International Journal of Advanced Culture Technology Vol.8 No.3 275-279 (2020)
DOI https://doi.org/10.17703/IJACT.2020.8.3.275

IJACT 20-9-35

Prediction of Childhood Asthma Using Expectation Maximization and Minimum Description Length Algorithm

[1]Hyo Seon Kim, [2]Jong Suk Park, [3]Dong Kyu Nam, [4*]Yong Gyu Jung

[1]*Master's student, Dept. of Medical IT Marketing, Eulji University, Korea*
[2,3]*{Deputy CEO, CEO}, PURIUM Co. Ltd., Korea*
[4]**Professor, Dept. of Medical IT, Eulji University, Korea* (corresponding author)
[1]20180737@eulji.ac.kr, [2]dorim604@daum.net, [3]namdongkyu.kor@gmail.com, [4]ygjung@eulji.ac.kr

Abstract

Due to the recent rapid industrialization worldwide, the number of pediatric asthma patients is increasing. And the fine dust containing heavy metals is linked to the characteristics of high toxic lead due to the increase heating in factory operation and automobile driving. It is the reason of arsenic increasing. In the treatment of pediatric asthma patients, drug administration, oral drug entry, and HMPC (Home Management Plan of Care) are used. In this paper, we analyze the relationship between the onset of asthma and the method of prescription for specific childhood asthma in the United States using EM (Expectation Maximization) and MDL (Minimum Description Length) algorithms. And the association is also analyzed by comparing the nature of specific congestion between the past prevalence of digestive asthma and the recent prevalence of environmental pollution.

Keywords: *Childhood Asthma, EM, Expectation Maximization, MDL, Minimum Description Length, HMPC, Home Management Plan of Care*

1. Introduction

In recent years, rapid industrialization is progressing around the world, and heavy metal-containing fine dust with high toxic lead and arsenic properties, especially from automobiles and plants, is increasing. Therefore, each country increases the cost of investment at the national level to reduce environmental pollution. Children who are sensitive to the environment are more susceptible to respiratory diseases caused by fine dust and the like than adults. Children's respiratory system is not fully developed and their immunity is weak. In particular, symptoms of cough and phlegm often appear due to inflammation of the bronchial allergic reaction, and children suffer from asthma while breathing. Asthma in childhood can increase cough, complaining of chest pain. In particular, asthma, lung and bronchial function deterioration that occurs with age due to narrowing of the bronchi are very serious. However, unlike adult asthma, children's asthma has a high cure rate, as well as exudates and genetic influences, so it can get better quickly depending on the treatment. Reducing the environmental causes of asthma in children, and the faster the treatment through accurate testing, the better the treatment process. You can also prevent progression to severe asthma and adult asthma with appropriate treatment right after recovery begins. In the process of treating asthma, liquids such as symptom relief, oral medications such as steroids are formulated, and when hospitalized, patient record condition monitoring is a method that can be systematically managed. In this paper, cases of congestion were clustered by region through EM algorithm by comparing the prevalence data of child asthma in the United States by feature variable. Treatment methods were compared and analyzed through the difference in congestion by region.

Manuscript Received: Aug. 11, 2020 / Revised: Aug. 24, 2020 / Accepted: Sep.1, 2020
+ Corresponding Author: ygjung@eulji.ac.kr Tel:031-740-7198
Dept. of Medical IT, Eulji University, Korea

II. Related Literature

2.1 Clustering

Assuming that one object has multiple properties, it can divide all objects with similar properties into multiple groups in cluster analysis. For this, it is necessary to determine the similarity between objects. The measure of similarity is determined by the Euclidean distance for each feature variable, and the Euclidean distance is defined as follows.

$$d(x_i,x_j) = \sqrt{\sum^{p}(X_{ai} - X_{aj})^2} = \sqrt{(x_i - x_j)^T(x_i - x_j)}$$

$$d(x_i \cdot x_j) = \sqrt{\mathrm{P}_{\Sigma\left(\mathrm{x_{a_i}} - \mathrm{x_{a_j}}\right)}2} = \sqrt{\left(\mathrm{x_i} - \mathrm{x_j}\right)^{\mathrm{T}}\left(\chi_{\mathrm{i}} - \mathrm{x_j}\right)}$$

The object data for Clustering point is not the state variable indicating the category to which an object belongs. It is different from that of the classification analysis. In other words, cluster analysis is a learning method by comparison method. If the cost is for too many and the class label of the data is not given, it is convenient that general rule is learning method of comparison. It is labeling to a particular class one by one to each data-consuming.

2.2 EM (Expectation Maximization) algorithm

The EM algorithm is not a learning algorithm limited to Gaussian mixed models, but a learning method commonly used to estimate parameters from other probabilistic models. In particular, in a probability model, it can be used to estimate parameters and cryptographic parameters together, and it has characteristics that differentiate it from other optimization algorithms. Also, as in the case of Gaussian mixed model, even if a specific model does not exist in the original and is hidden, EM algorithm can be applied by converting the model to encryption parameters.

