Similarity study of clinical data

Wenjun Yu, Kai Wang, Bin Hu and Yu Huang*

College of informatics, Huazhong Agricultural University, Wuhan, Hubei, China

*Corresponding author e-mail: yhuang@mail.hzau.edu.cn

Abstract. The classification of patients is very important for accurately determining the patient's disease and subsequent drug use. More and more clinical data provide a clue for the classification of various diseases. We use related software to generate simulation data based on ICD10 and propose a specific algorithm to classify patients, and find patients belonging to other categories in the classified patient group, which provides a theoretical basis for the improvement of various classification algorithms for diseases.

1. Introduction

With the great progress of similarity algorithms on patient clustering, accurately predicting the future development of patients' diseases will not only help to prevent diseases and provide drug use, but also help to provide early diagnosis to patients, so as to control the growing high-value medical care [1]. The accuracy of patient similarity algorithms is highly related to the preprocessing, structuring, system analysis and classification of clinical data. In fact, classification algorithms require the use of multi-dimensional statistical methods and algorithms. Patient similarity analysis is one of the main prediction methods based on medical big data. The main principle is to calculate the similarity of structured clinical data, classify patients, and then predict the condition of the target patient through the feature coefficients of each category [2-6].

In recent years, similar algorithms for clinical data have emerged continuously. These algorithms are mainly used for the prevention and diagnosis of breast cancer, lung cancer and other diseases. For example, in breast cancer screening, the use of related classification algorithms has achieved good prediction results [7], in the early prevention of lung cancer [8], the similarity algorithm provided by the relevant authors not only improves the existing prediction models, but also improves forecasting efficiency.

According to a small amount of clinical data provided by Huaxia-Hospital, we simulate the clinical data of 5000 patients and encode them in ICD10 code [9,10]. Some effective algorithms are used to calculate the similarity of structured clinical data and use this as a basis to classify patients. By comparison with manual classification, the accuracy rate reached 90.54%.

2. Method

2.1. Data filter and pre-process

The purpose of filtering clinical data is to select specific data types and data items to meet the purpose of similarity calculation. Because the clinical data has extremely large attribute items, patient data needs to be selected according to actual application requirement [11].

In this paper, we only used patient data encoded in ICD10 code to replace unstructured disease description text with ICD10 code. Patient similarity studies must be based on vast amounts of medical
data, and the standardized clinical data obtained from hospitals is limited, so we simulated clinical
data of 5000 patients coded with ICD10 (Figure 1). As Figure 1, the first column is the simulated
patient number, followed by the disease code, and the file format is csv (Comma-Separated Values).

2.2. Similarity algorithm selection and implementation

The distance between clinical concepts depends on their value attributes: numerical, Boolean,
coding and sequence. The simulated patient data in this paper is coding type. Therefore, we adopt an
information theory or a hierarchy-based method to estimate the distance, that is, calculate the
similarity between the two patients [12]. We used the correlation similarity algorithm of General
Semantic Graphs based on the ICD10 OWL file (OWL: The Web Ontology Language). The steps are
as follows:

Patient P contains m ICD10 codes (P₁, P₂...Pₘ), and Q contains n ICD10 codes (Q₁, Q₂...Qₙ):

1. For each Pᵢ (i = 1, 2, 3... m), the similarity with Q₁, Q₂, ..., Qₙ is calculated, and the maximum
   value Pᵢmax (i = 1, 2, 3... m) is taken.

2. For each Qⱼ (j=1, 2, 3...n), the similarity with P₁, P₂...Pₘ is calculated, and the maximum value
   Qⱼmax (j=1, 2, 3...n) is taken.

3. Finally sim(P, Q) = \frac{\sum_{i=1}^{m} Pᵢmₐₓ + \sum_{j=1}^{n} Qⱼmₐₓ}{m + n}.

The final result file is shown as follow (Figure 2):

![Figure 1. Simulated experimental data.](image1)

![Figure 2. Patient pairs similar values, a total of 12,500,000 value pairs.](image2)
2.3. Classification of patients

In specific clinical data, the most important classification levels are usually based on diagnostic, treatment and care data. The classification should be very targeted, and the relevant chemical components should be marked on the basis of laboratory test results [13].

We extract the corresponding coding composition vector data according to the patient data and the corresponding file of ICD10, then calculate the cosine similarity between the two vectors [14] and cluster the results. There are 12 clusters with at least 300 members that were generated through the clustering results. In all clusters except one cluster, 180 individual ICD10 codes have other characteristics, and the TF-IDF value of this characteristic accounts for 20-36% of the sum of all other TF-IDF values [15].

3. Result

From the above result similarity document, we can use the proposed algorithm to classify structured clinical data. We use CS (Cosine Similarity) [14] to calculate the similarity between all vector pairs. When manually checking the cluster tree, we modify the value at 0.5, resulting in about 12 clusters. There are at least 300 or more members in 12 categories, including 5,000 patients.

We evaluated the accuracy of the proposed algorithm by manually checking the disease codes of approximately 3800 active patients (Table 1). The verification set includes approximately 180 ICD10 codes.

