Advances in Computer, Signals and Systems (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2023.071108
ISSN 2371-8838 Vol. 7 Num. 11

Tree-Based Prediction of Influential Factors and Information Mining

Xin Tan*

Hubei University of Education of Mathematics and Statistics, Wuhan, Hubei, 430205, China
tan134945@foxmail.com
**Corresponding author*

Keywords: Traffic Flow; Large-scale Gathering Activities; Traffic Congestion; Macro Base Map; VISSIM Simulation

Abstract: In minimally invasive gastrointestinal surgery (IPI), local sedative and analgesic drugs are required, and a new type of drug, "R-drug", has yet to be studied non-intervention ally. This paper analyzes and explores the vital signs, adverse effects and patient satisfaction of IPI based on the real performance data of new and traditional sedative drugs in clinical trials. In this paper, we first cleaned, coded and normalized the data, then based on multivariate visualization analysis, we found that there were significant differences between different drug groups regarding each adverse reaction, and we conducted chi-square test on different drug groups regarding each adverse reaction, and we found that there were significant differences between different drug groups regarding intra-operative adverse reactions, and only "nausea and vomiting" and "abdomen and vomiting" were found in the post-operative adverse reactions. Among the postoperative adverse reactions, only "nausea and vomiting" and "abdominal distension and abdominal pain" showed significant differences. Regarding the prediction of adverse reactions, this paper up-sampled the dataset and built a model based on the K nearest neighbor algorithm, and the classification AUC of the model on the tested dataset was above 0.92, and the confusion matrix and ROC diagram were made to visualize the specific testing of the model.

1. Introduction

New drug research is a key part of clinical research. When a new drug is put into use, it usually needs to go through two phases: biological and clinical trials. In order to understand the characteristics of the new drug, it is necessary to study and analyze the adverse reactions of the patients during and after surgery, the vital signs of the patients after surgery, and the satisfaction of the patients and their related problems, and to make a prediction on the above aspects based on the analysis of the experimental data, which will provide an important reference for the selection of the drug, and a prediction basis for the physicians and patients[1-2].

Understanding the factors that contribute to drug innovation is important both for improving health care and for the future of organizations engaged in drug discovery, research, and development. This study aims to help promote drug innovation by identifying the working patterns of new drugs and categorizing these drugs based on innovativeness [3-4]. This paper provides a comprehensive analysis

of this data and discusses the potential factors contributing to the observed trends. The public's desire for new therapies, their rising costs, and the government's increasingly important role as a player for innovative new medicines have focused on the rapidly rising cost of new drug development - now thought to be more than $800 million - and underscored the need for efficient use of resources [5].

Therefore, this paper firstly determines whether there is a significant difference between the new drug group and the original drug group according to the intraoperative and 24-hour postoperative adverse reactions; then it establishes a mathematical model based on the patient's basic information and the type of sedative drug to predict whether the patient will have an adverse reaction during the intraoperative and 24-hour postoperative period.

2. The fundamental of data analysis

2.1 Data pre-processing

In order to improve the quality of the data and facilitate better data analysis, there are missing values and outliers are removed or interpolated, and by looking at the data in Annex I, there are missing values, and the numerical characteristics are used to fill in the missing values with the mean value [6]. For some models that require input features for computation, feature deflation allows the model to be more accurate for classification or regression prediction. In the prediction of the adverse effects on patients during and 24h after surgery, in order to improve the accuracy of the KNN model prediction, the irrelevant data were deleted first, and due to the large difference in the proportion of some features, the "never smoked", "occasional smoker: smoked cigarettes more than four times a day" and "frequent smoker: smoked cigarettes more than four times a day" were first analyzed. Because of the large difference in the proportion of some features, the features "never smoked", "occasional smoker: smoked cigarettes more than four times a day" and "frequent smoker: smoked cigarettes more than one time a day" were first clustered into the two categories of "smoker" and "non-smoker" [7]. In order to facilitate the analysis and transform the categorical data into numerical data, this paper mainly adopts the unique thermal coding to transform the categorical variables into numerical variables, and then normalizes the numerical variables to make the values of different features comparable, and binarizes some binary categorical variables (mapped as 0 and 1) for subsequent modeling. For the data in Annex 1 first extracted data on intraoperative adverse reactions (cough, body movement, intraoperative other) and 24 hours postoperative adverse reactions (nausea and vomiting, lethargy and fatigue, dizziness and lightheadedness and headache, abdominal distension and abdominal pain, and other discomforts), and labeled those with the occurrence of the condition as "1", and those without as "0". Data visualization of the raw data revealed that the number of intraoperative and postoperative 24-hour periods with no adverse reactions far exceeded the number of adverse reactions, so the RandomOverSampler function in the imblearn library was used to upsample the data to balance the distribution of categories in the dataset by increasing the number of samples from a few categories. Based on the intraoperative and postoperative 24-h adverse reactions, in order to determine whether there is a significant difference between the new drug group and the original drug group, this paper next focuses on qualitative (data visualization) and quantitative (chi-square test) aspects [8].

