VOICE DISORDER IDENTIFICATION BY USING MACHINE LEARNING TECHNIQUES

Seminar Report

Submitted in partial fulfillment of the requirements for the award of degree of

BACHELOR OF TECHNOLOGY

In

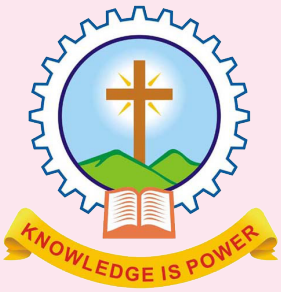
COMPUTER SCIENCE AND ENGINEERING

of

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

Submitted By

SAHEER S



Department of Computer Science & Engineering
Mar Athanasius College Of Engineering Kothamangalam

VOICE DISORDER IDENTIFICATION BY USING MACHINE LEARNING TECHNIQUES

Seminar Report

Submitted in partial fulfillment of the requirements for the award of degree of

BACHELOR OF TECHNOLOGY

In

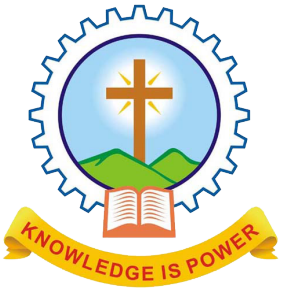
COMPUTER SCIENCE AND ENGINEERING

of

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

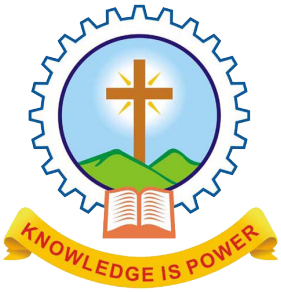
Submitted By

SAHEER S



Department of Computer Science & Engineering
Mar Athanasius College Of Engineering Kothamangalam

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
MAR ATHANASIUS COLLEGE OF ENGINEERING
KOTHAMANGALAM



CERTIFICATE

This is to certify that the report entitled **Voice Disorder identification by using Machine Learning Techniques** *submitted by* **Mr. SAHEER S**, *Reg.No.***MAC15CS052** *towards partial fulfillment of the requirement for the award of Degree of Bachelor of Technology in Computer science and Engineering from APJ Abdul Kalam Technological University for December 2018 is a bonafide record of the seminar carried out by him under our supervision and guidance.*

.....	.....	.....
Prof. Joby George	**Prof. Neethu Subash**	**Dr. Surekha Mariam Varghese**
Faculty Guide	***Faculty Guide***	***Head of the Department***

Date:

Dept. Seal

ACKNOWLEDGEMENT

First and foremost, I sincerely thank the God Almighty for his grace for the successful and timely completion of the seminar.

I express my sincere gratitude and thanks to Dr. Solly George, Principal and Dr. Surekha Mariam Varghese, Head Of the Department for providing the necessary facilities and their encouragement and support.

I owe special thanks to the staff-in-charge Prof. Joby george, Prof. Neethu Subash and Prof. Joby Anu Mathew for their corrections, suggestions and sincere efforts to co-ordinate the seminar under a tight schedule.

I express my sincere thanks to staff members in the Department of Computer Science and Engineering who have taken sincere efforts in helping me to conduct this seminar.

Finally, I would like to acknowledge the heartfelt efforts, comments, criticisms, co-operation and tremendous support given to me by my dear friends during the preparation of the seminar and also during the presentation without whose support this work would have been all the more difficult to accomplish.

ABSTRACT

The use of mobile devices in the healthcare sector is increasing significantly. Mobile health systems can contribute to make patient care faster, better, cheaper.Dysphonia, an alteration of the voice quality that affects about one person in three at least once in his/her lifetime. Voice disorders are rapidly spreading, although they are often underestimated. Mobile health systems can be an easy and fast support to voice pathology detection. The key contribution of this paper is to investigate and compare the performance of several machine learning techniques useful for voice pathology detection. All analyses are performed on a dataset of voices selected from the Saarbruccken voice database. The results obtained are evaluated in terms of accuracy, sensitivity, and receiver operating characteristic area. They show that the best accuracy in voice diseases detection is achieved by the support vector machine or the decision tree.

Contents

List of Figures

List of Abbreviation

MFCC	Mel-Frequency Cepstral Coefficients
SVD	Saarbruecken Voice Database
SVM	Support Vector Machine
DT	Decision Tree
BC	Bayesian Classification
LMT	Logistic Model Tree
ROC	Receiver Operating Characteristics
HNR	Harmonic to Noise Ratio
PCA	Principal Component Analysis
SMO	Sequential Minimal Optimization

Introduction

The introduction of mobile devices for data transmission or disease control and monitoring has been a main attraction of research and business communities. They offer, in fact, numerous opportunities to realise efficient mobile health (mhealth) systems. These solutions can allow patients and doctors to access medical records, clinical audio-visual notes and drug information anywhere and at any time from their mobile devices, such as a tablet or smartphone, to monitor several conditions[1]. M-health solutions can also be used in other important applications such as the detection and prevention of specific diseases, decision making and the management of chronic conditions and emergencies, improving the quality of patient care and reducing the costs of healthcare. Several pathological conditions can be detected and monitored, such as the well known and widespread cardiovascular diseases. In recent years, probably also due to the diffusion of the Internet of Things (IoT) and cloud technologies, there has been a development of monitoring systems in an unobtrusive, portable and easy way using wearable sensors and wireless communications.These systems are able to achieve health data monitoring and analysis, helpful for patients suffering from cardiovascular diseases or for their physical therapy.

