WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Analyzing the Various Respiratory Problem of Children Using Data Mining Techniques

[1] K.Nithyanandan, [2] Dr.S.Prakasam

[1]Research Scholar, Department of Computer Science & Applications Sri Chandrasekharendra Saraswathi Viswa Maha Vidyalaya University, Kanchipuram – 631 561, Tamilnadu, India
[2]Associate Professor, Department of Computer Science & Applications Sri Chandrasekharendra Saraswathi Viswa Maha Vidyalaya University Kanchipuram – 631 561, Tamilnadu, India

A B S T R A C T

Abstract— To understand the impact of diseases in the children in various age groups and comprehend its effects on various organs in the children. To harness the power of data mining tool in analyzing huge reservoir of data set to derive better immunization programmes for a particular age and to optimum use of vaccination according to a disease pattern among children. It helps in combating the diseases and providing knowledge about the kind of diseases the children are prone to in their children. In an increasing era of child population in the county, children care is of foremost importance and health of every child should be continuously monitored. Various factors contribute to the vulnerability of child's health, Lifestyle changes of the Indian youth, is the root problem to the deterioration of health issues High toxic contents in the junk foods and chocolates that the children consume is a cause for concern. Environmental changes and bad food habits result in declining of immune system levels in children leading to easy target for viruses and infections. Outdoor air pollution and continuous exposure to ambient air pollutants like particulate matter are among the leading contributors to adverse respiratory health outcomes all over the world. This association between air pollution and the impairment of respiratory functions is evident from number of epidemiological studies. Health risk from particulate pollution is especially high for some risk groups such as children and elderly persons, and those with diseases of lungs. However, there are still many issues to be clarified before we know the real causal relationship between air pollution and health effects. Specific air pollutants have not been identified as causes of health effects. This specific study has been conducted with an objective to evaluate the effects of air pollution on respiratory symptoms and diseases of children.

Keywords: Air Pollution, respiratory illness, Decision Tree, Data mining, Classification

1 Introduction

Clean air is considered to be a basic requirement of human health and well-being. The effects of air pollution on health have shown effects ranging from minor eye irritations to upper respiratory symptoms, chronic respiratory diseases, cardiovascular diseases and lung cancer, that may result in hospital admission and even death The impacts of air pollution on human health can be assess in terms of a reduction in average life expectancy, additional premature deaths, absent in work place or school, hospital admissions and the increase use of medication and days of restricted activity. In developed countries, legislation and guidelines regarding the concentrations of air pollutants in ambient air has been established based on the epidemiological, toxicological and clinical evidence. However, in developing

* *Corresponding author*
E-mail address: nithik_01@yahoo.co.in

countries, studies on this matter have been ignored thus allowing levels that have been proved to have serious effects on human health, especially on children and elderly exposed to them. Assessment on air pollution behavior and their impacts on health will help decision makers to understand better its effects, as well as the benefits that could be achieved through the application of control measures. The causes of respiratory illness and air pollution depend on various factors including the pollutant emissions, atmospheric chemical processes, topography, meteorological conditions and solar radiation the complex mechanism of air pollution formation and respiratory effects makes it even more complex and difficult to control. In order to understand it is necessary to apply an intelligent approach that can describe the complex relationship between air pollution concentrations and the many variables that cause or hinder the respiratory effects. The complexity makes applying the conventional statistical analysis to air quality and respiratory illness as inefficient task as it mostly based on basic linear principles. Though the statistical methods may provide reasonable results, but these are essentially incapable of capturing the important knowledge of the complexity and non-linearity of the pollution-adverse impacts relationships. Therefore, it is expected that it will underperform when use to model the relationship between air pollution and the health effects that extremely non-linear. In the past few years, the collection of air quality and clinical data has generated an urgency need for new techniques and tools that can intelligently and automatically.

2 Research Objectives

To find the factors that influences spread of various respiratory diseases in children.
To find and formulate a strategy for the classifying the diseases age-wise and a means to correlate disease pertaining to an organ to an age.
What refinement needs to be done to existing algorithm for deriving an association between a disease with children's age.
How data mining tool can be used to ascertain the most vulnerable organs for a given age in children from the collection of dataset.

