Comparison of Stacking-based Classifier Ensembles using Euclidean and Riemannian Geometries

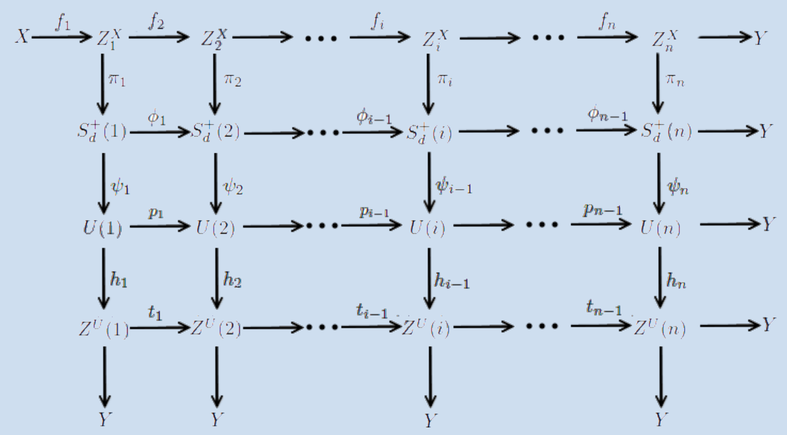
Vitaliy Tayanov, Adam Krzyżak, Ching Y. Suen
Concordia University, Montreal, Canada

Introduction

The objective of this research is to implement a generalized classifier stacking which is a simple classifier stacking and cascades of classifier ensembles in Riemannian geometry. To this end we:

- Built a homotopy diagram which shows all data transformation to be accomplished in linear space and on Riemannian manifolds.
- Developed an algorithm that exploits classifier interactions to build new patterns which are at the same time points on the Riemannian manifold.
- Carried out multiple experiments showing advantaged of application of Riemannian geometry for classifier stacking. The experiments have been done for different data sets, especially for Gesture Recognition dataset.

Homotopy background

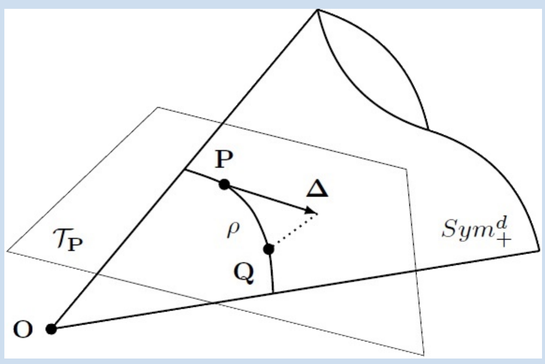


Homotopy diagram of data transformations

Riemannian manifolds

The space of $d \times d$ SPD matrices Sym_+^d is an open convex cone

$$Sym_+^d = \bigcap_{x \in R^d} \{\mathbf{P} \in Sym^d : \mathbf{x}^T \mathbf{P} \mathbf{x} > 0\} \quad (1)$$



A pair of Gauss maps for the log Eucliden metric can be written as follows

$$\exp_{\mathbf{P}}(\mathbf{\Delta}) = \exp(\log(\mathbf{P}) + \Delta) = \mathbf{Q}$$
$$\log_{\mathbf{P}}(\mathbf{Q}) = \log(\mathbf{Q}) - \log(\mathbf{P}) = \Delta \quad (2)$$

Geodesic on Sym_+^d can be computed as

$$d(\mathbf{P}, \mathbf{Q}) = ||\log(\mathbf{Q}) - \log(\mathbf{P})||_F \quad (3)$$

Data sets

Gesture Phase Segmentation data set

data set	size	features	classes	Tr, %	Ts, %
gesture-raw1	1747	18	5	50	50
gesture-raw2	1264	18	5	50	50
gesture-raw3	1834	18	5	50	50
gesture-raw4	1073	18	5	50	50
gesture-raw5	1424	18	5	50	50
gesture-raw6	1111	18	5	50	50
gesture-raw7	1448	18	5	50	50

Data sets of general character from UCI repository

dataset	size	features	classes	Tr, %	Ts, %
balance	625	4	3	50	50
bupa	345	6	2	50	50
gamma	19200	10	2	50	50
german	1000	24	2	50	50
heart	270	13	2	50	50
mfeat-mor	2000	6	10	50	50
mfeat-zer	2000	47	10	50	50
pima	768	8	2	50	50
segment	2310	19	7	50	50
sonar	208	60	2	50	50
spambase	4601	57	2	50	50

