A Large-Scale Manifold Learning Approach for Brain Tumor Progression Prediction

Loc Tran[1], Deb Banerjee[1], Xiaoyan Sun[1], Jihong Wang[4], Ashok J. Kumar[4], David Vinning[4], Frederic D. McKenzie[3], Yaohang Li[2], and Jiang Li[1]

[1] Departments of ECE, [2]CS, [3] MSVE, Old Dominion University, Norfolk, VA 23529
[4] Diagnostic Imaging, University of Texas MD Anderson Cancer Center, Houston, TX 77030

Abstract. We present a novel manifold learning approach to efficiently identify low-dimensional structures, known as manifolds, embedded in large-scale, high dimensional MRI datasets for brain tumor growth prediction. The datasets consist of a series of MRI scans for three patients with tumor and progressed regions identified. We attempt to identify low dimensional manifolds for tumor, progressed and normal tissues, and most importantly, to verify if the progression manifold exists - the bridge between tumor and normal manifolds. By mapping the bridge manifold back to MRI image space, this method has the potential to predict tumor progression, thereby, greatly benefiting patient management. Preliminary results supported our hypothesis: normal and tumor manifolds are well separated in a low dimensional space and the progressed manifold is found to lie roughly between them but closer to the tumor manifold.

1 Introduction

With the rapid advancement of diagnostic imaging technology, multi-dimensional, large-scale, and heterogeneous medical datasets are generated routinely in clinical imaging exams. For instance, MR diffusion tensor imaging (DTI) has become a routine component of the brain MR imaging exams in many institutions. Although this new imaging modality together with the traditional T1, T2 or FLAIR weighted MRI scans has provided additional information and has shown potential for better brain tumor diagnosis, interpreting these large-scale, high-dimensional datasets simultaneously is challenging [1, 2]. Due to the fact that significant correlations exist among these multi-dimensional images, we hypothesize that low-dimensional geometry data structures (manifolds) are embedded in the high-dimensional space. Those manifolds might be hidden from human viewers because it is challenging for human viewers to interpret high-dimensional data. The hidden manifolds correctly extracted from the high-dimensional space may provide particularly useful information for brain cancers studies. For example, one may investigate the residence of cancer and normal tissues on the manifolds to derive rules to accurately classify cancer regions. Moreover, the bridge manifolds connecting the cancer and normal tissue manifolds may provide hints for identifying cancer progression trajectory, which can be used for predicting future tumor growth.

K. Suzuki et al. (Eds.): MLMI 2011, LNCS 7009, pp. 265–272, 2011.

Many manifold learning algorithms essentially perform an eigenvector analysis on a data similarity matrix whose size is n by n, where n is the number of data samples. The memory complexity of the analysis is at least $O(n^2)$, which is not feasible for very large datasets in terms of both computational and storage requirements for a regular computer. To solve this problem, statistical sampling methods are typically used to sample a subset of data points as landmarks, a skeleton of the manifold is then identified based on the landmarks and the remaining data points can be inserted into the skeleton by a number of methods such as Nystrom approximation, column-sampling and locally linear embedding (LLE) [3-5]. To keep a faithful representation of the original manifold, effective sampling should be considered. Undersampling will distort true embedded geometry structures and thus lead to subsequent manifold learning failure while oversampling may introduce unnecessary noise. For example, the landmark MDS performs poorly for randomly chosen landmarks if the data is noisy (contains outliers) [6]. Also, data may sometimes collapse to a central point in the low dimensional space if certain "important" samples are missing [5].

Mathematical models have been studied in recent years for glioma tumor growth prediction. These models are usually classified into three major types: microscopic, mesoscopic and macroscopic. Microscopic models describe the growth process in the sub-cellular level, concentrating on activities that happen inside the tumor cell. Mesoscopic approaches focus on interactions between tumor cells and their surrounding tissue while macroscopic approaches focus on tissue level processes considering macroscopic quantities such as tumor volume and blood flow [7]. Most of the macroscopic methods use a reaction-diffusion model based on a diffusion equation introduced by Murray [8]. These models usually consist of a set of parameters which are estimated from data and have shown predictive values [9-10].

