Journal of Computer Science 7 (7): 991-996, 2011
ISSN 1549-3636

A Novel Linear-Polynomial Kernel to Construct Support Vector Machines for Speech Recognition

[1]Balwant A. Sonkamble and [2]D.D. Doye
[1]Pune Institute of Computer Technology, Pune, Maharashtra State, India
[2]Department of E and TC, SGGSIE and T, Nanded, Maharashtra State, India

Abstract: Problem statement: To accept the inputs as spoken word utterances uttered by various speakers, recognize the corresponding spoken words and initiate action pertaining to that word. **Approach:** A novel Linear-Polynomial (LP) Kernel function was used to construct support vector machines to classify the spoken word utterances. The support vector machines were constructed using various kernel functions. The use of well known one-versus-one approach considered with voting algorithm. **Results:** The empirical results compared by implementing various kernel functions such as linear kernel function, polynomial kernel function and LP kernel functions to construct different SVMs. **Conclusion:** The generalization performances based on the One-versus-One approach for speech recognition were compared with the novel LP kernel function. The SVMs using LP kernel function classifies the spoken utterances very efficiently as compared to other kernel functions. The performance of the novel LP kernel function was outstanding as compared to other kernel functions.

Key words: Hidden Markov Model (HMM), neural network, Empirical Risk Minimization (ERM), kernel function, voting algorithm, Modified Fuzzy-Hyper sphere Neural Networks (MFHNN), Support Vector Machines (SVM), hyperplane, Vapnik-Chervonenkis (VC), Linear-Polynomial (LP)

INTRODUCTION

From last several years, the speech recognition research playing a leading role in more number of applications. Many new techniques emerged including Modified Fuzzy-Hyper sphere Neural Networks (MFHNN), Neural Networks (Doye *et al*., 2002; Solaimani, 2009), Hidden Markov Models (Ping *et al*., 2009), Bayesian Networks (Mansouri *et al*., 2011) and Dynamic Time Warping decade to decade to increase the performance of the speech recognition systems but Hidden Markov Model (HMM) (Rabiner and Juang, 1993; Doye *et al*., 2002) is among the most successful state of art tools widely used but still speech recognition systems are far away to achieve high-performance as well as accuracy. The HMM are originally a generative models because the decisions are based on the likelihood estimation of the currently evaluated pattern. Thus, HMM requires additionally discriminative approaches to discriminate the speech samples. The limitation of HMM, is the loss of performance due to the mismatch between training and testing conditions.

The Support Vector Machine (SVM) (Clarkson and Moreno, 1999; Scholkopf and Smola, 2002) is emerged as a new machine learning technique for pattern classification. The SVMs are based on the discriminative approach which discriminates the patterns by finding the global minima. The SVM uses Structural Risk Minimization (SRM) principle to construct linear and nonlinear classifiers with Vapnik-Chervonenkis (VC) dimension (Vapnik, 1998; Cristianini and John, 2000). The VC dimension controls the capacity of the learning machine. The linear and nonlinear approaches are used to construct the SVMs. In linear methods inner products called dot products are considered for generating the optimal separating hyperplane for classifying the two classes where as in non linear approach dot products are replaced by kernel functions (Burges, 1998; Scholkopf and Smola, 2002) to construct the optimal separating hyperplane.

In this study we propose to use novel kernel function called the Linear-Polynomial (LP) (Cristianini and John, 2000) kernel including the description about the construction of linear support vector machine and the construction of nonlinear support vector machine

Corresponding Author: Balwant A. Sonkamble, Pune Institute of Computer Technology, Pune, Maharashtra State, India

along with description of the novel LP kernel function. The description of the classification approach also described and lastly, explained the detailed experimental results obtained by comparison with basic kernel function.

