



Wayne State University


Wayne State University Theses

1-1-2015

Predictive Analytics For Disease Condition Of Patients In Emergency Department

Azade Tabaie
Wayne State University,

Follow this and additional works at: https://digitalcommons.wayne.edu/oa_theses

 Part of the [Computer Sciences Commons](#), [Industrial Engineering Commons](#), and the [Medicine and Health Sciences Commons](#)

Recommended Citation

Tabaie, Azade, "Predictive Analytics For Disease Condition Of Patients In Emergency Department" (2015). *Wayne State University Theses*. 478.

https://digitalcommons.wayne.edu/oa_theses/478

This Open Access Thesis is brought to you for free and open access by DigitalCommons@WayneState. It has been accepted for inclusion in Wayne State University Theses by an authorized administrator of DigitalCommons@WayneState.

**PREDICTIVE ANALYTICS FOR DISEASE CONDITION OF PATIENTS IN
EMERGENCY DEPARTMENT**

by

AZADE TABAIE

THESIS

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

2015

MAJOR: INDUSTRIAL ENGINEERING

Approved By:

Advisor

Date

© COPYRIGHT BY

AZADE TABAIE

2015

All Rights Reserved

DEDICATION

TO MY BELOVED HUSBAND,

MY GREAT PARENTS,

MY WONDERFUL CLOSE FRIENDS

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Professor Chinnam for the continuous support of my Master study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Master study.

Besides my advisor, I would like to thank the rest of my thesis committee: Dr. Dalkiran, Dr. Murat, and Dr. Dong, for their insightful comments and encouragement which helped me to widen my research from various perspectives.

Last but not the least; I would like to thank my family: my parents and specially my husband for supporting me spiritually throughout writing this thesis and my life in general.

Table of Contents

| | |
|--|------------|
| ACKNOWLEDGEMENTS | III |
| LIST OF TABLES | V |
| CHAPTER 1: INTRODUCTION | 1 |
| CHAPTER 2: RELATED WORKS | 3 |
| CHAPTER 3: FEATURE SELECTION BASED APPROACH | 5 |
| 3.1. RELATED WORKS IN FEATURE SELECTION | 5 |
| 3.2. MODEL | 7 |
| 3.2.1. GINI-INDEX (GI) | 9 |
| 3.2.2. CHI-SQUARE (CHI) | 10 |
| CHAPTER 4: EXPERIMENT SETTINGS | 12 |
| 4.1. DATA COLLECTION | 12 |
| 4.2. CLASSIFIER | 13 |
| 4.2.1. SUPPORT VECTOR MACHINES | 13 |
| 4.2.2. L_1 REGULARIZED LOGISTIC REGRESSION MODEL | 14 |
| 4.2.3. RANDOM FOREST | 14 |
| CHAPTER 5: RESULTS AND DISCUSSION | 15 |
| CHAPTER 6: FUTURE RESEARCH | 18 |
| REFERENCES | 19 |
| ABSTRACT | 22 |
| AUTOBIOGRAPHICAL STATEMENT | 24 |

LIST OF TABLES

| | |
|---|-----------|
| <u>TABLE 1: THE NINETEEN MOST GENERAL ICD9 CATEGORIES</u> | <u>8</u> |
| <u>TABLE 2: MEMBERSHIPS AND NONMEMBER SHIPS</u> | <u>10</u> |
| <u>TABLE 3: CATEGORICAL VITAL SIGNS</u> | <u>13</u> |
| <u>TABLE 4: PREDICTION ACCURACY BASED ON DIFFERENT CHI-SQUARE ALPHA VALUES</u> | <u>15</u> |
| <u>TABLE 5: CLASSIFIERS' RESULTS</u> | <u>16</u> |
| <u>TABLE 6: TOP THREE SVM OUTPUTS FOR EACH PATIENT RECORD BASED ON SVM RANKED PROBABILITY FEATURE</u> | <u>16</u> |
| <u>TABLE 7: CHI-SQUARE FEATURE SELECTION AFTER ASSIGNING PROBABILITY THRESHOLD</u> | <u>17</u> |
| <u>TABLE 8: ADOPTED GINI-INDEX FEATURE SELECTION AFTER ASSIGNING PROBABILITY THRESHOLD</u> | <u>17</u> |

CHAPTER 1: INTRODUCTION

Emergency departments provide extraordinary important services to public health of the society. Physicians, nurses and staff in ED provide 24 hours of emergency care in every day of a year without discriminating patients by economic or social status. ED crowding jeopardizes patients' health and impairs their quality of care and satisfaction.