3 Methodology

Data mining is the process of discovering interesting knowledge, such as patterns, associations, changes, anomalies and significant structures, from large amounts of data stored in databases, data warehouses, or other information repositories. Mining association rules is one of the techniques involved in the process mentioned above and used in this paper. Association rules are the discovery of association relationships or correlations among a set of items. Association rule mining search for the interesting relationships among attributes in the database. Association rules are similar to classification rules except that they can predict any attribute and not just the class, and this allows them to predict combination of the attributes. Different association rules express different regularities that underlie the dataset, and they generally will predict different things. Because of so many interesting association rules can be derived from even a tiny dataset, interest is restricted to those that apply to a reasonably large number of instances and have a reasonably high accuracy on the instances to which they apply to. The coverage of an association rule is the number of instances for which it predicts correctly. This if often called its support. Its accuracy often called confidence is the number of instances it predicts correctly expressed as a proportion of all instances to which is applies.
For example, in this research, the rule Air Pollutant (T, "LESS") ==> Respiratory Illness (T, "LESS"), means if air pollutant, is less then, T, respiratory illness is less. The accuracy is the proportion of the days when air pollutant is less than the mean air pollution also has respiratory illness less than the mean respiratory illness, expressed in percentage or fraction. It is usual to specify minimum support (coverage) and the confidence (accuracy) values and to seek only those rules whose support and confidence are at least equal to these specified minima. Rules that satisfy both minimum support threshold and minimum confidence threshold are called strong. Generally, support and confidence values are expressed between 0% to 100% rather than 0 to 1.0. There are two methods for mining the form of association rules which is the Boolean association rules. One is a basic algorithm for finding frequent item sets and another one is the frequent pattern growth methods which adopts a divide and conquer strategy. Apriori algorithm for mining frequent item sets for Boolean association rules is used in the present study. The algorithm employs an iterative approach known as level-wise approach where k item sets are used to explore (k+1) item sets. In environmental, particularly on-air pollution association rules are useful to summarize pollutants levels into groups (categorized) and to build model for patient prediction.
In this study, it involves five major phases, namely (i) data selection, (ii) pre-processing, (iii) data mining; (iv) testing and evaluation; and (v) knowledge discovery, as shown in the framework of this study. Based on the framework, the stage of pre-processing and data preparation is done in two steps, which were during cleaning and integration of data collection and data selection and transformation. Pre-processing is done so that the generated rules at the end of the study will be the certainty and reliable rules as a knowledge based. In this stage, several phases have been carried out, which were the data integration, data cleaning, attribute selection and data reduction. Data cleaning is required when there are incomplete attributes or missing values in data. It involved filling the missing values, smoothing noisy data, identifying outliers and correcting the data inconsistency. Data integration combines data from multiple sources to form a coherent data store. Metadata, correlation analysis, data conflict detection and resolution of semantic heterogeneity contribute towards smooth data integration. Data transformation converts the data into appropriate forms for data mining that depends on the mining technique. In the case of developing a knowledge-based model, data are required to be discretized. This is because the rough

classification algorithm only accepts categorical attributes. Discretization involves reducing the number of distinct vales for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values.

Algorithm:

Step 1: Enter the text
Step 2: Predicting system will check for the condition.
Step 3: System predicts the values based on the user answers.
Step 4: The range of the risk is determined based on the predicted value.
Step 5: If the value is ≤18 the risk is considered as a low risk. If the value >18 and ≤ 21 the risk is considered as an intermediate risk. If the value is >21 and ≤28 is considered as a high risk. If the value is > 28 is considered as a very high risk.
Step 6: The user data is stored in data base.
Step 7: The result is obtained with the reference values of the data base.

Rules for Decision Tree

A decision tree is a flow chart like tree structure. ach branch represents an outcome of the test and each leaf node holds a class label. The top most node is the root node. The attribute value of the data is tested against a decision tree. A path i traced from root to leaf node, which holds the class prediction for that data. Decision trees can be easily converted into classification rules. This decision tree is used to generate frequent patterns in the dataset. The data and item sets that occur frequently in the data base are known as frequent patterns. The frequent patterns that is most significantly related to specific cancer types and are helpful in predicting the cancer and its type is known as Significant frequent pattern. Using these significant patterns, generated by decision tree the data set is clustered accordingly and risk scores are given.

If symptoms = none and risk score x = <35 then result = you may not have cancer, tests = do simple clinical tests to confirm.
If symptom= related to chest and shoulder and risk score x >= 40 then result = you may have cancer, cancer type may be= chest, tests = take CT scan of chest.
If symptom= related to head and throat and risk score x >= 40 then result = you may have cancer, cancer type = oral, tests = biopsy of tongue and inner mouth.
Else symptom= other symptoms and risk score x >= 40 then result = you may have cancer, cancer type = leukemia, tests = biopsy of bone marrow.
Else if symptom= related to stomach d risk score x >= 45 then result = you may have cancer, cancer type = stomach, tests = endoscopy of stomach.
If symptom= related to breast and shoulder and risk score x >= 45 then result = you may have cancer, cancer type = breast, tests= mammogram and PET scan of breast.