MATERIALS AND METHODS

The Hidden Markov models are most successful techniques for modeling a speech by determining the speech sound representation. The Classification problems are treated as complex problems. In classification problem, the main task is to classify the problem directly by estimating the decision surfaces. Many researchers have proved that, the Support Vector Machines (SVM) are the most efficient and popular generalized linear classifiers used for data classification. The Support Vector Machines are the machine learning techniques developed by Vapnik in 1960's can perform static classification tasks. The SVMs are applied successfully for solving pattern recognition problems due to its discriminative nature. The SVMs are also called hyperplane classifiers because it constructs optimal separating hyperplane to discriminate between two classes. The learning machines are constructed by nonlinearly mapping from input vector space to a high dimensional vector space called feature space. The SVMs are not designed to handle temporal structure of data. The SVMs has very good generalization ability that improves the system robustness of speech recognition tasks in noisy environment. The SVMs key property is to minimize the empirical classification error and maximize the geometric margin simultaneously. Hence it is also known as maximum margin classifier.

Construction of linear support vector machines: The learning machines can construct optimal predictor through a set of functions. Risk minimization means minimizing the functional from a given training data that is minimizing the optimal parameterization. The Empirical Risk Minimization (ERM) (Vapnik, 1998; Daniels and Ejara, 2009) is a kind of risk minimization commonly used as optimization procedure in machine learning. The optimization process depends on the loss functions because prior joint probability distribution is not known. The risk can be determined as a mean error computed from the fixed number of training data. Here, the risk is defined as measures of quality of a chosen function. ERM is computationally simpler than attempting to minimize the actual risk but due to non measuring capacity the machine, if the complexity of a machine increases then a machine over fits the data.

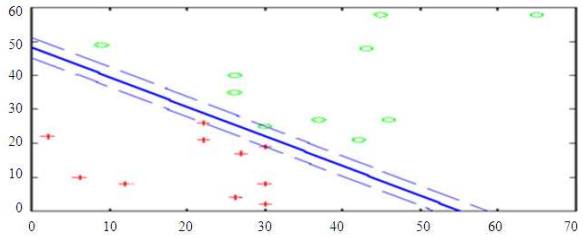

Fig. 1: Optimal separating hyperplane for seaparating two classes linearly

In the structural risk minimization, the optimal function not only depends on the loss functions for calculating the expected risk but also depends on its structure. Here, the risk is determining through VC dimension which measures the capacity of a learning machine by computing the upper bound.

The SRM principle is implemented by constructing the SVMs. The linear classifier in separable case the two datasets can be perfectly mapped. The separating hyperplane is called linear hyperplane separates the given datasets by maximizing the margin. The SVMs are constructed by constructing the binary classes. Consider binary classification problems by assuming the training data, given below:

$$(x_1, y_1), (x_{2,} y_2), ..., (x_1, y_1) \quad (1)$$

Here:

x_1 = The input patterns

y_1 = Outputs labeled by +1 and –1

The goal is to find the linear decision function f(x) and the separating hyperplane H, where H: x.w + b = 0 and f(x) = sgn(x.w+b)**.** Where bthe distance of the hyperplane from the origin is also referred as bias and w is the normal to the decision region also referred as weights. The value of H is calculated using quadratic programming approach. Figure 1 shows the optimal separating hyperplane to find the decision boundaries between the two classes. The margin of the SVM is defined as the distance from the separating hyperplane to the closest two classes. The margin is equal to 2/||w|| inequalities. Here, distance between the dotted lines is called margin and the data points appeared on the dotted lines are called support vectors.

The optimal hyperplane is obtained by applying scaling on the parameters w and b because scaling avoids variance among the data values. The existence of optimality is guaranteed by the Karush-Kuhn-Tucker (KKT) (Vapnik, 1998) theorem. The main feature of the optimal hyperplane is to maximize the margin while minimizing the empirical risk. For separating the two

realistic datasets linear classifier in non separable case considers the slack variable to find the misclassification errors.

