Rank-based Discriminative Feature Learning for Motor Imagery Classification in EEG signals

Byung Hyung Kim, Jin Woo Choi, Sungho Jo
School of Computing, KAIST
Daejeon, Republic of Korea
{bhyung, rayoakmont, shjo}@kaist.ac.kr

2021 9th International Winter Conference on Brain-Computer Interface (BCI) | 978-1-7281-8485-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/BCI51272.2021.9385305

***Abstract*—Existing deep feature learning methods usually compute semantic similarity on an embedding space over the average of the extracted features, relying on delicately selected samples for fast convergence. These deep learned features suffer from inter- and intra-class variations since they are spread across the feature space. In this paper, we present a rank-based feature learning method by exploiting the structured information among features for better separating non-linear data. By exploring Riemannian manifolds' geometric properties, the proposed approach models natural second-order statistics such as covariance and optimizes the dispersion using the distribution of Riemannian distances between a reference sample and neighbors and builds a ranked list according to the similarities. Experiments demonstrate significant improvement over state-of-the-art methods on three widely used EEG datasets in motor imagery task classification. Furthermore, the proposed method jointly enlarges the inter-class distances reduces the intra-class distances for learned features.**

***Index Terms*—BCI, Discriminative, Feature, Riemann, Ranking, z-Score**

I. Introduction

Learning semantic distance on a non-linear embedding space using deep neural networks has been focused with the development of deep metric learning algorithms in various areas [1], [2]. The metric learning is usually measured over the average of the extracted deep features, which tend to be scattered across the feature space where a Euclidean distance metric is used to measure the distance between paired examples. Consequently, the deeply learned features are vulnerable to inter- and intra- class variation problems. Particularly, many scientific fields study data with an underlying structure that is a non-Euclidean space. Hence, directly applying the state-of-the-art Euclidean-based approaches often result in poor or less informative performance. Furthermore, Euclidean distance cannot preserve the correlation and the drawback limits to understand non-stationary data. To overcome this problem, we focus on developing a novel method on non-Euclidean space.

In this paper, we devise a new non-linear rank-based feature learning method on Riemannian manifolds for EEG-based motor imagery tasks [3]–[8] . The proposed approach provides a measure of dispersion using the distribution of Riemannian distances and rejects epochs whose covariance matrices lie out of a region of acceptability defined by a z-score threshold. Our method aims to pull positive points closer than the potato-shaped region of acceptability (z-score) and push negative points out of the boundary. We find that our method on a Riemannian neural network [9] with rank perspective has better performance than current Euclidean-based approaches for learning discriminative non-linear embeddings. Our approach can further jointly reduce the intra-class distances and enlarge the inter-class distances for the learned features, and preserve the correlations of the non-linear EEG features.

II. Background

A. Riemannian Geometry in Metric Learning

A Riemannian manifold is a differential manifold equipped with a varying inner product smoothly on each tangent space. Riemannian metrics of the manifold are the family of inner products on all tangent spaces. Several metrics have been presented to capture its non-linearity. The two most widely used distance measures as true geodesic distances induced by Riemannian metrics are the log-Euclidean distance [10]

$$\delta_L(\Sigma_1, \Sigma_2) = \|\log(\Sigma_1) - \log(\Sigma_2)\|_F \tag{1}$$

and the affine-invariant distance (AIM) [11].

$$\begin{aligned}\delta_R(\Sigma_1, \Sigma_2) &= \|\log(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2})\|_F \\ &= (\sum_{c=1}^{C} \log^2 \lambda_c)^{1/2},\end{aligned} \tag{2}$$

where λ_c, $c = 1, \ldots, C$ are the eigenvalues of $\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}$.

Most existing methods learn the SPD distance measure in a more discriminative space such as the Euclidean tangent space [12]–[14] and the reproducing kernel Hilbert space (RKHS) [15]. Vemulapalli and Jacobs [12] proposed a Mahalanobis-based metric learning method on the tangent space. Huang *et al.* [13] introduced the LEM learning (LEML) approach to transform the matrix on the tangent space to other tangent spaces. Besides, several metric learning methods [16], [17] combine the discriminative abilities of multiple types of manifold representations into RKHS. Quang *et al.* [15] generalized the LEM between two finite-dimensional SPD matrices to infinite-dimensional covariance matrices in RKHS by Hilbert–Schmidt operators. To overcome the inaccurate approximation of the Euclidean space and preserve the SPD manifold structure, Harandi *et al.* [18] proposed projecting the high-dimensional SPD matrix into a low-dimensional manifold and learning a metric in the new manifold.

