Journal of Scientific Research & Reports
3(23): 2960-2971, 2014; Article no. JSRR.2014.23.004
ISSN: 2320–0227

SCIENCEDOMAIN *international*
www.sciencedomain.org

Application of Data Mining Techniques to Audiometric Data among Professionals in India

J. Majumder[1] and L. K. Sharma[1*]

[1]National Institute of Occupational Health (ICMR), Ahmedabad-380016, India.

Authors' contributions

This work was carried out in collaboration between both authors. Author JM designed the study, wrote the protocol, and wrote the first draft of the manuscript. Author LKS managed the literature searches, analyses of the study, and contributed in drafting the manuscript. Both authors read and approved the final manuscript.

Article Information

DOI: 10.9734/JSRR/2014/12700
Editor(s):
(1) William Ebomoyi, Department of Health Studies College of Health Sciences Chicago State University, USA.
Reviewers:
(1) Anonymous, Campus Universitário Marília, Brazil.
(2) Anonymous, University of the Witwatersrand, Johannesburg, South Africa.
(3) Anonymous, Federal University, Nigeria.
Complete Peer review History: http://www.sciencedomain.org/review-history.php?iid=664&id=22&aid=6134

Original Research Article

Received 14th July 2014
Accepted 5th September 2014
Published 16th September 2014

ABSTRACT

Aims: Noise induced hearing loss (NIHL) is among the principal occupational health hazard. To illustrate that, in order to enrich the database on audiometric status and fast dissemination of knowledgebase, data mining techniques are imperative tools.
Study Design: A cross sectional study design was used.
Place and Duration of Study: Pure tone audiometric data of both ears of drivers that have 10 years working experience and office workers from Kolkata City, India were recorded.
Methodology: The data were subjected to both unsupervised and supervised learning techniques, in turn, in order to train the classifier that determines the clusters for newly generated cases. Expectation Maximization (EM), k-means, Linear Vector Quantization

**Corresponding author: E-mail: sharmalk@icmr.org.in, lksharma@daad-alumni.de;*

(LVQ), and Self Organization Map (SOM) unsupervised learning techniques were utilized.
Results: Silhouette Plot (SP) validation showed that 93.3% of the considered cases for the left ear and 85.8% for the right ear were correctly classified. These metadata were further subjected to supervised learning algorithm to achieve a high level correctly classified result, in which, each cluster bears its class label. Naïve Bays Classifier (NBC) recorded, as accurate (98.8%) for both left and right ears. The high accuracy of supervised learning algorithms, cross validated with 10-fold cross validation tends to predict the class of audiometric data whenever a newly generated data are introduced.
Conclusion: This feasibility of using machine learning and data classification models on the audiometric data would be an effective tool in the hearing conservation program for individuals exposed to noisy environments in their respective workplaces.

Keywords:Hearing threshold; cluster analysis; unsupervised learning; supervised learning; cross validation.

1. INTRODUCTION

Researches on mining of medical data have been carried out extensively by various researchers [1,2]. This is because the generation of data is substantially more than the knowledge produced. Data mining, especially clustering techniques allows examining the similarity and dissimilarity among the groups. Clustering techniques include statistical and artificial neural network approaches enable to group a set of data in such a way that they are more similar to each other than those in other groups.

Looking into the perspective that noise induced hearing loss (NIHL) is irreversible [3] and rated among the most prevalent occupational health hazard [4], it is extensively studied using different analytical methods and techniques. For enriched database and faster dissemination of knowledgebase, data mining techniques are now-a-days applied extensively on the audiometric data [5-10]. In recent years, analytical techniques have been successfully applied in conjunction with Artificial Neural Networks (ANN) to better address the cause of NIHL.

ANN particularly Back-propagation neural network (BPNN) has been an effective tool to predict hearing loss in humans [11]. Anwar et al. [9] described the results of statistical and neural data mining of audiology patient records, with the aim of looking for factors influencing which patients would most benefit from being fitted with a hearing aid. The patients were clustered on the basis of similar audiograms using k-means clustering and self organizing map which yielded that automatic textual labeling addresses the heterogeneous character of medical audiology records. In another study, Anwar and Oakes [10] used data mining techniques on audiology patient records for the choice of hearing aid type. They utilized principal component analysis (PCA) which yielded four main audiogram types as per the type of hearing aid chosen. Oakes et al. [6] in a study concluded that audiometric data of individual records are difficult to examine and data mining technique would discover the relationship between the data and processed heterogeneous data, as well as audiology records.

