Mining Patients Medical Records Based On MIMIC-II Clinical Database

CAN YANG

April 2018

1 INTRODUCTION

Mining in MIMIC-II(Medical Information Mart for Intensive Care II) database as a clinical task for association of medicine or disease with patients. This report gives an overview of medicine and disease information using some data mining method. But very detailed personal results will not be included in this report. So different medicines or diseases clustering can be identify similar cases by different patients medical history by a deep learning framework in NLP tool in order to calibrate the missing or fuzzy data. We are mainly talking about the medicine type exploration and verification. And the results of disease are mentioned as well. The association of specific diseases with drugs will be an interesting search, so this will be the future work of current research.

2 METHOD

Some clinical notes from doctor are missing for some reasons[2]. This become a challenge of the data analysis, so NLP techniques which is word2vec is used for vector space to find similarities in this project. Each medicine and disease is represented by number which called code in MIMIC-II database. After generation of the medical vectors which are some disordered numbers, it can be convert to corpus as string as a input to this two layers neural network. This model can be implemented by compressed vector format with different dimension sizes. It is an unsupervised learning framework depending on the context and Skip-gram predicts surrounding words given the current word[1]. In order to visualizse result, t-SNE is used here to reduce the dimension to 2 or 3 dimension space. The steps can be simplified as Figure 1.

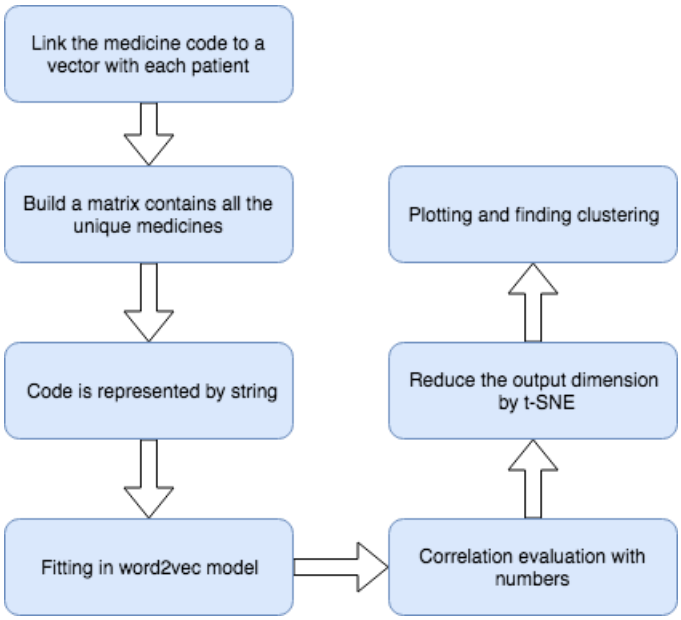


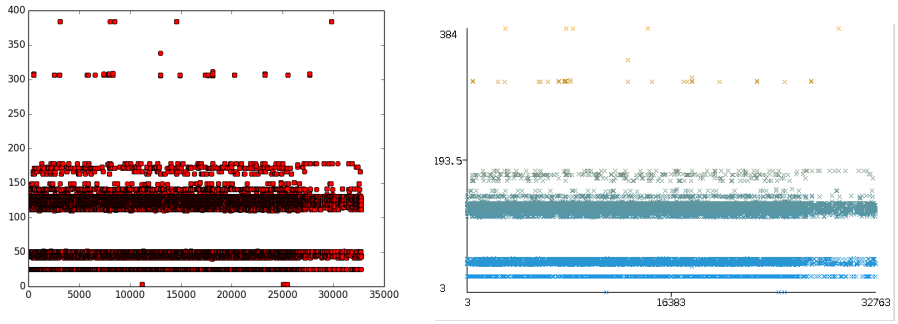
Figure 1: Simple Steps

3 IMPLEMENTATION

A large amount of medical data of ICU is collected from mimic2. According to various subsets, medicine and disease these two directions are chosen to analysis in this report. Information of them based on different patients in different ICU rooms are observed.

3.1 Medications

During the data cleansing process, the missing data of "NULL" is deleted. Visualised raw data can provide rough ideas about the data set. A simple plot by Weka and python can be found in Figure 2. The medicine code between 100 to 150 are used at high frequency. And the upper part in the figure are less. These two areas can be further checked.



(a) Plot by Python (b) Plot by Weka

Figure 2: Relations between Patients and Medicine

For same subject ID and same ICU ID, a subset is created to store the information for each patient. So, in total, a list of list will be generated. A medicine analysis is handled first. There are some natural numbers that represent medicine. Their names can be found in other subsets and will be merged later.

