Data Correlation based Feature Selection Model for Children’s Growth and Development Assessment

Changhua Zhao*, ChunqiaoMi, Yuanquan Shi

School of Computer Science and Engineering, Huaihua University, Huaihua 418000, China

*lala_zhao@163.com

Abstract: The paper focuses on children’s data present in Electronic Health Records (EHRs) starting from raw data gathered from Southwest Medical Data Center in China. We designed a system based on workflow to analyze features that affect children’s growth. The system is able to relate health care data to diagnosis codes and with additional information integrated and correlated to EHRs data. Finally, we propose prediction models based on Recursive Feature Elimination (RFE), which can identify features that are important to detect children's growth and find correlations among features.

Keywords: Feature Selection, Children’s Assessment, Electronic Health Record

1. Introduction

Electronic health record (EHR) is a digital record of health information. It is a patient-centered real-time record, which enables human health information to be provided to authorized users immediately and safely. EHRs data are usually generated through the routine provision of clinical care, and are "big data" in terms of capacity (number of discrete data points available), speed (rate of new data accumulation) and diversity (numerous data elements available for inquiry) [1]. These aspects, together with its unique clinical relevance, make EHRs an ideal data for disease prediction and risk prediction [2]. Especially for children, the most important thing is to meet their needs. The features included in EHRs data are designed to describe the needs of healthcare providers who treat children by combining best practice clinical standards with federal information technology standards [3].

Pediatricians and other providers pay little attention to information about child health care [4]. Although children's EHRs in some departments have been summarized, including 547 functions in 26
subject areas, which are found to be valuable, they need to be further refined and prioritized [5]. There are regional differences in children's health care in Southwest China, both in terms of examination and feature selection. Southwest China is a region of the People's Republic of China, defined by government departments including Chongqing, Sichuan, Yunnan, Guizhou and Tibet Autonomous Region.

In this paper, we use the embedded method of logistic regression to select the features that affect children's health. And through the regional comparison, we find out which features are affected by the region. Finally, it provides reference for doctors.

2. Related Works
The relationship between children's growth and disease can be determined by EHR. It can compare the results of children health examination with the therapeutic effect. For example, EHRs data values for delivery date, visit date, gender, height and weight can be used to provide information about the patient's BMI or to determine the prevalence of obesity in children in a given population [6]. Clinical research is based on statistical analysis of a large number of clinical data sets, and it is usually necessary to extract and combine different EHR data. There are many data extraction technologies. For example, matching is used to search for known or common patterns (such as text strings, dictionary entries) in the search data corpus. Most of these methods only involve matching literal meaning, and lack of internal substantive connection. The use of computer-based methods to integrate clinical data and extract information has also been widely recognized, for example, the combination of proteomics and clinical data [7]. In EHRs, data is usually recorded in text. In the literature, there are many tools, such as Medex, that can extract drug information from clinical records [8]. In our project, the original data is in text format, so we developed a system that extracts information from EHRs and stores the data in the database.

Feature selection can improve the interpretability of the classifier, reduce the computational complexity and improve the prediction performance. However, the stability of the sample is also affected by the selection of the sample.

In our project, geographic information is mapped based on the home address domain provided by and included in SouthwestCh (Southwest Medical Data Center). We analyzed data from different sources, specifically EHRs, SouthwestCh, and geographic data. Our data has been consolidated and standardized to retrieve meaningful information from many different medical fields, increasing its value.

3. Data and Methods
We have EHRs data about the growth and development of children, which comes from Southwest of China where is a more relatively-poor area. We want to know what features will influence the growth and development of children, and we also hope through feature selection analysis lend well to predictive modeling and early risk stratification. There are 22 features in the original data, they are: CSSC, YTZ3, YSC3, YTZ6, YSC6, YTZ9, YSC9, YTZ12, YSC12, YYCS12, FYJQTJB, GLBYWYY, FT, WYQK, SL, TL, SG, ZY, FX, DDSTCD, ZJPJ, ZD. First of all, the data was cleaned. The irrelevant factors in the data and the number of rows with too much missing data were deleted. Then, the columns and rows with little missing data were averaged. After a series of processing of the data, we
begin to analyze the data and make corresponding graphs. Based on the trends in the graph, we analyze the influencing factors that affect children's growth and development and discuss corresponding solutions.

