Journal of
Information Technology & Software Engineering

Research Article Open Access

Application of Data Mining Techniques in the Analysis of Acoustic Sound Characteristics

Mojtaba Talafi Daryani[1], Hossein Khabiri[1] and Zahra Yamini[2]*,

[1]*MA student of Business Management, Faculty of Management, University of Tehran, Iran*
[2]*MA graduated of Information Technology, Faculty of Management, University of Alzahra*

Abstract

In recent years, the analysis of acoustic characteristics of speech and sound has been one of the areas that data mining has found its way through. The present research study is also related to this topic which aims to detect the gender of the speaker by using the acoustic feature of his voice. In this research, the data set includes 3,168 recorded voice samples gathered from female and male speakers. Through acoustic analysis, 20 characteristics along with the desired labels were extracted and prepared for the data mining process. Finally, using Python programming language tools, 6 different techniques were used to construct an appropriate problem-solving model. These techniques were: support vector machines, logistic regression, random forest, regression and classification trees, adaptive boosting, and K-nearest neighbor. The accuracy of the models was also compared with each other. The obtained results revealed that the accuracy of all these techniques was sufficiently high (above 90%) for solving the problem and the model made by them had the necessary efficiency for classification. Moreover, the obtained model was specifically evaluated through decision tree and some principles and rules existing in the model were extracted. As a result, it was revealed that average fundamental frequency measured across the audio signal is the key characteristic of the sound for the evaluation of the voice gender to the extent that it plays a key role in data classification.

Keywords: Data mining; Sound mining; Speech analysis; Acoustic analysis; Classification algorithms

Introduction

In recent years, we have witnessed the use of data mining techniques for discovering hidden patterns in the massive (dense) dataset and thus solving problems in various fields of sciences. Among these issues we can refer to: diagnosis of various diseases in the field of medical sciences, customer relationship management in the field of marketing and business management, forecasting the price of a stock or index in the field of financial and economic sciences, and forecasting or analyzing the results of elections in the field of social and political sciences.

Over time, we have gradually observed that data mining techniques have made their way through more unknown fields of science, such that, relying on them, we can remove and solve problems that we couldn't even imagine to solve them before. For instance, the use of data mining techniques for voice and speech analysis is one of these cases, and the present research has also been done in this regard.

Speech is the most common way of communication among humans. The issue of communication between humans and their surrounding environment through the sound and the dominance of human over machines through this intermediary has always been a fascinating subject for researchers. This branch of research, once referred to as a dream, has now become a reality that is getting increasingly wider and its hidden angles become more and more apparent.

This scope and breadth has developed to the extent that in the general issue of data mining we have witnessed the emergence of such branches as speech data mining, voice data mining, conversation data mining, and audio mining. In general, these are all called speech data mining methods [1].

In recent years, we have also witnessed a number of studies that have proved the application of this topic in various fields of science. The study conducted by Hammerling et al. [2] that addresses the application of sound data mining fort the assessment of laryngeal pathology is an example of this case.

The aim of the present research study is study one of the problems (issues) of speech analysis based on the recognition of the speaker's gender with regard to the sound characteristics extracted from his speech. To achieve this aim, various data mining techniques such as decision trees, support vector machines, logistic regression and other methods have been used [3].

At the literature section, the theoretical concepts of the techniques have been described, and in the materials and methods section, the details of the adopted techniques have been mentioned. It is hoped that this research will reveal the hidden functions of speech analysis more than ever and will attract the attention of different researchers to this field of research as it has been to some extent hidden from their attention.

Research Literature

Data mining

Data mining, also known as knowledge discovery in the database, is an analytical process used in various disciplines to examine the meaningful relationships between variables in large datasets. The analysis of massive data flows leads to the discovery of valuable knowledge and theoretical concepts that help organizations to improve their operations and make quick and intelligent decisions [4].

Data mining techniques and classification

The most commonly used technique in data mining is classification. Classification algorithms allow the user to classify a dense dataset by a model and in the form of predefined classes. Some of these algorithmic

***Corresponding author:** Zahra Y, MA of Information Technology Management, University of Alzahra, Tehran, Iran, Tel: 98 21 8804 4040; E-mail: Yaminiparvaneh@yahoo.com

Received April 28, 2018; **Accepted** May 23, 2018; **Published** May 30, 2018

Citation: Daryani MT, Khabiri H, Yamini Z (2018) Application of Data Mining Techniques in the Analysis of Acoustic Sound Characteristics. J Inform Tech Softw Eng 8: 238. doi:10.4172/2165-7866.1000238

models are decision trees, random forest, neural networks, Bayesian classification and support vector machines, K nearest neighbor, adaptive boosting, and classification based on association rules [5].

