Li et al. BMC Medical Genomics 2015, 8(Suppl 3):S4
http://www.biomedcentral.com/1755-8794/8/S3/S4

BMC Medical Genomics

RESEARCH Open Access

Patient classification of hypertension in Traditional Chinese Medicine using multi-label learning techniques

Guo-Zheng Li[1,2†], Zehui He[1†], Feng-Feng Shao[2], Ai-Hua Ou[1*], Xiao-Zhong Lin[3]

From IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2014)
Belfast, UK. 2-5 November 2014

Abstract

Background: Hypertension is one of the major risk factors for cardiovascular diseases. Research on the patient classification of hypertension has become an important topic because Traditional Chinese Medicine lies primarily in "treatment based on syndromes differentiation of the patients".

Methods: Clinical data of hypertension was collected with 12 syndromes and 129 symptoms including inspection, tongue, inquiry, and palpation symptoms. Syndromes differentiation was modeled as a patient classification problem in the field of data mining, and a new multi-label learning model BrSmoteSvm was built dealing with the class-imbalanced of the dataset.

Results: The experiments showed that the BrSmoteSvm had a better results comparing to other multi-label classifiers in the evaluation criteria of Average precision, Coverage, One-error, Ranking loss.

Conclusions: BrSmoteSvm can model the hypertension's syndromes differentiation better considering the imbalanced problem.

Background

Hypertension is one of the major risk factors of cardiovascular diseases. It contributes to one half of the coronary heart disease and approximately two thirds of the cerebrovascular disease burdens [1]. There are over 972 million hypertension patients in the world [2].Traditional Chinese Medicine (TCM) has been playing an important role on treating hypertension, and it lies primarily in "treatment based on syndrome differentiation of the patients". Traditionally, syndrome differentiation is performed by TCM practitioner should have solid theoretical foundation and plentiful experiences.

In the field of data mining, syndrome differentiation can be regarded as a patient classification problem which can be solved with specific data mining and machine learning techniques. It has become a fast developing field with the accumulating of clinical data [3-6].

In traditional classification problems, one case would be only classified to one category (i.e. label) which is called single label classification. While in TCM, one patient may have more than one syndromes which should be multi-label classification problems in the data mining field. Multi-label learning has been used in TCM field and got better results comparing with conventional learning methods. Liu et al. compared the performance of Multi-label-KNN and KNN on a coronary heart disease dataset. Li et al. had investigated the contribution of symptoms to syndromes diagnosis by using fusion symptoms with ML-KNN classifier [7]. Li et al. and Shao et al. proposed embedded multi-label feature selection method MEFS [8] and wrapper multi-label feature selection method HOML [9], respectively, to get better performance for the multi-label classification.

* Correspondence: ouaihua2@163.com
† Contributed equally
[1]Department of Big Medical Data, Guangdong Provincial Hospital of Chinese Medicine, The Second Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou, China
Full list of author information is available at the end of the article

Multi-label classification was mainly motivated by the tasks of text categorization and medical diagnosis in the past. The existing methods for multi-label classification can be grouped into two main categories: a) problem transformation methods, and b) algorithm adaptation methods. The problem transformation methods transform the multi-label classification problem either into one or more single-label classification or regression problems, and there have been many learning algorithms depending on transformation methods. The algorithm adaptation methods could extend specific learning algorithms to deal with multi-label data directly [10].

In classification, a dataset is said to be imbalanced when the number of cases which represents one class is much smaller than the ones from other classes [11]. Furthermore, the class with the lowest number of cases is usually the class of interest from the point of view of the learning task [12]. This phenomenon is of great interest as it turns up in many real-world classification problems, such as risk management [13], fraud detection [14], and especially medical diagnosis [15-19].