B. z-Score on Riemannian Manifolds

For a z-score for each epoch t, the measure of dispersion z_t of the Riemannian distance d_t is computed as follows:

$$z_t = \frac{\log(d_t/\mu_t)}{\log(\sigma_t)}, \tag{3}$$

where $d_t = \delta_R(\Sigma_t, \bar{\Sigma}_{t-1})$ is the Riemannian distance between the current covariance matrix Σ_t and the reference matrix $\bar{\Sigma}_{t-1}$, which is the geometric mean of $i \in [0, I]$ numbers of Σ_i. The reference matrix $\bar{\Sigma}$ can be defined by minimizing the dispersion of Σ on the manifold $\mathcal{M}$ such as the sum of squared distances:

$$\bar{\Sigma} = \underset{\Sigma \in \mathcal{M}}{\arg\min} \sum_{i=1}^{N_I} \delta_R^2(\Sigma_i, \Sigma), \tag{4}$$

where N_I can be the maximum number of iterations. The mean μ and the standard deviation σ are

$$\mu_t = \exp(\frac{1}{t}\sum_{i=1}^{t} \log(d_i)), \tag{5}$$

$$\sigma_t = \exp(\sqrt{\frac{1}{t}\sum_{i=1}^{t} (\log(d_i/\mu_i))^2}. \tag{6}$$

Then, according to an objective statistical criterion based on z-score threshold z_{th}, we set the threshold to 2.5 to define the hull of acceptability, rejecting around 0.6% of data under Gaussian assumption. We should note that Riemannian distances are not normally distributed, but empirically follow a non-negative highly right-skewed distribution. Hence, we modeled them in Eq. (3), (5), and (6) by a log-normal or chi-squared distribution [19].

III. Proposed Approach

Given a input matrix Σ_i and the associated label $c = y_i \in C$, we first define the hull of acceptability by estimating a reference matrix $\bar{\Sigma}$ with positive pairs which are the same classes as the anchor matrix Σ_i. Hence, adding a superscript c to index the C classes in Eq. (3) $\sim$ (6), our method aims to pull positive samples from the same class with c closer than a predefined threshold z_{th} and push negative samples out of the boundary, separating the positive and negative sets by a margin m as follows:

$$\begin{aligned} L(\Sigma_i, \Sigma_j, y_i; f) = (1 - y_{ij})[z_{th} - z_j^c)] \\ + y_{ij}[z_j^c - (z_{th} + m)]_+, \end{aligned} \tag{7}$$

where $y_{ij} = 1$ if $c = y_i = y_j$ and $y_{ij} = 0$ otherwise. m is the margin between the positive and negative boundaries. z_j^c is the z-score of Σ_j for a class c with the AIR distance between two points $\delta_R(\Sigma_j, \bar{\Sigma}_{i-1}^c)$ in Eq. (2).

We rank all sample points according to their similarities to a given query Σ_i. We were motivated by the fact that high retrieval quality does not depend on the actual distances, but rather on the ranking order of the features from similar examples. Our strategy retrieves samples on the class level since instance-based sampling cannot guarantee that each example has at least one neighbor in the same mini-batch. That is, we focus on less trivial samples which have non-zero losses in violation of the pairwise similarity for the retrieval problem. In each class $c \in C$, we denote the sets of non-trivial positive $\hat{\mathcal{P}}_i^c$ and negative $\hat{\mathcal{N}}_i^c$ samples with respect to a query Σ_i as

$$\hat{\mathcal{P}}_i^c = \{\forall \Sigma_j | j \neq i \wedge c = y_i = y_j, z_j^c > z_{th}\}, \tag{8}$$

$$\hat{\mathcal{N}}_i^c = \{\forall \Sigma_j | c = y_i, y_i \neq y_j, z_j^c < z_{th} + m\}. \tag{9}$$

Since a perfect clustering can be achieved if and only if all distance to negative examples are larger than a boundary z_{th}, all samples from the same class are grouped into a hypersphere with z_{th} diameter. To pull all non-trivial positive points in $\hat{\mathcal{P}}$ together and learn a class hypersphere, we minimize:

$$L_P(\Sigma_i, y_i; f) = \frac{1}{|\hat{\mathcal{P}}_i^c|} \sum_{\Sigma_j \in \hat{\mathcal{P}}_i^c} L(\Sigma_i, \Sigma_j, y_i; f). \tag{10}$$

