Regularized Logistic Regression Fusion for Speaker Verification

Ville Hautamäki[1], *Kong Aik Lee*[1], *Tomi Kinnunen*[2], *Bin Ma*[1], *and Haizhou Li*[1]

[1]Human Language Technology Department, Institute for Infocomm Research, Singapore.
[2]School of Computing, University of Eastern Finland, Finland.
{vmhautamaki,kalee,mabin,hli}i2r.a-star.edu.sg tkinnu@cs.joensuu.fi

Abstract

Fusion of the base classifiers is seen as the way to achieve state-of-the art performance in the speaker verfication systems. Standard approach is to pose the fusion problem as the linear binary classification task. Most successful loss function in speaker verification fusion has been the weighted logistic regression popularized by the FoCal toolkit. However, it is known that optimizing logistic regression can overfit severely without appropriate regularization. In addition, subset classifier selection can be achieved by using an external 0/1 loss function on the best subset. In this work, we propose to use LASSO based regularization on the FoCal cost function to achive improved performance and classifier subset selection method integrated into one optimization task. Proposed method is able to achieve 51% relative improvement in Actual DCF over the FoCal baseline.

Index Terms: logistic regression, regularization, compressed sensing, linear fusion, speaker verification

1. Introduction

Speaker verification is the task of accepting or rejecting an identity claim based on a person's voice sample [1]. Classification can be done on either *base classifier* level or at the level of *ensemble*, which is then called *classifier fusion*. In fusion, binary classifier is trained on the base classifier scores to make the accept or reject decision. The base classifiers might utilize, for instance, different speech parameterizations (e.g. spectral, prosodic or high-level features), classifiers (e.g. Gaussian mixture models [2] or support vector machines [3]) or channel compensation techniques (e.g. joint factor analysis [4] or nuisance attribute projection [5]).

In this paper, we consider linear classifier as a fusion device for the base classifer scores. Loss function used for optimizing linear classifier parameters, i.e. the weight vector $\boldsymbol{w}$ and the bias b, play an important role as to how well learned classifier generalizes to unseen data [6]. It is well known that 0/1-loss, where classification error is directly optimized, can lead to a serious overfit. In addition, finding the global optimum of 0/1-loss is an NP-complete computational problem [7]. The *hinge loss*, also known as *maximum margin*, and *logistic regression* have been proposed for tackling these deficiencies, by optimizing an upper bound of the 0/1-loss instead of the classification error itself.

Logistic regression loss defines an unconstrained convex programming problem, implying that the global optimum can be found easily by iterative schemes [6]. In addition, logistic regression loss has similar generalization properties as the maximum margin in the SVM. Logistic regression has been applied to the speaker verification score fusion task [8]. It was later popularized by the *fusion and calibration* (FoCal) toolkit [9]. It has subsequently been found to be useful linear fusion training methodology in a number of independent studies (e.g. [9, 10, 11]) and is chosen here as a reference method.

Overfitting on the training data is still possible, even though upper bound is optimized instead of 0/1-loss. To avoid overfit, regularization techniques can be applied. Most common one is the quadratic regularization, also known as the *ridge regression* [12]. Regularization forces parameter shrinkage, where the greater the Lagrange coefficient λ is, smaller the norm $\|\boldsymbol{w}\|$ will be. Smaller norm implies better generalizability. Reason for it is also easy to see, as higher norm means that some classifiers are given a large weight based on the training data. Effectiveness of these classifiers might not be realized on the evaluation data.

When the ensemble has a large number of classifiers, it is expected that some of them will not play any role in a successful ensemble. Thus, it would be beneficial to remove some unpredictable classifiers from the ensemble to reduce the variance of the prediction error [13]. We have recently studied whether FoCal-based fusion could be improved by computing optimal weights and bias for all subsets and selecting the subset that yilds the best performance on the training set [14]. We noticed, in oracle experiments, that classifier selection can significantly improve performance if suitable selection criterion is utilized. Our proposal was to use 0/1-loss as the selection criterion but this turned out not to generalize so well.

