International Journal of Grid and Distributed Computing
Vol. 9, No. 9 (2016), pp.317-326
http://dx.doi.org/10.14257/ijgdc.2016.9.9.27

Cloud Based Data Mining Model for Asthma Diagnosis

Abhinav Hans[1], Sukhdeep Kaur[2] and Navdeep Singh[3]

[1]Department of Computer science, CT Group of Institutions, Jalandhar, Punjab, India
[2]Department of Computer science, DAVIET, Jalandhar, Punjab, India
[3]Department of FCS, GNA University, Phagwara, Punjab, India
abhinavhans@gmail.com, Sukhdeepkaur343@gmail.com, Navvdeep.singh@gmail.com

Abstract

The potential of cloud computing for overriding the needs for deploying various infrastructures for running a server based services brought up a revolutionary change in the way the traditional demands of the people use to be handled. Cloud computing provides the software's on pay per use option which make it easier for the people who are not able to meet the economic requirements. Since the whole scenario is beneficial to big industries like facebook, google, orkut etc , various order fields are also getting dependent on cloud computing. Since tons of data is uploading every second to the cloud server does needs to be mined properly for efficient data storage.in this paper we try to integrate the data preprocessing technique with data classification technique to mine big data's of asthma based patients.

Keywords*: Cloud computing, Data preprocessing, Data Classification, Cloud Computing*

1. Introduction

''Cloud" computing has been receiving much attention as an alternative to both specialized grids and to owning and managing one's own servers[1].with development of new technologies like wireless sensor network, grid computing, soft computing etc cloud computing among all is considered to be the most powerful technology developed. The property of providing the software's on lease allows the cloud to overcome the other technologies. The benefit of prepaid service of cloud computing allowed the resources to be accessible by any type of the user anywhere and anytime. Various factors like scalability, low recovery cost, less maintenance, huge data storage provisions, speedy deployment and many more factors make cloud the most powerful approach. Since the cloud is associated not only with the information technology, but also with many other fields that in the human health, sales and management files too. If health care could be provided remotely and the patient could be monitored continuously without having to go to hospitals, expensive medical costs could be saved [2].

Extracting a quite useful knowledge from various numbers of data sources that gives a necessary information by separating patterns,symbols,attributes etc is known as data mining. Data analyses process analyzes the various data resources to create and put them into separate classes of datasets. Data mining is a multi-disciplinary field which is a combination of machine learning, statistics, database technology and artificial intelligence. This technique includes a number of phases: Business understanding, Data understanding, Data preparation, Modeling, Evaluation, and Deployment. Various social media websites are totally dependent on data mining process as the amount of data that gets uploaded every single hour is in tons of terabytes so it makes data mining process to be very much significant. Although not any social media websites but many different sections like

ISSN: 2005-4262 IJGDC

hospitals ,IT industries, online shopping etc. are also bringing data mining as a significant approach in there data maintenance job . Association, Classification, Clustering, Neural Network and Regression are the various data mining techniques used for mining the data.

2. Asthma Overview

Type of asthma is determined by the source of bronchial hypersensitivity: allergic asthma (atopic, extrinsic, caused by immunologic stimulus of an antigen), intrinsic (non-allergic, induced by infection, physically or chemically), exercise induced, drug induced asthma, occupative asthma and asthmatic bronchitis [3].

According to the earlier definition of the ATS (American Thoracic Society), sounds are considered "continuous" if their duration is longer than 250 ms; otherwise they are considered "discontinuous" [10]. High-pitched continuous sounds (dominant frequency above 400 Hz) and rhonchi as low-pitched continuous sounds (dominant frequency of 200 Hz or less) is considered as wheeze according to the ATS.

According to the new definition of CORSA (Computerized Respiratory Sound Analysis) guide- lines, the dominant frequency of wheeze is usually above 100 Hz and the duration greater than 100ms [10].