If, on the one hand, health monitoring systems for cardiovascular diseases are so celebrated, on the other hand, there are other little known and often underestimated disorders, such as dysphonia, that could benefit from m-health solutions. Dysphonia is a disorder that occurs when the voice quality, pitch and loudness are altered. About 10suffer from this disorder, caused mainly by unhealthy social habits and voice abuse. Unfortunately, a large number of individuals with voice disorders do not seek treatment.Therefore, m-health systems could be an efficient support for the diagnosis and screening of voice disorders.

Clinical voice pathology detection is performed through the execution of several procedures, such as the acoustic analysis. It consists of an estimation of appropriate parameters extracted from voice signal to evaluate any possible alterations of the vocal tract, according to the guidelines of the SIFEL protocol (Societa Italiana di Foniatria e Logopedia), developed by

the ‘ Italian Society of Logopedics and Phoniatrics, following the instructions of the Committee for Phoniatrics of the European Society of Laryngology. It is a non-invasive examination in clinical practice, complementary to other medical tests, such as the laryngoscopic examination based on the direct observation of the vocal folds.

Several acoustic parameters are estimated to evaluate the state of health of the voice. Unfortunately, the accuracy of these parameters in the detection of voice disorders is, often, related to the algorithms used to estimate them. For this reason the main effort of researchers is oriented to the study of acoustic parameters and the application of classification techniques able to obtain a high discrimination accuracy.Recently, speech pathology has focused interest on machine learning techniques.

In this work, we want to discuss the application of machine learning algorithms and features selection methods capable of discriminating between pathological and healthy voices with a better accuracy. In detail, we evaluate the pathology recognition using the information data of patients, such as age and gender, and different features extracted from the voice signals. The adopted parameters are those estimated in the clinical acoustic analysis, such as the Fundamental Frequency (F0), jitter, shimmer and Harmonic to Noise Ratio (HNR). In addition, other parameters, the Mel-Frequency Cepstral Coefficients (MFCC), the first and second derivatives, are used due to their wide application both in machine learning techniques and in voice disorders classification as reported in several studies . The performances are evaluated in terms of accuracy, sensitivity, specificity and ROC area for each considered machine learning methods.

Existing methods

Speech or, in general, the voice signal is used in several kinds of application ranging from emotion recognition to patient healthcare state recognition.

Several m-health solutions adopt these signals to estimate the state of voice health, as well as systems that use voice signals to evaluate emotional condition. Voice pathology detection has, often, been achieved through specific machine learning techniques, and over recent years, several approaches have been developed to improve the performance in terms of accuracy in the discrimination between healthy and pathological voices.

These studies are focused on the identification of parameters to measure the voice quality and new techniques able to detect voice disorders.

Among several machine learning techniques existing in literature, Support Vector Machine (SVM) has been widely used in voice signal processing. Godino et al, for example,focused on the classification of pathological and healthy voices based on MFCC to train and test an SVM classifier. These have obtained a good accuracy (95numerosity of the used dataset composed of only 173 pathological and 53 healthy voices selected by the Massachusetts Eye and Ear Infirmary voice and speech lab (MEEI) database should be obsverved. Additionally, important information, as for example the pathologies of the selected voices, is not available in this work.

The SVM technique was also used in to estimate the presence of dysphonia, investigating four types of pathology: chronic laryngitis, cysts, Reinkes edema and spasmodic dysphonia. The authors proposed an algorithm based on the use of MFCC and Linear Discriminant Analysis (LDA) as a dimensionality reduction method. This algorithm identifies the presence of a pathology with a discrete accuracy (86However, it was tested on a very limited dataset. In fact, only 70 pathological and 40 healthy voices were selected by the Saarbruechen Voice Database (SVD).

The MFCC parameters were considered in other numerous studies, such as subjects with nodules, edema and unilateral vocal fold paralysis were analysed with not so encouraging

results (77.90suffering from spasmodic dysphonia were selected. Unfortunately, the performance of the algorithms, Gaussian Mixture Models (GMM) and SVM, was tested on a limited datasets composed of voices extracted from the MEEI database.

El Emary et al.instead, classified the speech signal by estimating not only the MFCC but also jitter and shimmer. The detection of voice suffering from neurological disorders was performed using the GMM algorithm on a very small subset of the SVD database containing only 38 pathological and 63 healthy voices. an algorithm based on Daubechies discrete wavelet transform, linear prediction coefficients and last squares support vector machine (LS-SVM) was used to identify laryngeal pathologies. The experiments were carried out using a private database.

Another private dataset collected in the Busan National University Hospital was used in the study . Wang et al. classified pathological voices using Hidden Markov Models (HMM), GMM and SVM. The voice disorders considered in this study included vocal polyps, vocal cord palsy, nodules, cysts, edema, laryngitis and glottic cancer. In several studies existing in literature private databases were used . In the first case the data adopted to test the developed system were captured at the Christie and Withington Hospitals in Manchester. Only 77 abnormal speech signals were used to train and test the proposed artificial neural network, and the authors did not specify the pathologies considered. The voices of the subjects were collected at the Phoniatric Department of the University Hospital of Sofia to detect people suffering from laryngeal pathology via the K nearest neighbours algorithm and linear discriminant analysis.

Most of the features and algorithms were trained and tested using limited databases, including a few types of disorders and a few voices. In many cases, the databases used are not available and their results cannot be compared. Moreover, the authors often use only the MFCC as the signal features, not considering the characteristic parameters indicated by clinical protocols, such as the SIFEL protocol, to evaluate the voice quality and the possible presence of disorders.