4 Agewise Classification Of Respiratory Diseases

Respiratory problem is a global health problem and the prevalence is increasing all over the world. It's a lung disease with following characteristics:

•Airway obstruction which is reversible either spontaneously or with treatment
•Airway obstruction which is reversible either spontaneously or with treatment
•Airway inflammation
•Airway Hypersensitivity

Respiratory problem is a disease that causes the airways of the lungs to tighten and swell. It is common among children and teenagers. These respiratory problems are affected through the lungs are not getting sufficient air to breathe and the child may cough or wheeze during an attack.

Classifying respiratory Diseases:

According to the age group of 3-5, 5-10, and 8- 12 yrs. we classify the respiratory diseases are intrinsic and extrinsic. The classification is further extended to group as severe or general. In the. arff file format various symptoms like wheezing, coughing, shortness of breath, chest tightness. If symptom= related to pelvis and lower hip and risk score x >=55 then result = you may have cancer, cancer type = cervix, tests = do pap smear test Based on the above mentioned rules and the calculated risk scores the severity of cancer is known as

well outputs the result in the form of decision tree indicating the best possible treatment. When we are actually able to classify the respiratory diseases according to the severity present.

The Various kind of Respiratory Diseases are as follows :

- Asbestosis. Asthma.
- B. Bronchiectasis. Bronchitis.
- C. Chronic Cough. Chronic Obstructive Pulmonary Disease (COPD) Common Cold.

Croup. Cystic Fibrosis.

- H. Hantavirus.
- I. Idiopathic Pulmonary Fibrosis. Influenza.
- L. Lung Cancer.
- P. Pandemic Flu. Pertussis. Pleurisy. Pneumonia. Pulmonary Embolism.
- R. Respiratory Syncytial Virus (RSV)

This study of assessing respiratory effects was focused with following objectives

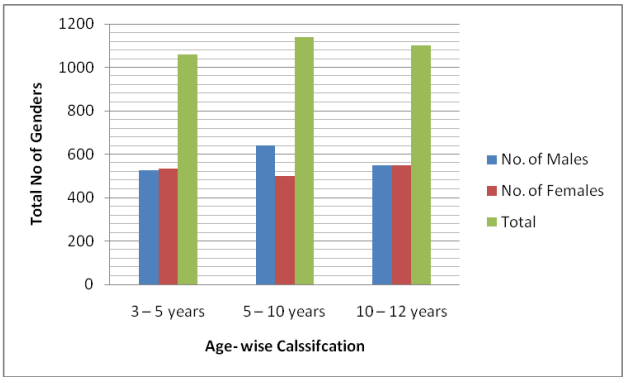
- To determine the effects of ambient air pollutants on respiratory health of children in rural and urban area.
- To determine the prevalence of respiratory symptoms and diseases (Asthma, Chronic Cough, Hanta Virus, Pneumonia etc.) among children.
- To understand the impact of diseases in the children of various age groups
- To find out the various respiratory diseases for the children are affected at winter and summer seasons
- To find out the children who are more affected male or female

5. Metrics and research hypotheses

- To understand classifier's behavior, we should calculate those metrics, we use hypothesis below:
- • True positive (TP) is the number of positive samples correctly predicted.
- • True negative (TN) is the number of negative samples correctly predicted
- False negative (FN) is the number of positive samples wrongly predicted.
- False positive (FP) is the number of negative samples wrongly predicted as positive.
- It is a visualization tool which is commonly used to present the accuracy of the classifiers in classification.

Table 1 : Proportion of patients with respiratory symptoms among all patients (aged 3 to 12 years) who visited primary health care facilities for various reasons of respiratory problems

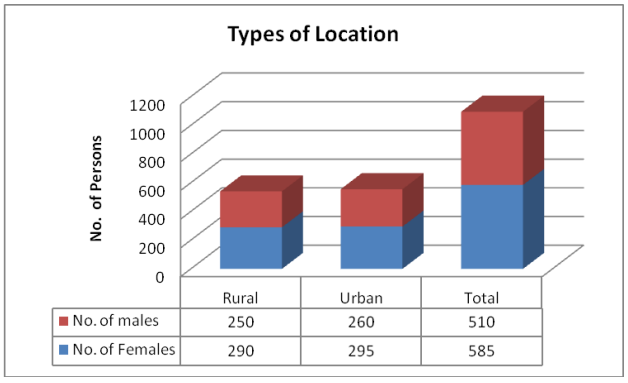
Gender	3 – 5 years	5 – 10 years	10 – 12 years
No. of Males	525	640	550
No. of Females	535	500	550
Total	1060	1140	1100



Respiratory symptoms are among the major causes of consultation at primary health care centers. Surveys in children from rural and urban areas. The number of primary health care facilities, involving 3300 respiratory patients, showed that the proportion of patients with respiratory symptoms, among those over 12 years of age, who visited primary health care centers ranged from 8.4% to 37.0% (Table1).

