Decision Tree-Based Data Stratification Method for the Minimization of the Masking Effect in Adverse Drug Reaction Signal Detection

Jianxiang Wei 1,2,*, Lu Cheng 3, Pu Han 1, Yunxia Zhu 3 and Weidong Huang 1,2

Abstract: Data masking is an inborn defect of measures of disproportionality in adverse drug reactions signal detection. Some improved methods which used gender and age for data stratification only considered the patient-related confounding factors, ignoring the drug-related influencing factors. Due to a large number of reports and the high proportion of antibiotics in the Chinese spontaneous reporting database, this paper proposes a decision tree-stratification method for the minimization of the masking effect by integrating the relevant factors of patients and drugs. The adverse drug reaction monitoring reports of Jiangsu Province in China from 2011 to 2018 were selected for this study. First, the age division interval was determined based on the statistical analysis of antibiotic-related data. Secondly, correlation analysis was conducted based on the patient’s gender and age respectively with the drug category attributes. Thirdly, the decision tree based on age and gender was constructed by the J48 algorithm, which was used to determine if drugs belonged to antibiotics as a classification label. Fourthly, some performance evaluation indicators were constructed based on the data of drug package inserts as a standard signal library: recall, precision, and F (the arithmetic harmonic mean of recall and precision). Finally, four experiments were carried out by means of the proportional reporting ratio method: non-stratification (total data), gender-stratification, age-stratification and decision tree-stratification, and the performance of the signal detection results was compared. The experimental results showed that the decision tree-stratification was superior to the other three methods. Therefore, the data-masking effect can be further minimized by comprehensively considering the patient and drug-related confounding factors.

Keywords: decision tree; data masking effect; adverse drug reaction; signal detection; data stratification

1. Introduction

Adverse drug reaction (ADR) refers to the harmful effects and negative reactions of qualified drugs without any relation to the purpose of the drug under normal usage and normal dosage, that is, discomfort symptoms or pathogenic reactions [1]. Spontaneous reporting system (SRS) is the main data source of risk reassessment of post-marketing drugs in various countries. ADR signal detection is the main work of pharmacovigilance, which is to explore the relationship between drug and adverse event (AE) by using statistical analysis or data-mining methods. The current methods of signal detection used in many countries are based on disproportionality analysis (DPA). These methods are mainly used to calculate whether the reported frequency of adverse reactions of a certain drug in the database is higher than the expected reported frequency of all drugs, and to qualitatively measure the correlation between drugs and adverse reactions. Methods include proportion reporting ratio (PRR) [2], reporting odds ratio (ROR) [3], information component...
DPA has been widely used in various countries, and plays a positive role in the pharmacovigilance of post-marketing drugs. However, DPA has an inherent defect, the masking effect, which can be caused by data deviation, competition deviation, confounding factors, low data quality and so on [7]. The most common masking effect is due to overreporting. That is, assuming that there is a large number of reports of drug A and adverse reaction B in the data, the ADR combination may reduce the contact strength generated by the combination of drug A and other adverse reactions, or the contact strength generated by the combination of other drugs and adverse reaction B, so that a part of the valuable signals will be masked and the detection efficiency will be reduced. Many scholars have proposed methods to address this issue. The commonly used method for the minimization of the masking effect is to adopt a data stratification strategy, which is to stratify the data according to the different classifications of certain variables that need to be controlled, and then estimate the relationship between a certain exposure factor and a certain AE. Dodd et al. [8] investigated the impact of age stratification and age adjustment on the performance of PRR and EBGM based on pediatric data from the US FDA adverse event reporting system. They thought that stratification could reveal new associations, and therefore suggest recommendations as to when drug use is age-specific or when an age-specific risk is suspected. Zeinoun et al. [9] evaluated the impact of stratification, the comparator dataset and the potential for masking, and conducted a semi-quantitative assessment by comparing the changes in the disproportionality scores and the number of vaccine-event pairs that exceeded an arbitrary threshold as a measure of the impact of any of these choices. The results showed that stratification by age and region has a significant impact. Hopstadius et al. [10] studied the impact of potential confounding factors based on stratification—such as gender, age, reporting quarter—and compared the changes in IC values before and after stratification. Mickael et al. [11] combined the method of removing the masking factor and the stratification of the confounding factor, and proposed a method based on the competition index (ComIn) to identify the disproportionate strength of competitors. They compared the competition factor with the masking factor (MF) and the masking ratio (MR), and found that the ComIn can minimize the competition bias. However, when stratifying confounding factors, these researchers only considered the two major confounding factors, age and gender, and ignored the influence of drug category. Therefore, the improvement effect of stratification was not obvious in the results.

