An Automatic Disease Diagnosis Method Based on Big Medical Data

Xu Luo, Yonghu Chang
Research Center of Big Medical Data
Zunyi Medical University
Zunyi, Guizhou, P. R. China 563000
Email: {luoxu, changyonghu}@zmc.edu.cn

Jun Yang
Department of Information Science and Engineering
Wuhan University of Science and Technology
Wuhan, Hubei, P. R. China 430081
Email: yangjun@wust.edu.cn

***Abstract*— Medical diagnoses automation based on huge data is of great significance to doctor-patient contradiction regulation and the optimizing allocation of health resources. This paper presents an intelligent automatic disease diagnosis system. Based on a large number of confirmed disease cases, the method of probability and statistics is used to find the relations between the symptoms and different diseases, the confirmed symptom sets of different diseases are obtained first, and then the diagnoses are given by examining that how many symptoms are both in the symptoms provided by patients and the confirmed symptom sets of different diseases. The automatic disease diagnoses system has been implemented preliminarily and the system is proved having some value in clinical application.**

Keywords*— *big medical data; automatic disease diagnosis system; probability and statistics

I. Introduction

With the fast development of information technology, the explosive growth of information produces a huge amount of data. We are in the ocean of data and almost all the things in daily life are related to data. There is huge amount of environmental data, financial data, medical data, shopping data and communication data. Data from various fields reflects the habits and customs of people, social laws and natural laws, and can be used in both social research and science research. Data is an important strategic resource like natural resource and human resource[1][2]. Medical industry produces large amounts of medical data every year, the general medical institutions produce 1 TB - 20 TB data every year and the data volume of some big hospitals have even reached 300 TB - 1PB every year .

In China, due to lack of high-quality medical resources, high working intensity of doctors and inadequate participation of patients during treatment, there are many doctor-patient conflicts. In order to solve the related problems, a lot of online hospitals appeared. Due to lack of detailed diagnostic data, the diagnoses of online doctors often biase, and there are some people with ulterior motives giving the wrong diagnoses sometimes.

*This work is supported by National Natural Science Foundation of China under Grant 61463053 to X. Luo, and also supported by Nature Science Foundation of Hubei Province under Grant 2015CFC839 to Y. Sheng.

With the advent of the era of big data, medical science and IT(Information Technology) is connected more and more closely. The healthcare data contains a lot of information, and the medical intelligence can be implemented through the data mining , that is to realize the facilitation when seeking medical treatment, automation of disease diagnoses and informatization of health care.

There are some related literatures about how medical data being used. In [3], it is explained how big data influences the medical profession in the United States from several aspects, and it is also pointed out that the impact of big data on the health care industry is still at the initial stage and most of the potential is still unclaimed. The paper [4] analyses how big data offer new solutions to the medical problems such as disease diagnoses and epidemic predictions, and how big data change the traditional diagnoses methods. In [5], it is pointed out that understanding how proposed medical devices will interface with humans is a major challenge that impacts design, approval and regulation of innovative new devices, and the limitations can be overcome through advancements in data-driven, simulation-based medical device design and manufacturing. In [6], the authors present a comprehensive method for rapidly processing, storing, retrieving, and analyzing big healthcare data. Based on NoSQL (not only SQL), a patient-driven data architecture is suggested to enable the rapid storing and flexible expansion of data. The paper [7] describes the initial work for Bigdata processing framework for MCPS (Medical Cyber Physical Systems) that combines the real world and Cyber world with dynamic provisioning and fully elastic system for decision making in health care application.

Current researches mostly discuss how the big data plays a role in medicine and can change the existing model of medical care. And most works only concern data storage, transmission and access, lacking specific technical discussions and theoretical researches about what can be obtained from the relationships in data. This paper presents an automatic disease diagnoses method based on big data. The basic idea is: Based on a large amount of medical data, use the method of mathematical statistics to discover the relationships between the symptoms and diseases in a large number of confirmed diagnoses cases. Implement the diagnoses automatically according to the relationships based on the provided data of patients.