Clustering: Clustering is another data mining technique that involves identifying clusters and grouping similar objects in each cluster. If it is stated that the classification techniques reclassified as the supervised learning methods then it should be admitted that clustering techniques are categorized as unsupervised learning methods. Although in this section, researchers mainly focus on partitioned algorithms such as K-means, but clustering also involves other methods such as: Hierarchical clustering algorithms like BIRCH, CURE, Grid-based clustering algorithms such as STING, Wave Cluster, model-based clustering algorithms like COBWEB, and density-based clustering algorithms such as DBSCAN [5] (Table 1).

Regression: Regression is a technique that is used for predictive modeling. The purpose of regression analysis is determining the best model that determines how a variable is associated with one or more other variables. Since in the real world, the forecast requires the integration of various and complex aspects of the data set, to complement it, a combination of different models is used. Among these combinatorial algorithms, we can refer to classification and regression trees.

The various regression methods used are logistic regression, Linear regression, Multivariate linear regression, Nonlinear regression, and Multivariate nonlinear regression [5].

Data mining steps: It should be noted that the implementation of data mining techniques is just one of the steps of the series of stages involved in the knowledge discovery process in the database. In addition, there are steps that it seems important to pay attention to them. Figure 1 shows the stages of knowledge discovery in the database. These stages are as follows:

a) ***Data selection*****:** This step involves studying the scope of application and selection of the datasets. The aim of studying the scope of application is to determine the project aims through understanding a business problem. At this stage, it is necessary to recognize and detect the minimum size, the required characteristics and appropriate time interval for the dataset [6].

b) ***Data preparation*****:** This step involves operations such as clearing the data by deleting useless data, making decision about the missing data, and..... Moreover, in this stage, it is possible to take into consideration issues related to database management such as data type, missing values pattern, etc. [6].

c) ***Data conversion*****:** This step involves processing the data to convert it into a format suitable for applying data mining algorithms. Ordinary processes that can be mentioned at this stage are: feature selection, data normalization, data aggregation and data discretization. To normalize the data, the mean value is subtracted from each value and then the result is divided by the standard deviation. Some algorithms are compatible either with quantitative data or qualitative data. So we sometimes need to change the data type [6].

d) ***Data mining*****:** This step involves discovering patterns in the dataset prepared in the previous steps. In this step, various algorithms are evaluated to determine the best way to achieve a particular purpose [6].

e) ***Interpretation and evaluation of the results*****:** This step involves interpreting the discovered patterns and evaluating their application and significance with respect to the scope of application. At this stage, for example, one can conclude that some of the selected characteristics can be ignored because they do not have any influence on the results and applied analysis. Therefore, it is possible to repeat the process after modifying the dataset [6].

Research Background

Very few studies have been done in the field of speech analysis and voice recognition, using a data mining technique. It may be admitted that the lack of sufficient data sets in this domain on the one hand and lack of the familiarity of the researchers with the topic of speech analysis and data mining functions on the other hand has led to insufficient research studies in this field. One of the studies that partly relates to this study is the research carried out by Buyukyilmaz and Cibikdiken [7].

This research has been done on the dataset of the present study. Using the Python programming language tool, this study has attempted to complete and provide multi-layer perceptron neural networks algorithm. This method is a predictive artificial neural network model that corresponds the input dataset to appropriate output. Also, this method consists of several layers of nodes in a directed graph and

