

# Mining Health Examination Records - Disease Detection

Miss. Rituja A. Bibave<sup>1</sup> Dr. Baisa L. Gunjal<sup>2</sup>

<sup>1</sup>ME Student <sup>2</sup>Professor

<sup>1,2</sup>Department of Computer Science & Engineering

<sup>1,2</sup>Amrutvahini COE, Sangamner, Maharashtra India

*Abstract*— Normally, health examination is an important method which can be used in multiple countries to identify the health records. To identify the risk factors which are warning and prevention in many diseases is important. This is the major challenge to classify this risk factors used in unlabeled data which contains the dataset. Health state condition can changes rapidly from healthy to very-ill. So, unlabeled data contains records of such health examination. There is no specific base for differentiating the state of health process. To identify and classify the risk prediction in unlabeled data multiple algorithms are used and implemented. To propose a graph-based, semi-supervised learning algorithm called SHG Health (Semi-supervised Heterogeneous Graph on Health) is used for risk predictions. So many efficient health learning technique is available to recognize any unlabeled dataset. The algorithms used to predicate the risk factors is based on both real health examination datasets show more effectiveness and efficiency.

**Key words:** Health Examination Records, Heterogeneous Graph Extraction, Semi-Supervised Learning

## I. INTRODUCTION

Today, huge amount of data collected to provide a rich base for electronic health records (EHRs) for risk analysis and prediction. An EHR stored healthcare information digitally about an individual. This digital information contains healthcare information such as analysis, various laboratory tests, records diagnostic reports, used medications, procedure used, patient recognize data, various allergies and diseases. Health Examination Records (HERs) is special type of EHR. An HER update its dataset from annual health check-ups. For example, governments in some countries such as Australia, U.K. implement periodic health examination as major part of their aged care programs. Healthcare system specifies difference in HER and EHR. Only limited and small sets of measures which are considered as necessary are collected and stored in patients EHR. Whereas, HERs collect and stored dataset for regular surveillance and preventive purposes, also maintain this dataset in manageable format. HER used both current and past HERs for early warning and preventive intervention. By risk, too mean unwanted outcomes such as normal ill and very ill. Identifying the participants which cause various diseases earlier is important for early warning and preventive techniques. This is the major challenge to classify risk factors used in unlabeled data which contains the datasets. Health condition can change rapidly from healthy to very-ill. So, unlabeled data contains records of such health examination. There is no special base facility for differentiating state of health process.

## II. LITERATURE REVIEW

### A. *Extraction of interpretable multivariate patterns for early diagnostics*

Leveraging temporal observations to predict a patient's health state at a future period is a very challenging task. Providing such a prediction early and accurately allows for designing a more successful treatment that starts before a disease completely develops. Information for this kind of early diagnosis could be extracted by use of temporal data mining methods for handling complex multivariate time series [2]. A temporal data mining method is proposed for extracting interpretable patterns from multivariate time series data, which can be used to assist in providing interpretable early diagnosis. The problem is formulated as an optimization based binary classification task addressed in three steps. First, the time series data is transformed into a binary matrix representation suitable for application of classification methods. Second, a novel convex-concave optimization problem is defined to extract multivariate patterns from the constructed binary matrix [2].

### B. *Stabilized sparse ordinal regression for medical risk stratification*

The recent wide adoption of Electronic Medical Records (EMR) presents great opportunities and challenges for data mining. The EMR data is largely temporal, often noisy, irregular and high dimensional. This paper constructs a novel ordinal regression framework for predicting medical risk stratification from EMR [4]. First, a conceptual view of EMR as a temporal image is constructed to extract a diverse set of features. Second, ordinal modeling is applied for predicting cumulative or progressive risk. The challenges are building a transparent predictive model that works with a large number of weakly predictive features, and at the same time, is stable against re-sampling variations. Our solution employs sparsity methods that are stabilized through domain-specific feature interaction networks. We introduces two indices that measure the model stability against data re-sampling. Feature networks are used to generate two multivariate Gaussian priors with sparse precision matrices (the Laplacian and Random Walk). We apply the framework on a large short-term suicide risk prediction problem and demonstrate that our methods outperform clinicians to a large-margin, discover suicide risk factors that conform with mental health knowledge, and produce models with enhanced stability [4].

### C. *Predicting the risk of exacerbation in patients with chronic obstructive pulmonary disease using home tele-health measurement data*

Chronic obstructive pulmonary disease (COPD) is responsible for significant morbidity and mortality worldwide. Recent clinical research has indicated a strong association between physiological homeostasis and the onset

of COPD exacerbation. Thus the analysis of these variables may yield a means of predicting a COPD exacerbation in the near future. However, the accuracy of existing prediction methods based on statistical analysis of periodic snapshots of physiological variables is still far from satisfactory, due to lack of integration of long term and interactive effects of the physiological variables. Therefore, developing a relatively accurate method for predicting COPD exacerbation is an outstanding challenge [3]. In this paper, a regression-based machine learning technique was developed, using trend pattern variables extracted from COPD patients longitudinal physiological records, to classify subjects into low-risk and high-risk categories, indicating their risk of suffering a COPD exacerbation event. Experimental results from cross validation assessment of the classifier model show an average accuracy of 79.27 percent using this method [3].