Algorithm

Assume that we have L classes and we have a vector of prediction probabilities for each classifier that the given probe belongs to some of L classes. Those predictions are conditional probabilities $p(y = c_\ell|X)$, where c_ℓ is the ℓth class, $\ell = 1, ..., L$. Let us assume that we have T classifiers in ensemble. Then we compose a tensor $\mathbf{T}$ of size $T \times T \times L$, where for each class $C_\ell, \ell = 1, ..., L$ we have a CPPM $\mathbf{A}^\ell(x)$ $T \times T$ with elements $a_{ij}^\ell, \{i, j\} = 1, ..., T$:

$$a_{ij}^\ell(x) = p_i(y = c_\ell|X)p_j(y = c_\ell|X) = h_i^\ell(x)h_j^\ell(x), i \neq j; a_{ij}^\ell = p_i(y = c_\ell|X) = h_i^\ell, i = j. \quad (4)$$

Since a space of SPD matrices is a tangent space the orthonormal coordinates of a tangent vector $\mathbf{y}$ in this space at point $\mathbf{X}$ are given by

$$\mathrm{vec}_{\mathbf{X}}^\ell(\mathbf{y}) = \mathrm{vec}_{\mathbf{I}}^\ell\left(\mathbf{A}^\ell\right), \quad (5)$$

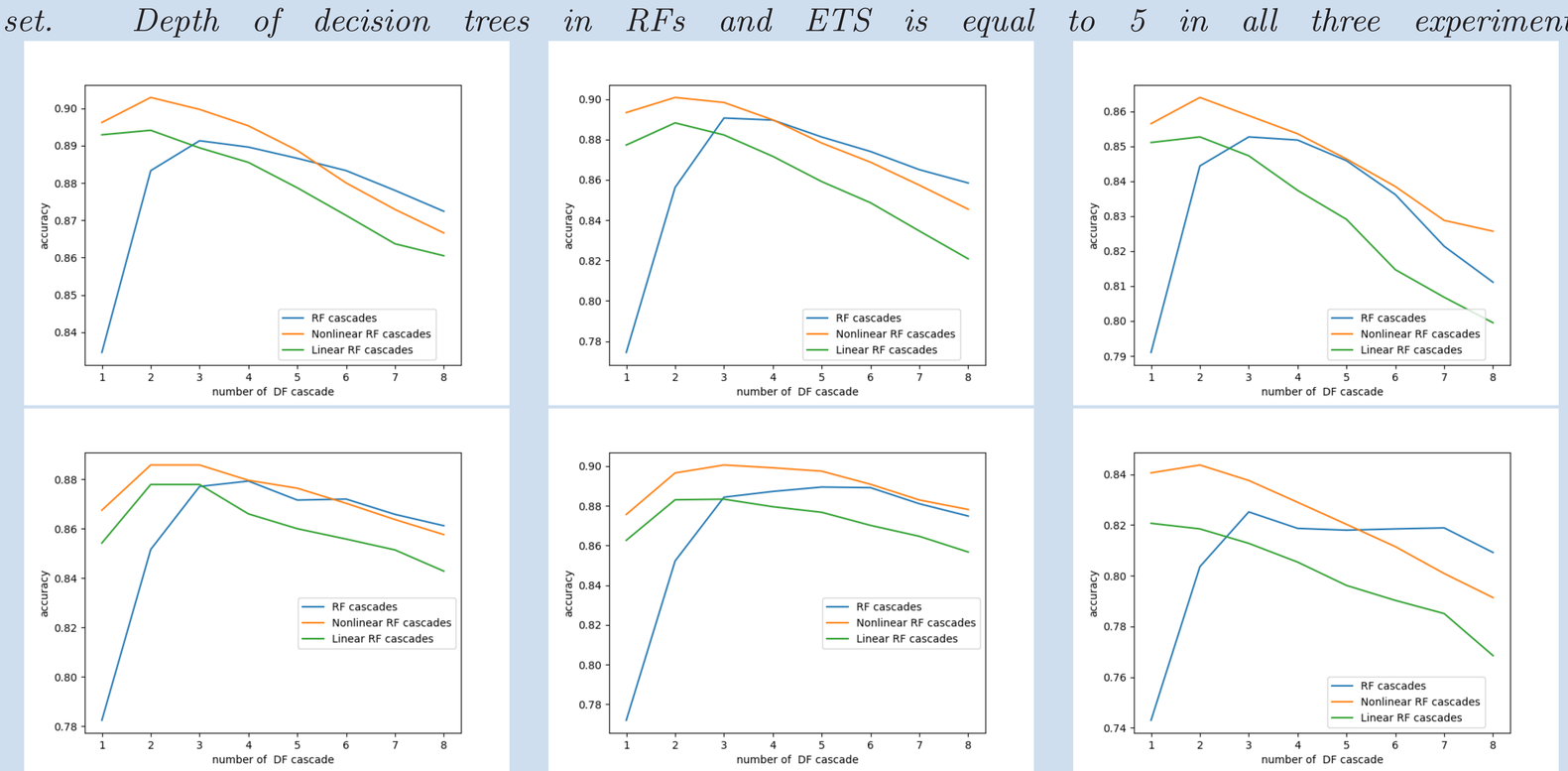
where $\mathrm{vec}_{\mathbf{I}}(\mathbf{y}) = \left[y_{1,1}, \sqrt{2}y_{1,2}, \sqrt{2}y_{1,3}, ...y_{2,2}, \sqrt{2}y_{2,3}, ...y_{d,d}\right]^T$. If want to take into the consideration the geometry of the R-manifold we have to compute the orthonormal coordinates in a flattened space by vectorizing the projection matrix decomposed using Singular Value Decomposition (SVD):