Classification is an important subject in data mining. In recent years, researchers have begun to use the decision tree model to classify datasets. In order to verify the performances of data mining methodology in the evaluation of cardiovascular risk in athletes, and whether the results may be used to support clinical decision making, Barbieri et al. [12] used resampling to balance positive/negative class ratios, and used a decision tree and logistic regression to classify individuals according to risk, so as to improve balance in the classification of medical datasets. The results showed that resampling by decision tree can be effectively applied to biomedical data in order to optimize clinical decision making, and—at the same time—minimize the amount of unnecessary examinations.

Since the mass production and use of penicillin by American pharmaceutical companies in 1942, hundreds of antibiotics have been synthesized. Antibiotic resistance affects the development of the world economy and threatens public health. Antibiotic-induced reactions account for half of spontaneous reports of adverse events in China [13]. Due to the high proportion of antibiotics in the Chinese spontaneous reporting database (CSRD), this paper proposes a decision tree-stratification method for the minimization of the masking effect in ADR signal detection by integrating the relevant factors of patients (age, gender) and drug category (whether or not antibiotics).
2. Methods

2.1. Data Source

The ADR monitoring reports of Jiangsu Province in China between January 1, 2011 and 31 December 2018 were selected for this study, including 754,882 reports. The original dataset includes the fields of drug category, drug name, ADR name, gender and age. The object to be predicted in this study is the combination of drugs and adverse reactions. Due to the lack of age, gender and other information in some reports, a study dataset is obtained after deleting the records with missing information and standardizing the terms of drug name and ADR name. The study dataset contains a total of 751,606 ADR records, which included 1252 drugs, 1262 ADRs and 64,846 drug-event combinations (DECs).

A reference database was established to evaluate the performance of signal detection, including 53,774 kinds of DECs collected from the package insert of drugs.

2.2. Stratification Strategy

The traditional method based on data stratification selects only a single confounding factor, such as age or gender. The reason gender can become a confounding factor is that men and women have many differences in physiological organs and body structures, such as height and weight, hormone secretion, fat distribution, etc., which can change the efficacy of drugs and affect the adverse reactions to drugs. The same is true for age. Due to the large proportion of antibiotic-related reports in CSRD and the complex relationship between age, gender and antibiotics (for example, metronidazole is mainly used for female gynecological diseases, and quinolone is mainly used for the elderly), this paper proposes a stratification method based on a decision tree by integrating the relevant factors of patients and drugs. The specific steps include:

(1) Determining the age division intervals by using the cumulative distribution of antibiotic-related reports based on the patient’s age;
(2) $\chi^2$ was used to analyze the correlation between age and drug category (“Antibiotics” or “Non-antibiotics”), as well as gender and drug category;
(3) Data stratification was conducted by a classification algorithm, based on a decision tree by using drug category as the class label, and the two confounding factors of “gender” and “age” as the stratification conditions;
(4) DPA was performed on the multiple datasets generated by the decision tree;
(5) The performance of this method is compared with that of non-stratification, gender-stratification and age-stratification. Classification performances were assessed by means of precision, recall and F-measure.

The overall research framework is shown in Figure 1.

