M.S. THESIS

Discriminative Probabilistic Pattern Mining using Graph for Electronic Health Records

전자의료기록을 위한 그래프 기반 확률적 판별 패턴 마이닝

AUGUST 2019

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Evgenii Li

Discriminative Probabilistic Pattern Mining using Graph for Electronic Health Records

전자의료기록을 위한 그래프 기반 확률적 판별 패턴 마이닝

지도교수 김선

이 논문을 공학석사 학위논문으로 제출함

2019 년 06 월

서울대학교 대학원

컴퓨터 공학부

이예브게니

이예브게니의 공학석사 학위논문을 인준함

2019 년 07 월

위 원 장	박근수	(인)
부위원장	김선	(인)
위 원	전화숙	(인)

Abstract

Electronic Health Records (EHR) contains plenty of useful information about patient's medical history. However, EHR is highly unstructured data and amount of it is growing continuously, that is why there is a need in a reliable data mining technique to group and categorize clinical notes. Although, many existing data mining techniques for group classification use frequent patterns generated based on frequencies of keywords, these patterns do not possess strong enough distinguishing characteristics to show the difference between datasets to classify complex data such as clinical notes in EHR. Also, these techniques encounter scalability and computational cost problems when used on large EHR dataset. To address these issues, we introduce discriminative probabilistic pattern mining algorithm that uses a graph (DPPMG) to generate the subgraphs of frequent patterns for classification in electronic health records.

We use co-occurrence, a combination of binary features, which is more discriminative than individual keywords to construct discriminative probabilistic frequent patterns graph for clinical notes classification. Each co-occurrence has a weight of log-odds score that is associated with its discriminative power. The graph, which reflects the essence of clinical notes is searched to find discriminative probabilistic frequent subgraphs. To discover the discriminative frequent subgraphs, we start from a hub node in the graph and use dynamic programming to find a path. The discriminative probabilistic frequent subgraphs discovered by this approach are later used to classify clinical notes of electronic health records.

Keywords: Discriminative Pattern Mining, Frequent Pattern Mining, Electronic Health Records

Student Number: 2017-29155

Contents

Abstract **i**

Contents **ii**

List of Figures **iii**

Chapter 1 Introduction and Motivation **1**

Chapter 2 Background **4**

2.1 Frequent Pattern Based Classification 4

2.2 Discriminative Pattern Mining 5

2.3 Electronic Health Records 6

Chapter 3 Related Work **8**

Chapter 4 Overview and Design **10**

Chapter 5 Implementation **12**

5.1 Dataset 12

5.2 Keyword Extraction and Filtering 15

5.3 Co-occurrence Generation and Graph Construction 16

5.4 Dynamic Programming to Discover Optimal Path 17

Chapter 6 Results and Evaluation **20**

6.1 Choosing Starting Hub Node 20

6.2 Qualitative Analysis 22

6.3 Discriminative Power of the Probabilistic Frequent Patterns . . . 24

Chapter 7 Conclusion **26**

Bibliography **27**

요약 **33**

Acknowledgements **34**

List of Figures

Figure 2.1 Electronic Health Records System 7

Figure 4.1 Workflow diagram 11

Figure 5.1 Number of pneumonia clinical notes 14

Figure 5.2 Number of sepsis clinical notes 14

Figure 6.1 Choosing the starting hub node 21

Figure 6.2 Percentage of patients who have keywords 23

Figure 6.3 Percentage of correctly classified pneumonia patients . . 24

Figure 6.4 Percentage of correctly classified sepsis patients 25

Chapter 1

Introduction and Motivation

Over the last decade many large hospitals adopted electronic health records systems [12, 25], which provide simple maintenance and easy access to patient's information. Many studies found secondary use for EHR data. In particular, patient data included in EHR contains many features and information about certain diseases that can provide a valuable insight into identifying differences between groups of patients. These differences are of a great importance for better understanding the reasons which lead to group differences of patients with same disease.

Our goal is to find a set of frequent patterns that possess distinguishing characteristics within the graph. We believe that these discriminative probabilistic frequent subgraphs can reflect the sense of the clinical note by their co-occurrences. Traditional frequent pattern mining algorithm, when used to our dataset, generates all possible frequent patterns. A large number of frequent patterns create unnecessary computational costs during mining, and patterns are not discriminative enough for efficient classification. The time and space

required to generate and store these disconnected frequent edges have negative impact on the overall performance. It can take a long time to complete due to exponential growth of combination of items.

Frequent patterns that are solely based on support information and do not include meaningful semantics can produce a large number of irrelevant itemsets. Especially when the support is set to minimum, the produced mining results are unacceptably large, but only few itemsets are of real concern. So, classification based on patterns that do not possess enough distinguishing characteristics is not effective for EHR data.

The other problem is computational cost. Creating all possible frequent patterns often generates a large number of frequent patterns, and the memory runs out. It also requires extensive mining to discover the frequent patterns. This mining becomes extensive based on minimum support. If the number of clinical notes is large, the cost of generating these frequent patterns becomes enormous, even with high minimum support. Processing millions of patterns for a feature selection, which is a common scale for pattern mining algorithms in dense datasets such as EHR, is computationally expensive and time consuming. In this case, the performance of the algorithm degrades drastically. So, it is highly inefficient to wait for a long time for mining algorithm to complete, and then use feature selection on all possible patterns.

The lack of discriminative power and computational cost of frequent pattern mining motivated us to investigate an alternative approach. Instead of generating a set of all possible frequent patterns, we suggest to construct a graph of features, generated from co-occurrences of selected keywords. Each co-occurrence has a weight associated with their log-odds score that is computed from probability of co-occurrence's patterns. This leads to our proposal of discriminative probabilistic frequent pattern mining using graph. It integrates feature selection

mechanism that uses dynamic programming to construct a chain or subgraph of probabilistic discriminative patterns starting from a hub node.

Association rule mining [1] is the data mining method in machine learning for discovering the rules that may govern associations and interesting relations between itemsets. Using co-occurrence patterns for association rule mining to classify EHR, to the best of our knowledge, is a new approach. Most of current classification techniques in data mining depend on the frequency of keywords or the bag-of-words approaches [18]. In these models, a text file is represented in terms of a vector whose elements are the keywords with associated frequencies. This is not sufficient to represent the concept of clinical notes and it gives an ambiguous result in many cases.

