RESEARCH ARTICLE OPEN ACCESS

A Survey: Classification Techniques

ShamlaMantri[1], Ria Agrawal[2], ShraddhaBhattad[3], AnkitPadiya[4], RakshitRathi[5]
Department of Computer Science and Engineering
MAEER's MIT College of Engineering, University of Pune
Kothrud, Pune-411038
Maharashtra-India

ABSTRACT
Clinical depression has become a leading cause of mental illness and also major cause of suicides and deaths. Thus, early detection of depression among the individuals will help to reduce the problems of mental illness and deaths. The signs of depressions can be studied by sampling the speech signals of the individuals. Depressive disorder affects the acoustic qualities of their speech; hence depression can be detected by analyzing the acoustic properties of speech by firstly pre-processing the speech signal and then features can be extracted and this speech is then classified by using classification techniques, as depressed or controlled.

Keywords:- Clinical depression, depressive disorders, acoustic properties, pattern recognition, classification, depression detection.

I. INTRODUCTION

Clinical depression belongs to the group of affective (mood) disorders in which emotional disturbances consist of prolonged periods of excessive sadness marked by reduced emotional expression and physical drive [5].From psychological point of view, emotions expressed in his/her speech show major signs of person being depressed or normal. So analyzing the speech signals can help to detect the depression, as the acoustic qualities of speech gets affected by depression.Pre-processing, feature extraction and then classifying using classification technique to speech signal help to detect the depression.. Classification of speech signal gives us the exact results of the person being depressed or controlled.But due to signal processing technologies, a vast amount of data can be extracted, processed and can be stored.However, all the data obtained from the raw data can be useless if it do not make sense. Therefore, extracting useful data (patterns and trends) from raw data and presentation of the useful data to obtain results is a crucial step.

This process of generalizing decisions based on patterns obtained from raw data is known as pattern recognition.
Basically, two pattern recognition techniques are used:
1) Gaussian Mixture Model.
2) Support Vector Machine.

II. GAUSSIAN MIXTURE MODEL

Gaussian Mixture Model is widely used model for the distribution of continuous variables is known as the Gaussian (or normal) distribution. For a variable with a single dimension *x*, the Gaussian distribution is of the form [2]:

$$N(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right] \quad (1)$$

Where μ is the mean and σ^2 is the variance of the normal distribution. The normal distribution is denoted by$N$ (x |μ , σ^2) with the argument in the function standing for probability of x given mean μ and variance σ^2.Translating Eq. (1) from a single variable to a D-dimension vector **x**, theGaussian distribution is written as [2]:

$$N(x|\mu,\Sigma) = \frac{1}{|\Sigma|^{1/2}(2\pi)^{D/2}} \exp\left[-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)\right] \quad (2)$$

Where **x** is a D-component column vector, μ is the D-component mean vector, Σ is the D by D covariance matrix, | Σ | and $\Sigma^{-1}$are its determinant and inverse respectively.

$(x - \mu)^t$ denotes the transpose of (**x** - μ).

Distribution of different random variables can be modelled using Gaussian distribution. However, this has severe limitations when it comes to modelling on real datasets [3]. To improve these limitations, we consider a linear combination of group of *M* Gaussian densities of the form:

$$p(x) = \sum_{m=1}^{M} w_m N(x|\mu_m, \Sigma_m) \quad (3)$$

Where w_m is the mixing coefficients and *M* is the total number of Gaussian mixtures. From Eq. (3) which is termed mixture of Gaussians, each Gaussian density N (x |μ_m, Σ_m) has its own mixing coefficient, mean μ_m and covariance Σ_m. Note that these individual Gaussian components are normalized so that [2]:

$$\sum_{m=1}^{M} w_m = 1 \quad (4)$$

where $0 \leq w_m \leq 1$. We therefore see that the mixing coefficients satisfy the requirements to be probabilities.
From the sum and product rules, the marginal density is given by [2]:

$$p(\mathbf{x}) = \sum_{m=1}^{M} p(m)\, p(\mathbf{x}|m) \quad (5)$$

Which is equivalent to Eq. (3), whereby $p(m) = w_m$ is the probability of the *mth*mixture component and the density$p(\mathbf{x}|m) = N(\mathbf{x}|\mu_m, \Sigma_m)$ is known as the probability of **x** conditioned on *m*. Now if we want to find the Gaussian mixture component from which vector **x** come, we can reverse the conditional probability by using Bayes' theorem [2]:

$$p_{posterior}(\mathbf{x}) \equiv p(m|\mathbf{x}) = \frac{p(m)\, p(\mathbf{x}|m)}{p(\mathbf{x})} = \frac{w_m N(\mathbf{x}|\mu_m, \Sigma_m)}{\sum_t^M w_t N(\mathbf{x}|\mu_t \Sigma_t)} \quad (6)$$