S.no	Method	Description
1	Support vector machines	In this method, classification is done through the construction of an extraordinary scheme in a multi- dimensional space which separates the samples with different labels and classes. This technique supports the operations and tasks related to regression and classification. It also has the ability to evaluate the continuous and stratified variables. This method is based on the concept of decision making scheme which includes decision borders [3].
2	Logistic regression	In this method, for making linear models appropriate, a model that provides an exact and accurate estimate of the class should be adapted to the data. Logistic regression uses linear models for classification and linear regression for estimation of the target numerical value [10].
3	Random forest	Random forest is one of the decision tree subset methods that generates a large number of random trees on the sub-sets of data sets for problem solving and by averaging, improves the accuracy of the results [9].
4	Regression and Classification trees	This method is one of the decision tree algorithms, which uses classification trees for classification of dependent variables and adopts regression trees for prediction of the response variables [5].
5	Adaptive Boosting	Boosting is a method which is based on achieving a very precise and accurate rule for prediction through the combination of a large number of weak and inaccurate rules and principles. The adaptive boosting is the first boosting applied algorithm and one of the most commonly used ones on a variety of issues [11].
6	K-nearest neighbor	This method is based on learning through training samples. Each sample represents a point in n-dimensional space. All training samples are stored in a n-dimensional spatial pattern. When an unknown sample is given, the k-nearest neighbor classifier would search the spatial pattern, looking for the k training sample that is the closest one to the unknown sample. Proximity is defined based on Euclidean distance34.After finding this k data which is similar to the training sample, the label of the unknown sample is selected based on the majority vote. Moreover, through assigning weight to the attributes it would be possible to change the degree of their involvement in the calculation of similarity [14].

Table 1: Theoretical concepts of the used data mining techniques.

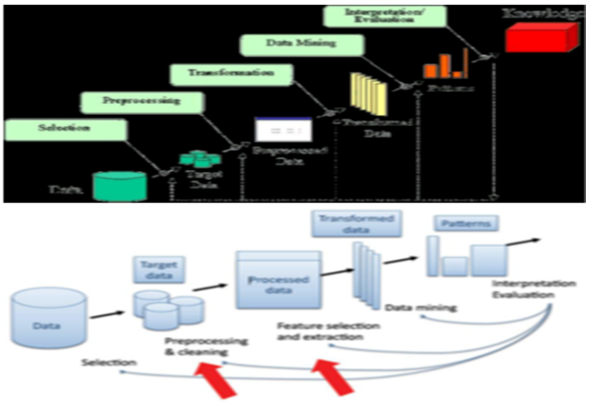

Figure 1: Knowledge discovery steps in the database.

each layer is completely connected to the next layer. The researchers eventually managed to classify the intended datasets with a precision of 96.74%.

Materials and Methods

Data selection

In this study, a dataset entitled Gender Identification via Voice was selected from the Kegel site data set. This dataset has been prepared with the intention to identify male or female voices based on their voice features. The dataset contains 3,168 recorded audio samples from male and female speakers. Audio samples are prepared through the audio analysis process (acoustic analysis) and finally, 20 attributes are extracted with the desired labels. These audio features listed in the dataset file can be seen in Table 2.

Data mining instruments

In the present study, all the processes, methods and techniques used to solve the intended problem have been performed, using the Python programming language and its functional libraries in the Jupyter notebook environment.

Preparation and conversion of the data

The data preprocessing stage is usually neglected, while it is an important stage for the implementation of data mining techniques. At this stage, you can use any method that prepares the raw data for subsequent processing and reorients it for easy and effective use [8,9]. Since real-world data may not have the quality required to start data mining, the implementation of the data preparation and conversion steps is necessary.

Because the present dataset is free of duplicate data or missing values, there is no need to perform operations such as duplicate data deletion, decision making on the missing data, and so on. One of the major goals we wanted to achieve at this stage was to increase our understanding of the datasets and available features, because such understanding plays a significant role in improving the quality of the knowledge discovery process. In this regard, we have been trying to increase our knowledge and information about data, using various methods.

As some examples that we used them for this purpose, we can refer to the study of data type, data format, the study of the statistical factors of the data of each characteristic (number, mean, standard deviation, maximum and minimum values, mean, first and third quartiles), the study of correlation between the characteristics and the data visualization methods (scatter plots, box plots, histograms, etc.). Also, at this stage, data normalization and character selection processes were performed for the use of data mining techniques and improving the quality of the results.

S.no	Description	Characteristic
1	Mean Frequency (in KHZ)	Meanfreq
2	Standard deviation of frequency	Sd
3	Median frequency (in KHZ)	Median
4	First quantile (in KHZ)	Q25
5	Third quantile(in KHZ)	Q75
6	Interquartile range(in KHZ)	IQR
7	Skewness	Skew
8	Kurtosis	Kurt
9	Spectral entropy	Sp.ent
10	Spectral flatness	Sfm
11	Mode frequency	Mode
12	Frequency centroid	Centroid
13	Average of fundamental frequency measured across acoustic signal	Meanfun
14	Minimum of fundamental frequency measured across acoustic signal	Minfun
15	Maximum of fundamental frequency measured across acoustic signal	Maxfun
16	Average of dominant frequency measured across acoustic signal	Meandom
17	Minimum of dominant frequency measured across acoustic signal	Mindom
18	Maximum of dominant frequency measured across acoustic signal	Maxdom
19	Range of dominant frequency measured across acoustic signal	Dfrange
20	Modulation index	Modinex
21	Male or female	Label

Table 2: Audio features listed in the dataset file.

