doi: 10.13241/j.cnki.pmb.2014.27.001

·基础研究·

Study of Relationship between Trace Elements in Soil and Leukemia Death Rate based on Adaboost Method

Isabel Cristina Echeverri Ocampo[1], CHEN Ding-ying[2], NIU Bing[1△], CHEN Fu-xue[1△]

(1 Shanghai Key Laboratory of Bio-Energy Crops, College of Life Science, Shanghai University, 99 Shang-Da Road, Shanghai, 200444, China; 2 Shanghai Institute of Measurement & testing,716 Yishan Road, Shanghai, 200233, China)

ABSTRACT Objective: This study was conducted to explore the relationship between trace elements level and leukemia based on data mining methods. **Methods:** In this article, the relationship between leukemia death rate and trace elements in soil of 29 provinces was studied using the correlation-based feature subset (CfsSubset)-Adaboost method. **Results:** Thirteen trace elements were found related to leukemia, and Arsenic (As) was found to have impact on leukemia death rate. By using AdaBoost method with the thirteen selected trace elements, the prediction model yields an accuracy rate of 89.7% for the 10-folds cross validation test. It is also found that As has an important impact on leukemia death rate. **Conclusion:** Trace elements in soil have correlation to leukemia death rate, especially the trace elements Arsenic. Our study results could provide an assistant for investigation of the relationship between leukemia death rate and trace elements, and even could be regarded as a potential supplement for the prevention and clinical treatment of leukemia.

Key words: Leukemia; Trace element; Feature selection; Adaboost

Chinese Library Classification (CLC): R733.7; R151.3; Q-31 Document code: A

Article ID: 1673-6273(2014)27-5201-03

Introduction

Leukemia is one of the most common cancers in adults. The etiology and pathogenesis of leukemia is not yet identified, but in recent years some experiments indicate that trace elements including zinc (Zn), copper (Cu), selenium (Se), chromium (Cr), lead (Pb), cobalt(Co),nickel (Ni), manganese (Mn) and barium (Ba) in patients' hair or blood are often abnormal when people get leukemia. On the other side, arsenic trioxide (As_2O_3)and Lithium Carbonate (Li_2CO_3)has been applied to treat leukemia. Hence, it is speculated that the pathogenesis of leukemia may be related to these trace elements. In contrast, investigating these relationships may be helpful for the prevention of leukemia[1,2].

China has a vast territory, and the content of trace elements in soil are various greatly from different regions as the soil is a major source of trace elements. Recent years, China carried out the project on the study of relationship between cancer epidemiology and soil environmental chemistry. Many efforts have been made to analysis the trace element of soil and the Chinese provinces, municipalities make the trace element analysis of soil throughout the region, in fact by district [3], and count the cancer death rate of different provinces and cities[4].Consequently, using mining data algorithms can summarizes the relationship between trace elements and various cancers. In this study, we report the use of Adaboost to study the relationship between leukemia and trace elements in soil.

Author introduction: Isabel Cristina Echeverri Ocampo(1986-), female, postgraduate, Research field: Bioinformatics, Tel:66137038, E-mail: cris_echeverri@hotmail.com

△Corresponding author: NIU Bing, E-mail: bingniu@shu.edu.cn; Chen Fu-xue, E-mail: gfxchen@shu.edu.cn

(Received:2014-01-12 Accepted:2014-02-10)

1 Methods and Data Files

1.1 Data

The data are derived from reference [3] which collected several trace elements (Al, As, B, Ba, Be, Ca,Cd, Co, Cr, Cs, Cu, Fe, Ge, Hg, K, Li, Mg, Mn, Na, Ni, Pb, Rb, Sr, Ta, Ti, Zn, Zr and Se) in soil of 29 provinces and Leukemia death rate of these regions. The death rate less than 2×10^{-5} is defined as positive samples and others as negative samples.

1.2 Theory of Correlation-based Feature Subset (CfsSubset) selection method

If there are n possible features in the original data set, there will be 2n possible combinations of features for the subset. The most rigorous way to find the best subset is to try them all, which is impossible due to the massive amount of calculations involved. CfsSubset selection algorithm is a heuristic feature-selection method for evaluating the worth of a subset of attributes by considering the individual predictive ability of each of the features along with the degree of redundancy between them[5].

The details of CfsSubset selection algorithm are as follow:

Step one: the merit of a matrix of features-class and feature-feature correlation was calculated according to equation (1).

In equation (1), Dn is the heuristic "merit" of a feature subset D, is the mean feature-class correlation, and is the average feature-feature inter-correlation. Eq (1) is the heart of the CfsSubset

selection algorithm, which could be used to evaluate the ability to predict, and the degree of redundancy of a subset of features.