In contrast to the ridge regression, other approach is to regularize via the sum of absolute values $\lambda \sum_i |w_i|_1$, which is called *least absolute shrinkage and selection operator* (LASSO) [13]. It shrinks all coefficients, where some are forced to exactly zero. By regularizing weighted logistic regression with LASSO constraint, one can simultaneously optimize fusion weights and perform classifier subset selection. The combination of ridge regression and LASSO, with two separate Lagrange coefficients (λ_1, λ_2) leads to a third regularization technique known as *Elastic-Net* [15], which is believed to be sharp on the zeroing capability and at the same time smoother than the LASSO type of regularization. In addition, with Elastic-Net control of the norm of the weight vector can be more fine-grained than using LASSO. The reason is that Lagrange coefficient for LASSO can be fixed while progressively increasing the Lagrange coefficient for ridge regression, thus decreasing smoothly the norm of the weight vector.

In this work, we propose to use LASSO and Elastic-Net regularization techniques to simultaneously achieve generalizable fusion device and classifier subset selection. By doing so we have proposed a method to train the subset selector by optimizing the weighted logistic regression loss.

The work of T. Kinnunen was supported by Academy of Finland.

2. Classifier Fusion

2.1. Problem Setup

We assume that, during the development phase, one has access to a development set $\mathcal{D} = \{(\mathbf{s}_i, y_i), i = 1, 2, \ldots, N_{\text{dev}}\}$ of base classifier score vectors $\mathbf{s}_i \in \mathbb{R}^L$, with $y_i \in \{+1, -1\}$ indicating whether the corresponding speech sample originates from a target speaker ($y_i = +1$) or from a non-target ($y_i = -1$). Using $\mathcal{D}$, the goal is to find the best parameters $(\mathbf{w}^*, \theta^*)$ of a linear combiner $f_{\mathbf{w},\theta}(\mathbf{s}) = \mathbf{w}^t\mathbf{s} + \theta$ so that a classification error measure is minimized. We adopt the *detection cost function* (DCF) used in the NIST speaker recognition evaluations,

$$C_{\text{det}}(\theta) = C_{\text{miss}}P_{\text{miss}}(\theta)P_{\text{tar}} + C_{\text{fa}}P_{\text{fa}}(\theta)(1 - P_{\text{tar}}), \quad (1)$$

where P_{tar} is the prior probability of a target (true) speaker, C_{miss} is the cost of a miss (false rejection) and C_{fa} is the cost of a false alarm (false acceptance). These application-dependent cost parameters can also be summarized as a single cost parameter, *effective prior*:

$$P = \text{logit}^{-1}(\text{logit}(P_{\text{tar}}) + \log(C_{\text{miss}}/C_{\text{fa}})), \quad (2)$$

where $\text{logit}\, P = \log P - \log(1 - P)$. It is possible to minimize DCF directly (e.g. [16]) or to optimize a surrogate cost such as effective prior weighted logistic regression cost [17].

2.2. Baseline system

As the baseline method, we use FoCal and as the loss function we optimize C_{wlr}. We use iterative gradient descent method to minimize the following effective-prior weighted logistic regression (WLR) objective [17],

$$\begin{aligned} C_{\text{wlr}}(\boldsymbol{w}, \boldsymbol{s}) \quad &= \quad \frac{P}{N_t}\sum_{i=1}^{N_t} \log\left(1 + e^{-\boldsymbol{w}^t\boldsymbol{s}_i - \theta'}\right) \\ &+ \quad \frac{1-P}{N_f}\sum_{j=1}^{N_f} \log\left(1 + e^{\boldsymbol{w}^t\boldsymbol{s}_j + \theta'}\right), \quad (3) \end{aligned}$$

where the two sums go through the N_t target score vectors $\boldsymbol{s}_i$ and the N_f non-target score vectors $\boldsymbol{s}_j$, respectively. We will also do the standard bias encoding, by adding one extra element containing 1 to $\boldsymbol{s}$. Global bias can then be extracted from the corresponding position in the weight vector. Here, P is the effective prior defined in subsection 2.1 and $\theta' = -\,\text{logit}(P)$ is the decision threshold which is determined from the pre-set cost parameters P_{tar}, C_{miss} and C_{fa}.

