DEFINING THE STATES OF THE PATIENTS WITH ERYSIPELAS DISEASE USING DIFFERENT DIMENSIONALITY REDUCTION TECHNIQUES

R.R. Kotyubeev, D.A. Zhurman
Scientific supervisor: S.V. Axyonov
Tomsk Polytechnic University
E-mail: rrk8@tpu.ru

Introduction

Erysipelas are the fourth most common in the world among infectious diseases and, in most cases, can be cured. Erysipelas most often affects infants and the elderly but can affect any age group The main symptoms of this disease are skin rash, fever, pressure, palpitations, headaches, loss of sleep and appetite.

Thus, making predictions to analyze the further state of a patient with erysipelas will be crucial. It helps doctors define how many days the patient will be in hospital. The research may be useful for medical insurance companies, as it will allow you to correctly determine the condition of the patient, and thereby adjust the amount of insurance payments for the treatment of the patient

The aim of this work is to consider the states of patients which treated in hospital with erysipelas through the treatment course based on the digitalized patients' history.

Because the input data represented as texts, TF-IDF will be used for information retrieval. To analyze the states of patients dimensionality reduction.

Data Description

The data is represented as a collection of the electronic medical records. All patients in the data were treated in hospitals with one diagnosis – Erysipelas. The dataset consists of 58 such records.

Table 1. A part of a patient's history

1	' жалоба на повышение температура, отечь правый голень. на кожа правый голень эритема, отечь тестовидный консистенция '
2	' жалоба на повышение температура, отечь правый голень. на кожа правый голень эритема, отечь тестовидный консистенция'
3	' жалоба на боль правый голень. на кожа правый голень эритема угасать, отечь тестовидный консистенция уменьшиться в объём кожный покров - с больший количество пигментный пятно, влажный'
4	' жалоба на боль правый голень. на кожа правый голень эритема угасать, отечь тестовидный консистенция уменьшиться в объём кожный покров - с больший количество пигментный пятно, влажный'
5	' жалоба на боль правый голень. на кожа правый голень эритема угасать, отёка нет, кожный покров - с больший количество пигментный пятно, влажный

Table 1 represents 5 days from 11 days of a patient's history after lemmatization. Some days repeated because the states of patients didn't change according to a doctor notes.

TF-IDF

TF-IDF stands for term frequency–inverse document frequency [1]. It reflects how important a word is to a document in a collection or corpus.

Term frequency defined as the number of times a word appears in a document, divided by the total number of words in that document:

$$tf_i(t_i, d_i) = \frac{n_{d_i}}{total_{d_i}}, \qquad (1)$$

where the number of times n_{d_i} token (word or n-gram) t_i appears in a document d_i, divided by the total number of words in that document.

The inverse document frequency is a measure of how much information the word provides, i.e., if it's common or rare across all documents:

$$idf(t_i, D) = \ln\frac{N_D}{n_t}, \qquad (2)$$

where N_D is total number of documents in the corpus D, n_t is the number of documents with token t in t_i.

TF-IDF for a token in a document is calculated by multiplying two term frequency and inverse document frequency:

$$tfidf(t_i, d_i, D) = tf_i(t_i, d_i) \cdot idf(t_i, D). \; (3)$$

Lemmatization and Stop words

Lemmatization is a method of morphological analysis, which boils down to reducing the word form to its original vocabulary form (lemma).

