Jurnal
Teknologi

Full Paper

Performance Analysis of Feature Selection Method Using ANOVA for Automatic Wheeze Detection

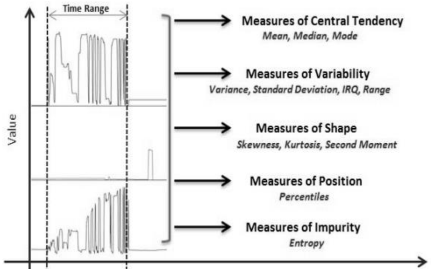
Syamimi Mardiah Shaharum[a*], Kenneth Sundaraj[a], Khaled Helmy[b],

[a]Al-Rehab Research Group, Universiti Malaysia Perlis (UniMAP), Kampus Pauh Putra, 02600 Perlis, Malaysia
[b]Hospital Tuanku Fauziah, Jalan Kolam, 01000 Kangar, Perlis, Malaysia

Article history
Received
19 June 2015
Received in revised form
26 June 2015
Accepted
10 July 2015

*Corresponding author
mimishaharum@yahoo.com

Graphical abstract



Abstract

In this work, we show that the classification performance of a high-dimensional features data can be improved by applying feature selection method. One-way ANOVA were utilized and to evaluate the performance measure of the feature selection method, Artificial Neural Network (ANN) was used. From the results obtained, it can be concluded that ANN performance using feature that undergo feature selection method produce a better classification accuracy compared to the ANN performance using feature that did not undergo feature selection method with 93.33% against 80.00% accuracy achieved. Therefore can be conclude that feature selection is a process that is crucial to be done in order to produce a good performance rate.

Keywords: Neural network, one-way ANOVA, statistical features, wheeze detection

Abstrak

Dalam karya ini , kami menunjukkan bahawa prestasi klasifikasi daripada ciri dimensi data yang tinggi boleh bertambah baik dengan menggunakan kaedah pemilihan ciri. ANOVA sehala telah digunakan untuk kaedah pemilihan ciri, dan Rangkaian Neural Buatan (ANN) telah digunakan untuk menilai dan mengukur prestasi keseluruhan. Daripada keputusan yang diperolehi, dapat disimpulkan bahawa prestasi ANN menggunakan ciri yang menjalani kaedah pemilihan ciri menghasilkan ketepatan pengelasan yang lebih baik berbanding dengan prestasi ANN dengan menggunakan ciri-ciri yang tidak menjalani kaedah pemilihan dengan 93.33% berbanding 80.00% ketepatan dicapai. Oleh itu boleh di simpulkan disini bahawa pemilihan ciri adalah satu proses yang sangat penting yang perlu dilakukan untuk menghasilkan kadar prestasi keseluruhan yang lebih baik.

Kata kunci: Rangkaian neural tiruan, ANOVA sehala, ciri statistic, pengesanan bubar

1.0 INTRODUCTION

There has been a relatively long history of automatic wheezes detection starting from some prepositions given in the 90's [1]. Wheezes are continuous adventitious sounds that are existing in the breathing sound and they have been of such an interest to the researchers due to the fact that the presence of wheezes is relatively related to respiratory diseases such as asthma, Chronic Obstructive Pulmonary Disease (COPD) and bronchitis. According to the American Thoracic Society (ATS), the duration of wheeze is commonly longer than 100 ms and shorter than 250 ms and the presence of wheezes is often related with a partial airway obstruction [2][3].

Generally, a stethoscope is used in diagnosing and monitoring a patient and although the stethoscope is reliable and accurate, there are also some disadvantages in using a stethoscope for auscultation [4][5][6][7]. It lacks a method of recording, and offers no quantitative description that can be observed for later assessment [5][7][8]. Therefore, the development of an automated auscultation system is crucial in assisting the physician as well as the patient in both hospital and home care, as it offers massive advantages in terms of acquisition, analysis and storage of the breathing sounds [9][10][11].

In developing the automated system, Artificial Intelligence (AI) is rapidly applied nowadays. The appropriate selection of the classifier to be implemented is one of the most important tasks in any decision-making system [12]. During the last 20 years, Artificial Neural Network (ANN) is one of the many artificial intelligence types that proved to have a great impact on the interpretation of medical data. This is due to the capabilities of ANN in learning a complex dataset and in constructing weight matrices in order to represent the learned patterns, as it is a mathematical model that imitates the function of the human brain neurons [13]. Some recent applications that are related to lung sound classification, such as Amjad et al., applied ANN in their work and obtained 89.28% accuracy in classifying wheeze [8], while ANN has also been extensively used as a classifier in analyzing EEG signals. For example, in EEG signals research ANN is implemented to analyze anesthesia depth monitoring, Parkinson disease and epileptic seizures.

