Chinese Journal of Electronics
Vol.26, No.3, May 2017

Prediction of Drug-Drug Interactions Based on Multi-layer Feature Selection and Data Balance*

YUE Kejuan[1,2,3], ZOU Beiji[1,2] ,WANG Lei[1,2], LI Xiao[1,2], ZENG Min[1,2] and WEI Faran[1,2]

(1. *School of Information Science and Engineering, Central South University, Changsha 410083, China*)

(2. *Center for Ophthalmic Imaging Research, Central South university, Changsha 410083, China*)

(3. *School of Information Science and Engineering, Hunan First Normal University, Changsha 410205, China*)

Abstract — Drug-drug interactions (DDIs) occur when two drugs react with each other, which may cause unexpected side effects and even death of the patient. Methods that use adverse event reports to predict unexpected DDIs are limited by two critical yet challenging issues. One is the difficulty of selecting discriminative features from numerous redundant and irrelevant adverse events for modeling. The other is the data imbalance, *i.e.*, the drug pairs causing adverse effects are far less than those not causing adverse effects, which leads to poor accuracy of DDIs detection. We propose a multi-layer feature selection method to select discriminative adverse events and apply an over-sampling technique to make the data balanced. The experimental results show that the validation accuracy of positive DDIs on the Canada Vigilance Adverse Reaction Online Database increases to two times, and 110 DDIs are identified on the drug interactions checker of Drugs.com in USA.

Key words — Adverse event reports, Drug-drug interactions (DDIs), Feature selection, Data balance.

I. Introduction

Drug-drug interactions (DDIs) may have adverse effects on the patients and even lead to death of the patients[1−3]. Predicting unexpected DDIs as early as possible is of great significance to clinical practice. Some algorithms make use of drugs' pleiotropic interactions to predict off target effects[4−9]. However, these effects are not necessarily adverse. Some DDIs can be discovered by analyzing molecular targets and metabolizing enzymes of drugs. For example, when two drugs are metabolized by the same enzyme (*e.g.*, CYP3A4), it may lead to unexpected blood levels[10−13]. Santiago Vilar predicted potential DDIs by establishing interaction profile similarity matrix of known DDIs[14]. These approaches are based on theory rather than clinical data, which make them less reliable.

Quantitative signal detection methods are primarily developed to detect drug adverse event signals from Adverse event reporting system(AERS) in clinical practice[15−16], but limited by underreporting of the unexpected events[17−20]. In order to address the issue of underreporting, Tatonetti *et al.* proposed a machine learning framework to predict DDIs[21], but there are still two problems to be resolved: 1) For feature selection, a Fisher's exact test is used to determine the enrichment of each feature, which only considers the associations between features and label variables but not the correlations of features. This may miss effective features and reduce the accuracy of DDIs prediction. 2) The data set for modeling is imbalanced, *i.e.*, the number of positive samples (really having adverse effects) is much less than that of the negative samples (having no adverse effects), which will lead to low prediction accuracy for positive drug pairs. It is disadvantageous for the detection of DDIs, because the cost of misclassifying positive drug pairs as negative is much greater than misclassifying negative drug pairs as positive. In this paper, we propose a multiple-layer feature selection method to select discriminative features with lower computational complexity, and then balance the data set with an oversampling technique. The experimental results demonstrate that our method ensures high accuracy for true positive drug pairs without severely jeopardizing the overall accuracy, and more predicted DDIs are also identified.

II. The Proposed Method

Based on framework provided by Tatonetti[21], we investigate drug interactions related to cholesterol. Fre-

*Manuscript Received July 08, 2016; Accepted Oct. 17, 2016. This work is supported by the National Natural Science Foundation of China (No.61573380, No.61672542), and the Scientific Research Fund of Hunan Provincial Education Department (No.13C143).