The issue of ED crowding occurs in almost every states in America. The reported crowding in hospitals results in patients in hospital hallways, long waiting times and full occupancy of ED beds. ED crowding has several potential unfavorable effects including patients and staff frustration, lower patient satisfaction and poor health outcomes [1].

During the last two decades, numerous researchers with different backgrounds and perspectives have conducted research on the ways to combat ED crowding. They brought together the fundamental methodologies and techniques from different disciplines such as statistical modeling, artificial intelligence and data mining to facilitate medical decision making which can lead to a better patient flow in ED.

The very initial interaction between clinicians and a patient is recorded on nurse triage notes (TN) which contain details of the reason for patient's visit including specific symptoms and incidents. TN and vital signs measured by triage nurse determine the complexity of the patient's condition. If a minor illness or injury occurred, patient would be treated by nurse practitioners under ED physicians' supervision. This process called fast track system which allows the main ED area to focus on more severe patient condition [2]. The final decision should be made by physicians so patients have to wait to be seen in order to find out whether they need to be admitted in the hospital or be discharged.

The goal of this study is to build a decision support system capable of detecting disease condition and patterns in early stage of ED in order to speed up the decision making process. However, lack of consistency in vocabularies of TN is a major constraint in this study which heavily affects the classifiers' performance.

In this study, we propose a feature selection-based approach to predict patient disease from early TN and measured vital signs. We tested our model on almost 8000 patient records from VA Medical Center (VAMC) in Detroit to show its capability in performing robust classification which can be used by surveillance system.

The rest of this thesis is organized as follows: Section 2 reviews the literature related to our work. Section 3 introduces the feature selection based approach. Section 4 presents the experiment results and the related discussion. Finally, section 5 explains our future work to strengthen this study.

CHAPTER 2: RELATED WORKS

ED crowding has been the center of numerous public and academic studies [3-4]. Researchers with different backgrounds suggested models in wide range of expertise.

Patients with more complex needs are more likely to be admitted in ED, however, patients are also likely to be on board when ED is crowded, mainly because the ability to safely discharge patients is weakened [5]. Moskop et al. [6] vastly investigated concepts, causes and moral consequences of ED crowding. After the investigation, they implemented a “reversed triage system” to safe early inpatients discharge. The system identifies inpatients who there is a small risk of complications in their health condition after being early discharged. This early discharge opens spaces for patients with more severe conditions so that they do not need to wait to be admitted [7-8]. Reversed triage system can have a significant effect on combating restricted inpatient capacity while facing disaster patient conditions every day.

Research points out if the decision on admitting a certain patient can be made during triage time or soon after, it will improve the patient flow in ED and advance the necessary arrangements. Extracting temporal information from clinical narratives is one way to compromise ED crowding. Patient chief complaint (CC), known as the reason of seeking emergency care, is a critical information which can lead to patient prioritization and determination of patient flow in ED. Patient chief complaint codes and nurse triage notes are the two crucial types of Electronic Medical Records (EMRs) in ED. Exploring information from these medical narratives has attracted quite a lot of researchers’ attention.

Patient chief complaint is the first data achieved in ED. Despite its valuable effect on decision making, it does not have a standard terminology which jeopardizes the performance of any decision support system. To compromise this inconsistency, Travers and Haas [9] developed

the construction of concept-oriented nursing terminologies from the actual language used by triage nurses.

Chapman et al. [10] proposed a computer script named CoCo to retrieve information from ED triage chief complaints. CoCo is a naïve Bayesian classifier. It was trained on 10000 patient chief complaint codes which were manually classified with seven syndromic categories. In learning process, CoCo has learned to probabilistically associate certain words with certain syndromic categories.

From a public health point of view, to shorten the time it may take to identify an outbreak or a community-wide health issue, Mahalingam et al. [11] proposed a system to detect gastrointestinal syndrome. The offered system classify an ED record based on chief complaint, triage nurse's note and patient's initial temperature at early phase in ED.