$$\mathrm{vec}_{\mathbf{X}}^\ell(\mathbf{y}) = \mathrm{vec}_{\mathbf{I}}^\ell\left(Proj(\mathbf{A}^\ell)\right) = \mathrm{vec}_{\mathbf{I}}^\ell\left(\mathbf{U}^\ell \log(\mathbf{S}^\ell)(\mathbf{U}^\ell)^T\right). \quad (6)$$

.

Experiments

Classification accuracy as a function of the number of cascades of RFs (first row) and ETs (second row) plotted for three experiments from Gesture Phase Segmentation data set. Depth of decision trees in RFs and ETS is equal to 5 in all three experiments.



Learning classifier predictions using different classifier stacking techniques: means and standard deviations of prediction accuracy (shown in %) for data sets from UCI repository (left table) and Gesture Phase data sets (right table).

Method/Dataset	Bupa	Mfeat-mor	Mfeat-zer	Segment	Sonar
A number of decision trees in a RFs are equal to 100 and its depth is equal to 5					
RF	$71,74 \pm 3,48$	$69,32 \pm 0,84$	$73,94 \pm 0,90$	$92,74 \pm 1,12$	$77,50 \pm 2,48$
RF_{max}	–	–	–	$93,05 \pm 3,84$	$78,94 \pm 3,76$
$RF - nonlinear_{max}$	$70,35 \pm 3,46$	$68,49 \pm 2,15$	$72,54 \pm 1,66$	$94,90 \pm 1,24$	$79,52 \pm 2,98$
$RF - linear_{max}$	$69,71 \pm 2,82$	$67,83 \pm 2,26$	$72,42 \pm 1,12$	$94,37 \pm 1,00$	$79,71 \pm 2,21$
$SVM - stacking$	$70,99 \pm 3,16$	$69,98 \pm 0,76$	$73,88 \pm 0,93$	$93,80 \pm 1,09$	$79,42 \pm 2,16$
$SVM - R - nonlinear$	$70,81 \pm 2,94$	$69,36 \pm 0,77$	$61,86 \pm 3,33$	$92,63 \pm 1,32$	$78,94 \pm 2,60$
$SVM - R - linear$	$69,07 \pm 3,71$	$69,95 \pm 0,73$	$76,39 \pm 0,67$	$94,05 \pm 1,36$	$76,92 \pm 3,44$
$kNN - stacking$	$70,70 \pm 3,72$	$68,58 \pm 0,72$	$74,08 \pm 0,98$	$94,88 \pm 0,88$	$78,37 \pm 2,33$
$kNN - R - nonlinear$	$70,76 \pm 3,94$	$68,22 \pm 0,70$	$69,16 \pm 1,04$	$94,84 \pm 0,95$	$78,37 \pm 1,98$
$kNN - R - linear$	$65,41 \pm 2,62$	$68,74 \pm 0,93$	$73,62 \pm 0,76$	$94,67 \pm 0,84$	$76,54 \pm 4,28$
$MLP - stacking$	$43,37 \pm 8,90$	$11,99 \pm 3,46$	$12,43 \pm 3,02$	$22,96 \pm 6,57$	$52,88 \pm 11,43$
$MLP - R - nonlinear$	$58,26 \pm 9,59$	$11,70 \pm 2,85$	$10,68 \pm 0,99$	$18,50 \pm 5,21$	$52,69 \pm 10,94$