This paper deals with the pure tone audiometric data of both ears collected on office workers, drivers with less than 10 years of experience and drivers with more than 10 years of experience. These data were undergone unsupervised and supervised learning technique

for training the classifier for predicting the cluster of the new generated unseen cases. This would be an effective tool toward hearing conservation programs.

2. MATERIALS AND METHODS

2.1 Audiometric Data

The study was a prospective cross sectional design wherein the subjects were recruited on the basis of self responded questions regarding their work profile and work experience. The pure tone audiometric data of three groups of men viz. office workers engaged in inspection and administrative work (N=30), automobile drivers with less than 10 years driving experience (N=30) and automobile drivers with more than 10 years driving experience (N=30) from Kolkata City, India were taken into consideration [12]. Audiometric testing consisted of air conduction, pure-tone, hearing threshold measurement at frequencies of 0.125, 0.25, 0.5, 1, 1.5, 2, 3, 4, 6, 8 and 10 kHz of both ears by using Arphi Audiometer Model 700 MK IV. The bone conduction audiometric testing was not performed due to feasibility issues.

2.2 Unsupervised Learning Techniques

Unsupervised learning techniques such as Expectation Maximization (EM), k-means, Linear Vector Quantization (LVQ), and Self Organization Map (SOM) were applied. EM is a probabilistic, two-step iterative optimization technique. Step (E) estimates probabilities and step (M) finds an approximation to the mixture model. Advantageously, EM utilizes an automatic cluster labeling algorithm that determines the number of clusters by itself. The k-means algorithm works for compact and hyper-spherical clusters using a known squared error-based clustering algorithm [13]. LVQ and SOM are Artificial Neural Network (ANN) based unsupervised learning techniques [14,15]. LVQ, a known prototype based clustering method, describes clusters, using a centre and some similarities (e.g. in sizes and shape parameters) [16]. LVQ also adapts parameters in order to fit the clusters to a given data set. It forms a quantized approximation of the distribution of an input data set using a finite number of reference vectors which are stored in the connection weights of neural network with two layers and trained through competitive learning [16].

Each unit in the lattice (neuron) and adjacent neurons are interconnected, which gives the clear topology of how the SOM network can be visualized in two-dimensional lattice structure. Input patterns are usually fully connected to all neurons via adaptable weights. Hence, during the training process, neighboring input patterns are projected into the lattice corresponding to adjacent neurons [13]. SOM is also used for supervised learning [17].

2.3 Supervised Learning Techniques

Naïve Bayes (NB) is composed of directed acyclic graphs with only one parent (representing the unobserved node) and several children (corresponding to observed nodes) with a strong assumption of independence among child nodes in the context of their parent [18]. The independence model (Naive Bayes) is based on estimated probabilities, where larger probability indicates the class label value. Instance-based (IB) [19] learning algorithms are lazy-learning algorithms as they delay the induction or the generalization process until the classification is performed. Lazy-learning algorithms require less computation time during the training phase than NB but more computation time during the classification process. Back

Propagation Network (BPN) is a Multi Layer Perceptron Learning (MLP) method capable of classifying non-linear input data, it uses extended gradient-descent based delta learning rule known as back propagation. During classification, the signal at the input units propagates all the way through the net to determine the activation values at all the output units. Each input unit has an activation value that represents some feature external to the net [18]. Radial Basis Function (RBF) network is a three-layer feedback network, where each hidden unit implements a radial activation function and each output unit implements a weighted sum of hidden unit outputs. Further, RBF Network (RBFN) can also be implemented wherein a normalized Gaussian radial basis function network is the basis of process.

3. RESULTS AND DISCUSSION

The audiometric data of volunteers from three different groups of occupation with different exposures are depicted in Figs. 1 and 2. It was observed that the mean hearing threshold levels were the lowest for office workers and highest for automobile drivers with more than 10 years of driving experience. The results indicated that hearing threshold levels increased with increase in driving experience of automobile drivers. Also, it was observed from Figs. 1 and 2 that the mean hearing threshold levels at all tested audiometric frequencies is higher in the left ear as compared to the right ear for all the three groups of volunteers.