Machine learning algorithm such as ANN are used in various fields but unfortunately, the probability of overfitting of a learning algorithm which increases with the number of features[14]. Therefore, feature selection techniques are powerful tools to avoid overfitting by decreasing the dimensionality of a data [15]. Feature selection methods work by identifying a subset of "meaningful" features from a set of the original features. They can be subdivided into filter, wrapper and also embedded methods [16][17]. Among all this various types of feature selection, we utilized the filter methods as it thus not depend on any specific classification method and thus very suitable to be used in any classification method that favor [16]. One of the common ways and is going to be used in this research is by using analysis of variance (ANOVA).

This paper aimed to show that the classification performance of a high-dimensional features data can be improved by applying feature selection method. The rest of the paper is organized as follows: Section 2 explains the data preparation details used for this project, while in section 3 an overview of the proposed method is introduced. This is followed by the end results and discussion in section 4, and the paper concludes in section 5.

2.0 DATA PREPARATION

The data used in this paper is collected directly from asthmatic patients in Hospital Tuanku Fauziah, Kangar, Malaysia. Ethical approval was provided and obtained by the Medical research and Ethics Committee (MREC) of Malaysia. A total of 15 normal respiratory sound and 15 wheeze data samples were used for this work. All the collected signals were sampled at 10 kHz. Each of the samples was segmented into numb of 500 samples. The sounds were filtered by a 4th order Butterworth bandpass filter with a cutoff frequency from 150 to 2000 Hz.

3.0 OVERVIEW OF PROPOSED METHOD

In this section, explanation on the method used will be discussed thoroughly.

3.1 Statistical Features Extraction

Feature extraction is a crucial method in data classification. The features extracted are a transformation of the signal into features that can be considered as its compressed representation. Due to the musical property of wheeze, researchers preferred to extract features in the frequency domain as, by doing so, the distinct frequency peaks can be clearly observed. As for the features themselves, a statistical based approach will be implemented. 11 statistical features extracted are mean, median, mode, variance, standard deviation, Interquartile Range (IQR), skewness, kurtosis, second moment, percentiles and entropy. These features were chosen based on their function contribution to the signals and they are grouped in order to clearly define this contribution. Figure 1, shows the main categorizations for some of the statistical features used for this paper.

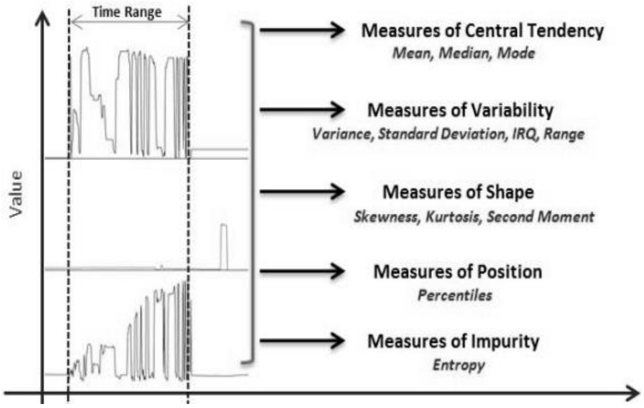


Figure 1 Group statistical features [18].

As shown in Figure 1, the statistical features selected represent the features that will measure the central tendency, shape and also variability of the signal. The measure of shape represented by skewness can be clearly defined as the measure of the asymmetry of the signal distribution while kurtosis as the measure of 'peakedness' of the signal distribution [8]. As variability is defined the measure of the amount of changes in the signal distribution [8].

3.2 Statistical Analysis with ANOVA

ANOVA is a statistical test to confirm whether or not there is a significant difference between the means of several data sets. Moreover, ANOVA examines the variance of data set means and compared to within class variance of the data sets themselves [21]. In this work, ANOVA test was performed using the software "IBM SPSS Statistics 20". The significance threshold was set to $p<0.05$ for all of the case studied.

3.3 Classification with Artificial Neural Network

After the feature selection process, we are now ready to start with the classification process. Artificial neural network (ANN) is a method that tries to mirror the brain functions in a computerized way by restoring the learning mechanism as the basis of human behavior. ANN can learn the relationship between input and output during the data training. Artificial neural network is a nonlinear informational processing device that is built from interconnected neurons. Each input is multiplied by a connection weight. The product and biases are summed and transformed through a transfer function (consisting of algebraic equations) to generate the final output. The process of combining the signals and generating the output of each connection is represented as weight [19]. ANN classification works by the training and the testing process applied to it. The network of the ANN consists of three main layers that are the input, hidden and output layer. In training the ANN network, back propagation (BP) procedures will be used.

