

Prediction of Allergy Symptoms among Children in Taiwan Using Data Mining

Hui-Fuang Ng^a, Halim Fathoni^a, I-Chi Chen^{b*}

a Dept. of Computer Science and Information Engineering, Asia University

b Dept. of Health Care Administration, Oriental Institute of Technology

** Correspondence: fl001@mail.oit.edu.tw*

Abstract

Allergic diseases have increased in children in many countries with modern living conditions. This study used data mining techniques to construct predictive models for prediction of allergy symptom among children in Taiwan from survey data. Among three predictive models, the decision tree model gives the best overall result, with a classification accuracy of 70.3%, followed by neural networks (69.6%) and support vector machine (66.4%). Results indicate that children exposed to environment that contains allergy inducing agents are key factors that trigger allergy symptoms. Family factors, especially heredity factors, are also important factors of children allergy.

Keywords: Allergy Symptoms, Data Mining, Children.

exposure to household pets must be controlled for those with hyperactive airways [4].

Healthy lifestyle is easier to maintain when healthful patterns of behavior are learned early in life. Environment and socio-cultural factor contributes to the person's ability to develop better health. For children, their parents make most of the decisions related to their health practices or behaviors. Therefore, it is important for the parents to understand what factors can influence the allergy symptoms in children.

Data mining is intelligent science of mining and extracting previously unidentified but potentially valuable information embedded in data. Data mining search valuable information by applying statistical, machine learning and pattern recognition. This study used data mining techniques as a tool to construct predictive models of allergy symptom which draws inferences from survey data to identify predictors of allergy symptoms among children in Taiwan. The results of this study could help health professional to intervene early in addressing allergy symptoms among children, and could be used as guidelines for developing strategies for allergy prevention.

1. Introduction

During the past 50 years, allergic diseases have increased in children in many countries with modern living conditions. A recently study found increased allergic sensitization in a subgroup of adults living on busy road [1]. Another research reported that overweight also increased risk of allergy [2]. The U.S. Center for Disease Control (CDC) has classified indoor air pollution as a factor of high environmental risk [3]. An investigation of risk factors for wheezing among people with nonspecific bronchial hyperactivity found that

2. Materials and Methods

2.1. Dataset

The data source used in this study was from the 2005 National Health Interview Survey (NHIS) in Taiwan. We focused on data of allergy symptoms in children of age under 12 years old. The selected dataset consisted of over 30 predictor variables covering the following groups of factors:

- personal factors: age, gender, height, weight, birth status;
- health behavior factors: personal hygiene, diet habit, preventive care utilization;
- living condition factors: second hand smoke, carpet, burning of joss sticks, present of cockroaches or rats, etc.;
- family factors: household income, parent's education, parents have allergies, siblings have allergies, number of siblings, etc.;
- allergy inducing factors: season change, air pollution, pets, dust, emotion, foods, smoke, etc.

After the preprocessing stage, which includes missing value removal and under-sampling on the majority class to make the classes even, the final dataset consisted of 1030 records. The dependent variable was a binary categorical variable with 0 representing no allergy and 1 representing allergy. Table 1 shows the distribution of the dependent variable, where about 45% of the records were children with allergy symptoms and the rest were children without allergy symptoms.

2.2. Predictive Models

This study used three different types of prediction models: neural networks, decision tree, and support

vector machine (SVM). These models are commonly used in data mining analysis.

Table 1 Distribution of dependent variable of the dataset.

Category	Frequency	Percentage
0 (No Allergy)	568	55.1 %
1 (Allergy)	462	44.9 %
Total	1030	100%

a. Neural Networks

Artificial neural networks are commonly known as biological inspired, highly sophisticated analytical techniques. A neural network is a mathematical model that makes predictions based on a series of input descriptor variables. Neural networks comprise of series of independent processors or nodes. These nodes are connected to other nodes and organized into a series of layers. There are three series of layer in neural networks: the input layer, the hidden layer, and the output layer [5]. We used the multi-layer perceptron network with back-propagation learning algorithm for our analysis.

b. Decision Tree

A decision tree can be learned by splitting the source data set into subsets based on an attribute value test. Splitting process in decision tree uses mathematical algorithms (e.g. information gain, Gini index, chi-squared test) to identify a variable and corresponding threshold. This splitting process is repeated on each derived subset in a recursive manner until complete tree

at each leaf node is constructed. The objective of the splitting process is to find a variable-threshold pair that maximizes the homogeneity of the resulting two or more subgroups of sample [6].

c. Support Vector Machine

The support vector machine (SVM) is a training algorithm for learning classification and regression rules from the data. Classification in SVM, the input data will be transformed to a high dimensional feature space in which the input data become more separable. Then, SVM will construct a separating hyperplane that has the maximum distance from the data of the two classes on each side of the hyperplane [6].

2.3. Performance Evaluation

In this study we evaluated the performance of the data mining models on the NHIS dataset by using the classification accuracy defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP, TN, FP and FN denote true positives, true negatives, false positives and false negative, respectively.

