International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 3 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT1953165

Mining Sequential Risk Patterns for Early Assessment of COPD

Saniya PK*, Bineesh V.
Computer Science and Engineering, MEA Engineering College, Perinthalmanna, Kerala, India

ABSTRACT

Nowadays, chronic diseases have been among the major concerns in medical fields since they may cause heavy burden on healthcare resources and disturb the quality of life. Chronic Obstructive Pulmonary Disease (COPD) is one kind of popular chronic diseases. COPD takes long period to evolve from mild symptoms (Stage I) to severe illness (Stage IV) and death. Earlier the disease is detected, the better the scope for effective treatment and improved control of symptom development. Therefore, the early of COPD is beneficial for better treatment. The paper examines the novel system for early appraisal on chronic illnesses by mining sequential risk patterns with interim data from diagnostic clinical records utilizing sequential rules mining and classification modelling systems. The system consists of four phases namely data pre-processing, risk pattern mining, classification modelling. SPADE algorithm and CBS algorithm used for risk pattern mining and classification. Decision tree algorithm is compared with the SPADE algorithm, and SPADE showing a better accuracy when comparing with decision tree.

Keywords : COPD, Sequential pattern, SPADE, CBS, Decision tree, Classification

I. INTRODUCTION

The computerization of our public has significantly upgraded our abilities for both creating and gathering data from different sources. A huge measure of data has overflowed every part of our lives. Data mining can be viewed because of the natural evolution of information technology. The database and data management industry evolved in the development of several critical functionalities: data collection and database creation, data (including data storage and retrieval and database transaction processing), and advanced data analysis (involving data warehousing and data mining). The early development of data collection and database creation mechanisms served as a prerequisite for the later development of effective mechanisms for data storage and retrieval, as well as query and transaction processing. Nowadays numerous database systems offer query and transaction processing as common practice. Advanced data analysis has naturally become the next step [7].

Nowadays, chronic diseases have been among the major concerns in medical fields since they may cause heavy burden on healthcare resources and disturb the quality of life. Chronic Obstructive Pulmonary Disease (COPD) is one kind of popular chronic diseases. The World Health Organization (WHO) predicts that COPD will become the third leading cause of death round the world in 2030. COPD takes long period to evolve from mild symptoms (Stage I) to severe illness (Stage IV) and death. Tobacco smoking, air pollution, occupational dust, etc often cause it. However, COPD is difficult to be diagnosed in the early stage due to the complex causes. Consequently, patients diagnosed with COPD are often on terminal illness and cannot be adequately treated. So earlier the disease is detected, the better the scope for effective treatment and improved control of symptom development [2]. So the early assessment of COPD is beneficial for better treatment.

Existing studies on analysis of chronic diseases have involved adopting different data mining methods. C. Y. Chin *et al.* [3] proposed a novel framework, which combines association rule mining and classification techniques to detect Rheumatoid Arthritis (RA). Soni et al. [4] presented an overview about using data mining to predict heart disease and analyzed supervised classification methods K-NN for heart disease prediction. Asha et al. [1] proposed a framework which combines Classification and Association Rule Mining(CARM) method and CBA(Classification Based Association) for detecting tuberculosis (TB).

However, the existing papers as mentioned above have leave unnoticed the important factors of sequence patterns for making prediction on health risks. Most of the related studies affected the data of physiological measurements, which have to be recorded for a long time. As a result, a small number of clinical observations limit the number of risk patterns. Besides, they didn’t consider the risk patterns in form of sequence patterns, which hold important factors for making prediction in health risks.

The major offerings of this work lies in that sequential risk patterns can be pull out effectively through the proposed framework for strengthen opportunities in early assessment of chronic diseases like COPD. Besides, the uncovered patterns may reveal valuable insights for further inspection by medical researchers to discover new markers and better treatment for chronic diseases.

The organization of this document is as follows. The remainder of this paper is as follows. In Section II, a detailed explanation of the paper, the description of the dataset is mentioned. In Section III, the result analysis and discussions are explained and finally in section IV, the research is concluded with the future scope.