DOI:10.1049/cje.2017.04.005

quency matrices for single drug and drug pairs are constructed respectively (Fig.1 and Fig.2). In Fig.1, each row corresponds to one drug (*i.e.*, a sample), each column corresponds to an adverse effect (*i.e.*, a feature), and the element of the matrix represents the reported frequencies of the adverse events. The last column is the class label identified by manual curation, 1 indicates that the drug is expected to cause cholesterol related adverse effect and 0 is not. In Fig.2, each row corresponds to a drug pair (*i.e.*, a sample), each column corresponds to an adverse effect (*i.e.*, a feature), and the element of the matrix represents the reported frequencies of the adverse events. The last column is the class label to be predicted.

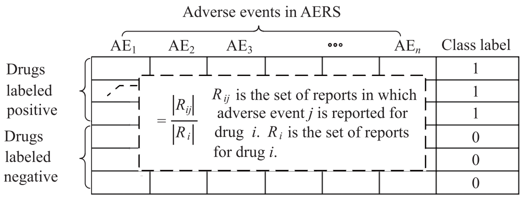
Adverse events in AERS
AE_1
AE_2
AE_3
•••
AE_n
Class label
Drugs labeled positive
Drugs labeled negative
$= \frac{|R_{ij}|}{|R_i|}$
R_{ij} is the set of reports in which adverse event j is reported for drug i. R_i is the set of reports for drug i.
1
1
1
0
0
0

Fig. 1. The frequency matrix for the single drug

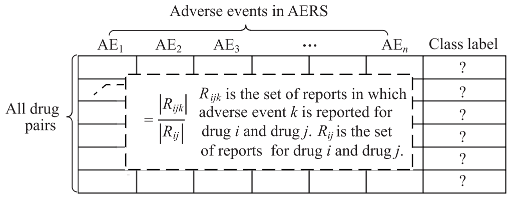
Adverse events in AERS
AE_1
AE_2
AE_3
···
AE_n
Class label
All drug pairs
$= \frac{|R_{ijk}|}{|R_{ij}|}$
R_{ijk} is the set of reports in which adverse event k is reported for drug i and drug j. R_{ij} is the set of reports for drug i and drug j.
?
?
?
?
?
?

Fig. 2. The frequency matrix for the drug pairs

After constructing frequency matrices, we propose a multiple-layer feature selection method to select predictive features and an oversampling method to balance the data set. Finally we learn a logistic classifier on single-drug frequency matrix and apply it to drug-pairs frequency matrix to predict potential DDIs. The work is illustrated by Fig.3.

1. Multiple-layer feature selection method

The task of feature selection is to choose a subset of features most relevant with the class label[22]. Least absolute shrinkage and selection operator (LASSO) is a method for linear model estimation with L1-regularization and also used for feature selection by shrinking some coefficients to zero[23]. The time complexity of the ordinary LASSO is O (np min $\{n, p\}$), here n is the number of samples and p is the number of features. For DDIs prediction model in this paper, the number of the features is more than 7000, which brings the problem of computational complexity. So we combine LASSO with Correlation based feature selector (CFS) to a multiple-layer method so that we can select discriminative features with lower computational complexity. First we use CFS to get a smaller feature subset, namely reduce the value of p, and then we apply LASSO to select features from the subset. Fig.4 illustrates the multiple-layer feature selection method. The details are listed as follows:

Step 1: Construct frequency matrices for single drug and drug pairs respectively, and take the adverse events frequency as the features.

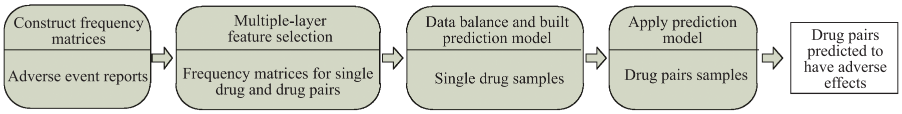
Construct frequency matrices
Adverse event reports
Multiple-layer feature selection
Frequency matrices for single drug and drug pairs
Data balance and built prediction model
Single drug samples
Apply prediction model
Drug pairs samples
Drug pairs predicted to have adverse effects