![Figure 1. Overall research framework.](image-url)
2.3. Decision Tree

The decision tree is a widely used technology in classification algorithms. Compared with other algorithms, the classification accuracy of the decision tree is competitive, and the efficiency is also very high. The classification model representation obtained by this algorithm is in the form of a tree. Among them, the more commonly used algorithm is the C4.5 [14]. The J48 algorithm is the application of the C4.5 algorithm in Waikato Environment for Knowledge Analysis (WEKA) [15]. Based on the recursive strategy from top to bottom, the algorithm uses information-gain ratio as attribute separation [16], searches for a property field with a maximum amount of information, establishes a decision tree root node, and then generates a branch for each possible attribute value, dividing instances into multiple subsets, where each subset corresponds to a branch of the root node. The process repeats recursively on each branch. Recursion stops when all instances have the same classification or when the Gini value is less than a certain point with no new leaf nodes generated.

In the design of the algorithm, a good pruning process is considered and added, making it easy for users to understand the classification rules and which has good accuracy in data processing. It has attracted the attention of data mining researchers and solved many practical application problems.

2.4. Signal Detection Method

The calculation of DPA is based on the $2 \times 2$ contingency table shown in Table 1. If a specific DEC in the database is significantly higher than the background frequency in the entire database and reaches a certain threshold, it is considered a positive signal.

Table 1. Two-by-two contingency table.

|                      | The Target ADR | The Other ADRs |
|----------------------|----------------|----------------|
| the target drug      | A              | B              |
| the other drugs      | C              | D              |

A represents the number of reports caused by the target drug and the target ADR, B represents the number of reports of the other ADRs caused by the target drug, C represents the number of reports of the target ADR caused by the other drugs, and D represents the number of reports of the other ADRs caused by the other drugs.

The PRR method is adopted for ADR signal detection. Based on Table 1, the calculation formula is as follows:

$$\text{PRR} = \frac{A}{A + B} \cdot \frac{C}{C + D}$$  \hspace{1cm} (1)

A positive signal is an output if $\text{PRR} \geq 2$.

2.5. Performance Evaluation

As an objective standard, the reference database is used for performance evaluation. If the signal result is positive and appears in the reference database, it is denoted as a true positive (TP), otherwise it is a false positive (FP). If the signal result is negative and appears in the reference database, it is denoted as a false negative (FN). Precision ($P$) is the proportion of true positive in all predicted positives, and can be defined as follows:

$$P = \frac{TP}{TP + FP}$$  \hspace{1cm} (2)

Recall ($R$), the proportion of true positive in all actual positives, is defined as follows:

$$R = \frac{TP}{TP + FN}$$  \hspace{1cm} (3)
F-Measure ($F$) is the arithmetic harmonic mean of Precision and Recall, is defined as follows:

$$F = \frac{2 \cdot P \cdot R}{P + R} \quad (4)$$

The larger the $F$ value, the higher the performance overall, and the more ideal the effect of minimizing data masking.

3. Results

3.1. Data Analysis

Due to the high proportion of antibiotics-related reports in CSRD, this paper analyzes the correlation between age and drug category, and gender and drug category. The proportion of ADR reports for Antibiotics and Non-antibiotics in the study dataset is given in Table 2.

Table 2. The proportion of reports for antibiotics and non-antibiotics.

| Drug Category | Number of Reports (Proportion) |
|---------------|--------------------------------|
| Non-antibiotics | 392,113 (52.17%)             |
| Antibiotics     | 359,493 (47.83%)             |
| Total           | 751,606                       |

From Table 2, we can see that the proportion of reports for antibiotics accounted for 47.83% of the total reports.

Previous related literature does not have a unified division for the confounding factor of age, they are all subjective individual divisions [17]. Therefore, this paper uses cumulative distribution graphs of the antibiotics-related ADR reports for division of age range, where the length of the age interval is set to five years. The resulting cumulative distribution diagram is shown in Figure 2.

It can be seen from Figure 2 that the reported number of antibiotics before and after the age of 60 tends to be flat, while the reported number between 20 and 60 years old has increased significantly. Therefore the age of patients in the data set was discretized into three values: younger than 20 years old is “Young”; 20–60 years old is “Middle”; and older than 60 years is “Senior”.