Wheezes are continuous adventitious sounds, which are superimposed on normal breath sounds and often associated with bronchial airway obstruction. There are many circumstances leading to wheezing. They include all mechanisms narrowing airway caliber such as bronchospasm, mucosal edema, external compression by a tumor mass, or dynamic airway obstruction [14].

3. Literature Survey

Now a days Asthma is the most ascent disease in which age factor doesn't matters i.e. it can be in people of any age group. Growing popularity increments the concern of personal disease administration, commercial expense and workload, on the both sides i.e. of patient's side and healthcare systems side. In [11] author has presented stepwise the background of asthma in medical terms followed by information about the Pathology and symptoms later. After that author has highlighted some the drawbacks of the existing techniques for managing asthma by emphasizing on showing the importance disease management techniques in traditional way. A tele-monitoring technique on glide paths to asthma is done.

Asthma is a very common disease and can give the very harmful effects if it not taken seriously. By taking its serious impact on health, a continuous monitoring is must to check the body and respiration behavior of the patient. The most countable factor is the environment in which they breathe. So In [12] author proposes a development of a rule-based asthma system. So according to it the patients are given various suggestions on the possibilities of occurring an asthma attack according to patient's current body conditions and the environment in which they breathe. The system is based on questioning process to the patient and answer given by patients defines the patient's present health condition and the environmental condition in which they living in.

This research work lights on the data mining procedure in which diabetes disease can be predict on the basis of medical record history of patient. Diabetes is very common disease that can happen in any age group. It is a serious disease that makes serious impact on heart, kidneys, nervous system, blood line and vessels. But mining the data of diabetes patient in efficient manner is a critical issue. The Pima Indians Diabetes Data Set is used in this paper; which collects the information of patients with and without having diabetes. For data mining procedure modified J48 so its accuracy rate can be increased. The data mining tool WEKA has been used as an API of MATLAB for generating the J-48 classifiers. Experimental results showed a significant improvement over the existing J-48 algorithm[10].

Data mining is the process that involves the symmetric analysis of data sets at large scale and data mining in agricultural soil datasets is exciting and modern research area. In this research work[7], Steps for building a predictive model of soil fertility have been explained. This paper aims at predicting soil fertility class using decision tree algorithms in data mining. Further, it focuses on performance tuning of J48 decision tree algorithm with the help of meta-techniques such as attribute selection and boosting.

Database used for storing large amount of data and data mining the process for maintaining the large amount of data of same. For classification of data and products, the technique of decision tree in used to get valuable result. And these results can be used for analysis and future prediction. In [8] paper the

author made an objective to present the enhanced decision tree algorithm that classifies the data. The tree classifiers used in this work are ID3, J48, NBTree on a large amount of data. Then the efficiency and performance of existing algorithms is examined and compared with new enhanced decision tree algorithm (NEDTA).

4. Proposed Approach

There are various numbers of approaches that works on asthma patients but are not capable enough to overcome the various problems of data missing values and classification problems. So in our proposed approach we try to integrate the data preprocessing approach had the classification approach to build da powerful setup for data mining on the asthma patients. The cloud based asthma detection system allows users to upload data from different parameters automatically with the help of body sensors and classifies as asthmatic and non-asthmatic. The system consists of (1) data pre-processing (2) attribute Extraction (3) data classification.

4.1 Data preprocessing: Present time health care gadgets allow patients to upload the data recorded by equipment automatically. Most of the time when connectivity is weak or there is a noise component present the data may get corrupted or there the available data may have some values missing which are necessary for pre-processing. Therefore, to deal with these problems we use data interpolation techniques non uniform data into uniform sample data. Based on the assumption that asthma profiles at the same time on different days are usually similar to each other, we interpolate missing values using Lagrange Two-Dimensional Interpolation Method [16]. 2D-Lagrange interpolation is based on 1D-Lagrange interpolation. In this method, one of the variables is forced to be constant and, with another variable, the Lagrange polynomials can be written by using the given data. Then, this value can be complicated for the final form of 2D-Lagrange interpolation. The result is a 2D-Lagrange polynomial whose functional agents are replaced by Lagrange polynomials