In this paper, we developed a different brain tumor progress prediction method using a data-driven manifold learning approach. We first selected a set of landmarks from a large dataset based on an importance function learned from the data, based on which we learned a manifold skeleton using the local tangent space alignment (LTSA) algorithm [11]. We then inserted the remaining data points into the skeleton using the LLE method [12]. There are several parameters to be optimized including the number of landmarks needed and the number of neighbors in the LTSA algorithm. We defined two cost functions for optimizing the parameters. We applied the method to MRI datasets from three brain tumor patients aiming to predict tumor progression by searching for the “bridge” manifold.

2 Method

2.1 Data Preparation

The MRI data of three brain tumor patients were collected using various MRI scans including FLAIR, T1-weighted, post-contrast T1-weighted, T2-weighted, and DTI. Five scalar volumes were also computed from the DTI volume including apparent

diffusion coefficient (ADC), fractional anisotropy (FA), max-, min-, and middle-eigenvalues yielding a total of ten image volumes for each visit of every patient. Each patient went through a series of visits over a time span of two years. For each patient, a rigid registration was utilized to align all volumes to the DTI volume at the first visit using the vtkCISG toolkit [13]. After registration, each pixel location can be represented by a ten dimensional feature vector corresponding to the ten MRI scans. We then selected two visits denoted as "visit 1" and "visit 2" with expanded tumor regions in visit 2 for our experiments. A radiologist defined the tumor regions on the post-contrast T1-weighted and FLAIR scans, respectively. We also defined normal regions far away from the tumor regions for training purposes. Figure 1 shows example MRI slices overlaid on the defined tumor and normal regions.

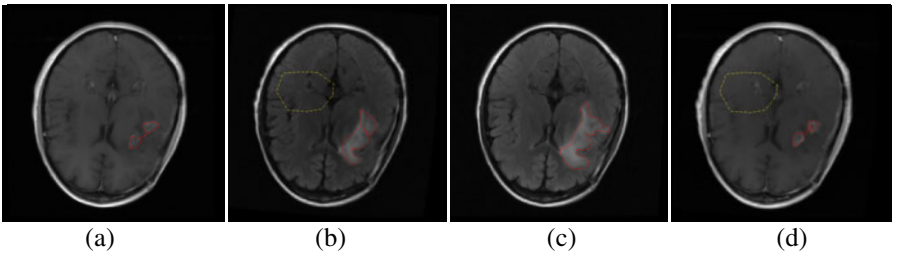


Fig. 1. Tumor and normal regions defined for patient A where the red tumor regions are labeled by a radiologist and the yellow polygon denotes normal regions. (a-b) FLAIR images at visit 1 and 2. (c-d) Post-contrast T1 images at visit 1 and 2.

2.2 Proposed System

There are roughly 65k (each slice contains 256x256 pixels) high dimensional data points in one MRI slice. We propose to use the LTSA algorithm [11] to learn the low dimensional manifolds at visit 1. Applying the LTSA algorithm to all data points is not feasible for a regular PC as we discussed above. Therefore, we developed an advanced sampling technique to select a set of landmarks based on which we learned a manifold skeleton. We then inserted the remaining data points into the skeleton using the LLE algorithm [11]. By combining the learned manifold at visit 1 with those tumor and normal regions defined at visits 1 and 2, we designed methods for predicting tumor progression and attempted to identify the "bridge" manifold, which is associated with the progressed tissues from visit 1 to visit 2. The system diagram of the proposed framework is shown in Fig. 2 and will be described as below.