Taking into consideration the fact that noise induced hearing loss and presbycusis are additive in the permanent threshold shift of the exposed individual [20], the hearing handicap percentage due to presbycusis was calculated as 2% for office workers, 7.1% for drivers with less than 10 years of experience and 8.7% for drivers with more than 10 years of experience [21,22]. Having said that, the mean age of the volunteers in the three occupational groups was 34.6±3.3, 32.7±1.9 and 36.0±3.2 years respectively at the time of the audiometric test. The data, as recorded did not follow the normal distribution curve. Therefore, Spearman's rho test was performed and correlation between years of exposure and hearing handicap percentage due to presbycusis was found to be highly significant ($\rho=0.620$, $p<0.001$).

3.1 Unsupervised Learning

The EM, LVQ, SOM and k-means clustering algorithms were performed on the left and right ear audiometric data of office workers, drivers with less than 10 years experience and drivers with more than 10 years experience. EM algorithm identified 3 and 4 number of clusters for left and right ear data, respectively. The same number of clusters was considered for the remaining algorithms. Silhouette Plot (SP) was used to validate the result of clustering algorithms. Figs. 3 and 4 depict the SP for left and right ear data, respectively. It depicts that for all the four clustering algorithms applied, at least 93.3% of data for the left ear and 85.8% of data for right ear were correctly classified Table 1. Further the average of silhouette values calculated for left and right ear audiometric data are depicted in Table 1.

The results were cross validated on the three categories of workers and the results for left and right ear are shown in Tables 2 and 3 respectively. It was observed that data from the similar groups of occupation were assimilated in similar clusters leaving behind fractional data that are scattered around different clusters. These metadata were then subjected to supervised learning as inconclusive results from the unsupervised learning algorithm.