All of our contributions can be specified into two parts:

- We propose discriminative probabilistic frequent pattern mining using graph for electronic health records. Our algorithm avoids not only generating a large number of indiscriminative patterns, but also reduces the problem size by constructing a graph of frequent patterns from discriminative probabilistic co-occurrences.

- Instead of mining a set of frequent patterns from a set of all possible patterns, our mining approach discovers subgraphs of discriminative patterns based on their log-odds score, starting from hub nodes in a graph.

The rest of the thesis is organized as follows. Chapter 2, describes the background literature. Chapter 3, describes related work. We give overview of our approach in Chapter 4, and explain the implementation in Chapter 5. The results and evaluation are described in chapter 6. Finally, we conclude in Chapter 7 and suggest the future direction of the research.

Chapter 2

Background

2.1 Frequent Pattern Based Classification

Frequent pattern based classification is data mining classification model that usually includes 3 steps: 1) frequent itemset mining 2) feature selection and 3) model learning. On the first step, itemsets or frequent patterns are generated, later on which we employ a feature selection algorithm to find a set of discriminative patterns. Then those discriminative patterns are represented in the form of training set in the feature space. At last, a classification model is constructed. However, there is a significant computational drawback in this approach, because both frequent pattern mining and feature selection steps could potentially generate exponentially growing combination of items.

Apriori [2] is the most popular frequent pattern based algorithm. Apriori is an iterative algorithm which generates frequent itemsets by scanning the whole dataset in the first iteration. And then frequent itemsets are extended by one item on each iteration. Each step can take a very long time to generate

a set of frequent patterns with millions of itemsets. On dense datasets, the complete mining result is unacceptably large and only small number of them are discriminative enough for classification. Employing feature selection on such a large mining results is inefficient as well.

A goal of feature selection is to select discriminative patterns with distinguishing characteristics. However, any feature selection algorithm applied on such a large dataset could also take a very long time to complete. Even if a linear algorithm is employed, it could still run slowly. In our experiment with a low minimum support, frequent pattern mining algorithm generates over millions of itemsets and feature selection never finishes or crashes.

2.2 Discriminative Pattern Mining

Unlike frequent pattern mining which is about finding itemsets based on the frequency of features only, discriminative pattern mining challenges a task of finding interesting patterns that occur with disproportionate frequency in datasets with various class labels. Discovering distinguishing features and differences between datasets with class labels is a valuable task in data mining, which is used mainly for group difference detection and classifier construction.

Discriminative pattern mining in recent years has drawn much attention among data mining and machine learning researches. A lot of research on discriminative patterns appear under different definitions such as contrasts sets [4], emerging patterns [10] and subgroups [16, 29]. According to [4, 28], contrast set mining aims at discovering patterns that capture prominent differences in frequency among different groups of subjects. Emerging pattern mining detects patterns that capture frequency growth change from one class to another [10, 19]. Discriminative patterns can identify the differences between two or

more datasets, which is a great value for building powerful classifiers and describing different classes. Discovery of such patterns contributes considerably in a wide range of applications, such as the patient risk detection in medicine, finding of overexpressed genes in microarray data analysis, and discovery of distinguishing features in customer relationship management [24].

2.3 Electronic Health Records

Recently electronic health records have been dragging a lot of interest from researchers. Originally it is mainly designed for archiving patient's clinical information and healthcare administration. However, researchers discovered a secondary use of EHR for wide range of clinical tasks for improving healthcare system [6, 13]. Many data mining and machine learning techniques are used in EHR research. Figure 2.1 describes how electronic health records are written by caregivers in different types of clinical notes. These clinical notes are then stored in EHR database.

EHR systems are storing data regarding each patient, such as demographic information, history of diagnoses, radiological images, laboratory tests and results, clinical notes, and many more [5]. In applications of clinical informatics, EHR systems has been used for various tasks, such as medical concept extraction [21, 14], patient trajectory modeling [11], disease inference [32, 3], clinical decision support systems [17], and many more.

EHR usage at hospitals can improve patient care system by minimizing errors, increasing efficiency, and improving care coordination. Depending on functionality, EHR systems can be categorized by EHR without clinical notes, EHR with clinical notes, and comprehensive systems [12].

EHR systems collect and keep the data in several formats:

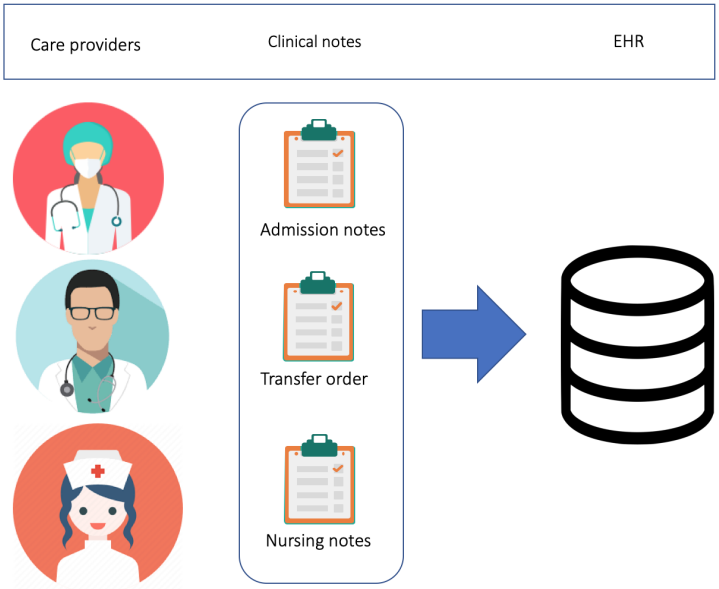

Figure 2.1: Electronic Health Records System

1. Simple numerical information such as patient age and weight
2. Medical codes such as ICD-9 or CPT codes
3. Categorical information such as patient gender and marital status
4. Natural language text such as nursing notes
5. Time series information about vital signs and laboratory tests

Chapter 3

Related Work

Classification based on the frequent patterns has a relation to associative classification. Classifier for associative classification are constructed based on high support association rules [26, 23]. Prediction is made based on combination of support and confidence measures of the rules.