II. METHODS AND MATERIAL

A novel framework is proposed for mining successive risk patterns and early appraisal of Chronic Obstructive Pulmonary Diseases (COPD). The principle work process of our framework is as appeared in Figure 1, which incorporates a few stages to be specific data pre-processing, risk pattern mining, classification modelling. The diagnostic records are utilized as input data, which are cleaned by expelling noises through the period of data pre-processing, incorporating characterizing COPD understanding with its diagnostic criteria and arrangement cleaning. The cleaned data are then imported to the consecutive pattern-based classification calculation for successive risk pattern mining and classification model structure. Along these lines, the proposed system empowers the distinguishing proof of risk patterns that are fundamentally connected with COPD and will help doctors to analyse potential COPD patients in early stages.

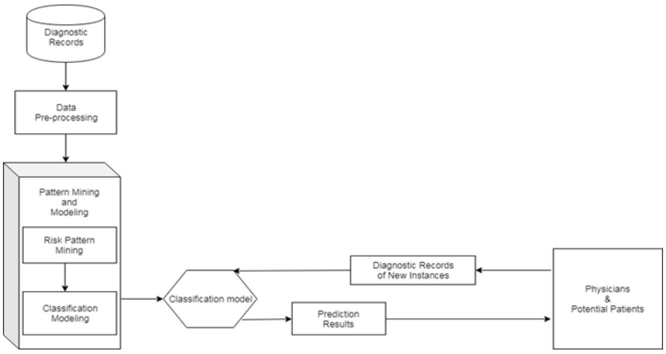


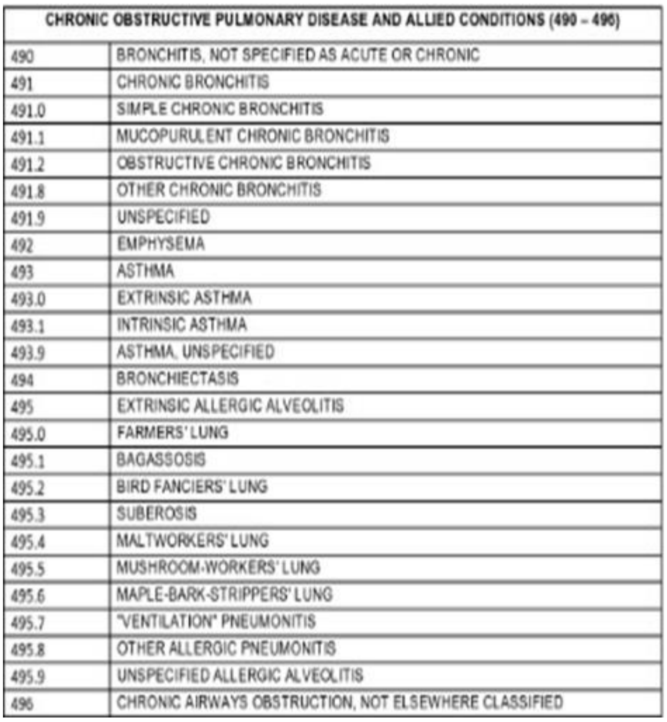
Figure 1: Proposed framework

A. Dataset Collection

The datasets originate from a broad certifiable database named National Health Insurance Research Database (NHIRD) as one of foremost materials in Taiwan. The records in NHIRD was recorded in

International Classification of Diseases (ICD) and we receive the data arrangement of the International Classification of Diseases Ninth Revision, Clinical Modification (ICD-9) in this examination, which is the standard encoding diagnostic instrument for the study of disease transmission, health the board and clinical[2]. NHIRD covers diagnostic records, medical clinic affirmations, prescriptions, disease profiles, and so on., and the diagnostic records are recorded with three ICD-9 codes all things considered in one visit. Figure 2 shows the ICD-9 codes of COPD diseases.

It is essential to have a severe meaning of patients with COPD. Medicinal specialists as given in Figure 3 for the clinical use to characterize confirmed patients, where the blue circle insignias Non-COPD-related ICD-9 codes and the red circle images COPD-related ICD-9 codes, gave COPD-related ICD-9 codes. For our motivations, the meaning of a COPD understanding is one whose diagnostic records incorporate COPD-related ICD-9 codes multiple times in various visits. The region of the data is extracted before the period of definite diagnosis.