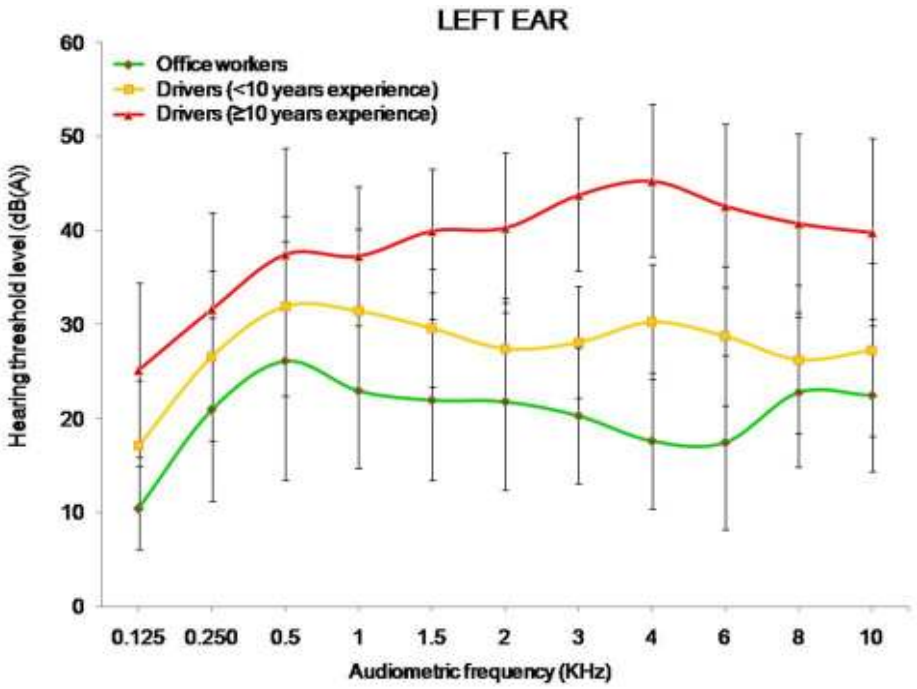

Fig. 1. Trend of audiometric data of left ear among the studied volunteers

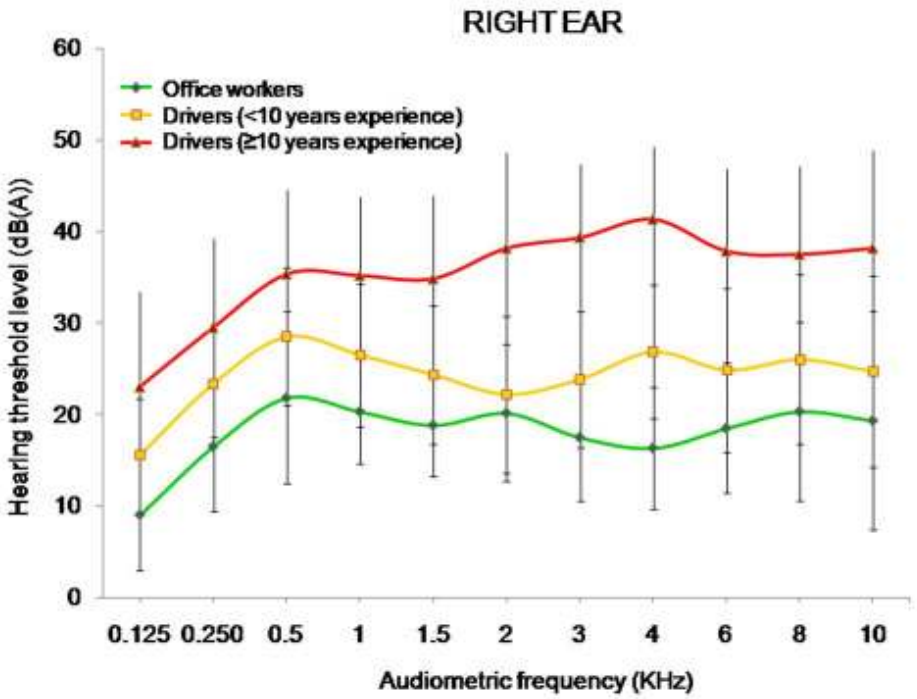

Fig. 2. Trend of audiometric data of right ear among the studied volunteers

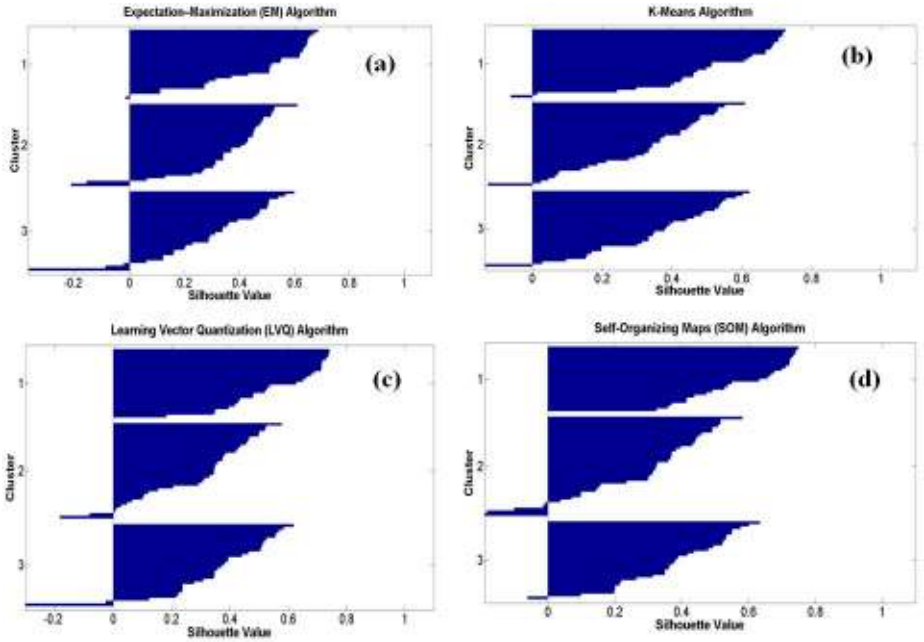

Fig. 3. Silhouette plot for left ear data cluster; a) EM algorithm, b) k-means algorithm, c) LVQ algorithm, d) SOM algorithm