$MLP - R - linear$	$61,69 \pm 8,19$	$17,38 \pm 3,26$	$12,28 \pm 2,32$	$24,65 \pm 4,95$	$50,38 \pm 9,65$
$CNN - R - nonlinear$	$\mathbf{73,78 \pm 3,57}$	$73,00 \pm 0,48$	$77,28 \pm 0,97$	$95,97 \pm 0,49$	$\mathbf{81,44 \pm 2,43}$
$CNN - R - linear$	$73,02 \pm 3,66$	$\mathbf{73,44 \pm 0,68}$	$\mathbf{77,71 \pm 0,75}$	$\mathbf{96,38 \pm 0,62}$	$80,67 \pm 1,90$
A number of decision trees in the ETs are equal to 100 and its depth is equal to 5					
ET	$64,01 \pm 5,08$	$69,50 \pm 0,66$	$74,61 \pm 1,30$	$89,79 \pm 1,14$	$78,85 \pm 3,10$
ET_{max}	$69,24 \pm 3,01$	–	–	$91,66 \pm 1,05$	$79,71 \pm 3,43$
$ET - nonlinear_{max}$	$69,07 \pm 3,73$	$62,65 \pm 4,95$	$72,54 \pm 1,39$	$86,05 \pm 4,16$	$79,33 \pm 4,33$
$ET - linear_{max}$	$68,31 \pm 2,10$	$62,01 \pm 3,46$	$73,64 \pm 0,93$	$90,82 \pm 3,18$	$78,17 \pm 2,82$
$SVM - stacking$	$67,73 \pm 2,84$	$69,71 \pm 0,57$	$72,83 \pm 0,69$	$91,55 \pm 1,21$	$79,90 \pm 3,14$
$SVM - R - nonlinear$	$68,43 \pm 3,53$	$65,40 \pm 1,99$	$38,65 \pm 3,56$	$89,84 \pm 1,32$	$79,52 \pm 2,85$
$SVM - R - linear$	$59,30 \pm 3,04$	$69,88 \pm 0,69$	$78,40 \pm 0,61$	$92,85 \pm 1,15$	$76,35 \pm 3,31$
$kNN - stacking$	$66,86 \pm 3,41$	$67,93 \pm 1,00$	$75,89 \pm 0,83$	$95,46 \pm 0,97$	$79,33 \pm 3,73$
$kNN - R - nonlinear$	$67,03 \pm 2,97$	$67,43 \pm 1,04$	$70,95 \pm 1,80$	$95,28 \pm 0,68$	$79,33 \pm 2,99$
$kNN - R - linear$	$63,31 \pm 1,48$	$67,71 \pm 1,41$	$76,02 \pm 0,98$	$95,30 \pm 1,17$	$76,83 \pm 3,83$
$MLP - stacking$	$40,06 \pm 7,72$	$14,08 \pm 4,75$	$14,46 \pm 4,32$	$19,04 \pm 6,39$	$48,75 \pm 6,79$
$MLP - R - nonlinear$	$59,77 \pm 6,28$	$11,26 \pm 2,20$	$10,31 \pm 0,69$	$20,44 \pm 5,34$	$50,77 \pm 6,84$
$MLP - R - linear$	$58,37 \pm 8,77$	$12,76 \pm 2,49$	$11,62 \pm 2,30$	$16,73 \pm 3,90$	$57,79 \pm 1,40$
$CNN - R - nonlinear$	$\mathbf{71,10 \pm 2,47}$	$73,50 \pm 0,84$	$78,52 \pm 0,89$	$95,42 \pm 0,69$	$81,25 \pm 3,14$
$CNN - R - linear$	$67,97 \pm 2,09$	$\mathbf{73,53 \pm 0,97}$	$\mathbf{79,14 \pm 0,99}$	$\mathbf{95,66 \pm 0,68}$	$\mathbf{81,64 \pm 2,90}$