In this study, a new classification method named BrSmoteSvm is built for hypertension syndromes differentiation. The BrSmoteSvm works on both multi-label data and class-imbalanced problem. It is a combination of Binary Relevance (BR), Synthetic Minority Over-sampling Technique (SMOTE) [16] and Support Vector Machine (SVM) [17]. Firstly, BR algorithm is used to transform the multi-label classification problem into single-label classification. And it is found class-imbalance on the single-label situation. Then, SMOTE is applied to decrease the effect of the class-imbalance problem. At last, SVM is used as the binary classifier to differentiate the syndromes.

The rest of this paper is arranged as follows. Section 2 describes the materials and the methods of this study. Section 3 presents the results and discussion of our experiment. Section 4 presents the conclusions.

Methods

Materials

The study dataset originated from the hypertension patients who visited the in-patient and out-patient departments of Internal Medicine, Nerve Internal Medicine and Health Management Center of the Guangdong Provincial Hospital of Chinese Medicine and Li Wan District Community Hospital in Guangzhou of China during November 2006 to December 2008. This study was approved by the ethics committee of the Guangdong Provincial Hospital of Chinese Medicine, China. Informed written consent was obtained from each participant prior to data collection. In total, 908 cases were collected with 13 syndromes and 129 TCM symptoms from inspection symptoms, tongue symptoms, inquiry symptoms, palpation symptoms and other symptoms.

Four cases were excluded from the analysis because of missing answers on features. And one syndrome were excluded because of its nonnumeric value to make sure the smooth application of data mining methods. Finally, we got 904 cases with 12 syndromes and 129 symptoms. Table 1 shows the number of cases (D); the number of features (M); the number of labels (|L|); the Label Cardinality (LC), which is the average number of single-labels associated with each example defined by $\mathrm{LC(D)} = \frac{1}{|D|}\sum_{i=1}^{|D|}|Y_i|$; the Label Density (LD), which is the normalized cardinality defined by $\mathrm{LC(D)} = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i|}{L}, L = \bigcap_{i=1}^{|D|} Y_i$; the number of Distinct Combinations (DC) of labels. |D| represents the number of examples and $|Y_i|$ represents the label number of the i case.

Computational methods

In multi-label classification, each case could have several syndromes. The cases are associated with a subset of labels $Y \subseteq L$ where L is the set of possible labels. Following is a brief introduction of the algorithms used in this study.

1) SMOTE

SMOTE is used to decease the influence of the class-imbalanced problem. It is an over-sampling approach in which the minority class is over-sampled by creating "synthetic" examples. The main idea of SMOTE can be described as follows.

Step 1: Compute the k nearest neighbors for each minority class instance. Randomly choose N of the k nearest neighbors of each minority class instance saved as Populate.

Step 2: Take the difference of the feature vector between each minority class instance and its nearest neighbors in Populate. Multiply this difference by a random number between 0 and 1, and add it to the feature vector of each minority class instance.

The synthetic examples generated by SMOTE cause the classifier to create larger and less specific decision regions rather than smaller and more specific regions. More general regions are now learned for the minority class samples rather than those being subsumed by the majority class samples around them.

Table 1. Description of the datasets

Dataset	Domain	N	M	|L|	LC	LD	DC
hypertension	medical	904	129	12	0.86	0.07	57

2) SVM

SVM is used as the binary classifier in BR. The original SVM algorithm was invented by Vladimir N. Vapnik and the current standard incarnation (soft margin) was proposed by Vapnik and Corinna Cortes in 1995. The basic SVM takes a set of input data and related label, and for each given input, two possible classes forms the output, making it a non-probabilistic binary linear classifier. Given a set of training instances, each marked as belonging to one of two classes, an SVM training algorithm builds a model that can assign new instances into one class or the other. An SVM model is a representation of the instances as points in space, mapped so that the instances of the separate classes are divided by a clear gap and the gap is as wide as possible. The test instances are then mapped into that same space and predicted to belong to a class based on which side of the gap they fall on. The above describes that SVM performs a linear classification. In addition, SVM can also efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

3) BrSmoteSvm

The main idea of BrSmoteSvm is described as follows. In each fold of the 10-fold cross validation, BR, a problem transformation method is used. The basic idea of BR is to decompose the multi-label learning problem into q independent binary classification problems, where q is the number of label and each binary classification problem corresponds to a possible label in the label space [18]. Therefore, for any multi-label training example, each instance will be involved in the learning process of q binary classifiers. Then SMOTE is applied to training data to decrease effect of the class-imbalanced problem. In the end, SVM is used as the binary classifier. After the 10-fold cross validation, we get the predicted label set.