The BP algorithm is based on the error correction rule. Error propagates via a forward and backward pass where weight is fixed and adjusted. Finally, a set of outputs is produced as the actual response of the network [1]. Strictly, a BP algorithm involves three stages; Feed forward; Back propagation of error; and Weight update.

There are four reasons for using an ANN as a classifier: 1) ANN has a simple structure for physical implementation, 2) Weights representing the solution are found by iteratively training, 3) ANN can easily map complex class distributions and 4) The generalization property of the ANN produces appropriate results for the input vectors that are not present in the training set [12]. A further advantage of ANN is that it is able to approximate a complex non-linear mapping. It is also very flexible with respect to incomplete, missing and even noisy data. Moreover, ANN can be implemented in parallel with hardware [20].

4.0 RESULTS AND DISCUSSIONS

Statistical validation of normal and wheeze features from the implementation of one-way ANOVA can be observed in Table1. The statistical significance was set to $p<0.05$ for all of the case studied. From Table 1 can clearly observe that not all of the features extracted were significant. Mean, median, mode, standard deviation, interquatile range, and percentiles were significant with $p<0.05$ while variance, skewness, kurtosis, second moment and entropy had no significance with $p>0.05$. Therefore, these feature will undergone feature selection process with them be divided to two group of features with one group of features were consist with only the selected features (mean, median, mode, standard deviation, interquatile range and percentiles) while the other group consist of all the 11 original features extracted. This two group of features will then be fed to ANN separately.

Table 1 Statistical validation of normal and wheeze features.

Features	**p-value**
Mean	**0.036**
Median	**0.010**
Mode	**0.030**
Variance	0.104
Standard deviation	**0.031**
Interquartile Range	**0.027**
Skewness	0.057
Kurtosis	0.089
Second moment	0.104
Percentiles	**0.017**
Entropy	0.082

As for the evaluation of the ANN performance, several parameters needed to be adjusted so as to produce an optimized network. Network architecture of 6 and 11 inputs presenting the extracted features and two outputs representing normal and wheezes data were used. It contains sigmoid hidden neurons and linear output neurons for the classification process. The network was first tested and optimized by using features without undergoing feature selection methods first.

The three types of parameters are the number of neurons in the hidden layer, the learning rate and finally the number of epochs. For each of the various parameters tested, the results were collected and recorded and are well presented in Figures 2, 3 and 4.

From Figure 1, it was found that the best number of neurons is 10 for both of the ANN tested with the accuracy of 93% and 80% for ANN with ANOVA selection features and ANN without ANOVA selection features respectively. It can also be observed that the performance of the ANN seems to be degraded with the increased number of neurons. This is due to the fact that the larger number of neurons in the hidden layer can degrade the performance of the ANN by over training the ANN. Thus the best number of neurons selected is 10 for both of ANN tested.

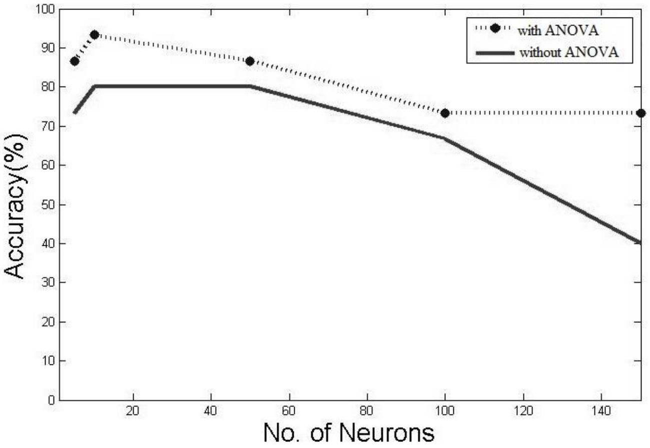


Figure 2 ANN accuracy for a varying number of neurons.

