Multi-Level Audio Classification Architecture

Jozef VAVREK, Jozef JUHAR

Department of Electronics and Multimedia Telecommunications, Faculty of Electrical Engineering and Informatics, Technical University of Kosice, Park Komenskeho 13, 042 00 Kosice, Slovak Republic

jozef.vavrek@tuke.sk, jozef.juhar@tuke.sk

DOI: 10.15598/aeee.v13i4.1454

Abstract. *A multi-level classification architecture for solving binary discrimination problem is proposed in this paper. The main idea of proposed solution is derived from the fact that solving one binary discrimination problem multiple times can reduce the overall miss-classification error. We aimed our effort towards building the classification architecture employing the combination of multiple binary SVM (Support Vector Machine) classifiers for solving two-class discrimination problem. Therefore, we developed a binary discrimination architecture employing the SVM classifier (BDASVM) with intention to use it for classification of broadcast news (BN) audio data. The fundamental element of BDASVM is the binary decision (BD) algorithm that performs discrimination between each pair of acoustic classes utilizing decision function modeled by separating hyperplane. The overall classification accuracy is conditioned by finding the optimal parameters for discrimination function resulting in higher computational complexity. The final form of proposed BDASVM is created by combining four BDSVM discriminators supplemented by decision table. Experimental results show that the proposed classification architecture can decrease the overall classification error in comparison with binary decision trees SVM (BDTSVM) architecture.*

Keywords

Audio data classification, binary discrimination architecture, support vector machine.

1. Introduction

Currently, the enormous amount of audio-visual data is available on the internet and various audio-visual databases. There is a need to manage the audio and video content. Various content-based techniques have been implemented in order to process these data automatically. The efficient processing of audio data is inevitable for applications like automatic speech recognition, classification and retrieval. A big effort is directed towards content-based analysis of audio data containing various, hard to discriminate, acoustic classes. The aim of this paper is therefore focused on content-based classification of BN audio data utilizing an efficient classification architecture. Moreover, we built the classification system also with intention to use it for refinement the acoustic models for each particular audio class and lower the word error rate of the automatic speech recognition (ASR) system.

In general, there are six acoustic classes with frequent occurrence in BN audio stream, namely pure speech, speech with environment sound, environment sound, speech with music, music and silence [1]. Each individual class is characterized by unique acoustic properties and random occurrence in audio stream. So far, various different approaches to discrimination of multiple classes have been investigated and compared. Some of them focus on the automatic selection of the most efficient features and the optimal thresholds, utilizing rule-based classification algorithms [2], [3]. Typically some long-term statistics, such as the mean or the variance, and not the features themselves, are used for the discrimination. Other works point to the importance of using appropriate classification architecture employing robust machine learning classifiers [4], [5]. Other authors also reported superior position of SVM in many classification tasks, especially in case of classifying audio stream with various acoustic events [6], [7], [8]. That was the main reason why we decided to use the SVM classifier as the core element in our binary discrimination architecture supplemented by sufficient features. The aim was to minimize miss-classification error and increase the overall classification accuracy for broadcast news audio data. The fundamental principles are based on a successive discrimination amendment where each block tries to correct what previous one miss-classified. The total number

of binary classifiers needed to discriminate N classes is $n \times N$, where n refers to the number of discriminators for one classification level (in our proposal $n = 4$). It follows a basic stepwise classification condition, where combination of n classifiers is aligned in such configuration which outputs only two decision values $+1$ and -1. It is considered as empirical approach.

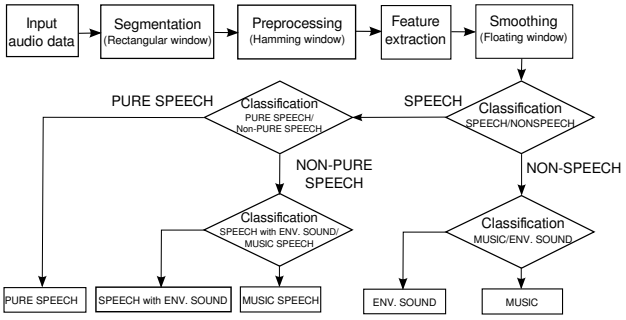


Fig. 1: Proposed classification architecture.