Sampling based on Local Tangent Space Variation (LTV): To keep a faithful representation of the original manifold, landmarks should be carefully selected from the original data. Ideally, landmarks should be the smallest subset that can preserve the geometry in the original data. Fig. 3 shows a toy dataset to illustrate the basic

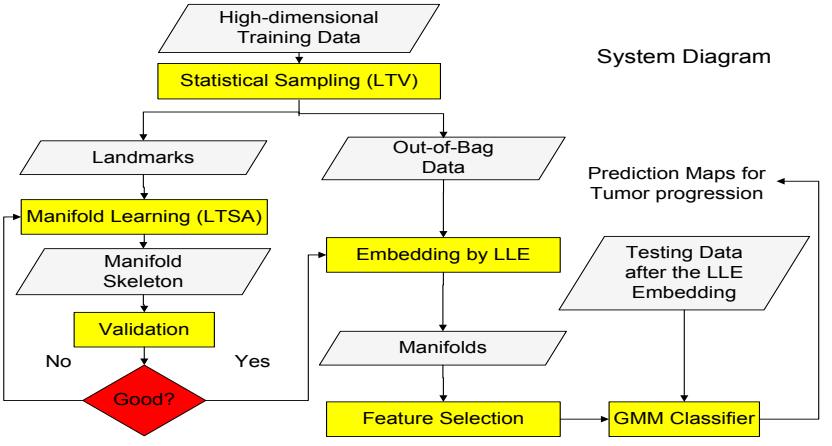


Fig. 2. System diagram for the manifold learning and tumor progression prediction

idea of LTV. Heuristically, to preserve the data structure after sampling, we should keep more data points near area 'A' rather than area 'B' in Fig. 3 because data structures near 'A' change more abruptly. Based on this observation, we assigned an importance value for each of the points by computing the local tangent space variation for it. For each data point in the dataset, we found its *k*-nearest neighbors and performed a local principle component analysis on the *k*-nearest neighbors including itself. We then identified the eigenvector (spans the tangent space) corresponding to the largest eigenvalue as the red arrows shown in Fig. 3. For each data point, it has *k* such eigenvectors and we computed the mean value of angles between its eigenvector and the eigenvectors of all of its *k*-nearest neighbors. We then normalized the importance values across all data points such that they sum to one. We then sampled the dataset to obtain a set of landmarks based on the importance values.

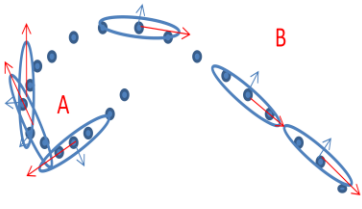


Fig. 3. The Concept of Local Tangent Space Variation

Validation of the Learned Manifold Skeleton: Successful manifold learning depends on many factors such as parameter choices for the learning algorithm and numerical stability of the algorithm. In this study, we need to determine the number of landmarks to be selected and the number of neighbors to be used by the LTSA algorithm. We proposed two cost functions as follows,

$$SF = \frac{\sum_{i,j,i\neq j}(d_M(i,j) - d_m(i,j))^2}{\sum_{i,j,i\neq j} d_M^2(i,j)} \text{ and } Acc = \frac{No.\ of\ correctly\ classified\ training\ data}{Total\ no\ .of\ training\ data}$$

where *SF* stands for the stress function and *Acc* represents the training accuracy after manifold learning, $d_M(i,j)$ denotes the geodesic distance between data points *i* and *j* and $d_m(i,j)$ represents their Euclidean distance. Intuitively, if the nonlinear manifold in high dimensional space is successfully unfolded, the Euclidean distance between points *i* and *j* will be the same as that of the geodesic distance in the low dimensional space. If a set of labeled training dataset is provided, which is the case in our study, the *Acc* value is a good criterion to verify the manifold learning. Otherwise, the stress function can be used in an unsupervised manner requiring no label information.

