International Journal of Biology and Life Sciences
ISSN: 2957-9511 | Vol. 3, No. 3, 2023

Study on Key Technologies of Cancer Surveillance and Early Warning based on Data Mining

Zijun Wu

School of Physical Science, University of Chicago, Chicago, 60637, USA

Abstract: Early detection and accurate judgment are crucial during cancer diagnosis and treatment. Hence, it is of great significance to establish an effective cancer early warning system. However, the arrival of big data provides new opportunities for cancer surveillance and early warning, and data mining technology, as an important data processing method, is being widely used in the field of cancer surveillance and early warning. By analyzing a large number of medical images and medical records, the author identifies some potential risk factors which can improve the early diagnosis rate and treatment effect. Moreover, this paper compares the current status and trends of research on data mining, analyzes the key technologies of cancer surveillance and early warning based on data mining, and summarizes the related research results with the evaluation of each method.

Keywords: Data Mining; Cancer; Surveillance; Early Warning; Algorithm.

1. Introduction

As a common chronic disease, the high morbidity and mortality of cancer make it one of the important issues that need to be addressed by the current medical community. Hence, how to effectively carry out cancer surveillance and early warning has become an urgent problem to be solved. Currently, the traditional cancer surveillance methods mainly rely on doctors' physical examination and various auxiliary detection means, such as X-ray and CT scans. [1] However, although these methods can detect some obvious symptoms or abnormalities, it is difficult to diagnose cancer in its early stages in time. In addition, as doctors partly have limited expertise and time pressure, traditional methods may inevitably have missed or misdiagnosed cases. In this situation, the application of data mining technology has great potential. By analyzing a large number of medical records, some regular characteristic values can be extracted from them, thus achieving accurate prediction and early warning of cancer.[2] Meanwhile, data mining can also help us better understand the knowledge of tumor cell growth mechanisms and treatment effects. In conclusion, this paper will focus on the research status and development trend of key technologies for cancer surveillance and early warning based on data mining and put forward corresponding solutions and suggestions to make some contributions to cancer prevention and treatment.

2. Current Status of Research on Key Technologies for Cancer Surveillance and Early Warning based on Data Mining

2.1. Data Mining Technology

Data mining is the analysis of observed data sets (often very large) to discover unknown relationships and summarize the data in novel ways that are understandable and valuable to the data owner. That is, data mining refers to the extraction or mining of knowledge from large amounts of data, which is also called knowledge discovery in data. However, data mining is an interdisciplinary technique that covers statistics, database technology, machine learning, pattern recognition, artificial intelligence, visualization techniques, etc. The relationships and abstractions derived through the data mining process are called models or patterns, such as linear equations, rules, clusters, graphs, tree structures, and cyclic patterns expressed as time series.

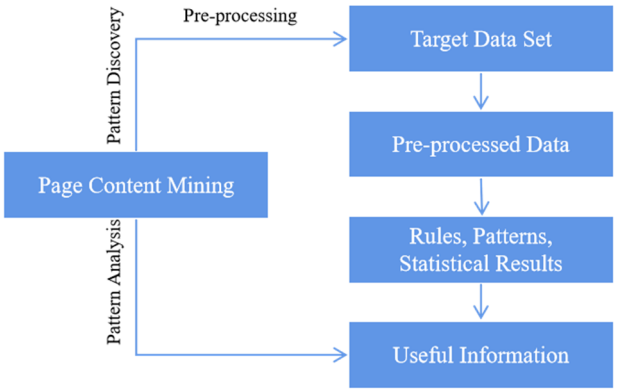

Figure 1. Data mining technology [3]