3. Regularized Logistic Regression

We extend the weighted logistic regression in Eq. (3), by adding a regularization term. It leads to minimizing [6],

$$C_{\text{wlr}}(\boldsymbol{w}, \boldsymbol{s}) \quad \text{s.t.} \quad J(\boldsymbol{w}) \leq t, \quad (4)$$

where $J(\boldsymbol{w})$ is either $\frac{1}{2}\|\boldsymbol{w}\|_2^2$, which is called ridge regression or $\|\boldsymbol{w}\|_1$, which is known as LASSO. The user specified parameter t indicates the intended amount of parameter shrinkage. The Lagrange coefficients will give us, in the case of LASSO, the following expression,

$$C_{\text{wlr}}(\boldsymbol{w}, \boldsymbol{s}) + \lambda\|\boldsymbol{w}\|_1. \quad (5)$$

The larger the value of λ, the more the norm $\|\boldsymbol{w}\|$ will be shrunk [13]. If the optimization is based on Eq. (5), the correspondence between λ and shrinkage threshold t can be found using a binary search the values of λ. In each iteration, we select one λ value and optimize weights using it, output is then the norm of the weights. Final weight vector is the one whose norm is closest to the target t, but does not violate it.

Table 1: Selection of the three datasets used in this study. We focus on the core-condition itv-tel subset with female trials.

Dataset	Usage	Data source	# Trials
Trainset	To train fusion parameters	NIST SRE 2008 itv-tel subset	263 t, 27315 f
Evalset 1	To compare fusion methods, cross-validate regularization params.	NIST SRE 2008 itv-tel subset	283 t, 27195 f
Evalset 2	To validate results	NIST SRE 2010 itv-tel subset	801 t, 30254 f

Elastic-Net, on the other hand, is based on the idea that we can combine both regularizers into one constrainted optimization problem,

$$C_{\text{wlr}}(\boldsymbol{w}, \boldsymbol{s}) + \lambda_1\|\boldsymbol{w}\|_1 + \frac{\lambda_2}{2}\|\boldsymbol{w}\|_2^2. \quad (6)$$

As can be seen, (6) is a generalized variant of both LASSO and ridge regression. That is, one can always find such regression parameter setup where, in terms of performance, Elastic-Net will at least not lose to LASSO or ridge regression. However, whereas LASSO and ridge regression have only one regression parameter, now we need to search 2-d space. In this work we use methodology where LASSO parameter is first fixed and then ridge regression parameter is found using cross validation as with LASSO and ridge regression methods.

Depending on the chosen regularization method, there are different strategies to optimize (4). Since logistic regression using quadratic regularization is differentiable, it can be efficiently optimized using standard packages [6]. Unfortunately, situation is not so simple for LASSO regularization. In [13], a *quadratic programming* (QP) solution was proposed by rewriting the constraints in (4) in a more convenient form. However, more recent techniques are faster in practice, for that reason we apply *projectionL1* algorithm [18] that optimizes the Lagrangian form Eq. (5). We apply the same method to Elastic-Net. Since, a sum of two convex functions is still convex, we can minimize $C_{\text{wlr}}(\boldsymbol{w}, \boldsymbol{s}) + \frac{\lambda_2}{2}\|\boldsymbol{w}\|_2^2$, given $\lambda_1\|\boldsymbol{w}\|_1$ as the constraint.

4. Corpora, Metrics and Base Classifiers

We utilize the two most recent NIST SRE corpora, namely, NIST 2008 and NIST 2010, in our experiments[1]. The usage of each corpus is shown in Table 1. To avoid any evaluation bias from pooling of incompatible subcondition scores (see [19]), we focus mostly on the female trials[2] of the interview-telephone (itv-tel) sub-condition in the core task. The NIST 2008 trial list was split into two disjoint parts without speaker overlap. The first part, *trainset*, is used for training the score warping parameters (S-cal was used as precalibration method) and fusion weights. The second part, *evalset 1*, as well as *evalset 2* based on the NIST 2010 data, serve for evaluation purposes.

