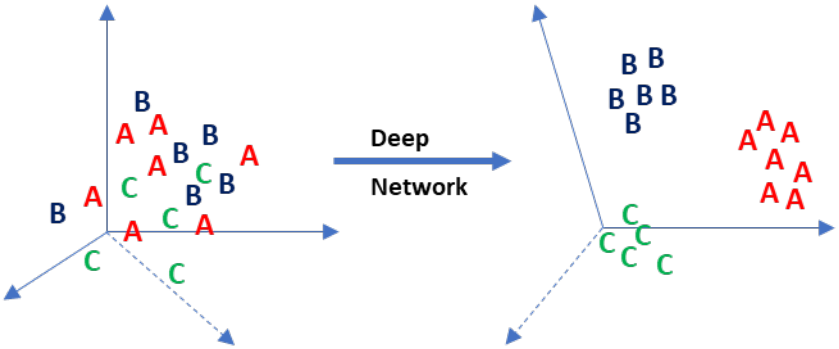
Data Science Research Series

Deep Kernel Method



BACKGROUND

Kernel method is a branch of machine learning that uses a kernel function to implicitly map the data to a feature space where modeling is "easier". For example, in classification, instances from the same classes are closer in the feature space and different-class data points are further. One disadvantage of using kernel methods is the need of optimizing the kernel functions – both the choice of function and its hyper-parameters – which can be a non-trivial task. Hyper-parameter tuning algorithms like Grid-search are not always efficient in terms of running time and model accuracy. Moreover, wrong choices of kernels or hyper-parameters may significantly drop the model's performance. Lastly, kernel functions like RBF or polynomial, even though with tuned hyper-parameters, are not learnable with respect to data.

The deep kernel tries to learn the similarity function directly from the data using a deep neural network. The proposed method lifts the burden of optimizing the kernel function from users and improves the model's performance.

APPROACH

A deep kernel is a deep neural network which simultaneously takes two data points as input, and outputs their similarity. The generated kernel value is the probability of two data points having the same labels (belonging to the same classes). Let $y^{(i)}, y^{(j)}$ be the classes of two data points $x^{(i)}, x^{(j)}$, then

$$K\left(x^{(i)}, x^{(j)}\right) = P\left(y^{(i)} = y^{(j)}\right)$$

The training data for the deep kernel is generated as:

- $X_{i,j} = \left\{x_1^{(i)}x_1^{(j)},\ \ldots x_d^{(i)}x_d^{(j)},\ e^{-\left|x_1^{(i)}-x_1^{(j)}\right|},\ \ldots\ e^{-\left|x_d^{(i)}-x_d^{(j)}\right|}\right\}$
- $\begin{cases} Y_{i,j} = 0 \text{ if } y_i \neq y_j \\ Y_{i,j} = 1 \text{ if } y_i = y_j \end{cases}$

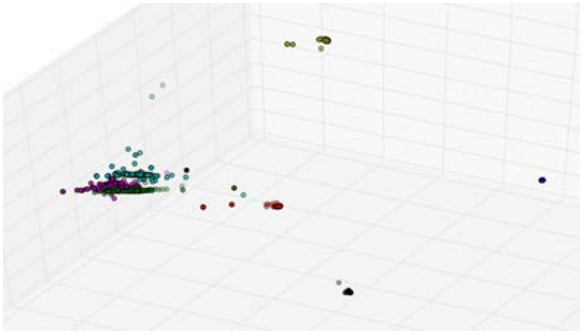
An unsupervised version of the deep kernel is developed for visualization of high dimensional data set. In other words, it maps high dimensional data to a 3D space with minimal changes in the data cluster structure. As data labels are not available in the unsupervised case, K-Means clustering is utilized to generate "pseudo-labels" for the data. The cluster labels are used in the process of training the deep kernel. The trained kernel is finally used in kernel PCA to map the data to a 3D space for visualization.

RESULTS

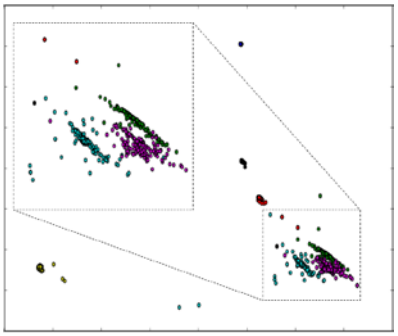
The supervised deep kernel is tested in classification and compared to the RBF kernel (with optimized hyper-parameters). In all experiments, the data is split into 70% training and 30% testing. A SVM model is used on top of both kernels. The reported results are computed from the testing set. The accuracy rates are in the below table.