Experimental design and evaluation

In our experiment, 10-fold cross validation is utilized to test the accuracy of the classification. Let 700 cases of the data be training set, and 204 cases be testing set. In order to validate performance of BrSmoteSvm, it is compared with other popular multi-label classifiers.

1) ML-KNN. The number of neighbors is set to 10 and the smoothing factor is set to 1 as recommended.

2) Random k-Labelsets (RAKEL) [19]. J48 is used as the base learner; the number of models is set to 5; the size of subsets is set to 8.

3) Instance-based learning and logistic regression (IBLR) [20]. The number of nearest neighbors is set to 10.

4) Ensemble of Classifier Chains (ECC). J48 is used as the base learner for each Classifier Chains model; the number of models is set to 10.

5) A lazy multi-label classification method based on the KNN (BRKNN) [21]. The number of the nearest neighbors is set to 10.

At last, for SMOTE, N is set to fixed value 10, and k is chosen from {10, 12, 14, 16, 18, and 20}; then, k is set to fixed value 16, and N is chosen from {6, 8, 10, 12, 14, and 16} to evaluate the robustness of our method.

Let × denote the domain of cases and let Y={1,2,...,Q} be the set of labels. The purpose of the learning system is to output a multi-label classifier h: X→2y for the given training set through optimizing some specific evaluation metric. In other word, a successful learning system would output larger values for labels in Y_i than those not in Y_i for the given instance xi and its label set Y_i. For example, $f(x_i,y_i)>f(x_i,y_j)$ for any y_i in Y_i and y_j not in Y_i.

The real-value function $f(.,.)$ can be transformed to a ranking function $rank(.,.)$, which maps the outputs of $f(x_i,y)$ for any y in Y to {1,2,...,Q} such that if $f(x_i,y_i) > f(x_i,y_j)$ then $rank(x_i,y_i) < rank(x_i,y_j)$. For a test set $S=\{(x_1, Y_1),(x_2,Y_2),...,(x_p,Y_p)\}$, the following criteria are used in this study:

1) Hamming loss: defined as:

$$\text{hamming loss}\,(h) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{Q}\left|h(x_i)\,\Delta Y_i\right|$$

where Δ stands for the symmetric difference between two sets. Note that when | Yi | = 1 for all instances, a multi-label system is in fact a multi-class single-label one and the hamming loss is 2/Q times the usual classification error. Hamming loss is used to evaluate how many times an instance-label pair is misclassified. The smaller the value of Hamming loss (h), the better the performance.

2) One-error: defined as:

$$\text{one error}\,(f) = \frac{1}{p}\sum_{i=1}^{p}\left[\left[\arg\max_{y\in Y} f(x_i,y)\right]\notin Y_i\right].$$

Note that, for single-label classification problems, the one-error is identical to ordinary classification error. One-error is used to evaluate how many times the top-ranked label is not in the set of proper labels of the instance. The smaller the value of one-error (f), the better the performance.

3) Coverage: defined as:

$$\text{coverage}\,(f) = \frac{1}{p}\sum_{i=1}^{p}\max_{y\in Y_i} rank_f(x_i,y) - 1,$$

evaluates how far we need, on the average, to go down the list of labels in order to cover all the proper labels of

the instance. It is loosely related to precision at the level of perfect recall. The smaller the value of coverage (f), the better the performance.