### III. PROBLEM STATEMENT

The Problem is to determine how to handle Automatically Learning Disease Problem as per Patient Queries by using Semi-Supervised Algorithm.

### IV. PROPOSED SYSTEM

To propose a graph-based, semi-supervised learning algorithm called SHG-Health or risk predictions to classify a progressively developing situation with the majority of the data unlabeled. An efficient iterative algorithm is designed and the proof of convergence is given. Extensive experiments based on both real health examination datasets and synthetic datasets are performed to show the effectiveness and efficiency of our method. Firstly, health examination records are represented as a graph that associates all relevant cases together. This is especially useful for modeling abnormal results that are often sparse. Secondly, multi-typed relationships of data items can be captured and naturally mapped into a heterogeneous graph. Particularly, the health examination items are represented as different types of nodes on a graph, which enables our method to exploit the underlying heterogeneous sub graph structures of individual classes to achieve higher performance. Thirdly, features can be weighted in their own type through a label propagation process on a heterogeneous graph. These in-class weighted features then contribute to the effective classification in an iterative convergence process.

### V. SYSTEM ARCHITECTURE

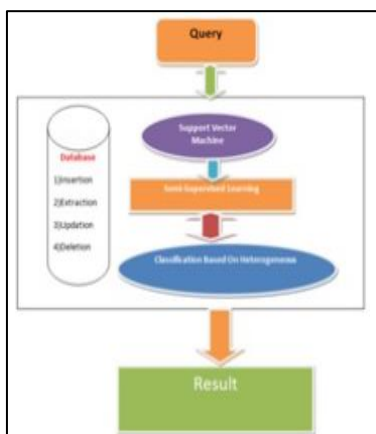


Fig. 1: System Architecture

Following figure represents the system architecture of proposed system. We have created a desktop system for end user that mines the risk factors for early prevention.

#### A. Algorithm

##### 1) SVM-Algorithm

- Step 1. Start.
- Step 2. . First select the query Q (Z).
- Step 3. Remove unwanted word from query.
- Step 4. Vector model process, Divide neural word from one site and non-neural word from one site.
- Step 5. Matching Relevance word->Semi- Supervised Learning.

##### 2) SHG-Algorithm

- Step 1. Start.
- Step 2. Binarization
- Step 3. Node Insertion
- Step 4. Node Typing
- Step 5. Link Insertion
- Step 6. Result.

### VI. MATHEMATICAL MODEL

S is the system such that

$$S = I, F, O$$

Where,

I is the input to the system

F is system functions

O is Systems output

Function F can be defined as:

$$F=Q, Sv, U, P, D$$

Where,

- Q(Z)= Query
- S(Z)= solved Disease Problem
- U(Z)= Un-Solved Disease Problem
- P(Z)= Patient
- D(Z)= Doctor

Such that,

- Q(Z) = q1,q2,q3.....
- S(Z) = s1,s2,s3,.....
- U(Z) = u1,u2,u3.....
- P(Z) = p1,p2,p3.....
- D(Z) = d1,d2,d3.....

Patient check own query in this technique through Semi-Supervised Learning Algorithm. Easily Known which type of disease occurs which disease is very easily recover. Also, easily Crack which doctor treatment is best.

#### A. Success Condition

Easily Kwon disease which is solved unsolved.

#### B. Failure Condition

Disease data doses not available then problem to recognized disease.

