Identifying fiber bundles with regularised k-means clustering applied to the grid-based data

Vladimir Nikulin (Team: VladN)

Department of Mathematics, University of Queensland
{v.nikulin@uq.edu.au}

Abstract. Based on our experimental evaluations against training data, we have found that the sets of three to five 3D-points (key-sets) may be sufficient in order to identify and separate different bundles. These points may be found using regularized k-means algorithm, where the main target of the regularization is to prevent oversmoothing (creation of the superbig clusters), and to avoid small and unstable clusters. We cannot apply k-means algorithm directly to the huge original data, and grid-based preprosessing step is a very important in order to reduce the size of the data without loss of an essential information. In order to define the distance between any fiber and key-set, we consider both as an unsorted datasets of 3D-points. We find maximum distance between the point within key-set and the given fiber, where the distance between any fixed 3D-point and the fiber is a minimum distance according to the points within the fiber. Using above technique, we achieved perfect separation of the first 8 given bundles. Further, we decided to extend the methodology of the supervised classification to the case of unsupervised classification. Firstly, we found 20 centroids using regularized k-means algorithm. Then, we considered all possible key-sets of three centroids (all possible combinations of three (without repetitions) from 20). Those key-sets, which did not attract sufficient number of fibers were excluded from further consideration. Those key-sets, which attract large numbers of the same fibers were united. As a consequence, some bundles were represented by several key-sets. Finally, we identified 48 bundles (including given 8) plus the corresponding key-sets. Using this information, we can compute for any BranScan the matrix of distances between fibers and bundles. The final classifier works in the following way: 1) fill first 8 bundles; 2) identify 50000 fibers with biggest distances (label zero); 3) all remaining fibers to be directed to the nearest bundle with index grater than 8.

1 Introduction

Quantitative characterization of neuronal fiber pathways is of significant neurological and clinical interest. Quantification of neural fibers can provide valuable insight into the progression of brain disease, while visualisation of neuronal fiber

pathways can assist the analysis of fiber connectivity and potentially guide brain surgery [1].

The brain segmentation represents a very complex and challenging problem. Fiber pathways connecting the same functional regions of the brain form a natural anatomical group (bundle). Fiber bundling is a typical clustering problem. Note that the fiber bundles in the human brain take various sizes and shapes. The measure used to define the spatial proximity between curves is of fundamental importance for clustering. It is not easy (first of all in terms of the computational time) to compare different fibers directly taking into account that they have different lengths and structures. As a solution for this problem, we propose to consider an intermediate key-sets with several very important 3D-points. Depending on the proximity to the one particular set we can make a conclusion whether or not two different curves are similar.

2 Task 1: supervised classification

Table 1. Some statistical characteristics and optimisation parameters related to the training data (Brain1Scan1).

Bundle	1	2	3	4	5	6	7	8	Others
Size:	2800	816	11041	1804	1076	324	858	310	230971
Selected:	5000	2000	10000	5000	1500	2000	2000	900	200000
High:	337	114	1990	238	182	140	392	140	2109
Low:	13	6	9	9	4	5	3	1	13
k-means:	5	2	4	5	5	3	5	4	20
thresh:	13	5	20	13	13	18	11	15	-
class:	0.8746	0.4963	0.5494	0.9191	0.7063	0.9414	0.8671	0.6129	0.9721

The training 3D data (BrainScan1) include $250,000$ fibers, where the first 8 bundles (or tracks) were identified by the Organisers of the Challenge (the sizes of the particular bundles are given in the row "Size", see Table 1). Accordingly, we split BrainScan1 set into 9 subsets. For any particular subset we found minimal 3D cube containing this subset, and applied 3D grid with step size $\tau = 1$.

Remark 1. The step size τ represents a very important smoothing parameter. Value $\tau = 1$ was chosen without consideration whether or not it is an optimal.

As a next step, we computed number of occurrences for any particular grid-point (used rounding to the nearest grid-point according to the Euclidian distance). As an outcome, we obtained 9 tables $A_i, i = 1, \ldots, 9$, with the following 4 columns: 1) number of occurrencies; 2-4) coordinates X-Y-Z of the grid-points.

We sorted matrices of the grid-points A_i in a decreasing order according to the first column and selected points which are frequent enough (see Table 1: rows "Selected" -numbers of points selected, and "Low" -level of occurrences suitable for the selection). Let us denote by B_i the matrices of 3D points as an outcome of this preprocessing step.

Remark 2. The latest transformation represents a very significant data reduction: the numbers in the row "Selected" indicate numbers of 3D points in difference to the numbers of fibers in the row "Size".

Further, we applied regularised k-means clustering algorithm [2] with Euclidean distance in order to reduce B_i to a few centroids, which will be used for the identification of the bundles. There are two major problems here: stability of clustering and meaningfulness of centroids as cluster representatives. On the one hand, big clusters impose strong smoothing and possible loss of very essential information. On the other hand, small clusters are, usually, very unstable and noisy. Accordingly, they can not be treated as equal and independent representatives. To address the above problems, we applied regularisation to prevent the creation of super big clusters, and to attract data to existing small clusters.

