Zhang *et al. J Transl Med* (2019) 17:2
https://doi.org/10.1186/s12967-018-1758-2

Journal of Translational Medicine

RESEARCH

Open Access

Prediction of postoperative complications of pediatric cataract patients using data mining

Kai Zhang[1,2], Xiyang Liu[1,3,4*], Jiewei Jiang[1,2], Wangting Li[2], Shuai Wang[4], Lin Liu[1], Xiaojing Zhou[5] and Liming Wang[1,3,4]

Abstract

Background: The common treatment for pediatric cataracts is to replace the cloudy lens with an artificial one. However, patients may suffer complications (severe lens proliferation into the visual axis and abnormal high intraocular pressure; SLPVA and AHIP) within 1 year after surgery and factors causing these complications are unknown.

Methods: Apriori algorithm is employed to find association rules related to complications. We use random forest (RF) and Naïve Bayesian (NB) to predict the complications with datasets preprocessed by SMOTE (synthetic minority oversampling technique). Genetic feature selection is exploited to find real features related to complications.

Results: Average classification accuracies in three binary classification problems are over 75%. Second, the relationship between the classification performance and the number of random forest tree is studied. Results show except for gender and age at surgery (AS); other attributes are related to complications. Except for the secondary IOL placement, operation mode, AS and area of cataracts; other attributes are related to SLPVA. Except for the gender, operation mode, and laterality; other attributes are related to the AHIP. Next, the association rules related to the complications are mined out. Then additional 50 data were used to test the performance of RF and NB, both of then obtained the accuracies of over 65% for three classification problems. Finally, we developed a webserver to assist doctors.

Conclusions: The postoperative complications of pediatric cataracts patients can be predicted. Then the factors related to the complications are found. Finally, the association rules that is about the complications can provide reference to doctors.

Keywords: Random forest, Naïve Bayesian, Association rules mining, Genetic feature selection, Medical decision making system

Background

The big data era for medicine is coming. Traditional statistical methods cannot effectively discover hidden laws from numerous medical records, whereas data mining [1] and machine learning [2], the most promising techniques, can tackle this problem. To investigate the application of data mining and machine learning on clinical records, we focus on pediatric cataracts, which can cause children blindness [3]. We obtained the detailed diagnosis and treatment information of 321 patients from one of the largest pediatric cataract databases (CCPMOH) [4] and evaluated the disease development in patients.

For children under age of two, pediatric cataracts prevent the light from penetrating into eyes and also delay the development of optic nerve [5]. Moreover, it is difficult for both patients and their parents to realize that children are already developing this disease, therefore, subsequent treatment is usually unfortunately delayed. Even if treatment is thankfully given, some unexpected complications (severe lens proliferation into the visual axis and abnormal high intraocular pressure) still arise after such treatment of replacing cloudy lens with artificial ones. Worse still, why and how these complications turn up are unknown according to the present studies. This study aims to adopt some data mining and machine

*Correspondence: xyliu@xidian.edu.cn; xyliuxidian@163.com
[1] School of Computer Science and Technology, Xidian University, No.2 South Taibai Rd, Xi'an 710071, China
Full list of author information is available at the end of the article

learning techniques to predict these complications automatically within 1 year after surgery and determine which factors are more related to the complications. Because the heterogeneity of different patients is common in many diseases [6–8] and datasets that cover this disease are rare, this problem is difficult to tackle. Doctors and researchers required to circumvent this problem.

Many achievements have been made regarding the application of data mining and machine learning in the field of medical records. For example, Dheeraj Raju et al. used random forest to explore the factors influencing pressure ulcers and built a predictive model that is useful for doctors to predict the complication of this disease with the data collected over time [9]. Sónia Pereira et al. used obstetric and pregnancy factors to predict the appropriate delivery type to provide service of high quality, and to decide how to take care of pregnant women and newborn babies [10]. Aljumah et al. collected data from WHO (World Health Organization) to study the effectiveness of different treatment types with consideration of the age factor [11]. The results show that younger patients should not take medicine immediately to avoid side effects and older patients must take medicine to relieve sickness. Somanchi et al. applied SVM (support vector machine) and logistic regression to predict cardiac arrest in the future with the data from medical records. The result obtained from these methods is better than the results of previous methods and tools [12]. All of the above examples show that data mining is a powerful tool to study the medical data and the application of data mining techniques will produce some desirable results.

Inspired by the above researches, we attempted to mine meaningful knowledge from medical records of pediatric cataracts patients and then developed a medical decision making system to help doctors with predicting the complications of pediatric cataract patients. To study which factors contribute to the complications, some pairs of discriminant association rules are mined out using Apriori algorithm with preprocessed dataset. The antecedents of the association rules are combinations of various factors. The consequents of these association rules are whether a patient will have complications or a specific type of complication. According to the characteristics of the dataset, we used a simple discretization method and SMOTE (synthetic minority oversampling technique) to preprocess the dataset. Subsequently the number of positive samples (samples with complications or specific complication) was close to the number of negative samples (samples without complication or specific complication). Then the random forest and naïve Bayes classifier were used to predict whether patients would be affected by complications of different levels. Experimental result shows that random forest and Naïve Bayesian achieve accuracies which are above 76% and 75% in three binary classification problems [whether a patient suffers from complications (severe lens proliferation into the visual axis (SLPVA) and abnormal high intraocular pressure (AHIP)]; whether a patient suffers from the SLPVA; whether a patient suffers from AHIP). Then the number of trees of the random forest is investigated to find the most suitable number of trees to optimize the performance of random forest. The genetic feature selection was employed to find out which factors exactly caused the complications. Besides, we used additional 50 data to test the performance of random forest and Naïve Bayesian classifier, and both of them reach 65% for three classification problems. Finally, we exploited the random forest and association rules mining in the present research to construct an automatic complication prediction system for pediatric cataract patients so that patients with different degrees of complication can be warned and treated earlier.