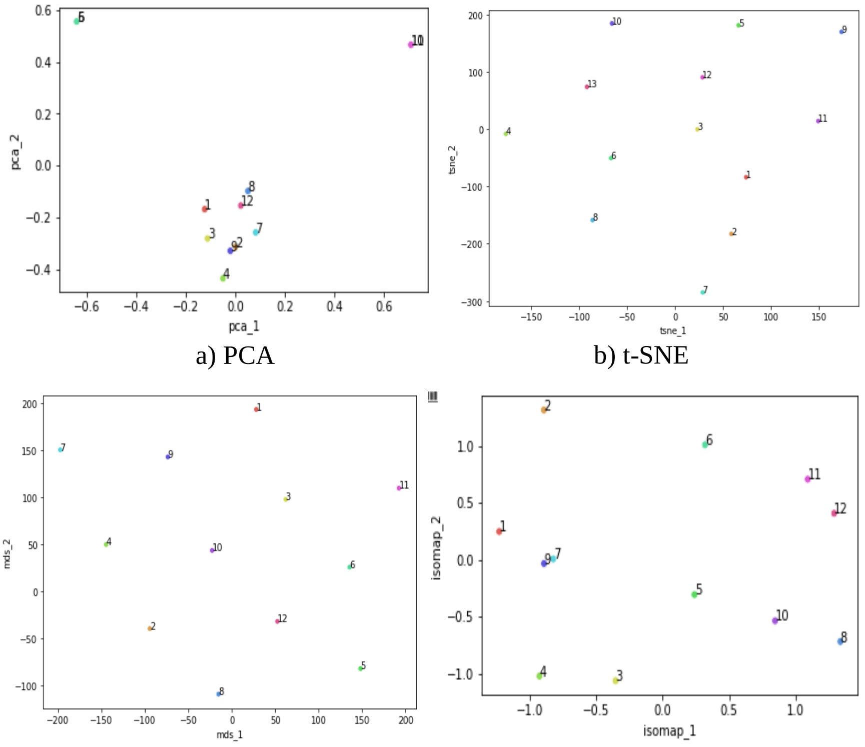
The lemmatization method is used in search algorithms in the process of schematizing web documents when they are indexed.

As a result of lemmatization, inflectional endings are discarded from the word form and the main or dictionary form of the word is returned.

For example, in Russian, the dictionary form for:

- Nouns - nominative case, singular (with swords - a sword);
- Verbs - infinitive form;
- Adjectives - singular, nominative, masculine.

Moreover, a text may consists of some words which doesn't provide valuable information. For instance, it can be function words or repeated words in the text.



a) PCA b) t-SNE

c) MDS d) ISOMAP

Fig. 1. The dimensionality reduction results of different techniques

Research Methodology

To analyze the results different dimensionality reduction techniques was used. Dimensionality reduction is necessary to reduce numbers of TF-IDF components.

PCA (a), t-SNE (b), ISOMAP (c), MDS (d) [2, 3, 4, 5] will be used. Such techniques helps to visualize commonalities between the data omitting the other components.

To build the expected results some libraries for the Python will be used. Some of them observed in [6]. There are:

1. *Scikit-learn* library provides tools for transforming data and dimensionality reduction.
2. *Seaborn* library will be used to plot the results of dimensionality reduction.
3. *Pymorphy2* [7] library will be used to make lemmatization. It provides lemmatization of Russian words.
4. *NLTK* library provides stop words.

Results

Fig. 1 shows how PCA, t-SNE, ISOMAP, MDS were used to proceeded the TFIDF scores of a patient. The number on the plots is the day that patient was in hospital (overall it is 12 days).

Smooth and continuous states of the patient was expected but none of these techniques didn’t show such results besides different parameters values for each technique. Any pattern of changing states wasn’t found. Such results can be explained because of these techniques omits a lot of components in a small data. A patient’s history did not contain large information that can be applied with dimensionality reduction.

Reference

1. Rajaraman, A.; Ullman, J.D. (2011). "Data Mining". Mining of Massive Datasets. pp. 1–17
2. Medium / Understanding Principal Component Analysis // URL: https://medium.com/@aptrishu/understanding-principle-component-analysis-e32be0253ef0. – (accessed 14.12.2019).
3. DataCamp / Introduction to t-SNE // URL: https://www.datacamp.com/community/tutorials/introduction-t-sne. – (accessed 14.12.2019).
4. Towards Data Science / Decomposing on-linearity with ISOMAP // URL: https://towardsdatascience.com/decomposing-non-linearity-with-isomap-32cf1e95a483. – (accessed 14.12.2019).
5. PaperspaceBlog / Multi-Dimension Scaling (MDS) // URL: https://blog.paperspace.com/dimension-reduction-with-multi-dimension-scaling/. – (accessed 14.12.2019).
6. Рашка С. Python и машинное обучение / пер. с англ. А. В. Логунова. М.: ДМК Пресс, 2017. – 133 с.: ил
7. Морфологический анализатор pymorphy2 / pymorphy2 // URL: https://pymorphy2.readthedocs.io/en/latest/ (accessed 14.12.2019).
8. Natural Language Toolkit / NLTK // URL: https://www.nltk.org/ (accessed 14.12.20)