Proposed method

In this study we analysed the accuracy in the discrimination of pathological from healthy voices of the main machine learning techniques to identify the most reliable one. The idea has been to integrate the best one in a valid m-health system, where the voice signal can be acquired by a mobile device, such as a smartphone or tablet, processed in real-time to extract the voice features, and analyzed by using the machine learning classifier to detect the presence or not of a voice disorders, as shown in Fig.3.1.

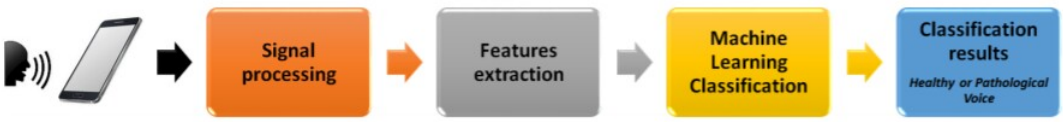


Fig. 3.1: The flowchart of a possible m-Health system for the voice health state classification

In detail, we have evaluated the performance of SVM, the principal adopted technique in literature in relation to the Kernel function, and of some other machine learning algorithms used to identify the presence of voice disorders.

The analysis has been performed using the WEKA tool, one of the most commonly used tools for data mining tasks, selected for the data analysis due to its efficiency, versatility and affordability.

In the following subsections we introduce the dataset used in this study, the features extracted from the voice signal and used for the classification, and the machine learning techniques compared.

3.1 The database

In our research, we have selected a subset of voice samples from the Saarbruecken Voice Database (SVD)[2]. SVD database is a collection of 2041 voice recordings, containing voices from healthy and pathological individuals, published online by the Institute of Phonetics of the University of Saarland. All the recordings are sampled at 50 kHz and their resolution is 16-bit. In total, there are 1354 pathological voices (627 male and 727 female), suffering from 71 different diseases, distinguished between functional and organic disorders. The remaining 687 healthy voices are 259 male and 428 female.

This collection consists of recordings of vowels /a/,/i/,/u/ and an appropriate sentence. To evaluate the patients voice quality the use of vowels is preferable because they avoid linguistic artifacts and are used in many voice assessment applications. In relation to voice disorder detection and identification problems, in clinical practice, the vocalization of the vowel /a/ is used.

In our experimental tests, to perform the experiments on a well-balanced database containing both pathological and healthy voices, we have selected a total of 1370 /a/ vocalizations. In detail, we have chosen:

- 685 pathological voices (257 male and 428 female)
- 685 healthy voices (257 male and 428 female).

More details of the selected recordings are indicated in Table I, in which we have reported the number (No) of considered voices for each age and gender, and the percentage calculated in relation to the whole dataset. The lower number of male samples than female ones is related to the higher incidence of voice disorders in female subjects than in males. We have used all the available healthy and pathological voices from the SVD.

DETAILS OF THE VOICE SIGNALS USED IN THIS STUDY.

Category	Gender	Age Group	No.	%
Healthy	Female	17-29	359	26.20 %
		30-39	27	1.98 %
		40-49	15	1.09 %
		50-59	13	0.95 %
		60+	14	1.02 %
Healthy	Male	17-29	138	10.07 %
		30-39	62	4.53 %
		40-49	37	2.70 %
		50-59	9	0.66 %
		60+	11	0.80 %
Pathological	Female	17-29	58	4.24 %
		30-39	94	6.86 %
		40-49	85	6.20 %
		50-59	87	6.35 %
		60+	104	7.59 %
Pathological	Male	17-29	23	1.68 %
		30-39	24	1.75 %
		40-49	43	3.14 %
		50-59	74	5.40 %
		60+	93	6.79 %
Total	*Female*	*17-60+*	*856*	*62.48%*
	Male	*17-60+*	*514*	*37.52%*

Fig. 3.2: Table 1

It detailly explained in the Fig.3.2. All samples contain the recording of the vowel /a/, the signal required by thehttps://www.overleaf.com/project/5bface3136079412dbb05992 SIFEL protocol to evaluate the voices state of health. To test the capability of the considered algorithms, we have selected the pathological voices from all types of pathology existing in the database. There are organic voice disorders, such as chronic laryngitis or Reinkes edema, and functional dysphonia as a hyperfunctional or hypofunctional one.

3.2 Features used for the classification

Feature extraction is an important task that allows an improvement of the analysis and classification. The choice of which features of the speech signal to use in our study was made by taking into account two considerations[3]. On the one hand, we have used the main parameters adopted by the specialist during the clinical evaluation; on the other, we have chosen the main features used in several correlated studies existing in literature concerning the use of machine learning techniques for the voice classification. In detail, the parameters used in clinical practice are:

- Fundamental Frequency (F0): this represents the rate of vibration of the vocal folds constituting an important index of laryngeal function. It is at the basis of the other parameters calculated in the acoustic analysis and most noise estimation methods
- Jitter: this describes the instabilities of the oscillating pattern of the vocal folds, quantifying the cycle-to-cycle changes in fundamental frequency.
- Shimmer: this indicates the instabilities of the oscillating pattern of the vocal folds, quantifying the cycle-to-cycle changes in amplitude.
- Harmonic to Noise Ratio (HNR): this quantifies the ratio of signal information over noise due to turbulent airflow, resulting from an incomplete vocal fold closure in speech pathologies. The parameters used in other correlated studies are:
- Mel-Frequency Cepstral Coefficients (MFCC): these coefficients try to analyse the vocal tract independently of the vocal folds that can be damaged due to voice pathologies. In this work, the experiments were conducted using 13 MFCC coefficients.
- First and second derivatives of cepstral coefficient: these are useful to investigate the properties of the dynamic behaviour of the speech signal.