Table 2. Confusion matrix, where the rows from 2 to 10 represent "truth", and the columns from 2 to 10 represent "report". Simulation experiments against Brain1Scan1 with the target to optimise regulation parameters, see rows "k-means" and "thresh" in Table 1.

N	1	2	3	4	5	6	7	8	Others	Total
1	2611	0	0	0	0	0	0	0	189	2800
2	0	455	0	0	0	0	0	0	361	816
3	0	0	8356	0	0	0	0	0	2685	11041
4	0	0	0	1804	0	0	0	0	0	1804
5	0	0	0	0	820	0	0	0	256	1076
6	0	0	0	0	0	311	0	0	13	324
7	0	0	0	0	0	0	744	0	114	858
8	0	0	0	0	0	0	0	244	66	310
Others	162	54	2290	146	60	6	0	54	228199	230971
Total	2773	509	10646	1950	880	317	744	298	231883	250000

Let us denote by C_i sets of centroids (key-sets) as an outcome of k-means algorithm. Combined with the following distance:

$$D_i(\mathbf{f}) = D(\mathbf{f}, C_i) = \max_{q \in C_i} \min_{x \in \mathbf{f}} \sqrt{\|x - q\|^2}, \tag{1}$$

key-sets $C_i, i = 1, \ldots, 8$, may be used as a classifier for the Task 1.

The classifier, which we employed during this Challenge works in the following way. Without loss of generality, we shall assume that the values $D_i(\mathbf{f})$ are

sorted in an increasing order. Then, we shall check sequentially the conditions

$$D_i(\mathbf{f}) \leq h_i, i = 1, \ldots, 8,$$

where $h_i > 0$ are threshold parameters.

In the case if the current condition is correct, we shall stop further checking and shall make decision that the fiber $\mathbf{f}$ belongs to the corresponding bundle. In the case if all 8 conditions are incorrect we shall make decision that the fiber $\mathbf{f}$ belongs to the "Others".

In order to run the simulations, special code in C was developed. Using this code we optimised the sizes k_i of the key-sets C_i (see row "kmeans" in Table 1), and the threshold parameters h_i, (see row "thresh" in Table 1). The time duration of one run against Brain1Scan1 was about 40 min.

Remark 3. Using these parameters, we observed remarkable outcome: perfect separation of the bundles 1 - 8, see Table 2. The corresponding classification rates (CR) (CR is the ratio of the number of correctly classified fibers to the total number of fibers within the particular bundle) are presented in the row "class" of Table 1 with an average of 0.771.

3 Task 2: unsupervised classification

We decided to extend result of Remark 3 to the unclassified set of "Others". Firstly, using the same regularised k-means algorithm as in Section 2, we found 20 centroids. In order to identify any particular bundle, we decided to consider the combinations of three different centroids. The total number of such combinations is 1140. Next, based on the row "thresh" of Table 1 we selected threshold parameter $H = 18$. With the parameter H, we computed for any particular combination of three centroids (key-set):

1. number of fibers (vector v), which are sufficiently close;
2. similarity matrix S: s_{ij} - number of common fibers, which are sufficiently close to the pair of combinations i and j.

Using condition $v_i \geq 400$ we reduced the number of combinations from 1140 to 216. Next, we unite some of the selected combinations according to the criterion $s_{ij} \geq 100$. As a result, we identified 40 different bundles. Therefore, the total number of bundles (including 8 given) was 48.

4 The final classifier

Let us consider an application of the system described in Sections 2 and 3 to the independent set Brain1Scan2 with 250,000 fibers. Using the parameters of the above sections we can compute matrix of the distances with sizes $[250000 \ \ by \ \ 48]$. The *referenceFiber* for the bundle i was selected according to the matrix of distances as a closest fiber to the bundle i.

The bundles 1-8 had higher priority and we filled them first. It may be done either according to the threshold parameters or, simply, select required number the most closest fibers.

All fibers, selected to the first 8 bundles, were excluded from further consideration. Then, we computed minimum distances using only bundles from 9 to 48 for the remaining fibers. Let us denote the corresponding matrix by D. We sorted set D according to the distances in an increasing order and 50000 fibers with biggest distances were given label zero. After that, all the fibers without label, were directed automatically to the nearest bundle.

Remark 4. After declassification of the 50000 fibers with biggest distances, some particular bundles may become empty or nearly empty. In our case we excluded from the final report 10 bundles as meaningless. The list of 38 bundles was reported.

5 Concluding remarks

It is a well known fact that for various reasons it may not be possible to theoretically analyze a particular algorithm or to compute its performance in contrast to another. The results of the proper experimental evaluation are very important as these may provide the evidence that a method outperforms existing approaches. Data mining competitions are very important.

References

[1] Ding, Z., Gore, J., Anderson, A.: Classification and quantification of neuronal fiber pathways using diffusion tensor mri. Magnetic Resonance in Medicine **49** (2003) 716–721

[2] Nikulin, V., McLachlan, G.: Regularised k-means clustering for dimension reduction applied to supervised classification. In: CIBB Conference, Genova, Italy. (2009)