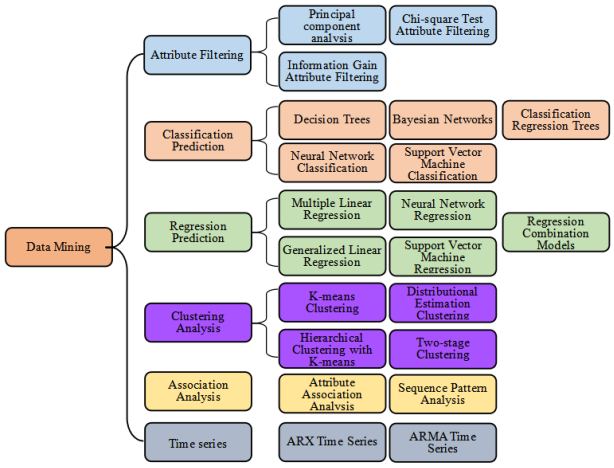

Figure 2. Common data mining technology methods [4]

2.2. Advantages and Prospects of Data Mining Technology Application in Cancer Surveillance

2.2.1. Limitations of Traditional Methods

Traditional cancer surveillance and early warning methods mainly rely on the experience and judgment of doctors or specialists, but this method has many potential problems. Firstly, as doctors or specialists partly have limited professional knowledge, it is difficult to understand all possible cases comprehensively; secondly, as doctors or specialists have limited time and energy, they inevitably cannot discover new cases in time, especially for acute diseases and cancers; finally, as doctors or specialists are under pressure at work and other factors, it could potentially lead to imprecise judgments. Hence, all these factors cause traditional cancer surveillance and early warning methods have great shortcomings.

2.2.2. Application Advantages of Data Mining Technology

With the development of medical technology, big data has become an important part of modern medicine. Moreover, data mining has been widely used in the medical field as an important data analysis method. By deeply mining and analyzing a large amount of medical data, some potential disease risk factors can be discovered, thus providing clinicians with more accurate diagnoses and treatment plans. In addition, many scholars have also started to explore how data mining technology can be applied to the medical field. In addition to being used for disease prediction, data mining can also help clinicians better understand the changing course of a patient's condition. For example, the possibility of certain diseases can be inferred by deep mining patients' medical records to identify abnormal values or pattern changes in their body index or symptoms, etc.

2.2.3. Typical Examples of Data Mining Technology in Cancer Surveillance and Early Warning

Currently, a series of data mining projects have been carried out to explore the regularities and trends in medical data. For example, the National Institutes of Health (NIH) in the United States has conducted extensive data mining efforts using large-scale genomics databases and has identified many new genetic variants associated with risk factors for a variety of diseases. Moreover, in the field of cancer detection, research institutions have started to adopt data mining technology to assist in cancer surveillance and early warning efforts. For example, the Cancer Surveillance Registry (CSR) at the University of California, Los Angeles, is one of the typical data mining systems. The system's core functions include: Collecting, storing, and managing a variety of information about cancer; analyzing and modeling this information using statistical methods; and generating reports for clinical practitioners to refer to. In addition, there are other data mining systems such as NCI-SEER, ACS Cancer Survivorship Study, and so on. All of these systems can provide strong support for cancer surveillance and early warning. [5]

2.2.4. Prospects of Data Mining Technology in Cancer Surveillance and Early Warning

Currently, many technical means based on data mining have emerged for cancer surveillance and early warning. Among them, the most common is the machine learning-based method. This method can train models to automatically identify tumor cells and other abnormalities and predict whether a patient has a certain disease. Moreover, deep learning-based methods are widely used, such as convolutional neural networks (CNN) and recurrent neural networks (RNN), etc. [6] These methods are capable of learning feature representations from large amounts of data, which in turn achieves accurate classification and detection of tumor histomorphology, imaging presentation, and other aspects. In addition to traditional machine learning methods, other emerging data mining techniques have been introduced into the field of cancer surveillance and early warning. For example, text mining-based methods can be used to extract keywords and semantic relationships in patient medical records to assist doctors in making more accurate diagnoses; social network-based methods can find people with similar symptoms or disease history within a community to better understand disease trends.