The overall system for classification of BN audio data is illustrated in Fig. 1. It utilizes the fundamental principles of audio processing techniques applied in BDT classification strategy. The input audio stream is firstly segmented into segments with duration 200 ms and 100 ms overlapping by using a simple rectangular window. Each segment is further divided into the overlapped frames, using Hamming window with length 50 ms and 25 ms frame shift, in order to avoid spectral distortions. The segmented audio signal is pre-emphasized by a FIR filter in order to emphasize higher frequencies in speech signals. All the features are calculated within each individual frame in time, frequency or cepstral domain. The variance of feature values is then calculated within each individual segments. Such long-term statistics reduce computational complexity and the influence of the signal's variability. Each extracted parameter is then smoothed by averaging the values using successive floating window with length 1 s. The process of smoothing can help to alleviate the influence of the abrupt changes between several adjacent coefficients within the feature vector that represents only one audio class and, as a consequence of that, reduces the miss-classification error.

Each block of classification is represented by one binary discriminator that performs discrimination of input feature vectors using corresponding decision function. Easy to separate and most general classes like speech and non-speech are classified on the first level of topology. The other classes, namely music/environment sound and pure speech/non-pure speech, are classified in the next step, processing the audio data from previous level. The last level performs classification of the two most difficult to discriminate classes: speech with environment sound/speech with music. The output value of decision function assigns the final class label for the actual vector.

Section 2. provides a description about binary decision algorithm and main discrimination principles applied in BDASVM. Section 3. discusses experimental setup and finally Section 4. gives our conclusions and shows future directions.

2. Binary Discrimination Architecture

The core element of proposed BDA is the SVM classifier [9]. We decided to use it in our classification topology also for its generalization ability and superior performance in various pattern classification tasks.

A decision function of the SVM is modeled by separating hyperplane:

$$d(\vec{x}_m, \vec{w}, b) = \langle \vec{w} \cdot \vec{x}_m \rangle + b = \sum_{i=1}^{N} w_i x_i + b, \qquad (1)$$

where the input audio data are represented in the form of feature matrix $\mathbf{X} : \{\vec{x}_1, \ldots, \vec{x}_M\}$ with dimension $M \times N$, where $\vec{x}_m = (x_1, \ldots, x_N)$, $m \in [1 : M]$ are input vectors with dimension N and class label $y_m = \pm 1$. N defines number of coefficients per frame (or segment) and M gives the overall number of frames (segments). After successful training stage the learning machine produces the output $D(\vec{x}_m)$, given as:

$$D(\vec{x}_m) = sign(d(\vec{x}_m, \vec{w}, b)), \qquad (2)$$

using weights $\vec{w}$ and bias b, obtained from the process of learning (training).

Algorithm 1 BDSVM algorithm

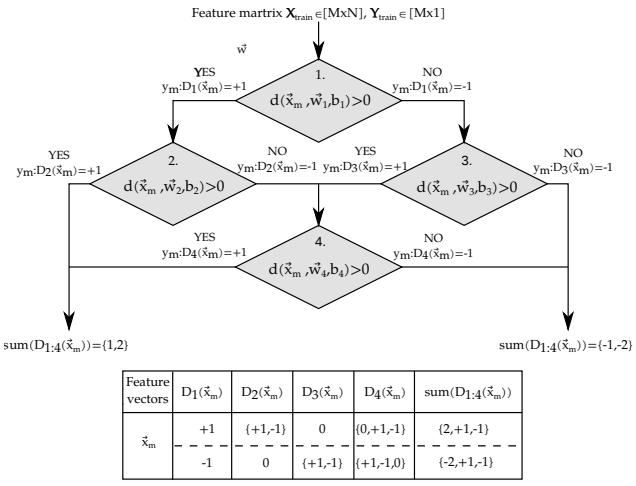
Require: Input feature matrices $\mathbf{X}_{train}$, $\mathbf{X}_{test}$ and label vector $\mathbf{Y}_{train}$.
Ensure: Output label vector $\mathbf{Y}_{predict}$.