CHRONIC OBSTRUCTIVE PULMONARY DISEASE AND ALLIED CONDITIONS (490 – 496)	
490	BRONCHITIS, NOT SPECIFIED AS ACUTE OR CHRONIC
491	CHRONIC BRONCHITIS
491.0	SIMPLE CHRONIC BRONCHITIS
491.1	MUCOPURULENT CHRONIC BRONCHITIS
491.2	OBSTRUCTIVE CHRONIC BRONCHITIS
491.8	OTHER CHRONIC BRONCHITIS
491.9	UNSPECIFIED
492	EMPHYSEMA
493	ASTHMA
493.0	EXTRINSIC ASTHMA
493.1	INTRINSIC ASTHMA
493.9	ASTHMA, UNSPECIFIED
494	BRONCHIECTASIS
495	EXTRINSIC ALLERGIC ALVEOLITIS
495.0	FARMERS' LUNG
495.1	BAGASSOSIS
495.2	BIRD FANCIERS' LUNG
495.3	SUBEROSIS
495.4	MALTWORKERS' LUNG
495.5	MUSHROOM-WORKERS' LUNG
495.6	MAPLE-BARK-STRIPPERS' LUNG
495.7	"VENTILATION" PNEUMONITIS
495.8	OTHER ALLERGIC PNEUMONITIS
495.9	UNSPECIFIED ALLERGIC ALVEOLITIS
496	CHRONIC AIRWAYS OBSTRUCTION, NOT ELSEWHERE CLASSIFIED

Figure 2: ICD-9 Codes of COPD Diseases

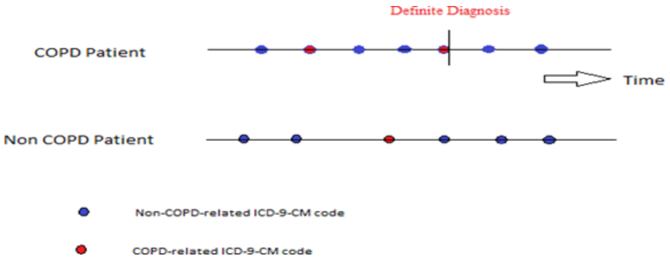


Figure 3: Definition of COPD

B. Data Pre-processing

Data pre-processing is a huge stage to avoid the blunders or inclination in the analytical results. Since there exist some diagnostic (ICD-9) codes that show up comprehensively without novel properties to our target, the investigative codes with repeat higher than 1 percent in the patient's demonstrative records are considered as normally unmistakable ones and filtered through by insinuating and confirmation of restorative authorities. The dataset may include noises such as inconsistent notation. Since each doctor has her/his custom about composition therapeutic records, the explanatory (ICD-9) codes many exist in different sorts of associations in decisive records[2]. In order to ensure that the therapeutic records are correct, we take a gander at the decisive records and clear the symptomatic (ICD-9) codes with coding blunders. Likewise, we extricate the date of patient visits and the diagnostic records are organized consecutively.

C . Sequential Risk Pattern Mining

For the step of sequential risk pattern mining, we integrate the techniques of sequential pattern mining. The SPADE algorithm is employed in our work, which is an efficient algorithm for mining sequential patterns.

SPADE (Sequential PAttern Discovery using Equivalence classes) a new algorithm for discovering the set of all frequent sequences. SPADE not just limits I/O costs by reducing database scans, yet in addition limits computational expenses by utilizing productive inquiry plans. The vertical id-list based methodology is additionally harsh toward data-slant. A broad arrangement of examinations demonstrates that SPADE out-performs previous methodologies by a factor of two, and by a request of greatness on the off chance that we have some extra off-line data. Besides, SPADE scales straightly in the database estimate, and various other database parameters. The main steps include the computation of the frequent 1-sequences and 2- sequences, the decomposition into prefix-based equivalence classes, and the enumeration of all other frequent sequences via minimum support [6].

Computing F1: Given the vertical id-list database, all frequent 1- sequences can be computed in a single database scan. For each database item, we read its id-list from the disk into memory. We then scan the id-list, incrementing the support for each new cid encountered.

Computing F2: Let N = |Z| be the number of frequent items, and A the average id-list size in bytes. A naive implementation for computing the frequent 2-sequences requires G) id-list intersections for all pairs of items. The amount of data read is A- N-(N-1)/2, which corresponds to around N/2 data scans.