According to this theorem function is defined as:-

$$f(X_1, \dots, X_m) = \sum_{e_i \cdot 1 \leq \mathrm{n}} \alpha_{e_i} \mathrm{X}^{e_i} \tag{1}$$

Where
α_{e_i} are the coefficients$f$, $X = (X_1, \dots, X_m)$ is m-tuple of independent variables of f, e_i=($e_{1i}, \dots, e_{mi}$) is an exponent vector of nonnegative integers of ordering partition of integer between 0and and inclusive, $e_i . 1 = \sum_{j=1}^{m} e_{ji}$ is usual vector dot product,$\mathrm{X}^{e_i} = \prod_{J=1}^{M} X_J^{e_{ji}}$.

According to langrage the $f\ can\ be\ defined\ as\ f = \sum_{i=1}^{p} f_i l_i(X)$ where $l_i(X)$ Is the multinomial function in the independent variables $X_1, \ldots, X_m$ with the property that when X = i[th] data value, or X=$X_i(X_1, \ldots, X_m) = (x_{1,i}, \ldots, x_{m,i})$ then $l_i(X_i) = 1$ and $l_i(X_i) = 0\ (j \neq i)$.

Consider the system of linear equation:

$$f_i = \sum_{e_i \cdot 1 \leq n} \alpha_{e_i} X^{e_i} \tag{2}$$

Where $1 \leq i \leq \rho$. From this system construct the sample matrix M = $[X_i^{e_j}]$:

$$M = \begin{pmatrix} X_1^{e_1} & \cdots & X_1^{e_\rho} \\ \vdots & & \vdots \\ X_i^{e_1} & \cdots & X_i^{e_\rho} \\ \vdots & & \vdots \\ X_\rho^{e_1} & \cdots & X_\rho^{e_\rho} \end{pmatrix}$$

If M is singular, then the coefficients of f are not uniquely determined, in which case f is clearly not unique. Therefore, f is unique if and only if its sample matrix is nonsingular. On the Other hand, characterizing the geometric configuration of the ρ points so that det (M) = 0 appears to be an intricate problem

Let Δ = det(M). Now make the substitutions X_j = X in M; this gives the following matrix M_j (X):

$$M_j(X) = \begin{pmatrix} X_1^{e_{1i}} & \cdots & X_1^{e_\rho} \\ \vdots & & \vdots \\ X^{e_{1i}} & \cdots & X^{e_\rho} \\ \vdots & & \vdots \\ X_\rho^{e_{1i}} & \cdots & X_\rho^{e_\rho} \end{pmatrix} \leftarrow jth\ row$$

LetΔ_j (X) = det(M_j (X)). Next, make the substitutions X = X_i in M_j (X) (i ≠ j); this gives the following matrix $(M_j)_i$

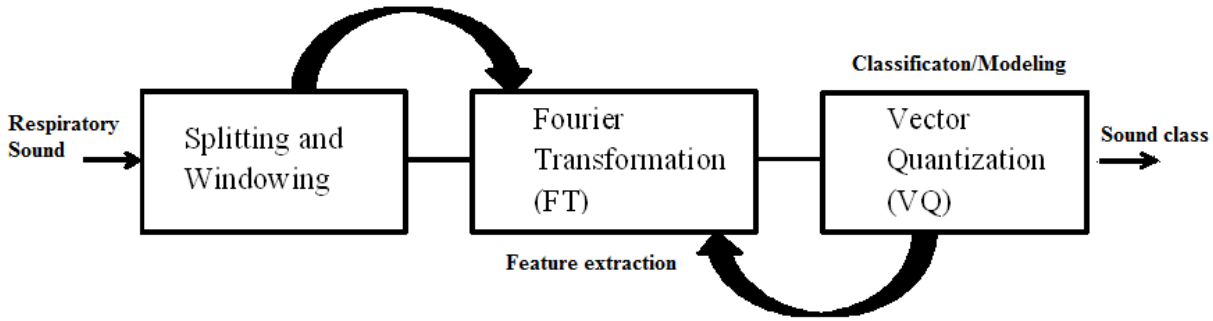

Figure 3. Respiratory Sound Classifier Obtained through Feature Extraction Classification Technique