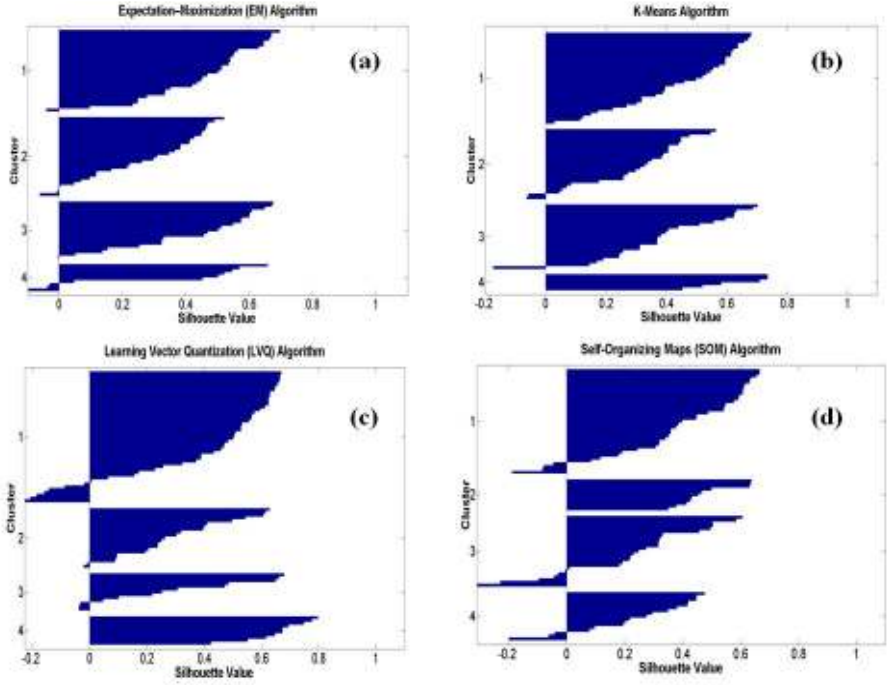

Fig. 4. Silhouette plot for right ear data cluster; a) EM algorithm, b) k-means algorithm, c) LVQ algorithm, d) SOM algorithm

Table 1. Correctly classified audiometric data and mean Silhouette value applying four algorithms

S. no.	Algorithm	% correctly classified data		Mean silhouette value	
		Left ear	Right ear	Left ear	Right ear
1	EM	93.3	92.2	0.37	0.37
2	k-Means	94.4	96.7	0.38	0.39
3	LVQ	95.6	87.8	0.33	0.39
4	SOM	93.3	85.8	0.40	0.39

Table 2. Data distribution of left ear on different unsupervised learning algorithms with respect to occupation

Cluster	Cluster no.	Office workers	Drivers (<10 yrs exp)	Drivers (≥10 yrs exp)	Total
EM	Cluster 1	**24**	3	0	27
	Cluster 2	5	**24**	3	32
	Cluster 3	1	3	**27**	31
k-means	Cluster 1	**20**	7	0	27
	Cluster 2	9	**21**	3	33
	Cluster 3	1	2	**27**	30
LVQ	Cluster 1	**20**	5	0	25
	Cluster 2	9	**22**	4	35
	Cluster 3	1	3	**26**	30
SOM	Cluster 1	**20**	7	0	27
	Cluster 2	9	**21**	3	33
	Cluster 3	1	2	**27**	30

Table 3. Data distribution of right ear on different unsupervised learning algorithms with respect to occupation

Cluster	Cluster no.	Office workers	Drivers (<10 yrs exp)	Drivers (≥10 yrs exp)	Total
EM	Cluster 1	**23**	7	0	30
	Cluster 2	6	**18**	5	29
	Cluster 3	1	4	**16**	21
	Cluster 4	0	1	9	10
k-means	Cluster 1	**25**	9	0	34
	Cluster 2	5	**16**	5	26
	Cluster 3	0	4	**20**	24
	Cluster 4	0	1	5	6
LVQ	Cluster 1	**28**	17	1	46
	Cluster 2	2	**10**	9	21
	Cluster 3	0	1	**12**	13
	Cluster 4	0	2	8	10
SOM	Cluster 1	**25**	**12**	0	37
	Cluster 2	1	**7**	3	11
	Cluster 3	0	3	**22**	25
	Cluster 4	4	**8**	5	17

3.1.1 Data trend

As observed from the unsupervised learning results in Table 1, 95.6% of the left ear audiometric data by LVQ technique and 96.7% of right ear data by k-means technique were correctly clustered. For left ear, the office workers had better threshold of hearing as compared to the drivers with less than as well as more than 10 years of experience. The mean hearing threshold at 4000 Hz was found to be 30 dB(A) for drivers with less than 10 years of experience and 45 dB(A) for drivers with more than 10 years of experience. The trend of audiometric data for the left ear with different occupational groups is shown in Fig. 5. Cluster-wise distribution of the occupational groups for the left ear reveals that similar trend audiometric data has been placed in three different clusters. For the right ear, the mean hearing threshold at 4000 Hz followed the similar trend being 27 dB(A) for drivers with less than 10 years of experience and 41 dB(A) for drivers with more than 10 years of experience Fig. 6.

