Random Forest-Supervised Manifold Alignment

Jake S. Rhodes[†]
Department of Statistics
Brigham Young University
Provo, Utah, USA
rhodes@stat.byu.edu

Adam G. Rustad[†]
Department of Computer Science
Brigham Young University
Provo, Utah, USA
arusty@byu.edu

Abstract—Manifold alignment is a type of data fusion technique that creates a shared low-dimensional representation of data collected from multiple domains, enabling cross-domain learning and improved performance in downstream tasks. This paper presents an approach to manifold alignment using random forests as a foundation for semi-supervised alignment algorithms, leveraging the model's inherent strengths. We focus on enhancing two recently developed alignment graph-based by integrating class labels through geometry-preserving proximities derived from random forests. These proximities serve as a supervised initialization for constructing cross-domain relationships that maintain local neighborhood structures, thereby facilitating alignment. Our approach addresses a common limitation in manifold alignment, where existing methods often fail to generate embeddings that capture sufficient information for downstream classification. By contrast, we find that alignment models that use random forest proximities or class-label information achieve improved accuracy on downstream classification tasks, outperforming single-domain baselines. Experiments across multiple datasets show that our method typically enhances cross-domain feature integration and predictive performance, suggesting that random forest proximities offer a practical solution for tasks requiring multimodal data alignment.

Index Terms—manifold alignment, manifold learning, random forest proximities, multimodal learning, representation learning

I. Introduction

Data collected from an individual observation can come from multiple sources (multi-modal) which may be inherently connected, such as a document translated into multiple languages [1]. However, challenges arise when attempting to use multiple data domains in a single model, especially if the data lack clear connections, or when there is no originating source from which both domains are sampled [2].

Recent machine learning tools are capable of improving classification and regression problems by incorporating multiple domains in the learning paradigm [3]. Most recent multimodal models are large and computationally expensive deep learning models [4] which are well formulated for tasks such as generative natural-language processing or image generation [5], but are often overkill for smaller, scientific projects. For moderate data sets for scientific tasks, the field of manifold alignment forms a means of generating a shared representation across multiple domains for subsequent tasks.

Manifold alignment aims to find a shared, low-dimensional representation of data sources, enabling the learning of inter-domain relationships while preserving the structure of each domain. Semi-supervised manifold learning involves using known connections, or anchors, to help uncover relationships between domains. In some instances, it is possible to learn richer information by combining insights from multiple sources formed into a single representation, such as a manifold. For example, a classification model can be enhanced by integrating relevant information from multiple data domains. However, we show that many existing manifold alignment methods often yield representations on which predictive models perform poorly compared to single-domain baselines.

In this paper, we overcome this weakness by employing random forests [6] to initialize manifold learning algorithms. We show that using this initialization often improves the accuracy of subsequent supervised models when trained on the resulting joint embedding. We build on recent diffusion- and shortest-path methods [7], named, respectively, MASH (Manifold Alignment via Stochastic Hopping) and SPUD (Shortest Path on the Union of Domains). We show that the embedding quality, as measured by post-alignment classification accuracy, is generally improved over existing, semi-supervised alignment methods.

II. Manifold Alignment Background

Manifold alignment refers to methodologies aimed at discovering a common, low-dimensional structure, or manifold embedding, that forms a common or shared representation of data from multiple sources. The representation can be subsequently used to analyze relationships across different domains or enhance performance in downstream machine-learning tasks like label prediction. In some cases, alignment is aided by known connections or anchors, represented by data points shared across multiple domains [2]. In others, alignment relies on discrete, shared data labels [8, 9]. These examples belong to a branch known as semi-supervised manifold alignment. Here, we assume that knowledge between points of other modalities or domains can be distilled through known anchor points. At a high level, we assume that point nearness in one modality transfers across domains, extending the idea that "the neighbor of my neighbors is also my neighbor" to multiple domains.

This concept was explored in [7]. In the paper, the authors described two graph-based manifold alignment methodologies

that build new graph edges through either diffusion [10] or shortest paths. They showed that both the diffusion method, Manifold Alignment via Schochastic Hopping (MASH), and the shortest-path method, Shortest Paths on the Union of Domains (SPUD) improved alignment performance compared to state-of-the-art methodologies measured using two independent metrics: The fraction of samples closer than the true match (FOSCTTM) [11] and cross-embedding classification (CE) or label transfer [8]. We build upon this method by defining weighted graphs via random forest proximities to initialize the graph structure. In doing so, the aligned embedding more frequently improves classification accuracy over that derived from a single domain. In contrast, we show that most manifold alignment methods typically do not enhance the learnable information of classification tasks.