Table 2: Types of Location that the children are affected from respiratory diseases.

Types of Location	No. of Males	No. of Females
Rural	250	290
Urban	260	295
Total	510	585



Finding of Field Study:

The data included primary data collection from children to assess the lung function among children, flow meters were used and readings were recorded. 3300 children were selected in this study. Data was analyzed using frequency tables, cross tab analysis and chi-square test to show significance. There is a significant effect of ambient air pollution on respiratory symptoms of female children with high prevalence of the symptoms in the study area which is the industrial area than the control area. In this disease shows that close to 40% of children had dry cough in study area, the percentage of which in control area was 31.3%. More than 33% of children in study area had night cough. More children in study area 7.9% had sore throats as compared to the children in control area 3.9% More children in study area 5.2% had presence of wheeze as compared to the children in control area 5.5%. More children in study area 6% had asthma as compared to the children in control area. Phenomena were very low. There will be a significant difference between male and female children.

Tools for the study:

In this study children are affected from respiratory diseases; they are analyzed through the SPSS and Weka software tool. Using SPSS software frequency and Chi squared test are performed. Using Weka software major data mining tasks are Attribute selection, Association, Clustering, Classification performed

6.Conclusion

This paper has given a promising and valuable contribution especially to the respiratory problem management. It is the first attempt using the association rules in trying to understand the air pollution formation to its effect to the respiratory illness, thus to solve environmental issue. The knowledge model obtained can be used as a decision support system to gain sets of knowledge that is useful in terms of preventing the elevated exposure of the hazardous air pollutants that gives more impact on the respiratory illness. From the association rules base knowledge as well, we can know what are the best combinations or associations of the air pollutants that contributes to higher health risk, so that more action plans can be done in resolving the problem. Association rule data mining produces knowledge that is understandable and can be interpreted easily. This system is validated by comparing its predicted results with the patient's prior medical record and also this is analyzed using WEKA tool. This prediction system is available in online, people can easily check their risk and task appropriate action based on their risk status.

REFERENCES

[1] Choi, K., Inou, S. and Shinozaki, R. 1997. Air pollution, temperature, and regional differences in lung cancer mortality in Japan. Archaeology Environmental Health, Vol. 52, pp. 160-168.

[2] Creason, J., Neas, L., Walsh, D. and Sheldon, L. 2001. Particulate matter and heart rate variability among elderly retirees. Journal of Exploratory Analytical Environmental Epidemiology, Vol. 11, pp. 116-122.
[3] EEA. 2007. Europe Environment: The 4th Assessment. European Environment Agency, Copenhagen.
[4] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R.1996. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press.
[4] Ritu Chauhan "Data clustering method for Discovering clusters in spatial cancer databases" International Journal of Computer Applications (0975-8887) Volume 10-No.6, November 2010.
[5] Dechang Chen "Developing Prognostic Systems of Cancer Patients by Ensemble Clustering" Hindawi publishing corporation, Journal of Biomedicine and Biotechnology Volume 2009, Article Id 632786.
[6] Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems) Hardcover – Import, 25 Jul 2011 by Jiawei Han (Author),
Micheline Kamber, (Author), Jian Pei Professor(Author)
[7] Data Mining and Analysis: Fundamental Concepts and Algorithms Hardcover – Import, 12 May 2014 by Mohammed J.Zaki (Author), Wagner Meira Jr(Author)
[8] Understanding Machine Learning: From Theory to Algorithms Hardcover – Import, 19 May 2014 by Shai Shaley-Shwartz (Author), Shai Ben-David(Author)
[9] Bayesian Data Analysis, Third Edition (Chapman & Hall/CRC Texts in Statistical Science) Hardcover – Import, 5 Nov 2013 by Andrew Gelman(Author), John B.Carlin (Author), Hal.Stern, (Author)
[10] Data Analysis Using Regression and Multilevel/Hierarchical Models (Analytical Methods for Social Research) Paperback – 18 Dec 2006 by Andrew Gelman(Author), Jemmofer Hill (Author)
[11] Mitchell T.M., Machine learning, McGraw-Hill, 1997.
[12] Pal S.K. and Mitra P., Pattern Recognition Algorithms for Data Mining, CRC Press, 2004.
[13] Tan P.-N., Steinbach M. and Kumar V., Introduction to Data Mining, Addison Wesley, 2006.
[14] Webb A., Statistical Pattern Recognition, Wiley, 2002.
[15]Witten I. and Frank E., Data Mining: Practical machine learning tools and techniques with Java implementations, Morgan Kaufmann Publishers, 2000.