In order to perform the process of selecting an effective characteristic, we also examined the value and importance of each attribute and, as a result, by selecting the valuable attributes, we created a new characteristic category for the implementation of data mining techniques in the next step. In Figure 2, the importance of each attribute can be observed. It should also be noted that for the implementation of data mining techniques in the next step, a portion of the data (20%) was isolated and was not included in the design of the model so that it could be used in the model testing phase.

Data mining

After the dataset was prepared for the implementation of data mining techniques and knowledge discovery process, and after the necessary understanding of the existing features and relationships between them was obtained, different data mining techniques were used to compare their results for solving the desired problem in this study. In the present study, the techniques of support vector machines, logistic regression, random forest, classification and regression trees, adaptive boosting and K nearest neighbor were used. The obtained results and comparisons between these techniques, presented in the finding section, are significant. Also, in implementing each of these techniques, different tricks such as cross-validation and grid search have been used to achieve better results. In a cross-validation method, the dataset is divided into several sections (usually 5 or 10 parts), then the construction and testing stages of the model are repeated based on the number of sections.

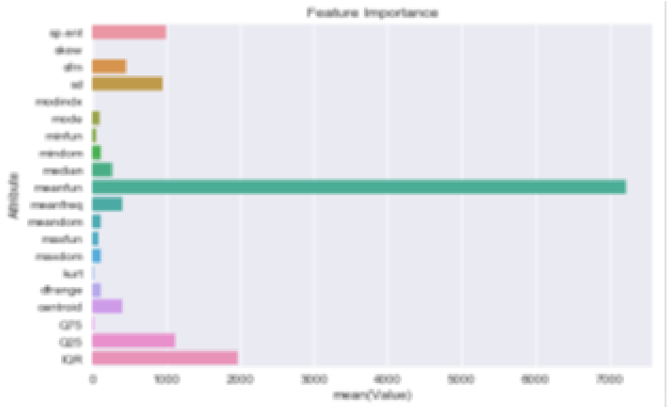

Figure 2: The value and importance of characteristics.

Each time, a part is selected as the testing data, and the other parts are combined as training data for constructing the model. Finally, to measure the validity of the model, the mean accuracy of the repeated steps and its variance are taken into consideration [10-12]. The grid search is also a trick by which we can, in some techniques (methods whose parameters are adjustable), obtain the optimal parameter in order to achieve the best result. For example, we can point out that using the grid search method, one can obtain the best possible depth for a decision tree in order to obtain the desired accuracy. Table 3 illustrates the data mining techniques employed, as well as the complementary techniques used to make each technique appropriate and improve the accuracy of its results.

Results

After various data mining techniques were used to solve the desired problem, the accuracy of each method as well as the underlying parts of the receiver operating characteristic curves were calculated and collected for reviewing the results and comparing them with each other. The receiver operational characteristic curve is a measure of efficiency in classification issues. The more the level below these curves, the more efficient the final performance of the model will be. Table 4 compares the accuracy of the methods used.

As it can be observed in Table 4, all data mining techniques used for solving the intended problem in the present research study have the required accuracy for the construction of the classification model. Since the degree of accuracy of each method and their underlying level of receiver operating characteristic curve are very close to each other (the accuracy higher than 90%), the comparison of techniques in this research study does not lead to a unique, particular or significant result. Moreover, one of the findings that is of significant importance is paying attention to the obtained decision and the rules extracted from it that would be of considerable value, mainly as an obtained model for solving the intended problem that is of classification type problems. Table 3 shows the obtained decision tree. It should also be noted that this decision tree is categorized as a regression and classification type tree that entropy method has been used to draw it (Figure 3).

As it can be observed in Table 3, the decision tree involves valuable If……..Then……..principles that can be used for the classification of the data and detection of the voice gender in the application area. To mention a few examples, the following principles are presented:

If meanfun ≤ 0.1418 and IQR ≥ 0.0825, then with the probability of 0.985, the voice belongs to a man

s.no	Complementary methods	Data mining techniques
1	Selection of Kernel function, cross Validation, Grid Search	Support vector machines
2	Character Selection, cross Validation, Grid Search	Logistic regression
3	Character Selection, cross Validation, Grid Search	Random forest
4	Character Selection, cross Validation, Grid Search	Regression and classification trees
5	Character Selection, cross Validation, Grid Search	Adaptive boosting
6	Character Selection, cross Validation	K-nearest neighbor

Table 3: Techniques and methods of data mining.

