Manifold Learning-based Data Sampling for Model Training

Shuqing Chen[1], Sabrina Dorn[2], Michael Lell[3],
Marc Kachelrieß[2],Andreas Maier[1]

[1]Pattern Recognition Lab, FAU Erlangen-Nürnberg
[2]German Cancer Research Center (DKFZ), Heidelberg, Germany
[3]University Hospital Nürnberg, Paracelsus Medical University, Nürnberg, Germany
shuqing.chen@fau.de

Abstract. Training data sampling is an important task in machine learning especially for data with small sample size and data with nonuniform sample distribution. Dividing data into different data sets randomly can cause the problem that, the training model covers only parts of the sampled cases and works inaccurately for weakly sampled cases. Recent research showed the benefit of manifold learning techniques in medical image processing. In this work, we propose a manifold learning based approach to improve the data division and the model training. We evaluated the proposed approach using an atlas registration framework and a deep learning framework. The final segmentation results using methods with and without data balancing were compared. All of the final segmentations were improved after implementing the manifold learning based approach into the frameworks. The largest improvement was 24.4%. Thus, the proposed manifold learning based approach is effective for the model training.

1 Introduction

To model common properties and variations from many individuals over an entire task domain is important for many machine learning methods, such as atlas registration and deep learning. These methods can be used for varied applications, e.g. multi-organ segmentation in the field of medical image processing. Model training is an important part of such methods. A well-trained model works for most cases within the domain. However, due to the limited data sample, size and nonuniform sample distribution, problems arise in practical implementations. For many applications, the data has to be sampled into different balanced sets before starting the modeling. With small sample size or nonuniform sample distribution, random data selection may not work robustly.

Aljabar et al. [1] discussed the power of manifold learning in the field of medical image processing. Using manifold learning techniques, high dimensional medical data can be converted to lower dimensional representation while respecting the intrinsic geometry [2], so that it is possible to facilitate the application of machine learning techniques such as clustering and regression. Manifold learning techniques can be used in the preprocessing to extract features for registration [3]. They can also improve segmentation successfully by allowing the propagation of multiple atlases to a diverse set of images [4]. Furthermore, manifold learning techniques can be used for clinical classification [1]. However, few research showed the benefit of the manifold learning to reduce the problem caused by data selection in many machine learning methods. In this paper, we proposed one manifold learning based approach to solve the data selection problem and to improve the model training.

2 Materials and Methods

In order to avoid the bias due to the data selection and to keep the distribution of the data sets (i.e. training data set, test data set as well as validation data set if required) similar, three steps were employed to improve the data selection:

- *Data Representation*: Initially, the high-dimensional volumetric medical data will be projected into a low-dimensional (e.g. 2-D) visualization plane using manifold learning techniques. Each point in such visualization plane will denote a volumetric image. To improve the performance, the data should be resampled to the same image spacing before using manifold learning.
- *Data Clustering*: Afterwards, the data points can be divided into different classes using clustering techniques.
- *Data Selection*: Finally, the training data set, the validation data set and the test dataset can be built by selecting data samples from each class randomly.

To show the effectivity of this datatype-independent approach in general machine learning methods, the evaluation is designed with the multi-organ segmentation using atlas registration on computed tomography (CT) images [5] and the multi-organ segmentation using deep learning on dual-energy CT (DECT) images [6]. For our cases, the clusters are well distributed with less overlap on 2-D space. Thus, data projection on 2-D space is sufficient.

The method of the multi-organ segmentation using atlas registration is described in [5]. In this method, the data selection for the atlas modeling should focus on the inter-subject organ shape variation. Therefore, the data selection approach is applied after the step of the affine registration, in order to avoid the effect of the position variance. The atlas construction part described in [5] is improved with data selection in following steps. First of all, the data is cropped to get the region of interest (ROI) for the redundancy reduction of the atlas. Then a reference volume is selected which is the most similar to the mean volume of all samples. Subsequently, affine registration is used to reduce the variation of the rotation, the translation, the scaling and the shearing. The results after the affine registration are then split into training data set and test data set using the manifold learning based approach mentioned previously. Average volume and atlas is constructed finally after the fine alignment based on B-Spline registration as described in [5].