CHAPTER 3: FEATURE SELECTION BASED APPROACH

Our proposed model to predict patient ICD9 codes utilizes information in EMRs including nurse triage note and patient vital signs measured soon after triage stage. This surveillance system enables triage nurse to decide on severity of patient condition and make the arrangements if he/she needs to be admitted before being visited by ED physician. Employing nurse triage note introduces several challenges in building the model. The fast-paced ED environment results in potentially error-prone entries which lead to serious term inconsistencies. The most important concern in dealing with free language text is its unstructured feature. This inconsistency jeopardizes the result of any predictive model accuracy. To combat the potential errors in ED records, we developed a library consists of abbreviations and word substitutions which does the necessary substitution in order to make the words more consistent, i.e. “chest pain” modified to “cp”. This library can partially cover the lack of generalization in nurse triage note.

3.1. RELATED WORKS IN FEATURE SELECTION

Utilizing EMRs and text classification becomes more and more important particularly in dealing with medical domain data. There are two serious challenges in medical text classification; high dimensional feature space and class-imbalanced data set. Not all the features carry significant information; therefore, selecting features can reduce noises and collecting representative features able to discriminate between class labels improves the prediction accuracy. On the other hand, class-imbalanced data highly affect the performance of any classifier. Under this condition, even the most powerful classifiers have poor results.

Filter feature selection methods are the ones that aggressively omit low-informative features to improve the generalization of the classification model and prevent overfitting.

Valuable efforts have been done in feature selection area. Yang and Pedersen [12] proposed Chi-Square feature selection (CHI) which measures the lack of independency between each pair of features and class labels. Chi square value of each feature will be compared to critical value of a chi-square distribution with one degree of freedom to find informative variables. CHI is not capable of presenting the dependency between the variables which is considered as its weakness. Yang and Pedersen [12] also presented Information Gain (IG) method which depends on time-consuming and complex probability and conditional probability computations. The authors proved that CHI and IG of a feature are strongly correlated and they are among best feature selection methods for high-dimensional class-imbalanced data sets in which non-informative features are aggressively removed to achieve smaller feature space and better classification performance.

Features are classified in two categories based on their memberships; positive features which accounts for membership and negative features which accounts for non-membership. To shed more lights on the definition of positive and negative features, considering Table (2), a_{ij} and b_{ij} represent positive features of class j and the other two, c_{ij} and d_{ij} represent negative features of class j .

To include negative features which have undeniable effect on the classification performance, Zheng et al. [13] proposed a framework to optimize the combination of positive and negative features in order to achieve the best classification result. If we select features from a balanced data set, the number of selected positive and negative features will be equal.

Peng [14] suggested mutual information (MI) based feature selection method which measures the dependency between the density of a variable and the density of a class label. The challenge in this method is that the density function of the variables, targets and their joint

density function are required which is normally hard to achieve from data. MI can be computed from the expression (1):

$$I(t, c) = \log P_r(t|c) - \log P_r(t) \quad (1)$$

3.2. MODEL

The primary motivation behind predicting disease condition of patients in Emergency Department (ED) is to shorten the patients' waiting time and improve decision making process on severity of patients' condition. Building a powerful decision support system for surveillance during triage time can highly contribute to the quality of care, health outcome and patient's satisfaction. In this article, we propose a feature selection based model employing nurse triage notes and vital signs that can automatically predict ICD9 code assigned to each patient prior to the visit time.

Each ICD9 code, International Classification of Disease 9th edition, stands for a specific kind of disease. There are almost 14000 ICD9 codes available in medical records but due to the lack of observation we predict 19 most general categories. Table (1) denotes the 19 general ICD9 categories which are the target classes in our study. During our experiment explained in section 4 each ICD9 category is identified by its lowest number in the related range.

| ICD9 Range | Disease |
|-------------------|--|
| 001 – 139 | Infectious And Parasitic Diseases |
| 140 – 239 | Neoplasms |
| 240 – 279 | Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders |
| 280 – 289 | Diseases Of The Blood And Blood-Forming Organs |
| 290 – 319 | Mental Disorders |
| 320 – 389 | Diseases Of The Nervous System And Sense Organs |
| 390 – 459 | Diseases Of The Circulatory System |
| 460 – 519 | Diseases Of The Respiratory System |
| 520 – 579 | Diseases Of The Digestive System |
| 580 – 629 | Diseases Of The Genitourinary System |
| 630 – 679 | Complications Of Pregnancy, Childbirth, And The Puerperium |
| 680 – 709 | Diseases Of The Skin And Subcutaneous Tissue |
| 710 – 739 | Diseases Of The Musculoskeletal System And Connective Tissue |
| 740 – 759 | Congenital Anomalies |
| 760 – 779 | Certain Conditions Originating In The Perinatal Period |
| 780 – 799 | Symptoms, Signs, And Ill-Defined Conditions |
| 800 – 999 | Injury And Poisoning |
| V01 – V91 | Supplementary Classification Of Factors Influencing Health Status And Contact With Health Services |
| E000 – E999 | Supplementary Classification Of External Causes Of Injury And Poisoning |