If meanfun ≤ 0.1418 and IQR ≤ 0.0626, then with the probability of 0. 996, the voice belongs to a woman.

Discussion and Conclusion

The aim of the present research study was to show the function and application of data mining in acoustic analysis of the voice. In this regard, we succeeded in obtaining significant results. As it was explained, various data mining techniques can be used for the detection of voice gender such that the models resulting from these techniques have the required accuracy for the classification and labeling of the data. As an example, we studied the decision tree obtained through data modeling and observed that for the detection of gender based on voice, the average of measured fundamental frequency across acoustic signal, compared with other characteristics, is of more importance such that in the decision tree, this characteristic is considered as the main node for data classification. The results obtained from this study are comparable with other research studies conducted in this area [13]. Acknowledges the fact that the fundamental frequency is the key guide for the detection of the voice of males and females. The point that fundamental frequency plays a significant role in data classification is also observable in Table 4 that is a box plot (Figure 4).

One of the points that should be taken into consideration is that, based on the obtained results from the conducted research studies, the fundamental frequency of the voice of men and women are considerably different. That is why in this issue, the fundamental frequency plays an important role in classifying the data. The fundamental frequency range for men is between 80-200 Hz and for women is between 150 to 300 Hz [14].

With this point in mind, if we once again look at the decision tree whose simplified form has been presented in Figure 5, we find that the

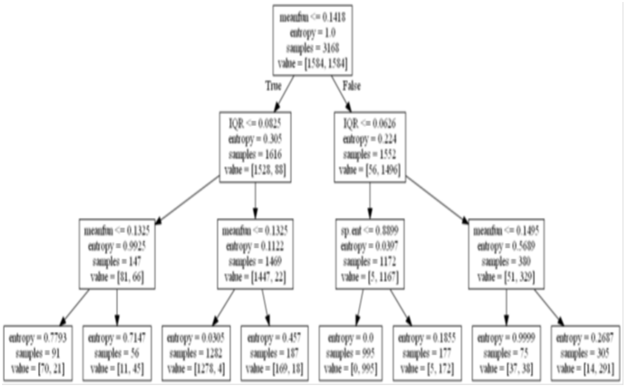

Figure 3: Regression and classification decision tree.

	The level under ROC curve	Accuracy	Data mining techniques
1	0.977970	0.977918	Data mining techniques
2	0.977970	0.971577	Support Vector Machines (linear kernel function)
3	0.976362	0.971552	support Vector Machines (Linear Kernel Function) + cross Validation
4	0.976362	0.975138	support Vector Machines (RBF Kernel Function) + cross Validation
5	0.977970	0.981452	support Vector Machines (Linear Kernel Function) + cross Validation+ grid search
6	0.969991	0.970032	Losistic regression
7	0.969991	0.968377	Logistic regression+cross validation
8	0.963680	0.963722	Logistic regression +attribute selection
9	0.963680	0.968402	Logistic Regression + cross Validation + Attribute Selection
10	0.971539	0.972770	Logistic Regression + cross Validation + Grid Search
11	0.963680	0.973560	Logistic regression + cross validation + Grid search + attribute selection
12	0.971599	0.971609	Random forest
13	0.971599	0.958906	Random forest + cross validation
14	0.973206	0.973186	Random forest + attribute selection
15	0.973206	0.973076	Random forest + cross validation+ attribute selection
16	0.973147	0.981452	Random forest + cross validation+grid search
17	0.974993	0.979874	Random forest + cross validation+grid search**+** attribute selection
18	0.971778	0.971609	Regression and classification trees
19	0.971778	0.957463	Regression and classification trees **+** cross validation
20	0.968383	0.954264	Regression and classification trees **+** cross validation **+** attribute selection
21	0.960464	0.970008	Regression and classification trees **+** cross validation **+** grid search
22	0.962191	0.974349	Regression and classification trees **+** cross validation **+** grid search **+** attribute selection
23	0.974814	0.974763	Adaptive boosting
24	0.974814	0.968500	Adaptive boosting **+** cross validation
25	0.969991	0.971625	Adaptive boosting **+** cross validation **+** attribute selection
26	0.976362	0.980268	Adaptive boosting **+** cross validation **+** grid search
27	0.968443	0.972386	Adaptive boosting **+** cross validation **+** grid search **+** attribute selection
28	0.965168	0.965300	K-nearest neighbor
29	0.965168	0.946303	K-nearest neighbor **+** cross validation
30	0.981066	0.981073	K-nearest neighbor **+** attribute selection
31	0.981066	0.966790	K-nearest neighbor **+** cross validation **+** attribute selection

Table 4: Accuracy and the level under ROC curve of the applied data mining techniques.