3. Analysis of Key Technologies for Cancer Surveillance based on Data Mining

The key nodes in the processing of data mining technologies are as follows:

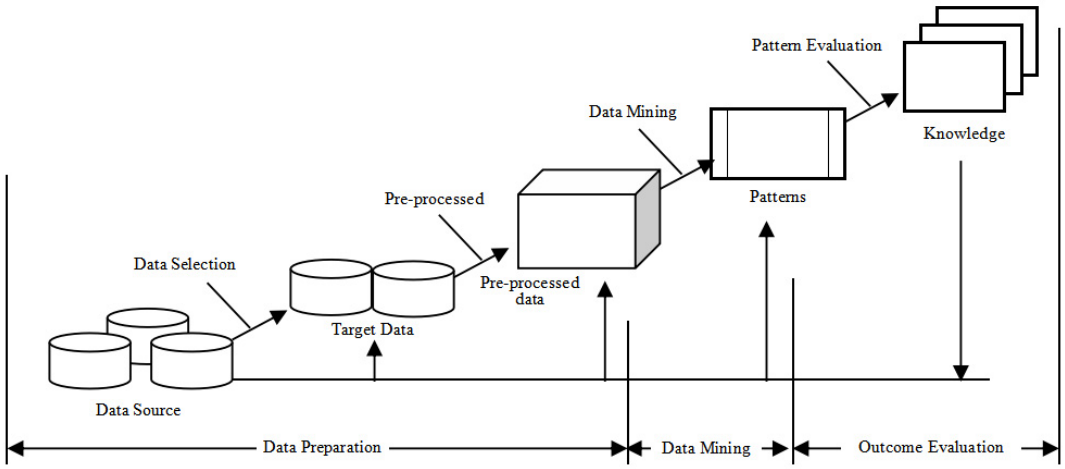

Figure 3. Processing process of data mining technology [7]

3.1. Data Preparation

During cancer surveillance and early warning, the acquisition of data is very important. Moreover, adequate preparation of data is required to achieve effective data

processing and analysis. Specifically, data preparation mainly includes the following aspects: Firstly, the type and source of data to be collected should be clarified. Different types of data have different roles and meanings for cancer surveillance and early warning. For example, clinical data can provide information about the basic conditions and treatment plans of patients, while genomic data can reveal the genetic characteristics of tumor cells by detecting the DNA sequences in them. Secondly, the quality of the data needs to be considered. Since cancer surveillance and early warning involve a large amount of data, the quality control of the data is particularly important. Finally, visualization and interactivity of the data need to be considered. By presenting the data in the form of charts or graphs, the trends and regularities of the data can be displayed more intuitively, thus better supporting the decision-making process. Hence, choosing the appropriate data preparation method for the practical application is necessary.

3.2. Data Preprocessing

Data preprocessing aims to convert the raw data into a data format suitable for subsequent analysis and modeling and to remove noise or abnormal values to improve the accuracy and reliability of the model. Moreover, the data needs to be cleaned and normalized during data preprocessing. Among them, cleaning refers to removing invalid or irrelevant data points, such as missing values, duplicate values, abnormal values, etc.; normalization transforms multiple variables into a standardized scale for further processing and analysis. In addition, feature extraction and dimensionality reduction operations on the data are required to reduce the data size and lower the computational complexity. Finally, the effectiveness and accuracy of data preprocessing are evaluated by cross-validation. For the characteristics of cancer data, common data preprocessing algorithms include principal component analysis (PCA), factor analysis (FA), and association rule learning (ARL). These algorithms can effectively discover commonalities and correlations in the data, thus helping us to understand better and predict the development process of tumors. In conclusion, data preprocessing is one of the important parts of cancer big data analysis, and its results directly affect the subsequent modeling and decision-making.

3.3. Main Methodologies of Data Mining

3.3.1. Association Rule Mining

Association rule mining is a common data mining algorithm whose main aim is to extract relevant features and patterns from a large amount of data for the aim of early diagnosis and prediction of diseases. The basic idea of association rule mining is to establish a knowledge base by finding commonalities and differences in a data set. Specifically, it can divide the dataset into several subsets and then use some known statistical models or inference rules to determine whether there is an association relationship between the data in each subset. If there is an obvious association relationship between the data in a subset, then this subset can be regarded as a meaningful knowledge unit and can be used for further analysis and modeling. Moreover, in practical applications, association rule mining usually requires the use of specially designed tools and software programs. Some of the more common ones are ApacheMahout, Weka, RapidMiner, etc. These tools all provide a series of functional modules and operating interfaces, enabling researchers to quickly set up relevant experimental platforms and start data mining work in a relatively short time. For association rule mining in the cancer field, the most common method is the support vector machine (SVM) based association rule mining algorithm. The core idea of the algorithm is to create a high-dimensional spatial representation of the training samples and then use an SVM classifier to determine the distance between the points. This method is effective in capturing potential associations in the data while avoiding the problem of overfitting.