Table 1. The nineteen most general ICD9 categories

Some of these ICD9 categories have been observed rarely but some of them dominate patients' diseases in ED. Consequently, electronic medical records in ED are highly imbalanced in terms of assigned ICD9 categories. Plot (1) signifies this particular attribute in our data set.

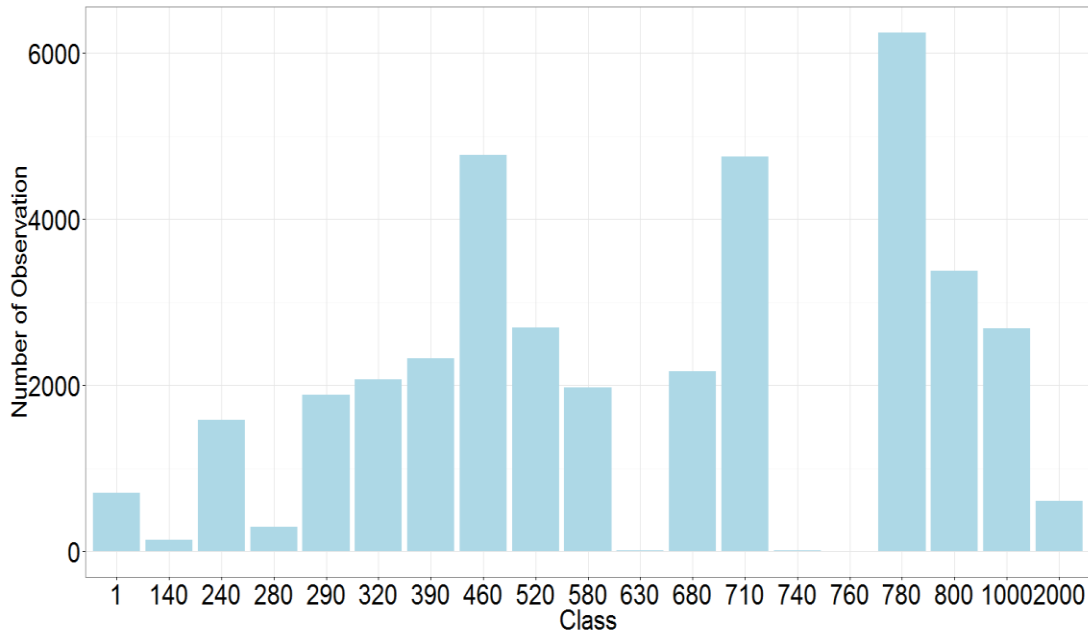


Figure 1: ICD9 categories distribution in ED patient records

Several researchers have evaluated the accuracy of assigned ICD9 codes. MacIntyre found discrepancy between final diagnosis and the codes [8]. Consequently, it is not always possible to capture the most accurate final diagnosis through ICD9 codes but they still comprise valuable information to categorize patients' condition during ED early stage.

Our feature selection based model considers the two most important challenges in developing predictive models from medical domain data sets. The challenges are dealing with high-dimensional feature space and class-imbalanced data set. If the feature selection phase is not appropriately done, any classification algorithms' performance will be impaired. According to the literature, the following two feature selection methods have proved their ability to handle this crucial part of the model building very well.