It is important to note that, for some of the above mentioned features, there are no standard algorithms available for their calculation. This is a critical issue, because the more

accurate is the computation of each parameter, the more reliable is the voice analysis, namely the classification of the voice signal as healthy or pathological. For example, for the evaluation of the F0, there are several different methods proposed in literature, like Spectral Analysis , the Hilbert-Huang transform, the Robust Algorithm for Pitch Tracking (RAPT) , the Dynamic Programming Projected Phase-Slope Algorithm (DYPSA) , the Speech Transformation and Representation based on Adaptive Interpolation of weiGHTed spectrogram (STRAIGHT) method and the extraction based on the Autocorrelation Function of the speech signal.

However, in our study, we have used a new proposed methodology, that is an optimization and personalization of the Yin algorithm.

Concerning the jitter and shimmer features, In detail, the jitter was expressed as a percentage and it was calculated as the average absolute difference between consecutive periods, divided by the average period.

The shimmer, instead, was estimated as the average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20. It was expressed in decibels (dB).

The HNR was computed in dB as the mean difference between the harmonic peaks and the aperiodic components according to de Kroms algorithm.

Finally, the MFCC coefficients resulting from by the cepstral representation of the voice signal, were calculated by evaluating the Discrete Cosine Transform and the log compression of the voice samples in the frequency domain. These coefficients and their derivatives were extracted using the melcepst Matlab function of the VOICEBOX tool, a speech processing toolbox realised by the Department of Electrical and Electronic Engineering of Imperial College of London and used by several studies existing in literature.

In summary, each instance i of the database used in this study is constituted by the following information:

- Subject ID: a number value to identify the subject;
- age: measured by years from birth;
- gender: female or male;

- features: F0, jitter, shimmer, HNR, MFCC (from 1 to 13), first derivative and second derivative, calculated over the recording of the vowel /a/;
- class: pathological or healthy;

3.3 Machine learning classifiers

In order to make an exhaustive comparison, we have chosen different machine learning algorithms. Actually, each of them has been chosen as a representative of a class of algorithms based on similar characteristics. These techniques are:

- Support Vector Machine

  this is a discriminative classifier formally defined by a separating hyperplane that divides data belonging to different classes. The aim is to identify the class of belonging of the different data. Training a support vector machine requires the solution of a very large quadratic programming optimization problem. To resolve this problem the sequential minimum optimization (SMO) technique is used[4], which is able to divide the optimization problem into a series of smaller possible problems. The classification accuracy can be improve by selecting opportune form and parameters characteristic of the kernel function K(x,y). The most popular kernel function forms are polynomial and radial basis ones.

  In machine learning, support vector machines (SVMs, also support vector networks are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples

as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

More formally, a support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

Training a support vector machine requires the solution of a very large quadratic programming optimization problem. To resolve this problem the sequential minimum optimization (SMO) technique is used, which is able to divide the optimization problem into a series of smaller possible problems. Sequential minimal optimization (SMO) is an algorithm for solving the quadratic programming (QP) problem that arises during the training of support vector machines. SMO is widely used for training support vector machines and is implemented by the popular LIBSVM tool. The publication of the SMO algorithm in 1998 has generated a lot of excitement in the SVM community, as previously available methods for SVM training were much more complex and required expensive third-party QP solvers.

SMO is an iterative algorithm for solving the optimization problem described above. SMO breaks this problem into a series of smallest possible sub-problems, which are then solved analytically. Because of the linear equality constraint involving the Lagrange multipliers , the smallest possible problem involves two such multipliers.and this reduced problem can be solved analytically: one needs to find a minimum of a one-dimensional quadratic function. k is the negative of the sum over the rest of terms in the equality

constraint, which is fixed in each iteration.

The classification accuracy can be improve by selecting opportune form and parameters characteristic of the kernel function K(x,y). The most popular kernel function forms are polynomial and radial basis one.

In machine learning, kernel methods are a class of algorithms for pattern analysis, whose best known member is the support vector machine (SVM). The general task of pattern analysis is to find and study general types of relations (for example clusters, rankings, principal components, correlations, classifications) in datasets. In its simplest form, the kernel trick means transforming data into another dimension that has a clear dividing margin between classes of data. For many algorithms that solve these tasks, the data in raw representation have to be explicitly transformed into feature vector representations via a user-specified feature map: in contrast, kernel methods require only a user-specified kernel, i.e., a similarity function over pairs of data points in raw representation.

Kernel methods owe their name to the use of kernel functions, which enable them to operate in a high-dimensional, implicit feature space without ever computing the coordinates of the data in that space, but rather by simply computing the inner products between the images of all pairs of data in the feature space. This operation is often computationally cheaper than the explicit computation of the coordinates. This approach is called the ”kernel trick”. Kernel functions have been introduced for sequence data, graphs, text, images, as well as vectors.

Algorithms capable of operating with kernels include the kernel perceptron, support vector machines (SVM), Gaussian processes, principal components analysis (PCA), canonical correlation analysis, ridge regression, spectral clustering, linear adaptive filters and many others. Any linear model can be turned into a non-linear model by applying the kernel trick to the model: replacing its features (predictors) by a kernel function.