Method/Dataset	gesture-raw1	gesture-raw2	gesture-raw3	gesture-raw4	gesture-raw5	gesture-raw6	gesture-raw7
A number of decision trees in a RFs is equal to 50 and its depth is equal to 2							
RF_1	$74,45 \pm 1,36$	$69,07 \pm 1,44$	$64,25 \pm 2,01$	$49,94 \pm 3,75$	$54,48 \pm 1,48$	$62,52 \pm 3,48$	$58,49 \pm 2,37$
RF_{max}	$76,31 \pm 1,50$	$73,38 \pm 1,44$	$68,12 \pm 1,39$	$55,25 \pm 5,96$	–	–	–
$RF - nonlinear_{max}$	$77,94 \pm 1,42$	$75,31 \pm 1,41$	$69,65 \pm 1,98$	$61,34 \pm 3,55$	$57,35 \pm 2,25$	$64,62 \pm 2,93$	$60,33 \pm 1,94$
$RF - linear_{max}$	$77,62 \pm 1,75$	$75,01 \pm 2,14$	$69,28 \pm 1,96$	$65,72 \pm 2,73$	$57,51 \pm 2,02$	$63,42 \pm 3,85$	$60,55 \pm 2,57$
$CNN - R - nonlinear$	$85,46 \pm 1,21$	$84,29 \pm 1,73$	$86,93 \pm 0,91$	$80,88 \pm 2,94$	$\mathbf{80,62 \pm 1,78}$	$79,59 \pm 2,16$	$81,81 \pm 1,55$
$CNN - R - linear$	$\mathbf{85,69 \pm 1,23}$	$\mathbf{84,94 \pm 1,53}$	$\mathbf{87,19 \pm 1,02}$	$\mathbf{80,99 \pm 3,25}$	$80,46 \pm 2,00$	$\mathbf{80,54 \pm 2,15}$	$\mathbf{82,17 \pm 1,66}$
$SVM(RBF)$	$74,19 \pm 1,24$	$68,05 \pm 1,46$	$70,12 \pm 1,53$	$46,16 \pm 1,79$	$41,19 \pm 2,15$	$48,15 \pm 1,77$	$49,28 \pm 1,27$
$SVM(RBF) - stacking$	$80,41 \pm 1,49$	$77,91 \pm 2,23$	$81,25 \pm 1,69$	$69,78 \pm 3,07$	$69,44 \pm 3,12$	$73,33 \pm 2,01$	$74,59 \pm 1,76$
Number of decision trees in a RFs is equal to 50 and its depth is equal to 5							
RF_1	$83,46 \pm 0,93$	$85,35 \pm 1,12$	$85,81 \pm 0,71$	$77,17 \pm 2,66$	$77,44 \pm 2,11$	$79,10 \pm 1,98$	$79,16 \pm 1,29$
RF_{max}	$89,13 \pm 1,88$	$89,68 \pm 1,04$	$92,04 \pm 1,12$	$89,70 \pm 0,94$	$89,05 \pm 1,75$	$85,27 \pm 1,06$	$86,20 \pm 1,41$
$RF - nonlinear_{max}$	$90,29 \pm 1,52$	$89,98 \pm 0,86$	$92,16 \pm 1,14$	$89,81 \pm 1,10$	$90,08 \pm 1,45$	$86,40 \pm 1,54$	$87,00 \pm 1,08$
$RF - linear_{max}$	$89,41 \pm 1,22$	$89,29 \pm 0,92$	$91,48 \pm 1,18$	$89,35 \pm 1,25$	$88,82 \pm 1,69$	$85,27 \pm 1,20$	$85,88 \pm 1,23$
$CNN - R - nonlinear$	$89,91 \pm 1,39$	$89,46 \pm 0,83$	$92,34 \pm 1,00$	$89,29 \pm 1,02$	$89,31 \pm 1,27$	$86,29 \pm 1,31$	$86,34 \pm 0,84$
$CNN - R - linear$	$\mathbf{90,67 \pm 1,14}$	$\mathbf{90,33 \pm 1,16}$	$\mathbf{93,04 \pm 1,06}$	$\mathbf{90,73 \pm 1,31}$	$\mathbf{90,45 \pm 1,22}$	$\mathbf{87,27 \pm 1,53}$	$\mathbf{87,87 \pm 1,14}$
$SVM(RBF) - stacking$	$89,15 \pm 1,54$	$89,95 \pm 1,15$	$91,55 \pm 0,83$	$87,09 \pm 1,50$	$89,41 \pm 1,73$	$86,98 \pm 1,60$	$86,71 \pm 1,33$
Number of decision trees in an ETs is equal to 50 and its depth is equal to 2							