3.2.1. GINI-INDEX (GI)

The traditional Gini-index computes the impurities of each feature as presented in expression (2).

$$\text{Gini}(S) = \sum_{i=1}^m P_i^2 \quad (2)$$

This method was adopted through the time so that it can be employed in various fields. To apply Gini-Index theory to text feature selection, expression (3) was developed. Shang et al. [15] investigated the weaknesses of Gini-Index represented by expression (3). He proposed a novel form of Gini-index presented in expression (4) which can be used strongly in text feature selection.

$$\text{Gini}(W) = P(W)(1 - \sum_i P(C_i|W)^2) + P(\bar{W})(1 - \sum_i P(C_i|\bar{W})^2) \quad (3)$$

$$\text{GiniText}(t) = \sum_i P(W|C_i)^2 P(C_i|W)^2 \quad (4)$$

3.2.2. CHI-SQUARE (CHI)

CHI method measures the lack of independency between each pair of features and class labels through the expression (5) in which N stands for the total number of observations in data set. A, B, C and D are the simple and abbreviated forms of a_{ij} , b_{ij} , c_{ij} and d_{ij} which can be calculated from Table (2).

$$X^2(t_i, c_j) = \frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \quad (5)$$

| | Containing t_i | Not Containing t_i |
|------------------------|------------------|----------------------|
| Belonging to c_j | a_{ij} | b_{ij} |
| Not Belonging to c_j | c_{ij} | d_{ij} |

Table 2. Memberships and Nonmember ships

To find the corresponding chi-square value of each term, we may compute CHI_{avg} which can be found through the expression (6).

$$X_{\text{avg}}(t) = \sum_j \text{pr}(C_j) X^2(t, C_j) \quad (6)$$

Since chi-square value of each term must be compared with chi-square distribution with one degree of freedom to filter out non-informative features, we should find the optimized alpha value in chi-square distribution which gives us the critical value as the most proper threshold we are looking for. There is no specific rule to find the best alpha and critical value in chi-square feature selection method; therefore, we tried different alpha values and their corresponding thresholds to investigate how the prediction accuracy is changing. The critical value achieved the highest accuracy was selected as the feature selection threshold.

After selecting the most informative features which are able to strongly discriminate between class labels, we employ the most powerful classifiers with the ability of significantly handling multi-class classification. In this study, we applied Support Vector Machines, Random Forest and L_1 Regularized Logistic Regression Model to perform the 19-class prediction process.

CHAPTER 4: EXPERIMENT SETTINGS

To test the feature selection based model, we conducted an experiment to evaluate the predictive model's performance.

4.1. DATA COLLECTION

The data set of this experiment came from Veterans Affairs Emergency Department Integrated Software (EDIS) in Detroit from July 2011 to December 2013. The patient records include nurse triage notes, initial vital signs, and assigned ICD9 codes. Primarily, patient records without proper nurse triage note, vital signs or ICD9 code were filter out from data set. Standard preprocessing procedure including stop words and punctuation elimination has been applied on nurse triage notes which are in unstructured text format. Finally, the library of abbreviation and word substitution was employed to make the data more consistent.

To perform cross validation part in classification, we need to split data into training and testing data set. Since the data set is highly imbalanced in terms of 19 class labels, we decided to employ stratified sampling method to make sure that there are observations from all ICD9 codes in the training data set. Stratified sampling is a probability sampling technique wherein we divide the entire population into different subgroups or strata, and then randomly select the final subjects proportionally from the different strata.

To build this decision support system, our main input is nurse triage note to predict ICD9 categories but we added vital signs in order to see whether they are able to contribute to the model and improve the classifier performance. In this study, vital signs are treated as categorical variables not continuous ones. Each of them has three categories: high, normal and low. Table (3) presents the thresholds of these three vital levels according to NIH National Library of Medicine.

| | Low | Normal | High |
|--------------------------|-------|---------------------|-------|
| Pulse | <60 | 60 – 100 per minute | >100 |
| Blood Pressure Diastolic | <60 | 60 – 80 | >80 |
| Blood Pressure Systolic | <90 | 90 – 120 | >120 |
| Temperature (F) | <97.8 | 97.8 – 99.1 | >99.1 |
| Respiratory | <12 | 12 – 16 per minute | >16 |

Table 3. Categorical vital signs

4.2. CLASSIFIER

To test our model, we chose to employ the supervised classifiers which are proved to have the best performance in text classification. Since we perform multi-class classification, these classifiers should not be limited to binary-class classification. The most popular metric to evaluate a classifier's performance is accuracy which represents the proportion of true classified records to all testing records.

4.2.1. SUPPORT VECTOR MACHINES

Support Vector Machines (SVMs), Cortes and Vapnik [16], are supervised learning models with associated learning algorithms that analyze data and recognize patterns. This method has a very powerful performance in classification and regression problems. In binary classification, given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. SVM model is a representation of the observations mapped as points in space so that the observations of the separate categories are divided by a clear gap that is as wide as possible. New observations are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

In our research, we perform multi class classification. Under this condition, SVM consider each class versus the rest to complete the training and validation parts.