![Cumulative distribution graph](image_url)

Figure 2. Age cumulative distribution of spontaneous reporting of antibiotics.
Correlation analysis between gender and drug category was conducted based on Chi-square. The \( \chi^2 \) value is 343.42, which is far greater than the critical value 3.84 when the degree of freedom is 1 and the significance level is 95%. In the same way, the \( \chi^2 \) value between age and drug category based on Chi-square is 36,435.81, which is much larger than the critical value of 5.99 when the degree of freedom is 2 and the significance level is 95%. Therefore, the drug category is closely related to gender and age. The contingency table of drug category with gender and age is shown in Table 3.

### Table 3. The contingency table of drug category with gender and age.

|         | Male       | Female     | Young (\( \leq 20 \)) | Middle (21–60) | Senior (\( \geq 61 \)) |
|---------|------------|------------|------------------------|----------------|--------------------------|
| Antibiotics | 170,739    | 188,754    | 25,979                 | 225,676        | 140,458                  |
| Non-antibiotics | 194,618    | 197,495    | 70,145                 | 209,886        | 79,462                   |
| \( \chi^2 \) | 343.42     |            |                        | 36,435.81      |                          |

3.2. Decision Tree

The J48 classification algorithm in WEKA software is used to construct the decision tree (Figure 3).

![Decision Tree](image)

**Figure 3.** Decision tree.

This decision tree realizes the optimal division of data by using age and gender as conditions and drug category as a class label. The study data set is divided into four data sets:

1. The data meeting the condition “Age = Senior” are classified into “Non-antibiotics class”, including 219,920 reports. The accuracy rate is 63.87%.
2. The data meeting the condition “Age = Young” are classified into “Antibiotics class”, including 96,124 reports. The accuracy rate is 50.28%.
3. The data meeting the condition “Age = Middle” and “Gender = Male” are classified into “Non-antibiotics class”, including 193,708 reports. The accuracy rate is 54.42%.
4. The data meeting the condition “Age = Middle” and “Gender = Female” are classified into “Antibiotics class”, including 241,854 reports. The accuracy rate is 63.87%.

3.3. Performance Evaluation

PRR was used to detect signals in datasets (D1, D2, D3, D4) generated by non-stratification, gender-stratification, age-stratification and decision tree-stratification. Signal sets (S1, S2, S3, S4) were generated (See Figure 1). The comparison results are shown in Table 4.
Table 4. Comparison of detection results of different stratification methods.

| Method                  | TP    | FP    | FN    | R     | P     | F     |
|-------------------------|-------|-------|-------|-------|-------|-------|
| non-stratification      | 7293  | 28,781| 6661  | 52.26%| 20.22%| 29.16%|
| gender-stratification   | 8369  | 34,236| 5585  | 59.98%| 19.64%| 29.59%|
| age-stratification      | 8977  | 35,591| 4977  | 64.33%| 20.14%| 30.68%|
| decision tree-stratification | 9605  | 38,225| 4349  | 68.83%| 20.08%| 31.09%|

As can be seen from Table 4, F of the three stratification methods has been improved. Among them, the F value obtained by decision tree-stratification is the largest, and the F value of decision tree-stratification is 1.93% higher than that of non-stratification. In addition, the R obtained by decision tree-stratification is significantly improved, which is 16.57% higher than that of non-stratification.

4. Discussion

4.1. Discussion of Methods

Unlike other countries, China has a large population and is one of the largest manufacturers and users of antibiotics. The more kinds of antibiotics that are used, the more ADRs are produced [18]. In this study dataset, antibiotic-related reports accounted for 47.83% of the total reports. Of all 1262 different ADRs, 969 were caused by antibiotics, accounting for 77% of the total. The essence of the signal masking effect is that when a group of DECs are reported too frequently, other DECs associated with these drugs will produce signal delay or direct masking phenomenon [19]. Therefore, drug category is considered an important confounding factor in CSRD, which could cause the data masking effect. The method based on a decision tree is to minimize the signal masking effect by separating antibiotics from other drugs.