III. RANDOM FOREST PROXIMITIES

A random forest is a supervised learning model designed to predict data labels using an ensemble of randomized decision trees [6]. Each tree is trained on a bootstrap sample, uniformly selected with replacement, with included points referred to as "in-bag", and excluded points called "out-of-bag" or OOB.

Similarity measures can be derived from the terminal nodes between the in-bag points and a new observation. Since the similarities are determined by the decision space laid out by the terminal nodes, they can be viewed as a form of supervised similarity. Such similarity measures are often called random forest proximities. Random forest proximities define pairwise relationships between data points, leveraging supervised learning to reveal meaningful patterns guided by the supervision of the random forest model. They have proven to be useful in many applications. One of the original uses of random forest proximities is in data visualization, where they facilitate dimension reduction techniques such as multidimensional scaling (MDS) [12]. Random forest proximities have also been used to improve feature importance measures [13] and data imputation [14].

The connection between random forest regression models and a nearest-neighbor approach with an adaptive local metric was established in [15], framing the model's predictions in terms of local instance relationships. Building on this concept, Random Forest-Geometry- and Accuracy-Preserving Proximities (RF-GAP) [16] provided a novel measure that extends this relationship classification contexts and can perfectly replicate random forest out-of-bag predictions. Thus, the RF predictions are described in terms of instance-level attribution; that is, they reflect the influence of adjacent training points. In this way, the model predictions are driven by the importance of local instances in shaping the decision space.

RF-GAP proximities have shown superior performance in proximity-based tasks, including visualization [17], imputation, and outlier detection [16]. This improvement is likely due to their being intentionally designed to generate random forest OOB predictions, which design also prevents overfitting when used in downstream tasks. By encoding the relationships between data points through random forest proximities, we can establish a meaningful connection between domains through neighboring points.

A. Related Methods

Manifold alignment methods for semi-supervised learning focus on aligning two modalities or co-domains, $\mathcal{X}$ and $\mathcal{Y}$, where the partial correspondence between points is known. The correspondence is typically defined through points known to belong to both domains (i.e., x_i and y_j form alternative representations of the same point in $\mathcal{X}$ and $\mathcal{Y}$, respectively), although, in some cases, correspondence is inferred via known class labels [8, 9]. We briefly describe below existing semi-supervised alignment methods.

Semi-Supervised Manifold Alignment (SSMA) [18] uses constrained Laplacian Eigenmaps to align known correspondences between domains in low-dimensional embeddings. An alignment method based on Diffusion Maps was derived in [19], performing density normalization across domains and employing an affine transformation to match embeddings. However, limited implementation details preclude us from comparing this approach. Manifold Alignment via Procrustes Analysis (MAPA) [20] applies Laplacian Eigenmaps followed by Procrustes alignment on correspondences and generally outperforms SSMA.

Joint Laplacian Manifold Alignment (JLMA) builds on k-NN similarities to create a joint Laplacian across domains, embedding known correspondences in a shared coordinate system [2]. MAGAN [21], a Generative Adversarial Network-based approach, aligns domains using reconstruction, discriminator, and custom correspondence losses. Diffusion Transport Alignment (DTA) [22] combines diffusion processes with optimal transport to align domains via cross-domain transition matrices, outperforming over MAGAN, MAPA, and SSMA.

Kernel Manifold Alignment (KEMA) [9] and Manifold Alignment with Label Information (MALI) [8] are both semi-supervised methods that perform alignment through common data labels rather than known data anchors. KEMA is a kernelized version of SSMA, while MALI uses entropic optimal transport initialized with prior class probabilities, and has been shown to outperform KEMA in label transfer tasks [8].