Step two: Search the feature subset space with forward Best first[6].

During this process, greedy hill-climbing augmented with a backtracking facility are employed to search the feature. Best first method may start using three types of sets: I. start with the empty set of attributes and search forward; II. start with the full set of attributes and search backward; III. start at any point and search in both directions(backward or forward).

More information of CfsSubset selection algorithm can be found in our previous study[6,7].

1.3 AdaBoost Learner

AdaBoost is one of the most popular and important ensemble learning algorithms. The goal of this algorithm is to find a strong classifier from many weak classifiers[8-11]. An essential aspect of applying AdaBoost is to give greater weight to those samples that are difficult to be predicted correctly. A detailed description of AdaBoost algorithm can be found in our previous study[7, 12-14].

1.4 C4.5 decision tree

C4.5 is an extension of ID3 decision tree designed by Quinlan. C4.5 decision tree is a branch-test-based classifier and can be used for classification [36]. At each node of the tree, C4.5 chooses the feature of the training data that can classify the instances into subsets. The classification criterion is the normalized information gain. The feature with the best normalized information gain is chosen to make the decision. Compared to its earlier version, C4.5 is improved in the following aspects:

1). Handling both continuous and discrete attributes;

2). Handling training data with missing attribute values;

3). Pruning trees after creation;

4). Avoiding over fitting data.

1.5 Accuracy Measure

Generally speaking, the prediction performance of different discriminative methods is commonly evaluated by the function of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). In this study, we employed sensitivity (SN=TP/[TP+FN]), specificity (SP=TN/ [TN+FP]), and overall accuracy (ACC=[TP+TN]/[TP+TN+FP+FN]) to measure the prediction.

1.6 Implmentation

All the computations were run on a 1.86 GHz Intel (R) core (TM) 2 Dell PC with Windows 7 and Linux Ubuntu 9.04 desktop edition operation system.

We choose C4.5 as a weak learner. For AdaBoost, the number of iteration was 10; the weight threshold was set 100. For C4.5, the confidence factor was 0.24.

2 Results and Discussion

2.1 Selection of features

The selection of features is relatively tough work due to the redundancy of some features. In the past, it has been dependent on the experience of the researcher. Recently, some promising results have been reported on the problem of descriptor selection [8-11]. In this work, CfsSubset searching method was applied to the selection of descriptors. After computation, 13 trace elements (As, Cd, Ta, Hg, Cs, Rb, K, Ca, Ti, Mn, Pb, Mg and Na) were selected to build the model.

2.2 Calculation results

A re-substitution test is an examination for the self-consistency of a prediction method, which was performed in the current study. The prediction result of the re-substitution test for the model is 100%. Although the prediction accuracy of re-substitution test is high, it was not good enough to evaluate the generalization and reliability of this prediction model due to the problem of over-fitting. Hence, 10-folds cross validation test is used to further validate the generalization and reliability of the prediction model. During the process of 10-folds cross-validation analysis, the datasets are fed into the system and divided into N folds, a model is built with N-1 folds samples and the N-th fold is predicted. Each fold is left out from the model derivation and predicted in turn[28,36-52]. As a result, the accuracy of 10-folds cross validation test for the prediction model is 89.7 % (see Table 1).

To further evaluate the feature selection method, we also use the original 29 feature set to predict (see Table 1). From Table 1, it can be seen that the original 29 features achieve accuracy of 79.3 % for the 10-folds cross-validation test. After applying the approach, the accuracy of the 10-folds cross-validation test raised to 89.7 %. Moreover, it can be found that SP rose greatly from 37.5 % to 75 %. Herein, it can be concluded that some features in the original data set are disturbing and redundant, and these features has been successfully excluded after performing CFS method. Also, comparing to the original data set, the decreasing of the number of features in final subsets suggests that CFS approach could decrease the feature number and improve prediction.

Table 1 Prediction accuracy of different dataset of 10-folds cross-validation

Data set	Prediction accuracy(%)		
	SN	SP	ACC
Original data set	95.2	37.5	79.3
Optimal data set	95.2	75	89.7

2.3 Feature analysis

Leukemia death rate is extremely complex which combined effect of various factors, the impact of trace elements in the soil, is just one aspect. We rank the 13 selected features according to their relevance to the target (see Table 2).