$$(M_j)_i = \begin{pmatrix} X_1^{e_{1i}} & \cdots & X_1^{e_\rho} \\ \vdots & & \vdots \\ X_i^{e_1} & & X_i^{e_\rho} \\ \vdots & \cdots & \vdots \\ X_i^{e_1} & & X_i^{e_\rho} \\ \vdots & & \vdots \\ X_\rho^{e_{1i}} & \cdots & X_\rho^{e_\rho} \end{pmatrix} \leftarrow ith \leftarrow jth$$

Note that the i th row appears twice in $(M_j)_i$. That means det(($(M_j)_i$)= 0. In other words, when X = X_i then Δ_j (x_i) = 0 (i ≠ j). By construction, moreover, X = $x_i \Rightarrow \Delta_i$ (X) = Δ. Hence

$$l_i(X) = \frac{\Delta_i(X)}{\Delta} \tag{3}$$

and therefore

$$f = \sum_{i=1}^{\rho} f_i \frac{\Delta_i(X)}{\Delta} \tag{4}$$

4.2 Feature Extraction: The fundamental function of a feature selector is to extract the most useful information from the data, and reduce the dimensionality in such a way that the most significant aspects of the data are represented by the selected features [9]. The human operated system often consists of attribute variables which are not required for the job classification. Therefore, we used a feature extraction technique for extracting useful features from the huge amount of data. Specially, we have utilized the Fourier extraction technique for identifying Where w[n] is the short-time windowing function of size L, gathered at time, location m and N is the number of distinct frequencies. Power spectrum density is used for simplifying the complex Fourier transformation which is stated as:

$$P_s[m,k] = \frac{1}{N}|S[m,k]|^2 \tag{6}$$

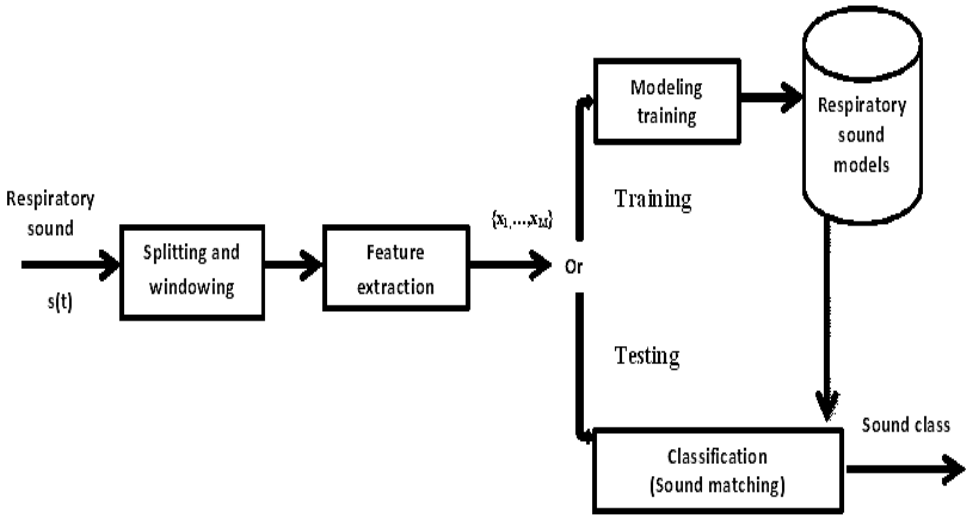

Figure 4. Block Diagram of Respiratory Sound Classifier

The windowed signal (frame) at sampling frequency Fe is characterized by N-point power spectrum, which integumenting the frequency domain [−*fS*/2, *FS*/2]. the advantage of the power spectrum is that it can accommodate the large size data (N/2 components) but due to this advantage it cannot be used directly as feature vectors. Therefore, to classify the asthmatic breath, RIETVELD AT [8] the power spectra must be from shorter intervals of almost 3 sec, defining the full breathing cycle.so in order to make shorter intervals the frequency range from 100-1300 hrz is divided into 25 bands of 46 hzs each hence 26 new components regenerated.to find the power spectra of these 26 components we find the average of the power spectra in each band. We have used R-FT method as the feature vector which is defined as:

$$x = \overline{[P_1}, \overline{P_2} \ldots, \overline{P_{26}}]^T \quad (7)$$

Where t is the transpose operation and *Pk is the* average power spectrum in the *k* th bandf.

4.3 Patient Health Condition Classification: The closing step is to analyze the health condition of the users based on the extracted data through the Fourier extraction technique. There are many classification accuracy metrics which are discussed in [17].In our model we have used vector quantization as a classifying technique. This classification mechanism is implemented in our system to assort the class tag of the users.In the underneath subsidiary, a brief description of the classification technique used in our system is presented. The process of mapping vectors from a large space into a finite space, number of regions is called Vector Quantization and each region is called clusters in that space. The clusters can be expressed by codeword [18] which is the centroid **ci .**The codewords aggregate the form the codebook C = {**c1**,**c2**, ..., **cN**} for defining a respiratory sound class. Classification system consists of two stages: training and recognition (see Fig. 4). Training section consists of the process of generating an acoustical model(codebook) for each respiratory sound class and are then stored in a database .so ,there are J *codebooks* C1,C2, ...,C*J provoked* for *J* reference sounds.TO generate the codebook many different algorithms exist among which the most broadly used is LBG (Linde–Buzo–Gray) algorithm [19]. In recognition section, each codebook is computed from the distorted sets of testing feature vectors Y = {**y1**, y**2**... y**M**} through

which an average quantization distortion *Qk* to the *J* th codebook (C*J*) is performed [31], according to