In the study data set, among the adverse reaction reports of Epirubicin, 1143 were female and 317 were male, as Epirubicin was mainly used in the treatment of female breast cancer. For Cefpiramide, there are 1851 cases for the young, 1299 cases for the middle, and only 380 cases for the senior. Chi-square analysis revealed a strong correlation between drug category and gender, as well as drug category and age. Therefore, previous studies were also used for reference in our method, and gender and age were considered as two confounding factors. The proposed method integrated the information of patients and drug categories, so it showed advantages in signal detection performance.

In addition, the age interval division in previous studies was subjective and there was no unified standard. An objective method was proposed to discretize age data based on the cumulative distribution of antibiotic-related reports with age.

4.2. Discussion of Results

In this study, four performance comparison experiments are conducted on the same dataset: non-stratification, gender-stratification, age-stratification and decision tree-stratification. The first three are previous methods and the last one is proposed in this study. Experimental results show that the proposed method improves the performance of signal detection in different degrees compared with the previous three methods. Specifically, compared with non-stratification, the R obtained by decision tree-stratification increased greatly from 52.26% to 68.83%, an increase of 16.57%, when the P remains basically unchanged. In addition, the value of F-measure increased from 29.16% to 31.09%, an increase of 1.93%. Moreover, the F-measure of our method was higher than that of age-stratification and gender-stratification, which proved the effectiveness of our method.

4.3. Limitations

First of all, the accuracy of the decision tree algorithm adopted in this paper is only 58.22%. If higher classification accuracy is needed, more classification attributes need to be added. However, this will also lead to excessive stratification, which is not good for minimizing the masking signal [20]. Therefore, optimizing the algorithm to improve
the accuracy of classification without adding more attributes is what we need to do in the future.

Secondly, while the signal detection method adopted in this paper was PRR, some other methods, such as ROR, IC and MHRA, also need further tested.

5. Conclusions

Data stratification can effectively reduce the data masking effect. The traditional methods were based on the patient’s age and gender and other confounding factors, ignoring the drug information. Because there were a large number of reports related to antibiotics in CSRD, we proposed a decision tree-stratification method for the minimization of the masking effect by integrating the relevant information of patients and drug categories, and achieved better performance in ADR signal detection.

Author Contributions: J.W. conceived of the study. L.C. participated in data processing and experimental simulation. J.W. and L.C. participated in data analysis, and drafting of the manuscript. P.H., Y.Z. and W.H. were involved in critically revising the manuscript and reviewing experiment results. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the Major Project of Philosophy and Social Science Research in Jiangsu Universities (2020SJZDA102).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All ADR spontaneous reporting data in this study is licensed by the CFDA. The datasets are not publicly available due to the policy of confidentiality of the CFDA but are available from the corresponding author on reasonable request and with permission of the CFDA.

Conflicts of Interest: The authors declare no conflict of interest.