4) Ranking loss: defined as:

$$\text{ranking loss}(f) = \frac{1}{p}\sum_{i=1}^{p}\frac{|D|}{|Y_i|\,|\overline{Y}_i|},$$

$$D = \{(y_1, y_2) f(x_i, y_1) \le f(x_i, y_2), \quad (y_1, y_2) \in Y_i \times \overline{Y_i}\},$$

where $\overline{Y}$ denotes the complementary set of Y in y. Ranking loss is used to evaluate the average fraction of label pairs that are reversely ordered for the instance. The smaller the value of ranking loss (f), the better the performance.

5) Average precision: defined as:

$$\textit{average precision}(f) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{|Y_i|}\sum_{y\in Y_i}\frac{|L_i|}{rank_f(x_i, y)},$$

$$L_i = \left\{y' | rank_f(x_i, y') \le rank_f(x_i, y), y' \in Y_i\right\},$$

which is used to evaluate the average fraction of labels ranked above a particular label $y \in Y$ which actually are in Y. The bigger the value of *average precision* (f), the better the performance.

Results and discussion

Comparison with other multi-label classifiers

The 10-fold cross validation was applied to test the accuracy of classification in which BrSmoteSvm was compared with other five multi-label classifiers. Results of 10-fold cross validation are shown in Table 2. In Table 2, the Average precision of BrSmoteSvm is 0.66, which is much higher than the results of other methods. For Coverage, One-error and Ranking loss, BrSmoteSvm also performs better than other methods. While, for Hamming loss, BrSmoteSvm is 0.09, which performs worse than other methods.

The reason of the large number of Hamming loss might be serious imbalance of the dataset. For most labels, there are only 20 to 70 positive cases, which means the ratio of the negative and positive cases is very high. On the other side, for the low number of the positive case, the classifier would be trained insufficiently producing bad performance of the testing data. Performance of machine learning methods is typically evaluated using predictive accuracy. It would be inappropriate when the data set is imbalance or the cost of different errors vary significantly. So, the simple predictive accuracy is inappropriate in this situation. In this study, SMOTE is applied to decrease the effect of the imbalance problem. The rate of detection positive cases would be improved, while the error rate for the negative cases be increased.

Another reasons could be that SMOTE might not be the best method dealing with the imbalance of the dataset, and the parameters for the algorithms used were not optimal. Further studies could focus on how to deal with the imbalanced problem and optimize the algorithms.

Furthermore, an experiment was conducted to compare the results with SMOTE and without SMOTE. The results are shown in Table 3. BrSmoteSvm +SMOTE represents with SMOTE, and BrSmoteSvm-SMOTE represents without SMOTE. It shows that the results with SMOTE are better than the results without SMOTE.

Stability of BrSmoteSvm

Two experiments were designed to validate the stability of BrSmoteSvm. The first one set N fixed as 10, and k was from {10, 12, 14, 16, 18, and 20} for SMOTE. The second one set k fixed as 16, and N was from {6, 8, 10, 12, 14, and 16}. The results of the two experiments are shown in Figure 1 and 2 using the evaluation criteria of Average precision, Hamming loss, Coverage, One-error and Ranking loss.

Figure 1 and 2 illustrate that:

1) The results of BrSmoteSvm vary with different k and N, but the change is small, indicating BrSmoteSvm is stable.
2) Whatever k and N values, BrSmoteSvm performs better than other methods in the evaluation of Average precision, Coverage, One-error and Ranking loss except for Hamming loss.