Embedding by LLE: Once the manifold skeleton is learned, we utilized the LLE algorithm [12, 5] to insert the remaining data points into the manifold skeleton as shown in Fig. 4, where the red dots are landmarks consisting the manifold skeleton, and the yellow square is a remaining data point to be embedded into the skeleton. We used three steps to perform this task. 1) Discovered *K* nearest landmarks in the original data space for the yellow square, 2) Computed a linear model that can best reconstruct the yellow square using the *K* landmarks and 3) inserted the yellow square into the skeleton by reusing the reconstruction weights in the linear model.

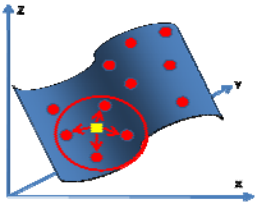


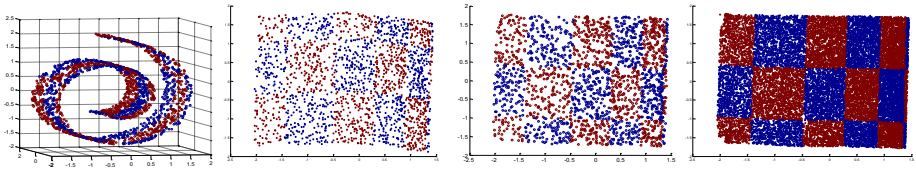
Fig. 4. Illustration of LLE embedding

Feature Selection and Classifier Training: The learned low dimensions are usually not equally important for the subsequent classification. We ranked those dimensions by the Fisher score [14] based on the data points in the labeled regions at visit 1. In our experiments, we found that the Fisher score for features beyond the 3rd one were two orders of magnitude lower than the highest score. Next, we trained a Gaussian mixture model (GMM) using the Expectation Maximization (EM) algorithm on the labeled data. We then produced a probability map for the MRI slices at visit 1 and the probability map was used to generate a binary classification by thresholding.

3 Experiments and Results

3.1 Results for a Simulated Dataset – The Swiss Roll

Figure 5 a-d) show results for the 'Swiss roll" dataset. Fig. 5a) is the original dataset having 2000 data points. Fig. 5b) shows the learned results based on 900 randomly selected landmarks (*SF* = 0.4527). Fig. 5c) is the result based on 900 landmarks selected based on the LTV concept (*SF* = 0.4293). Fig. 5d) illustrates the result for a very large Swiss Roll dataset having 20k data points. A direct manifold learning for this dataset using a regular PC is prohibitive, we utilized the manifold skeleton learned in Fig. 5c) and inserted all the 20k data points into the skeleton based on the LLE algorithm. Result in (b) is slightly worse than that in (c) as expected.

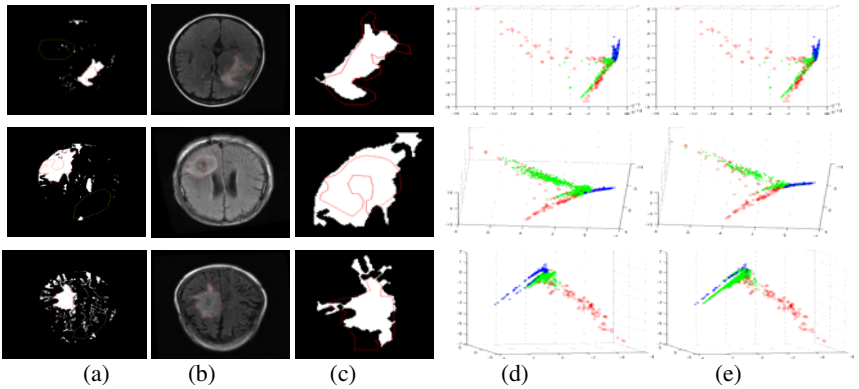


(a) (b) (c) (d)