To push the non-trivial negative points in $\hat{\mathcal{N}}$, beyond the boundary $z_{th} + m$, we minimize:

$$L_N(\Sigma_i, y_i; f) = \frac{1}{|\hat{\mathcal{N}}_i^c|} \sum_{\Sigma_j \in \hat{\mathcal{N}}_i^c} L(\Sigma_i, \Sigma_j, y_i; f). \tag{11}$$

We adopt the joint supervision of the two objective functions to enhance the discriminative power of deep features as follows:

$$L_{RP}(\Sigma_i, y_i; f) = L_P(\Sigma_i, y_i; f) + \lambda L_N(\Sigma_i, y_i; f), \tag{12}$$

where λ controls the balance between positive and negative sets. With the joint supervision, not only the inter-class features differences are enlarged, but also the variations of the intra-class feature are reduced.

In order to optimize the proposed method within mini-batches. Each mini-batch is randomly sampled from the whole training classes, emits one value, and the overall objective is the average of the values as follows:

$$\bar{L}_{RP} = \frac{1}{N} \sum L_{RP}(\Sigma_i, y_i; f), \tag{13}$$

where N is the batch size and geometric statistics are updated as

$$\mu_i = \exp((1 - \beta_\delta)\log(\mu_{i-1}) + \beta_\delta \log(d_i)), \tag{14}$$

$$\sigma_i = \exp(\sqrt{(1 - \beta_\delta)(\log(\sigma_{i-1})^2 + \beta_\delta(\log(d_i/\mu_i))^2)}, \tag{15}$$

$$\bar{\Sigma}_i = \bar{\Sigma}_{i-1}^{\frac{1}{2}}(\bar{\Sigma}_{i-1}^{-\frac{1}{2}} \Sigma_i \bar{\Sigma}_{i-1}^{-\frac{1}{2}})^{\beta_\delta} \bar{\Sigma}_{i-1}^{\frac{1}{2}}, \tag{16}$$

where $\beta_\delta \in [0, 1]$ defines the learning rate for adaptation in online implementations. A hyper-parameter N_I in Eq. (4) defines the number of positive covariance matrices used for initializing the region of acceptability to model an accurate estimate for the mean and the distribution of distances to it, which will significantly influence the retrieval performance. Otherwise, the region of acceptability will be inefficient to separate negative examples. For the calibration method to initialize the region of acceptability, we uniformly sample the N_I numbers of samples for each class c and estimate

TABLE I

COMPARISON WITH THE STATE-THE-OF-ART METHODS ON DIFFERENT DATASETS.

Method	EEGBCI		BCICIV	
	mAP	R@1	mAP	R@1
MDM	54.3	54.9	27.4	28.1
FgMDM	56.2	55.7	31.8	33.4
SPDNet-Softmax	66.1	64	42.4	42.3
SPDNet-Contrastive	53.1	54	33.4	35.3
SPDNet-Semihard	46.4	48.2	32.8	34.4
SPDNet-Random	49.6	50.4	30.5	32.8
Ours	**69.3**	**68.2**	**44.2**	**45.5**

the geometric statistics μ, σ, and $\bar{\Sigma}$ in Eq. (4) $\sim$ (6) before training.

IV. EXPERIMENT

We compared our approach against the following state-of-the-art methods as described in the Background section: MDM [20], FgMDM [21], SPDNet [9]. Because SPDNet basically has one softmax layer, we conducted a further evaluation of SPDNet using two popular metric loss functions: contrastive (SPDNet-contrast) and triplet (SPDNet-triplet) loss functions [1]. We report the mean average precision (mAP) and the Recall@K (R@K) on the state-of-the-art methods associated with EEG datasets described below.

A. EEG Datasets in Motor Imagery Tasks

We evaluated the proposed system for motor imagery classification tasks using two EEG datasets.

- BCICIV [22]: The BCI competition IV Database-Dataset IIa (BCICIV) consists of 22-channeled EEG signals at 250 Hz from nine subjects. The subjects were asked to imagine four different motor imagery tasks until a fixation cross disappeared from the screen at 6 s. The dataset consisted of 6 runs of 48 trials (12 for each class) during 2 sessions conducted on different days.
- EEGBCI [23]: The EEG Motor Movement/Imagery Dataset using the BCI2000 system (EEGBCI) contains 64 channeled EEG signals at 160 Hz over 1500 1- and 2-min recordings from 109 participants. During this experiment, we used EEG signals obtained in a two-class (hands vs. foot) motor imagery task (Task 4).

