International Journal of Scientific Research in Computer Science and Engineering
Research Paper
Vol.6, Issue.2, pp. 96-99, April (2018)
E-ISSN: 2320-7639

# Mining Health Data in Multimodal Data Series for Disease Prediction

**R. Anupriya[1*], P. Saranya[2], R. Deepika[3]**

[1]Computer science and engineering, Sree Sakthi Engineering College,Anna University, Coimbatore, India
[2]Computer science and engineering, Sree Sakthi Engineering College,Anna University, Coimbatore, India
[3]Computer science and engineering, Sree Sakthi Engineering College,Anna University, Coimbatore, India

**Corresponding author: anupriyag1771@gmail.com*

**Available online at: www.isroset.org**

Received: 28/Feb/2018, Revised: 10/Mar/2018, Accepted: 31/Mar/2018, Online: 30/Apr/ 2018

***Abstract***— Disease Prediction plays a major role in health care community. With data mining process, disease will be predicted from large number of data. Dataset may be structured or unstructured. If the dataset is unstructured then the latent factor model is used to convert unstructured to structured data and it is very complex to predict a disease using unstructured data. Therefore we use synthetic data, which is structured. We concentrate on different kind of diseases. We propose a convolutional neural network based multimodal disease risk prediction (CNN-MDRP). Here datasets are stored as HER records. K-means clustering algorithm is used to group the datasets. Semi-Supervised Heterogeneous algorithm is applied to grouped data to predict the disease.

***Key Words:*** Data Mining, Disease Prediction, HealthCare, Multimodal, K-means, SVM classification.

## I. INTRODUCTION

Data mining is the process of extracting data from large set of dataset. Usage of data mining in various fields like medical, marketing & so on. Multiple algorithms are used to segment the data. To enhance the value of existing information resources, the data mining techniques is used in software and hardware platforms. There are four stages of data mining and they are Data sources, Data gathering, Modeling and Deploying models. It is used in machine learning to predict the future outcomes. All the data should be assembled properly before using the data mining technique. The cost is low when using the data mining techniques.

Now-a-days, people get affected by various diseases due to surroundings, food. In the World, 75% of the people get affect by various diseases. In India, 80% of the people affected by Diarrhea, Malaria, HIV/AIDS, Typhoid, Diabetes, etc., the people spent more money to medical. And they don’t have precaution about the disease. People mostly prefer junk foods always and wherever they go. It may lead to disease. There may be communicable and non-communicable diseases. Currently people get various symptoms and they don’t have any awareness about disease that they get affected. So it is necessary to take necessary steps about the health of the people. To that with the help of data mining the patient will get precaution about the disease through their symptoms. If the patient provides necessary information about his/her health i.e.., symptoms then through various algorithm and so on to find the disease. For Example, if the patient has cough then it may be cancer or dengue; the diseases are predicted based on the two or more symptoms of the patient.

## II. RELATED WORK

In Existing Paper, a chronic disease is the main cause of death in china. As the disease is very dangerous in other countries. Percentage of death is increasing year by year. They collected a dataset from Central china hospital. As the data is of both structured and unstructured. They combine both data in order to predict disease. Latent Factor model is used to reconstruct missing data which is being collected from central china. Then using statistical knowledge the important disease is found in part of the region. Handling of structured data is to consult with experts to gather information.

The dataset used in existing contains real-life hospital data and it is stored in data center. To protect that, they have created a security access mechanism. They used Three year dataset (2013-2015). Structured data contains patient’s basic information like age, gender and their habits, etc. While unstructured contains illness of his/her and doctor narration. They focus on three dataset to find a conclusion.

*Structured data(S- data)
*Text data (T- data)
*Structured and text data(S&T- data)

© 2018, IJCSE All Rights Reserved

They used C++ languages to realize machine learning and deep learning algorithm. For S- data with discussion to doctor and person's analysis, they extract patient's demographics and some information about the disease and living habits. For T- data, Word embedding is to learn in the text. Machine learning and deep learning algorithm is used here. For S- data, algorithms like Conventional machine learning, K- nearest neighbor and Decision Tree are used to predict disease. For T- data, they proposed CNN-based unimodal disease risk prediction. For evaluation, TP, FP, TN and FN is denoted for true positive, false positive, true negative and false negative.