4.2.2. L₁ REGULARIZED LOGISTIC REGRESSION MODEL

L₁ regularized logistic regression model was introduced by Lokhorst [17]. First, we briefly describe general linear regression model (GLM) in which the response variable is being predicted by a linear combination of the predictors. This relationship can be presented in the expression (7) in which X and β denote the family of predictors and their coefficients respectively.

$$Y = X^T\beta \quad (7)$$

To estimate β , we may need to solve a set of non-linear expressions that satisfy the maximum likelihood criterion.

$$\hat{\beta} = \operatorname{argmax}_{\beta} L(y; \beta) \quad (8)$$

In the expression above, L stands for the likelihood function with respect to the response variable and its predictors. When insignificant predictors are present, we can impose a penalization on the L₁ norm of the coefficients to build the model based on more informative features. Then expression (8) will change to expression (9). In expression (9), $\lambda > 0$ denotes the regularization parameter.

$$\hat{\beta}(\lambda) = \operatorname{argmax}_{\beta} \{-\log L(y; \beta) + \lambda \|\beta\|_1\} \quad (9)$$

4.2.3. RANDOM FOREST

Random forest or random decision forests, proposed by Breiman [18], is an ensemble learning method for classification and regression that operates by constructing high number of decision trees at the training time and outputting the class that is the mode of the classes (in classification) or mean prediction (in regression) of the individual trees.

Since the number of training records in this study was high, almost 30000, the number of trees was determined to be 10 in order to prevent the complexity of the calculations.

CHAPTER 5: RESULTS AND DISCUSSION

Categorical vital signs did not have a significant effect on the classification process so none of them were selected by the feature selection algorithms. The reason could be their low frequencies comparing to selected words from nurse triage note. Table (4) reports the performance of the three classifiers in terms of prediction accuracy with chi-square feature selection considering different values of chi-square critical value. The results of alpha value equal to 0.25 seem to be the best so we consider that as the threshold of chi-square feature selection method.

| Alpha Value | Prediction Accuracy | | |
|--------------|---------------------|------------------------|---------------|
| | SVM | Regularized Regression | Random Forest |
| 0.25 | 0.523 | 0.53 | 0.512 |
| 0.2 | 0.522 | 0.53 | 0.51 |
| 0.15 | 0.523 | 0.527 | 0.511 |
| 0.1 | 0.522 | 0.522 | 0.508 |
| 0.05 | 0.517 | 0.512 | 0.502 |
| 0.025 | 0.503 | 0.497 | 0.497 |
| 0.01 | 0.492 | 0.487 | 0.487 |
| 0.005 | 0.487 | 0.481 | 0.482 |
| 0.001 | 0.457 | 0.456 | 0.454 |

Table 4. Prediction accuracy based on different chi-square alpha values

Final results of the two feature selection methods, Gini-index and chi-squared, are presented in Table (5). To show the effect of the feature selection process, we present the classifiers' accuracy when we simply select the most frequent words without any sophisticated feature selection method.

Clearly, there is not a specific rule able to identify the best feature selection method for this kind of data. So we try to make a comparison here. In selecting the most frequent features and Gini-index feature selection method, we assign the corresponding thresholds of these two methods in a way that the numbers of selected features are almost the same as chi-square feature

selection method with alpha equals to 0.25. In this way, by comparing the results of the classifiers in Table (5), the value of selecting features with CHI and GI is apparent.

| Threshold | Feature Selection Method | Number Selected Features | Random Forest | Regularized Regression | SVM |
|---|--------------------------|--------------------------|---------------|------------------------|-------|
| $\alpha = 0.25$ | CHI | 311 | 0.512 | 0.53 | 0.523 |
| Thresholds equivalent to CHI with $\alpha = 0.25$ | GI | 355 | 0.514 | 0.527 | 0.528 |
| | Most Frequent Words | 365 | 0.493 | 0.504 | 0.495 |

Table 5. Classifiers' results

By analyzing the selected features, we realize that there was a huge overlap between the representative words of several ICD9 categories which compromise the prediction accuracy. To overcome this issue, we think of a probabilistic way. As we know, SVMs tend to calculate the probability of each category to be predicted as the class label of each patient record. In a case that two or three ICD9 categories have similar key words, they might be predicted with close probabilities. According to the probability ranking rule of the SVM, this classifier will announce the category with the highest probability as the predicted class.