B. Experimental Setup

1) EEG Preprocessing: EEG signals are first band-pass filtered with a bandwidth of $7-30$Hz for the two datasets [24] with electrode-wise exponential moving standardization for normalizing the continuous data. Each EEG signal is represented by a channel $\times$ channel SPD matrix, which is calculated using a second-order statistics.

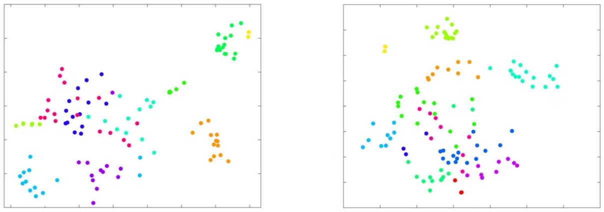

Fig. 1. The t-SNE result of the deep features learned by (a) our method and (b) the SPDNet method on EEGBCI dataset.

2) Training, Validation, and Testing Datasets: We conducted this experiment based on a 5-fold cross validation as follows. We split the full datasets into fifths for testing. From the remaining data (four-fifths of the total data), we used one-fifth of the remaining data for validation and four-fifths for training. Note that the training, validation, and testing data were subject-wise.

3) Network Configuration and Parameter Settings: For a fair comparison, all methods applied a batch size, learning rate, weight decay, and momentum of 16, 10^{-2}, 10^{-3}, and 0.9, respectively, as the training parameters. The initial weights were set to random semi-orthogonal matrices, and the rectification threshold ϵ was set to 10^{-4}. Early-stopping during validation with a fixed patience size was adopted to prevent an overfitting in learning the deep features. We configured a block pair of BiMap and ReEig with LogEig and FC layers for all variants of SPDNet. The sizes of the transformation matrices are set to 22×18 for BCICIV and 64×56 for EEGBCI, respectively. For other parameters of the compared methods, we empirically set the best parameters with the highest accuracy based on the original study.

C. Experimental Results

Table I summarizes the classification results on the two datasets. Overall, the proposed approach outperformed all the compared methods, which validates the effectiveness of ranking-based strategy for learning discriminative features to classify different motor imagery tasks. Among the baselines, the triplet loss always performed the worst. Furthermore, when triplet loss with hard negatives, the bad results imply the loss may waste gradient update on input features far away from the decision boundary. These results support the significance of mining examples. Pairwise or triplet based metric losses require careful sample mining and weighting strategies to obtain the most informative pairs of samples, particularly when considering mini-batches.

Fig. 1 shows the pairwise distances between the centers of each class and matrices within a class (intra-class distances), and with different classes (inter-class distances). This result indicates the discriminative power of our approach. The features learned by our model exhibit more clear discriminative structures, while the other methods present relatively vague structures.

V. Conclusion

In this paper, we proposed a rank-based discriminative feature learning method and showed the efficacy on classifying nonlinear EEG signals in motor imagery tasks. The proposed method achieves state-of-the-art performance, reducing the intra-class distances and enlarging the inter-class distances for learned features. Our next work will further study the nonstationary nature of brain activity as revealed by EEG. Hence, we will investigate the efficacy of the proposed method for discriminating EEG signals under various datasets in EEG-related tasks such as emotion recognition [25]–[30].

Acknowledgment

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2017-0-00432)

References

[1] K. Roth, T. Milbich, S. Sinha, P. Gupta, B. Ommer, and J. P. Cohen, "Revisiting training strategies and generalization performance in deep metric learning," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2020, pp. 8242–8252.

[2] B. H. Kim and S. Jo, "Real-time motion artifact detection and removal for ambulatory bci," in *Proceedings of the International Winter Conference on Brain-Computer Interface*, 2015.

[3] V. Jayaram and A. Barachant, "Moabb: trustworthy algorithm benchmarking for bcis," *Journal of Neural Engineering*, vol. 15, no. 6, p. 066011, 2018.

[4] I. Daly, D. Williams, A. Malik, J. Weaver, A. Kirke, F. Hwang, E. Miranda, and S. J. Nasuto, "Personalised, multi-modal, affective state detection for hybrid brain-computer music interfacing," *IEEE Transactions on Affective Computing*, vol. 11, no. 1, pp. 111–124, 2020.

[5] J. W. Choi, B. H. Kim, S. Huh, and S. Jo, "Observing actions through immersive virtual reality enhances motor imagery training," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 7, pp. 1614–1622, 2020.