3.2 Supervised Learning

The supervised learning methods (NB, IB, BPN, RBF, RBFN and SOM) were applied on the data set with correctly classified result of unsupervised algorithm as class label, using Weka 3.7 [23]. The results of LVQ and k-means were used, respectively for left and right ears. Each classifier was trained to build the classifier model. To validate the model and measure the accuracy of the classifier model, two methodologies were applied viz. test data set, and 10-fold cross validation [24,25]. The detailed results of 10-fold cross validation test for left and right ear audiometric data are shown in Table 4. The 10-fold cross validation estimator has a lower variance than a single hold-out set estimator (test data set validation), which is important if the amount of data available is limited. In case of a single hold-out set, 70% of data are used for training and 30% used for testing, the test set is considered as small, and there ought to be variation in the performance estimate for different samples of data, or for different partitions of the data to form training and test sets. However, with 10-fold validation, the variance is reduced by averaging over 10 different partitions to form 10 sub-sets; making the performance estimate less sensitive to the partitioning of data [24,25].

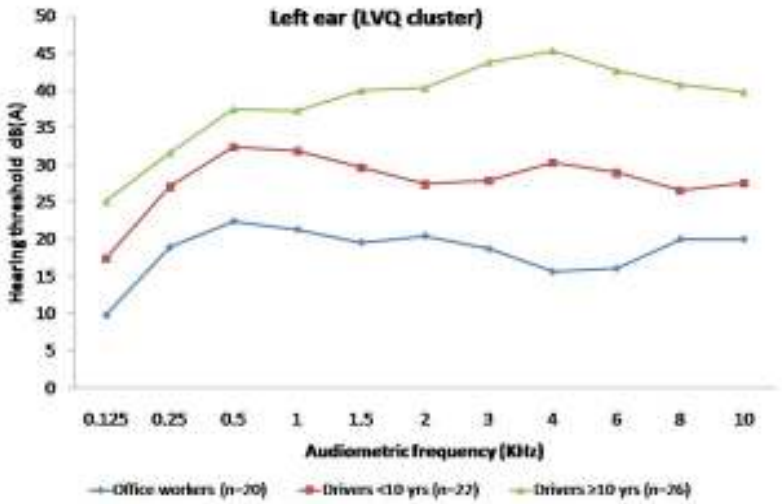

Fig. 5. Trend of left ear audiometric data with different occupational groups after LVQ cluster analysis

Table 4. Supervised learning classifier validation results for left and right ear audiometric data

Statistics	NB		IB		BPN		RBF		RFBN		SOM	
	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right
Correctly classified instances (%)	98.8	98.8	94.4	93.3	94.4	93.3	93.3	82.2	93.3	94.4	95.5	95.6
Kappa statistic	0.98	0.98	0.91	0.90	0.91	0.90	0.89	0.74	0.89	0.92	0.93	0.93
Mean absolute error	0.02	0.20	0.06	0.05	0.04	0.04	0.17	0.19	0.04	0.02	0.23	0.25
Root mean squared error	0.07	0.09	0.16	0.17	0.15	0.15	0.22	0.27	0.21	0.16	0.29	0.31
Relative absolute error (%)	3.6	5.8	13.3	16.4	9.8	12.2	37.4	53.8	10.0	7.7	52.3	70.5
Root relative squared error (%)	15.1	22.2	34.5	41.0	30.8	35.5	45.3	64.2	44.7	39.3	61.5	74.9
Case coverage (0.95 level) (%)	100	100	100	97.77	100	98.88	100	100	93.33	94.44	100	100