Log likelihood function is computed to find out the parameters of $w \equiv \{w_1, \ldots, w_m\}$, $\mu \equiv \{\mu_1, \ldots, \mu_m\}$, $\Sigma \equiv \{\Sigma_1, \ldots, \Sigma_m\}$ in Eq. (1.3) & Eq. (1.6) by using [2]:

$$\ln p(\mathrm{X}|\mathrm{w}, \mu, \Sigma) = \sum_{n=1}^{N} \ln[\sum_{m=1}^{M} w_m N(\mathrm{x}_m|\mu_m, \Sigma_m)] \quad (7)$$

Where $X=\{x_1, \ldots, x_N\}$.
Optimized parameters are to be found of each Gaussian mixture using iterative optimization techniques to maximize the likelihood function by using expectation maximization (EM) technique. A two stage process is involved in EM iterative method [2]. Current estimate of the latent variables is used to evaluate the expectation (E)-step of the log-likelihood. The second stage is the maximization (M)-step, the parameters is re-estimated by maximizing the expected log-likelihood found in the E-step. A repeated process is done until a convergence criterion is satisfied. [2]

The expectation maximization process is illustrated in the plots depicted in Fig 1. Plot (a) shows an example of data points in two-dimensional Euclidean space coloured in green. Plot (b) shows the first stage of the E-step where the mean μ_{init}, covariance Σ_{init} and mixing coefficient w_{int}parameters in each of the two Gaussian components (highlighted in blue and red) are initialized. Next, the posterior probabilities of the data points to each Gaussian component are evaluated. The data points highlighted in blue (8) indicates that the posterior probabilities are closer to the blue Gaussian component while the data points coloured in red (9) are closer to the red Gaussian component. The data points that have probabilities belonging to either Gaussian component are depicted by a pink triangle (:). Plot (c) shows the first M-step in re-estimating the new means, covariance and mixing coefficient of the Gaussian components from the data points that were newly assigned to either the red or the blue component based on their current posterior probabilities. The log likelihood probability is then maximized by repeating the cycle of the E and M steps until convergence criteria is met as illustratedinplots(d)–(i) [2].

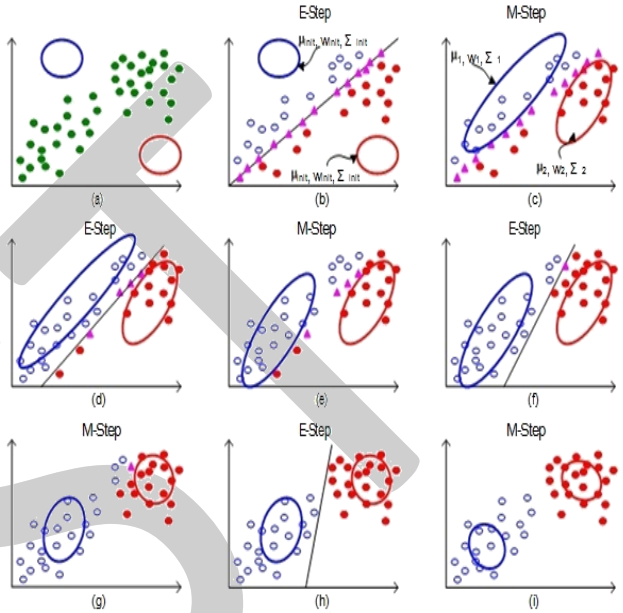

Figure 1

Figure 1: Illustration of the expectation maximization (EM) iterative optimization technique for a mixture of two Gaussian components (adapted from Bishop [3]). (a) Green points denote an example of a dataset in two-dimensional Euclidean space. (b) First stage, expectation (E) step: Initialization of the parameters mean μ_{init}, covariance Σ_{init} and mixing coefficient $w_{int}$and evaluating the posterior probabilities of the data points to each Gaussian component. (c) Second stage, maximization (M) step: Re-estimating the parameters using the current posterior probabilities and calculating the log likelihood. (d)-(i) show subsequent E and M steps through to the final convergence of the log likelihood.

III. SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) [4] is linear classifier. Compared to the other classifiers, Support Vector Machine yields good performance to algorithms in binary classification problems [3]. SVM with a two-classification problem using a simple form of linear classification [2].

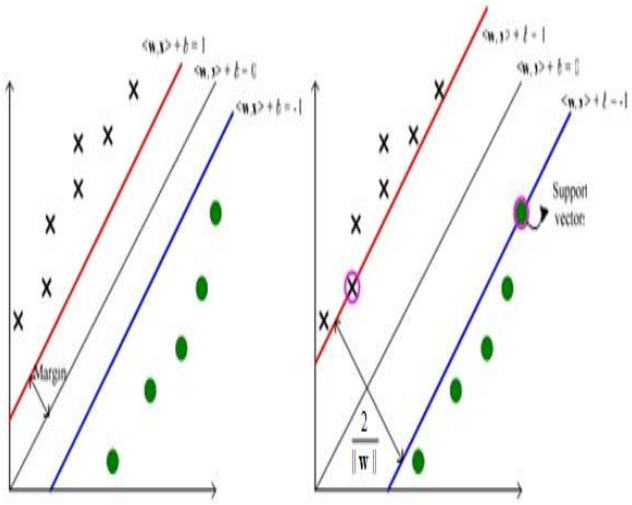

Figure 2: A maximal margin hyper plane with its support vectors highlighted in circles.