[6] J. W. Choi, S. Huh, and S. Jo, "Improving performance in motor imagery bci-based control applications via virtually embodied feedback," *Computers in Biology and Medicine*, vol. 127, p. 104079, 2020.

[7] Y.-J. Suh and B. H. Kim, "Riemannian embedding banks for common spatial patterns with eeg-based spd neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[8] B. H. Kim, Y.-J. Suh, H. Lee, and S. Jo, "Nonlinear ranking loss on riemannian potato embedding," in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, 2021, pp. 4348–4355.

[9] Z. Huang and L. Van Gool, "A riemannian network for spd matrix learning," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

[10] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Log-euclidean metrics for fast and simple calculus on diffusion tensors," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 56, no. 2, pp. 411–421, 2006.

[11] X. Pennec, P. Fillard, and N. Ayache, "A riemannian framework for tensor computing," *International Journal of Computer Vision*, vol. 66, no. 1, pp. 41–66, 2006.

[12] R. Vemulapalli and D. W. Jacobs, "Riemannian metric learning for symmetric positive definite matrices," *arXiv preprint arXiv:1501.02393*, 2015.

[13] Z. Huang, R. Wang, S. Shan, X. Li, and X. Chen, "Log-euclidean metric learning on symmetric positive definite manifold with application to image set classification," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2015, pp. 720–729.

[14] L. Zhou, L. Wang, J. Zhang, Y. Shi, and Y. Gao, "Revisiting metric learning for spd matrix based visual representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (ICCV)*, 2017, pp. 3241–3249.

[15] M. H. Quang, M. San Biagio, and V. Murino, "Log-hilbert-schmidt metric between positive definite operators on hilbert spaces," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 388–396.

[16] Z. Huang, R. Wang, S. Shan, L. Van Gool, and X. Chen, "Cross euclidean-to-riemannian metric learning with application to face recognition from video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2827–2840, 2017.

[17] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen, "Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild," in *Proceedings of the International Conference on Multimodal Interaction (ICMI)*, 2014, pp. 494–501.

[18] M. Harandi, M. Salzmann, and R. Hartley, "Dimensionality reduction on spd manifolds: The emergence of geometry-aware methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 1, pp. 48–62, 2017.

[19] M. Congedo, A. Barachant, and R. Bhatia, "Riemannian geometry for eeg-based brain-computer interfaces; a primer and a review," *Brain-Computer Interfaces*, vol. 4, no. 3, pp. 155–174, 2017.

[20] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass brain–computer interface classification by riemannian geometry," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pp. 920–928, 2011.

[21] ——, "Classification of covariance matrices using a riemannian-based kernel for bci applications," *Neurocomputing*, vol. 112, pp. 172–178, 2013.

[22] M. Tangermann, K.-R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. J. Miller, G. Mueller-Putz *et al.*, "Review of the bci competition iv," *Frontiers in Neuroscience*, vol. 6, p. 55, 2012.

[23] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "Bci2000: a general-purpose brain-computer interface (bci) system," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 1034–1043, 2004.

[24] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[25] M. Moghimi, R. Stone, and P. Rotshtein, "Affective recognition in dynamic and interactive virtual environments," *IEEE Transactions on Affective Computing*, vol. 11, no. 1, pp. 45–62, 2020.

[26] H. Becker, J. Fleureau, P. Guillotel, F. Wendling, I. Merlet, and L. Albera, "Emotion recognition based on high-resolution eeg recordings and reconstructed brain sources," *IEEE Transactions on Affective Computing*, vol. 11, no. 2, pp. 244–257, 2020.

[27] T. Song, W. Zheng, P. Song, and Z. Cui, "Eeg emotion recognition using dynamical graph convolutional neural networks," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 532–541, 2020.

[28] L. Piho and T. Tjahjadi, "A mutual information based adaptive windowing of informative eeg for emotion recognition," *IEEE Transactions on Affective Computing*, vol. 11, no. 4, pp. 722–735, 2020.

[29] H. Cai, X. Zhang, Y. Zhang, Z. Wang, and B. Hu, "A case-based reasoning model for depression based on three-electrode eeg data," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 383–392, 2020.

[30] S. Grissmann, M. Spüler, J. Faller, T. Krumpe, T. O. Zander, A. Kelava, C. Scharinger, and P. Gerjets, "Context sensitivity of eeg-based workload classification under different affective valence," *IEEE Transactions on Affective Computing*, vol. 11, no. 2, pp. 327–334, 2020.