HARMONY [27] is a ruled based classifier which directly mines classification rules. HARMONY uses an instance-centric rule generation approach and makes sure that one of the highest-confidence rules, covering the instance, is included in the rule set for each training instance.

Lazy associative classification [26] is another association rule-based classification method. It is based on non-eager or,according to the authors, lazy classification philosophy, where the classification is made on a demand-driven basis. This approach reduces the number of generated rules by concentrating on the test instance only.

Systematic exploration of frequency based classification, introduced in [7], is a method that selects a highly discriminative frequent itemsets to represent the

data in a feature space and based on this any learning algorithm can be used for model learning. First, a set of frequent itemsets are mined, then a feature selection is performed on the mining results to distinguish a compact set of highly discriminative itemsets. This method is shown to achieve high accuracy.

Direct discriminative pattern mining or DDPMine [8] is the mining approach that uses branch-and-bound search for directly mining discriminative patterns without generating the complete pattern set. It generates discriminative patterns sequentially on a progressively shrinking FP-tree by incrementally eliminating training instances. The instance elimination helps to reduce the problem size iteratively and speed up the mining process.

Chapter 4

Overview and Design

In this chapter, we present a brief overview of our algorithm components and design choices. Figure 4.1 denotes the step by step workflow. The steps are as follows:

- Medical notes extraction
- Word filtering
- Co-occurrence graph construction
- Discriminative log-odds weight assignment
- Finding starting nodes
- Finding discriminative probabilistic paths by dynamic programming

First, we extract electronic health records for a particular disease, in our case pneumonia. Then we divide the data into two classes and select candidate words, as not all the words in the dataset are discriminative. On the feature

generation step we construct co-occurrences using the candidate words. Each co-occurrence is a binary of single features. To compute the probabilities for co-occurrences in each class, we count their associative frequencies. We use co-occurrences to construct a graph, because they have more distinguishing characteristics than individual features. We give a weight to each co-occurrence based on their log-odds score computed from probabilities.

On the feature selection step, we choose a starting node and then construct a subgraph of discriminative probabilistic patterns. We discover new edge by using dynamic programming to choose the subgraph with edges of the highest sum of log-odds score. At last, we search for discriminative probabilistic frequent patterns in the graph, generated from those co-occurrences.

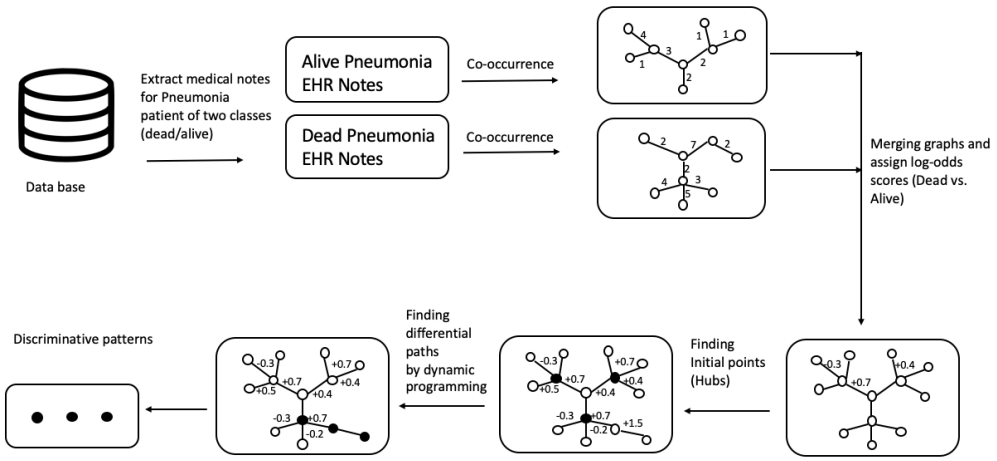

Figure 4.1: Workflow diagram

Chapter 5

Implementation

5.1 Dataset

In this chapter, we explain our dataset. MIMIC-III (Medical Information Mart for Intensive Care) [15] is a large database that contains information about patients admitted to critical care units at a hospital. Data includes vital signs, medications, laboratory measurements, observations and notes charted by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data, and more. MIMIC-III comprises of de-identified, clinical data of patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts.

MIMIC-III contains data of 53,423 distinct hospital admissions for adult patients. The age of patients is 16 years or above. Patients are admitted to critical care units between 2001 and 2012. The median age of adult patients is 65.8 years, 55.9% patients are male, and in-hospital mortality is 11.5%. The median length of an ICU stay is 2.1 days and the median length of a hospital stay is 6.9 days.

We use data of MIMIC-III that only contains pneumonia and sepsis patient information. The pneumonia data contains 1,419 patients, among which 1,166 patients from 'Alive' class label and 253 patients from 'Dead' class label. There are 26,152 clinical notes from 'Alive' class label and 10,706 clinical notes from 'Dead' class label. We divided the data into training set and test set for our experiments: The training set contains 27,712 clinical notes and the test set 9,146 clinical notes. All notes in both sets are randomly selected. The sepsis data contains 1,100 patients, among which 841 patients from 'Alive' class label and 259 patients from 'Dead' class label. There are 24,142 clinical notes from 'Alive' class label and 9,488 clinical notes from 'Dead' class label. We divided the data on training set and test set for our experiments: The training set contains 25,223 clinical notes and the test set 8,407 clinical notes. All notes in both sets are randomly selected.

Figures 5.1 and 5.2 describe the types of clinical notes and their associated number in pneumonia and sepsis datasets.