Table 2. Results of BrSmoteSvm and other multi-label classifiers using 10-fold cross validation

	BrSmoteSvm	MLKNN	BRKNN	ECC	IBLR	RAKEL
Average precision	0.66	0.53	0.51	0.51	0.51	0.46
Hamming loss	0.09	0.07	0.07	0.07	0.07	0.09
Coverage	1.11	2.21	2.46	2.41	2.34	2.89
One-error	0.47	0.75	0.75	0.76	0.76	0.78
Ranking loss	0.16	0.16	0.18	0.18	0.17	0.22

Table 3. Results of BrSmoteSvm with and without SMOTE

	BrSmoteSvm+SMOTE	BrSmoteSvm-SMOTE
Average precision	0.66	0.58
Hamming loss	0.09	0.07
Coverage	1.11	1.36
One-error	0.47	0.59
Ranking loss	0.16	0.19

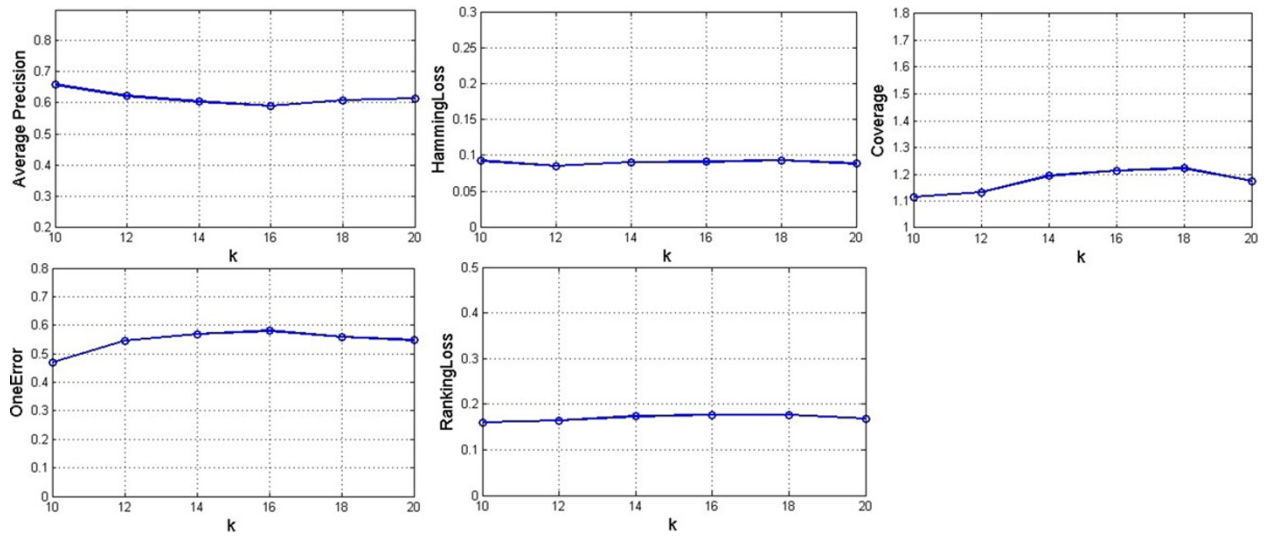

Figure 1 Results of BrSmoteSvm with different k values and fixed N value for SMOTE.