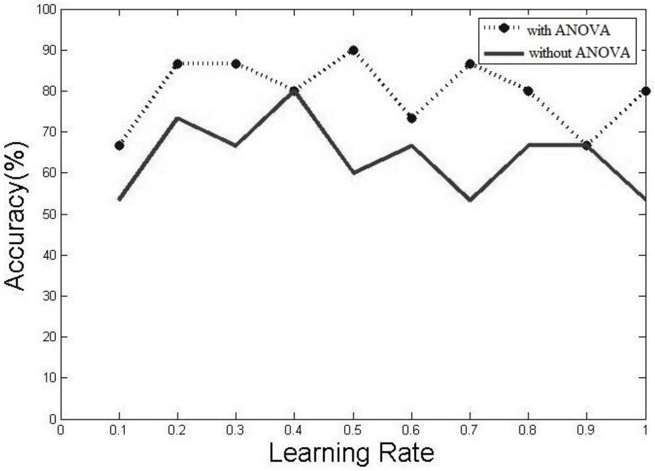


Figure 3 ANN accuracy for a varying learning rate.

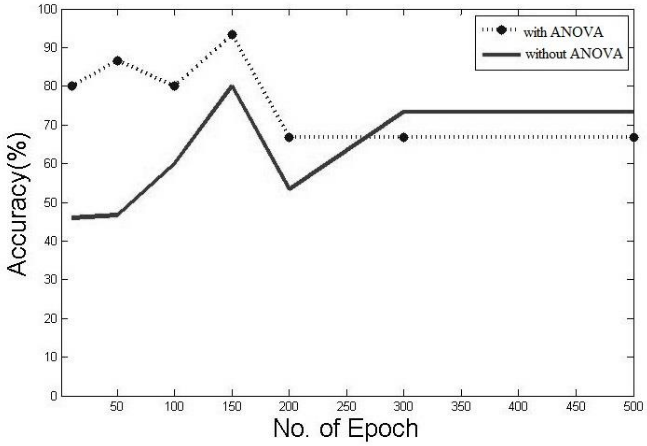


Figure 4 ANN accuracy for a varying number of epochs.

Figure 3 illustrates the results of finding the optimum learning rate for the network built. A fluctuation trend is seen for both the ANN tested. Thus, only the learning rate that produces the highest accuracy was selected. As for the learning rate, the value differs for both ANN tested. The highest accuracy reached for the ANN with ANOVA selection features was 90% with a 0.5 learning rate, while accuracy of only 80% was achieved for the ANN without ANOVA selection features with a learning rate of 0.4.

As for the performance evaluation of the number of epochs in Figure 4 once again the ANN with ANOVA selection features achieved the highest accuracy with 93.33% compared to the ANN without ANOVA selection features, which only achieved 80%. They both, however, achieved the best accuracy at a number of 150 epochs.

To sum up, the best ANN architecture is achieved by 10 hidden nodes, a 0.5 learning rate and number of 150 epochs resulting in 90% accuracy achieved for ANN with ANOVA selection features, while the best ANN architecture of 10 hidden nodes, a 0.4 learning rate and a number of 150 epochs resulted in 80% accuracy for ANN without ANOVA selection features.

The overall results of the classifiers with and without the use of feature selection method are shown in Table 2. Based on the results, the used of feature selection method produces a better result with a higher accuracy compared with the classifier that use data that not undergo the selection process, whilst, ANN with feature that undergo feature selection method produce a better classification accuracy compared to the ANN with feature that did not undergo feature selection method with 93.33% against 80.00% accuracy achieved.

Table 2 Overall performance comparison for ANN with and without feature selection method.

Classifier	Accuracy (%)	
	Without feature selection method	With feature selection method
ANN	80.00	93.33

5.0 CONCLUSIONS

Throughout this paper, the performances between ANN that uses features that undergo and did not undergo feature selection are compared. From the results obtained, it can be concluded that ANN with feature that undergo feature selection method produce a better classification accuracy compared to the ANN with feature that did not undergo feature selection method with 93.33% against 80.00% accuracy achieved. Our aim for future work is to strive to obtain a standard lung sound processing method in order to develop a computerized system to assist doctors and patients.

Acknowledgement

The authors like to gratefully thank the Director General of Health Malaysia for the permission to publish this paper. We would also like to thank the Medical Research and Ethics Committee (MREC) of Malaysia for approving and permitting real data to be collected in Hospital Tuanku Fauziah to be used in this study. (Ref. No.: NMRR-12-1188-14426).

References

[1] Pasterkamp, H., Kraman S. S., Wodicka G. R. 1997. Respiratory Sounds: Advances Beyond The Stethoscope. *American Journal Of Respiratory And Critical Care Medicine*. 156(3): 974-987.

[2] Le C. S., Belghith A., Collet C., Salzenstein F. 2009. Wheezing Sounds Detection Using Multivariate Generalized Gaussian Distributions. *IEEE International Conference on Acoustics, Speech and Signal Processing* . 541-544.