In our probabilistic solution, we consider the first, the first two and the first three ranked predicted class labels as the outputs of SVM for each patient record. The accuracy of prediction with SVM would be as follow:

| Feature Selection Method | Number Selected Features | SVM 1 | SVM 1,2 | SVM 1,2,3 |
|--------------------------|--------------------------|-------|---------|-----------|
| CHI | 311 | 0.523 | 0.698 | 0.78 |
| GI | 355 | 0.528 | 0.694 | 0.776 |

Table 6. Top three SVM outputs for each patient record based on SVM ranked probability feature

We may look at this ranked probability outputs from a more strict way by adding an acceptance threshold and accepting the prediction output if the probability of the predicted class label exceeds a constant threshold. In this way, we may have more reliable results in terms of prediction accuracy but this improvement comes with the price of losing many observations which their predicted class label did not exceed the threshold. Tables (7) and (8) represent the

relations among probability thresholds, predicted accuracy and the number of ignored patient records for both feature selection algorithms.

| Probability Threshold | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|------------------------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| Lost Observation (%) | 0 | 4.3 | 56 | 59.7 | 50.2 | 67.6 | 82.7 | 93 | 98.8 |
| Accuracy (%) | 52.3 | 53.5 | 15.4 | 30.7 | 65.5 | 70.7 | 76.9 | 83 | 84.6 |

Table 7. Chi-Square Feature Selection after assigning Probability Threshold

| Probability Threshold | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|------------------------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| Lost Observation (%) | 0 | 2.9 | 11.7 | 29.2 | 48.8 | 69 | 86 | 94.8 | 98.9 |
| Accuracy | 52.8 | 53.5 | 55.4 | 60.4 | 65.7 | 72.2 | 79.9 | 83 | 80.2 |

Table 8. Adopted Gini-Index Feature Selection after assigning Probability Threshold

CHAPTER 6: FUTURE RESEARCH

At this stage of our research, we have the best result that feature selection methods can provide because we employed the two available methods known to have the best result in class-imbalanced problems. While the results of classifiers are promising, we believe that taking different approaches can improve the result.

One of the best ways to improve the performance of the decision support system is building a customized library which benefits from external medical resources. Building this library will take a lot of time and effort since we need to go through the related documents of each category and identify their keywords. Then we should use the keywords to make the available features more strong in order to predict ICD9 code of patient records in the dataset. In this approach, we will have a set of more reliable features which hopefully make the accuracy better while improving the robustness of our decision support system.

Another approach is employing deep learning in dimensionality reduction part. Deep learning is a novel method in dimension reduction especially for medical domain data. We are willing to employ deep learning dimension reduction however representing a document to a deep learning architect is hard. Building the customized library which was mentioned earlier can help with presenting the dataset to the deep learning dimension reduction architect. This method has proved its effectiveness in recent years and can be a good way to help improving accuracy of the prediction.

REFERENCES

1. Olshaker JS (2009). Managing emergency department overcrowding. *Emergency Medicine Clinics of North America*. Volume 27, Issue 4, Pages 593-603.
2. Qiu Sh, Chinnam RB, Murat A, Batarse B, Neemuchwala H, Jordan W (2015). A cost sensitive inpatient bed reservation approach to reduce emergency department boarding times. *Health Care Management Science*. Volume 18, Issue 1, Pages 67-85.
3. Derlet RW, Richards JR, Kravitz RL (2000). Frequent overcrowding in U.S. emergency departments. *Academic Emergency Medicine*. Volume 8, Issue 2, Pages 151-155.
4. Powell ES, Khare RK, Venkatesh AK, Van Roo BD, Adams JG, Reinhardt G (2010). The relationship between inpatient discharge timing and emergency department boarding. *The Journal of Emergency Medicine*. Volume 42, Issue 2, Pages 186-196.
5. Boyle A, Beniuk K, Higginson I, Atkinson P (2012). Emergency department crowding: Time for interventions and policy evaluations. *Emergency Medicine International*. Volume 2012.
6. Moskop JC, Sklar DP, Geiderman JM, Schears RM, Bookman KJ (2009). Emergency department crowding, part 1—concept, causes, and moral consequences. *Annals of Emergency Medicine*. Volume 53, Issue 5, Pages 605-611.
7. Moskop JC, Sklar DP, Geiderman JM, Schears RM, Bookman KJ (2009). Emergency department crowding, part 2—barriers to reform and strategies to overcome them. *Annals of Emergency Medicine*. Volume 53, Issue 5, Pages 612-617.
8. MacIntyre CR, Ackland MJ, Chandraraj EJ, Pilla JE (1997). Accuracy of ICD-9-CM codes in hospital morbidity data, Victoria: implications for public health research. *Australian and New Zealand Journal of Public Health*. Volume 21, Issue 5, Pages 477-482.