3.1. Dataset Cleaning
The uncorrelated features in the data will reduce the accuracy of many models, especially the logistic regression model. Before processing data sets, we should clean up the data. So we start testing each feature column in the dataset file to remove features from some invalid data. In our dataset, we have 22 features, after cleaning we get 16 features, they are: {"CSSC", "YSC12", "YYCS12", "YTZ3", "YSC3", "YTZ6", "YSC6", "YTZ9", "YSC9", "YTZ12", "TL", "FYJQTJB", "GLBY WYY", "KQBJ", "KXWY", "SZFW"}.

Starting from the SouthwestCh dataset, with all hospitalized health data, we selected only children and matched these results with the EHRs dataset. Supported by pediatricians, we applied Recursive Feature Elimination (RFE) and sparse logistic regression to identify which features are important to the growth and development of children and if there are correlations among children.

3.2. Feature Selection
EHRs data sets usually contain a lot of noise, diversity and redundancy, which affect the assessment of patients' health status. In order to better predict the results, it is necessary to obtain features highly related to health status. For child care, this can save budget and time and improve the accuracy of scientific feeding. Feature selection technology based on regression can help to realize feature selection. It sorts all the available features, and finally gets the feature with the highest correlation.

Feature selection based on RFE and L1 are two common feature ranking methods. We implemented both methods using the R library. They are used to select representative growth and development features of children.

RFE eliminates features recursively with the minimum weight until the required number of features is reached. Firstly, the estimation is trained on the initial data set, and the weight is assigned to each feature. Next, the feature with the minimum absolute weight value is removed; the process is repeated continuously to obtain the highly relevant features. The estimator we used was logistic regression. The estimator is logistic regression,

\[
h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}
\]

(1)

\[
h_\theta(x) = P(y = 1|x; \theta)
\]

(2)

\[
1 - h_\theta(x) = P(y = 0|x; \theta)
\]

(3)

where \( x \) is the input features, \( y \) is the binary class label. By using gradient descent, the training process of logistic regression is to minimize the loss function:

\[
f(\theta) = -\frac{1}{m} \left[ \sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log \left(1 - h_\theta(x^{(i)})\right) \right]
\]
Where $x^{(i)}$ is the input features, $y^{(i)} \in \{0,1\}$ is binary-valued labels.

In order to fit parameters $\theta$, we minimize $J(\theta)$ to get $\theta$.

Another feature selection method is based on L1 regularized features. In order to cause sparsity, we use L1 norm penalized classification model. It can make most of the estimated parameters (weights) be zero, only a few are non-zero. We use sparse logistic regression, which is a sparse induced L1 norm regularized logistic regression. The loss function (4) will change as following:

$$f(\theta) = J(\theta) + \lambda \| \theta \|_2$$

Where $\lambda$ is the regularization parameter controlling the sparsity of the model. The larger the $\lambda$ is, the greater the penalty for model parameter $\theta$ is, so that the more features are given with zero coefficient, the less features are selected. The selected features can be sorted according to the absolute value $\theta$ of their weights. Experiments show that different $\lambda$ can produce different top features. Therefore, we use a set of regularization parameters to calculate the average weight of each feature.

4. Experiment

EHRs data is extracted from EHRs and SouthwestCh Dataset. They are extracted from our collaborative company (Southwest Medical Data Center). The dataset includes geographical information of children, the data comes from east of China which uses Google maps Geocoding API. It contains information about the child's physical examination at each stage, such as height, weight, growth and development, and illness. There are 22 features to describe. In SouthwestCh Dataset it includes 10,000 children identified by an univouque code that is the same present in the EHRs dataset and allows us to correlate each child with the corresponding health analysis. Doctors use a unified diagnostic code in this dataset. In this system, we convert all medical diagnostic terms and child care terms into alphabetic or numeric codes.

From the Southwest China Medical Big Data Centers, take out the indicators of the growth and development of newborn babies in a certain year and divide these indicators into weight and height at the 3rd, height and weight at the 6th month, height and weight at the 9th month and weight at the 12th month. Then combine them within a year, with or without rickets, and the number of diarrhea, branch inflammation, within 3 months, through the child's growth environment, and feeding conditions (1. breast milk within 4 months 2. animal milk 3. mix feeding) for analysis.