[1] `http://www.itl.nist.gov/iad/mig//tests/sre/`

[2] Female trials are somewhat more difficult than males. Similar rationale was taken, for instance, in [4].

For evaluation of the methods, we consider the detection cost function in (1), where the cost parameters are adopted from the NIST 2010 SRE evaluation plan, namely, $C_{\mathrm{miss}} = C_{\mathrm{fa}} = 1$ and $P_{\mathrm{tar}} = 0.001$. Decision is based on the threshold obtained from effective prior in Eq. (2). Our primary evaluation metric is the actual DCF (ActDCF).

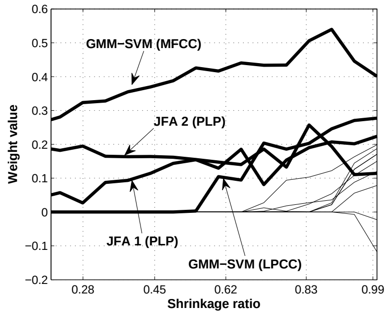

Figure 1: Weight evolution of the LASSO regularization as a function of normalized t.

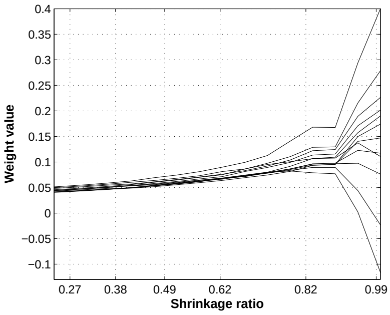

Figure 2: Weight evolution of the ridge regression regularization as a function of normalized t.

In this study we use the same ensemble setup as in our previous work [14]. We have twelve subsystems in total, all are based on different cepstral features and four different classifiers, as part of the of the I4U system. When subsystems share the same classifier and features, it means that the systems are independent implementations. For classifiers, we use the generative GMM-UBM-JFA [4] and the discriminative GMM-SVM approaches with KL-divergence kernel [20] and the recently proposed Bhattacharyya kernel [21]. We also include another recent method, feature transformation [22], as an alternative supervector for SVM. All of the methods are grounded on the *universal background model* (UBM) paradigm and they share similar form of subspace channel compensation, though the training methods differ. We used data from the NIST SRE 2004, SRE 2005 and SRE 2006 corpora to train the UBM and the session variability subspaces, and additional data from the Switchboard corpus to train the speaker-variability subspace for the JFA systems. Each base classifier has its own score normalization prior to score warping and fusion. To this end, we use T-norm and Z-norm with NIST SRE 2004 and SRE 2005 data as the background and cohort training data.

5. Experiments

It is instructive to show the evolution of the individual classifier weights as the function of threshold parameter t. In Figs. 1 and 2 we observe the fusion weights as a function of normalized shrinkage threshold $\hat{t} = t/\|\hat{\boldsymbol{w}}\|$, where $\hat{\boldsymbol{w}}$ is the unregularized solution. We see that $\hat{t}$ will tell how much of the unregularized norm is left after shrinkage. It can be noticed immediately that ridge regression tends to group all classifiers to similar weights as the norm is shrunk towards zero. Grouping effect and the lack of it in the LASSO is known in the general regression literature [15]. Ridge regression tends to group together classifiers that are correlated. LASSO, on the other hand, tends to select few classifers per group. Selection is evident in Fig. 1, as very quickly only four classifiers are left in the ensemble, namely GMM-SVM using MFCC and LPCC front-end and two JFA systems with PLP front-end (base classifiers $\{1, 2, 6, 7\}$). It is notable that even though both JFA base classifers use the same features, they are different implementations (even using different programming languages). The data sets used for learning factor loading matrices are also different.

