Research on the Application of Statistical Technology in Biomedical Information Data Processing Based on Data Science

Xinyi Yang

University of San Francisco, San Francisco, USA

23130002466@qq.com

Abstract. This paper shows various methods in the field of biomedical research. The correlation between data exploration and statistics in biomedicine. This paper describes the process of data acquisition in the field of biomedicine, and summarizes several common algorithms in data acquisition. Then a method based on fuzzy clustering is proposed for biomedical data analysis. Compared with the traditional Bayesian model, this model has higher efficiency.

Keywords: Data Statistics, Biomedical Science, Data Processing, Rough Set Model.

1. Introduction

This is through the analysis of massive observational data to find previously unknown relationships and rules. The vast amount of data accumulated by various industries is also increasing. However, with the rapid development of massive data, if only rely on conventional analysis means, it can often only get the surface information, but can not dig out the essential correlation and potential knowledge [1]. Contemporary data mining technology is gradually taking shape. Some scholars have proposed a new data acquisition technology that is both safe and feasible. Data mining is a kind of information technology that extracts reliable and efficient data from a large number of data, aiming to discover new, uncertain and valuable knowledge from a large number of random data to improve the effectiveness of data. Among them, correlation analysis, classification analysis, cluster analysis, statistical analysis, time series model, deviation analysis and so on are one of the most commonly used research methods. The traditional data mining process includes business knowledge, understanding, preparation, modeling analysis, evaluation and so on.

2. Medical data analysis based on deep learning

2.1. Identification of pathogenic microorganisms

Disease diagnosis is one of the major applications of deep learning in medicine. This method is based on the patient's clinical data, using deep neural network to analyze the patient's lesions and risk factors to achieve the auxiliary diagnosis of the patient's condition. Automated disease diagnosis can speed up the processing of data, can provide a basis for doctors, and its judgment is not easy to be affected by subjective factors. In addition to reducing the burden of doctors, it can also improve the efficiency of work and the accuracy of diagnosis [2]. The automatic diagnosis of diseases mainly includes the diagnosis of diseases, the classification of diseases, and the grade of diseases. Some studies have combined CNN and recurrent neural network to classify nuclear cataract according to ophthalmic diagnostic images. There have been studies using SAE model to build brain MRI image and apply it to clinical diagnosis. The results show that the correct recognition rate reaches 97%.

2.2. Medical image processing

Hospitals produce a large number of images, images usually contain a wealth of hidden information. Most of the existing methods rely on manual interpretation and analysis, which are not only inefficient, but also get little information. It does not maximize the role of the data. The high precision image analysis capability based on deep learning will bring new ideas for automatic analysis

of medical images. At present, medical image research based on deep neural network is mainly used in clinical image classification, key object recognition, automatic image segmentation and so on. Some studies have used algorithms based on deep learning to recognize cancer cells in images. The recognition accuracy is improved by about 7% compared with the conventional algorithm [3]. It has important application value to realize the automatic identification of cancer. There have been studies on the use of ultrasonic signals to detect the left endocardium, and the automatic processing of ultrasonic signals has obtained good results. Some researchers used CNN to perform an automated search of breast tumor images, and the results were more accurate than any previous algorithm, and won first place in that year's ICPR competition. Some researchers have used StackedISA model to automatically segment MRI images of the prostate. The model can automatically segment the prostate from MRI images, and use the abstract features obtained by deep learning methods to replace the artificially designed features in the past. The image segmentation accuracy is improved significantly. 3D imaging and automatic segmentation of knee cartilage using 3D CNN have been studied. The result is better than modeling based on 3D images. Researchers have used this network to extract plasmic nuclei from images of the uterine neck. The correct rate was 91.34%.

2.3. Medical data modeling

Modeling medical data using deep neural networks. Compared with the diagnosis of diseases, modeling problems are more difficult, but also more valuable, because it involves the target of complex systems, so a good modeling method has a broader application prospect. The mathematical modeling of the course of disease development contributes to the analysis and monitoring of related diseases and contributes to the prevention and treatment of diseases [4]. Previous studies have used DBN to simulate human emotion and conventional human emotion data to predict emotion, and the results have similar accuracy to the shallow level emotion model based on experience. Some researchers proposed the DL-PRO method based on SAE, and established the protein spatial structure prediction method, and obtained good experimental results.

