data for an efficient risk prediction**

ALGORITHM:

- **SHG- Semi-supervised Heterogeneous Graph on Health**
- **multimodal disease risk prediction (MDRP) algorithm**

INTRODUCTION

According to a report by McKinsey , 50% of Americans have one or more chronic diseases, and 80% of American medical care fee is spent on chronic disease treatment. With the improvement of living standards, the incidence of chronic disease is increasing. The United States has spent an average of 2.7 trillion USD annually on chronic disease treatment. This amount comprises 18% of the entire annual GDP of the United States. The healthcare problem of chronic diseases is also very important in many other countries. In China, chronic diseases are the main cause of death, according to a Chinese report on nutrition and chronic diseases in 2015, 86.6% of deaths are caused by chronic diseases.

Therefore, it is essential to perform risk assessments for chronic diseases. With the growth in medical data , collecting electronic health records (EHR) is increasingly convenient. Besides, first presented a bioinspired high-performance heterogeneous vehicular telematics paradigm, such that the collection of mobile users' healthrelated real-time big data can be achieved with the deployment of advanced heterogeneous vehicular networks. Chen et.al – proposed a healthcare system using smart clothing for sustainable health monitoring. Qiu et al. had thoroughly studied the heterogeneous systems and achieved the best results for cost minimization on tree and simple path cases for heterogeneous systems. Patients' statistical information, test results and disease history are recorded in the EHR, enabling us to identify potential data-centric solutions to reduce the costs of medical case studies. Qiu et al. proposed an efficient flow estimating algorithm for the telehealth cloud system and designed a data coherence protocol for the PHR(Personal Health Record)-based distributed system. Bates et al. proposed six applications of big data in the field of healthcare. Qiu et al. proposed an optimal big data sharing algorithm to handle the complicate data set in telehealth with cloud techniques. One of the applications is to identify high-risk patients which can be utilized to reduce medical cost since high-risk patients often require expensive healthcare. Moreover, in the first paper proposing healthcare cyber-physical system, it innovatively brought forward the concept of prediction-based healthcare applications, including health risk assessment. Prediction using traditional disease risk models usually involves a machine learning algorithm (e.g., logistic regression and regression analysis, etc.), and especially a supervised learning algorithm by the use of training data with labels to train the model.

In the test set, patients can be classified into groups of either high-risk or low-risk. These models are valuable in clinical situations and are widely studied. However, these schemes have the following characteristics and defects. The data set is typically small, for patients and diseases with specific conditions the characteristics are selected through experience. However, these pre-selected characteristics maybe not satisfy the changes in the disease and its influencing factors. With the development of big data analytics technology, more attention has been paid to disease prediction from the perspective of big data analysis, various researches have been conducted by selecting the characteristics automatically from a large number of data to improve the accuracy of risk classification , , rather than the previously selected characteristics. However, those existing work mostly considered structured data. For unstructured data, for example, using convolutional neural network (CNN) to extract text characteristics automatically has already attracted wide attention and also achieved very good results.

However, to the best of our knowledge, none of previous work handle Chinese medical text data by CNN. Furthermore, there is a large difference between diseases in different regions, primarily because of the diverse climate and living habits in the region. Thus, risk classification based on big data analysis, the following challenges remain: How should the missing data be addressed? How should the main chronic diseases in a certain region and the main characteristics of the disease in the region be determined? How can big data analysis technology be used to analyze the disease and create a better model?

EXISTING SYSTEM / PROBLEM DESCRIPTION:-

Prediction using traditional disease risk models usually involves a machine learning algorithm (e.g., logistic regression and regression analysis, etc.).In the test set, patients can be classified into groups of either high-risk or low-risk. These models are valuable in clinical situations and are widely studied. However, these schemes have typically small data set, for patients and diseases with specific conditions; the characteristics are selected through experience. However, these pre-selected characteristics maybe not satisfy the changes in the disease and its influencing factors .Most existing classification methods on healthcare data do not consider the issue of unlabeled data. They either have expert-defined low-risk or control classes, or simply treat non-positive cases as negative. It does not consider about years and years of dead data set. Hence the prediction accuracy is poor.