Fig. 5. Manifold learning results for swiss roll datasets

3.2 Results from the Marked FLAIR Slices at Visit 1

In this experiment, the tumor regions marked on the FLAIR slices at visit 1 were used as the ground truth for tumor and similarly sized normal regions were selected far away from the tumor regions. We first selected a set of landmarks from the marked tumor and normal regions and optimized the skeleton manifold learning based on the *Acc* criterion. Then we embedded all data points at visit 1 inside the skull into the skeleton followed by the Fisher score ranking to keep the top three dimensions. Finally, we trained a GMM model using the selected landmarks and applied the trained model to all data points inside the skull at visit 1 to obtain prediction maps. As shown in the first column in Fig. 6, the model provided a strongly localized region for the location of the tumor with noise in other regions. The brightness of a pixel corresponds to the probability from the GMM. Using a threshold of 0.5, we formed a classification mask. To remove the noise, we segmented out only the largest blob in



(a) (b) (c) (d) (e)

Fig. 6. Results based on the tumor regions defined on FLAIRs at visit 1. (a) GMM prediction overlaid on visit 1. (b) Original FLAIRs at visit 2. (c) Post-processed prediction results overlaid on slices at visit 2. (d). Scatter plot of the three selected dimensions, where red circles denote predicted tumor samples, blue crosses denote predicted normal samples, and green dots denote predicted tumor samples outside the tumor regions defined at visit 1 (i.e., predicted progression regions at visit 2). (e) Green dots denote actual progressed tumor samples. The progression region's ground truth was obtained by subtracting the tumor regions at visit 1 from those at visit 2. Note that only the information at visit 1 was utilized to train the prediction model.

the classification mask as shown in the third column. While the predicted tumor regions are usually beyond the marked regions, this expansion into the ambiguous area may provide the potential region of growth of the tumor. Column (b) shows the FLAIR image at visit 2 and we overlaid them on the processed binary map in column (c). It can be seen that the tumor regions at visit 2 are mostly covered by the probability map. For the low dimensional feature space plots in columns d-e, one is the predicted and another is the ground truth, there are clear separations between tumor and normal tissues (red and blue dots). The green dots in (d) denote predicted tumor tissues that fell outside of the tumor regions at visit 1. These points might be used to predict the tumor progression at visit 2 and can be considered to form a "bridge" between tumor and normal tissues. In (e), the green dots denote the actual progression points obtained by referring to the tumor regions defined at visits 1 and 2.

3.3 Results from the Marked Post-contrast T1 Slices at Visit 1

We obtained similar results in this experiment except that the predicted regions extend far beyond those defined at visits 1 and 2. This may due to the fact that the prediction may be dominated by the information in the FLAIR slices. We also computed average sensitivity and specificity for the three patients based on the defined tumor and normal regions and compared them with those without manifold learning as shown in Table 1. In the Table, sensitivities at visit 2 represent the accuracies of the predicted tumor regions based on information at visit 1 matching the tumor regions defined at visit 2, which can be interpreted as the prediction accuracy. We found that manifold learning can significantly improve the classification results.

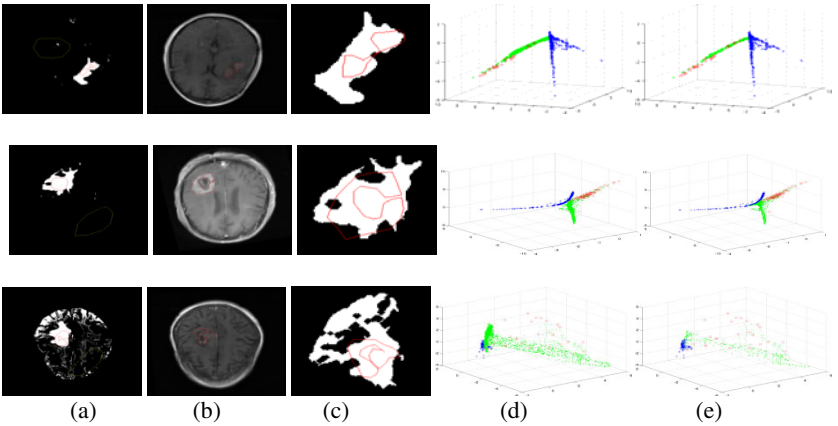


Fig. 7. Prediction results using the tumor regions defined on Post-contrast T1 slices at visit 1