978-1-4673-8611-1/15/$31.00 ©2015 IEEE

II. The Automatic Disease Diagnosis System

First, give the definitions and symbols being used in this paper:

Definition 1, dominant symptom: If there is an obvious relationship between symptom A and disease B, that is while the disease is diagnosed to be B and the symptoms of B always include symptom A, symptom A is a dominant symptom of disease B.

Definition 2, the confirmed symptom set: The dominant symptoms of disease B form the confirmed symptom set of B.

Definition 3, the correlation between the symptom A and disease B: The probability the symptom A appears in the disease cases of B, marked as $C(A,B)$. If there are K cases of disease B, and in the K cases there are M cases having symptom A, the value of $C(A,B)$ is M/K.

Let $Z_i, i = 1,2,3,...,I$ denote the symptoms. Let $B_j, j = 1,2,3,...,J$ denote the diseases. Let $Q(B_j) = B_j \sim \{Z_j^{(k)} | k = 1,2,3,...,K\}$ denote the confirmed symptom set of B_j. Let G be the symptom set which contains the symptoms the user provided. Let $Y(Q(B_j))$ be the number of elements in the confirmed symptom set of B_j and $Y(G)$ be the number of elements in G.

The automatic disease diagnosis method in this paper mainly consists of two parts; the first is how to get the "confirmed symptom sets", and then the specific diagnosis method.

Algorithm 1, obtaining the confirmed symptom sets:

Step1, compile all the symptoms and diseases in the cases. Get the symptom set $\{Z_i | i = 1,2,3,...,I\}$ and the disease set $\{B_j | j = 1,2,3,...,J\}$.

Step2, give the threshold $\beta(\beta \geq 0.6)$, and let $j = 1$.

Step3, calculate $C(Z_i, B_j)$ between Z_i and B_j, $i = 1,2,3,...,I$. If the value of $C(Z_i, B_j)$ is larger than β, put Z_i into the confirmed symptom set of B_j.

Step4, $j = j + 1$, if $j \geq N$, step out, or else, return to Step3.

In Algorithm 1, the confirmed symptom sets $Q(B_j)$, $j = 1,2,3,...,J$ are obtained. Base on the results, the automatic diagnosis algorithm is as follow:

Algorithm 2, automatic diagnosis:

Step1, in the system, list the symptoms according to the body parts where the diseases exist, including skin, head, neck, chest, waist, chest, stomach, back, arms and legs.

Step2, calculate $|Y(Q(B_j)) - Y(G)|$, $j = 1,2,3,...,J$. Order the values of $|Y(Q(B_j)) - Y(G)|, j = 1,2,3,...,J$ from the smallest to the largest. If the smallest value $|Y(Q(B_s) - Y(G)|$(s is a value in 1,2,3,...,J) is 0, it is can be deduced the disease the user consults is B_s, or not, go to Step3.

Step3, list the corresponding diseases of the smallest three values in $\{|Y(Q(B_j)) - Y(G)|, j = 1,2,3,...,J\}$ as a reference.

The result outputted in Algorithm 2 is the diagnosis result.

III. The Automatic Disease Diagnosis System

In this part, an example is given to illustrate the diagnosis method implementation, and then the developed system is described.

Example: There are two diseases: acute nasopharyngitis and normal rhinitis, marked as B_1 and B_2.