Data imputation is a method in health data process. In patient's data there will be of missing data, so it is necessary to fill structured data. Before it takes place we need to identify incomplete data to modify and to delete for data quality. Here latent factor model is used to explain in terms of latent variables. Assuming Rm×n where row designation m represent total number of patient and column designation n represent number of attribute of patient. They used stochastic gradient descent method to solve this. And finally they can able to fill missing values.

Next method is Convolutional Neural Network based Unimodal Disease Risk Prediction (CNN-UDRP). It is being performed under five steps.

Next method is Convolutional Neural Network based Multimodal Disease Risk Prediction (CNN-MDRP). They design CNN-MDRP algorithm based on CNN-UDRP. It is performed under two parts.

They use various algorithms to predict the chronic disease. By conclusion the disease prediction accuracy depends on features of hospital.

## III. METHODOLOGY

As we know data mining place a big part in medical field. Due to slight complication in finding the disease, we proposed a disease prediction project based on data mining. In our project, we use Convolutional Neural Network-Multimodal Disease Risk Prediction (CNN-MDRP) algorithm to predict the disease. The CNN-MDRP algorithm is used to find the risk for multiple diseases. Then k-means clustering algorithm is used to group the dataset, and a semi-supervised heterogeneous algorithm for predicting the diseases.

### 1. HER HEALTH DATA PROCESSING:

Dataset is a collection of multiple datas that is about the patients details. The datasets are collected to predict the disease based on symptoms of the patients. The dataset may be structured or unstructured. If dataset are collected from the hospitals then it is in the form of unstructured, and these unstructured dataset are converted into structured using the latent factor model. It is hard to find the risk from the unstructured data. So we used synthetic dataset which contains structured data only .The dataset are stored electronically and named as Health Examination Record (HER) Record.

### 2. WEIGHT CALCULATION:

The weight will be calculated for finding that, how much information the word gives. In this the weight will be calculated based on the TF, IDF.

- **Term Frequency:** The TF defines that, the count of the word that is how much time the word present in the document& the frequency of the each term is calculated. The weight of the term which is occurring in the document is proportional to the term frequency of the word.

**TF=No. of times token appears in the document/Total No. of Tokens in the document**

- **Inverse document frequency:** This means how much information the word provides.

  **IDF=Total no. of documents/No. of documents in U that contain t**

**Weight calculation: TF*IDF.**

### 3. CLUSTERING:

K-Means Clustering is used for cluster the dataset into three groups, though it is type of unsupervised learning, The main goal of this algorithm is to find a groups in the data, with number of groups represented by a variable *K*.

### 4. SEMI-SUPERVISED HETEROGENOUS (SVM Based):

An SVM model is a new classification approach for predicting the disease. SVM is semi-supervised learning technique and it tries to classify by finding a separation boundary. In this, the patient details are compared with the dataset. The symptoms of the patient are compared to the symptoms in the document. If any matching occurs, the prediction will be provided to the patient. This is performed algorithmically.

## IV. RESULTS AND DISCUSSION

- **HEALTHCARE DATASET:**

  The HER record consider only about the structured dataset. The hospital won't provide their patient details, so we collect the synthetic data. If the dataset is unstructured, then the latent factor model is used to convert into structured. The dataset is stored in database and named as HER (Health Examination Record) records.

- **CLUSTERING:**

  After collecting the dataset , the weight will be calculated for finding the frequency of the each word in the document. The weight will be calculated based on TF,IDF. After this, the k-means clustering algorithm is used to group the datasets. The Clustering is performed based on the frequency of the word. The result of this module is, the K number of the groups are created.

- **SEMI-SUPERVISED HETEROGENEOUS:**

  In this module, the patient will get awareness about the disease. By giving their symptoms they will know about, they affected by what kind of disease. The symptoms of the patient is compared to the symptoms in the dataset to find the disease. If any matching occur, then it will be informed to the patient. Finally the patient will get name of disease.

**Discussion:**

In this paper, the HER records are collected that is synthetic, which is structured only. From the HER dataset the weight will be calculated using the terms TF, IDF for each words to find the frequency of the words. The K-means clustering is used to group the dataset based on the weight. The number of groups are represented as "K". Finally by applying the SVM classification the patient will get information about the disease using the symptoms of the patient. There is some complexity when using the unstructured dataset, So we don't consider about the unstructured dataset.