IV. METHODS

Both MASH and SPUD [7] are typically initialized with a kernel matrix for each domain using the α-decaying kernel [23]. Instead, we initialize each model using RF-GAP proximities [16]. The RF-GAP proximities form a row-stochastic diffusion operator which encodes similarities within a given domain. The supervised information, however, is encoded in a way that does not overfit to the training data, which is important for downstream classification tasks. We refer the reader to [16] for more on this aspect. To learn cross-domain similarity measures, we assign the similarities between anchor points to be maximal. That is, $cross\_similarity(x_i, y_j) = max(p(x_i, x_k))$ for x_k in $\mathcal{X}$, where $p(.,.)$ is the normalized proximity, and y_j is the representation of x_i in $\mathcal{Y}$. We form a joint, cross-domain similarity matrix as

$$\mathcal{P} = \begin{pmatrix} P_{\mathcal{X}} & P_{\mathcal{X}\mathcal{Y}} \\ P_{\mathcal{Y}\mathcal{X}} & P_{\mathcal{Y}} \end{pmatrix}$$

with $P_{\mathcal{Y}\mathcal{X}} = P_{\mathcal{X}\mathcal{Y}}^T$.

MASH powers the cross-domain similarity, forming new graph connections between the domains. We apply an information distance [24] to the powered diffusion operator, $\mathcal{P}^t$ to perform a potential distance [23] upon which we can apply MDS to form an embedding.

RF-SPUD also starts by forming two RF-GAP similarity graphs, $P_{\mathcal{X}}$, and $P_{\mathcal{Y}}$. Anchor points are then leveraged to form connections between domains. However, instead of following a diffusion process, RF-SPUD learns the shortest paths between the domain graphs by following geodesic curves through anchor points, thereby forming additional cross-domain connections and learning long-range dependencies. A potential distance function [23] is applied to the learned connections and an embedding is formed using MDS. For further details, we refer the reader to [7].

V. EXPERIMENTAL RESULTS

We investigate the potential for manifold alignment models to produce alignments or embeddings capable of enhancing downstream classification. We compare RF-SPUD and RF-MASH to SSMA, MAPA, JLMA, MAGAN, DTA, KEMA, MALI, and MALI initialized using RF-GAP proximities (RF-MALI). We used 16 classification[1] files from the UCI Machine Learning Repository [25]. We removed observations with missing values and normalized all features from 0 - 1.

To simulate different modalities from each dataset, we perform feature-level splits (i.e., treat different sets of features as a different domain), and two types of feature distortions. In the **random** split setup, each domain was assigned a randomly chosen set of features without any overlap. The **skewed** split assigns domains according to feature relevance as assessed by a random forest model; the most relevant features formed one domain, while the remaining formed the other. For the **even** split, features were divided alternating between domains according to importance scores to balance prediction relevance. All remaining less relevant features are assigned randomly.

Two data-distortion domain adaptions were used for comparisons, as presented in [22]. For the first method, **distort**, Gaussian noise was added to the original features to form a second domain. For the second, random rotations via QR factorization were applied to the original features to produce the **rotate** domain adaptions.

Our experiments aim to identify alignment methods that produce embeddings with sufficient information to enhance downstream prediction tasks and to determine the contexts in which certain methods perform best. Specifically, we train classification models (random forests and k-nearest neighbors)

[1]Balance Scale, Breast Cancer, Crx, Diabetes, Ecoli (5 majority classes), Flare1, Glass, Heart Disease, Heart Failure, Hepatitis, Ionosphere, Parkinson's, Seeds, Iris, Tic-Tac-Toe, Cancer Data
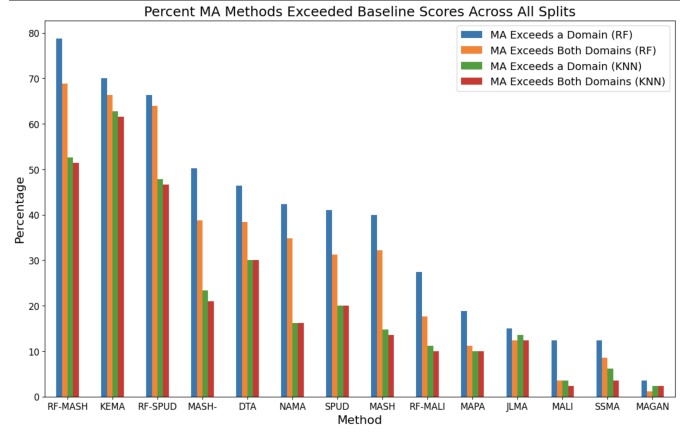

Fig. 1. We trained RF and k-NN models on each split separately across 5 split types, 16 publicly available datasets, and 3 repetitions each. The models were then trained on the aligned embedding from each method. The proportion of times the classification accuracy exceeded one or both individual models was tracked. Overall, RF-MASH performed best at, exceeding both baselines for random forest models. 30% of data points served as anchors.

on each domain separately, record validation scores, and then train these models across the aligned, joint embedding space. Ideally, a joint embedding should capture inter-domain relationships, enabling the models to exceed baseline accuracy.