3. Rough set model extension

3.1. Extended definition of upper and lower approximation

R is the relation $[x]_R$, Pawlak's element-based definition is obtained on the approximate space (U,R). When R is equivalence $[x]_R$, there is a grain-based definition on the approximate space $(U,\sigma(U/R))$. On the approximate space $(U,\sigma(U/R))$ there are sub-system-based definitions.

Grain based definition:

$$\overline{apr}(X)=\bigcup\{[x]_R\,|\,[x]_R\in\sigma(U/R),[x]_R\cap X\neq\phi\},$$

$$\underline{apr}(X)=\bigcup\{[x]_R\,|\,[x]_R\in\sigma(U/R),[x]_R\subseteq X\}.$$

Based on subsystem definition:

$$\overline{apr}(X)=\bigcap\{Y\,|\,Y\in\sigma(U/R),X\subseteq Y\},$$

$$\underline{apr}(X)=\bigcup\{Y\,|\,Y\in\sigma(U/R),Y\subseteq X\}.$$

The above three definitions describe the equivalence relation from the perspectives of element, particle and subsystem respectively [5]. These three equivalent definitions give different expressions set theory. In the element-based definition, if all the equivalent elements of an element x are in the set X, then x is lower $\underline{apr}(X)$ of X; If there is at least one equivalent element of x in X, then this element x is upper $\overline{apr}(X)$ of X. In the grain-based definition, equivalence classes contained in X forms the lower $\underline{apr}(X)$, and equivalence classes whose intersection with X is not empty forms the upper approximation $\overline{apr}(X)$.

3.2. Element-based extension model

Let $R \subseteq U \times U$ are the domain of discourse, then it defines an extended approximation space $apr = (U, R)$. The equivalence class $[x]_R$ of element x is regarded as a neighborhood of x, the neighborhood is obtained. Neighborhood relation $R_s(x)$ only requires reflexivity, not necessarily symmetry or transitivity.

The $R_s(x)$ is special case when A is an equivalence relation in the neighborhood model.

4. Medical data cleaning

4.1. Data cleaning framework

Data cleansing is a very important step. Due to the different types of data have their own needs, there is no unified standard for data cleaning. The concrete methods from domain to domain, subject to subject and project to project are put forward. However, data cleaning is generally the processing of unreasonable, incomplete and inaccurate data. The quality of data is improved by automatic identification of missing features and automatic completion of missing features in the data cleaning stage [6]. Figure 1 shows an architecture for cleaning data (image cited in an automated data cleaning method for Electronic Health Records by incorporating clinical) knowledge). This framework does not include algorithms for data exploration and provides guidelines for data exploration research. The corresponding data clearing method is proposed. This architecture allows the user to choose the appropriate algorithm architecture, so it is more efficient for noisy data. The basic principles of the data mining framework are as follows:

(1) A method of selecting appropriate features from source data using pre-processing techniques; (2) Adopt the appropriate clustering algorithm or diversity method according to the characteristics of the cluster; (3) clustering carry out similarity analysis of relevant information to obtain useful information.

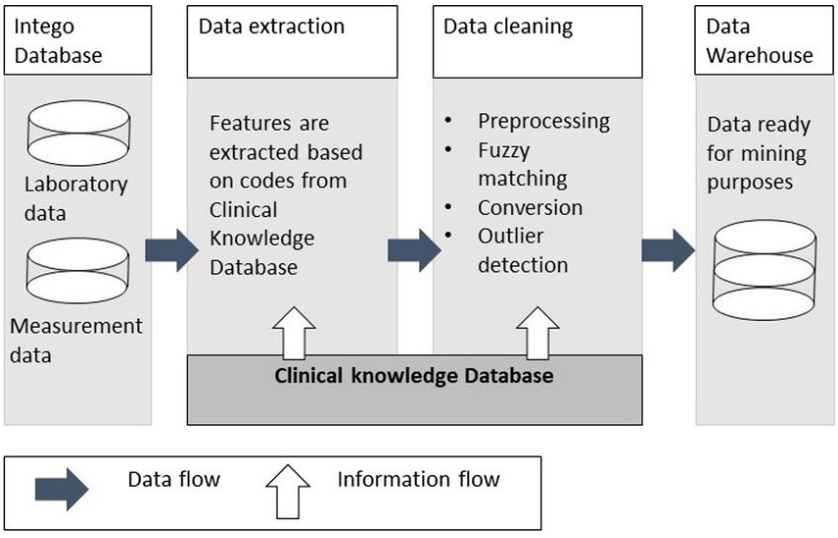


Figure 1. Data cleansing framework

4.2. Data Clearing

Inconsistencies and errors found during data collation can be corrected manually or automatically to improve data quality. While there are several commercially available tools that provide some capabilities for data cleansing, these capabilities are not perfect. Some scholars have given a complete data cleaning architecture. Data cleansing is mainly to provide an effective method for data warehousing or data mining technology. Some patterns or maps can be drawn from the data set. At the same time, an unsupervised step by step learning method is proposed to automate the data. This