Methods

The dataset used in the current research is from Zhongshan Ophthalmic Center, Sun Yat-sen University, which is the most professional ophthalmic hospital in China and has set up the state of art ophthalmology laboratory [13]. There are a total of 321 samples in the dataset, where 26 samples suffer from SLPVA and AHIP simultaneously; 69 patients only suffer from SLPVA; 84 patients only suffer from AHIP, and 194 patients have no complication after surgery. In addition, the detailed information about all the attributes and their possible values are shown in Table 1.

Moreover, the sixth, seventh and eighth attributes in Table 1 are marked as the evaluation standards of the severity of pediatric cataracts, which are proposed by pediatric ophthalmologist with over 5 years of working experience in Zhongshan Ophthalmic Center. These three evaluation standards can portray the severity based on the morphology of the cataracts, where the density of the cataract can be classified as dense or sloppy, the spreading area of cataract can be expressed as big or small, and the relative position of the cataract can be assessed by whether focus covers the central area of the lens. Previous work of our team completed the grading (severity evaluation) of pediatric cataracts from these three aspects and made some progress [14], but the accuracy still cannot reach 100%. Therefore, we invite pediatric ophthalmologist with over 5 years of working experience to assess the three attributes of these samples in terms of the slit lamp images of patients. The difference among the three grading standards can be illustrated by Fig. 1.

Table 1 The specification of attributes in dataset

NO. of Attribute (attribute name)	Values
1. Gender	Male, female
2. Secondary IOL placement	Yes, no (primary IOL placement)
3. Operation mode	Lens aspiration (I/A), lens aspiration with posterior continuous curvilinear capsulorhexis (I/A + PCCC), lens aspiration with posterior continuous curvilinear capsulorhexis and anterior vitrectomy (I/A + PCCC + A-Vit)
4. Laterality	Unilateral cataracts, bilateral cataracts
5. Age at surgery (AS)	1, 2, 3, 4, 5, 6, 7
6. Area of cataracts (AC)	Large, small
7. Density of cataracts (DC)	Dense, sloppy
8. Position of cataracts (PC)	Covering the central area of lens, not covering the central area of lens
9. Nystagmus	Yes, no
10. Microphthalmia	Yes, no
11. Microcornea	Yes, no
12. Persistent hyperplastic primary vitreous (PHPV)	Yes, no

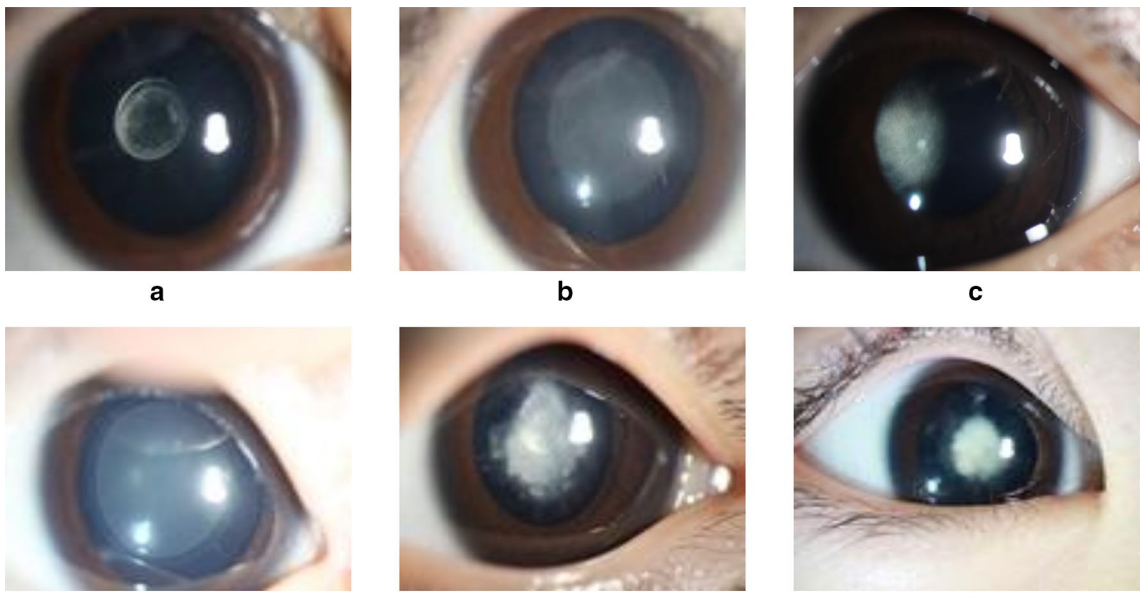

Fig. 1 Three grading standards of severity of pediatric cataracts. **a** the area of cataract is small; **b** the area of cataract is large; **c** the density of the cataract is sloppy; **d** the density of the cataract is dense; **e** the cataract covers the central area of lens; **f** the cataract does not covering the central area of lens

In addition, according to relevant literature [15], the surgery age is particularly important in the development of the illness. Finally, according to the expert knowledge of three experienced ocular doctors, the operation age for all of the patients is discretized into seven sections: [0, 3], [4, 6], [7, 9], [10, 12], [13, 18], [19, 24], [>24] (the unit is month) to facilitate classification. Except for the first and fifth attributes, the remaining attributes in Table 1 are recorded by doctors.

In current research, there are three main classification problems in this research: whether suffer from complications; whether suffer from SLPVA; whether suffer from AHIP. Therefore, the original dataset were decomposed into three sub datasets accordingly. Because of the classification problems in this research are non-numerical classification problems, random forest and Naïve Bayesian classifier which can tackle non-numerical classification problems were selected to complete classification tasks. Additionally, the mechanism of these two classification

methods are pretty easy to understand and there are so many packages can be directly used to implement.

Overview of methods

There are some common ways to tackle an imbalanced dataset, such as over sampling [16] (e.g., SMOTE [17]), under sampling [18] (e.g., clustering based under sampling [19]) and cost sensitive classification [20, 21]. Cost-sensitive method requires a cost matrix or a modified algorithm mechanism and architecture, which is more difficult for specific application. Therefore, SMOTE (Synthetic Minority Oversampling Technique) was adopted as an over sampling method to preprocess the dataset. Apriori algorithm and genetic feature selection were used to explore which combinations of factors are more likely to cause complication and which factors are more related with complications, respectively. Because the attributes in the dataset is non-numeric value, we choose Naïve Bayesian classifier and random forest to predict the post-operative complication of pediatric cataract patients.