We used a staggered grid search strategy for model parameter selection, testing each parameter set on a fixed 10-nearest neighbor model across each dataset to ensure comparable performance evaluations. We explored combinations of variables with the greatest impact on model performance and reported the best results from the search for each model.

In Figure 1, we compare baseline classification accuracy with the accuracy of models trained on aligned embeddings. For each dataset and split, we trained a random forest and k-NN model, recorded validation accuracy, and then trained each model using manifold embeddings for each alignment method, fitting on one domain and predicting on the other. The proportion of models that outperformed one or both baselines, Overall, RF-MASH, KEMA, and RF-SPUD. Generally, methods initialized with random forests outperformed their unsupervised counterparts.

While KEMA generally performs well at improving multi-modal classification accuracy over the baselines, we note that this method underperforms at aligning specific data points to their known correspondences in another domain. For comparison, we include models that use label information during the alignment process using the combined metric (CE - FOSCTTM) as introduced in [7] in Figure 3. In this Figure, we see that KEMA is outperformed by RF-SPUD and RF-MASH across all domain adaptations except for random rotations. For all other splits, RF-SPUD and RF-MASH are top performers.

VI. CONCLUSION

In this paper, we introduced two semi-supervised alignment methods (RF-SPUD and RF-MASH) initialized using RF-GAP proximities. Although most semi-supervised manifold
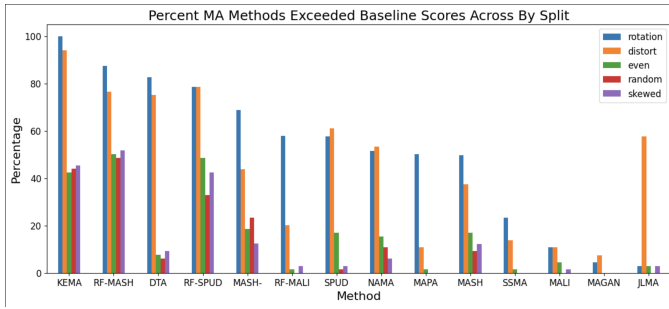
Fig. 2. The same datasets and configurations from Figure 1 were compared here. In this figure, we distinguish between the different split types used to simulate multiple domains. The RF-initialized methods tend to have more well-rounded results across all split types, though each method fails to exceed 50% of the baselines for the random and even splits. KEMA performed best at the rotation split but underperformed RF-MASH at all feature-level splits. DTA does well at each distort and rotate, but not well at feature-level splits.
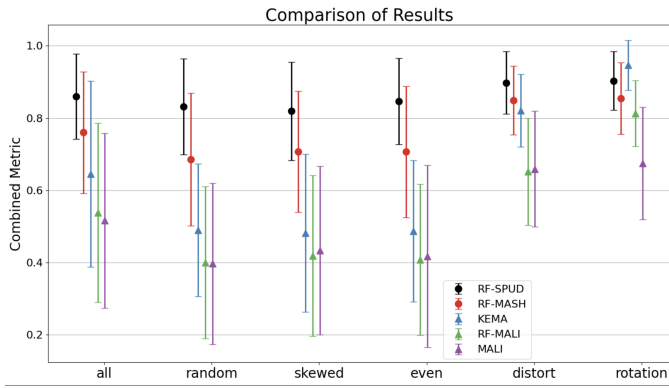

Fig. 3. Here we compare all methods that use label information in the alignment process. The combined metric (CE - FOSCTTM) [7], is computed and averaged across all datasets. Generally, RF-SPUD and RF-MASH have the best performance with the exception of KEMA on the rotation split.

alignment methods generally do not improve classification models trained on the resulting embeddings, we showed that proximity-initialized models usually enhance aligned manifold embeddings by retaining information sufficient for improving multi-modal classification tasks.