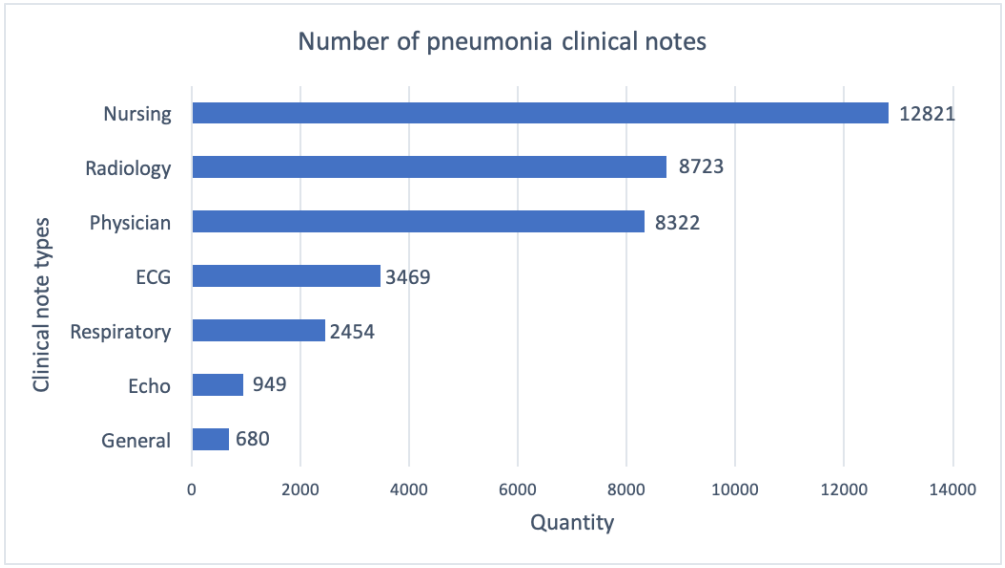

Figure 5.1: Number of pneumonia clinical notes

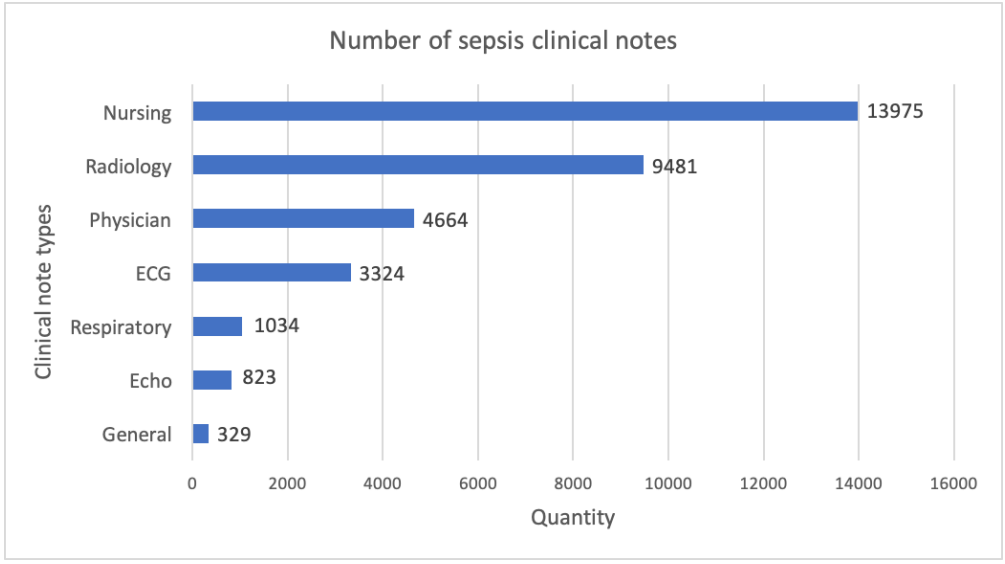

Figure 5.2: Number of sepsis clinical notes

5.2 Keyword Extraction and Filtering

Clinical text notes contain a lot of unstructured data. Some of data is written by hand and then converted to a digital format, so it often contains many grammar mistakes. Also, clinical notes in general contain many abbreviations, which require readers to know the expanded form to understand the semantics. Each abbreviation can have tens of possible explanations, which makes analysis of the text a challenging task. In our algorithm, we treat each abbreviation as a single pattern for the subgraph of itemsets, because we are not interested in the meaning of each individual word, but in co-occurrence semantic relation of binary patterns and their associated log-odds score.

All the patients' data in our database comes from Intensive Care Unit (ICU). This means that each patient is in a very critical health condition and has a high probability of dying. Before extracting key words, we divide patients into two groups and give them labels. The first group has a label of 'Alive' for the patients who survived from the disease after receiving treatment in the ICU. The second group has a label of 'Dead' for the patients who died after receiving treatment in the ICU. To extract the words from clinical notes, we tokenize all the words from each note of each patient's timeline in our database. Then we create a set of key words for each patient. We repeat the process for each group and count frequency of each unique word in the sets of words.

To filter the words and leave only meaningful ones, two parameters are used to define if a word is meaningful for our algorithm, which are:

- n_fold: minimum ratio of the posterior probability of a word in one group to the posterior probability of a word occurring in another group

- threshold: minimum frequency of a word

As a result of this procedure we generate a list of meaningful discriminative probabilistic keywords for each group of patients.

5.3 Co-occurrence Generation and Graph Construction

The co-occurrence pattern is a binary combination of features. In our case, each feature is a keyword from a list of filtered words. The co-occurrence patterns have more discriminative power than individual features, thus they have been extensively used in classification tasks, such as feature co-occurrence [22], multi-local features [9], compositional feature [30, 31], and high order feature [20]. We generate co-occurrences for each note of each patient in the database. Co-occurrences are generated only for the keywords from the lists that were selected on the previous stage. For each co-occurrence, we assign a log-odds score.

Assignment of log-odds score plays a central role for classification performance in our algorithm. The co-occurrence with positive log-odds score implies that binary combination of this features can be served as a classification rule for identifying notes for patients whose health condition is improving with a medical treatment. On the other hand, co-occurrence with negative log-odds score can be served as a classification rule for patients whose health condition is deteriorating and they have a higher probability for lethal outcome.

First step to compute the log-odds score is to compute the probability of co-occurrences. Then we find odds ratio and take the logarithm. The odds ratio for co-occurrences that appears only in one group of patients and does not in other is 1:0. In this case, logarithm expression diverges to infinity, thus we are adding α.