Apriori algorithm

Association rules mining [22] is a useful technology that is first used to analyze the shopping basket to find some shopping habits hidden in the shopping list of consumers and thus is used to promote sales. This method has been applied in many fields to discover the meaningful knowledge in specific problem, such as fault diagnosis of power transformers [22] and the detection of industrial intrusion [23].

The most commonly used algorithm to mine association rules is Apriori [23], which utilizes minimum support (the frequency of an itemset appeared in dataset for screening out the frequent item-set) to sift out the frequent k-item sets and then selects the association rules whose confidence is larger than the minimum confidence from frequent k-item set. In the first stage of the Apriori algorithm, with the assistance of a transcendental property that the nonempty subset of the frequent item set must be frequent, the Apriori algorithm finds the most common occurring mode (k-item set) whose support is larger than or equal to the minimum support that is set ahead until the $(k+1)$-item set produced from k-item set is empty; Then Apriori algorithm checks the confidence of every rule and retains only the rules with higher confidence. The confidence of an association rule is computed as Eq. (1).

$$Confidence(a \rightarrow b) = \frac{Frequency(a \cup b)}{Frequency(a)} \tag{1}$$

The association rules will come into being as many implications with the format "$a \rightarrow b$" using the Apriori algorithm, where a and b are the antecedents and consequents of association rules, respectively. The minimum support is set as 0 to let all items combine freely to produce association rules without considering their support and confidence. Because the aim is finding out the combinations of attributes which will or will not bring complication, the rules whose consequents are whether a patient suffers from complications (SLPVA and AHIP) are sifted out preliminarily. At last, the pairs of association rules with the same antecedent and different consequents will be sifted out again with Eq. (2), which means that the confidence of the pair of association rules should be far from each other mutually, where *threshold* is a parameter.

$$\frac{Confidence(a \rightarrow b)}{Confidence(a \rightarrow b')} \geq Threshold, \quad b = \neg b', \quad b \in \{1, 0\} \tag{2}$$

where, $a \rightarrow b$ and $a \rightarrow b'$ are a pair of association rules with the same antecedent and different consequents. 1 and 0 refer to suffering from complications (SLPVA and AHIP) and not suffering from complication (SLPVA and AHIP), respectively. The association rules mining is widely applied in biomedical fields, such as microbial energy prospection [24], pollution epidemiology [25].

SMOTE

SMOTE is a commonly used over-sampling technique for tackling imbalanced dataset. It used the k-nearest neighbor of each minority sample to produce more minority samples to offset the imbalance between majority class and minority class. The new minority samples is the linear combination of a minority sample and one of its nearest neighbor. Non-numerical minority data can also be preprocessed with SMOTE in a similar manner. In current research, we choose SMOTE as a comparison.

Naive Bayesian classifier

Naive Bayes classifier [26–28] based on Bayes theorem and assumes all attributes are independent is a simple yet useful pattern recognition method. Given sample (x, y) is a sample to be classified in multi-class classification problem, where $x = [x_1, \ldots, x_s]$, $y \in \{y_1, \ldots y_o\}$, S and o are the attribute vector, label, the number of attributes and the number of classes, respectively. Naive Bayes classifier compute probabilities $P(y_1|x), \ldots, P(y_o|x)$, and then classify the sample into the class with largest probability. These probabilities could be estimated with Bayes theorem [29, 30] which is shown as Eq. (3), then the label of a new sample can be computed with Eq. (4).

$$P(y_j|x) = \frac{P(x|y_j)}{P(x)} \quad j = 1, \ldots, o \tag{3}$$

$$Label = \max_j p(y_j|x) \quad j = 1, \ldots, o \tag{4}$$

Because of the assumption that all attributes are independent and same denominator of Eq. (3) for all classes, the probabilities could be converted to be Eq. (5):

$$\begin{aligned} P(x|y_j) &= p(y_j) = P(x_1|y_j)\ldots P(x_S|y_j)p(y_j) \\ &= \prod_{i=1}^{S} P(x_i|y_j)p(y_j) \quad j = 1,\ldots,o \end{aligned} \tag{5}$$

Naive Bayes classifier has been applied in a plenty of fields, such as detection of cardiovascular disease risk's level [31], identification of hot spots in protein structures [32].

Random forest

Random forest is an effective ensemble classifier that originates from decision tree, and this classifier can effectively avoid overfitting, which is a common issue with normal decision trees. Therefore, the decision tree is introduced first.

Decision tree [33, 34] is a type of classifier that regards the dataset to be an entire set, yet recursively divides this set into subsets as well according to a certain standard. During the latter process, all subsets are divided to the extent that they have no attributes to divide further or all samples in every subset belong to a uniform category. Decision tree has been applied in many fields, such as industrial safety [34], power [35] and financial behavior prediction [36].

There are three common standards to partition the dataset into subsets, namely, information gain, gain ratio and Gini index [37].

Random forest [38, 39] is an excellent ensemble learning algorithm that combines many decision trees with partial samples and partial attributes that are randomly selected from the whole dataset. Finally, the label of a testing sample is decided by voting in accordance with all decision trees of the random forest. Therefore, overfitting problem can be avoided effectively. The current research adopted this algorithm to predict whether a patient will have complications and to classify the patients whether suffer from a specific complication. Random forest is also popular and widely applied in many disciplines, such as prediction of double-strand DNA breaks [40], localization of prostate cancer [39] and computational bioinformatics [41].