3.3.2. Cluster Analysis

In the field of cancer science, cluster analysis is a common data mining algorithm. It can aggregate similar data points together to form a cluster or group, thus discovering potential anomalies and regular changes. Moreover, during cancer surveillance, cluster analysis can be used to classify and categorize a large number of medical images to improve diagnostic accuracy and efficiency. Specifically, cluster analysis aims to determine whether two samples belong to the same category by calculating the distance between them. Common clustering algorithms include the K-mean clustering method, hierarchical clustering method, central vector method, and other types. Among them, the K-mean clustering method is the most common and effective method. This method first calculates the Euclidean distance between each sample and all other samples, then selects the sample with the minimum distance to represent the new category and repeats the process until a predetermined number is reached. The clustering results can be presented graphically or visually to understand the data structure and characteristics better.

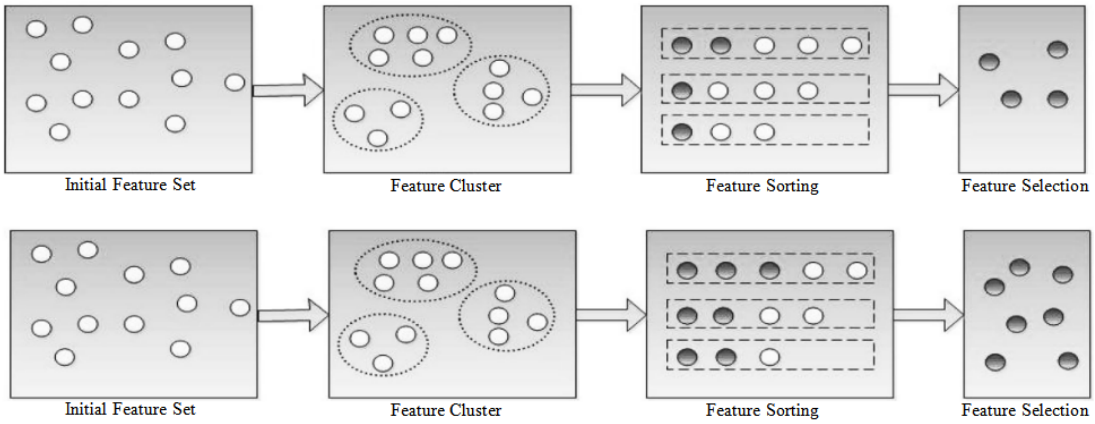

Figure 4. Cluster analysis [8]

3.3.3. Decision Tree

The decision tree is a commonly used machine learning algorithm that is widely used in the medical field for disease prediction and classification tasks. The decision tree is a non-parametric model that represents the relationship between input variables by constructing a series of nodes and branches. The basic idea is to perform feature selection and partitioning operations at each node and then pass the results to the next node for further processing. Finally, a complete tree structure is obtained to classify or predict new samples. In the context of cancer early warning, decision trees can effectively analyze a large number of medical records and extract useful information. For example, decision trees can be used to identify different types of cancer cells and their characteristics, thus providing doctors with more accurate diagnostic recommendations. In addition, decision trees can be used to predict whether a patient will develop cancer or other health problems.