For the computation of co-occurrence probabilities, we use their associated

frequencies. Frequency of co-occurrence for a group of patients labeled as 'Alive', we define as $f(A)$, and we define $f(D)$, for a group of patients labeled as 'Dead', and n is a pseudo count. Probability of co-occurrence for two group of patients we denote as $P(A)$ and $P(D)$. Log-odds score for each co-occurrence is defined as

$$\log \frac{P(A) + \alpha}{P(D) + \alpha}$$

Therefore, where:

$$P(A) = \frac{f(A) + n}{f(A) + f(D) + 2n}$$

$$P(D) = \frac{f(D) + n}{f(A) + f(D) + 2n}$$

$$\alpha = \frac{n}{f(A) + f(D) + n}$$

This gives the final formula:

$$\log \frac{P(A) + \alpha}{P(D) + \alpha} = \log \frac{\frac{f(A)}{f(A)+f(D)+2n} + \frac{n}{f(A)+f(D)+2n}}{\frac{f(D)}{f(A)+f(D)+2n} + \frac{n}{f(A)+f(D)+2n}}$$

We construct a graph of co-occurrences by merging them with one another, where each node is one of the words and edge is a co-occurrence relation between two words. The edge weight in the graph is assigned according to the associated co-occurrence log-odds score. Each node in a graph can have both negative and positive edge's weights.

5.4 Dynamic Programming to Discover Optimal Path

In our algorithm, the optimal path problem is the problem of finding a path between two nodes in a graph such that the sum of the weights of edges is maximized or minimized. The path is the subgraph of discriminative probabilistic frequent patterns, where a node corresponds to a word from co-occurrence and each edge is a co-occurrence weighted by the log-odds score.

We want to discover the path that has a maximum sum of log-odds scores, if we mine discriminative probabilistic patterns for classification of patients whose health condition is improving. On the contrary, to find discriminative probabilistic patterns for patients whose health condition is deteriorating, we want the sum of log-odds scores to be minimized.

The usual choice to find the path in a graph would be to use greedy algorithm, such as Dijkstra algorithm. The algorithm makes the optimal choice at each step as it attempts to find the overall optimal way to solve the entire problem.

Greedy algorithm seeks to find the path with the largest sum by selecting the largest available weighted node at each step. In other words, for greedy algorithm approach the optimal solution can be reached by choosing optimal choice at each step. So, greedy algorithms work on problems for which it is true that, at every step there is a choice that is optimal for the problem on that step, and after the last step, the algorithm generates the optimal solution of the complete problem. However, it fails to find the globally optimal solution because they do not consider all the data. The choice made by a greedy algorithm may depend on choices it has made so far, but it is not aware of future choices it could make.

To find the globally optimal solution we use dynamic programming. The principle of dynamic programming is using 'memoization' or, in other words, simply store the results of sub-problems, so that we do not have to re-compute them when needed later. This optimization reduces time complexity from exponential to polynomial. By using dynamic programming, we can always find the path with optimal sum of log-odds scores in the graph every time we add a new edge.

For the starting node, we choose a hub node with the number of edges that

greatly exceeds the average. We count all hubs in the graph and list them by giving each hub a rank according to the maximum value of the sum of log-odds scores. More details about starting hub node is described in the next chapter.

Chapter 6

Results and Evaluation

In this chapter, we are explaining experimental results of our discriminative probabilistic frequent pattern mining using graph approach.

6.1 Choosing Starting Hub Node

To make our frequent patterns have high discriminative power, we selected keywords with discriminative power from clinical notes by including pre-processing stage where we filter insignificant keywords. This way we discard keywords that are most frequent in the dataset and have very abstract concept, such as "Patient"," Cash", "Wallet", etc. We also discard keywords that are least frequent and represents very specific concept, such as "Death", "Grave" and etc. Concepts that are very abstract appear in many clinical notes while specific concepts tend to appear in a very small number of clinical notes. Therefore, very-abstract and very-specific concepts are not good candidates for generation of discriminative patterns for classification. We only consider keywords that are

filtered with *n_fold* and *threshold* parameters to make the frequent patterns more meaningful. The reduced number of keywords used to generate the discriminative patterns also have a positive impact on the computation cost as less combination of itemsets needs to be created.

In the discriminative probabilistic frequent patterns generation part, each discovered path in a graph is checked to make sure it follows the dynamic programming optimality constraints. For the starting node, we choose hub nodes with the highest sum of log-odds score. Figure 6.1 shows keywords with the highest sum of log-odds score of connected edges for each class label.

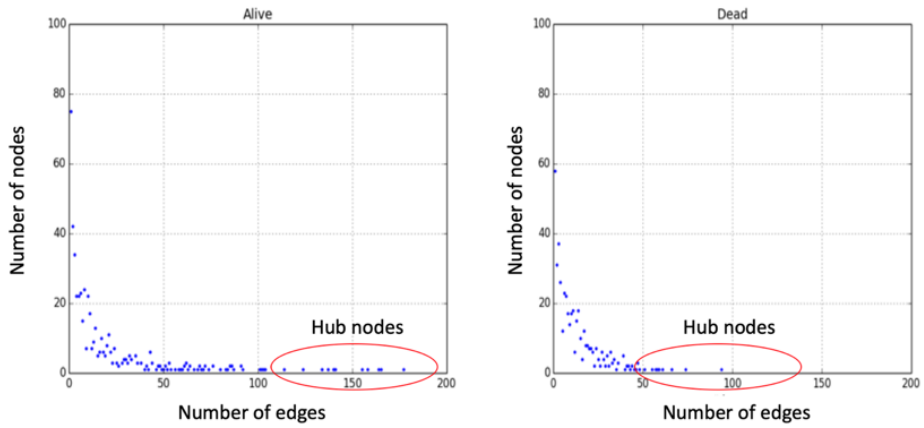

Figure 6.1: Choosing the starting hub node

6.2 Qualitative Analysis

Our discriminative probabilistic frequent patterns mining using graph approach shows that the distinguishing characteristics of itemsets from related patterns are higher, than itemset of not related patterns. For example, one of the discovered paths from graph of discriminative probabilistic frequent patterns with a class label 'Alive' is: [OOB, cooperative, arch, sinus]. Figure 5 represents the percentage of patients who have the keywords in their electronic health records. In this itemset we have one abbreviation, which is OOB and stands for "out of bed", charting notation indicating that a patient has become ambulatory. The word "arch" in the context of medical notes stands for aortic arch, the portion of main artery. And the word "sinus" in the context of medical notes stands for sinus rhythm. The semantics and frequency of each single keyword in the itemset does not strongly imply improvement or deterioration of health condition for a patient. Thus, it is not effective to process the classification only by a single pattern.