First, model "word2vec" is used here which will sign a vector to each unique element. In order to get vector space of each medicine, it is converted to string as shown in Figure 4. Then for each vector, the string can be decoded to a substring as Figure 3.

```
decode [[u'43'], [u'128'], [u'48'], [u'141'], [u'118'], [u'120'], [u'112'], [u'25'], [u'131'], [u'45'], [u'25'], [u'121'],
  [u'112'], [u'120'], [u'51'], [u'126'], [u'128'], [u'130'], [u'120'], [u'128'], [u'50'], [u'148'], [u'43'], [u'128'],
  [u'45'], [u'128'], [u'43'], [u'131'], [u'131'], [u'138'], [u'120'], [u'120'], [u'51'], [u'124'], [u'118'], [u'128'],
  [u'120'], [u'128'], [u'45'], [u'131'], [u'51'], [u'25'], [u'50'], [u'117'], [u'50'], [u'127'], [u'43'], [u'131'], [u'25'],
  [u'128'], [u'121'], [u'115'], [u'131'], [u'45'], [u'25'], [u'120'], [u'124'], [u'118'], [u'43'], [u'25'], [u'131'],
  [u'128'], [u'43'], [u'120'], [u'123'], [u'131'], [u'127'], [u'117'], [u'25'], [u'131'], [u'118'], [u'141'], [u'128'],
  [u'43'], [u'123'], [u'131'], [u'131'], [u'118'], [u'124'], [u'45'], [u'123'], [u'117'], [u'25'], [u'50'], [u'131'], [u'50'],
  [u'25'], [u'131'], [u'149'], [u'124'], [u'43'], [u'120'], [u'51'], [u'121'], [u'122'], [u'128'], [u'43'], [u'50'], [u'45'],
  [u'128'], [u'43'], [u'131'], [u'124'], [u'118'], [u'25'], [u'124'], [u'126'], [u'119'], [u'128'], [u'121'], [u'131'],
  [u'45'], [u'120'], [u'25'], [u'120'], [u'120'], [u'118'], [u'124'], [u'126'], [u'125'], [u'131'], [u'128'], [u'149'],
  [u'43'], [u'128'], [u'25'], [u'45'], [u'120'], [u'122'], [u'43'], [u'131'], [u'118'], [u'121'], [u'45'], [u'25'], [u'50'],
  [u'112'], [u'115'], [u'128'], [u'120'], [u'126'], [u'131'], [u'25'], [u'25'], [u'121'], [u'142'], [u'120'], [u'51'],
  [u'128'], [u'25'], [u'121'], [u'125'], [u'131'], [u'50'], [u'112'], [u'45'], [u'42'], [u'120'], [u'124'], [u'118'], [u'51'],
  [u'126'], [u'131'], [u'120'], [u'124'], [u'118'], [u'25'], [u'25'], [u'126'], [u'120'], [u'120'], [u'42'], [u'114'],
```

Figure 3: Translate to substring

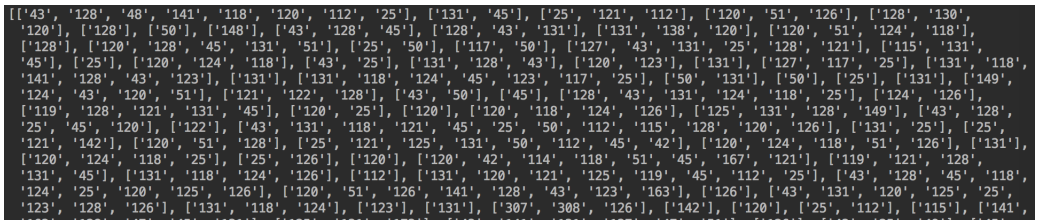


Figure 4: Translate to Vector

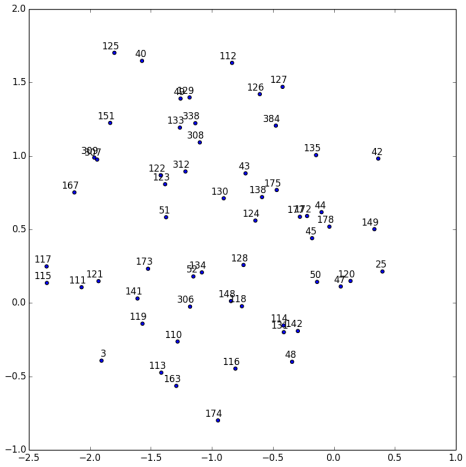
After fitting the data in the model, a high dimensional matrix is created and each dimension corresponds to one medicine. There are 59 different kinds of medicines in this dataset. If the frequency smaller than the mini count, it can be negligible. And another parameter that can be considered in this model is the dimension size. And there are different modes or different sizes of display clusters later. Then, in order to visualise the data, we reduced the high-dimensional matrix to two dimensions using t-SNE. Figure 5 shows the distribution of the result with numbers.