Construction of Nonlinear support vector machines (Sonkamble and Doye, 2008): The non-linear classifiers can handle the decision boundaries in the complex nonlinear data very efficiently. The use of kernel functions is essential to construct optimal hyperplanes of non-linear classifiers**.** The kernels are positive-definite functions to map data into high dimensional spaces which increases the computational power of linear machines. The key advantages of the kernels are firstly, it incorporates prior knowledge of the problem by defining a similarity measure between two data points. Secondly, kernel function finds the kernel matrix which contains all the information about the input space and thirdly, the number of operations required is not necessarily proportional to the number of features. There are various kinds of kernel functions used commonly for speech classification. The kernel functions should satisfy the mercer's condition which shows the symmetry property. The mapping is achieved through a replacement of the inner product:

$$x_i.x_j \rightarrow \Phi(x_i).\Phi(x_j)$$

The functional form of the mapping Φ(x), does not need to be known since it is implicitly defined by the choice of kernel:

$$k(x_i.x_j) \rightarrow \Phi(x_i).\Phi(x_j)$$

Each choice of kernel will define a different type of feature space and the resulting classifiers will perform differently on test data, though good generalization. For an SVM with RBF kernels the resulting architecture is an RBF network (Cortes and Vapnik, 1995; Mahi and Izabatene, 2011). However, the method for determining the number of nodes and their centers is quite different from standard RBF networks with the number of nodes equal to the number of support vectors and the centers of the RBF nodes identified with the support vectors themselves.

Formation of New kernels: There are different kernel functions commonly used for classification. In this case we are proposing two novel kernel functions by combining the linear kernel function with polynomial kernel function called Linear-Polynomial (LP) Kernel (Kurtz, 1991; Tan and Wang, 2004) function which formalized as follows:

$$k(x_i.x_j) \rightarrow k_1(x_i.x_j) + k_2(x_i.x_j)$$

Where:

$k_1(x_i.x_j) = (x_i.x_j)$ = A linear kernel function

$k_2(x_i.x_j) = (x_i.x_j + 1)^d$ = A polynomial kernel function

The construction of the optimal hyperplane is of the form:

$$f[x,\alpha_0] = \sum_{i=1}^{l} \alpha_i^0 y_i (x_i.x) + b \quad (2)$$

Here, b-indicates threshold as a constant and (x_i,x) indicates inner product of two input vectors as well as ℓ-indicates number of data pairs. The maximum-margin separating hyperplane called optimal hyperplane which reduces the generalization errors. The objective function of our optimization problem is the form:

$$D(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j (x_i.x_j) \quad (3)$$

Such that:

$$\alpha_i \geq 0 \text{ and } \sum_{i=1}^{l} \alpha_i \alpha_j = 0, i = 0,1,...,l \quad (4)$$

where, α_i are the Lagrange multipliers which define the weights of the model as $w_i = \alpha_i y_i$.

The construction of decision functions are depends on generating the inner product in a feature space which are nonlinear in their input space as given below:

$$f(x,\alpha) = \text{sign}\left(\sum_{\text{sup portvectors}} y_i \alpha_i^0 k(x_i.x) + b \right) \quad (5)$$

and are equivalent to linear decision functions in the feature space $z_1(x),...,z_\gamma(x),...$

$$f(x,\alpha) = \text{sign}\left(\sum_{\text{sup portvectors}} y_i \alpha_i^0 \sum z_\gamma(x_i) z_\gamma(x) + b \right) \quad (6)$$

The kernel function can be represented as $k(x_i, x_j)$ which generates the inner product for the feature space. The commonly used kernel functions are.

The Linear Kernel function is represented by the inner product given by the equation:

$$k(x_i.x_j) = (x_i.x_j) \quad (7)$$

The polynomial kernel has more number of hyperparameters than the RBF kernel which influences the complexity of model selection. The Polynomial Kernel function is generated for finding the inner product given by the equation:

$$k(x_i.x_j) = (x_i.x_j + 1)^d \quad (8)$$

Here, d is the polynomial degree which is a positive integer.