algorithm can realize the automatic collection of cluster data [7]. The clustering method aims to produce several subgroups so that the data within the same subgroup are as close as possible. Instead of trying to differentiate the data within the same cluster.

4.3. A step-by-step undirected learning algorithm

In this unsupervised learning process, there is no external teacher or critic to monitor the learning process (Figure 2 cited in What Are Graph Neural Networks). Unsupervised machine learning falls into two categories: reinforcement and competitive. Reinforcement learning is evaluation-based, based on improving external knowledge. Competitive learning refers to a competitive learning process that can obtain information from a large amount of information. The author improves the Hebbian criterion based on the traditional unsupervised method and introduces a data-based competitive learning method.

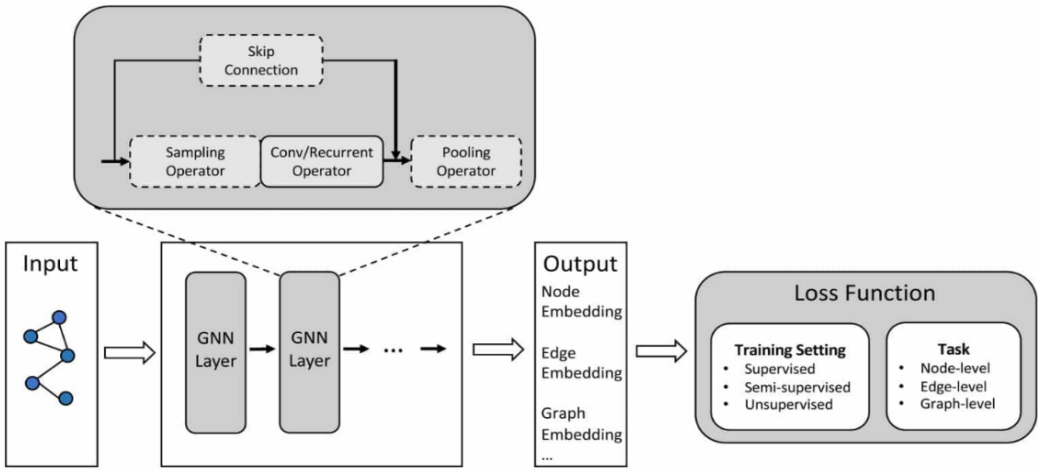


Figure 2. Undirected learning based on block graphs

5. System inspection

The accuracy actual data. Firstly, the proposed algorithm is compared with the Bayesian network model. For each generated CPD, the initial statistics can be normalized to obtain the corresponding statistics. Firstly, the original data is copied and modified in each experiment, and a subset of the original data is constructed. Secondly, the error term is added to the subset, and the missing term is obtained by eliminating the useful term. Finally, it is transformed into Markov carpet model. In the experiment on the corrupted data set, a rough cleaning method was used to construct the patterns of the two algorithms for the sake of fairness [8]. The accuracy and quality shown in Figure 3 and Figure 4 include two different algorithms. The method has strong destruction resistance. The reason for this is mainly because from start to finish, the method is combined with Markov carpets. At present, ANOVA method based on Rough sets has better ANOVA effect than ERACER method, but due to the large differences among samples, there is no conclusion at present. The algorithm introduced in this paper is basically the same as the rough set algorithm at the bottom of the damaged data, but it is more sensitive to the overall error. Each side is independent and depends on how the majority of cases are voted. If most of the correlation data is error, this error may be transmitted to the Markov carpet and the result deviates from the original prediction. However, the method presented in this study is very useful for many programs that are easy to operate and have fewer bugs. The accuracy of the proposed algorithm is 95%, which is superior to the Bayesian network model. The method described in this paper is superior to the Bayesian method in data cleaning.