As the Figure 6, it shows different distribution. We can find the relations or association among different medicines. From Figure 6a, Figure 6c and Figure 6d, “Esmolol” and ”Diltiazem” are two similar medicines which are blockers. And all of “Doxacurium”, “Fentanyl Drip” and “Pancuronium” are anesthetic since there are similar we can find from Figure 6a and Figure 6c. If we look for “Dobutamne” and it neighbor medicine in these four figures, there are “Natrecor”, “Dopamine”, “Epinephrine”, “Integrelin”, “Nitroglycerine-k” and “Amrinone”. In this class, it is mainly used to treat heart-related diseases. Also, they are in the same cluster for four subfigures of Figure 6. As is shown in Figure 6a, Figure 6b, Figure 6c and Figure 6d, the closest point to “Insulin” is “Nicardipine”,and they are two treatments of blood. “Insulin” is for high blood sugar and “Nicardipine” is for high blood pressure.

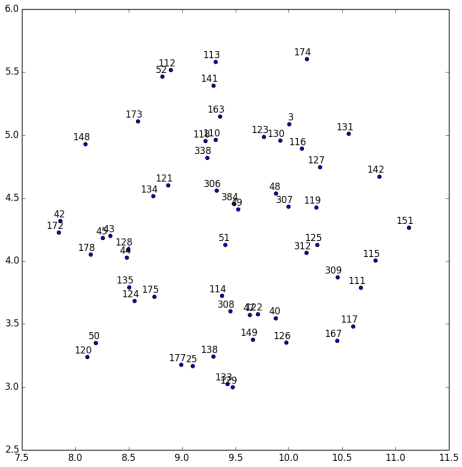
However, it is not always good result in theses four figures. For example, “Dobutamne” and “Dobutamine Drip” are almost two similar drug, but they are never close to each other in these four figures. The reason is that they are too similar to apply to one patient. So If we observe the surrounding drugs “Dobutamine Drip”, “Aggrastat” and “Procainamide” are close to it in Figure 6b. And they are used to treat heart failure or cardiogenic shock. “Narcan” and “Neosynephrine-k” are neighbors in Figure 6c and Figure 6d, because these two are medicine for the maintenance of blood pressure. In addition, even “Dobutamine” is not the same class with anesthetic drugs, but they always use together for surgery since Dopamine helps wake up anesthesia after surgery.

In total, these results in Figure 6 make sense in some samples. There are still points or clusters in the zone that I can get reasonable information since the dataset is too small. In this experiment, clusters can be found based on

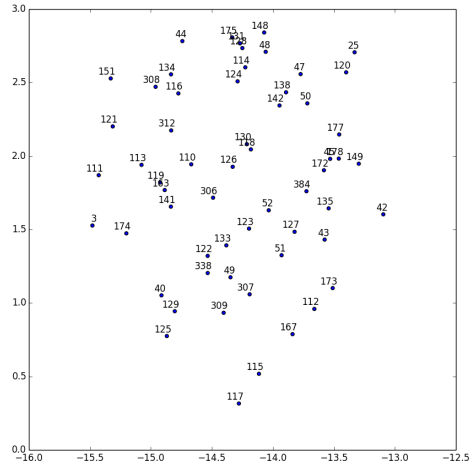
each patient's usage record. By switching to contextual form and searching for relationships in the content, we can gain some knowledge about the medicine class. And among these four figures, the result of Figure 6b is not good enough. Compare to this, others can provide me more reliable information. Setting to 30 dimensions give the best performance for this dataset.