[3] Riella, R. J., Nohama P., Maia J. M. 2009. Method For Automatic Detection Of Wheezing In Lung Sounds. *Brazilian Journal of Medical and Biological Research*. 42(7): 674-684.

[4] Taplidou, S. A., Hadjileontiadis L. J. 2007. Wheeze Detection Based On Time-Frequency Analysis Of Breath Sounds. *Computers In Biology And Medicine*. 37(8): 1073-1083.

[5] Jané, R., Salvatella D., Fiz J. A., Morera J. 1998. Spectral Analysis Of Respiratory Sounds To Assess Bronchodilator Effect In Asthmatic Patients. *Proceedings of the 20th Annual International Conference of the IEEE in Engineering in Medicine and Biology Society*. 3203-3206.

[6] Alsmadi, S. S., Kahya Y. P. 2002. Online Classification Of Lung Sounds Using DSP. *Proceedings of the Second Joint in Engineering in Medicine and Biology 2002*. (2): 1771-1772.

[7] Güler, E. Ç., Sankur B., Kahya Y. P., Raudys S. 2005. Two-Stage Classification Of Respiratory Sound Patterns. *Computers in Biology and Medicine*. 35(1): 67-83.

[8] Hashemi, A., Arabalibiek H., Agin K. 2011. Classification Of Wheeze Sounds Using Wavelets And Neural Networks. *International Conference on Biomedical Engineering and Technology*. (127).

[9] Sello, S., Strambi S. K., De M. G., Ambrosino N. 2008. Respiratory Sound Analysis In Healthy And Pathological Subjects: A Wavelet Approach. *Biomedical Signal Processing and Control*. 3(3): 181-191.

[10] Van Der Heijden M., Lucas P. J., Lijnse B., Heijdra Y. F., Schermer T. R. 2013. An Autonomous Mobile System For The Management Of COPD. *Journal Of Biomedical Informatics.*

[11] Wisniewski M., Zielinski T. 2011. Tonal Index In Digital Recognition Of Lung Auscultation. *Conference Proceedings in Signal Processing Algorithms, Architectures, Arrangements, and Applications 2011*. 1-5.

[12] Dokur Z., Ölmez T. 2003. Classification Of Respiratory Sounds By Using An Artificial Neural Network. *International Journal Of Pattern Recognition And Artificial Intelligence*. 17(04): 567-580.

[13] Nadiatun Z. S., Mashor P. M. D., Nor Hazlyna H., Fatimatul Anis B., Rosline H. 2012. Classification Of Blasts In Acute Leukemia Blood Samples Using K-Nearest Neighbour. *IEEE 8th International Colloquium on Signal Processing and its Application*. 461-465.

[14] Grünauer, A., & Vincze, M. 2015. Using Dimension Reduction to Improve the Classification of High-dimensional Data. arXiv preprint arXiv:1505.06907.

[15] Andrew Y. Ng. 2004. Feature Selection, L1 Vs. L2 Regularization, And Rotational Invariance. *In Proceedings of the Twenty-first International Conference on Machine Learning*, *ICML '04*, pages 78–, New York, NY, USA. ACM.

[16] Dunja Mladenić. 2006. Feature selection for dimensionality reduction. In Subspace, Latent Structure and Feature Selection. 3940 of Lecture Notes in Computer Science. 84–102. Springer Berlin Heidelberg.

[17] Guyon, I. and Elisseeff, A. 2003. An Introduction To Variable And Feature Selection. *Journal of Machine Learning Research: Special Issue on Variable and Feature Selection*. 3: 1157–1182.

[18] Esmael B., Arnaout A., Fruhwirth R. K., Thonhauser G. 2011. Automated System For Drilling Operations Classification Using Statistical Features. *11th International Conference on Hybrid Intelligent Systems*. 196-199.

[19] Mohanraj M., Jayaraj S., Muraleedharan C. 2012. Applications Of Artificial Neural Networks For Refrigeration, Air-Conditioning And Heat Pump Systems—A Review. *Renewable and Sustainable Energy Reviews. 16*(2): 1340-1358.

[20] Ali A. 2012. A Concise Artificial Neural Network in Data Mining. *International Journal of Research in Engineering & Applied Sciences.* 2(2): 418-428.

[21] Van den Broek, E. L., Lisý, V., Janssen, J. H., Westerink, J. H., Schut, M. H., & Tuinenbreijer, K. 2010. Affective Man-Machine Interface: Unveiling Human Emotions Through Biosignals. *Biomedical Engineering Systems and Technologies*. 21-47. Springer Berlin Heidelberg.