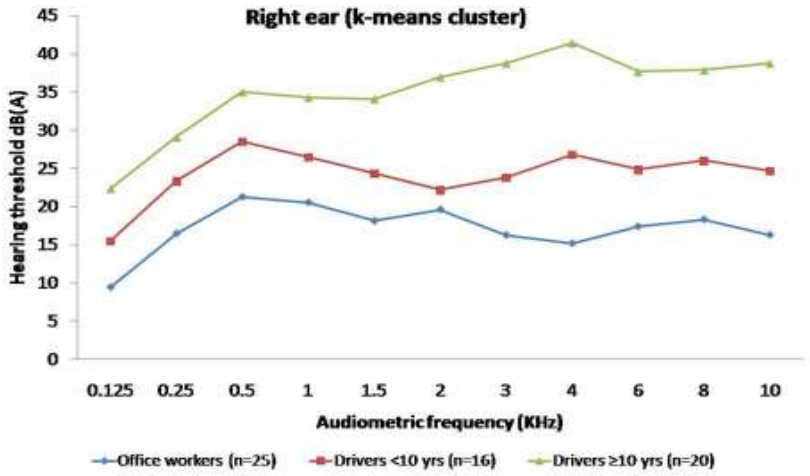

Fig. 6. Trend of right ear audiometric data with different occupational groups after k-means cluster analysis

3.3 Discussion

It is noted that nearly all automobile vehicles are right hand drive in Kolkata city of India, the left ear of the drivers is near to the engine whereas the right ear faces the outside environment. Having said this, the data reported shows that the left ear is more affected at all tested frequencies than the right ear. This is probably due to the additive effect of the engine noise and the environmental sound that are localized around the left ear. Literature also reported that for left hand drive vehicles, drivers infrequently have a greater degree of hearing loss in their left ear. Although, the left ear is more affected in the Indian drivers, a dip in the hearing threshold at 4000 Hz frequency is very prominent in both the ears in all the groups that there is maximal the 4000 Hz sound hair cells normally reside [26].

This paper is an attempt to compare the techniques of unsupervised algorithm (EM, k-means, LVQ and SOM) with those of the supervised learning one as we work on the three categories of audiometric data. SOM happens to be used under both learning algorithms because it studies as well as clusters and classifies unstructured and structured portions of audiology records [6]. The results of four different clustering algorithms in this study are shown in Table 2, which reflect that formation of clusters are similar in k-means, LVQ and SOM for the left ear is in accordance with earlier study [10]. However, the formation of clusters with the right ear audiometric data showed variable distribution.

The supervised algorithm requires data samples with their respective class labels [18], hence the inconclusive metadata and their class labels were supplied to it. Naïve Bayes Classifier recorded as accurately classified (98.8%) for ears data. Except for the RBF classifier's result for the right ear data, Kappa statistics revealed that results from all other classifiers are well above the acceptance region of 0.8 [27]. This result verifies that the classification of the data was good enough, irrespective of the fact that the RBF classifier for right ear allows drawing of tentative conclusions [27]. In a study [7], auditory brainstem classified using Naïve Bayes Classifier (NBC), Support Vector Machine Multilayer Perceptron and KStar algorithms. NBC achieved the accuracy of 83.4% and specificity of 86.3% [7], which is considerably lower than the results obtained by each one of the algorithm

in our study. The high accuracy of supervised learning algorithms, cross validated with 10-fold cross validation as shown Table 4 tends to predict the class of audiometric data when a new data without a class label is applied.

4. CONCLUSION

Our research findings concerning the use of unsupervised and supervised learning of audiometric data can be used to train the classifier for predicting the cluster of the newly generated (or unseen) cases.