Most kernel algorithms are based on convex optimization or eigenproblems and are statistically well-founded. Typically, their statistical properties are analyzed using statistical learning theory (for example, using Rademacher complexity).

The polynomial kernel is a kernel function commonly used with support vector machines (SVMs) and other kernelized models, that represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models.

Intuitively, the polynomial kernel looks not only at the given features of input samples to determine their similarity, but also combinations of these. In the context of regression analysis, such combinations are known as interaction features. The (implicit) feature space of a polynomial kernel is equivalent to that of polynomial regression, but without the combinatorial blowup in the number of parameters to be learned. When the input features are binary-valued (booleans), then the features correspond to logical conjunctions of input features.

- Decision Tree (DT): this technique is used to classify categorical data in which the learned function is represented by a decision tree[5]. Decision trees are easy to interpret, capable of working with missing values and categorical and continuous data, characteristics of the medical field. We have used J48, an implementation of algorithm C4.5 , the most popular tree classifier

  Decision tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. An example is shown in the diagram at right. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

  A decision tree is a simple representation for classifying examples. For this section, assume that all of the input features have finite discrete domains, and there is a single target feature called the ”classification”. Each element of the domain of the classification is called a class. A decision tree or a classification tree is a tree in which each internal (non-leaf) node is labeled with an input feature. The arcs coming from a node labeled with an input feature are labeled with each of the possible values of the target or output

feature or the arc leads to a subordinate decision node on a different input feature. Each leaf of the tree is labeled with a class or a probability distribution over the classes.

A tree can be ”learned” by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node has all the same value of the target variable, or when splitting no longer adds value to the predictions. This process of top-down induction of decision trees (TDIDT) is an example of a greedy algorithm, and it is by far the most common strategy for learning decision trees from data.

In data mining, decision trees can be described also as the combination of mathematical and computational techniques to aid the description, categorization and generalization of a given set of data.

Decision trees used in data mining are of two main types:

- Classification tree analysis is when the predicted outcome is the class to which the data belongs.
- Regression tree analysis is when the predicted outcome can be considered a real number (e.g. the price of a house, or a patient’s length of stay in a hospital).

The term Classification And Regression Tree (CART) analysis is an umbrella term used to refer to both of the above procedures, first introduced by Breiman et al. in 1984. Trees used for regression and trees used for classification have some similarities - but also some differences, such as the procedure used to determine where to split.

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan’s earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. Authors of the Weka machine learning software described the C4.5 algorithm as ”a landmark decision tree program that is probably the machine learning workhorse most widely used in practice to date”.

In pseudocode, the general algorithm for building decision trees is;

- Check for the above base cases.
- For each attribute a, find the normalized information gain ratio from splitting on a.
- Let abest be the attribute with the highest normalized information gain.
- Create a decision node that splits on abest.
- Recur on the sublists obtained by splitting on abest, and add those nodes as children of node.

- Bayesian Classification (BC): this approach named after Thomas Bayes, who proposed the Bayes Theorem[6]. The classification is achieved by evaluating the probabilistic model that represents a set of random variables and their conditional dependencies identified, respectively, as nodes and strings . The major advantage is the easy interpretation of the results and the robustness in dealing with missing data.

  naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

  Naive Bayes has been studied extensively since the 1950s. It was introduced under a different name into the text retrieval community in the early 1960s and remains a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. With appropriate pre-processing, it is competitive in this domain with more advanced methods including support vector machines. It also finds application in automatic medical diagnosis.

  Naive Bayes classifiers[7] are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

In the statistics and computer science literature, naive Bayes models are known under a variety of names, including simple Bayes and independence Bayes. All these names reference the use of Bayes' theorem in the classifier's decision rule, but naive Bayes is not (necessarily) a Bayesian method.

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness, and diameter features.

For some types of probability models, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods.

Despite their naive design and apparently oversimplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. In 2004, an analysis of the Bayesian classification problem showed that there are sound theoretical reasons for the apparently implausible efficacy of naive Bayes classifiers.Still, a comprehensive comparison with other classification algorithms in 2006 showed that Bayes classification is outperformed by other approaches, such as boosted trees or random forests.

An advantage of naive Bayes is that it only requires a small number of training data to estimate the parameters necessary for classification.[citation needed]

Despite the fact that the far-reaching independence assumptions are often inaccurate, the

naive Bayes classifier has several properties that make it surprisingly useful in practice. In particular, the decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one-dimensional distribution. This helps alleviate problems stemming from the curse of dimensionality, such as the need for data sets that scale exponentially with the number of features. While naive Bayes often fails to produce a good estimate for the correct class probabilities, this may not be a requirement for many applications. For example, the naive Bayes classifier will make the correct MAP decision rule classification so long as the correct class is more probable than any other class. This is true regardless of whether the probability estimate is slightly, or even grossly inaccurate. In this manner, the overall classifier can be robust enough to ignore serious deficiencies in its underlying naive probability model.

- Logistic Model Tree (LMT): this technique combines logistic regression models with tree induction. It consists of a standard decision tree structure with logistic regression functions at the leaves. SimpleLogistic class

  a logistic model tree (LMT) is a classification model with an associated supervised training algorithm that combines logistic regression (LR) and decision tree learning[8].