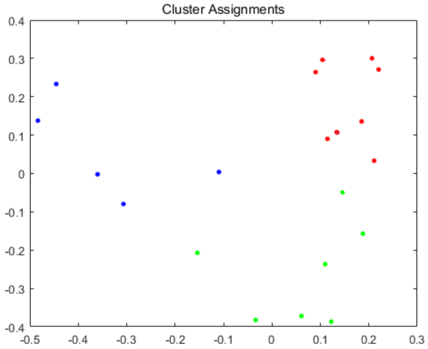

Fig. 1. Data representation and clustering using LLE and k-Means for atlas registration data. Colors denote classes, numbers denote volume indexes.

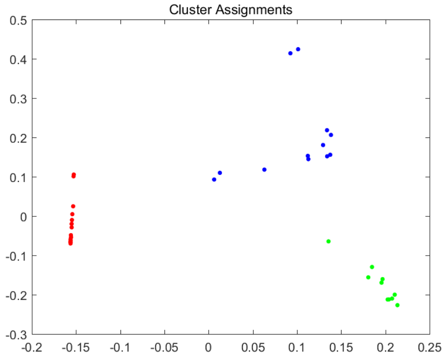

Fig. 2. Data representation and clustering using LLE and k-Means for deep learning data. Colors denote classes, volume indexes are ignored for clarity.

A 3-D cascaded fully connected network [7] was used for the multi-organ segmentation using deep learning technique [6]. In this network, the input training data will be augmented with more translation and rotation. Thus, unlike the atlas registration, it is not required to remove the position variation for the data selection. The data selection approach can be applied directly to the original data set without any preprocessing. Training data set, validation data set and test data set are selected using the proposed selection approach. The whole framework is kept in same as the description in [6].

3 Results

The Matlab toolbox provided by van der Maaten et al. [8] was used for manifold learning in our implementation. Dice coefficient was utilized to measure the performance of the segmentation result.

To show the effect of the data selection on atlas registration, 20 VISCERAL non-enhanced CT volumes [9] were used for the evaluation. Fig. 1 plots the 20 CT volumes by characterizing the shape variation using the manifold learning approach. Each point denotes one CT volume. Locally linear embedding (LLE) [10] was chosen to reduce the dimensionality because LLE showed the best performance by experimenting the different manifold learning methods provided in the toolbox. To construct balanced training, test and validation sets, the data was clustered into k classes that should be evenly represented in each set. k-Means was used here for the data clustering with $k = 3$, because $k = 3$ is most reasonable for these volumes. This is indicated by color in Figs. 1 and 2. A 10-fold cross validation was tested for 6 target organs including left and right lung, liver, spleen, as well as left and right kidney. To construct balanced sets, one volume was selected randomly from each class as test volume, in total 3 test

Table 1. Comparison of atlas registration on CT images without and with data selection.

	Right Lung	Left Lung	Right Kidney	Left Kidney	Liver	Spleen
Without Data Selection						
Avg.	0.960	0.957	0.794	0.731	0.900	0.813
Std. Dev.	0.014	0.015	0.140	0.214	0.034	0.104
With Data Selection						
Avg.	0.965	0.960	0.834	0.821	0.912	0.842
Std. Dev.	0.009	0.010	0.080	0.121	0.024	0.051
Improvements of the average						
Diff. Avg.	0.005	0.003	0.040	0.090	0.012	0.029

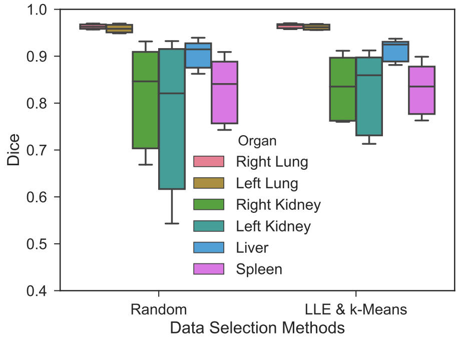

Fig. 3. Comparision test for atlas registration with/without data selection

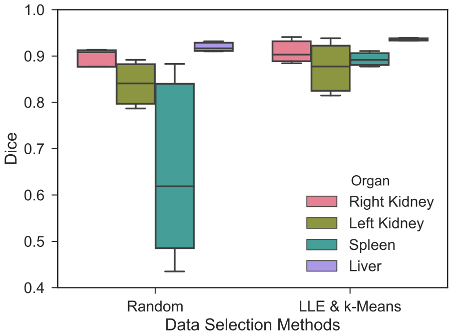

Fig. 4. Comparison test for U-Net with/without data selection

volumes were selected as test data set. The remaining 17 volumes were used as training data set. Test data sets were segmented using the segmentation method mentioned in [5]. For comparison, a 10-fold cross validation was tested using the 20 CT volumes for the same target organs based on the atlas construction method described in [5]. The same reference volume was used for the registrations, but training data set and test data set was selected randomly. The amount of the data sets were kept in same, i.e. 17 for training and 3 for the test. The results were summarized in Table 1 and shown in Fig. 3. The final segmentation of all target organs was improved with the data selection approach. Right lung and left lung has slight improvement with 0.5% and 0.3%, respectively. Liver has small improvement with about 1%. Other organs have a significant improvement from around 3% to around 9%. The distributions of the Dice coefficients are also converged significantly.