The LP kernel function can be represented as combined kernel functions of linear and polynomial kernels which is formulated as below:

$$k(x_i.x_j) = k_1(x_i.x_j) + k_2(x_i.x_j) \Rightarrow$$
$$k(x_i.x_j) = ((x_i.x_j) + (x_i.x_j + 1)^d) \quad (9)$$

The decision function can be constructed in the form of:

$$f(x,\alpha) = \text{sign}(\sum_{\text{sup portvectors}} y_i\alpha_i k(x_i.x) + b \quad (10)$$

The Radial Basis Function (RBF) kernel nonlinearly maps input samples into a higher dimensional space, which can handle the relation between class labels and attributes is nonlinear. The RBF kernel is not efficient when the number of features is very large as compared to other kernel functions. The Radial Basis Function kernel can represent as:

$$k(x_i.x_j) = \exp(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}) \quad (11)$$

The decision discriminative decision function is determined by the following equation:

$$D(x) = \sum_{i=1}^{l} y_i y_j \alpha_i \alpha_j k(x_i.x) + b = \sum_{i=1}^{1} w_i k(x_i.x) \quad (12)$$

This gives a decision about the classes to discriminate among them.

Classification approach: The speech recognition problem is a multiclass classification problem where as SVMs are efficiently solve binary classification problem. There are two approaches to solve multiclass problem by using SVM. First, One-versus-One (Ganapathiraju *et al*., 2000) classification approach also called pair wise classification by simply constructs for each pair of classes a classifier which separates those classes and second, One-versus-All classification approach (Osuna *et al*., 1997; Chin, 1999) by constructing for each class a classifier which separates that class from the reminder of data. All data with the exception of one row is used to train the learning algorithm.

One-versus-One classification approach: One-versus-One is one of the most commonly used successful approaches for discrimination of classes. The classifiers required according this approach is equal to k(k-1)/2 classifiers. Where k = 10, that is 45 classifiers are constructed. The One-versus-One classification approach is also called pair-wise classification approach where only pair-wise data points can be considered to discriminate between the two classes. The main feature is that, it reduces the generalization error rate by reducing the number of support vectors hence is faster than the One-versus-All approach. A voting scheme algorithm used with fixed weights to cast one vote in favor the class. This algorithm force to choose among one class. These votes are distributed uniformly so that we can classify the correct classes of the speech signals by considering the highest score. The One-versus-One approach requires more memory space as well as requires more time for training.

RESULTS AND DISCUSSION

The database was collected from 5 Indians. The database was collected for 10 digits uttered by 15 times. The speech features are extracted and obtained LPC Coefficients. The speech signals were sampled at 8 KHz divided into a sequence of data blocks, each block spanning 20ms and separated by 10 ms. The speech features are extracted and obtained LPC Coefficients and these LPCC were used as a data points for training the SVM. The number speech samples used for training were 50 from each digit and rests of the samples were used as testing data for speech signal classification. We have constructed various SVMs using linear kernel function, Polynomial kernel function and the proposed LP kernel function. When LP kernel function used to construct nonlinear support vector machines, it gives very good performance as compared to linear kernel function as well as polynomial kernel function. The observation is that, it maximizes the margin with small fraction value increased by 0.01 to 0.001 as compared to the margin obtained by polynomial kernel function. Hence, the LP kernel reduces generalization error drastically so it discriminates the data points very accurately as compared to polynomial kernel function. Table 1 shows the training performance of the polynomial kernel function for calculating the support vector while the training performance of the LP kernel function is shown in Table 2. The compared training performance graph is also shown Fig. 2. The LP kernel also finds better number of support vectors as compared to polynomial kernel function.