9. Travers DA, Haas SW (2003). Using nurses' natural language entries to build a concept-oriented terminology for patients' chief complaints in the emergency department. *Journal of Biomedical Informatics*. Volume 36, Issue 4-5, pages 260-270.
10. Chapman WW, Dowling JN, Wagner MM (2005). Classification of emergency department chief complaints into 7 syndromes: A retrospective analysis of 527,228 patients. *Annals of Emergency Medicine*. Volume 46, Issue 5, pages 445-455.
11. Mahalingam D, Mostafa J, Travers D, Haas S, Waller A (2012). Automated syndrome classification using phase emergency department data. *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium* Pages 373-378.
12. Yang Y, Pedersen JO (1997). A comparative study on feature selection in text categorization.
13. Zheng Z, Wu X, Srihari R (2004). Feature selection for text categorization on imbalanced data. *ACM SIGKDD Explorations Newsletter - Special issue on learning from imbalanced datasets*. Volume 6, Issue 1, Pages 80-89.
14. Peng H (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions*. Volume 27, Issue 8, Pages 1226-1238.
15. Shang W, Huang H, Zhu H, Lin Y, Qu Y, Wang Z (2007). A novel feature selection algorithm for text categorization. *Expert Systems with Applications*. Volume 33, Issue 1, Pages 1-5.
16. Cortes C, Vapnik V (1995). Support-Vector Networks. *Machine Learning*. Volume 20, Pages 273-297.
17. Lokhorst J (1999). The lasso and generalized linear models, Technical report, University of Adelaide.

18. Breiman L (2001). Random Forests. *Machine Learning*. Volume 45, Issue 1, Pages 5-32.

ABSTRACT**PREDICTIVE ANALYTICS FOR DISEASE CONDITION OF PATIENTS IN
EMERGENCY DEPARTMENT**

by

AZADE TABAIE**December 2015****Advisor:** Prof. Ratna Babu Chinnam**Major:** Industrial Engineering**Degree:** Master of Science

Emergency Departments (EDs) in hospitals are experiencing severe crowding and prolonged patient waiting times. The reported crowding in hospitals shows patients in hospital hallways, long waiting times and full occupancy of ED beds. ED crowding has several potential unfavorable effects including patients and staff frustration, lower patient satisfaction and poor health outcomes. The primary motivations behind this study are shortening the patients' waiting time and improving patient satisfaction and level of care.

The very initial interaction between clinicians and a patient is recorded on nurse triage notes which contain details of the reason for patient's visit including specific symptoms and incidents. Triage notes and vital signs measured by triage nurse determine the complexity of the patient's condition. If a minor illness or injury occurred, patient would be treated by nurse practitioners under ED physicians' supervision. This process called fast track system which allows the main ED area to focus on more severe patient condition. The final decision should be

made by physicians so patients have to wait to be seen in order to find out whether they need to be admitted in the hospital or be discharged.

In this study, we propose a decision support system based on nurse triage notes and vital signs that can automatically predict ICD9 code assigned to each patient prior to the visit time. We tested the model on 8000 patient records from VA Medical Center in Detroit for ICD9 classification and measured performance in terms of accuracy.

AUTOBIOGRAPHICAL STATEMENT

Azade Tabaie received her Bachelor's degree in Applied Mathematics at Amirkabir University of Technology in Tehran, Iran. Currently, she works as research assistant in Big Data and Business Analytics Group at Wayne State University. She has been studying towards her Master's degree in Industrial Engineering at Wayne State University since August 2013. Her research interest includes text mining, healthcare data analytics, artificial intelligence, statistics and predictive modeling.