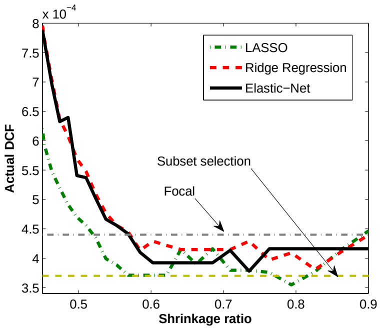

Figure 3: ActualDCF results for the Evalset 1, as a function of normalized t.

In Fig. 3 we see recognition results for the Evalset 1 (subset of the NIST SRE 08). The Elastic-Net needs to optimize two constraints, one for LASSO and the other for ridge regression. We selected the LASSO Lagrange coefficient as the one that gave the best performance on the Evalset 1 for the rest of the experiments. Then Elastic-Net shrinkage percentage was varied as in the other methods. We notice that 20% to 40% shrinkage by LASSO gives the best results. Ridge regression is slightly better than baseline and Elastic-Net can obtain easily the same performance as LASSO.

The subset ensemble [14] outperforms all but Elastic-Net in one threshold location. Subset ensemble contains base classifiers $\{1, 2, 3, 4, 6\}$. However, as we can see in Fig. 4 and Table 2, subset ensemble does not generalize well to NIST SRE 2010 data set. It is interesting to note that using threshold $t \leq 0.6$ for LASSO, we obtain 4 base classifer ensemble that is more generalizable and has 3 classifiers in common with the subset ensemble.

In Fig. 4 we see recognition results for Evalset 2 (NIST SRE 2010). Significant improvement over the baseline can be achieved using any of the regularization methods. Ridge regression and Elastic-Net obtain the best performance.

It is worth to note that improvement of regularization over baseline is more apparent for Evalset 2 than on Evalset 1. Reason for the behaviour is that fusion weights are optimized on the Trainset, which is disjoint subset of the same NIST SRE 2008

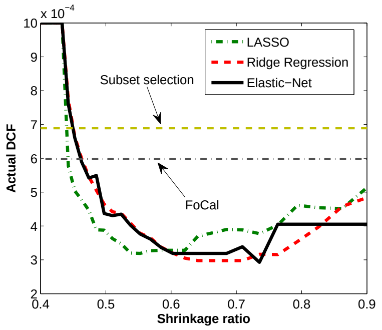

Figure 4: ActualDCF validation results for the Evalset 2, as a function of normalized t.

Table 2: Performance comparison of five fusion methods on Evalset 2.

Fusion	EER (%)	MinDCF (×1000)	ActDCF (×1000)
Subset ensemble	2.62	0.374	0.689
FoCal	2.58	0.306	0.598
Ridge regression	2.24	0.284	0.398
LASSO	2.50	**0.278**	0.461
Elastic-Net	**2.20**	0.281	**0.293**

as Evalset 1. We conlude that regularization does help to tackle the overfitting problem.

In Table 2 summary of validation results are shown for Evalset 2. All fusion device training is done on Trainset, λ parameters are tuned with Evalset 1 and then applied to Evalset 2. We note that all the regularization techniques improve on FoCal baseline. LASSO obtains best minDCF and, using Elastic-net, large improvement is obtained in ActDCF over the baseline. It is also notable that calibration error in Elastic-Net is very small.

6. Conclusion

We have proposed the use of LASSO based regularization to tackle the overfitting problem in optimizing the weighted logistic regression loss function. In addition to performance improvement, the proposed method achieves classifier selection at the same time in one optimization task. Using Elastic-Net regularization we achieve 51% relative improvement in ActDCF, 8% relative improvement in minDCF and slight improvement in EER over the baseline FoCal result. In future, it would be interesting to study automatic selection of the regularization parameters, for example via Bayesian approach.

7. References

[1] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: from features to supervectors,” *Speech Communication*, vol. 52, no. 1, pp. 12–40, January 2010.

[2] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker verification using adapted gaussian mixture models,” *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, January 2000.

[3] W. Campbell, J. Campbell, D. Reynolds, E. Singer, and P. Torres-Carrasquillo, “Support vector machines for speaker and language recognition,” *Computer Speech and Language*, vol. 20, no. 2-3, pp. 210–229, April 2006.