Table 1: Generalization performance for one-versus-one classifier using polynomial kernel function for training data

1	2	3	4	5	6	7	8	9	0
-	5.8	7.7	5.8	7.7	7.7	7.7	7.7	9.6	7.7
-	-	5.8	7.7	5.8	5.8	3.8	5.8	5.8	7.7
-	-	-	5.8	7.7	9.6	7.7	7.7	7.7	9.6
-	-	-	-	7.7	3.8	5.8	5.8	5.8	7.7
-	-	-	-	-	7.7	7.7	9.6	9.6	7.7
-	-	-	-	-	-	5.8	9.6	7.7	7.7
-	-	-	-	-	-	-	7.7	7.7	9.6
-	-	-	-	-	-	-	-	7.7	9.6
-	-	-	-	-	-	-	-	-	9.6

Table 2: Generalization performance for one-versus-one classifier using LP kernel function for training data

1	2	3	4	5	6	7	8	9	0
-	5.8	7.7	5.8	7.7	7.7	7.7	7.7	7.7	9.6
-	-	5.8	7.7	5.8	5.8	3.8	5.8	5.8	7.7
-	-	-	5.8	7.7	9.6	7.7	7.7	7.7	9.6
-	-	-	-	5.8	3.8	5.8	5.8	5.8	7.7
-	-	-	-	-	7.7	7.7	9.6	9.6	7.7
-	-	-	-	-	-	5.8	9.6	7.7	7.7
-	-	-	-	-	-	-	7.7	7.7	9.6
-	-	-	-	-	-	-	-	7.7	9.6
-	-	-	-	-	-	-	-	-	9.6

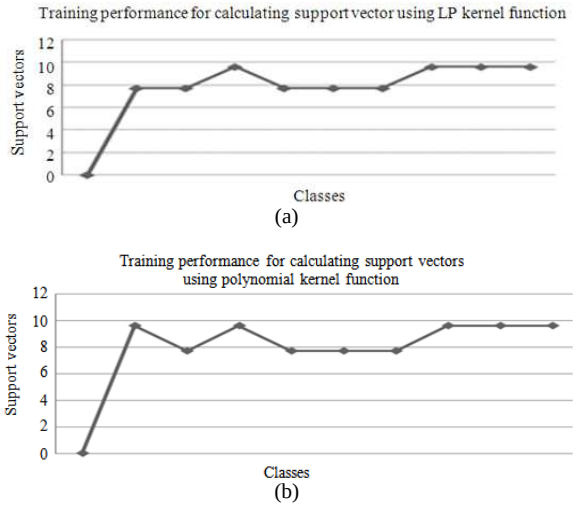

Fig. 2: The comparative training performance graph

The existing kernel functions such as linear and polynomial kernel functions are implemented in addition to the implementation of LP kernel function for the construction of support vector machines.

CONCLUSION

In this study, the proposed novel LP kernel function outperforms as compared to polynomial kernel function and linear kernel functions. This kernel function can be considered as a more suitable for classifying the nonlinear signals. The speech classification experiment conducted shows more accurate results as compared to polynomial kernel function by discriminating the decision boundaries between two speech data points. We considered more accurate approach as a One-versus-One to achieve better performance as compared to One-versus-All approach. In future work, the LP kernel functions can be compared with RBF kernel functions.

REFERENCES

Burges, C.J.C., 1998. A Tutorial on support vector machines for pattern recognition. Knowl. Discovery Data Min., 2: 121-167. DOI: 10.1023/A:1009715923555

Chin, K.K., 1999. Support vector machines applied to speech pattern classification. Master's Thesis, Engineering Department, Cambridge University. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.27.257

Clarkson, P. and P.J. Moreno, 1999. On the use of support vector machines for phonetic classification. Proceeding of the IEEE International Conference. on Acoustics, Speech and Signal Processing (ICASSP), Phoenix, Arizona, USA., pp: 585-588. DOI: 10.1109/ICASSP.1999.759734

Cortes, C. and V. Vapnik, 1995. Support vector networks. Mach. Learn., 20: 1-25. DOI: 10.1234/12345678

Cristianini, N. and S.-T. John, 2000. An introduction to support vector machines and other kernel-based learning methods. 1st Edn., Cambridge University Press, John Shawe-Taylor, Royal Holloway, London, ISBN: 9780521780193, pp: 204.