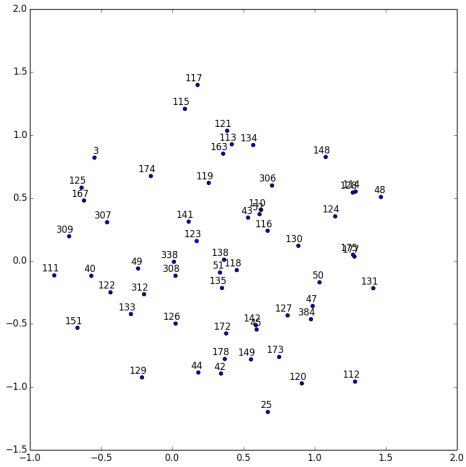
(a) 30 dimension of feature vector



(b) 10 dimension of feature vector

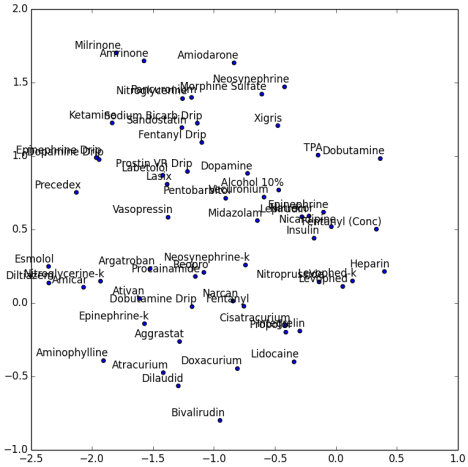


(c) 40 dimension of feature vector

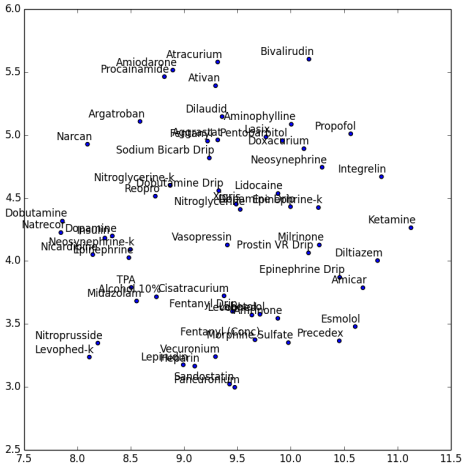


(d) 50 dimension of feature vector

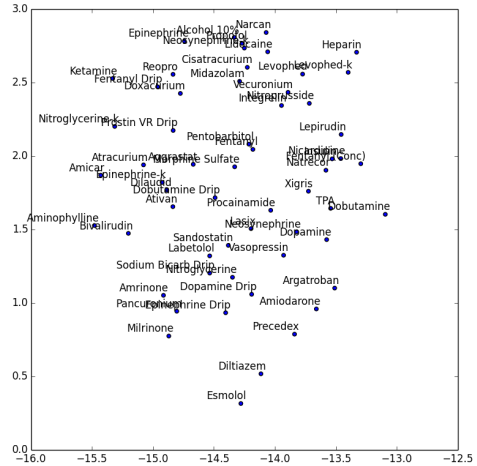
Figure 5: Medicine Result Marked by Number