We here focus on the use and analysis of child growth and development data in EHRs. We figured out the growth and development data collected by the medical big data centers in the southwestern region, analyzed the factors that affect the growth and development of children in southwestern China, and analyzed some of the growth conditions of children during the growth and development stages, such as diarrhea, branch inflammation, and rickets and other influencing factors. Through the analysis of various factors such as infant feeding, living environment, weather, and climate, we discuss and propose key solutions to affect the growth and development of the baby to ensure that the baby grows up as healthy and happy as possible.

Our system is applied to a real clinical scene, extracting data from anonymous EHRs, and then combining these data with the geographic location of Southwest China and patients. Firstly, the relevant features of children's health care are extracted. Then the obtained features are associated with
geographical features to prove that there is a significant correlation between some diseases and geographical location.

5. Results
According to the method discussed above, we rank the importance of each feature to children's growth and development. RSS, Cp, BIC, and adjusted Rsq are displayed for each possible model that contains a subset of 16 variables in the subset. The best model of black boundary tracking for a given number of variables.

![Different Features selection result in RSS, Cp, BIC, and adjusted Rsq](image)

**Fig. 1**

RSS, Cp, BIC, and adjusted Rsq are displayed for the best model for each size of the children's dataset (bottom border in Fig.1). Cp and BIC are the estimation of test MSE. After 10 features are selected, BIC estimation value of test error increases. After including 8 features, the changes of the other three graphs are quite small. This shows that there is almost no difference in accuracy between the 8 features and 10 features models. So we chose eight features. Because n>p means the features we use Backward algorithm will be better. The lowest BIC was 8 features model: YYCS12, YTZ3, YSC3, YSC6, YTZ9, TL, FYJQTJB, GLBYWYY, KQBJ, KXWY.

6. Conclusion
Health care institutions generate a lot of data. These original data need to be automatically stored, extracted and analyzed to obtain useful information. In this project, we designed a system to extract
features from EHRs, analyze children's health status and predict the relationship between children's health status and regions. The model we designed can well determine which health test features are important for the detection of children's growth and development. Whether there is correlation between children's growth and development characteristics. In addition, we found that there is a strong correlation between geographical regions and children's growth and development.

The future work includes the accuracy of the first eight features found by the model in Southwest China, and further mining out more regional features. Through this model, doctors are reminded to automatically grow and develop child recognition.

Acknowledgments
The authors are very thankful that this study is supported by the projects No. 19YJC880064 granted by the General program of Humanities and Social Sciences Research of the Ministry of Education, China, No. 19B447 granted by the scientific research project of the Education Department of Hunan Province, China, No. XSP20YBZ140 granted by the program of Hunan Provincial Social Science Achievements Evaluation Committee, China, No. 18C0980 granted by the program of Hunan Provincial Department of Education, China. This work is also supported in part by the Huaihua University Double First-Class initiative Applied Characteristic Discipline of Control Science and Engineering. We are also very grateful for the helpful suggestions and constructive comments given by the reviewers and editors.

References
[1] Baro E, Degoul S, Beuscart R and Chazard E2015 Toward a Literature-Driven Definition of Big Data in Healthcare J. BioMed research international pp1-9
[2] Sutherland et al 2016 Utilizing electronic health records to predict acute kidney injury risk and outcomes: workgroup statements from the 15th ADQI Consensus Conference J. Canadian Journal of Kidney Health and Disease 3 11
[3] Wald JS, Haque SN, Rizk S, Webb JR etc 2018 Pediatrics J. Enhancing Health IT Functionality for Children: The 2015 Children’s EHR Format 141(4) pp2016-3894
[4] Lehmann CU, O’Connor KG, Shorte VA and Johnson TD 2015 Pediatrics, Use of electronic health record systems by office-based paediatricians 135(1)
[5] Miller D, Noonan K, Fiks AG and Lehmann CU 2015 Pediatrics, Increasing pediatrician participation in EHR incentive programs 135(1)
[6] Canino G, SuoQ L and Veltri P 2016 Feature Selection Model for Diagnosis, Electronic Medical Records and Geographical Data Correlation Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 10 pp616-621
[7] GulloF, Ponti G, TagarelliA, TradigoG and VeltriP 2012A time series approach for clustering mass spectrometry data J. Journal of Computational Science 3(5) pp344-355
[8] XuH, StennerS P, DoanS, JohnsonKB, WaitmanLR and DennyJC 2010 Medex: a medication information extraction system for clinical narratives J. Journal of the American Medical Informatics Association 17(1) pp19-24