for all $\vec{x} \in \{\mathbf{X}_{train}, \mathbf{X}_{test}\}$ **do**
  scale($\vec{x}$);
end for
for (m=0, m<length($\mathbf{X}_{train}$), m++) **do**
  $y_m = +1$, $y_{m+1} = -1$;
end for
for all $(\vec{x}, \vec{y}) \in \{\mathbf{X}_{train}, \mathbf{Y}_{train}\}$ **do**
  cv($\vec{x}$, $\vec{y}$, $\log_2 C$ [0 6 2], $\log_2 g$ [0 6 2], v 5);
  svm_train($\vec{x}$, $\vec{y}$, $best\_C$, $best\_g$);
end for
for all $\vec{x} \in \{\mathbf{X}_{test}\}$ **do**
  svm_predict($\vec{x}$, model);
end for

Proposed solution for binary decision (BD) algorithm utilizing SVM is stated in Alg. 1. It follows

the basic principles applied in the training process. Thus, the main task is to find optimal parameters for binary decision function on each level of classification, using training data. The optimal setting is understood as the process of finding parameters for kernel function and the penalty parameter in the process of training. The input of the BDSVM algorithm is represented by a feature matrix corresponding to training data $\mathbf{X}_{train}$, with dimension $M \times N$. Scaling values in the range of 0 – 1 ensures function scale. It helps to eliminate big differences between coefficients and can be considered as some kind of smoothing. Each feature vector then takes the values $y_m = +1$ and $y_{m+1} = -1$ keeping the same number of vectors for both classes. Reordered feature vectors and labels are assigned as $\mathbf{X}'_{train}$, $\mathbf{Y}'_{train}$. This step helped us to optimize process of cross-validation and alleviate overfitting of classifier. Cross-validation technique, also known as leave-one-out cross-validation) [10], is then applied in order to find optimal parameters of kernel function (g) and penalty parameter (C). After several initial experiments, we decided to use 5-fold cross-validation and RBF kernel function as the best choice. The parameters C and g were adjusted exponentially, taken the values 2^0, 2^2, 2^4 and 2^6.

AUC (Area Under the Curve) [11] parameter was used as the main evaluation criterion during the cross-validation. The highest value of AUC signifies the most optimal (best) parameters C and γ. The optimal parameters are then used to generate *model* by using *svm_train* function. Acoustic model is consequently applied in the process of prediction the class labels for testing data ($\mathbf{Y}_{predict}$).



Feature vectors	$D_1(\vec{x}_m)$	$D_2(\vec{x}_m)$	$D_3(\vec{x}_m)$	$D_4(\vec{x}_m)$	$sum(D_{1:4}(\vec{x}_m))$
$\vec{x}_m$	+1	{+1,-1}	0	{0,+1,-1}	{2,+1,-1}
	-1	0	{+1,-1}	{+1,-1,0}	{-2,+1,-1}

Fig. 2: Binary discrimination architecture.

The proposed BDASVM topology is depicted in Fig. 2. After several initial experiments, we found out that the number of input vectors on the fourth level was insufficient (very small) for training the SVM classifier. This unwanted effect was caused by a high discrimination power on the first level of classification.

There was a need to decrease discrimination ability on the first level and increase on the fourth level in order to suppress this effect. Partial solution was to divide training set at the input of BDASVM into two parts $\mathbf{X}_{train\_train}$ and $\mathbf{X}_{train\_test}$ with equal size. Feature vector matrix $\mathbf{X}_{train\_train}$ was used for training the SVM and $\mathbf{X}_{train\_test}$ for testing on the first level of discrimination. The second level of discrimination enter the feature vectors classified to the class $+1$: $\mathbf{X}_{train\_test_{+1}}$. On the contrary, the third level enter the feature vectors classified to the class -1 on the first level of classification: $\mathbf{X}_{train\_test_{-1}}$. The whole training set of feature vectors $\mathbf{X}_{train}$ was used for training the SVM on the fourth level, regardless the testing vectors which enter the classifier from level two and three. The whole set of testing data $\mathbf{X}_{test}$ enters the BDASVM in testing phase and the values of decision functions are written to the decision table.