Fig. 3. The data flow diagram of the method

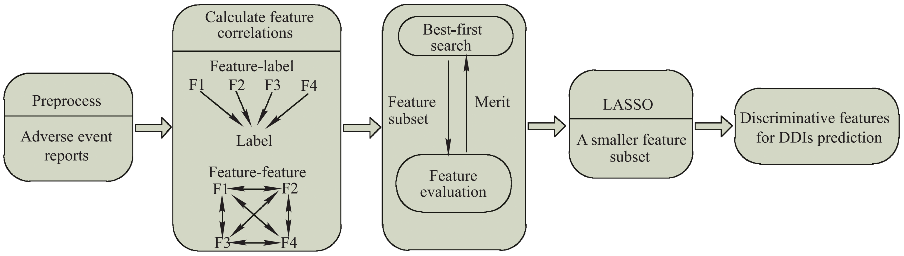

Fig. 4. The procedure of multiple-layer feature selection

Step 2: Calculate feature-label and feature-feature correlations.

Step 3: Use the best-first strategy to search the space of feature subsets, and iteratively expand the subset according to the rule of best Merit value.

Step 4: Use LASSO to select discriminative features from a smaller feature subset of the previous step.

Step 2 and Step 3 are implemented based on the CFS algorithm that measure the merit of feature subsets using a evaluation function[24]. The function is defined as Eq.(1).

$$Merit_{s_k} = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{1}$$

Here, k is the number of features, s_k is a feature subset consisting of k features, $\overline{r_{cf}}$ is the average value of all feature-label correlations, and $\overline{r_{ff}}$ is the average value of all feature-feature correlations. The numerator of equation indicates the predictive ability of features and the denominator indicates the redundancy among the features. A feature subset having greater feature-label correlations and less feature-feature correlations will achieve the best Merit value.

The objective of LASSO is defined as Eq.(2):

$$\arg\min_{\boldsymbol{\beta}}\{\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j)^2\} \text{ subject to} \sum_{j=1}^{p}|\beta_j|\leq t \tag{2}$$

Here $\boldsymbol{\beta} =< \beta_0, \beta_1, \cdots, \beta_p >$ is the coefficient vector, n is the size of samples and p is the number of features, x_{ij} is the j^{th} feature of the i^{th} sample, y_i is the observation of the i^{th} sample, t is a parameter that is used to control the amount of shrinkage. With the decrease of t value, some coefficients will be set to zero, which means the corresponding features are pruned from the model. These pruned features are irrelevant (or less relevant) to the observations.

2. Data balance

In our prediction model the negative samples (2526 drugs having no cholesterol related effects) are much more than the positive samples (26 drugs having cholesterol related effects). In that case, classifiers with good overall accuracy may provide low accuracy for positive drugs. It is disadvantageous for the detection of DDIs, because the cost of misclassifying positive drug pairs as negative is much greater than misclassifying negative drug pairs as positive. To deal with imbalance data there are sampling methods and algorithmic methods[25]. The algorithmic methods exploit cost sensitive learning mechanism to improve existing algorithms, which consider the costs of misclassifying samples and make the classifier more accurate for the classification of minority samples. But in many practical situations, it is very difficult to describe the misclassifying costs exactly and an unreasonable cost will decrease the accuracy of the classifier[26]. Sampling methods consist of oversampling and undersampling. In the prediction of DDIs, the quantity of minority samples is too small, so the undersampling methods will cause a serious loss of majority samples information. Therefore, we use Synthetic minority over-sampling technique (SMOTE) to balance the data set, which creates artificial data based on the feature space similarities between existing minority examples. For a sample $\boldsymbol{x}_i$ belonging to the minority class, we randomly choose one of the K-nearest neighbors in the minority class (K is adjustable), compute the vector difference and multiply it with a random number between [0,1], then add this vector to $\boldsymbol{x}_i$[27].It is defined as Eq.(3).

$$\boldsymbol{x}_{new} = \boldsymbol{x}_i + (\hat{\boldsymbol{x}}_i - \boldsymbol{x}_i) \times \delta \tag{3}$$