Background: 1). In the available data, the symptoms associated with "acute nasopharyngitis" are: headache, dizziness, runny nose, stuffy nose, fever and weak limbs, marked as $B_1 - Z_1^{(1)}$, $B_1 - Z_1^{(2)}$,$B_1 - Z_1^{(3)}$,$B_1 - Z_1^{(4)}$,$B_1 - Z_1^{(5)}$,$B_1 - Z_1^{(6)}$. The frequency of the symptoms appear in the cases of "acute nasopharyngitis" are: 40%, 70%, 90%, 90%, 70%, 60% respectively, marked as $C(Z_1^{(1)}, B_1) = 0.4$,$C(Z_1^{(2)}, B_1) = 0.7$,$C(Z_1^{(3)}, B_1) = 0.9$,$C(Z_1^{(4)}, B_1) = 0.9$,$C(Z_1^{(5)}, B_1) = 0.7$,$C(Z_1^{(6)}, B_1) = 0.6$. Mark the symptom set related to B_1 as $B_1 - \{Z_1^{(k)} | j = 1,2,3,...,6\}$. In the available data, the symptoms associated with "normal rhinitis" are: rhinalgia, runny nose, stuffy nose, marked as $B_2 - Z_2^{(1)}$,$B_2 - Z_2^{(2)}$,$B_2 - Z_2^{(3)}$. The frequency of the symptoms appear in the cases of "normal rhinitis" are 40%, 50%, 90% respectively, marked as $C(Z_2^{(1)}, B_2) = 0.4$, $C(Z_2^{(2)}, B_2) = 0.5$, $C(Z_2^{(3)}, B_2) = 0.9$. Mark the symptom set related to B_2 as $B_2 - \{Z_2^{(k)} | k = 1,2,3\}$. 2). The symptoms provided by the user are: dizziness, runny nose, stuffy nose, fever and weak limbs.

Get the dominant symptom sets: give the threshold $\beta = 0.6$. As $C(Z_1^{(1)}, B_1) < \beta$, $C(Z_1^{(2)}, B_1) > \beta$, $C(Z_1^{(3)}, B_1) > \beta$, $C(Z_1^{(4)}, B_1) > \beta$, $C(Z_1^{(5)}, B_1) > \beta$, $C(Z_1^{(6)}, B_1) >= \beta$, the confirmed symptom set of "acute nasopharyngitis" is $B_1 - \{Z_1^{(k)} | k = 2,3,4,5,6\}$that is {dizziness, runny nose, stuffy nose, fever and weak limbs}. As $C(Z_2^{(1)}, B_2) < \beta$, $C(Z_2^{(2)}, B_2) > \beta$, $C(Z_2^{(3)}, B_2) > \beta$,the confirmed symptom set of "normal rhinitis" is $B_2 - \{Z_2^{(k)} | k = 2,3\}$, that is {runny nose, stuffy nose}.

Give the diagnosis: G={dizziness, runny nose, stuffy nose, fever and weak limbs}. There are 5 symptoms both in the symptoms the user provided and $B_1 - \{Z_1^{(k)} | k = 2,3,4,5,6\}$, and there are 2 symptoms both in the symptoms the user provided and $B_2 - \{Z_2^{(k)} | k = 2,3\}$. As $|Y(Q(B_1)) - Y(G)| = 0$ and $|Y(Q(B_2)) - Y(G)| = 3$. It is can be deduced that the disease the user consults is "acute nasopharyngitis" according to Algorithm 2.

System development: Build the system platform in accordance with the above method. Fig.1 and Fig.2 show the interfaces which is provided for the users to choose the body parts where the symptoms happen. Fig.3 shows the interface which is provided for the users to choose the symptoms which they have. Fig.4 is the interface which provides the diagnoses. In the system implement, the experimental data is provided by affiliated hospital of zunyi medical university. At present, the test result in our experiments is good, and the diagnoses accuracy of normal diseases can approximate to 80%.