[4] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, “A study of inter-speaker variability in speaker verification,” *IEEE T. Audio, Speech & Lang. Proc.*, vol. 16, no. 5, pp. 980–988, July 2008.

[5] A. Solomonoff, W. Campbell, and I. Boardman, “Advances in channel compensation for SVM speaker recognition,” in *Proc. ICASSP 2005*, Philadelphia, Mar. 2005, pp. 629–632.

[6] C. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer Science+Business Media, LLC, 2006.

[7] S. Ben-David, N. Eiron, and P. Long, “On the difficulty in approximately maximizing agreements,” *Journal of Computer and System Sciences*, vol. 66, no. 3, pp. 496–514, 2003.

[8] S. Pigeon, P. Druytsa, and P. Verlinde, “Applying logistic regression to the fusion of the NIST’99 1-speaker submissions,” *Digital Signal Processing*, vol. 10, no. 1–3, pp. 237–248, January 2000.

[9] N. Brümmer, L. Burget, J. Černocký, O. Glembek, F. Grézl, M. Karafiát, D. Leeuwen, P. Matějka, P. Schwartz, and A. Strasheim, “Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006,” *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2072–2084, September 2007.

[10] L. Ferrer, K. Sönmez, and E. Shriberg, “An anticorrelation kernel for subsystem training in multiple classifier systems,” *J. of Machine Learning Research*, vol. 10, pp. 2079–2114, 2009.

[11] T. Kinnunen, J. Saastamoinen, V. Hautamäki, M. Vinni, and P. Fränti, “Comparative evaluation of maximum *a Posteriori* vector quantization and Gaussian mixture models in speaker verification,” *Pattern Recognition Letters*, vol. 30, no. 4, pp. 341–347, March 2009.

[12] A. Hoerl and R. Kennard, “Ridge regression: Biased estimation for nonorthogonal problems,” *Technometrics*, vol. 12, pp. 55–67, 1970.

[13] R. Tibshirani, “Regression shrinkage and selection via the LASSO,” *Journal of the Royal Statistical Society, Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[14] F. Sedlak, T. Kinnunen, V. Hautamäki, K. Lee, and H. Li, “Classifier subset selection and fusion for speaker verification,” in *ICASSP 2011*.

[15] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *Journal of Royal Statistical Society, Series B*, vol. 67, pp. 301–320, 2005.

[16] W. Campbell, D. Sturim, W. Shen, D. Reynolds, and J. Navratil, “The MIT-LL/IBM 2006 speaker recognition system: High-performance reduced-complexity recognition,” in *Proc. ICASSP 2007*, vol. IV, 2007, pp. 217–220.

[17] N. Brümmer and J. Preez, “Application-independent evaluation of speaker detection,” *Computer Speech and Language*, vol. 20, pp. 230–275, April-July 2006.

[18] M. Schmidt, G. Fung, and R. Rosales, “Fast optimization methods for L1 regularization: A comparative study and two new approaches,” in *ECML 2007*, Warsaw, Poland, September 2007.

[19] D. Leeuwen, “A note on performance metrics for speaker recognition using multiple conditions in an evaluation,” Research note, June 2008, http://sites.google.com/site/sretools/cond-weight.pdf?attredirects=0.

[20] W. Campbell, D. Sturim, and D. Reynolds, “Support vector machines using GMM supervectors for speaker verification,” *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, May 2006.

[21] C. You, K. Lee, and H. Li, “GMM-SVM kernel with a Bhattacharyya-based distance for speaker recognition,” *IEEE Trans. Audio, Speech and Language Processing*, vol. 18, no. 6, pp. 1300–1312, August 2010.

[22] D. Zhu, B. Ma, and H. Li, “Joint MAP adaptation of feature transformation and gaussian mixture model for speaker recognition,” in *Proc. Int. conference on acoustics, speech, and signal processing (ICASSP 2009)*, Taipei, Taiwan, April 2009, pp. 4045–4048.