In our approach the keywords in co-occurrences are related to each other. And the larger log-odds score, the stronger the relation. Thus, the combination of such discriminative patterns can reflect the semantics of classified data. The word "cooperative' itself can appear in different context and combinations in clinical notes, but when it meets in combination with the word "OOB", it has a positive meaning. Example sentence from nursing type of clinical note: "Patient was OOB for 5 hrs, he tolerated this well. This morning, he was cooperative ...". This co-occurrence relation gives an implication that patients condition is improving.

The combination of "arch" and "sinus" appears in a patient's laboratory test clinical notes. For example: "Normal diameter of aorta at the sinus". These

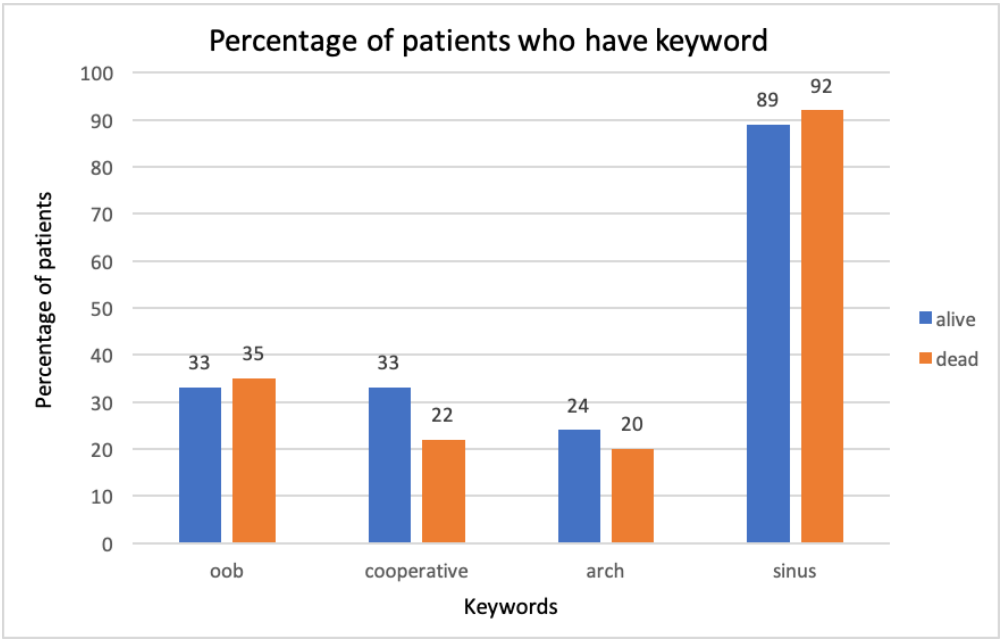

Figure 6.2: Percentage of patients who have keywords

patterns can be interpreted as the test for aortic arch was processed for the patient and sinus rhythm implied for normal electrical activity to flow within the heart.

The whole itemset [OOB, cooperative, arch, sinus] can be interpreted as "the patient condition allows him to walk out of bed by himself and he is being cooperative with medical care. Moreover, the patient's sinus rhythm and aortic test implies on normalized heart rate".

Similar interpretation would not be possible with the itemset generated based on the frequency of the keywords. As the frequent patterns in such itemset don't have semantic relation among each other.

The complex data with many repeated words, abbreviations, and syntactic mistakes made by the writer of clinical notes in EHR cannot be efficiently classified based only on frequency of keywords. However, understanding of key-

words semantics helps to distinguish the group difference in the data better and improve the classification performance.

6.3 Discriminative Power of the Probabilistic Frequent Patterns

To show efficiency of discriminative probabilistic frequent patterns, we compare it with two other methods, Apriori and DDPMine. The Apriori data mining algorithm generates frequent patterns based on the frequency of keywords. DDPMine generates discriminative patterns sequentially on a progressively shrinking FP-tree by incrementally eliminating training instances.

For the fairness of the experiment, all algorithms use the same set of preprocessed keywords for generating itemsets of frequent patterns.