Given N number of data points with d-dimensional features (variables) R^d, where $\mathbf{x} \in \mathbf{R^d}$ is the d-dimensional input vector and $q \in \{-1, 1\}$ is a class for binary classification. The decision function is of the form [2]:

$$q = sign\ (\boldsymbol{w}.\ \boldsymbol{x} + b) \qquad (8)$$

Where b is bias and w is the weight vector and q_i is associated label, $b \in R$ and $w \in R^d$ are the parameters that control the function of the decision boundary given by Eq. (8). The linear data set is linearly separable in input space is assumed.

Margin is defined as the smallest distance between the decision boundary and any of the samples as illustrated in Fig. 2 (left).Margin is used to form linear classifier in SVM. To maximize the distance of the margin between two parallel hyper-planes which separates two groups (classes) is the main objective of margin. This is illustrated in Fig.2 (right) where the linear classifier defined by the hyper-plane $\langle \mathbf{w}, \mathbf{x} \rangle + b = 0$ is midway between the separating hyper-planes. The support vector that is boundary is highlighted by the circles in Fig. 1.2 (right). This margin can be computed as $2 / \|w\|$ [2].

If the training data set is not linearly separated we can we can map the non-separable data from the input space to a higher dimensional feature space whereby the data is transformed to be linearly separable in which the linear models can be used. So we can write the equation as [2]:

$$q = sign\ (w.\ \phi(\boldsymbol{x}) + b) \qquad (9)$$

where $\phi : \mathbf{x} \rightarrow \phi(\mathbf{x})$; is a non-linear map from the input space to some feature space. This means that we can build non-linear machines in two steps: first a fixed non-linear mapping transforms the data into a feature space $\phi(\mathbf{x})$, and then a linear machine is used to classify the data in the feature space.

The decision rule can be evaluated using just inner products between the test point and the training point by expressing equation (9) as a linear combination of training points [2].

$$q = \sum_{i=1}^{l} a_i q_i \ \langle \phi(\mathbf{x}_i)\phi(\mathbf{x}) \rangle + b \qquad (10)$$

Exact separation of the training data in the original input space $\mathbf{x}$ will be given by the resulting support vector machine, although the corresponding decision boundary is nonlinear. In real world datasets, exact separation of the training data can lead to poor generalization due to the class-condition distributions overlap. We need to modify the support vector to allow some of the training points to be misclassified to find the solution in searching for the maximum margin classifier of the following optimization problem [2]:

$$\text{Zminimize} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{l} \xi_i$$

$$\text{subject to } q_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0 \qquad (11)$$

Where $b \in \boldsymbol{R}$ and $w \in \boldsymbol{R^d}$ $x_i \in \boldsymbol{R^d}$ is the i^{th} data sample with a d-dimensional feature vector, $q_i \in \{-1, 1\}$ is the class labels, and l is the number of training points. The function ϕ maps the training vectors x_i into a higher dimensional space. The first constraint dictates that points with equivalent labels are on the same side of the line. The slack variable ξ_i allows data to be misclassified while being penalized at rate C in the objective function in Eq. (11). Therefore this allows SVM to handle non-separable data in real-world situations.

IV. CONCLUSION

In this paper, we briefed various types of classification techniques used for depression analysis. We have discussed these classification techniques and also explained their functionality in detail. We can conclude that the GMM (Gaussian Mixture Model) when used provide less accurate results than SVM (Support Vector Machine). The both linear and non-linear features of speech signal can be classified using the SVM (Support Vector Machine) classifier.

ACKNOWLEDGMENT

We wish to express our sincere gratitude to Prof Dr. V. M. Wadhai, Principal and Prof. Dr.PrassanaJoeg H.O.D of Computer Engineering Technology, MITCOE PUNE for guiding us in this survey. We also thank to our friends and other MITCOE staff members for guidance and encouragement in carrying out this work. We would like to thank them for their valuable guidance throughout the preparation of this.

REFERENCES

[1] Lu-Shih Alex Low*, Namunu C. Maddage, Margaret Lech, Lisa B. Sheeber, and Nicholas B. Allen, "Detection of Clinical Depression in Adolescents Speech During Family Interactions", ieee transactions on biomedical engineering, vol. 58, no. 3, march 2011.

[2] Low Lu-Shih Alex, "Thesis Detection of Clinical Depression in Adolescents Using Acoustics Speech Analysis", May 2011.

[3] BISHOP C. M., Pattern recognition and machine learning. New York: Springer, 2006. [1]

[4] BOSER, B. E., GUYON, I. M., and VAPNIK, V. N., "A training algorithm for optimal margin classifiers," Proceedings of the fifth annual workshop on Computational learning theory, New York, NY, USA: ACM, pp. 144-152, 1992.

[5] J. O. Cavenar, H. Keith, H. Brodie, and R. D. Weiner, Signs and Symptomsin Psychiatry.Philadelphia: Lippincott Williams & Wilkins, 1983.