  Logistic model trees are based on the earlier idea of a model tree: a decision tree that has linear regression models at its leaves to provide a piecewise linear regression model (where ordinary decision trees with constants at their leaves would produce a piecewise constant model). In the logistic variant, the LogitBoost algorithm is used to produce an LR model at every node in the tree; the node is then split using the C4.5 criterion. Each LogitBoost invocation is warm-started[vague] from its results in the parent node. Finally, the tree is pruned.

  The basic LMT induction algorithm uses cross-validation to find a number of LogitBoost iterations that does not overfit the training data. A faster version has been proposed that uses the Akaike information criterion to control LogitBoost stopping.

  In statistics, the logistic model (or logit model) is a widely used statistical model that, in its basic form, uses a logistic function to model a binary dependent variable; many more

complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model; it is a form of binomial regression. Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail, win/lose, alive/dead or healthy/sick; these are represented by an indicator variable, where the two values are labeled ”0” and ”1”. In the logistic model, the log-odds (the logarithm of the odds) for the value labeled ”1” is a linear combination of one or more independent variables (”predictors”); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled ”1” can vary between 0 (certainly the value ”0”) and 1 (certainly the value ”1”), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from logistic unit, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probit model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a constant rate, with each dependent variable having its own parameter; for a binary independent variable this generalizes the odds ratio.

Logistic regression was developed by statistician David Cox in 1958. The binary logistic regression model has extensions to more than two levels of the dependent variable: categorical outputs with more than two values are modelled by multinomial logistic regression, and if the multiple categories are ordered, by ordinal logistic regression, for example the proportional odds ordinal logistic model. The model itself simply models probability of output in terms of input, and does not perform statistical classification (it is not a classifier), though it can be used to make a classifier, for instance by choosing a cutoff value and classifying inputs with probability greater than the cutoff as one class, below the cutoff as the other; this is a common way to make a binary classifier. The coefficients are generally not computed by a closed-form expression, unlike linear least squares.

- Istance-based Learning algorithms: these algorithms use specific instances to achieve the classification predictions. The algorithms used are k-nearest neighbour one (k-NN) , where the classification is based on k nearest neighbours of a new instance (Ibk in WEKA) and K* , an instance-based classifier that uses an entropybased distance function to classify data (kStar in WEKA)[8]

  Instance-based learning (sometimes called memory-based learning) is a family of learning algorithms that, instead of performing explicit generalization, compares new problem instances with instances seen in training, which have been stored in memory.

  It is called instance-based because it constructs hypotheses directly from the training instances themselves. This means that the hypothesis complexity can grow with the data: in the worst case, a hypothesis is a list of n training items and the computational complexity of classifying a single new instance is O(n). One advantage that instance-based learning has over other methods of machine learning is its ability to adapt its model to previously unseen data. Instance-based learners may simply store a new instance or throw an old instance away.

  Examples of instance-based learning algorithm are the k-nearest neighbor algorithm, kernel machines and RBF networks. These store (a subset of) their training set; when predicting a value/class for a new instance, they compute distances or similarities between this instance and the training instances to make a decision.

  To battle the memory complexity of storing all training instances, as well as the risk of overfitting to noise in the training set, instance reduction algorithms have been proposed.

  In pattern recognition, the k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

  In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the

object is simply assigned to the class of that single nearest neighbor.

In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors.

k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms.

Both for classification and regression, a useful technique can be used to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor.

The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

A peculiarity of the k-NN algorithm is that it is sensitive to the local structure of the data.

It is important to remark here that other classification techniques are not reported in this study due to the poor performance achieved during our experiments.

3.4 Performance evaluation

Cross-validation was used in our experiments, to overcome the problem of overfitting and to make the predictions more general.In detail we have made reference to a 10-fold cross-validation, dividing the training set into k=10 smaller sets. For each of the k folds, a model is trained using k-1 of the folds as the training data, while the resulting model is validated on the remaining part of the data. The performance of the selected machine learning classification techniques was evaluated in terms of accuracy, sensibility, specificity and ROC area by using the following measurements:

- True Positive (TP): the voice sample is pathological and the algorithm recognizes this;
- True Negative (TN): the voice sample is healthy and the algorithm recognizes this;

- False Positive (FP): the voice sample is healthy but the algorithm recognizes it as pathological;
- False Negative (FN): the voice sample is pathological but the algorithm recognizes it as healthy

The accuracy, that is the percentage of correctly classified instances, is defined as:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad \text{(Equ:3.1)}$$

while the sensitivity and the specificity, that represent respectively the tests ability to detect positive results or the identification of negative results, are defined as:

$$Specificity = TN/(TN + FP) \quad \text{(Equ:3.2)}$$

$$Sensitivity = TP/(TP + FN) \quad \text{(Equ:3.3)}$$

The ROC area is a measure of the goodness of a classification algorithm evaluated by plotting a curve representing the sensitivity versus the complementarity of specificity (1- specificity) and measuring the area under this curve (AUC). The AUC can be interpreted as the average value of sensitivity for all the possible values of specificity. The maximum (AUC=1) means that the algorithm is perfect in the classification between diseased and non-diseased voices. On the other hand, AUC=0 means that the algorithm incorrectly classifies all subjects with diseases as negative and all healthy subjects as pathological.

3.4.1 Features selection

Attribute selection is an important task that allows the improvement of the dataset analysis to identify the redundant and/or irrelevant features to optimize memory space and time machine computing speed. For these reasons, in our study, we have chosen to test the machine learning classification techniques over the overall database, and, additionally, over three different subset of the database chosen by selecting some of the calculated features applying the following features selection methods:

- InfoGainAttributeEval algorithm : this calculates the information gain for each feature. The results can vary from 0 (no information) to 1 (maximum information). The information gain obtained for each considered feature is shown in Fig.3.3. The best value is achieved by the age, followed by two MFCC coefficients[9], the HNR, jitter, the second derivative, and others. For our experimental tests, we excluded all features that did not produce an information gain, this is equal to 0, while those features with an information gain greater than 0 were considered.