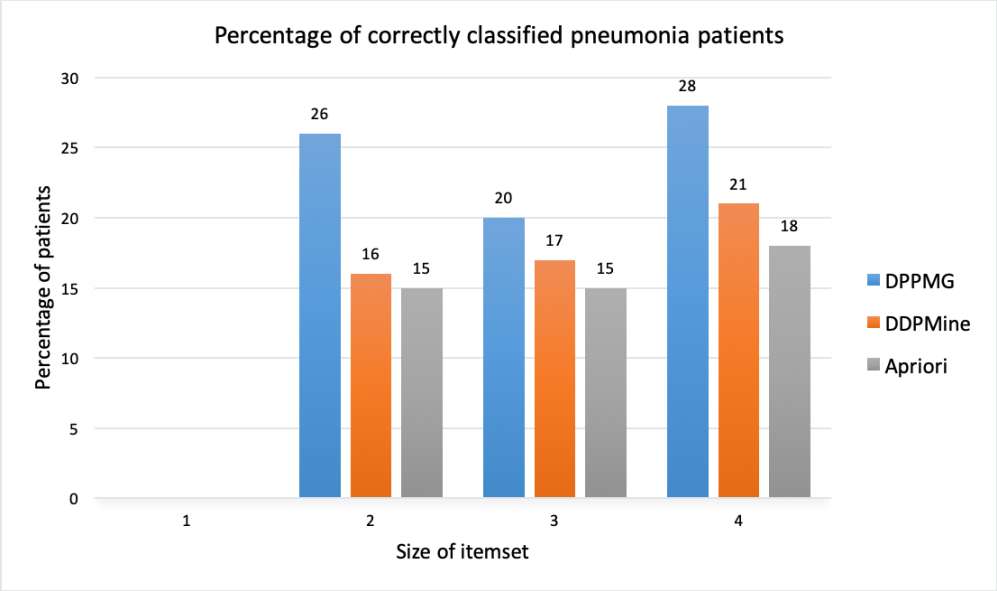

Figure 6.3: Percentage of correctly classified pneumonia patients

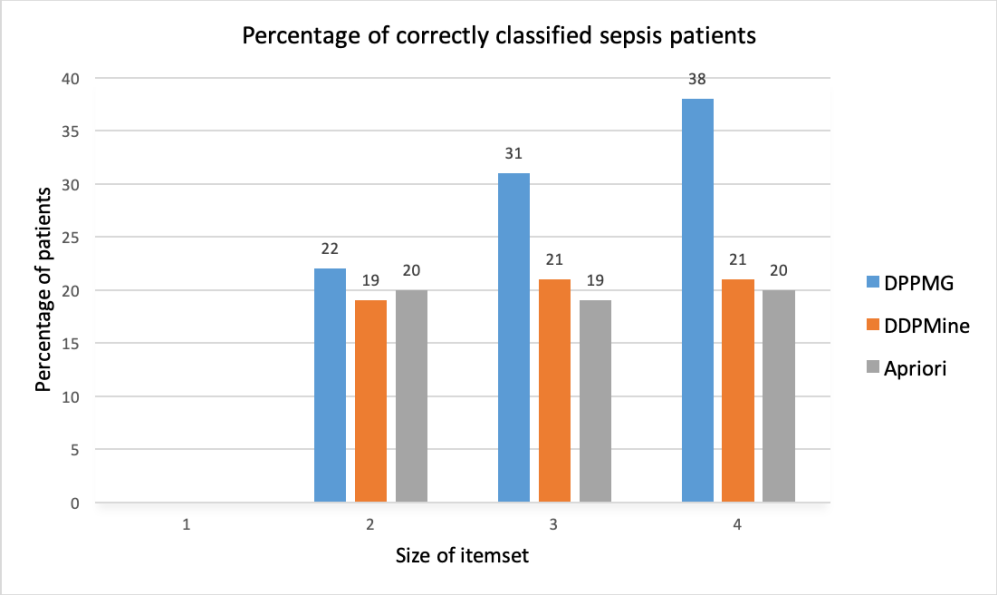

Figure 6.4: Percentage of correctly classified sepsis patients

Figures 6.3 and 6.4 represent percentage of correctly classified patients for Apriori, DDPMine and DPPMG for different sizes of discovered itemsets.

Our approach is more effective in classification of clinical notes from EHR, than Apriori and DDPMine algorithms and it's discriminative power grows with the size of itemset. We can conclude that the more related patterns included in the itemset, the more distinguishing DPPMG becomes.

Chapter 7

Conclusion

Discriminative pattern mining is of a great value among data mining techniques used for classification. Many ideas have been explored to find effective algorithms to employ discriminative patterns for classification. Even so, producing more intuitive classifying algorithm for group difference detection remains a challenging task. In our work, we developed discriminative probabilistic frequent pattern mining algorithm by employing dynamic programming with a graph mining technique. Traditional frequent pattern mining techniques mostly depend on keywords frequency only, whereas our technique relies on co-occurrences and their associated log-odds score, which has more discriminative power than individual keywords. We use dynamic programming to discover the frequent subgraphs in the clinical notes graph. Experimental results show that our algorithm can be successfully used to find discriminative patterns that occur with disproportionate frequency in datasets with various class labels. The classification based on pattern's sum of log-odds scores performs better than the traditional frequency based approach.

We started our research by using an association rule mining algorithm for finding discriminative subgraphs. The traditional frequent mining algorithm can discover the set of all frequent patterns. However, it does not fit to a non-traditional domain like graphs. When it is directly used in mining, it mostly produces set of disconnected itemsets. We have modified our approach to ensure that it follows the connectivity constraint for all frequent patterns discovered. As a result, our discovered discriminative patterns are always connected subgraphs, which discriminative power is associated with the sum of log-odds scores of its edges. We focused on the classification of the electronic health records based on the senses of discovered discriminative probabilistic frequent patterns. We believe that the subgraphs discovered by our approach reflect the concept of electronic health records better than frequent patterns generated by keywords frequencies. Experimental results support the efficiency of our approach in classification of electronic health records.

For the future work, our approach can be extended in number of ways. We can generate some optimization techniques for computing the combinations of larger single paths which can improve our algorithm performance. Additionally, as many electronic health records are kept in time-series data format we plan to improve our approach into intelligent system that can extract useful discriminative frequent patterns that include the notion of time from time-series data.

Bibliography

[1] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM, 1993.

[2] R. Agrawal, R. Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[3] P. C. Austin, J. V. Tu, J. E. Ho, D. Levy, and D. S. Lee. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *Journal of clinical epidemiology*, 66(4):398–407, 2013.

[4] S. D. Bay and M. J. Pazzani. Detecting group differences: Mining contrast sets. *Data mining and knowledge discovery*, 5(3):213–246, 2001.

[5] G. S. Birkhead, M. Klompas, and N. R. Shah. Uses of electronic health records for public health surveillance to advance public health. *Annual review of public health*, 36:345–359, 2015.

[6] T. Botsis, G. Hartvigsen, F. Chen, and C. Weng. Secondary use of ehr: data quality issues and informatics opportunities. *Summit on Translational Bioinformatics*, 2010:1, 2010.

[7] H. Cheng, X. Yan, J. Han, and C.-W. Hsu. Discriminative frequent pattern analysis for effective classification. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 716–725. IEEE, 2007.

[8] H. Cheng, X. Yan, J. Han, and S. Y. Philip. Direct discriminative pattern mining for effective classification. In *2008 IEEE 24th International Conference on Data Engineering*, pages 169–178. IEEE, 2008.

[9] O. Danielsson, S. Carlsson, and J. Sullivan. Automatic learning and extraction of multi-local features. In *2009 IEEE 12th International Conference on Computer Vision*, pages 917–924. IEEE, 2009.

[10] G. Dong and J. Li. Efficient mining of emerging patterns: Discovering trends and differences. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 43–52. Citeseer, 1999.