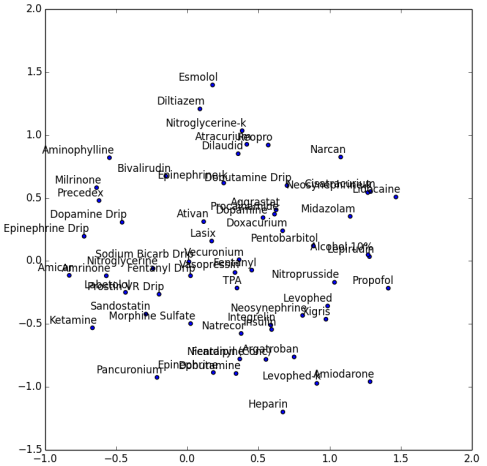
(a) 30 dimension of feature vector



(b) 10 dimension of feature vector



(c) 40 dimension of feature vector



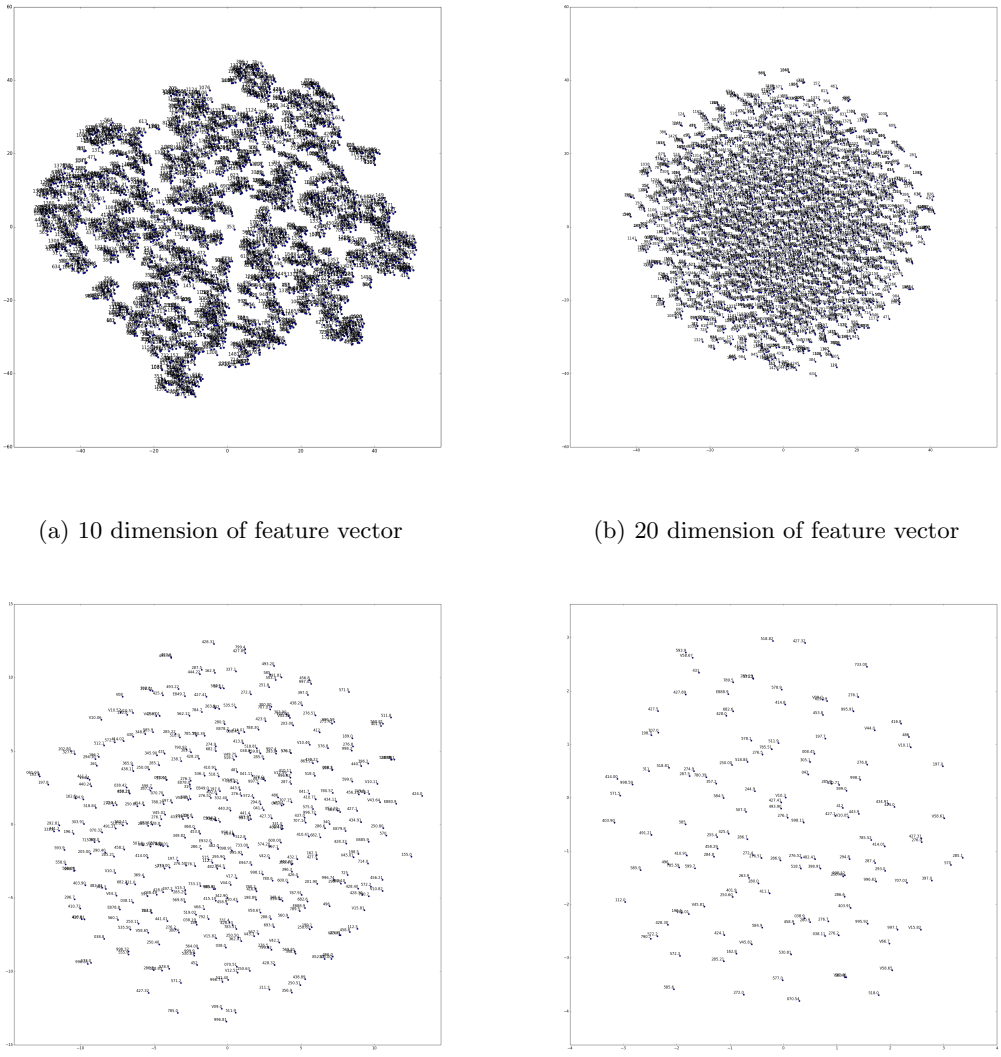
(d) 50 dimension of feature vector

Figure 6: Medicine result Marked by Name

3.2 Diagnosis

After the medicine analysis, the same method is used for disease analysis. Therefore, the disease of each patient living in the same ICU room is stored in a vector. And find the relevance between these disease to get some ideas. The result is

shown in Figure 7.



(a) 10 dimension of feature vector

(b) 20 dimension of feature vector

(c) 20 dimension with 20 min count

(d) 20 dimension with 50 min count

Figure 7: Disease Result

The difference between Figure 7a and Figure 7b is the size of dimensions. The first figure is 10 and another is 20, so it shows different distributions. The rest of figures are got from the different “min count”. The value of “min count”

are 20 and 50 respectively. If some diseases only appear very few times, for a clear overview it can be ignored. According to the evaluation of medicine case, the result of my "Word2vec" model is not very reliable. Unfortunately, we will not talk about how to choose good model parameters to optimize performance in this paper. And the result of the disease is not very clear enough here. Figure 7a shows a rough review of the disease types.

4 Discussion And Future Direction

In this experiment, a large data set of an ICU is handled and there are some detailed records of patients. The information related to medicine and disease are targets for this paper. A "Word2Vec" method which two layers neural network is used during this procedure and plot data with "t-SNE" which can reduce the dimensions. Different size of dimensions are fitting in the model respectively and 30 dimensions among 59 kinds of medicines provides the best performance. With the exception of Anesthetics, most of these drugs are distributed in the same category. Some interesting knowledge are found, for instance Dopamine and Anesthetics normally used together. But some results are not make sense. So after this unsupervised algorithm provide similarity of medicine, some another method should add for evaluation or verification. In addition, a same method is applied on the disease set and a simple conclusion are shown by figures among different parameters.

For future work, the relationship between medicine and disease can be further discovered. Using same parameter of model to evaluate matching performance between them. Also, there is more potential space for disease in this research.

References

[1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[2] Benjamin Shickel, Patrick Tighe, Azra Bihorac, and Parisa Rashidi. Deep EHR: A survey of recent advances on deep learning techniques for electronic health record (EHR) analysis. *CoRR*, abs/1706.03446, 2017.