4 Conclusion and Future Work

In this paper, we showed a possible nonlinear transition between normal and tumor brain tissues in high dimensional MRI datasets. Using the landmark sampling, we reduced the sample size of the MRI dataset so that conventional nonlinear dimensionality reduction techniques can be performed. There is a distinct separation

Table 1. Average sensitivities and specificities for the three patients

	Sensitivity at Visit 1	Specificity at Visit 1	Sensitivity at Visit 2
FLAIR with LTSA	**0.957**	**1.000**	**0.770**
FLAIR w/o LTSA	0.798	0.976	0.651
Post-contrast T1 with LTSA	**0.987**	**0.997**	**0.876**
Post-contrast T1 w/o LTSA	0.957	0.976	0.677

between normal and tumor tissues in the low dimensional space. We also showed that the points belonging to the tumor progression tend to accumulate between the normal and abnormal clusters. Our future work includes testing this method on more patients.

References

[1] Pauleit, D., Langen, K.J., et al.: Can the Apparent Diffusion Coefficient be Used as A Noninvasive Parameter to Distinguish Tumor Tissue from Peritumoral Tissue in Cerebral Gliomas? J. Magn. Reson. Imaging (20), 758–764 (2004)

[2] Bode, M.K., Ruohonen, J., et al.: Potential of Diffusion Imaging in Brain Tumors: A Review. Acta Radiol. (47), 585–594 (2006)

[3] Deshpande, A., Rademacher, L., et al.: Matrix approximation and projective clustering via Volume Sampling. In: Symposium on Discrete Algorithms, vol. (2), pp. 225–247 (2006)

[4] Drineas, P., Mahomey, M.W.: On the Nystrom Method for Approximating a Gram Matrix for Improved Kernel-Based Learning. JMLR (6), 2153–2175 (2005)

[5] Wang, S., Yao, J., et al.: Improved classifier for computer-aided polyp detection in CT colonography by nonlinear dimensionality reduction. Med. Phys. 35(4), 1377–1386 (2008)

[6] Silva, V., Tenenbaum, J.B.: Global Versus Local Methods in Nonlinear Dimensionality Reduction. In: NIPS, vol. 15, pp. 721–728 (2003)

[7] Hatzikirou, H., Deutsch, A., et al.: Mathematical Modelling of Glioblastoma Tumor Developement: A Review. Mathematical Models and Methods in Applied Sciences 15(11), 1779–1794 (2005)

[8] Murray, J.: Mathematical Biology. Springer, Heidelberg (1989)

[9] Atuegwu, N.C., et al.: The integration of quantitative multi-modality imaging data into mathematical models of tumors. Physics in Medicine and Biology 55, 2429–2449 (2010)

[10] Cobzas, D., Mosayebi, P., Murtha, A., Jagersand, M.: Tumor Invasion Margin on the Riemannian Space of Brain Fibers. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009. LNCS, vol. 5762, pp. 531–539. Springer, Heidelberg (2009)

[11] Zhang, Z., et al.: Principal manifolds and nonlinear dimensionality reduction via local tangent space alignment. SIAM Journal of Scientific Computing 26(1), 313–338 (2004)

[12] Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500), 2323–2326 (2000)

[13] Hartkens, T., Ruechert, D., et al.: VTK CISG Registration Toolkit: An open source software package for affine and non-rigid registration of single- and multimodal 3D images. In: Workshop Bildverarbeitung fur die Medizin, pp. 409–412 (2002)

[14] Duda, R., Hart, P.: Pattern Classification and Scene Analysis, pp. 114–129 (1973)