Daniels, K. and D.D. Ejara, 2009. Impact of information asymmetry on municipal bond yields: An empirical analysis. Am. J. Econ. Bus. Admin., 1: 11-20. DOI: 10.3844/ajebasp.2009.11.20

Doye, D.D., U.V. Kulkarni and T.R. Sontakke, 2002. Speech recognition using modified fuzzy hypersphere neural network. Proceedings of the International Joint Conference on Neural Networks, May 12-17, Honolulu, HI, USA., pp: 65-68. DOI: 10.1109/IJCNN.2002.1005443

Ganapathiraju, A., J. Hamaker and J. Picone, 2000. Hybrid SVM/HMM architectures for speech recognition. Proceedings of the 6th International Conference on Spoken Language Processing, Oct. 16-20, Beijing, China, pp: 504-507. http://www.isca-speech.org/archive/icslp_2000/i00_4504.html

Kurtz, M., 1991. Handbook of Applied Mathematics for Engineers and Scientists. 1st Edn., McGraw Hill, New York, 608. DOI: 10.1036/0070356858

Mahi, H. and H.F. Izabatene, 2011. Segmentation of satellite imagery using RBF neural network and genetic algorithm. Asian J. Applied Sci., 4: 186-194. http://198.171.234.225/fulltext/emailthisarticle.php?doi=ajaps.2011.186.194&org=

Mansouri, M., A. Ganguly and A. Mostashari, 2011. Evaluating agility in extended enterprise systems: A transportation network case. Am. J. Eng. Applied Sci., 4: 142-152. DOI: 10.3844/ajeassp.2011.142.152

Osuna, E., R. Freund and F. Girosi, 1997. An improved training algorithm for support vector machines. Proceeding of the IEEE Workshop on Neural Networks for Signal Processing, Amelia Island, Florida, USA., pp: 276-285. DOI: 10.1109/NNSP.1997.622408

Ping, Z., T. Li-Zhen and X. Dong-Feng, 2009. Speech recognition algorithm of parallel subband HMM based on wavelet analysis and neural network. Inform. Technol. J., 8: 796-800. http://docsdrive.com/pdfs/ansinet/itj/2009/796-800.pdf

Rabiner, L.R. and B.H. Juang, 1993. Fundamentals of Speech Recognition. 1st Edn., Prentice Hall, Englewood Cliffs, New Jersey, USA., ISBN: 10: 0130151572, pp: 496.

Scholkopf, B. and A. Smola, 2002. Learning with Kernels. 1st Edn., MIT Press, Cambridge, Mass, USA., ISBN-10: 9780262194754, pp: 644.

Solaimani, K., 2009. A study of rainfall forecasting models based on artificial neural network. Asian J. Applied Sci., 2: 486-498. DOI: 10.3923/ajaps.2009.486.498

Sonkamble, B.A. and D.D. Doye, 2008. An overview of speech recognition system based on the support vector machines. Proceeding of the International Conference on Computer and Communication Engineering, May 13-15, IEEE Xplore, Kuala Lumpur, pp: 768-771. DOI: 10.1109/ICCCE.2008.4580709

Tan, Y. and J. Wang, 2004. A support vector machine with a hybrid kernel and minimal vapnik-chervonenkis dimension. IEEE Trans. Knowl. Data Eng., 16: 385-395. DOI: 10.1109/TKDE.2004.1269664

Vapnik, V.N., 1998. Statistical Learning Theory. 1st Edn., John Wiley and Sons, New York, USA., ISBN-10: 0471030031, pp: 736.