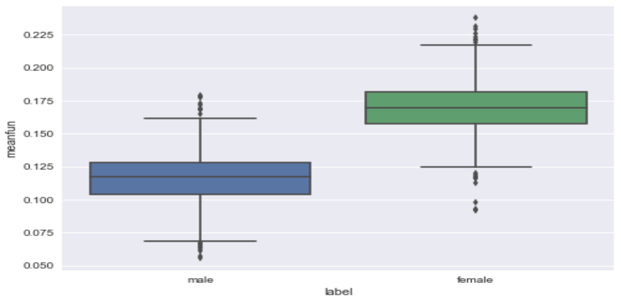

Figure 4: The box plot of the fundamental frequency.

fundamental frequency for data categorization involves the condition of being larger or smaller than 14.1 kHz. Given the intervals mentioned, this value of 140 Hz can be relied upon for detection of the sound of men and women.

Finally, it must be acknowledged that the recognition of the gender of a speaker by using sound characteristics is only one of the simple applications in the field of acoustic science, which has been implemented by data mining techniques.

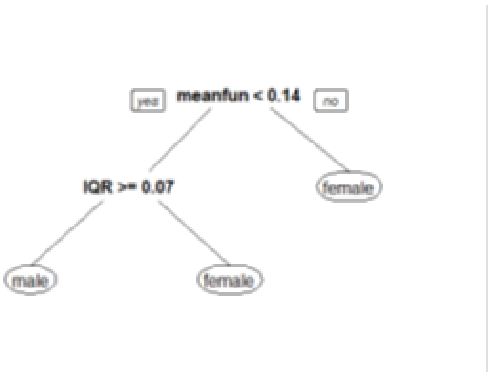

Figure 5: The obtained simplified decision tree.

Certainly, the sound acoustic analysis will have more applications and usages in various fields of science, particularly in interdisciplinary sciences, some of which can be observed nowadays. As an instance, we can discuss the analysis of speaker's emotions or identification of laryngeal diseases, using audio signals. It is suggested that researchers focus their attention on these topics more than ever to make an important contribution to exploring acoustic data.

References

1. Senthildevi KA, Chandra E (2012) Data mining techniques and applications in speech processing-A Review. IJARS 1: 1-8.
2. Hemmerling D, Skalski A, Gajda J (2016) Voice data mining for laryngeal pathology assessment. Computers in Biology and Medicine 69: 270-276.
3. Fatima, Ikbal Khan J (2016) Classification of data mining techniques & tools: A survey. IJIRAS 3: 396-399.
4. Jha A, Dave M, Madan S (2016) A reviews on the study and analysis of big data using data mining techniques. IJLTET 6: 94-102.
5. Maksood FZ, Achuthan G (2016) Analysis of data mining techniques and its applications. IJCA 140: 6-14.
6. Sharma S, Mittal H (2016) Data mining unblocking the intelligence in data. JNCET 6: 22-28.
7. Buyukyilmaz M, Cibikdiken AO (2016) Voice gender recognition using deep learning. Advances in Computer Science Research 58: 409-411.
8. Bharat V, Shelale B, Khandelwal K, Navsare S (2016) A review paper on data mining techniques. International Journal of Engineering Science and Computing 6: 6268-6271.
9. Biau G, Scornet E (2016) A random forest guided tour. Test 25: 197-227.
10. Provost F, Fawcett T (2013) Data science for business: what you need to know about data mining and data-analytic thinking. O'Reilly Media.
11. Schapire RE (2013) Explaining adaboost. Empirical Inference, pp: 37-52.
12. Poon MSF, Ng ML (2015) The role of fundamental frequency and formants in voice gender identification. Speech, Language and Hearing 18: 161-165.
13. Ashby M, Maidment J (2005) Introducing phonetic science. Cambridge University Press, Cambridge.
14. Gorade SM, Deo A, Purohit P (2017) A study of some data mining classification techniques. IRJET 4: 3112-3125.