Given a database D, which consists of sequences for sequential pattern mining, a frequent sequence will be found if it occurs more than minimum support according to user-specified thresholds minimum support. In addition, a frequent sequence is maximal if it is not a subsequence of any other frequent sequence. The support count of a sequence α is denoted as Sup(α), which indicates the number of α occurring in D. Once we obtain the frequent sequences, we can generate them to produce rules. After generating rules, a rule R: (X⇒ Y) may be produced and accompanied Sup(X) and Conf(R), where Conf(R) is the confidence of the sequential rule and is denote as shown in formula:

$$Conf(X \to Y) = \frac{Sup(X \to Y)}{Sup(X)}$$

D . Classification Modelling

After the period of sequential rule generation, a rich arrangement of sequential rules could be separated and afterward CBS (Classify By Sequence) calculation is utilized to assemble the classification model by utilizing the numerous rules mined. The CBScalculation comprises two stages: the main stage for mining classifiable sequences, and the second stage for utilizing the sequences discovered in stage one to fabricate a classifier. The idea of this technique is to separate the time sequence highlights of each class bunch in creating the classification rules. Along these lines, the classifier worked by utilizing the extricated sequences can be progressively precise[5].

Phase I of CBS: CSP-Miner algorithm: This phase, we discover all classifiable sequential patterns (CSP) for temporal data classification. First, we discover all frequent items of the whole dataset as length 1 frequent sequences. We built a root containing these items as leaves. Then, with this architecture, CBS feature mining can discover all possible sequences.

Phase II of CBS: Classifier Builder algorithm: In this section, we state how to build a classifier using all CSPs discovered by CSP-Miner. Before building a classifier, we need to remove the CSPs having no classification information. After pruning, the remaining CSPs are the features of the class they belong. So, we can performthe classification using these CSPs. Because each CSP has its own class, it shows that the instance containing this CSP has some possibility of being classified as a class of CSP.

When performing classification on a temporal dataset, we classify the new occurrence as the class having the most scores from the CSPs that have a place with it. So as to accelerate the CSP coordinating for new examples, we utilize the prefix tree design once more. The class scores of each CSP are recorded in a tree hub. By following the prefix tree, we can without much of a stretch look for all CSPs contained by another temporal occasion. Meanwhile, we sum the scores of each class on the traced nodes. Finally, we can classify the temporal occurrence as the class with most astounding score.

E . Decision Tree Algorithm

Tree based learning algorithms are considered to be one of the best and mostly used supervised learning methods. Tree based methods empower predictive models with high accuracy, stability and ease of interpretation. Unlike linear models, they map non-linear relationships quite well. They are adaptable at solving any kind of problem at hand (classification or regression). Decision Tree algorithms are referred to as CART (Classification and Regression Trees). The decision tree algorithm tries to solve the problem, by using tree representation[8].

In our sample data there are five important attributes that contribute to the likelihood of getting sick with COPD disease. These attribute are gender, age, genetic causes, Forced expiratory volume(FEV), chest pain, asthma, constant cough, barrel chest, blue finger, chronic bronchitis, Emphysema and chronic airways obstruction. These attributes represent the risk factors that have impact on developing COPD disease. The sample dataset that we use consists of 27 tuples (rows) that contain various values of these attributes depending on patients data. The data sample is already classified and the classlabel is the last column to the right, which is COPD disease (COPD). The data shows that there are two distinct classes, namely, *COPD = yes* for those who have been diagnosed with COPD and *COPD = no* for those who have not.

III.RESULTS AND DISCUSSION

We strengthen our proposed work by defending some of the mined rules from a medical perspective. Our collective medical specialists show that there exist some astounding risk patterns identified with COPD and some scarcely conceivable risk patterns which are disconnected to COPD. For example, smoking is a barely possible risk pattern, since people have smoking habit with the age increasing in their life. Thus, the risk pattern is frequently recorded and it can not be a reasonable risk pattern without other evidences. We can clearly observe from our analysis that for the dataset the number of symptoms and diseases are most important element for the assessment of COPD. SPADE and decision tree algorithms are used for the early assessment of COPD. SPADE algorithm is found to be most suitable algorithm for risk pattern mining and classification. Accuracy of decision tree is low compared to SPADE. If more than two COPD diseases in the span of five year have diagnosed the patient, then it is clear that the patient have high chances of having COPD. In the case of symptoms, more number of symptoms gives the more accurate information about the disease. To determine the best classifier and improve the accuracy of the model, the 10-fold cross validation is used for the training set. We can see that the accuracy of SPADE is roughly 80 above, but for decision tree the accuracy is low.