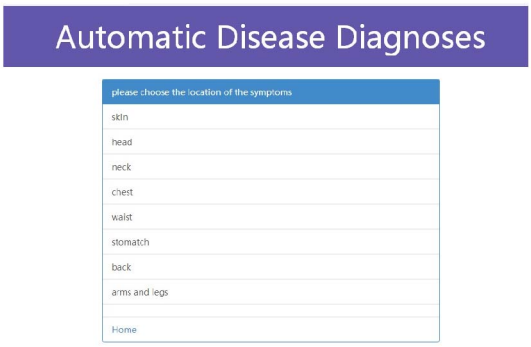

Fig. 1. Choosing the body part where the symptoms happen(1)

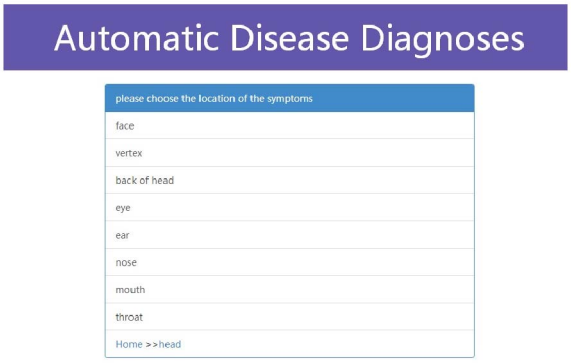

Fig. 2. Choosing the body part where the symptoms happen(2)

IV. Conclusion

The diagnoses automation based on big medical data analysis has important significance to doctor-patient contradiction regulation and the optimizing allocation of health resources. This paper puts forward a disease diagnosis system based on the method of mathematical statistics. The system has been implemented preliminary. The later work is enriching and improving the system, and applying it into clinical practice further.

References

[1] H. Wu, "Big Data is Like Oil," *CHUANGXINKEJI*, vol. 11, no. 7, pp. 6-7, December. 2013.

[2] G. Li, "The Scientific Value of the Study on Big Data," *Communications of the China Computer Federation*, vol. 8, no. 9, pp. 8-15, September. 2012.

[3] B. Kayyali, D. Knott and S. Kuiken, "The Big-data Revolution in US Health Care: Accelerating Value and Innovation," [Online]. Available: http://www.mckinsey.com/insights/health systems and services/the big-data revolution in us health care, [Accessed: Jun. 31, 2015].

[4] B. Cowan, "Big Data Medical Imaging," [Online]. Available: http://www.nihi.auckland.ac.nz/sites/nihi.auckland.ac.nz/files/pdf/informatics/bigdata,[Accessed: Jun. 31, 2015].

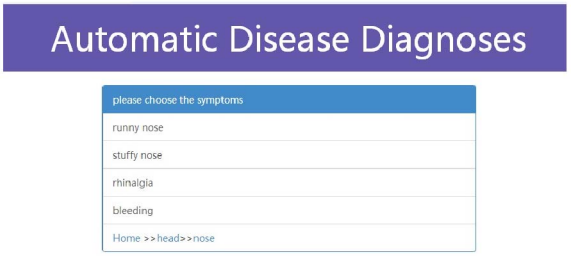

Fig. 3. Choosing the symptoms

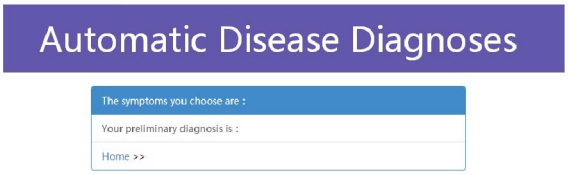

Fig. 4. The interface providing the diagnosis

[5] A. Erdman, D. Keefe and R. Schiestl, "Grand Challenge: Applying Regulatory Science and Big Data to Improve Medical Device Innovation," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 3, pp. 700–706, March. 2013.

[6] C. Lin, L. Huang, S. Chou, C. Liu, H. Cheng, I. Chiang, "Temporal Event Tracing on Big Healthcare Data Analytics," in *Proceedings of 2014 IEEE International Congress on Big Data*, June. 2014, pp. 281–287.

[7] S. Don, D. Min, "Medical Cyber Physical Systems and Bigdata Platforms," in *The Fourth Workshop on Medical Cyber-Physical Systems*, April, 2013.