The steps of the general EM algorithm are

1) Given data set $\{x_1 x_2 .. x_n\}$ can define the parameter which probability model $p(x, z|\theta)$.
2) The initial value of each parameter, $\theta^{(0)}$ Arbitrarily set.
3) [E-step] parameters given iteration step, $\theta^{(r)}$ Using Q Function is obtained.

$$Q\left(\theta, \theta^{(r)}\right) = E_z[lnP\ (X, z|\theta]) = \sum_z lnp\ (X, z|\theta)p(z|X, (r))$$

4) Repeat until the E-step until convergence of the parameters or the desired Q value is obtained
5) [M-step] E-step in the resulting Q calculate the parameters that maximize the function.

$$\theta^{(r+1)} = argm\ a\ x_\theta Q\left(\theta, \theta^{(\gamma)}\right)$$

2.3 MDL (Minimum Description Length) Algorithm

MDL which means Minimum Description Length takes the minimum lower bound of the optimal theory of a single data set. It defines the amount of information needed to be added by the theory, the information related to the exception. If that's the best, a decent way is to compare all such cases on the same basis. In other words, you are running a training set for your village, and you don't need another validation set and you are looking for an evaluation system. At this point, assuming that T is expressed based on the training set E, the theory of T requires L. Vito specific [T]. Since you are only interested in correctly predicting the class label, I assume you mean the congested

class label in training set *E*. Given a theory, it can be coded with a specific number of bits in the training set itself. $L[T]$ is actually given as the loss function of the information obtained by adding all of the target elements of the training set. Accordingly, the result of adding the explanatory length and full theory of the training set is:

$$L[T] + L[E|T]$$

A notable relationship exists between probability theory and basic MDL principles. The conditional probability of Bayes' rule can be written as

$$p_r[T|E] = \frac{p_r[E|T]p_r[T]}{p_r[E]}$$

By taking the negative logarithm function follows

$$l\ o\ g\ p_r[T|E] = -l\ o\ g\ p_r[E|T] - l\ o\ g\ p_r[T] + l\ o\ g\ p_r[E]$$

Using an algebraic function is equivalent to maximizing the probability value immediately. Now the number of bits required for encoding will immediately have a logarithmic value of negative probability. It is independent of the learning algorithm at the end of the above formula [*E*], and intentionally does not rely solely on the training set. Hence the probability value. Choosing a theory that maximizes [*TE*] satisfies the same relationship as choosing a theory with the minimum value of the following equation.

$$L[E|T] + L[T]$$

III. Experimental Process

3.1 Childhood Asthma

Childhood asthma is an allergic disease in which the typical bronchi in childhood is narrowed, and symptoms such as shortness of breath and cough appear repeatedly. Very sensitive to non-specific external stimuli: children, allergens, patients with bronchial asthma, loosening, vomiting. If you cough with wheezing along with complement radish and shortness of breath, the causative symptoms appear. Difficulty breathing may occur depending on the patient, sputum, or cough, but most people have these symptoms together. It is a common symptom that the causative agent disappears from the environment and soon disappears, sometimes lasts long, and can be repeated. In addition, bronchial asthma in pediatric patients has a persistent allergic inflammatory reaction not only with symptoms, but also without symptoms. Childhood asthma is similar to adult asthma, and the mechanism of triggering factors differs in prognosis and diagnosis. In children with a lot of asthma, symptoms of asthma often go away after puberty. If asthma is not managed, it can occur annually, and lung function declines continue to appear, making lung damage unrecoverable or limited in daily activities. Therefore, not only when irritability is severe, but without proper treatment, it can become a chronic disease that can plant mental stress such as substances that cause allergies, cold, cold, cigarette smoke, soot, exercise, etc. Even if only symptoms appear and there are no symptoms, they should be managed with continuous treatment.

3.2 Experimental data description

This data used child asthma from public data in the United States. The attribute variable names and data values are organized as shown in Table, and it shows the progress of pediatric asthma prescription hospitals in the United States, and includes patient information, various hospitals, and probability data on whether prescription Na is included in the number of pediatric asthma patients. Has been.

3.3 Preprocess

Pre-process the data for accurate analysis of the data. Some attributes are excluded because they do not help in

data analysis. Specifically removed by footnote and hospital informant numbers, zip codes, and phone numbers. Data for null values for which the instance has no value was deleted. The attributes used for actual data analysis include state name, Reliever Medication, Systemic Corticosteroid, and A Home Management Plan of Care.