[11] S. Ebadollahi, J. Sun, D. Gotz, J. Hu, D. Sow, and C. Neti. Predicting patient's trajectory of physiological data using temporal trends in similar patients: a system for near-term prognostics. In *AMIA annual symposium proceedings*, volume 2010, page 192. American Medical Informatics Association, 2010.

[12] J. Henry, Y. Pylypchuk, T. Searcy, and V. Patel. Adoption of electronic health record systems among us non-federal acute care hospitals: 2008-2015. *ONC data brief*, 35:1–9, 2016.

[13] P. B. Jensen, L. J. Jensen, and S. Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395, 2012.

[14] M. Jiang, Y. Chen, M. Liu, S. T. Rosenbloom, S. Mani, J. C. Denny, and H. Xu. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association*, 18(5):601–606, 2011.

[15] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[16] W. Klösgen. Explora: A multipattern and multistrategy discovery assistant. In *Advances in knowledge discovery and data mining*, pages 249–271. American Association for Artificial Intelligence, 1996.

[17] G. J. Kuperman, A. Bobb, T. H. Payne, A. J. Avery, T. K. Gandhi, G. Burns, D. C. Classen, and D. W. Bates. Medication-related clinical decision support in computerized provider order entry systems: a review. *Journal of the American Medical Informatics Association*, 14(1):29–40, 2007.

[18] D. D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning*, pages 4–15. Springer, 1998.

[19] J. Li, K. Ramamohanarao, and G. Dong. The space of jumping emerging patterns and its incremental maintenance algorithms. In *ICML*, pages 551–558, 2000.

[20] D. Liu, G. Hua, P. Viola, and T. Chen. Integrated feature selection and higher-order spatial feature extraction for object categorization. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[21] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*, 17(01):128–144, 2008.

[22] T. Mita, T. Kaneko, B. Stenger, and O. Hori. Discriminative feature co-occurrence selection for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(7):1257–1269, 2008.

[23] T. M. Mitchell and M. Learning. Mcgraw-hill science. *Engineering/Math*, 1:27, 1997.

[24] P. K. Novak, N. Lavrač, and G. I. Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, 10(Feb):377–403, 2009.

[25] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2017.

[26] A. Veloso, W. Meira Jr, and M. J. Zaki. Lazy associative classification. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 645–654. IEEE, 2006.

[27] J. Wang and G. Karypis. Harmony: Efficiently mining the best rules for classification. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, pages 205–216. SIAM, 2005.

[28] G. I. Webb, S. Butler, and D. Newlands. On detecting differences between groups. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 256–265. ACM, 2003.

[29] S. Wrobel. An algorithm for multi-relational discovery of subgroups. In *European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 78–87. Springer, 1997.

[30] J. Yuan, J. Luo, and Y. Wu. Mining compositional features for boosting. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[31] J. Yuan, J. Luo, and Y. Wu. Mining compositional features from gps and visual cues for event recognition in photo collections. *IEEE Transactions on Multimedia*, 12(7):705–716, 2010.

[32] D. Zhao and C. Weng. Combining pubmed knowledge and ehr data to develop a weighted bayesian network for pancreatic cancer prediction. *Journal of biomedical informatics*, 44(5):859–868, 2011.

요약

전자의료기록(Electronic Health Records)의 임상 노트에는 환자의 병력에 대한 유용한 정보가 많이 포함되어 있다. 그러나 임상 노트는 체계화되지 않은 데이터이며 그 양은 나날이 증가하고 있다. 따라서 임상 노트를 그룹화하고 분류하기 위한 신뢰할 수 있는 데이터 마이닝 기술이 필요하다. 기존의 데이터 마이닝 기술은 키워드의 빈도를 기반으로 생성된 빈발 패턴(frequent patterns)을 이용하여 그룹 분류 작업(classification)을 수행한다. 하지만 이러한 빈발 패턴은 전자의료기록의 임상 노트와 같이 복잡한 데이터의 분류를 위해 필요한 충분히 강력하고 명확하게 구별되는 특징을 갖고 있지 않다. 또한 빈발 패턴 기반 기술은 대규모 전자의료기록 데이터에 적용될 때 확장성과 계산 비용의 문제에 직면한다. 따라서 본 연구에서는 이러한 문제점을 해결하기 위해 확률적 판별 패턴 마이닝(discriminative probabilistic pattern mining) 알고리즘을 소개한다. 확률적 판별 패턴 마이닝 알고리즘에서는 전자의료기록의 임상 노트를 분류하기 위해 그래프 구조를 도입하여 빈발 패턴의 부분 그래프를 생성하게 된다. 본 연구에서는 판별력을 높이기 위해 개별 키워드를 사용하는 대신 이진 특성 조합에서의 동시 출현(co-occurrence)을 사용하여 임상 노트 분류를 위한 빈발 패턴 그래프를 구성한다. 각각의 동시 출현은 판별력(discriminative power)에 따른 log-odds 값으로 그 가중치를 갖는다. 임상 노트의 본질을 반영하는 그래프를 찾기 위해 확률적 판별 부분 그래프 검색을 수행하며 그래프의 허브(hub) 노드에서 시작하여 동적 프로그래밍(dynamic programming)을 사용하여 경로를 찾는다. 이러한 방법으로 검색한 빈발 부분 그래프를 이용하여 전자의료기록의 임상 노트에 대한 분류 작업을 수행하게 된다.

주요어: Discriminative Pattern Mining, Frequent Pattern Mining, Electronic Health Records

학번: 2017-29155

Acknowledgements

First of all, I would like to thank my professor Sun Kim for his guidance and support in performing this research. I am thankful for giving me the opportunity to be a part of Bio & Health Informatics Lab. During my stay in the lab professor Sun Kim was always giving me the right directions and advice whenever I needed it.

Second, I would like to thank Hongryul Ahn and Sungjoon Park for sharing with me all their knowledge and experience when we worked together on projects in the lab. I am also grateful for their valuable comments on this thesis.

I would also like to thank Inyoung Kim and Chai-Jin Lee for their help and support.

I would also like to thank all the other lab members for making my time at the lab enjoyable.

Finally, I must express my very profound gratitude to my family and my girlfriend for providing me with continuous support and encouragement during my time in Korea and at the lab. This accomplishment would not have been possible without them.