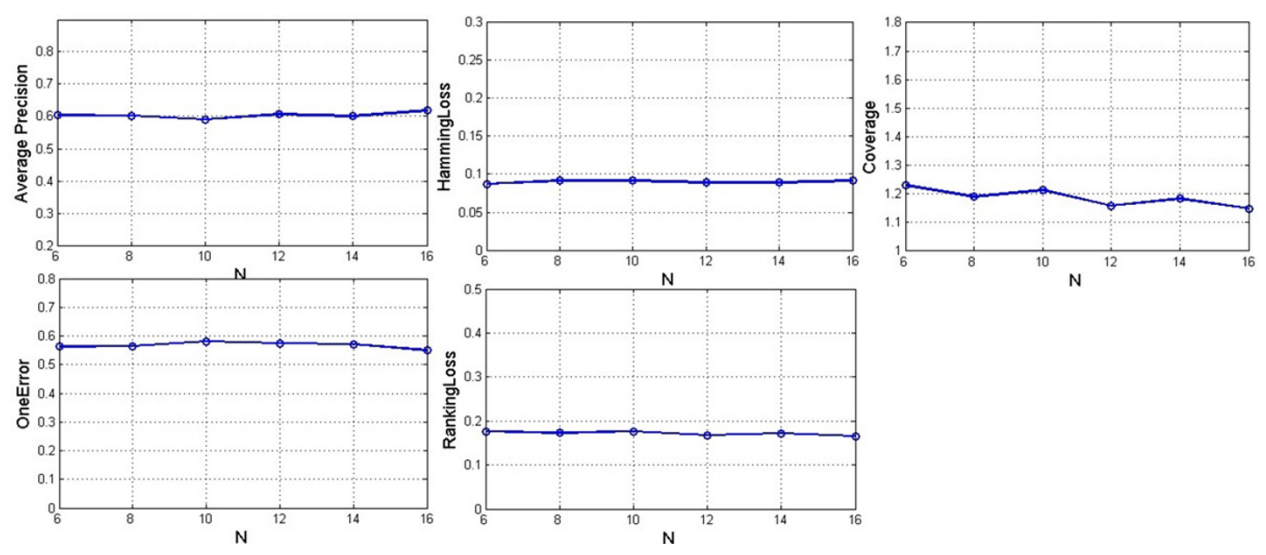

Figure 2 Results of BrSmoteSvm with fixed k value and different N values for SMOTE.

3) When k and N are both set to 10, BrSmoteSvm obtains the best performance.

Conclusions

Pattern classification is important in TCM for specific disease like hypertension. However, there are multi-labels of syndromes in patients, and the numbers of patients under each syndromes are so skew that classification performance is reduced. BrSmoteSvm is proposed by combining multi-label learning and SMOTE, to help overcome the effects of multi-labels and skew numbers of patients of syndromes. Results of experiments showed that BrSmoteSvm improves the performance of the previous works. Multi-label learning and imbalance learning techniques are necessary to process the medical data sets with above problems.

Further work may focus on novel combination of multi-label learning and imbalance learning techniques to improve the accuracy of classification.

Abbreviations used

BR: Binary Relevance

SMOTE: Synthetic Minority Over-sampling Technique

SVM: Support Vector Machine

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
GZL and ZH contributed to the design of the study, the critical revision of the manuscript. FFS performed the statistical analysis and drafted the manuscript. AHO and XZL planned and monitored the data collection procedures. All authors read and approved the final manuscript.

Acknowledgements
This work was supported by the Natural Science Foundation of China under grant nos. 61105053 and 61273305. Publication costs for this article were funded by National Science Foundation of China under grant no. 61273305. This article has been published as part of BMC *Medical Genomics* Volume 8 Supplement 3, 2015: Selected articles from the IEE International Conference on Bioinformatics and Biomedicine (BIBM 2014): Medical Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcmedgenomics/supplements/8/S3.

Authors' details
[1]Department of Big Medical Data, Guangdong Provincial Hospital of Chinese Medicine, The Second Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou, China. [2]Department of Control Science and Engineering, Tongji University, Shanghai, China. [3]Department of Cardiology, Guangdong Provincial Hospital of Chinese Medicine, The Second Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou, China.