2.2 Exploring quantitative differences based on chi-square tests

Next, the chi-square test was utilized to quantitatively explore whether there is a significant difference between the two groups of drugs by the emergence of adverse reactions during and 24 hours after surgery respectively, for the categorization problem can be listed in the cross-tabulation [9], it was found that the total sample is greater than 40 and the theoretical frequency of each category

is greater than 5 available chi-square test:

1) original hypothesis H0: there is no significant difference between the two drugs;

2) alternative hypothesis H1: there is a significant difference between the two drugs;

First calculate the test statistic:

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{O_{ij}-E_{ij}}{E_{ij}}. \quad (1)$$

Next calculate the degrees of freedom:

$$df = (m-1)\times(k-1). \quad (2)$$

Values were calculated using the cumulative distribution of the chi-square distribution:

$$p = 1 - F(\chi 2, df). \quad (3)$$

If the test statistic is greater than the critical value (p-value greater than the significance level of 0.05), the original hypothesis cannot be rejected and there is no significant difference between the two drugs; if the test statistic is less than the critical value (p-value less than the significance level of 0.05), the original hypothesis is rejected and there is a significant difference between the two drugs.

2.3 KNN-based Adverse Reaction Prediction

Let the number of samples in each labeled minority category be k [10-11]. There are:

$$P\left(x\middle|y=1, D_{up}\right) = \frac{\sum_{i=1}^{k}[x_i=x]}{k}. \quad (4)$$

Wherex_i denotes the features of the few class samples obtained from the ith sampling. The up-sampled dataset is then divided into a training set and a test set, a KNN model is built to fit the training set, and then the test set is used to evaluate the performance of the model.

There are three main basic elements of the model used in the nearest neighbor method - the distance metric, the choice of $L3(\vec{x}_i, \vec{x}_j)$ values and the classification decision rule decision. In the nearest neighbor method, when the training set, distance metric, $L3(\vec{x}_i, \vec{x}_j)$values and classification decision rule are determined, for any new input instance, the class to which it belongs is uniquely determined. This is equivalent to dividing the feature space into a number of subspaces based on the above elements and determining the class to which each point in the subspace belongs [12].

More optimal processing generally uses kd trees. To find k nearest neighbor samples more efficiently, a k-d tree is used to organize the training dataset. kd tree is a binary tree structure where each node represents a training sample. A kd tree can be constructed by dividing the space recursively for the training samples. When querying the k nearest neighbors of a test sample, the k training samples closest to the test sample can be found quickly by traversing the k-d tree.

Table 1: GridSearchCV tuning results for KNN models

parameters	range of values	optimal outcome
algorithm	[auto,ball__tree,kd__tree,brute]	auto
leaf_size	np.arange(10,51,10)	10
weights	[uniform,distance]	uniform
n__neighbors	np.arange(1,11,1)	1
p	[1,2,3]	3

Therefore, the k-d tree can be regarded as the data structure that implements the nearest neighbor search in the k-nearest neighbor method model, which can greatly improve the prediction efficiency of the model. In order to increase the accuracy of the model further, make the hyperparameter

optimization of the KNN model based on GridSearchCV implementing grid search with cross-validation. The parameter selection and final results are shown in Table 1.

3. Results

3.1 Results of the chi-square test

As can be seen from the table 2, the p-values of the chi-square test results for "cough", "body movement" and "other intraoperative reactions" are all less than 0.05, so it is considered that there is a significant difference between the two drugs in terms of intraoperative adverse reactions. The p-value of the chi-square test results of "whether nausea and vomiting" and "whether abdominal distension and pain" were less than 0.05, so there was a significant difference between the two drugs. The p-value of "whether dizziness, lightheadedness and headache", "whether drowsiness and fatigue" and "whether other discomfort" are all greater than 0.05, so there is no significant difference between the two drugs.

Table 2: Calculated results of the chi-square test

name	chi-square value	p	freedom
whoosh	13.514691	0.000236	1
body movement	7.278738	0.006977	1
Other intraoperative adverse reactions	12.815257	0.000343	1
nausea and vomiting	11.929730	0.000552	1
drowsiness and fatigue	0.003899	0.950210	1
Dizziness and headaches	2.333728	0.126598	1
bloating and abdominal pain	7.536825	0.006045	1
Other discomfort	0.226608	0.634050	1

3.2 The final computational result of the KNN implementation

The model was initialized based on the Scikit-Learn library according to the most parameters above, with "choking", "body movement", "other intraoperative adverse reactions", "nausea and vomiting", "drowsiness and fatigue", "dizziness and headache", "abdominal distension and abdominal pain", "Other postoperative discomfort" eight as labels to generate the dataset, divide the dataset in a ratio of 4:1, and use the training set to train the initialized KNN model. The KNN model was used to predict the feature space of the test set, and the common evaluation indexes of the classifier were used to determine whether the model could be used to predict the adverse reactions of the patients during the operation and 24 hours after the operation according to the basic information of the patients and the types of sedative drugs, and the results were obtained as follows in Table 3:

In order to make the results more intuitive, this paper is based on Scikit-Learn to analyze the test results using the confusion matrix and ROC plot, and calculate the AUC value predicted by KNN on the test set, and found that the number on the diagonal line is relatively large, which indicates that the model's classification effect is better, but the prediction for the "No" category is slightly less effective than the "Yes" category. The ROC curve shows a high rate of true positives, and in summary, the KNN model performs well. In this paper, we believe that this model can be used to predict the occurrence of adverse reactions during the operation and 24 hours after the operation.