Here, $\boldsymbol{x}_i$ is a existing minority sample, $\hat{\boldsymbol{x}}_i$ is a K-nearest neighbor belonging to the minority class for $\boldsymbol{x}_i$, δ is a random number between $[0, 1]$, $\boldsymbol{x}_{new}$ is a new artificial sample along the line segment connecting $\boldsymbol{x}_i$ to $\hat{\boldsymbol{x}}_i$. By SMOTE, synthetic samples representing the characters of positive drugs are added to the data set, which will help to improve the prediction accuracy of positive samples.

III. Experiments and Results

1. Data sources

We downloaded 1854669 adverse event reports (the first quarter of 2004 to the first quarter of 2009) from AERS of the U.S. Food and Drug Administration (FDA). We use only two kinds of reports. One is those reporting exactly single drug at least 10 times (2552 single drugs), and the other is those reporting two drugs at least 5 times (6341 drug pairs). The total number of the events is 7109. Three databases are used to measure the performance of our method: 1) MedEffect, Canada Vigilance Adverse Reaction Online Database (from 1965 to 2013). 2) The Veterans Affairs (VA), a list of significant and critical DDIs from the Veterans Affairs Hospital in Arizona. 3) Drugs.com, the most popular, comprehensive and up-to-date source of more than 24,000 prescription drugs information online in USA.

2. Feature selection

Based on CFS, we search the feature space using the best-first strategy, and terminate if consecutive 5 nodes show no improvement, and get a subset of 28 features. Then we continue to use LASSO to select features highly correlated from the 28 features. The tuning parameter is the fraction of L1 norm of the coefficients, which is used to constrain the coefficients. When it is set to 0.15, the cross validation error is the minimum. Five features are screened out, *i.e.*, aspiration pleural cavity abnormal, blood triglycerides increased, coronary artery reocclusion, myalgia and rhabdomyolysis. In order to compare the performance of different features, we train the logistic classifier using 28 features and 5 features respectively on the FDA single drug data set, and adopt 10-fold cross validation. Four measures are used to evaluate the performance. True positive rate (TPR) measures the proportion of positive samples correctly identified, true negative rate (TNR) measures the proportion of negative samples correctly identified. Accuracy measures the proportion of correction for all the samples. Area under the roc curve (AUC) measures the performance of the classifier. Experimental results in Table 1 demonstrate that the model using 5 features performs better than that using 28 features.

Table 1. Multiple-layer feature selection method

	TPR	TNR	Accuracy	AUC
28 features	0.115	0.994	0.985	0.608
5 features	0.269	0.997	0.989	0.867

3. Model validation for data balance

MedEffect database is used for validating the performance of data balance. Since there is no accepted gold standard for identification of drug interactions, we label drug pairs as positive if at least one of the drug pairs is related to the adverse event, which is called strategy of Known single effects (KE). These drug pairs may do not indicate drug interactions, but build confidence that the model can detect true adverse event signals.

We build 3 logistic classifiers respectively using 3 features proposed by Tatonetti[21], 5 features (ours without data balance), and 5 features with data balance (ours with data balance). Then we validate them on MedEffect by the strategy of KE. Synthetic positive samples are created step by step (doubling each time) until the numbers of positive samples and negative samples are approximately equal. Fig.5 shows the True positive rate before and after data balance. K is the number of nearest neighbors used for SMOTE. The number of minority class samples increases to N times by creating synthetic examples. From Fig.5, we can see that with the increase of synthetic samples the true positive rate is improved greatly, which demonstrates that our method with data balance achieve higher accuracy for positive than Tatonetti's method[21] and our method without data balance.