Genetic feature selection

Next, genetic feature selection is used to find out which factors are more related with the complications of pediatric cataract patients. Genetic algorithm (GA) [42] can be used to select attributes that are useful for classification or more related to complications in the current research, where the fitness evaluation function of GA is the accuracy returned by classifier (random forest in this paper). In the flow of GA, the real factors which are associated with labels are searched out. The number of iteration steps and the size of population are 50 and 30, respectively. Genetic feature selection has been applied in many fields, such as image classification [43], text categorization [44], image feature extraction [45] and signal processing [46].

Experimental settings and evaluation indicator

Because this classification problem (prediction of postoperative complication of pediatric cataracts patients) is a multi-label learning problem [47], common solution for this type of classification problem is to decompose the problem into several binary classification ones. At the same time, the dataset is imbalanced, so the samples of the minority class are fed into SMOTE to produce more minority samples. Consequently, there are three corresponding datasets for three binary classification problems (whether a patient suffers from complications; whether a patient suffers from SLPVA; whether a patient suffers from AHIP). The samples having complications (SLPVA and AHIP) are positive samples and the samples without complications (SLPVA and AHIP) are negative samples. Three-fold cross validation was adopted to permit a fair comparison of these methods.

All of the methods were implemented with MATLAB R2016a on a personal computer with an Intel 2.80 GHz i5 processor, 8G RAM. The objective of our study was to predict the complication of pediatric cataracts patients; the measures that were employed to evaluate the performance are shown as Eqs. (6–11):

$$Accuracy = (TP + TN)/(P + N) \tag{6}$$

$$Precision = TP/(TP + FP) \tag{7}$$

$$Sensitivity\ (TPR,\ Recall) = TP/(TP + FN) \tag{8}$$

$$FNR\ (false\ negative\ rate) = 1 - Sensitivity = FN/(TP + FN) \tag{9}$$

$$Specificity = TN/(TN + FP) \tag{10}$$

$$FNR\ (false\ positive\ rate) = 1 - Sensitivity = FP/(TN + FP) \tag{11}$$

where P and N are the number of positive samples and negative samples, respectively; TP indicates the number of positive samples classified into the positive class; FN denotes the number of positive samples classified as negative samples; TN is the number of negative samples recognized as negative samples; and FP refers to the number of negative samples identified as positive samples. In addition, the ROC (receiver operating characteristics) curve, which indicates how many positive samples

are recognized conditioned on a given false positive rate and AUC (area under curve) which means the area of the zone under ROC curve are also adopted to assess the performance [48].

Results

At first, we provide a part of the association rules mined out by Apriori algorithm in Additional file 1: Tables S1–S3. Then the postoperative complications of pediatric cataracts patients were predicted with random forest and Naïve Bayesian classifier. Then the contributing factors for the postoperative complications were found out with genetic feature selection. Finally, we used additional 50 data to verify the performance of random forest and Naïve Bayesian classifier, the accuracies of them are over 65% for three classification problems.

Results of association rules mining

In order to find out the combination of factors which can cause complications with bigger probability, the Apriori algorithm is used to mine out the association rules about the complications.

Classification performance of RF and NB

The prediction of postoperative complication of pediatric cataracts patients is carried out with random forest and naïve Bayes classifier. The evaluation indicators for random forest and Naïve Bayes classifier with three different datasets that are preprocessed with SMOTE or original datasets in terms of three binary classification problems (whether a patient suffers from complications; whether a patient suffers from SLPVA; whether a patient suffers from AHIP) are shown in Table 2, where the format of these data is $\mu \pm \delta$ (μ is the mean value, δ is the standard deviation). The ROC curves combining AUC values in three binary classification problems with datasets preprocessed with SMOTE are shown in Fig. 2. The number of trees in random forest influences the classification performance and we investigate the relationship between them. Three classification problems were repeatedly solved with different number of trees and their accuracies are shown as Fig. 3.

Results of genetic feature selection

Next, we applied genetic algorithm (GA) combined with random forest to study which factors affect the complication, where the classification accuracy is used as the fitness of GA. Because GA is a type of probabilistic heuristic algorithm that involves random factors, the experiment was repeated ten times and the best result was selected as the final result. The experimental result shows that the accuracy for the first binary classification problem (whether suffers from complications) can reach 0.783 without using the gender and age at surgery (AS) attributes. The accuracy of the second binary classification (whether suffers from SLPVA) problem can reach 0.795 without the secondary IOL placement, operation mode, age at surgery and area of cataracts (AC) attributes. The accuracy of the third binary classification problem (whether suffer from AHIP) can reach 0.836 without gender, operation mode and laterality.

Table 2 Performance indicators in three binary classification problems

Method		Accuracy	FNR	FPR
Problem 1: Whether a patient suffers from complications				
Random forest	–	0.757 ± 0.025	0.414 ± 0.031	0.128 ± 0.013
	SMOTE	0.762 ± 0.019	0.231 ± 0.013	0.220 ± 0.037
Naïve Bayesian	–	0.748 ± 0.025	0.465 ± 0.042	0.887 ± 0.023
	SMOTE	0.751 ± 0.032	0.270 ± 0.043	0.208 ± 0.044
Problem 2: Whether a patient suffers from SLPVA				
Random forest	–	0.810 ± 0.014	0.621 ± 0.089	0.071 ± 0.023
	SMOTE	0.753 ± 0.069	0.257 ± 0.054	0.258 ± 0.044
Naïve Bayesian	–	0.782 ± 0.014	0.155 ± 0.043	0.449 ± 0.100
	SMOTE	0.782 ± 0.043	0.244 ± 0.065	0.267 ± 0.025
Problem 3: Whether a patient suffers from AHIP				
Random forest	–	0.838 ± 0.024	0.580 ± 0.050	0.015 ± 0.014
	SMOTE	0.813 ± 0.016	0.228 ± 0.055	0.265 ± 0.025
Naïve Bayesian	–	0.847 ± 0.033	0 ± 0	0.321 ± 0.043
	SMOTE	0.816 ± 0.037	0.225 ± 0.047	0.265 ± 0.074

Additional testing

We used additional 50 samples to test performance of RF and NB, the accuracies reach 65% for three classification problems, respectively. The detailed information about additional testing is shown in Table 3. Figure 4 shows the ROC curves for additional testing. There are 27 samples do not suffer from complications; 18 samples suffer from SLPVA; 11 samples suffer from AHIP and five samples suffer both SLPVA and AHIP.