The maximum classification accuracy was achieved by adding weighted factor w_C to each discrimination level. We defined it as the value of penalty parameter C divided by number of training vectors belonging to particular class. Thus, for class $+1$: $w_{C_+} = C/num_{+1}$ and for class -1: $w_{C_-} = C/num_{-1}$. In a certain way, weighted factor helps to decrease the influence of overfitting by adding higher weight to the vectors belonging to the minor class and lower weight to the vectors that belong to the major class. More detailed description about the implementation into the SVM training algorithm can be found in [12].

3. Experiments

The classification performance of basic BDSVM (BDTSVM) and proposed BDASVM topology was evaluated on KEMT-BN1 database with total duration about 65 hours of the Slovak TV broadcast audio stream [13]. Only part of the database was considered in our experiments, namely 49 min for training (PS: 10.19 min, MS: 9.26 min, SES: 9.41 min, M: 11.7 min, B: 9.06 min) and 46.2 min for testing (PS: 9.16 min, MS: 9.44 min, SES: 9.25 min, M: 9.04 min, B: 9.31 min). Our aim was to extract maybe smaller amount of audio data, but more accurate, with equal size for each audio class in order to avoid the overfitting of classifier. Silent parts and all other audio events were extracted manually using only available word level transcription.

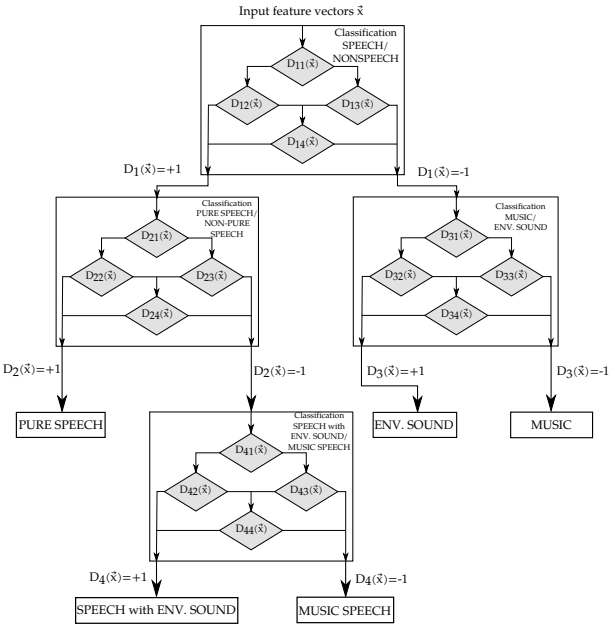
Specifically, the following parameters have been extracted in the process of feature extraction [1], [14] and [15]: Mel-frequency Cepstral Coefficients 13 (MFCCs), Variance Mean of Filter Bank Energy 1 (VMFBE),

Tab. 1: The classification performance of basic BDSVM and proposed BDASVM architectures.

Topology	PS	MS	SES	M	B	Avg	PT
			Acc [%]			Acc [%]	[min]
BDSVM	85.69	54.46	48.63	72.75	77.83	**67.87**	44.13
BDASVM	85.94	53.29	48.94	72.85	80.74	**68.35**	48.37

Variance of Acceleration MFCCs 1 (VAMFCCs), Band Periodicity 1 (BP), Spectral Flux 1 (SF), Spectral Centroid 1 (SC), Spectral Spread 1 (SS) and Spectral roll-off 1 (ROLLOFF). Each coefficient was firstly extracted on frame level and the variance of 7 coefficients was then computed within each segment.

The last step in our experimental work was to implement proposed BDASVM architecture into the overall classification system (Fig. 1). The architecture is depicted in Fig. 3. Each block of classification is represented by one BDA module for discrimination S-NS on the first level, PS-NPS on the second level, M-B on the third level and finally MS-SES on the fourth level. Input audio data are represented by feature vectors $\vec{x}$ with dimension $1 \times N$, as a part of the feature input matrix $\mathbf{X}$. The output gives the information about the particular audio class in audio stream. During the testing phase, depending on the output values of decision functions $D_i(\vec{x}), i \in [1:4]$, label $+1$ or -1 is assigned to each input feature vector on each level of discrimination and saved into the decision table. Final class label is assigned according to the following criteria:



Input vectors	$D_1(\vec{x})$	$D_2(\vec{x})$	$D_3(\vec{x})$	$D_4(\vec{x})$	sum($D_{1:4}(\vec{x})$)
$\vec{x}$	+1	{+1,-1}	0	{+1,-1}	{2,+1,-1}
	-1	0	{+1,-1}	0	{0,-2}

Fig. 3: Implementation of BDA into classification architecture.