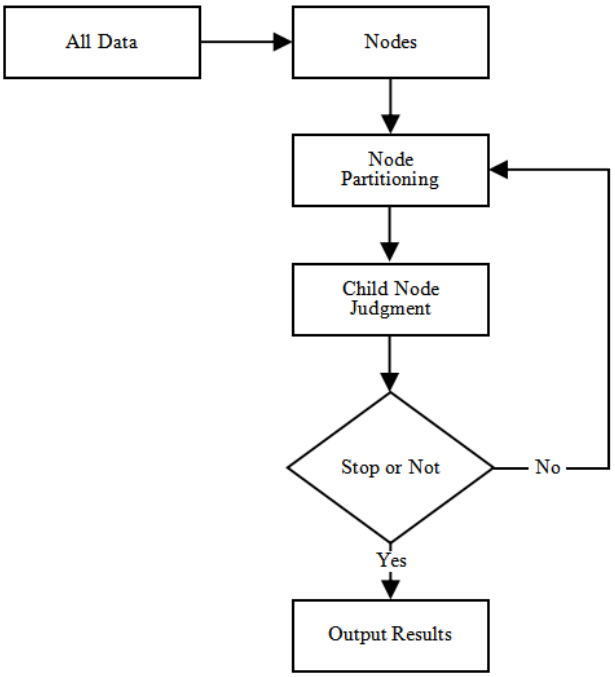

Figure 5. Decision tree steps

3.3.4. Neural Networks

Neural networks are an important tool in tumor diagnosis and treatment. Neural networks can learn feature representations from a large amount of medical image data to achieve recognition and classification of tumor tissue morphology. Moreover, neural networks are adaptive and robust, capable of processing complex nonlinear relationships and noise problems. In addition, neural networks can improve their performance by continuous learning. Hence, neural networks have become one of the important research directions in the field of cancer image analysis. Currently, the commonly used types of neural networks include convolutional neural networks (CNN), recurrent neural networks (RNN), and long and short-term memory models (LSTM). Among them, CNN is the most popular type of neural network, which can extract useful information and perform feature encoding by convolutional operations in multiple layers. RNN, on the other hand, is mainly used for time series data modeling tasks. LSTM, on the other hand, is a special RNN structure that can maintain state information on a long-time scale without losing short-term information. Furthermore, all these neural network types can be used to detect and classify tumor tissue morphology automatically, and some results have been achieved. In the future, with the development of deep learning and algorithm optimization, it is believed that the application area of neural networks will be more extensively.

3.3.5. Support Vector Machines

In the field of cancer science, a support vector machine (SVM) is a commonly used classification algorithm. It can classify samples into two classes and find an optimal hyperplane to distinguish the two classes. The basic idea of SVM is to find a maximum interval point, that is, to make the two classes of samples as far away from each other as possible by minimizing the distance. This method is applicable to nonlinear problems in high-dimensional space and has good generalization ability and robustness. In cancer prediction, SVM is widely used in image recognition, gene expression profiling, etc. Moreover, the selection of the SVM model needs to consider several factors, such as feature selection, parameter settings, and the size of the training set. For tumor image diagnosis, a convolutional neural network (CNN) is usually used as a preprocessing step, which is then converted to binary labels for SVM classification. In addition, SVM can be optimized and improved using deep learning methods. For example, recurrent neural networks (RNN) or long short-term memory networks (LSTM) are used to build multilayer structures to improve classification accuracy. In conclusion, SVM is a very useful data mining tool for the early detection of neoplastic diseases and treatment decision-making.