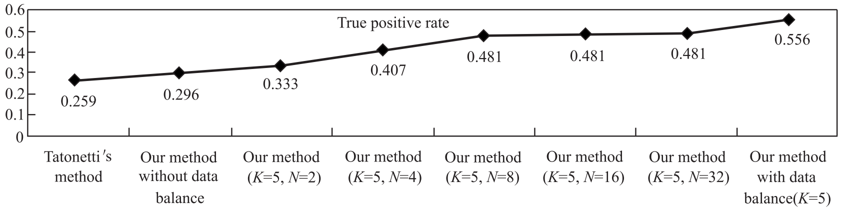

Fig. 5. True positive rate on MedEffect using the strategy of known single effects

Table 2 shows that our method with data balance makes the TPR twice of Tatonetti's[21], while the average precision only decreases 0.3%. It is significant for prediction of DDIs, because the cost of misclassifying positive drug pairs is more severe than that of misclassifying negative drug pairs.

Table 2. Validation on MedEffect using the strategy of known single effects

	TPR	Avg. precision	AUC
Tatonetti's method	0.259	0.974	0.744
Ours without data balance	0.296	0.967	0.753
Ours with data balance	0.556	0.971	0.754

4. DDIs prediction

We learn the logistic classifier on FDA single-drug samples and apply them to drug-pairs samples to predict putative interactions. These interactions are validated using three strategies: 1) Drug pairs whose effect can be explained by the strategy of KE. 2) Drug pairs already known to have clinically significant interactions according to the list of VA. 3) Drug pairs identified by Drugs.com. Table 3 shows the breakdown of the DDI predictions between known single drug effects, established DDIs of VA, and DDIs checked on Drugs.com. We can see that our method can discover more true positive drug pairs by the strategy of KE. By the strategy of VA, Tatonetti's method[21] and our method without data balance predicts the same 7 pairs of drugs, which are also predicted by our method with data balance. Moreover, the 7 pairs of drugs are all can be explained that at least one of the drug pairs has cholesterol-related adverse effects. In addition, we use Drug Interactions Checker on the website Drugs.com to check drug pairs that have not been validated on KE and VA. 6, 18, 110 drug pairs are validated respectively using Tatonetti's method[21], our method without data balance and our method with data balance, which account for 5.4%, 11.8%, 20.7% of the total predictions. These predictions are more significant because their interactions have been identified by clinical practice rather than by the strategy of KE (*i.e.*, at least one of the drug pairs is associated with the adverse event).

Table 3. Drug-pair predictions validated by three strategies.

	Total prediction	Known single effects		Known DDIs of VA		Drugs.com	
		numbers	percent	numbers	percent	numbers	percent
Tatonetti's method	111	75	67.6%	7	6.3%	6	5.4%
Ours without data balance	153	85	55.6%	7	4.6%	18	11.8%
Ours with data balance	531	191	36.0%	8	1.5%	110	20.7%

Table 4. Breakdown of drug-pair predictions into three groups.

	Total predictions	Known single effects	Known DDIs	Novel DDIs
Tatonetti's method	111	75	6	30
Ours without data balance	153	85(71)	18(6)	50
Ours with data balance	531	191(75)	111(6)	229

Table 4 and Fig.6 show the breakdown of the DDI predictions between known single drug effects, known DDIs, and novel interactions predicted by our method. Known DDIs indicates that none of the drug pairs has cholesterol-related adverse effects but there is an interaction between the two drugs known to the DDIs of VA and Drugs.com. In terms of known single drug effects, 71 pairs of drugs detected by Tatonetti's method[21] are also discovered by our method without data balance, and our method with data balance detect all the 75 drug pairs of Tatonetti's method[21]. In terms of known DDIs, 6 drug pairs detected by Tatonetti's method[21] are all identified by our methods.