To show the effect of the data selection on deep learning, 42 clinical DECT volumes were used for the evaluation. The data representation and the data clustering is illustrated in Fig. 2. The dimensionality was reduced using LLE. Each point denotes one DECT volume. The data points were clustered into 3

Table 2. Comparison of U-Net on DECT images without and with data selection.

	Right Kidney	Left Kidney	Liver	Spleen
Without Data Selection				
Avg.	0.905	0.856	0.919	0.652
Std. Dev.	0.020	0.047	0.015	0.188
With Data Selection				
Avg.	0.905	0.863	0.934	0.896
Std. Dev.	0.034	0.071	0.011	0.032
Improvements of the average				
Diff. Avg.	0.000	0.007	0.015	0.244

classes by using k-Means with $k = 3$. Like described in [6], the ratio 5:1:1 was used for the data selection. That means, 2 volumes from each class were selected randomly for validation and test. In total, validation data set and test data set was built with 6 volumes, respectively. The remaining 30 volumes were used as training data set. The segmentation of liver, spleen, as well as left and right kidney, were evaluated. 0.9 was taken as training weight and test weight. A comparison model was experimented using same framework and same condition but with randomly generated data sets. The results were summarized in Table 2 and presented in Fig. 4. The data selection approach improved all final segmentation. Right kidney and left kidney has slight improvement with 0.00% and 0.7%, respectively. Liver has small improvement with about 1.5%. Spleen has significant improvement around 24.4%.

4 Discussion

We proposed a manifold learning based approach to reduce the bias of the data selection. The proposed approach was implemented into an atlas registration framework and a deep learning framework. The comparison evaluation showed the benefit of the data selection approach. Both of the machine learning methods can be improved by adding the proposed approach.

The approach can be tested for more machine learning methods in future. Moreover, more manifold learning techniques can be investigated further. Furthermore, other clustering techniques can be used.

ACKNOWLEDGMENTS This work was supported by the German Research Foundation (DFG) through research grant No. KA 1678/20, LE 2763/2-1 and MA 4898/5-1.

References

1. Aljabar P, Wolz R, Rueckert D. Manifold learning for medical image registration, segmentation, and classification. Machine Learning in Computer-Aided Diagnosis.

2012; p. 351.

2. Maier A, Schuster M, Eysholdt U, et al. QMOS - A Robust Visualization Method for Speaker Dependencies with Different Microphones. J Pattern Recognit Res. 2009;4(1):32–51.
3. Wachinger C, Navab N. Manifold Learning for Multi-Modal Image Registration. Proc 11th BMVC. 2010 01; p. 1–12.
4. Wolz R, Aljabar P, Hajnal JV, et al. LEAP: Learning embeddings for atlas propagation. NeuroImage. 2010;49(2):1316 – 1325.
5. Chen S, Endres J, Dorn S, et al. A Feasibility Study of Automatic Multi-Organ Segmentation Using Probabilistic Atlas. Proc BVM. 2017; p. 218–223.
6. Chen S, Roth H, Dorn S, et al. Towards Automatic Abdominal Multi-Organ Segmentation in Dual Energy CT using Cascaded 3D Fully Convolutional Network;.
7. Roth HR, Oda H, Hayashi Y, et al. Hierarchical 3D fully convolutional networks for multi-organ segmentation;.
8. van der Maaten LJP, Postma EO, van den Herik HJ. Dimensionality Reduction: A Comparative Review; 2008.
9. Jiménez-del Toro OA, Dicente Cid Y, Depeursinge A, et al. Hierarchic Anatomical Structure Segmentation Guided by Spatial Correlations (AnatSeg–Gspac): VISCERAL Anatomy3. Proc Visc Chall ISBI. 2015 Apr; p. 22–26.
10. Roweis ST, Saul LK. Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science. 2000;290(5500):2323–2326.