All paragraphs must be indented. All paragraphs must be justified, i.e. both left justified and right-justified.

A. Analysis of the Proposed Algorithms

Table 1 shows the identification accuracy obtained for SPADE and decision tree with using different number of symptoms and diseases . We can see that SPADE performs better than decision tree and as the number of symptoms and diseases increases, accuracy also increases for SPADE but doesn‘t increase much for decision tree.

Accuracy is calculated with the help of cross validation using confusion matrix. Confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data. There are True Positive(TP), Condition Positive(P), True Negative(TN) and Condition Negative(N). TP is the number of instances correctly predicted as required. P is the real positive number in the data. TN is number of instances correctly predicted as not required and N is the number of real negative instances in the data.

$$\text{Accuracy} = \frac{\text{True positive} + \text{True negative}}{\text{Condition positive} + \text{Condition negative}}$$

Cross validation is one of a model validation technique for assessing how the results perform. It is mainly used when one wants to estimate how accurately a predictive model performs. One round of cross validation involves partitioning a sample of data into subsets performing the analysis on one subset and validating the analysis on other subset. Multiple rounds of cross validation are performed using different partitions and the results are combined.

TABLE I
COMPARISON TABLE OF THE CLASSIFIERS
W.R.TO SYPMTOMS

	Accuracy(%)	
Symptoms	**SPADE**	**Decision Tree**
n = 04	62	29
n = 07	77	38
n = 11	82	51

IV.CONCLUSION

We have proposed a novel framework for early appraisal on incessant infections like COPD by mining successive risk patterns from symptomatic clinical records utilizing consecutive standard mining and classification systems. Our proposed methodology at the same time considers the issues of consecutive risk pattern mining and classification while incorporating them together to be a ground-breaking prediction model with interpretable outcomes. We contrast the SPADE calculation and the choice tree calculation. In outline, in this work, the principle accentuation is set on finding the successive risk patterns, which are well interpretable to foresee the health risk of another obscure case. For future work, more data from different data sources will be considered.

V. REFERENCES

[1] Asha, T., Natarajan, S., and Murthy, K. B. (2011). Associative classification in the prediction of tuberculosis. In *Proceedings of the International Conference & Workshop on Emerging Trends in Technology*, pages 1327–1330. ACM.

[2] Cheng, Y.-T., Lin, Y.-F., Chiang, K.-H., and Tseng, V. S. (2017). Mining sequential risk patterns from large-scale clinical databases for early assessment of chronic diseases: a case study on chronic obstructive pulmonary disease. *IEEE journal of biomedical and health informatics*, 21(2):303–311.

[3] Chin, C. Y., Weng, M. Y., Lin, T. C., Cheng, S. Y., Yang, Y. H. K., and Tseng, V. S. (2015). Mining disease risk patterns from nationwide clinical databases for the assessment of early rheumatoid arthritis risk. *PloS one*, 10(4):e0122508.

[4] Soni, J., Ansari, U., Sharma, D., and Soni, S. (2011). Predictive data mining for med- ical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, 17(8):43–48.

[5] Tseng, V. S. and Lee, C.-H. (2009). Effective temporal data classification by integrating sequential pattern mining and probabilistic induction. *Expert Systems with Applications*, 36(5):9524–9532.

[6] Zaki, M. J. (1998). Efficient enumeration of frequent sequences. In *Proceedings of the seventh international conference on Information and knowledge management*, pages 68–75. ACM.

[7] Han, Jiawei, Micheline Kamber, and Data Mining. "Concepts and techniques." *Morgan Kaufmann* 340 (2001): 94104-3205.

[8] Safavian, S. Rasoul, and David Landgrebe. "A survey of decision tree classifier methodology." *IEEE transactions on systems, man, and cybernetics* 21.3 (1991): 660-674.

Cite this article as :

Saniya PK, Bineesh V., "Mining Sequential Risk Patterns for Early Assessment of COPD", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 3, pp. 496-502, May-June 2019. Available at doi : https://doi.org/10.32628/CSEIT1953165
Journal URL : http://ijsrcseit.com/CSEIT1953165