IV. Evaluation and Discussion

4.1 Experimental Result

The experiment was carried out using the EM algorithm, divided into 6 places to analyze the child's asthma using the method of maximum likelihood, the value of maximum likelihood came-measuring the hospital care procedure. You can see that the mean and standard deviation of Corticosteriod Medication congestion Reilever Medication converges to almost 1. In the case of the Home Management Plan of Care document, except for Cluster0, Cluster1, and Cluster3, compare prime numbers *89*, and crowded Cluster4 and Cluster2 so that there is a difference between average values such as *92,57,93,77,88.* I compared it and there is a big difference. As a result of the experiment, when the average value of HMPC (Home Management Plan of Care) in Cluster2 was high, 16 dogs and 16 dogs with a prevalence of 69% in the past were 8.6 or higher. Cluster4 cases, the past 8 cases were *4.4%~5.0%* prevalence of *62%.* The results suggest that urban prevalence in the past was higher, childhood asthma and historical regional HMPC prescription rates were nearly high, indicating that this is more effective, and that there is a relationship between prevalence and HMPC.

4.2 Evaluation and Discussion

Means and standard deviations were investigated in the treatment method. Average values converge with one corticosteroid drug and reliever drug. In the case of the Home Management Plan of Care, you can see that the average difference between the clusters is 10 or more. Cluster2 has 23 instances of *ID, MO, IL, FL, MI. , SC, VT, FL, NV, OH, KY, NJ,* etc., which account for 23% of *SD*, etc. As in the case of Cluster4, there are 13 instances of *WV, AL, VA, CT, AZ, GA, NC,* accounting for 13% of the total.

Table1. Cluster Instances of clusters

Cluster Name	Cluster Instances
Cluster0	1(1%)
Cluster1	6(6%)
Cluster2	23(23%)
Cluster3	3(3%)
Cluster4	13(13%)
Cluster5	55(54%)

In Cluster2 and Cluster4, the Home Management Plan of Care instance was assigned the 1980-2005 US Childhood Asthma Status. In the case of replacing Cluster2, it can be seen that 1980-2005 US Childhood Asthma Status, US Congested Home Management Plan document is 8 8.6%, Cluster4 is 4.4-8.5% than 16 instances (69%).

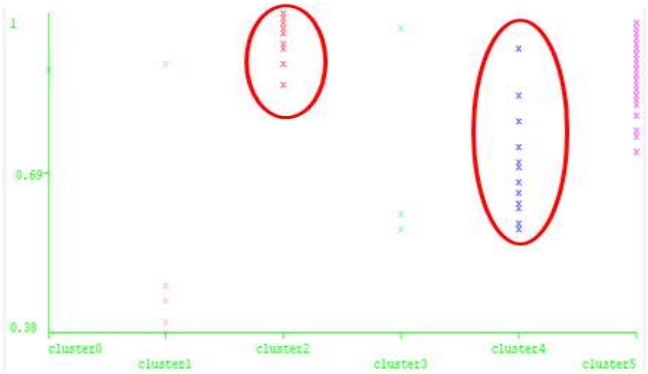

Figure1. The location of the data

V. Conclusion

In this paper, the prescribing probability of HMPC was high in hospital treatment measures-regional results, and the results of comparative experiments where children's asthma data were crowded were high, and the prevalence of childhood asthma was high in the past. As a result of treatment, many pediatric asthma patients who had a high prevalence in the past came to the conclusion that they prefer to prescribe HMPC unless the patient's condition is more systematically protected. Historical data tends to look a bit lacking, although experiments have been conducted to eliminate many factors that vary from hospital to hospital. Experiments should be attempted to cope with data such as environmental pollution, fine dust concentration, and missing data, as well as prevalence data by geographic location in the past.

References

[1] Ian H. Witten, Eibe Frank, Mark A. Hall, Data Mining Practical Machine Learning Tools and Techniques Third Edition, Morgan Kaufmann Publishers, 2011
[2] Jun-ho Lim, medical data mining using association rules, School of Computer & Information Technology Korea University, 2010
[3] Korea Research Society, the study of specimens correction and weight calculation of discharge patient survey, 2007
[4] Ltifi, Hela, et al. "A human-centred design approach for developing dynamic decision support system based on knowledge discovery in databases", Journal of Decision Systems, 2013
[5] Barnes, Sean, Bruce Golden, and Stuart Price. "Applications of agent-based modeling and simulation to healthcare operations management", Springer New York, 2013
[6] WEITZMAN, Michael, et al. Maternal smoking and childhood asthma. Pediatrics, 1990
[7] EGE, Markus J., et al. Exposure to environmental microorganisms and childhood asthma. New England Journal of Medicine, 2011
[8] MOFFATT, Miriam F., et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature, 2007
[9] STRUNK, Robert C., et al. Azithromycin or montelukast as inhaled corticosteroid–sparing agents in moderate-to-severe childhood asthma study. Journal of Allergy and Clinical Immunology, 2008
[10] BREHM, John M., et al. Serum vitamin D levels and severe asthma exacerbations in the Childhood Asthma Management Program study. Journal of Allergy and Clinical Immunology, 2010