Webserver of decision-making system

We developed and deployed a web based system to facilitate the prediction of complication levels for doctors and to show the association rules about complication. This system serves two functions: judging whether a patient suffers from complications or the specific type of complication using random forest as the flowchart shown in Fig. 5 and showing the association rules after user inputs parameter *threshold* for three types of problems. We provide two language versions (English and Chinese) for users and this web server is freely available at (English version) http://120.27.126.89:5001/option_xian (Chinese version) http://120.27.126.89:5001/option_xianch.

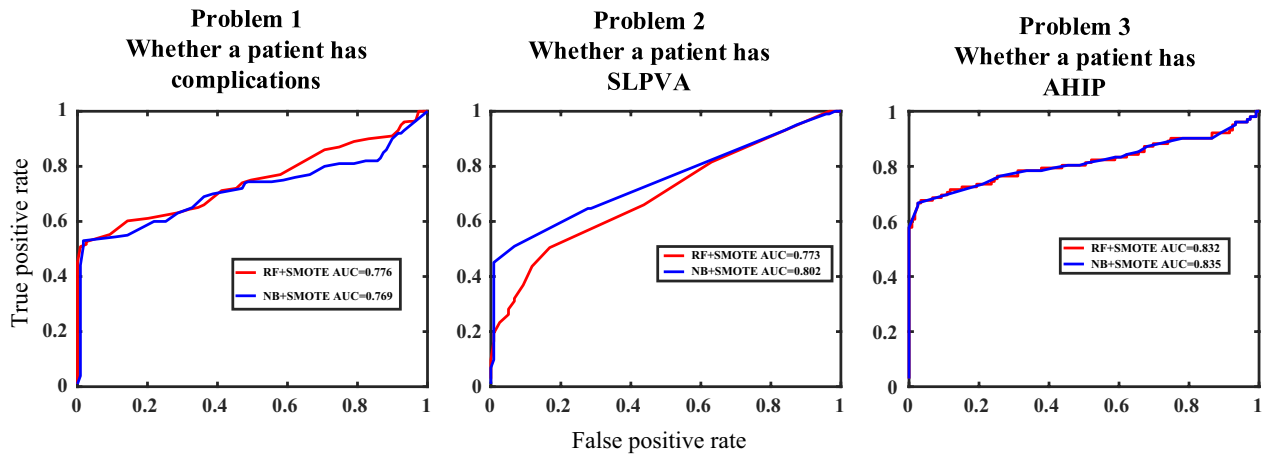

Fig. 2 ROC curves and AUC values in three binary classification problem. (*ROC* receiver operating characteristics curve, *AUC*: area under curve, *RF* random forest; *SMOTE* synthetic minority oversampling technique, *NB* Naïve Bayesian classifier)

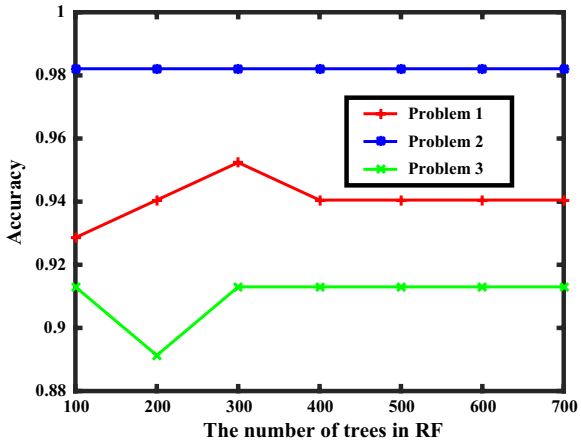

Fig. 3 The relationship between accuracy and the number of trees of random forest. (RF means random forest)

Discussion

The number of association rules that are sifted out decreases while the *threshold* increases. The antecedents, consequents and confidences of all association rules are summarized to be three categories of tables, contributing as a part of experimental result. The reason why these steps are adopted is that we want to obtain some discriminant association rules that can provide some evidence about complications and whose consequences is different, but the reliability of these rules need to be verified with more clinical data. Some discriminant combinations of attributes (antecedent) can reflect the power of attributes for classification, such as secondary IOL replacement and gender used for classifying whether suffers from complications. However, these association rules need to be testified with more clinical data. The relationship between threshold and the number of associations rules is shown as Fig. 6.

These association rules can provide some reference for doctors to predict postoperative complication of pediatric cataracts. For example, a male patient whose age is between 13 and 18 months will be more unlikely to suffer from AHIP after the surgery on the basis of *threshold* = 6, which is shown in Additional file 1: Table S3.

SMOTE algorithm[17] is a common over sampling method for imbalance dataset classification, it randomly sampling two samples from dataset and linearly combined them to be a new sample until imbalanced classes become balance. Here we used nominal SMOTE that is

Table 3 The performance of RF and NB for additional testing

Problem	Algorithm	Accuracy	Sensitivity	Specificity
Whether a patient suffers from complications	Random forest	0.700	0.625	0.769
	Naïve Bayesian	0.700	0.731	0.667
Whether a patient suffers from SLPVA	Random forest	0.720	0.667	0.722
	Naïve Bayesian	0.660	0.611	0.688
Whether a patient suffers from AHIP	Random forest	0.700	0.636	0.718
	Naïve Bayesian	0.660	0.545	0.692