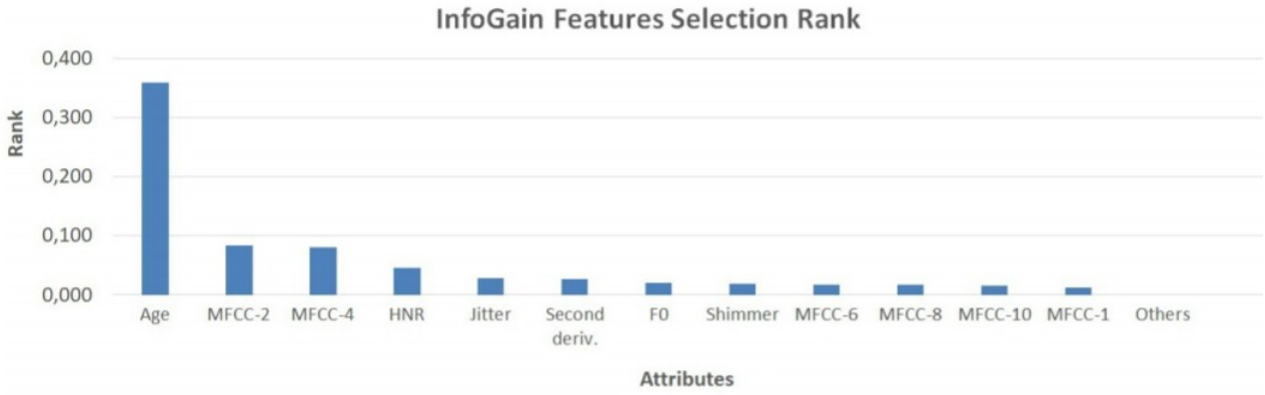


Fig. 3.3: Information gain estimated for each feature

- Correlation method : this assesses the predictive ability of each attribute, giving us the possibility of preferring sets of attributes that are highly correlated with the class. We have used 0.15 as our cut-off for relevant attributes.

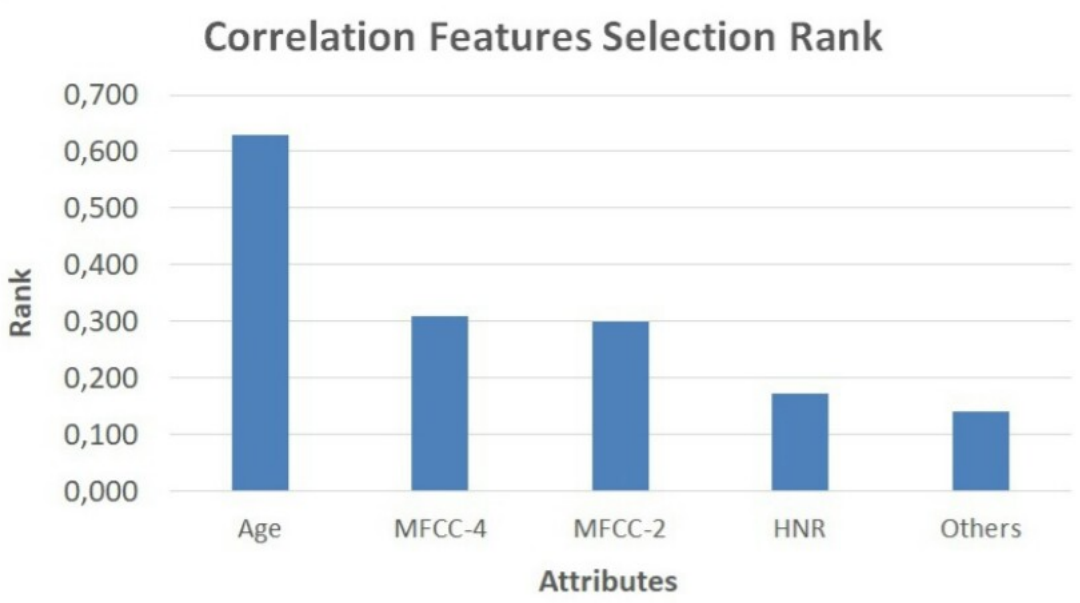


Fig. 3.4: Correlation rank obtained for each feature

The remaining attributes have been removed in accordance with the Fig3.4.

- Principal Components Analysis (PCA) method : Similarly to we used the PCA method[10] to select the most significant parameters. We selected the principal components which have obtained at least 50ranking.