REFERENCES

[1] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proceedings of Machine Translation Summit X: Papers*, Phuket, Thailand, 2005, pp. 79–86.

[2] C. Wang, P. Krafft *et al.*, "Manifold alignment," in *Manifold Learning: Theory and Applications*, Y. Ma and Y. Fu, Eds. CRC Press, 2011.

[3] M. Fattahi, M. H. Moattar *et al.*, "Locally alignment based manifold learning for simultaneous feature selection and extraction in classification problems," *Knowledge-Based Systems*, vol. 259, p. 110088, 2023.

[4] D. Ramachandra and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96–108, 2017.

[5] C. Zhang, Z. Yang *et al.*, "Multimodal intelligence: Representation learning, information fusion, and applications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 478–493, 2020.

[6] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[7] J. S. Rhodes and A. G. Rustad, "Graph integration for diffusion-based manifold alignment," 2024. [Online]. Available: https://arxiv.org/abs/2410.22978

[8] A. F. Duque Correa, M. Lizotte *et al.*, "Manifold alignment with label information," in *2023 International Conference on Sampling Theory and Applications (SampTA)*, 2023, pp. 1–6.

[9] D. Tuia and G. Camps-Valls, "Kernel manifold alignment for domain adaptation," *PLoS One*, vol. 11, no. 2, p. e0148655, Feb. 2016.

[10] R. R. Coifman and S. Lafon, "Diffusion maps," *Appl. Comput. Harmon. Anal.*, vol. 21, no. 1, pp. 5–30, 2006, special Issue: Diffusion Maps and Wavelets.

[11] J. Liu, Y. Huang *et al.*, "Jointly embedding multiple Single-Cell omics measurements," *Algorithms Bioinform*, vol. 143, Sep. 2019.

[12] A. Cutler, D. R. Cutler *et al.*, "Random forests," in *Ensemble Machine Learning: Methods and Applications*, C. Zhang and Y. Ma, Eds. Boston, MA: Springer US, 2012, pp. 157–175.

[13] L. S. Whitmore, A. George *et al.*, "Explicating feature contribution using random forest proximity distances," 2018.

[14] M. Kokla, J. Virtanen *et al.*, "Random forest-based imputation outperforms other methods for imputing lc-ms metabolomics data: a comparative study," *BMC Bioinf.*, vol. 20, no. 1, p. 492, Oct 2019.

[15] Y. Lin and Y. Jeon, "Random forests and adaptive nearest neighbors," *J. Am. Stat. Assoc.*, vol. 101, no. 474, pp. 578–590, 2006.

[16] J. S. Rhodes, A. Cutler *et al.*, "Geometry- and accuracy-preserving random forest proximities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 947–10 959, 2023.

[17] J. S. Rhodes, "Supervised manifold learning via random forest geometry-preserving proximities*," *2023 International Conference on Sampling Theory and Applications (SampTA)*, pp. 1–5, 2023.

[18] J. Ham, D. Lee *et al.*, "Semisupervised alignment of manifolds," in *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*. PMLR, 2005, pp. 120–127.

[19] S. Lafon, Y. Keller *et al.*, "Data fusion and multicue data matching by diffusion maps," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1784–1797, 2006.

[20] C. Wang and S. Mahadevan, "Manifold alignment using procrustes analysis," in *Proceedings of the 25th International Conference on Machine Learning*, ser. ICML '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 1120–1127.

[21] M. Amodio and S. Krishnaswamy, "Magan: Aligning biological manifolds," in *International Conference on Machine Learning*, 2018.

[22] A. F. Duque, G. Wolf *et al.*, "Diffusion transport alignment," in *Advances in Intelligent Data Analysis XXI*, B. Crémilleux, S. Hess *et al.*, Eds. Cham: Springer Nature Switzerland, 2023, pp. 116–129.

[23] K. R. Moon, D. van Dijk *et al.*, "Visualizing structure and transitions in high-dimensional biological data," *Nat. Biotechnol.*, vol. 37, no. 12, pp. 1482–1492, Dec 2019.

[24] A. F. Duque, G. Wolf *et al.*, "Visualizing high dimensional dynamical processes," in *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2019, pp. 1–6.

[25] M. Kelly, R. Longjohn *et al.*, "The uci machine learning repository," 2024, accessed: 2024-10-30.