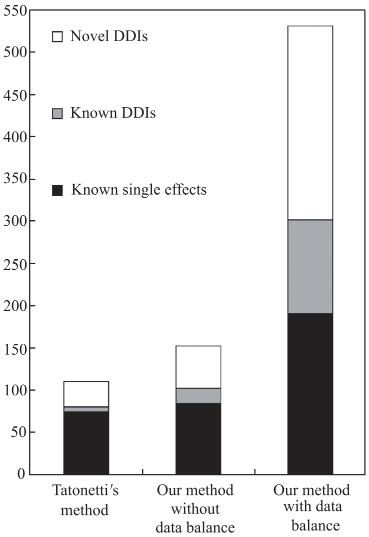

Fig. 6. Drug-drug interactions broken into three groups using Tatonetti's method and our method

IV. Conclusions

Compared with Tatonetti's work[21], the multiple-layer features selection method together with data balance can choose more discriminative features and make the prediction model work more effectively. The area of ROC for the validation on MedEffect has been improved, and the true positive rate also increases a lot.

We build a predictive model on the single-drug samples and apply it to the drug-pairs samples, which is evaluated by two strategies, *i.e.*, known single drug effects and known DDIs. The experimental results show that our method can detect all the drug pairs that are identified by Tatonetti's method[21]. Especially for the known DDIs, our method with data balance achieves maximum in the terms of numbers and detection rate, which is significant for the detection of DDIs, because the strategy of known DDIs represents the real interactions of two drugs. Moreover, our method predicts more novel interactions, which means offering more chances for follow-up research in Electronic medical record (EMR) and other clinical data.

However, there are still some limitations, for example, the single-drug samples need to be labeled manually, the drug name is non-unique (*e.g.* generic name and trade name), and the labeling of true positive drug pairs is incomplete, all these situations will influence the accuracy of validation or prediction.

References

[1] D.D.I. Quinn and R.O. Day, "Drug interactions of clinical importance", *Drug Safety*, Vol.12, No.6, pp.393–452, 1995.

[2] P.G. Mvan der Heijden, E.P. van Puijenbroek, S. Buuren, *et al.*, "On the assessment of adverse drug reactions from spontaneous reporting systems: the influence of underreporting on odds ratios", *Stat Med*, Vol.21, No.14, pp.2027–2044, 2002.

[3] E.L. Olvey, S. Clauschee and D.C. Malone, "Comparison of critical drug-drug interaction listings: the Department of Veterans Affairs medical system and standard reference compendia", *Clin Pharmacol Ther*, Vol.87, No.1, pp.48–51, 2010.

[4] N. Tatonetti, T. Liu and R. Altman, "Predicting drug side-effects by chemical systems biology", *Genome Biol*, Vol.10, No.9, pp.238.1–238.4, 2009.

[5] M. Campillos, M. Kuhn, A.C. Gavin, *et al.*, "Drug target identification using side-effect similarity", *Science*, Vol.321, No.5886, pp.263–266, 2008.

[6] J.C. Adams, M.J. Keiser, L. Basuino, *et al.*, "A mapping of drug space from the view point of small molecule metabolism", *Plos Comput Biol*, Vol.5, No.8, pp.e100047, 2009.

[7] M.J. Keiser, B.L. Roth, B.N. Armbruster, *et al.*, "Relating protein pharmacology by ligand chemistry", *Nat Biotechnol*, Vol.25, No.2, pp.197–206, 2007.

[8] M. Kuhn, C. von Mering, M. Campillos, *et al.*, "STITCH: Interaction networks of chemicals and proteins", *Nucleic Acids Res*, Vol.36, Suppl.1, pp.684–688, 2008.

[9] L. Xie, J. Li, L. Xie, *et al.*, "Drug discovery using chemical systems biology: identification of the protein-ligand binding network to explain the side effects of CETP inhibitors", *PLoS Comput Biol*, Vol.5, No.5, pp.e1000387, 2009.

[10] S.U. Mertens-Talcott, I. Zadezensky, W.V. De Castro, *et al.*, "Grapefruit-drug interactions: Can interactions with drugs be avoided?", *The Journal of Clinical Pharmacology*, Vol.46, No.12, pp.1390–1416, 2006.

[11] P.J. Neuvonen, M. Niemi and J.T. Backman, "Drug interactions with lipid-lowering drugs: Mechanisms and clinical relevance", *Clinical Pharmacology Therapeutics*, Vol.80, No.6, pp.565–581, 2006.