2.4. k-fold Cross Validation

To minimize the bias associated with the random sampling of the training data and testing data, we used the *k*-fold cross validation technique. This method splits data set into *k* subsets or folds, and the classification model is trained and tested *k* times. Each time one subset

is chosen as test set and the remaining subsets are used as training set. The process is repeated until all subsets are tested once [5]. The overall accuracy is simply the average across the *k* accuracy measures.

2.5. Feature Selection

Real data normally consist of large number of correlating features which may cause serious problems to data mining. The objective of feature selection is to select a minimal subset of features according to some criteria in order to remove irrelevant and redundant features and hence to improve the classification accuracy. In this study, we used the sequential forward selection approach in which the selection begins with one feature, and sequentially adds another feature one by one, until every feature is selected. The minimal subset of features that produces the best performance is then selected as predictor variables for the model. This process is performed separately for each model.

2.6. RapidMiner

We used the *RapidMiner* software to perform the data mining tasks. *RapidMiner* is an open-source data mining software tool which provides easy to use user interface and offers a huge amount of data mining analysis methods. The software can be downloaded freely from www.rapidminer.com.

3. Results

The classification accuracies of neural networks, decision tree, and SVM on predicting allergy symptoms in children under age 12 using the NHIS dataset are shown in Table 2. The classification accuracies are 69.6%, 70.3%, and 66.4% for neural networks, decision tree, and SVM respectively. The best result is achieved by the decision tree model, although the differences are not very significant. The average classification accuracy for the three models is about 69%.

Table 2 Classification accuracies of three predictive models.

Model	Accuracy (%)	Average (%)
Neural Networks	69.6	68.8
Decision Tree	70.3	
SVM	66.4	

Because the decision tree model produces the best overall result, we investigated the important of the key factors that cause allergy symptoms in children by computing the information gain of each factor in the decision tree model, and obtaining the relative rankings for the predictor variables. Table 3 shows the top ten most significant predictor variables based on information gain ranking. As we can see from Table 3, allergy inducing factors such as season change, foods, dust, emotion, and air pollution are the most significant factors for predicting allergy symptoms. The next important factors are the family factors, including heredity factors (parents and siblings have allergy), number of siblings in the family, and interestingly, parent's education level.

Table 3 Relative ranking of the top ten most significant predictor variables.

Predictor Variables	Relative Rank
Season change	1.00000

Foods	0.52618
Dust	0.52197
Emotion	0.39291
Air pollution	0.30902
Parents have allergies	0.23943
Number of siblings	0.08010
Siblings have allergies	0.06837
Smoke	0.05324
Parent's education	0.04031

4. Discussion and Conclusion

In this study, three data mining models: neural networks, decision tree, and SVM were used to construct predictive models for predicting allergy symptom among children in Taiwan based on the NHIS dataset. Among the three models, the decision tree model gives the best result, with a classification accuracy of 70.3%, followed by neural networks (69.6%) and SVM (66.4%). Experimental results also indicate that allergy inducing factors appears to be the most significant factors influencing allergy symptoms in children. Children exposed to environment that contains allergy inducing agents could trigger allergy symptoms; this finding is in line with the results reported in [1]. Family factors, especially heredity factors, also have an impact on children allergy. If the parents have allergy, it is likely that their children will also be more susceptible to allergens.

The average classification accuracy on predicting allergy symptoms using three data mining models and the NHIS dataset is about 69%, which is considered low for prediction purposes. Nevertheless, this result is not unexpected for a survey database that normally contains large variations which make it difficult to extract accurate predicting models.

5. Acknowledgement

This study is based on part of the data from the 2005 National Health Interview Survey (NHIS) in Taiwan. The authors are grateful to the National Health Research Institutes and the Department of Health of Taiwan for the NHIS database.

6. References

- [1] Nicolai T, Carr D, Weiland SK, et al., “Urban traffic and pollutant exposure related to respiratory outcomes and atopy in a large sample of children”, *European Respiratory Journal* 2003; 21:956-963.
- [2] Irei AV, Sato Y, Lin TL, et al., “Overweight is associated with allergy in school children of Taiwan and Vietnam but not Japan”, *The Journal of Medical Investigation* 2005; 53(1-2):33-40.
- [3] U. S. Center for Disease Control (CDC), “Strategic Plan 1999-2003”, 2000; <http://www.cdc.gov/nceh/publications/stratplan/default.htm>.
- [4] Dharmage S, Abramson M, Raven J, et al., “Why do only some of the young adults with bronchial hyperactivity wheeze?”, *Journal of Asthma* 1998; 35(5):391-399.
- [5] Larose D, *Discovering Knowledge in Data: An introduction to data mining*, Wiley-Interscience 2005; Hoboken, NJ.
- [6] Olson L and Delen D, *Advanced data mining techniques*, Springer 2008; Berlin Heidelberg.