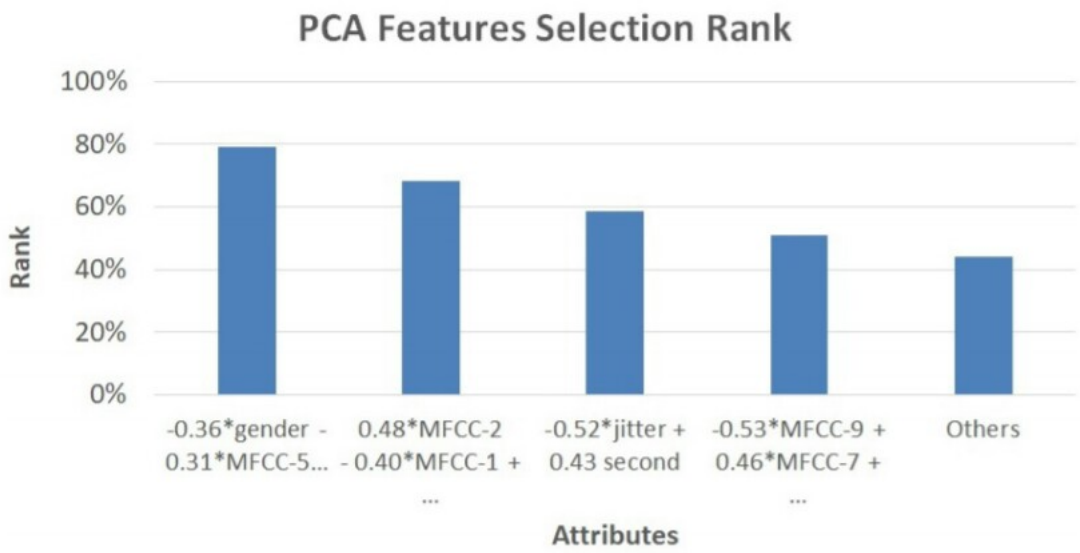


Fig. 3.5: PCA rank obtained in our study

As shown in Fig.3.5, we obtained four new parameters that are combination of several features.

The choice of these features selection methods was suggested by the wide use of such techniques in machine learning classifications to improve the overall quality of the patterns and/or the time required for the actual mining.

Conclusion

In recent years, the use of mobile multimedia services and applications in health-care sector has been increasing significantly. Mobile health applications allow people to access medical information and data of interest at any time and anywhere, useful for the monitoring and detection of specific diseases, such as dysphonia, a voice disorder often underestimated that affects a great percentage of people. Research on mobile automatic systems to estimate voice disorders has received considerable attention in the last few years due to its objectivity and non-invasive nature. Machine learning techniques can be a valid support to investigate new approaches to signal processing in an easy and fast way that can be implemented in an m-health solution. This study compares the performance of different voice pathology identification methods, taking into account the main machine learning techniques. Several techniques are applied such as the Support Vector Machine, Decision Tree, Bayesian Classification, Logistic Model Tree and Instance-based Learning algorithms. Moreover, in this work we focus on identifying appropriate voice signal features by using the comparative study of different classifiers. All analyses are performed on a wide dataset of 1370 voices selected from the Saarbruecken Voice Database.

The tests have been carried out over the overall dataset and over three different subset where we only have considered the selected features by three specific features selection methods. The results have shown that the best accuracy in voice pathology detection is achieved using the Support Vector Machine algorithm.

REFERENCES

[1] B. M. Silva, J. J. Rodrigues, I. de la Torre Dez, M. Lopez-Coronado, and K. Saleem, Mobile-health: A review of current state in 2015, Journal of biomedical informatics, vol. 56, pp. 265272, 2015.

[2] S. Naddeo, L. Verde, M. Forastiere, G. De Pietro, and G. Sannino, A real-time m-health monitoring system: An integrated solution combining the use of several wearable sensors and mobile devices. in HEALTHINF, 2017, pp. 545552.

[3] M. S. Hossain and G. Muhammad, Cloud-assisted industrial internet of things (iiot)enabled framework for health monitoring, Computer Networks, vol. 101, pp. 192202, 2016. Research: A State-of-the-Art Summary 4. Berlin, Germany: SpringerVerlag, 2015, pp. 18.

[4] G. Sannino, I. De Falco, and G. De Pietro, A supervised approach to automatically extract a set of rules to support fall detection in an mhealth system, Applied Soft Computing, vol. 34, pp. 205216, 2015.

[5] J. Mohammed, C.-H. Lung, A. Ocneanu, A. Thakral, C. Jones, and A. Adler, Internet of things: Remote patient monitoring using web services and cloud computing, in Internet of Things (iThings), 2014 IEEE International Conference on, and Green Computing and Communications (GreenCom), IEEE and Cyber, Physical and Social Computing (CPSCom), IEEE. IEEE, 2014, pp. 256263.

[6] R. Gravina and G. Fortino, Automatic methods for the detection of accelerative cardiac defense response, IEEE Transactions on Affective Computing, vol. 7, no. 3, pp. 286298, 2016.

[7] S. Iyengar, F. T. Bonda, R. Gravina, A. Guerrieri, G. Fortino, and A. Sangiovanni-Vincentelli, A framework for creating healthcare monitoring applications using wireless body sensor networks, in Proceedings of the ICST 3rd international conference on Body area networks. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2008, p. 8.

[8] R. H. G. Martins, H. A. do Amaral, E. L. M. Tavares, M. G. Martins, T. M. Goncalves, and N. H. Dias, Voice disorders: etiology and diagnosis, Journal of Voice, vol. 30, no. 6, pp. 761e1, 2016.

[9] A. R. Maccarini and E. Lucchini, La valutazione soggettiva ed oggettiva della disfonia. il protocollo sifel, ACTA PHONIATRICA LATINA, vol. 24, no. 1/2, pp. 1342, 2002.

[10] J. I. Godino-Llorente, P. Gomez-Vilda, N. S aenz-Lech on, M. Blanco- Velasco, F. Cruz-Roldan, and M. A. Ferrer-Ballester, Support vector machines applied to the detection of voice disorders, Lecture notes in computer science, vol. 3817, p. 219, 2005.