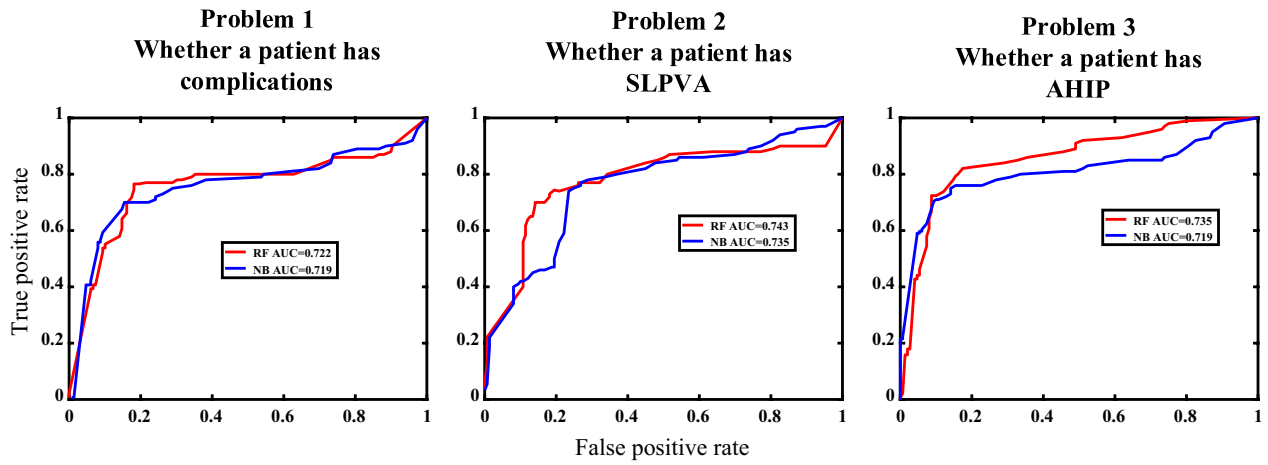

Fig. 4 ROC curves and AUC values for additional testing

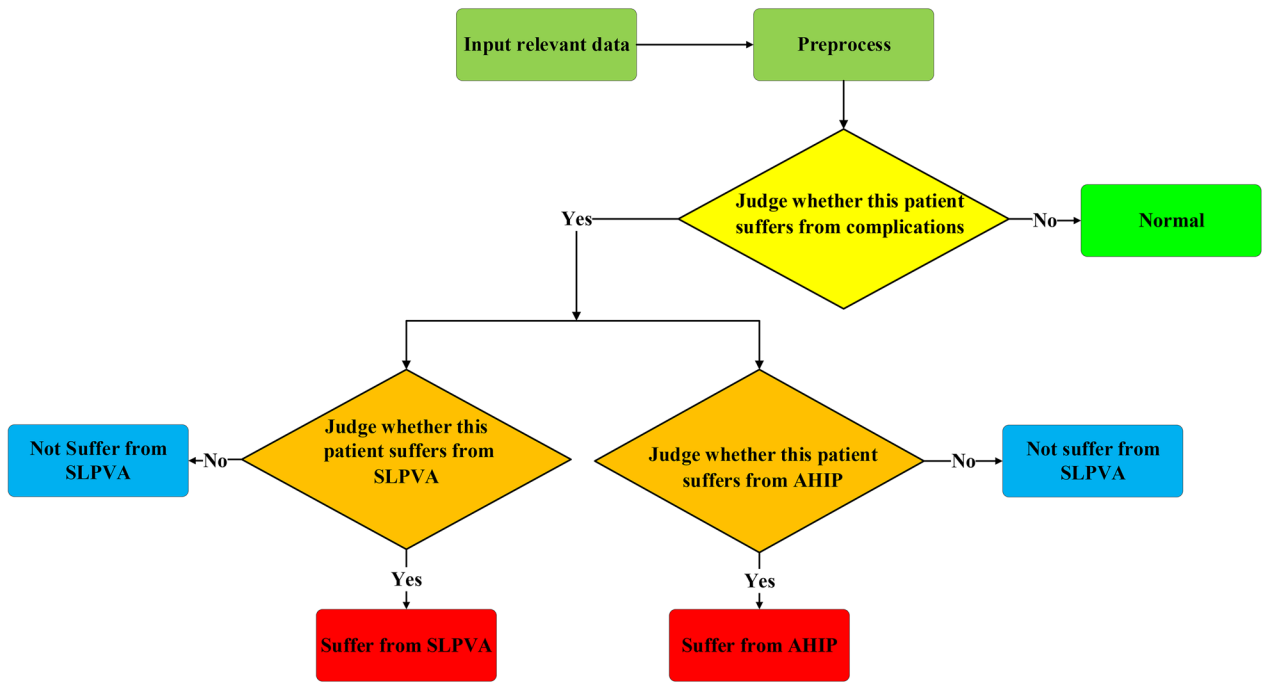

Fig. 5 Flowchart of postoperative complication prediction. (At first, the inputted data is preprocessed with data discretion method. Then one of the three models is applied to distinguish whether a patient suffers from complication. If the patient is not normal, then the remaining two models are used to judge whether the patient has two types of complication. If the patient is judged to be normal, the remaining two models will not be used)

the extension of SMOTE in nonnumeric variables. All data were divided into multiple parts for cross-validation and then were preprocessed by SMOTE. Validation dataset is not preprocessed with SMOTE.

Sensitivity and specificity can be computed with FNR and FPR, so values for some sensitivity and specificity indices are not shown. We also use original dataset to perform these classification problems. SMOTE can effectively improve the classification performance. The classification performance with original dataset is so imbalanced, while SMOTE can relieve this condition to a certain extent.