Audiometric data can now be easily and appropriately classified. The NBC recorded as more accurate for both left and right ear audiometric data. In terms of the overall classification accuracy, all the classifiers have shown consistent, relatively high performance except the RBF classifier for right ear, allowing the drawing of tentative conclusions. This indicates the feasibility of using machine learning and statistical classification models for the audiometric data. This would be an effective tool toward hearing conservation programs by classifying the audiometric status of individuals exposed to noisy environment, especially the workplace noise.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Cios K. Medical data mining and knowledge discovery. Berlin/Heidelberg: Springer Verlag; 2001.
2. Velemínská J, et al. Dental age estimation and different predictive ability of various tooth types in the Czech population: Data mining methods. Anthropological indicator; Report about the bio-anthropological literature. 2013;70(3):331–345.
3. Sliwinska-Kowalska M, Davis A. Noise-induced hearing loss. Noise Health. 2012;14(61):274-80.
4. Borchgrevink HM. Does health promotion work in relation to noise? Noise Health. 2003;5(18):25-30.
5. Marozas V, Lukoševičius A, Engdahl B, Svensson O, Sörnmo L. Otoacoustic emissions and pass/fail separation using artificial neural network. Ultragarsas. 2000;1(34):7-12.
6. Oakes M, Cox S, Wermter S. Data mining audiology records with the Chi-squared test and self organizing maps. 22nd British National Conference on Databases, D. Nelson et al., editors. University of Sunderland Press. 2005;123-30.
7. McCullagh P, Wang H, Zheng H, Lightbody G, McAllister G. A comparison of supervised classification methods for auditory brainstem response determination. Stud Health Technol Inform. 2007;129(Pt 2):1289-93.
8. Hardalaç F. Classification of educational backgrounds of students using musical intelligence and perception with the help of genetic neural networks. Expert Syst Appl. 2009;36(3):6708-13.
9. Anwar MN, Oakes MP, Wermter S, Heinrich S. Clustering audiology data. Paper presented at the 19th machine learning conference of Belgium and The Netherlands; Belgium; 2010.

10. Anwar M, Oakes MP. Data mining of audiology patient records: Factors influencing the choice of hearing aid type. BMC Med Inform Decis Mak. 2012;12(Suppl 1):6.
11. Rehman MZ, Nazri MN, Ghazali MI. Noise-Induced Hearing Loss (NIHL) prediction in humans using a modified back propagation neural network. Paper presented at the International Conference on Advanced Science, Engineering and Information Technology; Bangi. 2011;185-189.
12. Majumder J. An investigation into the auditory threshold profile of automobile drivers and office workers [dissertation]. Presidency College: University of Calcutta; 2002.
13. Xu R, Wunsch D. Survey of clustering algorithm. IEEE T Neural Networ. 2005;16(3):645-78.
14. Kohonen T. Self organization of very large documents: State of the art. Paper presented at the 8th International Conference on Artificial Neural Networks; London: Springer. 1998;1:6574.
15. Cox S, Oakes M, Wermter S, Hawthorne M. Audio mine: Medical data mining in heterogeneous audiology records. Int J Comput Int. 2004;1(1):112.
16. Borgelt C, Girimonte D, Acciani G. Learning vector quantization: Cluster size and cluster number. International Symposium on Circuits and Systems. 2004;5:808-811.
17. Salah M, Trinder J, Shaker A. Evaluation of the self organizing map classifier for building detection from Lidar data and multispectral Arial images. Spatial Science. 2009;54(2):15-34.
18. Kotsiantis SB. Supervised machine learning: A Review of Classification Techniques. Informatica. 2007;31:249-68.
19. Aha D, Kibler D. Instance-based learning algorithms. Mach Learn. 1999;6:37-66.
20. Macrae JH. Presbycusis and noise-induced permanent threshold shift. J Acoust Soc Am. 1991;90(5):2513-6.
21. Robinson DW, Sutton GJ. Age effects and hearing. A comparative analysis of published threshold data. Audiology. 1979;18:320-34.
22. NIOSH. Criteria for a recommended standard: Occupational noise exposure, revised Criteria 1998. U.S. Department of Health, Education, and Welfare, Public Health service, centres for DIsease control and prevention. National Institute of Occupational Safety and Health. DHHS (NIOSH). 1988;98-126.
23. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: An update. SIGKDD Explorations. 2009;11(1):10-8.
24. Elkan C. Evaluating classifiers. Department of Computer Science and Engineering, University of California, San Diego; 2012.
Available: http://cseweb.ucsd.edu/users/elkan/250B winter2012/classifiereval.pdf.
25. Song Q, Wang G, Wang C. Automatic recommendation of classification algorithms based on data set characteristics. Pattern Recogn. 2012;45(7):2672-89.
26. Rutka J. Discussion Paper on Hearing Loss. Veterans review and appeal board, Canada; 2011. Available: http://www.vrab-tacra.gc.ca/Publications/Discussion-Paper-on-Hearing-Loss.pdf.
27. Carletta J. Assessing agreement on classification tasks: The kappa statistic. Comput Linguist. 1996;22(2):249-54.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sciencedomain.org/review-history.php?iid=664&id=22&aid=6134