[12] D.G. Bailey, J. Malcolm, O. Arnold, *et al.*, "Grapefruit juice-drug interactions", *British Journal of Clinical Pharmacology*, Vol.46, No.2, pp.101–110, 1998.

[13] K.Y. Yap, W.L. Tay, W.K. Chui, *et al.*, "Clinically relevant drug interactions between anticancer drugs and psychotropic agents", *European Journal of Cancer Care*, Vol.20, No.1, pp.6–32, 2011.

[14] V. Santiago, U. Eugenio, L. Santana, *et al.*, "Detection of drug-drug interactions by modeling interaction profile fingerprints", *PloS One*, Vol.8, No.3, pp.e58321, 2013.

[15] A. Bate and S.J. Evans, "Quantitative signal detection using spontaneous ADR reporting", *Pharmacoepidemiol Drug Saf*, Vol.18, No.6, pp.427–436, 2009.

[16] W. DuMouchel, "Bayesian data mining in large frequency tables, with an application to the fda spontaneous reporting system", *Am Stat*, Vol.53, No.3, pp.177–190, 1999.

[17] A.M. Hochberg and M. Hauben, "Time-to-signal comparison for drug safety data-mining algorithms vs. traditional signaling criteria", *Clin Pharmacol Ther*, Vol.85, No.6, pp.600–606, 2009.

[18] W. DuMouchel and D. Pregibon, "Empirical Bayes screening for multi-item associations", *ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, San Francisco, USA, pp.67–76, 2001.

[19] G.N. Noren, R. Sundberg, A. Bate, *et al.*, "A statistical methodology for drug-drug interaction surveillance", *Stat Med*, Vol.27, No.16, pp.3057–3070, 2008.

[20] R. Harpaz, H.S. Chase and C. Friedman, "Mining multi-item drug adverse effect associations in spontaneous reporting systems", *BMC Bioinformatics*, Vol.11, Suppl.9, pp.S7–S9, 2010.

[21] N.P. Tatonetti, G.H. Fernald and R.B. Altman, "A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports", *J Am Med Inform Assoc*, Vol.19, No.1, pp.79–85, 2012.

[22] H. Yuan, S. Wang, Y. Li, *et al.*, "Feature selection with data field", *Chinese Journal of Electronics*, Vol.23, No.4, pp.661–665, 2014.

[23] S. Zhang, L. Zhang, K. Qiu, *et al.*, "Variable selection in logistic regression model", *Chinese Journal of Electronics*, Vol.24, No.4, pp.813–817, 2015.

[24] M.A. Hall, "Correlation-based feature selection for machine learning", *Ph.D. thesis*, The University of Waikato, New Zealand, 1999.

[25] Y. Zhai, S.P. Wang, N. Ma, *et al.*, "A data mining method for imbalanced datasets based on one-sided link and distribution density of instances", *Acta Electronica Sinica*, Vol.42, No.7, pp.1311–1319, 2014. (in Chinese)

[26] H.B. He and E.A. Garcia, "Learning from imbalanced data", *IEEE Transactions on Knowledge and Data Engineering*, Vol.21, No.9, pp.1263–1284, 2009.

[27] N.V. Chawla, K.W. Bowyer, L.O. Hall, *et al.*, "SMOTE: synthetic minority over-sampling technique", *Journal of Artificial Intelligence Research*, Vol.16, No.1, pp.321–357, 2002.

YUE Kejuan received the M.S. degree from North China University of Technology. She is a Ph.D. candidate of Central South University. Her research interests include data mining and image processing. (Email: yuekejuan@163.com)

ZOU Beiji (corresponding author) received the B.S., M.S., and Ph.D degrees from Zhejiang University in 1982, Qinghua University in 1984 and Hunan University in 2001 respectively. He is currently a Professor and served as the dean at the school of Information Science and Engineering at Central South University. His research interests include computer graphics and image processing. (Email: bjzou@csu.edu.cn)