References
1. Edwards, I.R.; Aronson, J.K. Adverse drug reactions: Definitions, diagnosis, and management. Lancet 2000, 356, 1255–1259. [CrossRef]
2. Slattery, J.; Alvarez, Y.; Hidalgo, A. Choosing thresholds for statistical signal detection with the proportional reporting ratio. Drug Saf. 2013, 36, 687–692. [CrossRef] [PubMed]
3. Bate, A.; Leufkens, H.G.M.; Lindquist, M.; Orre, R.; Egberts, T. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol. Drug Saf. 2002, 11, 3–10. [CrossRef]
4. Bate, A.; Lindquist, M.; Edwards, I.R.; Orre, R. A data mining approach for signal detection and analysis. Drug Saf. 2002, 25, 393–397. [CrossRef] [PubMed]
5. Hauben, M.; Bate, A. Decision support methods for the detection of adverse events in post-marketing data. Drug Discov. Today 2009, 14, 343–357. [CrossRef] [PubMed]
6. Martin, D.; Menschik, D.; Bryant-Genevier, M.; Ball, R. Data mining for prospective early detection of safety signals in the vaccine adverse event reporting system (VAERS): A case study of febrile seizures after a 2010–2011 seasonal influenza virus vaccine. Drug Saf. 2013, 36, 547–556. [CrossRef] [PubMed]
7. Almenoff, J.; Tonning, J.M.; Gould, A.L.; Szarfman, A.; Hauben, M.; Ouellet-Hellstrom, R.; Ball, R.; Hornbuckle, K.; Walsh, L.; Yee, C.; et al. Perspectives on the use of data mining in pharmacovigilance. Drug Saf. 2005, 28, 981–1007. [CrossRef] [PubMed]
8. Caster, O.; Norén, G.N.; Madigan, D.; Bate, A. Large-scale regression-based pattern discovery: The example of screening the WHO global drug safety database. Stat. Anal. Data Min. ASA Data Sci. J. 2010, 3, 197–208. [CrossRef]
9. Zeinoun, Z.; Seifert, H.; Verstraten, T. Quantitative signal detection for vaccines: Effects of stratification, background and masking on GlaxoSmithKline’s spontaneous reports database. Hum. Vaccines 2009, 5, 599–607. [CrossRef] [PubMed]
10. Hopstadius, J.; Norén, G.N.; Bate, A.; Edwards, I.R. Impact of stratification on adverse drug reaction surveillance. Drug Saf. 2008, 31, 1035–1048. [CrossRef] [PubMed]
11. Mickael, A.; Francesco, S.; Ismail, A.; Philip, R.; Nicholas, M.; Bernard, M.; Tubert-Bitter, P.; Pariente, A. A method for the minimization of competition bias in signal detection from spontaneous reporting databases. Drug Saf. 2016, 39, 251–260. [CrossRef]
12. Barbieri, D.; Chawla, N.; Zaccagni, L.; Grgurinović, T.; Sarac, J.; Čoklo, M.; Missoni, S. Predicting cardiovascular risk in Athletes: Resampling improves classification performance. Int. J. Environ. Res. Public Health 2020, 17, 7923. [CrossRef] [PubMed]
13. Li, Y.; Xu, J.; Wang, F.; Wang, B.; Liu, L.; Hou, W.; Fan, H.; Tong, Y.; Zhang, J.; Lu, Z. Over-prescribing in China, driven by financial incentives, results in very high use of antibiotics, injections, and corticosteroids. Health Aff. 2012, 31, 1075–1082. [CrossRef] [PubMed]
14. Quinlan, J.R. C4.5: Programs for Machine Learning; Morgan Kaufmann Publishers: San Francisco, CA, USA, 1993.
15. Garner, S.R. WEKA: The Waikato Environment for Knowledge Analysis; Department of Computer Science, University of Waikato: Hamilton, OH, USA, 1995.
16. Li, C.B.; Li, S.J. Data Warehouse and Data Mining Practice and Application; Publishing House of Electronics Industry: Beijing, China, 2014; pp. 328–330.
17. Zhao, L.; Ye, X.; Wang, C.; Qian, W.; Du, W.; He, J. Application of stratification analysis in adverse drug reaction signal detection. *Chin. J. Pharmacovigil.* 2011, 8, 158–160.
18. Walley, J.D.; Zhang, Z.; Wei, X. Antibiotic overuse in China: Call for consolidated efforts to develop antibiotic stewardship programmes. *Lancet Infect. Dis.* 2021, 21, 597. [CrossRef]  
19. Wang, H.-W.; Hochberg, A.M.; Pearson, R.K.; Hauben, M. An experimental investigation of masking in the US FDA adverse event reporting system database. *Drug Saf.* 2010, 33, 1117–1133. [CrossRef] [PubMed]  
20. Hopstitialus, J.; Norén, G.N.; Bate, A.; Edwards, I.R. Stratification for spontaneous report databases. *Drug Saf.* 2008, 31, 1145–1147. [CrossRef] [PubMed]