DRAWBACKS:-

- Identifying participants at risk based on their current and past health records is important for early warning and preventive intervention. By "risk", we mean unwanted outcomes such as mortality and morbidity.
- Our geriatric health examination (GHE) dataset do not have a dead dataset
- Most existing classification methods on healthcare data do not consider the issue of multimodal risk predictions. They either have expert-defined low-risk or control classes.

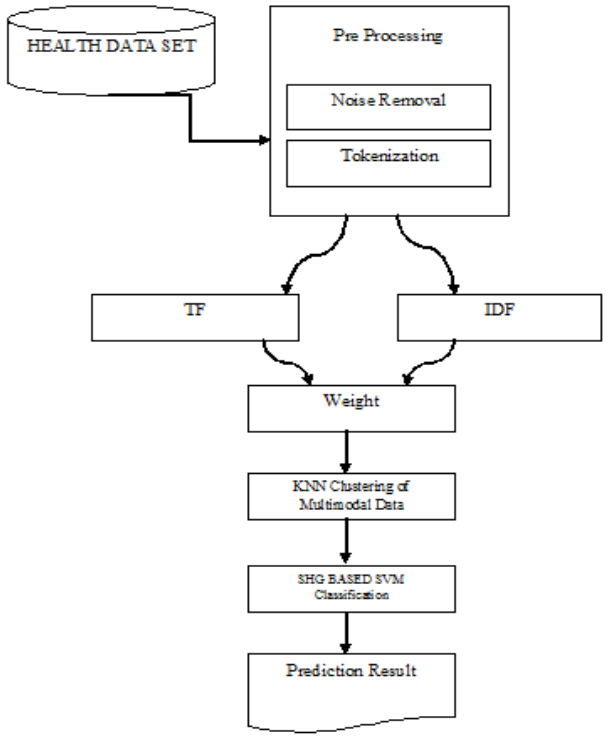
PROPOSED SYSTEM:-

Our scope is to implement a semi-supervised learning algorithm called SHG (Semi-supervised Heterogeneous Graph) for risk predictions of what will take place in the future to put in order a by degrees and to convert unknown patient profile into accurate possible risk prediction (possible disease based on the time period) system will work for both undiagnosed patient and the healthy one. With this system, people will be getting intimate precaution before even dealing with a disease. Hence, this system will lead to a healthy life. Our proposed methodology is applicable even in the presence of structured as well as unstructured data. . We additionally implement dead data set to efficiently characterize the region based health data for incrementing accuracy of risk prediction

ADVANTAGES:-

- It extracts features automatically for structured and unstructured data set
- It provides accurate possible risk predictions.
- It combines the structured and unstructured data in healthcare field to assess the risk of disease.
- It handle's a challenging multimodal classification problem

ARCHITECTURE DIAGRAM



1. DATA PREPROCESSING:

The dataset contains the huge amount of data. The data may be structured or unstructured in Dataset. If dataset will be unstructured means the preprocessing takes place. In preprocessing phase each and every transaction's are analyzed and determine the parameters are used in the transactions. Thus, the unstructured dataset is converted into structure dataset.

HER record

HUGE amounts of Electronic Health Records (EHRs) collected over the years have provided a rich base for risk analysis and prediction. An EHR contains digitally stored healthcare information about an individual, such as observations, laboratory tests, diagnostic reports, medications, procedures, patient identifying information, and allergies. A special type of HER is the Health Examination Records (HER) from annual general health check-ups. Identifying participants at risk based on their current and past HER's is important for early warning and preventive intervention. By "risk", we mean unwanted outcomes such as mortality and morbidity. With the help of HER record the data's are classified into two types

- Labeled Data
- Unlabeled Data

WEIGHT CALCULATION:

In this module, the weight has been calculated as area and disease wise, so we find the TF for area and disease and IDF for area and disease.

Term frequency:

In the case of the **term frequency** tf(*t*,*d*), the simplest choice is to use the *raw count* of a term in a document, i.e. the number of times that term *t* occurs in document *d*.

Inverse document frequency:

The **inverse document frequency** is a measure of how much information the word provides, that is, whether the term is common or rare across all documents. It is the logarithmically scaled inverse fraction of the documents that contain the word, obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient.