- PS: $sum(D_1(\vec{x}),\ D_2(\vec{x})) = 2$, if $D_1(\vec{x}) = 1$,
- M: $sum(D_1(\vec{x}),\ D_3(\vec{x})) = -2$, if $D_1(\vec{x}) = -1$,
- B: $sum(D_1(\vec{x}),\ D_3(\vec{x})) = 0$, if $D_1(\vec{x}) = -1$,
- MS: $sum(D_1(\vec{x}),\ D_2(\vec{x}),\ D_4(\vec{x})) = -1$, if $D_1(\vec{x}) = 1$ and also $D_2(\vec{x}) = -1$,
- SES: $sum(D_1(\vec{x}),\ D_2(\vec{x}),\ D_4(\vec{x})) = 1$, if $D_1(\vec{x}) = 1$ and also $D_2(\vec{x}) = -1$.

4. Results and Discussion

The classification performance for both types of evaluated architectures is given in Tab. 1. We used the classification accuracy *Acc* as the main evaluation criterion. It defines the number of correctly predicted frames to all tested frames within the particular audio class. Each individual value of *Acc*, except for *Avg Acc*, represents the average for parameterization on frame level, segmentation level with smoothing and without smoothing. Value of *Avg Acc* was obtained by averaging the overall classification accuracy for each class. Such interpretation of results helps to minimize redundant information about the parameterization technique and keeps the substantial information about system's performance. PT corresponds with processing time needed for classifying each testing feature vector into the particular class. The implementation of proposed BDASVM topology resulted in higher classification performance. The 0.48 % increase of *Avg Acc* and 4.24 s growth of PT was achieved in comparison with BDSVM. Relatively small increase in processing time was caused by loading all the testing data at once in the first step of classification and processing them on each level of discrimination. We assume that the main cause for relatively low enhancement in classification performance was a high influence of level-dependent miss-classification error in case of MS-BS discrimination problem. Level-dependent error propagates only within one BDASVM component. A possible cure for these drawbacks is the implementation of feature selection algorithm for each level of BDASVM. The aim of that algorithm is to reduce the number of training data by selecting the optimal data set on each level of discrimination. The comparison with other classification architectures, like one-against-one and one-against-all, willbe investigated in the future work as well.

The process of training and testing of the SVM was performed by LIBSVM software (http://www.csie.ntu.edu.tw/~cjlin/libsvm/). Features were extracted using our own sw implementation. The classification algorithms were running on HPC system with 24 nodes. Each one contains computing server IBM Blade System x HS22 with two six-core processor units Intel Xeon L5640 (2.27 GHz) and 48 GB RAM.

Acknowledgment

This publication is the result of the Project implementation: University Science Park TECHNICOM for Innovation Applications Supported by Knowledge Technology, ITMS: 26220220182, supported by the Research & Development Operational Programme funded by the ERDF (100 %).

References

[1] XIE, L., Z. H. FU, W. FENG and Y. LUO. Pitch-density-based features and an SVM binary tree approach for multi-class audio classification in broadcast news. *Multimedia Systems.* 2011, vol. 17, iss. 2, pp. 101–112. ISSN 1432-1882. DOI: 10.1007/s00530-010-0205-x.

[2] LAVNER, Y. and D. RUINSKIY. A Decision-Tree-Based Algorithm for Speech/Music Classification and Segmentation. *EURASIP Journal on Audio, Speech, and Music Processing.* 2009, vol. 2009, no. 2, pp. 405–411. ISSN 1687-4722. DOI: 10.1155/2009/239892.

[3] SONG, Y. and W. H. WANG and F. J. GUO. Feature extraction and classification for audio information in news video. In: *International Conference on Wavelet Analysis and Pattern Recognition.* Baoding: IEEE, 2009, pp. 43–46. ISBN 978-1-4244-3729-0. DOI: 10.1109/ICWAPR.2009.5207452.