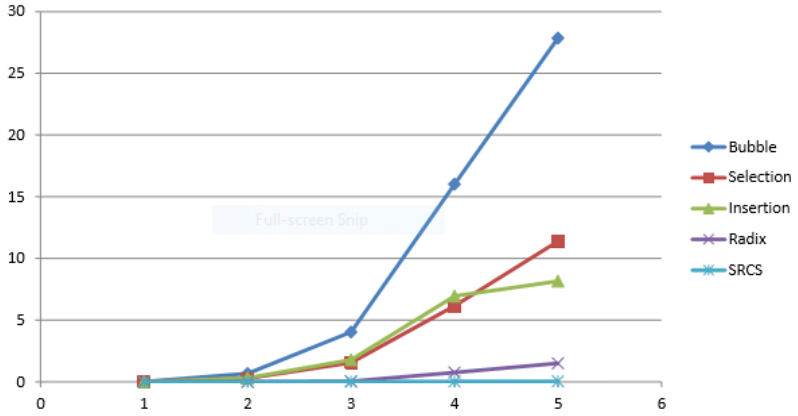


Figure 3. Comparison between ERACER algorithm and rough set algorithm

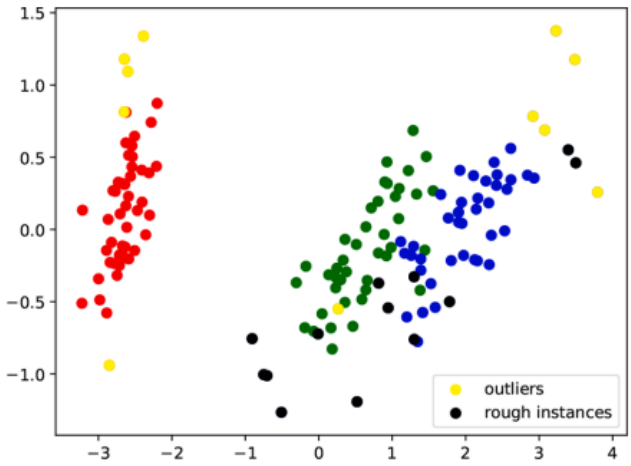


Figure 4. Experimental results

6. Conclusion

In biomedical big data, the structure and format of the data are also different due to its different sources. The same medical record, under the interpretation of different doctors, may have different results. In addition, due to the consideration of patient privacy, many individual information is deleted. The massive biomedical information effectively, it is necessary to clean and integrate it effectively, so as to obtain accurate, regular, structured and high-quality medical information. The experimental scheme adopted in the experiment is obviously better than the simple Bayesian network mode.

References

[1] Zheng Guangyong, ZENG Tao, LI Yixue. Application and prospect of cutting-edge information technology in biomedical big data. Genetics, Vol. 43(2021) No.10, p.924-929.

[2] Huang Xiuzhen, Zhou Tanping, LI Ningbo. Safe location scheme of pathogenic gene based on cryptography. Beijing Biomedical Engineering, Vol. 40(2021) No.2, p.107-112.

[3] Shao Tanbin, Yang Kecheng, Xia Min, et al. Spectral confocal microscopy imaging techniques and applications. Advances in Laser and Optoelectronics, Vol. 60(2023) No.12, p.15-22.

[4] CAI Xin-Tian, Zhang De-Lian, Bi Yun-Wei, et al. Analysis of key genes of renal damage in ANCA-associated vasculitis based on bioinformatics. Journal of Medical Research, Vol. 50(2021) No.1, p.54-59.

[5] Zeng Yu, Yang Xiaoyan, Zhang Wei. Application status of long short-term memory network in medicine, health and biomedicine. West China Medical Journal, Vol. 36(2021) No.1, p.66-72.

[6] Wang Wei, Cao Yinan, Liu Jiali, et al. Analysis of regulatory networks and related mechanisms in Alzheimer's disease based on multiple bioinformatics. Journal of Nanchang University: Medical Edition, Vol. 62(2022) No.3, p.100-129.

[7] Chen Cong, NI Leyi, Chen Yumei. Bioinformatics analysis of immune-related genes in ovarian endometrioid carcinoma. Zhejiang Med, Vol. 43(2021) No.4, p.77-82.

[8] Liang Liming, LAN Zhimin, Sheng Xiaoqi, et al. Fusing contour information and conditions to generate antagonistic retinal vessel segmentation. Journal of Biomedical Engineering, Vol. 38(2021) No.2, p.10-19.