Weight calculation: TF*IDF.

CLUSTERING:

We are using K means Clustering for Cluster into three groups data, here *K*-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable *K*. The algorithm works iteratively to assign each data point to one of *K* groups based on the features that are provided.

SHG Based SVM Classification:-

Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.

Our proposed SHG-Health algorithm can be seen as combining the advantages of GSSL and GNet- Mine for solving a practical clinical problem of risk prediction from longitudinal health examination data with heterogeneity and large unlabeled data issues. To solve the problem of health risk prediction based on health examination records with heterogeneity and large unlabeled data issues.

FEASIBILTY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

The feasibility study investigates the problem and the information needs of the stakeholders. It seeks to determine the resources required to provide an information systems solution, the cost and benefits of such a solution, and the feasibility of such a solution. The analyst conducting the study gathers information using a variety of methods, the most popular of which are:

- Interviewing users, employees, managers, and customers.
- Developing and administering questionnaires to interested stakeholders, such as potential users of the information system.
- Observing or monitoring users of the current system to determine their needs as well as their satisfaction and dissatisfaction with the current system.
- Collecting, examining, and analyzing documents, reports, layouts, procedures, manuals, and any other documentation relating to the operations of the current system.
- Modeling, observing, and simulating the work activities of the current system.

The goal of the feasibility study is to consider alternative information systems solutions, evaluate their feasibility, and propose the alternative most suitable to the organization. The feasibility of a proposed solution is evaluated in terms of its components. These components are:

- ECONOMICAL FEASIBILITY
- TECHNICAL FEASIBILITY
- SOCIAL FEASIBILITY
- OPERATIONAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity.

The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

OPERATIONAL FEASIBILITY

The ability, desire, and willingness of the stakeholders to use, support, and operate the proposed computer information system. The stakeholders include management, employees, customers, and suppliers. The stakeholders are interested in systems that are easy to operate, make few, if any, errors, produce the desired information, and fall within the objectives of the organizat

SYSTEM IMPLEMENTATION

Implementation is the process that actually yields the lowest-level system elements in the system hierarchy (system breakdown structure). The system elements are made, bought, or reused. Production involves the hardware fabrication processes of forming, removing, joining, and finishing; or the software realization processes of coding and testing; or the operational procedures development processes for operators' roles. If implementation involves a production process, a manufacturing system which uses the established technical and management processes may be required.

The purpose of the implementation process is to design and create (or fabricate) a system element conforming to that element's design properties and/or requirements. The element is constructed employing appropriate technologies and industry practices. This process bridges the system definition processes and the integration process.

System Implementation is the stage in the project where the theoretical design is turned into a working system. The most critical stage is achieving a successful system and in giving confidence on the new system for the user that it will work efficiently and effectively. The existing system was long time process.

The proposed system was developed using .Net The existing system caused long time transmission process but the system developed now has a very good user-friendly tool, which has a menu-based interface, graphical interface for the end user. After coding and testing, the project is to be installed on the necessary system. The executable file is to be created and loaded in the system. Again the code is tested in the installed system. Installing the developed code in system in the form of executable file is implementation.

SYSTEM DEVELOPEMENT

A Systems Development Life Cycle (SDLC) adheres to important phases that are essential for developers, such as planning, analysis, design, and implementation, and are explained in the section below. A number of system development life cycle (SDLC) models have been created: waterfall, fountain, and spiral, build and fix, rapid prototyping, incremental, and synchronize and stabilize. The oldest of these, and the best known, is the waterfall model: a sequence of stages in which the output of each stage becomes the input for the next.

The waterfall model is a popular version of the systems development life cycle model for software engineering. Often considered the classic approach to the systems development life cycle, the waterfall model describes a development method that is linear and sequential. Waterfall development has distinct goals for each phase of development. Imagine a waterfall on the cliff of a steep mountain. Once the water has flowed over the edge of the cliff and has begun its journey down the side of the mountain, it cannot turn back. It is the same with waterfall development. Once a phase of development is completed, the development proceeds to the next phase and there is no turning back.