RF_1	$74,98 \pm 0,75$	$68,97 \pm 1,00$	$62,99 \pm 0,92$	$45,90 \pm 2,21$	$49,90 \pm 2,95$	$58,62 \pm 3,01$	$55,40 \pm 3,11$
RF_{max}	$76,99 \pm 1,53$	$71,90 \pm 1,51$	$68,21 \pm 2,36$	$50,00 \pm 3,03$	$54,23 \pm 2,51$	$60,81 \pm 2,37$	$59,92 \pm 2,62$
$RF - nonlinear_{max}$	$77,87 \pm 1,21$	$73,96 \pm 1,64$	$70,92 \pm 1,58$	$55,51 \pm 4,56$	$56,76 \pm 1,59$	$63,21 \pm 2,55$	$61,15 \pm 2,46$
$RF - linear_{max}$	$77,84 \pm 0,78$	$73,58 \pm 1,07$	$68,78 \pm 1,48$	$61,15 \pm 3,75$	$55,74 \pm 2,59$	$63,41 \pm 1,86$	$60,69 \pm 2,38$
$CNN - R - nonlinear$	$87,14 \pm 1,66$	$85,92 \pm 1,43$	$87,21 \pm 0,68$	$81,73 \pm 2,34$	$83,89 \pm 2,03$	$80,40 \pm 2,43$	$\mathbf{84,03 \pm 1,51}$
$CNN - R - linear$	$\mathbf{88,07 \pm 1,67}$	$\mathbf{87,15 \pm 1,65}$	$\mathbf{87,75 \pm 0,54}$	$\mathbf{83,48 \pm 2,95}$	$\mathbf{83,89 \pm 1,69}$	$\mathbf{81,55 \pm 1,94}$	$83,49 \pm 1,39$
$SVM(RBF) - stacking$	$81,69 \pm 1,00$	$78,51 \pm 2,20$	$82,88 \pm 1,00$	$71,34 \pm 2,61$	$73,19 \pm 2,74$	$77,28 \pm 2,97$	$79,41 \pm 1,69$
Number of decision trees in an ETs is equal to 50 and its depth is equal to 5							
RF_1	$78,24 \pm 1,40$	$77,88 \pm 1,79$	$77,20 \pm 2,01$	$72,33 \pm 3,65$	$70,62 \pm 1,87$	$74,30 \pm 2,06$	$72,75 \pm 2,38$
RF_{max}	$87,93 \pm 0,95$	$87,93 \pm 0,86$	$88,94 \pm 1,39$	$88,70 \pm 1,36$	$85,52 \pm 1,53$	$82,52 \pm 1,73$	$84,46 \pm 1,83$
$RF - nonlinear_{max}$	$88,58 \pm 1,42$	$88,39 \pm 0,80$	$90,05 \pm 0,81$	$89,01 \pm 1,32$	$86,33 \pm 1,33$	$84,37 \pm 2,10$	$84,85 \pm 1,73$
$RF - linear_{max}$	$87,79 \pm 0,87$	$87,59 \pm 1,26$	$88,33 \pm 1,62$	$87,56 \pm 1,18$	$84,61 \pm 1,47$	$82,07 \pm 1,72$	$83,20 \pm 2,22$
$CNN - R - nonlinear$	$89,28 \pm 0,98$	$89,41 \pm 1,13$	$90,88 \pm 0,66$	$89,39 \pm 2,14$	$89,14 \pm 1,04$	$86,22 \pm 1,79$	$86,93 \pm 1,52$
$CNN - R - linear$	$\mathbf{90,90 \pm 0,86}$	$\mathbf{89,98 \pm 0,93}$	$\mathbf{92,06 \pm 1,05}$	$\mathbf{90,15 \pm 0,92}$	$\mathbf{90,00 \pm 0,98}$	$\mathbf{87,43 \pm 1,47}$	$\mathbf{87,89 \pm 1,35}$
$SVM(RBF) - stacking$	$87,49 \pm 1,41$	$87,63 \pm 0,93$	$88,17 \pm 1,22$	$85,49 \pm 1,85$	$86,18 \pm 1,19$	$85,00 \pm 1,65$	$85,46 \pm 1,60$

Conclusions

Our experiments confirm that application of Riemannian geometry for classifier stacking such as simple stacking or cascades of classifier ensembles is advantageous. Riemannian geometry is especially useful for nonlinear problems such as Gesture Phase data set. Riemannian manifolds allow to use less cascades in case of recursive classifier stacking realized via classifier cascades.