[4] CHEN, L., S. GUNDUZ and M. T. OZSU. Mixed type audio classification with support vector machine. In: *International Conference on Multimedia and Expo.* Toronto: IEEE, 2006, pp. 781–784. ISBN 1-4244-0367-7. DOI: 10.1109/ICME.2006.262954.

[5] THEODOROU, T., I. MPORAS and N. FAKOTAKIS. Automatic Sound Classification of Radio Broadcast News. *International Journal of Signal Processing, Image Processing and Pattern Recognition.* 2012, vol. 5, no. 1, pp. 37–47. ISSN 2005-4254.

[6] LU, L., H. J. ZHANG and S. Z. LI. Content-based audio classification and segmentation by using support vector machines. *Multimedia Systems.* 2003, vol. 8, iss. 6, pp. 482–492. ISSN 1432-1882. DOI: 10.1007/s00530-002-0065-0.

[7] HAHALE, P. M. B., M. RASHIDI, K. FAEZ and A. SAYADIYAN. A New SVM-based Mix Audio Classification. In: *40th Southeastern Symposium on System Theory.* New Orleans: IEEE, 2008, pp. 198–202. ISBN 978-1-4244-1806-0. DOI: 10.1109/SSST.2008.4480219.

[8] DHANALAKSHMI, P., S. PALANIVEL and V. RAMALINGAM. Classification of audio signals using SVM and RBFNN. *Expert Systems with Applications.* 2009, vol. 36, iss. 3, pp. 6069–6075. ISSN 0957-4174. DOI: doi:10.1016/j.eswa.2008.06.126.

[9] ABE, S. *Support vector machines for pattern classification.* 2nd ed. New York: Springer, 2005. ISBN 978-1-84996-098-4. DOI: 10.1007/978-1-84996-098-4.

[10] HSU, C. W., C. C. CHANG and C. J. LIN. A practical guide to support vector classification. *CSIE* [online]. 2003. Available at: `http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf`.

[11] FAWCETT, T. An introduction to ROC analysis. *Pattern Recognition Letters.* 2006, vol. 27, iss. 8, pp. 861–874. ISSN 0167-8655. DOI: 10.1016/j.patrec.2005.10.010.

[12] CHANG, C. C. and C. J. LIN. A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology.* 2011, vol. 2, iss. 3, pp. 1–27. ISSN 2157-6904. DOI: 10.1145/1961189.1961199.

[13] PLEVA, M., J. JUHAR and A. CIZMAR. Slovak Broadcast News Speech Corpus for Automatic Speech Recognition. In: *8th International Conference Research in Telecommunication Technology.* Liptovsky Jan: University of Zilina, 2007, pp. 10–12. ISBN 978-80-8070-735-4.

[14] XIONG, Z., R. RADHAKRISHNAN, A. DIVAKARAN and T. S. HUANG. Comparing MFCC and MPEG-7 audio features for feature extraction, maximum likelihood HMM and entropic prior HMM for sports audio classification. In: *International Conference on Multimedia and Expo.* Baltimore: IEEE, 2003, pp. 397–400. ISBN 0-7803-7965-9. DOI: 10.1109/ICME.2003.1221332.

[15] KOS, M., M. GRASIC and Z. KACIC. Online Speech/Music Segmentation Based on the Variance Mean of Filter Bank Energy. *EURASIP*

Journal on Advances in Signal Processing. 2009, vol. 2009, no. 50, pp. 1–13. ISSN 1110-8657. DOI: 10.1155/2009/628570.

About Authors

Jozef VAVREK was born in Kosice, Slovakia in 1985. In 2010 he graduated M.Sc. at the Department of Electronics and Multimedia Communications of the Faculty of Electrical Engineering and Informatics at the Technical University of Kosice. Four years later, he received Ph.D. degree in Telecommunications. His research is oriented on audio data classification, retrieving and digital speech and audio processing.

Jozef JUHAR was born in Poproc, Slovakia in 1956. He graduated from the Technical University of Kosice in 1980. He received Ph.D. degree in Radioelectronics from Technical University of Kosice in 1991, where he works as a full professor and head of the Department of Electronics and Multimedia Communications. He is author and co-author of more than 200 scientific papers. His research interests include digital speech and audio processing, speech/speaker identification, speech synthesis, development in spoken dialogue and speech recognition systems in telecommunication networks.