## Conclusion and Future Scope

The main conclusions of the study may be presented in a short Conclusion Section. In this section, the author(s) should also briefly discuss the limitations of the research and Future Scope for improvement.

## Abbreviations and Acronyms

| | |
|---|---|
| **CNN** | **- Convolutional Neural Network** |
| **UDRP** | **- Unimodal Disease Risk Prediction** |
| **MDRP** | **- Multimodal Disease Risk Prediction** |
| **HER** | **- Health Examination Record** |
| **TF** | **- Term Frequency** |
| **IDF** | **- Inverse Document Frequency** |
| **SVM** | **- Support Vector Machine** |

## Equations:

**TF=No. of times token appears in the document/Total No. of Tokens in the document**

**IDF=Total no. of documents/No. of documents in U that contain t**

**Figures and Tables**

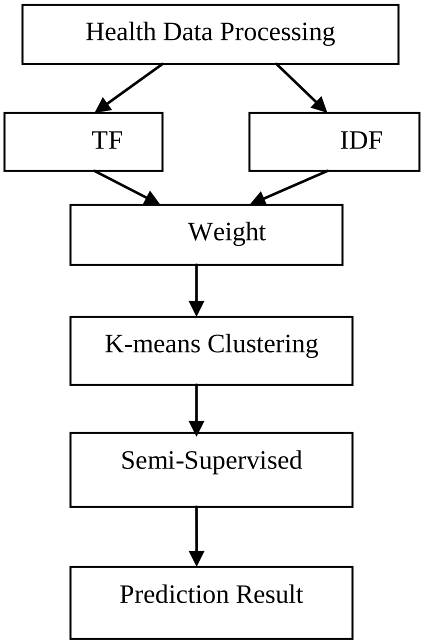

Fig.1. Proposed Architecture

### References

[1] Min Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang*, "Disease Prediction by Machine Learning over Big Data from Healthcare Communities", vol.5, Pages: 8869 – 8879,2017.

[2] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," Nature Reviews Genetics, vol. 13, no. 6, pp. 395–405, 2012.

[3] D. Oliver, F. Daly, F. C. Martin, and M. E. McMurdo, "Risk factors and risk assessment tools for falls in hospital in-patients: a systematic review," Age and ageing, vol. 33, no. 2, pp. 122–130, 2004.

[4] P. Groves, B. Kayyali, D. Knott, and S. V. Kuiken, "The'big data'revolution in healthcare: Accelerating value and innovation," 2016.

[5] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," Mobile Networks and Applications, vol. 19, no. 2, pp. 171–209, 2014.

[6] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, "A relative similarity based method for interactive patient risk prediction," Data Mining and Knowledge Discovery, vol. 29, no. 4, pp. 1070–1093, 2015.

[7] A. Singh, G. Nadkarni, O. Gottesman, S. B. Ellis, E. P. Bottinger, and J. V. Guttag, "Incorporating temporal ehr data in predictive models for risk stratification of renal function deterioration," Journal of biomedical informatics, vol. 53, pp. 220–228, 2015.

[8] J. Wan, S. Tang, D. Li, S. Wang, C. Liu, H. Abbas and A. Vasilakos, "A Manufacturing Big Data Solution for Active Preventive Maintenance", IEEE Transactions on Industrial Informatics, DOI: 10.1109/TII. 2017.2670505, 2017.

[9] W. Yin and H. Sch¨utze, "Convolutional neural network for paraphrase identification." in HLT-NAACL, 2015, pp. 901–911.

[10] Nori, H. Kashima, K. Yamashita, H. Ikai, and Y. Imanaka, "Simultaneous modeling of multiple diseases for mortality prediction in acute hospital care," in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015,pp. 855–864.

## AUTHOR BIOGRAPHY:

**R.Deepika** received her B.E degree in computer science and Engineering in R.V.S college of Engineering and technology,Coimbatore in 2012. She did her M.E degree in computer science and engineering in Avinashilingam Institute for Home science and Education for Women University Tamil Nadu, India . Now she is working as Assistant Professor in Sree Sakthi Engineering College.

**R.Anupriya** persuing B.E degree in Computer science And Engineering in Sree Sakthi Engineering College, Karamadai, Coimbatore.

**P.Saranya** persuing her B.E degree in computer science and Engineering in Sree Sakthi Engineering College, Karamadai, Coimbatore.