Table 1. Accuracy is the sum of true positives divided by the sum of effective values. In terms of accuracy, each ICD10 code only counts the specified number of times for each patient. If the ICD10 code corresponding to the patient is correct within the specified number of times, it is considered correct. The last line describes the accuracy of all series.

| Series | Incident precision | Data precision (ICD10) |
|--------|--------------------|------------------------|
|        | True | False | Precision | True | False | Precision |
| I      | 210  | 8     | 98.71%     | 230  | 3     | 96.33%    |
| II     | 180  | 20    | 93.75%     | 180  | 12    | 90.00%    |
| VI     | 230  | 109   | 92.23%     | 190  | 16    | 67.85%    |
| VII    | 140  | 8     | 98.15%     | 212  | 4     | 94.59%    |
| IX     | 452  | 27    | 68.75%     | 176  | 80    | 94.36%    |
| X      | 1167 | 50    | 89.41%     | 287  | 34    | 95.89%    |
| XI     | 898  | 15    | 78.51%     | 285  | 78    | 98.36%    |
| XII    | 686  | 38    | 80.56%     | 87   | 21    | 94.75%    |
| XIII   | 128  | 62    | 96.15%     | 125  | 5     | 67.37%    |
| XIV    | 313  | 11    | 98.51%     | 265  | 4     | 96.60%    |
| XIX    | 156  | 8     | 96.55%     | 28   | 1     | 95.12%    |
| XX     | 80   | 4     | 87.50%     | 14   | 2     | 95.24%    |
| All    | 4640 | 360   | 89.90%     | 2079 | 260   | 90.54%    |

4. Conclusion

As the EPR system becomes the storage of medical data today, our research goal shifts to the study of clinical data to improve the accuracy of disease classification and prediction [16]. We collect a variety of structured, semi-structured and unstructured data, based on these data we can use various text mining tools and algorithms. We believe that we use business processes throughout the process,
which can extract valuable patient information from the EPR system. In addition, we use the above algorithm to show how to use this information for disease classification and clustering, and how to apply it to disease prevention and prediction, to understand the source of disease from a system perspective [17].

5. Acknowledgments

This work was financially supported by the Fundamental Research Funds for the Central Universities (Program No. 2662015PY057) and the Horizontal Subject of Huazhong Agricultural University (Network clinical medical data analysis system).

References

[1] Jia Zheng, Zong Ruijie, Duan Huilong, et al. A Review on the Patient Similarity Analysis Based on Electronic Medical Records[J]. Chinese Journal of Biomedical Engineering, 2018, 37(3): 353-366.
[2] Hoogendoorn M, El Hassouni A, Mok K, et al. Prediction using patient comparison vs. modeling: A case study for mortality prediction[C]//2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2016: 2464-2467.
[3] Sharafoddini A, Dubin J A, Lee J. Patient similarity in prediction models based on health data: a scoping review[J]. JMIR medical informatics, 2017, 5(1): e7.
[4] Zhang P, Wang F, Hu J, et al. Towards personalized medicine: leveraging patient similarity and drug similarity analytics[J]. AMIA Summits on Translational Science Proceedings, 2014, 2014: 132.
[5] Ng K, Sun J, Hu J, et al. Personalized predictive modeling and risk factor identification using patient similarity[J]. AMIA Summits on Translational Science Proceedings, 2015, 2015: 132.
[6] Brown S A. Patient similarity: emerging concepts in systems and precision medicine[J]. Frontiers in physiology, 2016, 7: 561.
[7] Wang F, Sun J. PSF: a unified patient similarity evaluation framework through metric learning with weak supervision[J]. IEEE journal of biomedical and health informatics, 2015, 19(3): 1053-1060.
[8] Lowsky D J, Ding Y, Lee D K K, et al. AK-nearest neighbors survival probability prediction method[J]. Statistics in medicine, 2013, 32(12): 2062-2069.
[9] Li L, Cheng W Y, Glicksberg B S, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity[J]. Science translational medicine, 2015, 7(311): 311ra174-311ra174.
[10] Roque F S, Jensen P B, Schmock H, et al. Using electronic patient records to discover disease correlations and stratify patient cohorts[J]. PLoS computational biology, 2011, 7(8): e1002141.
[11] Wang F, Sun J. PSF: a unified patient similarity evaluation framework through metric learning with weak supervision[J]. IEEE journal of biomedical and health informatics, 2015, 19(3): 1053-1060.
[12] Girardi D, Wartner S, Halmerbauer G, et al. Using concept hierarchies to improve calculation of patient similarity[J]. Journal of biomedical informatics, 2016, 63: 66-73.
[13] Ma H, Wang Y, Sullivan-Halley J, et al. Use of four biomarkers to evaluate the risk of breast cancer subtypes in the women’s contraceptive and reproductive experiences study[J]. Cancer research, 2010, 70(2): 575-587.
[14] Lage K, Karlberg E O, Störrling Z M, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders[J]. Nature biotechnology, 2007, 25(3): 309.
[15] Robertson S E, Jones K S. Relevance weighting of search terms[J]. Journal of the American Society for Information science, 1976, 27(3): 129-146.
[16] Plönnick R M. The progression of electronic health records and implications for psychiatry[J]. American Journal of Psychiatry, 2010, 167(5): 498-500.
[17] Embi P J, Jain A, Clark J, et al. Development of an electronic health record-based Clinical Trial Alert system to enhance recruitment at the point of care[C]//AMIA Annual Symposium Proceedings. American Medical Informatics Association, 2005, 2005: 231.