Data Set	No. of Classes	RBF Kernel	Deep Kernel	Deep Kernel Improvement
Breast Cancer	2	0.97317	0.98049	↑ **0.7%**
Wine Quality	11	0.57292	0.59167	↑ **1.9%**
Segment	7	0.95844	0.97143	↑ **1.3%**
Cardiotocography	2	0.97179	0.99373	↑ **2.2%**
Pima Indians Diabetes	6	0.75325	0.78355	↑ **3.03%**
Breast Cancer Wisconsin (Diagnostic)	2	0.97661	0.98246	↑ **0.58%**

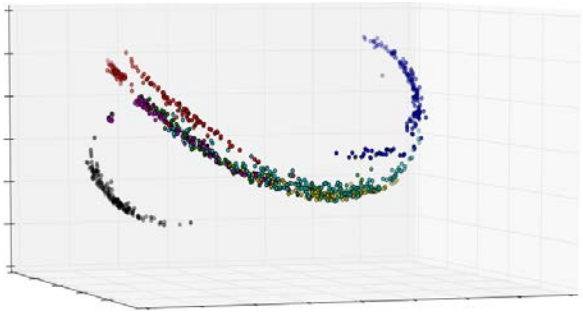
Additionally, the kernel can be used to visualize the feature space:



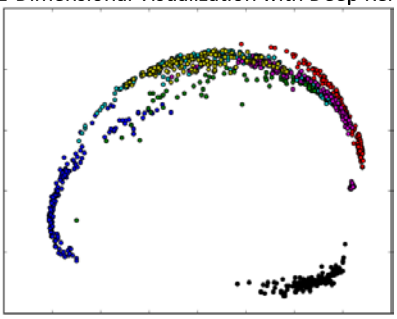
a) 3-Dimensional Visualization using Deep Kernel



b) 2-Dimensional Visualization with Deep Kernel



c) 3-Dimensional Visualization using RBF Kernel



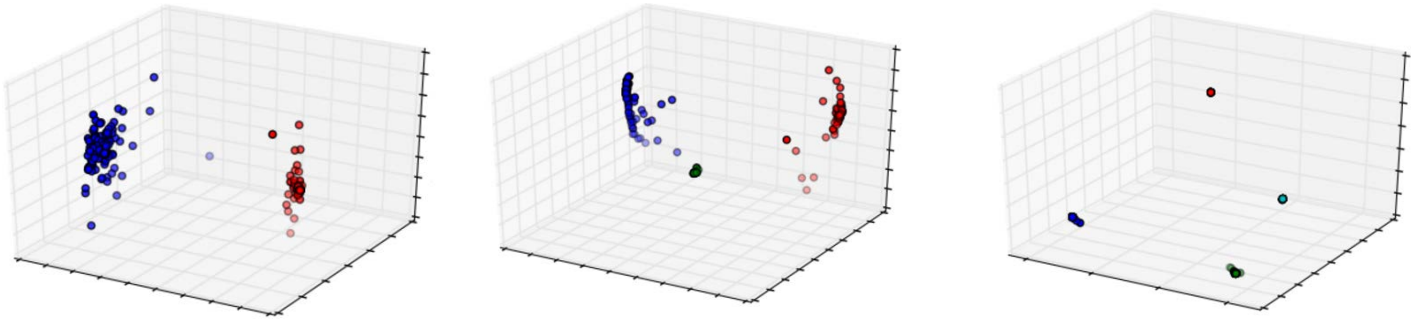
d) 2-Dimensional Visualization using RBF Kernel

which shows the advantage of the deep kernel: even in a 2D space, the deep kernel can still well separate all 7 classes in the Segment data while the RBF kernel fails to do so.

The unsupervised deep kernel is tested and compared to numerous methods of dimension reduction, including PCA, kernel PCA with RBF and polynomial kernel, entropy component analysis, deep architectures (deep belief network and stacked auto-encoder), etc. The V_{3D} of all methods is recorded and shown in the below table ($\boldsymbol{V_{3D}}$ measures the degree of structural information kept when performing dimension reduction).

Dataset	Number of Dimensions	Optimal Clusters	V_{3D}								
			PCA	DK-PCA	GK-PCA	PK-PCA	GK-ECA	PK-ECA	SVD	DBN	SAE
Ionosphere	34	2	0.984753	***1***	0.974991	0.974991	0.9743124	0.9743124	0.780634	0.708305	0.709306
Wine	13	3	0.93403	***1***	0.806865	0.806865	0.7808524	0.8068653	0.839379	0.660554	0.689654
Shuttle	9	4	1	***1***	1	1	1	1	0.700945	0.750053	0.792001
Ecoli	8	3	0.983225	***1***	0.944427	0.944274	0.4419419	0.9305372	0.729715	0.769727	0.814014
Breast Cancer	32	2	1	***1***	1	1	1	1	0.937836	0.94336	0.937836

As can be seen, the deep kernel outperforms all other methods with a V_{3D} of 1 in all cases, indicating no structural information is lost after reducing the dimension of the data to 3. Some visualization results are included below, the colors show the cluster labels from the original space.



a) Ionosphere b) Wine c) Shuttle

CONCLUSIONS

The Deep Kernel is a deep neural network designed to learn the similarity of data. It lifts the burden of optimizing kernel functions in kernel methods while improve model accuracy. Experiments show that the deep kernel outperforms the traditional RBF kernel in classification and dimension reduction for visualization.

CITATION FOR FULL ARTICLES

- Ying Xie, Linh Le, Jie Hao, *Unsupervised Deep Kernel for High Dimensional Data*, In Proceeding of the 30th IEEE International Joint Conference on Neural Networks
- Linh Le, Jie Hao, Ying Xie, Jennifer Priestley, *Deep Kernel: Learning the Kernel Function from Data Using Deep Neural Network,* In Proceedings of the Third IEEE/ACM International Conference on Big Data Computing, Applications and Technologies