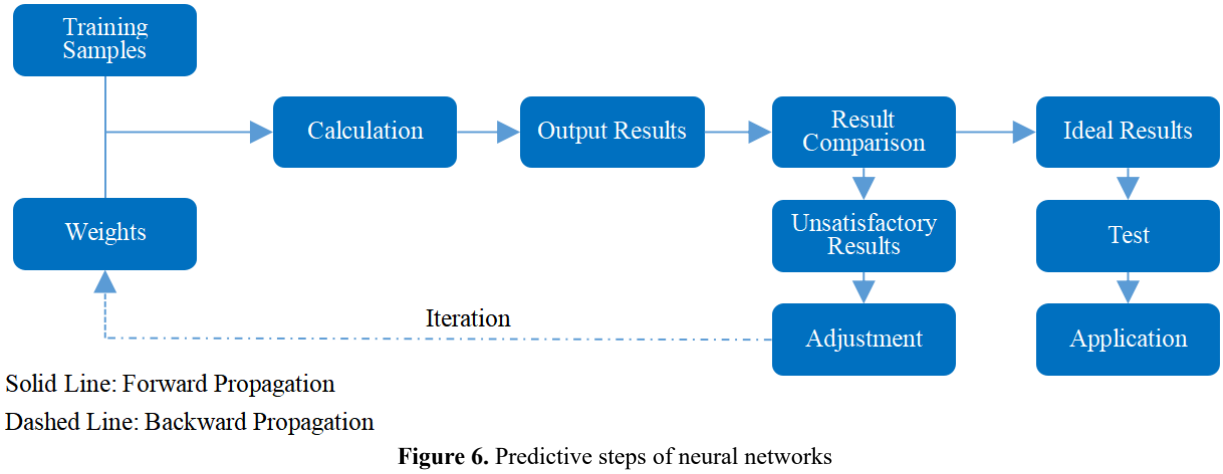

Figure 6. Predictive steps of neural networks

4. Conclusion

This paper mainly introduces the key technologies of cancer surveillance and early warning based on data mining and summarizes and evaluates the related research results. Moreover, through the study of existing literature, the author found that in the field of tumor diagnosis, traditional medical imaging methods can no longer meet the needs of clinicians; thus, advanced computer vision algorithms are needed to assist in disease detection and analysis. Meanwhile, with the advent of the era of big data, a large amount of medical image data has been collected, which can provide important support for tumor prediction. In this process, data mining is a very important technical means. It can extract useful information and features by processing and analyzing a large number of medical images, which can improve the accuracy and efficiency of diseases.

References

[1] Siqi Li,Diansen Chen,Wang Chen. Evaluation of Diagnostic Effect and Accuracy of Chest X-ray and CT in Patients with Pulmonary Tuberculosis[J]. Journal of Contemporary Medical Practice,2022,4(4).

[2] Rao Madhuri. Post-operative Lung Cancer Surveillance: The Highs and Lows of Computerized Tomographic Scanning. [J]. European journal of cardio-thoracic surgery: official journal of the European Association for Cardio-thoracic Surgery,2023.

[3] Liu J. From Statistics to Data Mining: A Brief Review [C]// Eliwise Academy. Proceedings of the 2020 International Conference on Computing and Data Science (CONF-CDS 2020).CONFERENCE PUBLISHING SERVICES, 2020:355-358.DOI: 10.26914/ c.cnkihy. 2020. 074580.

[4] Yang Fuli,Hou Xingzhe. Research on Smart Electric Meter Data Mining Technology Method for Line Loss Diagnosis of Low Voltage Station Area[P]. Proceedings of the 2019 International Conference on Precision Machining, Non-Traditional Machining and Intelligent Manufacturing (PNTIM 2019),2019.

[5] Czarnota Jenna,Gennings Chris,Colt Joanne S,De Roos Anneclaire J,Cerhan James R,Severson Richard K,Hartge Patricia,Ward Mary H,Wheeler David C. Analysis of Environmental Chemical Mixtures and Non-Hodgkin Lymphoma Risk in the NCI-SEER NHL Study.[J]. Environmental health perspectives,2015,123(10).

[6] Liu Jixin,Dai Pengcheng,Han Guang,Sun Ning. Combined CNN/RNN video privacy protection evaluation method for monitoring home scene violence[J]. Computers and Electrical Engineering,2023,106.

[7] Laouni Djafri. Dynamic Distributed and Parallel Machine Learning algorithms for big data mining processing[J]. Data Technologies and Applications,2022,56(4).

[8] Gao Y,Su F,Xiong J. Classification and study of glass based on cluster analysis[C]//Wuhan Zhicheng Times Cultural Development Co., Ltd..Proceedings of 2023 International Conference on Mathematical Modeling, Algorithm and Computer Simulation (MMACS 2023),2023:366-372.DOI: 10.26914/ c.cnkihy.2023.008062.