$$Q_k = \frac{1}{M}\sum_{i=1}^{M} \min_{1\le j\le N} d(x_i, c_j) \tag{8}$$

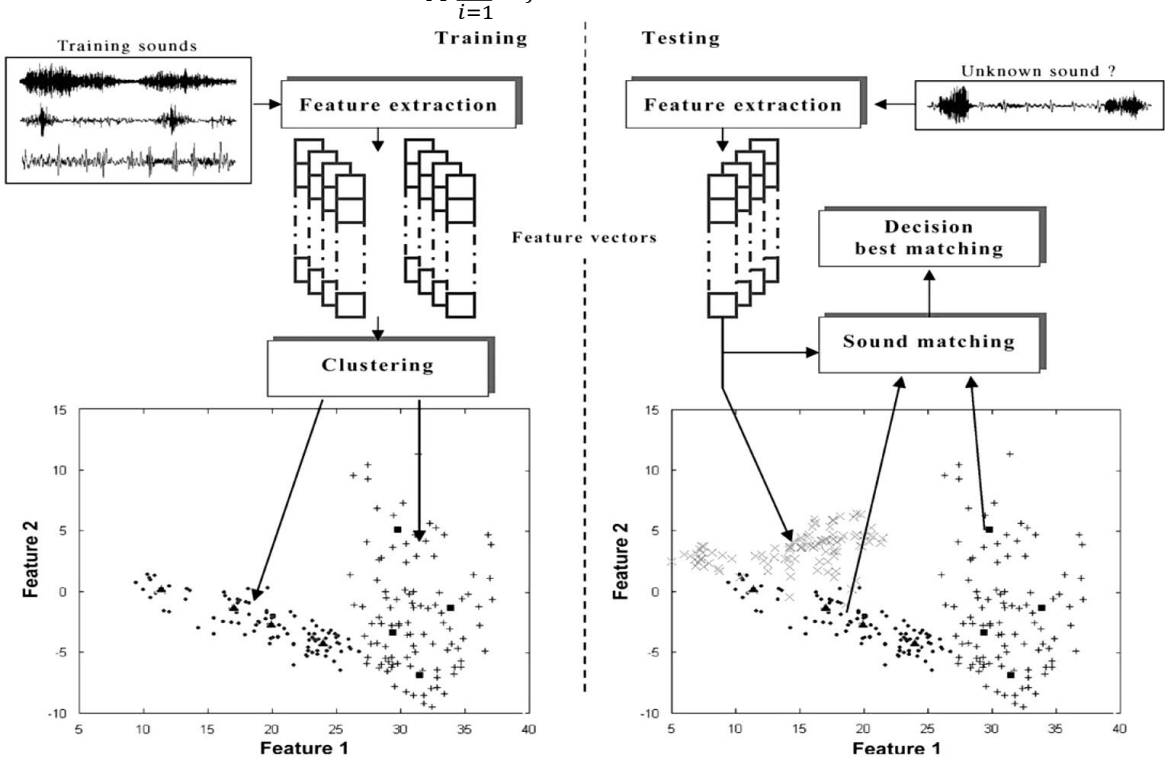

Figure 4. Block Diagram of VQ-based Classifier in the Feature Space [18]. The Feature Vectors of the Normal Sound (·) are Represented by Four (▲) Centroids, While those of the Wheezing Sound (+) are Modeled by Four Centroids (■). The Unknown Respiratory Sound (×) is Compared to the Codebook of these Classes

Where $d(\mathbf{x}i, \mathbf{c}j)$ is the distortion measure(usually a Euclidean distance) between the input vector x**i** and a centroid **c***j* of the *k* th code- book (C*k*). The unknown sound is then identified as the reference sound with the minimum average distortion measure

$$\hat{k} = arg \min_{1\le k\le K}\{Q_k\} \tag{9}$$

Where *K* is the number of the sound classes. In this study, unknown sound is classified segment-by-segment
($M = 1$).
Two class classifications can be done among this for that we attempt to classify respiratory sounds in two classes, $k \in \{normal, wheezing\}$. To achieve that, we propose to define a score function $Sc(X)$ by the difference of distortions:

$$Sc(X) = Q_{normal} - Q_{wheezing} \tag{10}$$

The classification decision is made by comparing the score function $Sc(X)$ to a threshold θ

$$Sc(X) > \theta \text{ (wheezing)} \tag{11}$$

And

$$Sc(X) < \theta \text{ (normal)}. \tag{12}$$

Eq. (9) is a particular case of Eq. (11) and Eq. (12), where $\theta = 0$.

5. Conclusion and Future Work

In this paper, we have presented an intelligent system for detection of user health data collected through advanced body sensors. It applies a Fourier transformation (FT) for feature extraction and Vector Quantization (VQ) for user health status classification. Also we have classified the user's health as asthmatic and non-asthmatic through our classification technique. The main emphasis was given to the user's health status classification, but a lot work can be done in the resource elasticity and load balancing of the system too, to make it more powerful approach. Our system can provide high classification with greater sensitivity and specificity measures which proving it to be highly useful for disease identification in the real life grand.