The advantage of waterfall development is that it allows for departmentalization and managerial control. A schedule can be set with deadlines for each stage of development and a product can proceed through the development process like a car in a carwash, and theoretically, be delivered on time. Development moves from concept, through design, implementation, testing, installation, troubleshooting, and ends up at operation and maintenance. Each phase of development proceeds in strict order, without any overlapping.

QUALITY ASSURANCE

Quality assurance comprises all those planned and systematic actions necessary to provide confidence that a structure, system or component will perform satisfactorily is service.

Quality assurance includes formal view of care, problem definition, corrective actions to remedy any deficiencies and evaluation of actions that to be taken.

The function of software quality that assures that the standards, processes, and procedures are appropriate for the project and are correctly implemented. This is an "umbrella activity" that is applied throughout the engineering process. Quality software is reasonably bug-free, delivered on time and within budget, meets requirements and/or expectations, and is maintainable.

The system is developed such that it ensures all the level of quality. It checks whether a user friendly environment is provided to the users and that there is a reliable, accurate and efficient flow of data within the system. The system also checks that due it contains the level of security required for the user. Hence as long as there is no hardware complaint, there is no problem with the software.

SYSTEM MAINTENANCE

Maintenance

The term "Software Maintenance" is used to describe software engineering activities. Maintenance activities involve making enhancements to software products, adapting to new environments and correcting problems. Software product enhancements may involve providing new functional capabilities, improving user displays and nodes of interaction, upgrading external documents and internal documentation or upgrading the performance characteristics of a system. Adaptation of software to a new environment may involve moving the software to a different machine, or for instance, modifying the software to accommodate a new telecommunication protocol or an additional disk drives. Problem correction involves modification and revalidation of software to correct errors.

Many activities performed during software development enhance the maintainability of a software product. They are:-

Analysis activities:

The analysis phase of software development is concerned with determining customer requirements and constraints and establishing feasibility of the product.

- Develop standards and guidelines
- Set milestones for supporting documents
- Specify quality assurance procedures
- Identify likely product enhancements
- Determine resources required for maintenance
- Estimate maintenance costs

Architectural Design Activities:

- Emphasize clarity and modularity as design criteria
- Design to ease likely enhancement
- Use standardized notations to document, data flows, functions, structure and interconnections
- Observe the principles of information hiding, data abstraction and top-down hierarchical decomposition

Detailed Design Activities

- Use standardized notations to specify algorithms, data structures and procedure interface specifications
- Specify side effects and exception handling for each routine

Implementation activities

- Use single entry, single exit constructs
- Use standard indentation of constructs
- Use simple, clear coding style
- Use symbolic constants to parameterize routines
- Provide margins on resources
- Provide standard documentation
- Follow standard internal commenting guidelines

Other activities:

- Develop a maintenance guide
- Develop a test suite
- Provide test suite documentation

CONCLUSION

In this paper, we propose a new convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm using structured and unstructured data from hospital. To the best of our knowledge, none of the existing work focused on both data types in the area of medical big data analytics. Compared to several typical prediction algorithms, the prediction accuracy of our proposed algorithm reaches 94.8% with a convergence speed which is faster than that of the CNN-based unimodal disease risk prediction (CNNUDRP) algorithm.

Textual Reference

1. Bill Hamilton, "**Programming SQL Server 2005**", O'Reilly Media Publisher, 2006.
2. Elias M.Award,"System Analysis and Design", Galgotia Publications, Second Edition.
3. Daniel Solis, "Illustrated C# 2008", Apress Publisher, 2008.
4. David B. Makofske, Michael J. Donahoo, Kenneth L. Calvert, "TCP/IP Sockets in C#", Academic Press Publishers, 2004.
5. Richard Blum, "C# Network Programming", John Wiley & Sons Publishers, 2006.
6. Robin Dewson, "**Pro SQL Server 2005",** Apress Publisher.
7. Roger S. Pressman,"Software Engineering", Fourth Edition, 2005.

Online Reference

- www.dotnetspider.com
- www.programersheaven.com
- www.sql-server-performance.com
- www.developerfusion.com
- www.winsocketdotnetworkprogramming.com