Table 3: Evaluation of KNN test results

label name	Test Options	Accuracy	Recall	F1 Score	Support Rate
whoosh	No	1.00	0.93	0.96	247
	Yes	0.93	1.00	0.96	230
body movement	No	1.00	0.92	0.96	224
	Yes	0.93	1.00	0.96	228
Other intraoperative adverse reactions	No	1.00	0.96	0.98	250
	Yes	0.93	1.00	0.96	230
nausea and vomiting	No	0.98	0.88	0.93	228
	Yes	0.88	0.98	0.93	210
drowsiness and fatigue	No	0.99	0.86	0.92	232
	Yes	0.86	0.99	0.92	207
Dizziness and headaches	No	1.00	0.87	0.93	221
	Yes	0.88	1.00	0.94	221
bloating and abdominal pain	No	1.00	0.92	0.96	224
	Yes	0.93	1.00	0.96	228
Other discomfort	No	1.00	0.97	0.99	245
	Yes	0.97	1.00	0.99	241

4. Conclusions

In this paper, statistical analysis and statistical inference of the data set from multiple perspectives, first of all, this paper requires about intraoperative and postoperative 24h adverse reactions to determine whether there is a significant difference between the new drug group and the original drug group. Secondly, this paper carries out data cleaning, feature scaling and feature coding for the sample data, and since the adverse reactions are categorical features after coding, the qualitative method based on multivariate visual analysis and the quantitative method based on chi-square test are used to explore the differences between different drug groups for the adverse reactions. In this paper, an effective classification model is required for the prediction of adverse reactions in patients. In this paper, based on the dataset after data preprocessing, the distribution of labels is firstly analyzed and the dataset with serious imbalance in the proportion of labels is up-sampled. Then the data is divided into training set and test set proportionally. Then the training set was feature extracted based on decision tree, and the trained decision tree model was used to realize the prediction of two major categories and eight groups of adverse reactions. Finally, the performance of the model on the test set is evaluated based on the confusion matrix, ROC plot.

References

[1] SHEN Qing, ZHANG Lianzeng. A new approach to bank credit risk identification: a combined SVM-KNN model[J]. Research on Financial Regulation, 2020, (07):23-37.

[2] Sarkar S , Vinay S , Maiti J .Text mining based safety risk assessment and prediction of occupational accidents in a steel plant[C]//International Conference on Computational Techniques in Information & Communication Technologies.IEEE, 2016.DOI:10.1109/ICCTICT.2016.7514621.

[3] Wang Dapeng, Yan Su, Wang Nan, et al. Analysis of intelligent fire protection industry based on chi-square test and rank sum test[J]. Fire Science and Technology, 2022, 41(11):1594.

[4] Haochen Wang, Changlun Zhang, Mingliang Lai. Research on deep learning based point cloud upsampling algorithm [J]. JournalofImageandSignalProcessing, 2023, 12:21.

[5] Cao Qian. Research on multispectral dimensionality reduction algorithm based on second-order polynomial regression and weighted principal component analysis[J].Optical Technique, 2023, 49(2):250-256.

[6] Quancheng Z , Jingbin H .Research on Data Mining of Physical Examination for Risk Factors of Chronic Diseases Based on Classification Decision Tree[J]. 2021.DOI:10.1109/ICSP51882.2021.9408682.
[7] K. Zhang, K. Zhang. Research on slope stability prediction based on LightGBM algorithm[J]. Chinese Journal of Safety Science, 2022, 32(7):113.
[8] Yang Qiang, Feng Yan, Guan Li, Wu Wenyu, Wang Sichen, Li Qiangyu. XBand Radar Attenuation Correction Method Based on LightGBM Algorithm[J].RemoteSensing, 2023, 15(3).
[9] Lee S , Choi K , Yoo D .Predicting the Insolvency of SMEs Using Technological Feasibility Assessment Information and Data Mining Techniques[J].Sustainability, 2020, 12(23):9790.DOI:10.3390/su12239790.
[10] Harsha P , Manikanta V , Kumara S S ,et al.Classification and Regression Tree - Based Analysis for the Prediction and Mapping of Funding Pattern[J].Srels Journal of Information Management, 2012.
[11] Shuchen Wu, Zongfeng Qi, Jianxun Li. Intelligent global sensitivity analysis based on deep learning[J]. Journal of Shanghai Jiao Tong University, 2022, 56(7):840.
[12] Ren Jiadong, Liu Xinqian, Wang Qian, He Haitao, Zhao Xiaolin. A multi-layer intrusion detection method based on KNN outlier detection and random forest [J]. Computer Research and Development, 2019, 56(03):566-575.