Published: 23 September 2015

References

1. Whitworth JA, World Health Organization, International Society of Hypertension Writing Group: **2003 World Health Organization (WHO)/ International Society of Hypertension (ISH) statement on management of hypertension.** *Journal of hypertension* 2003, **21(11)**:1983-1992.
2. Vilela-Martin JF, Vaz-de-Melo RO, Kuniyoshi CH, Abdo AN, Yugar-Toledo JC: **Hypertensive crisis: clinical-epidemiological profile.** *Hypertension Research* 2011, **34(3)**:367-371.
3. Liu GP, Li GZ, Wang YL, Wang YQ: **Modelling of inquiry diagnosis for coronary heart disease in traditional Chinese medicine by using multi-label learning.** *BMC complementary and alternative medicine* 2010, **10(1)**:37.
4. Li GZ, Sun S, You M, Wang YL, Liu GP: **Inquiry diagnosis of coronary heart disease in Chinese medicine based on symptom-syndrome interactions.** *Chinese medicine* 2012, **7(1)**:9.
5. Poon S K, Poon J, McGrane M, Zhou X, Kwan P, Zhang R, Liu B, Gao J, Loy C, Chan K, Sze DM: **A novel approach in discovering significant interactions from TCM patient prescription data.** *International journal of data mining and bioinformatics* 2011, **5(4)**:353-368.
6. Wang X, Li GZ: **Multilabel learning via random label selection for protein subcellular multilocations prediction.** *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2013, **10(2)**:436-446.
7. Li GZ, Yan SX, You M, Sun S, Ou A: **Intelligent ZHENG classification of hypertension depending on ML-kNN and information fusion.** *Evidence-Based Complementary and Alternative Medicine* 2012, doi: 10.1155/2012/837245.
8. Li GZ, You M, Ge L, Yang JY, Yang MQ: **Feature selection for semi-supervised multi-label learning with application to gene function analysis.** *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology* 2010, doi:10.1145/1854776.1854828.
9. Shao H, Li GZ, Liu GP, Wang YQ: **Symptom selection for multi-label data of inquiry diagnosis in traditional Chinese medicine.** *Science China Information Sciences* 2013, **56(5)**:1-13.
10. Tsoumakas G, Katakis I: *Multi-label classification: An overview* [http://www.lpis.csd.auth.gr/publications/tsoumakas-ijdwm.pdf].
11. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F: **A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches.** *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions* 2012, **42(4)**:463-484.
12. Chawla NV, Japkowicz N, Kotcz A: **Special issue on learning from imbalanced data sets.** *ACM Sigkdd Explorations Newsletter* 2004, **6(1)**:1-6.
13. Huang YM, Hung CM, Jiau HC: **Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem.** *Nonlinear Analysis: Real World Applications* 2006, **7(4)**:720-747.
14. Cieslak DA, Chawla NV, Striegel A: **Combating imbalance in network intrusion datasets.** *IEEE International Conference on Granular Computing* 2006, doi: 10.1109/GRC.2006.1635905.
15. Mazurowski MA, Habas PA, Zurada JM, Lo JY, Baker JA, Tourassi GD: **Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance.** *Neural networks* 2008, **21(2)**:427-436.
16. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP: **SMOTE: synthetic minority over-sampling technique.** *Journal of artificial intelligence research* 2002, **16(1)**:321-357.
17. Cortes C, Vapnik V: **Support vector machine.** *Machine learning* 1995, **20(3)**:273-297.
18. Boutell MR, Luo J, Shen X, Brown CM: **Learning multi-label scene classification.** *Pattern recognition* 2004, **37(9)**:1757-1771.
19. Tsoumakas G, Katakis I, Vlahavas I: **Random k-labelsets for multilabel classification.** *IEEE Transactions on Knowledge and Data Engineering* 2011, **23(7)**:1079-1089.
20. Cheng W, Hüllermeier E: **Combining instance-based learning and logistic regression for multilabel classification.** *Machine Learning* 2009, **76(2-3)**:211-225.
21. Spyromitros E, Tsoumakas G, Vlahavas I: **An empirical study of lazy multilabel classification algorithms.** In *Artificial Intelligence: Theories, Models and Applications*. Berlin: Springer Berlin Heidelberg;Darzentas J, Vouros GA, Vosinakis S, Arnellos A 2008:401-406.

doi:10.1186/1755-8794-8-S3-S4
Cite this article as: Li *et al.*: Patient classification of hypertension in Traditional Chinese Medicine using multi-label learning techniques. *BMC Medical Genomics* 2015 **8**(Suppl 3):S4.