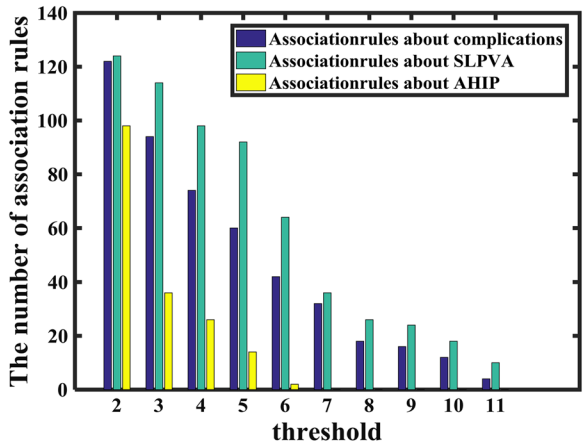

Fig. 6 The relationship between *threshold* and the number of association rules

Conclusion

To predict the postoperative complication of pediatric cataracts, SMOTE is employed to preprocess dataset and then two types of classifier (random forest and naïve Bayes classifier) was exploited to classify samples with or without complication (two types of complication). All average accuracies in solving three binary classification problems are over 91%, and one is even reaching 95%. Then real factors that are more related to the complications were identified with genetic feature selection. Besides, the association rules that hide in the dataset can also provide some evidence about complication to assist doctors in treatment. Finally, additional 50 data were used to test the performance of RF and NB, and both the accuracies of them reach 65% for three classification problems. All these verified methods and models were integrated into a medical decision making system to help doctors to predict the postoperative complication of Pediatric Cataract Patients. In the future, more information, such as medication during pregnancy and genetic information, can be put into dataset to predict complications more accurately. As a type of rare disease, the data of pediatric cataracts is not enough. Therefore, we hope the on-line system in current research can assist in collecting more data and support the multicenter study in the future.

Additional file

Additional file 1: Table S1. Association rules about whether a patient will have complications (*threshold* = 6). **Table S2.** Association rules about whether a patient will have the first type of complication (*threshold* = 7). **Table S3.** Association rules about whether a patient will have the second type of complication (*threshold* = 4).

Abbreviations

RF: random forest; NB: Naïve Bayesian Classifier; ROC: receiver operating characteristics curve; AUC: area under curve; SMOTE: synthetic minority over sampling technique.

Authors' contributions

XYL designed the research; KZ conducted the study; WTL collected the dataset; KZ were responsible for coding; SW analyzed and finished the experimental results; SW, LL, XJZ and KZ developed the complication prediction system; KZ, JWJ, XYL and LMW co-wrote the manuscript. All authors discussed the results and commented on the manuscript. All authors read and approved the final manuscript.

Author details

[1] School of Computer Science and Technology, Xidian University, No.2 South Taibai Rd, Xi'an 710071, China. [2] State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou 510060, China. [3] Institute of Software Engineering, Xidian University, Xi'an 710071, China. [4] School of Software, Xidian University, Xi'an 710071, China. [5] School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China.

Acknowledgements

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The data that support the findings of this study are available from the corresponding authors on request. The code are available to readers on request. The code and data in this study can be accessed freely at https://github.com/Hugo0512/Complications.

Consent for publication

Not applicable.

Ethics approval and consent to participate

The study was approved by the Ethics Committee of Xidian University and Zhongshan Ophthalmic Center of Sun Yat-sen University. Written informed consent was obtained from all study participants' parents or legal guardian.

Funding

This study was funded by the National Key Research and Development Program of China (2018YFC0116500); the NSFC (No. 91546101, 61472311 and 11401454); the Guangdong Provincial Natural Science Foundation (No. YQ2015006, No. 2014A030306030, No. 2014TQ01R573, No. 2013B020400003); the Natural Science Foundation of Guangzhou City (No. 2014J2200060); the State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University (No. 2015ykzd11, No. 2015QN01); the Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase); and the Clinical Research and Translational Medical Center for Pediatric Cataract in Guangzhou City; the Fundamental Research Funds for the Central Universities (No. BDZ011401, BDZ011401 and JB151005); and the National Defense Basic Research Project of China (jcky2016110c006).

Publisher's Note

Received: 1 October 2018 Accepted: 21 December 2018
Published online: 03 January 2019

References

1. Duggirala HJ, Tonning JM, Smith E, et al. Use of data mining at the Food and Drug Administration. J Am Med Inform Assoc. 2016;23:428.

2. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316:2402.
3. Resnikoff S, Keys TU. Future trends in global blindness. Indian J Ophthalmol. 2012;60:387–95.
4. Lin H, Lin D, Chen J, et al. Distribution of axial length before cataract surgery in chinese pediatric patients. Sci Rep. 2016;6:23862.
5. Daw NW. Visual development. US: Springer; 2006.
6. Jackson WS, Lindquist S. Illuminating aggregate heterogeneity in neurodegenerative disease. Nat Methods. 2007;4:1000–1.
7. Chen Z, Fillmore CM, Hammerman PS, Kim CF, Wong KK. Non-small-cell lung cancers: a heterogeneous set of diseases. Nat Rev Cancer. 2014;14:535–46.
8. Bedard PL, Hansen AR, Ratain MJ, Siu LL. Tumour heterogeneity in the clinic. Nature. 2013;501:355–64.
9. Raju D, Su X, Patrician PA, Loan LA, McCarthy MS. Exploring factors associated with pressure ulcers: a data mining approach. Int J Nurs Stud. 2015;52:102–11.
10. Pereira S, Portela F, Santos MF, Machado J, Abelha A. Predicting type of delivery by identification of obstetric risk factors through data mining. Procedia Comput Sci. 2015;64:601–9.
11. Aljumah AA, Ahamad MG, Siddiqui MK. Application of data mining: diabetes health care in young and old patients. J King Saud Univ Comput Inf Sci. 2013;25:127–36.
12. Somanchi S, Adhikari S, Lin A, Eneva E, Ghani R. Early prediction of cardiac arrest (code blue) using electronic medical records. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM; 2015. p. 2119–2126.
13. Lin H, Long E, Chen W, Liu Y. Documenting rare disease data in China. Science. 2015;349:1064.
14. Liu X, Jiang J, Zhang K, et al. Localization and diagnosis framework for pediatric cataracts based on slit-lamp images using deep features of a convolutional neural network. PLoS ONE. 2017;12:e0168606.
15. Mataftsi A, Haidich AB, Kokkali S, et al. Postoperative glaucoma following infantile cataract surgery: an individual patient data meta-analysis. Jama Ophthalmol. 2014;132:1059–67.
16. Barua S, Islam MM, Yao X, Murase K. MWMOTE–majority weighted minority oversampling technique for imbalanced data set learning. IEEE Trans Knowl Data Eng. 2014;26:405–25.
17. Verbiest N, Ramentol E, Cornelis C, Herrera F. Preprocessing noisy imbalanced datasets using SMOTE enhanced with fuzzy rough prototype selection. Appl Soft Comput. 2014;22:511–7.
18. Burges CJC. A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc. 1998;2:121–67.
19. Yen SJ, Lee YS. Cluster-based under-sampling approaches for imbalanced data distributions. Expert Syst Appl. 2009;36:5718–27.
20. Ibáñez A, Bielza C, Larrañaga P. Cost-sensitive selective naive Bayes classifiers for predicting the increase of the h-index for scientific journals. Neurocomputing. 2014;135:42–52.
21. Zidelmal Z, Amirou A, Ould-Abdeslam D, Merckle J. ECG beat classification using a cost sensitive classifier. Comput Methods Progr Biomed. 2013;111:570–7.
22. Yang Z, Tang WH, Shintemirov A, Wu QH. Association rule mining-based dissolved gas analysis for fault diagnosis of power transformers. IEEE Trans Syst Man Cybern Part C. 2009;39:597–610.
23. Khalili A, Sami A. SysDetect: a systematic approach to critical state determination for Industrial Intrusion Detection Systems using Apriori algorithm. J Process Control. 2015;32:154–60.
24. Shaheen M, Shahbaz M. An algorithm of association rule mining for microbial energy prospection. Sci Rep. 2017;7:46108.
25. Bellinger C, Mohomed Jabbar MS, Zaïane O, Osornio-Vargas A. A systematic review of data mining and machine learning for air pollution epidemiology. BMC Public Health. 2017;17:907.
26. Jiang L, Li C, Wang S, Zhang L. Deep feature weighting for naive Bayes and its application to text classification. Eng Appl Artif Intell. 2016;52:26–39.
27. Kim S-B, Han K-S, Rim H-C, Myaeng SH. Some effective techniques for naive bayes text classification. IEEE Trans Knowl Data Eng. 2006;18:1457–66.
28. Wu J, Pan S, Zhu X, Cai Z, Zhang P, Zhang C. Self-adaptive attribute weighting for Naive Bayes classification. Expert Syst Appl. 2015;42:1487–502.
29. Liu Y-F, Guo J-M, Lee J-D. Halftone image classification using LMS algorithm and naive Bayes. IEEE Trans Image Process. 2011;20:2837–47.
30. Marucci-Wellman HR, Lehto MR, Corns HL. A practical tool for public health surveillance: semi-automated coding of short injury narratives from large administrative databases using Naïve Bayes algorithms. Accid Anal Prev. 2015;84:165–76.
31. Miranda E, Irwansyah E, Amelga AY, Maribondang MM, Salim M. Detection of cardiovascular disease risk's level for adults using Naive Bayes classifier. Healthcare Inform Res. 2016;22:196–205.
32. Zhang H, Jiang T, Shan G. Identification of hot spots in protein structures using Gaussian network model and Gaussian Naive Bayes. Biomed Res Int. 2016;2016:4354901.
33. Arvind V, Köbler J, Kuhnert S, Rattan G, Vasudev Y. On the isomorphism problem for decision trees and decision lists. Theor Comput Sci. 2015;590:38–54.
34. Mistikoglu G, Gerek IH, Erdis E, Usmen PEM, Cakan H, Kazan EE. Decision tree analysis of construction fall accidents involving roofers. Expert Syst Appl. 2014;42:2256–63.
35. Senroy N, Heydt GT, Vittal V. Decision tree assisted controlled islanding. IEEE Trans Power Syst. 2006;21:1790–7.
36. Nasseri AA, Tucker A, Cesare SD. Quantifying StockTwits semantic terms' trading behavior in financial markets: an effective application of decision tree algorithms. Expert Syst Appl. 2015;42:9192–210.
37. Jiawei H, Micheline K, Jian P. Data mining: concepts and techniques. 3rd ed. China: China Machine Press; 2012. p. 217–21.
38. Pei S-C, Chen L-H. Image quality assessment using human visual DOG model fused with random forest. IEEE Trans Image Process. 2015;24:3282–92.
39. Qian C, Wang L, Gao Y, et al. In vivo MRI based prostate cancer localization with random forests and auto-context model. Comput Med Imaging Gr. 2016;52:44–57.
40. Mourad R, Ginalski K, Legube G, Cuvier O. Predicting double-strand DNA breaks using epigenome marks or DNA at kilobase resolution. Genome Biol. 2018;19:34.
41. Wu Q, Ye Y, Liu Y, Ng MK. SNP selection and classification of genome-wide SNP data using stratified sampling random forests. IEEE Trans Nanobiosci. 2012;11:216–27.
42. Yang Q, Wang M, Xiao H, et al. Feature selection using a combination of genetic algorithm and selection frequency curve analysis. Chemomet Intell Lab Syst. 2015;148:106–14.
43. Wang L, Zhang K, Liu X, et al. Comparative analysis of image classification methods for automatic diagnosis of ophthalmic images. Sci Rep. 2017;7:41545.
44. Ghareb AS, Bakar AA, Hamdan AR. Hybrid feature selection based on enhanced genetic algorithm for text categorization. Expert Syst Appl. 2016;49:31–47.
45. Nagarajan G, Minu R, Muthukumar B, Vedanarayanan V, Sundarsingh S. Hybrid genetic algorithm for medical image feature extraction and selection. Procedia Comput Sci. 2016;85:455–62.
46. Lu L, Yan J, de Silva CW. Feature selection for ECG signal processing using improved genetic algorithm and empirical mode decomposition. Measurement. 2016;94:372–81.
47. Zhang M-L, Zhang K. Multi-label learning by exploiting label dependency. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. New York: ACM; 2010. p. 999–1008.
48. Zhang K, Liu X, Liu F, He L, Zhang L, Yang Y, Li W, Wang S, Liu L, Liu Z, Wu X, Lin H. An interpretable and expandable deep learning diagnostic system for multiple ocular diseases: qualitative study. J Med Internet Res. 2018;20(11):e11144.