From Table 2, we can see that As is more closed to the death rate of leukemia. In recent year, several papers with respect to the pathology of arsenic trioxide and cancer have been published. The experiments show that: Although high concentrations of arsenic

Table 2 Ranking list of thirteen trace elements

No.	Trace element	No.	Trace element
1	As	8	Ca
2	Cd	9	Ti
3	Ta	10	Mn
4	Hg	11	Pb
5	Cs	12	Mg
6	Rb	13	Na
7	K		

trioxide is toxic and have carcinogenic effect, the low concentrations of arsenic trioxide was able to induce apoptosis in cancer cells, which became a pharmaceutically effective treatment of leukemia [12]Showed low concentrations of arsenic may be beneficial anticancer, in turn the literature has not yet seen the reports of Cd can cause cancer, however known Cd in the body have anticancer effects, combined with selenium, loss of activity or difficult to absorb selenium. This article summarizes the arsenic and mercury in the soil and Leukemia death rate relationship may wish to refer to the pathology results to be understood.

3 Conclusion

In this study, CFS-Adaboost was applied to explore the relationship between leukemia and trace elements in soil. As a result, thirteen trace elements were abstracted from original twenty-nine features. And As is found to have an important effect on leukemia. Based on these thirteen trace elements, a prediction model was built. Because of its high rate of self-consistency (100 %) and correct prediction rate in 10-folds cross validation test (89.7 %), it is expected that this method can be a promising assistant technique for modifying effect in clinical outcome of leukemia.

References

[1] Carpentier U, Hyers J, Thorpe L. Copper, zinc and iron in normal and leukemic lymphocytes from children [J]. Cancer Research, 1986,46 (2):1981

[2] Zhou CL, Li YJ, Wei AY. Trace elements and leukocythemia[J]. Trace elements and health research, 2002,19(2):76

[3] Central Station of Environmental Monitoring of China. The background values of element contents in China [J]. Beijing: Chinese Environmental Press, 1990

[4] Zhong HX. Epidemiology and prevention of cancer [J]. Guangzhou: Guangzhou Branch of Publishers for Popular Sciences, 1987

[5] Hall MA. Practical feature subset selection for machine learning [M]. Proceedings of the Twenty first Australian Computer Science Conference: Springer, 1998

[6] Korf RE, Chickering DM. Best-first minimax search [J]. Artificial Intelligence, 1996,84:299-337

[7] Niu B, Jin YH, Lu WC, et al. Predicting toxic action mechanisms of phenols using AdaBoost Learner [J]. Chemometrics and ntelligent Laboratory Systems, 2009,96(1):43-48

[8] Goldberg DE. Genetic algorithms in search optimization and machine learning[J]. Readomg, Mass: Addison-Wesley, 1989

[9] Hall MA. Practical feature subset selection for machine learning[M]. Proceedings of the Twenty first Australian Computer Science Conference: Springer, 1998

[10] Kohavi R, John G. Wrapper for Feature Subset Selection [J]. Artif. Intell, 1997,1-2:273-324

[11] Peng HC, Long FH, Ding C. Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy[J]. Ieee T Pattern Aanal, 2005,27(8):1226-1238

[12] Du KJ, Chen JY. Arsenic trioxide and the apoptotic effect of cancer cells[J]. Trace elements and health research, 2001,18(1):67-69

Adaboost 算法研究土壤微量元素含量与白血病的关系

Isabel Cristina Echeverri Ocampo[1] 陈丁滢[2] 钮 冰[1△] 陈付学[1△]

（1 上海大学生命学院 & 上海市能源作物育种及应用重点实验室 上海 200244；2 上海市计量测试技术研究院 上海 200233）

摘要 目的：本文使用数据挖掘方法研究土壤中微量元素和白血病的相关性。**方法：**使用 CFS-Adaboost 算法研究我国 29 个省、市、自治区白血病死亡率的统计数据和土壤中微量元素含量的对应关系。**结果：**从 29 种微量元素中发现了 13 种微量元素与白血病相关，其中砷(As)的相关性较为明显。基于该 13 种微量元素，建立了土壤中微量元素和白血病致死率的数学模型，该模型的预报准确率可达到 89.7 %。**结论：**土壤中微量元素的含量与白血病有一定关系，其中砷(As)元素含量与白血病死亡率较为密切，这与近年文献报导的少量氧化砷治疗白血病效果显著相符合。以上研究发现，可以为研究土壤中微量元素和白血病的关系提供参考，对白血病的防治工作具有一定意义。

关键词：白血病；特征筛选；微量元素；Adaboost 算法

中图分类号：R733.7；R151.3；Q-31 **文献标识码：**A **文章编号：**1673-6273(2014)27-5201-03

作者简介：Isabel Cristina Echeverri Ocampo (1986-)，女，硕士研究生，主要从事生物信息学研究，电话：66137038，
E-mail：cris_echeverri@hotmail.com

△通讯作者：钮冰：副教授，博士，E-mail: bingniu@shu.edu.cn；
陈付学：教授，博士，E-mail:gfxchen@shu.edu.cn

（收稿日期：2014-01-12 接受日期：2014-02-10）