References

[1] Arnon Rosenthal, Peter Mork , Maya Hao Li, Jean Stanford, David Koester, Patti Reynolds "Cloud computing: A new business paradigm for biomedical information sharing" Journal of Biomedical Informatics.
[2] Edmund Y. W. Seto, Annarita Giani, Victor Shia, Curtis Wang, Posu Yan, Allen Y. Yang, Michael Jerrett, Ruzena Bajcsy " A Wireless Body Sensor Network for the Prevention and Management of Asthma" Industrial Embedded Systems, 2009. SIES '09. IEEE International Symposium on 8-10 July 2009.
[3] Dinko Oletic" Wireless sensor networks in monitoring of asthma"
[4] Suraj Pandey, William Voorsluys, Sheng Niu, Ahsan Khandoker, Rajkumar Buyya" An autonomic cloud environment for hosting ECG data analysis services" Future Generation Computer Systems.
[5] Marina Zapater , Patricia Arroba , José L. Ayala , José M. Moyab, Katzalin Olcoz " A novel energy-driven computing paradigm for e-health scenarios" Future Generation Computer Systems.
[6] Giancarlo Fortino , Daniele Parisi , Vincenzo Pirrone , Giuseppe Di Fatta " BodyCloud: A SaaS approach for community Body Sensor Networks" Future Generation Computer Systems.
[7] Fabricio F. Costa" Social networks, web-based tools and diseases: implications for biomedical research " Drug Discovery Today Volume 18, Numbers 5/6 _ March 2013.
[8] S. Rietveld, M. Oud, E.H. Dooijes"Classification of asthmatic breath sounds: preliminary results of the classifying capacity of human examiners versus artificial neural network" Computers and Biomedical Research 32 (1999) 440–448.
[9] M. Dash, H. Liu "Consistency-based search in feature selection" Artificial Intelligence 151 (December (1–2)) (2003)155–176.
[10] A.R.A.Sovij¨arvi,L.P.Malmberg,G.Charbonneau,J.Vanderschoot,F.Dalmasso,C.Sacco,M.Rossi,J.E.Earis "Characteristics of breath sounds and adventitious respiratory sounds"EuropeanRespiratoryReview10(2000)591–596.
[11] T. Wang, K. Shao, Q. Chu "Automics: an integratedplatform for NMR-based metabonomics spectral processingand data analysis" BMC Bioinformatics 10 (2009) 83.
[12] J.S. Dhillion, B.C. Wunsche, C. Lutteroth " Leveraging Web 2.0and consumer devices for improving elderlies' health", HIKM11 (2011) 120
[13] J.S. Ash, M. Berg, E. Coiera, " Some unintended consequences of information technology in health care: the nature of patient care information system-related errors", Journal of the American Medical Informatics Association 11 (2004)104–112.
[14] N. Meslier, G. Charbonneau, J.L. Racineux, "Wheezes", European Respiratory Journal 8 (11) (1995) 1942–1948.
[15] Judith W Dexheimer, Thomas J Abramo, Donald H Arnold, MPH Kevin Johnson , MS Yu Shyr, Fei Ye, Kang-Hsien Fan Neal Patel , MS Dominik Aronsky , " Implementation and Evaluation of an Integrated Computerized Asthma Management System in a Pediatric Emergency Department: A Randomized Clinical Trial" International Journal of Medical Informatics.
[16] R. Bozorgmanesh a_, M. Otadi b, A. A. Safe Kordi c, F. Zabihi d, M. Barkhordari Ahmadi" Lagrange Two-Dimensional Interpolation Method for Modeling Nanoparticle Formation During RESS Process" Int. J. Industrial Mathematics Vol. 1, No. 2 (2009) 175-181.

[17] M. Sokolova, G. Lapalme, "A systematic analysis of performance measures for classification tasks", Information Processing and Management 45 (2009) 427–437.
[18] M. Bahoura, C. Pelletier, "New parameters for respiratory sound classification",Electrical and Computer Engineering, 2003, Canadian Conference on IEEE CCECE 2003, vol. 3, Montreal, Canada, May 4–7, 2003, pp. 1457–1460.
[19] Y. Linde, A. Buzo, R.M. Gray, "An algorithm for vector quantizer design", IEEE Transactions on Communications 28 (1) (1980) 84–95.
[20] T. Pham, M. Wagner, "Ambiguity reduction in speaker identification by the relaxation labeling process, Pattern Recognition 32 (7) (1999) 1249–1254.