## VII. EXPERIMENTAL SETUP

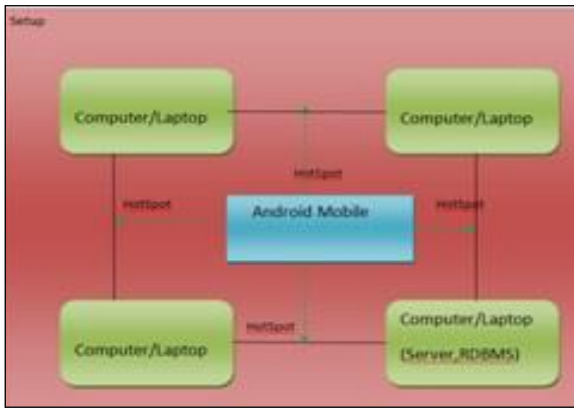


Fig. 2: Experimental Setup

## VIII. RESULT ANALYSIS

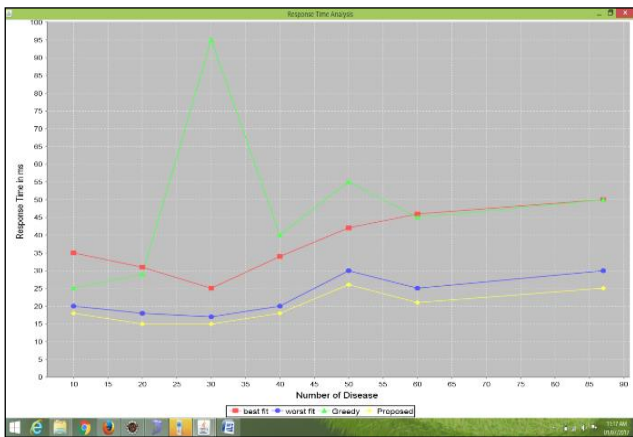


Fig. 3: Time Analysis of proposed vs. Existing System  
Above Fig show how much time too taken find label & unlabel disease. Proposed System is required minimum time as compared to Existing System.

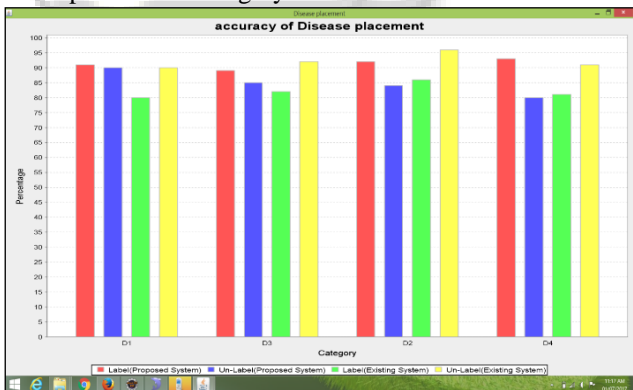


Fig. 4: Accuracy of proposed vs. Existing System  
Above Diagram shows accuracy of the label & unlabeled disease of proposed and existing System.

## IX. CONCLUSION AND FUTURE WORK

In enhancement technique, as per patient queries we can recognize which precaution needs to be taken. Confirmation about precaution can be given by mail or by own account, also can be recommended to relevant patients friend. Finally we conclude that how we can manage health care dataset and also help people live happily. In future work, by using SHG algorithm techniques with efficient classification techniques we can provide better service to the patients.

## ACKNOWLEDGMENT

A very firstly I gladly thanks to my project guide Dr. B. L. Gunjal, for her valuable guidance for implementation of proposed system. I will forever remain a thankful for their excellent as well as polite guidance for preparation of this report. Also I would sincerely like to thank to HOD of computer department Mr.R.L.Paikrao and other staff for their helpful coordination and support in project work. I also thankful to our principal Dr. M. A. Venkatehsh.

## REFERENCES

- [1] J. M. Toi, S. Q. Wang, and X. J. Yuan, Ensemble rough hypercuboid approach for classifying cancers, *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 3, pp. 381391, 2010.
- [2] M. F. Ghalwash, V. Radosavljevic, and Z. Obradovic, Extraction of interpretable multivariate patterns for early diagnostics, *IEEE International Conference on Data Mining*, pp.201210, 2013.
- [3] E. Kontio, A. Airola, T. Pahikkala, H. Lundgren-Laine, K. Junttila, H. Korvenranta, T. Salakoski, and S. Salantera, Predicting patient acuity from electronic patient records. *Journal of Biomedical Informatics*, vol. 51, pp. 813, 2014.
- [4] T. Tran, D. Phung, W. Luo, and S. Venkatesh, Stabilized sparse ordinal regression for medical risk stratification, *Knowledge and Information Systems*, pp. 128, Mar. 2014.
- [5] Q. Nguyen, H. Valizadegan, and M. Hauskrecht, Learning classification models with soft-label information. *Journal of the American Medical Informatics Association: JAMIA*, vol. 21, no. 3, pp. 5018, 2014.
- [6] L. Chen, X. Li, S. Wang, H.-Y. Hu, N. Huang, Q. Z. Sheng, and M. Sharaf, Mining Personal Health Index from Annual Geriatric Medical Examinations, in 2014 *IEEE International Conference on Data Mining*, 2014, pp. 761766.
- [7] G. J. Simon, P. J. Caraballo, T. M. Therneau, S. S. Cha, M. R. Castro, and P. W. Li, Extending Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus, *IEEE Transactions Knowledge and Data Engineering*, vol. 27, no. 1, pp. 130141, 2015.
- [8] S. Pan, J. Wu, and X. Zhu, CogBoost: Boosting for last Cost sensitive Graph Classification, *IEEE Transactions on Knowledge and Data Engineering*, vol. 6, no. 1, pp. 11, 2015
- [9] M. S. Mohktar, S. J. Redmond, N. C. ntoniades, P. D. Rochford, J. J. Preto, J. Basilakis, N. H. Lovell, and C. F. McDonald, Predicting the risk of exacerbation in patients with chronic obstructive pulmonary disease using home telehealth measurement data, *Artificial Intelligence in Medicine*, vol. 63, no. 1, pp. 5159, 2015.