Topological Data Analysis on Magnetic Resonance Image Biomarkers

Yaodong Du[1], Ming Zhang[*, 2, 3], Garrett Stonis[2], Shan Juan[*, 1]
[1]Department of Computer science, Pace University, NYC, NY
[2]Department of Computer Science & Networking, Wentworth Institute of Technology, Boston, MA
[3]Division of Rheumatology, Tufts Medical Center, Boston, MA
yd67578n@pace.edu, zhangm1@wit.edu, stonisg@wit.edu, jshan@pace.edu

***Abstract*—Big data provides us the source to extract new insights in various disciplines. In healthcare, big data may help discover the pathology of a disease and new effective treatments. However, the era of big data taxes the ability of many researchers to analyze and interact with data in biomedical research. Traditional statistical methods have limitations to infer relevant features for possibly complex data. Topological Data Analysis (TDA) provides a set of new topological and geometric tools to discover the hidden relations or key features from complex data. It may help identify the key risk factors related to a given disease and reduce the noise impact from other factors. In this paper, we use TDA to analyze the multidimensional data from MRI of knee osteoarthritis patients.**

Keywords—Topological Data Analysis, Knee Osteoarthritis, MRI, Biomarkers

I. INTRODUCTION

Big data has become the ubiquitous word of medical innovation. The rapid development of machine-learning techniques and artificial intelligence, in particular, has promised to revolutionize medical practice from the allocation of resources to the diagnosis of complex diseases [1]. However, there are challenges coming with understanding and analyzing medical big data. Due to the complexity of healthcare results from the diversity of health-related ailments and their comorbidities, the heterogeneity of treatments and outcomes, and the inherent difficulties in understanding large, high-dimensional, often, noisy data [2]. Traditional data analysis methods, despite their effectiveness, still have some drawbacks that might introduce unwanted biases or the need for ad hoc adjustments.

Topological Data Analysis (TDA) provides a general framework for analyzing data, with the advantages of being able to extract information from large volumes of high-dimensional data, while not depending on the choice of metrics, and providing stability against noise. TDA combines tools from algebraic topology and statistical learning to give a quantitative basis for the study of the "shape" of data [3]. Compared to machine learning methods such as artificial neural networks, which often requires a large number of training samples, TDA can work for a small data set.

As the most common form of arthritis, knee osteoarthritis (OA) is a major cause of motion limitation and physical disability in aged people. In 2000, over 13% of the US population (about 35 million people) endure OA disease in one or both joints with radiological evidence [4]. However, it's still unclear the pathology of OA disease, and there is no effective treatment that can alter OA progression [5].

Magnetic resonance imaging (MRI) is a non-invasive technology that generates three-dimensional (3D) images of intra-articular soft-tissue structures, which represent potential surrogate endpoints in OA. By using the 3D model to show small tissues and structures, MRI is used to identify risk factors of structural disease progression and to facilitate the testing efficacy of disease-modifying interventions [6, 7]. In addition, MRI has no radiation and is generally safe. However, 3D MRI generates a huge amount of data. In our study, each knee MRI contains 160 two-dimensional (2D) images and the resolution of each 2D image is 384 x 384 pixels. A lot of image biomarkers can be extracted from knee MRI including knee cartilage, bone marrow lesion, effusions, etc. Each biomarker has its own measuring unit and scale. Beyond that, those biomarkers are also impacted by other factors, such as age, body weight, gender, etc. It is challenging to use traditional statistical methods to study how the multidimensional biomarkers from MRI are related to OA disease. TDA provides a novel way to infer, analyze, and exploit the complex multidimensional data by grouping patients with similar conditions into a connected subnetwork. Such networks help us identify the key factors closely related to OA disease and reduce the impact of other factors (e.g., age, BMI).

II. DATA AND METHODS

A. Data Selection

We selected a sample of 200 participants from the Osteoarthritis Initiative (OAI) dataset. The 200 cases include all OA severity levels and have complete MRI, radiographic, and clinical data. We used semi-automated programs to quantify MRI biomarkers including: CDI, BML, and effusion. We assessed BML and CDI at the medial and lateral compartments for patella, tibia, and femur. 17 features were extracted in total.

1) Cartilage Damage Index (8 Features): Cartilage damage index (CDI) is a recently proposed cartilage quantification method that is much more efficient than the traditional manual segmentation of cartilage [8-10]. It quantifies osteoarthritis cartilage thickness through informative locations on knee MR images. These informative locations are selected from regions on the articular surface where cartilage denudation frequently happen. The knee joint is the most complex joint in the human body. It includes

[*] Corresponding Authors: Ming Zhang, zhangm1@wit.edu and Juan Shan, jshan@pace.edu

This research is supported by the National Science Foundation and Rheumatology Research Foundation.

Authorized licensed use limited to: Pace University. Downloaded on July 08,2020 at 23:57:25 UTC from IEEE Xplore. Restrictions apply.

femur, tibia, and patella cartilages and each cartilage is divided into medial and lateral compartments. In total, there are 6 compartments in a knee joint. The 8 CDI features are extracted from the 6 compartments and listed in Table I (CDI_total_FTP, femur_CDI_medial, femur_CDI_lateral, tibia_CDI_meidial, tibia_CDI_lateral, patella_CDI_medial, patella_CDI_lateral, femur_tibia_lateral).

2) Bone Marrow Lesion (7 features): Bone Marrow Lesion (BML) is characterized by excessive water signals in the marrow space on MRI. BMLs constitute a central component of a wide variety of inflammatory and non-inflammatory rheumatologic conditions affecting the musculoskeletal system. BML volume is measured on each of the 6 compartments. Together with the total of all the BML volumes, we have 7 features to represent BML (total_BML, femur_BML_medial, femur_BML_lateral, patella_BML, patella_BML_medial, patella_BML_lateral, tibia_BML_medial, tibia_BML_lateral).

3) Effusion (1 feature): Effusion occurs when excess synovial fluid accumulates in or around the knee joint. The volume of the effusion is used as a feature.

4) Clinical data (2 features): Besides image biomarkers, other factors may also impact the progression of OA disease. We have included age and BMI in our multidimensional data.

B. Methods

There are five steps to create TDA.

1) Create a Pearson correlation matrix using MRI biomarker and clinical data. Each of the MRI biomarkers and clinical data has its own measurement unit and scale. In the clinical study, cross-correlation is used to help identify the relations between risk factors and disease. Here we used cross-correlation to emphasize the clinical relations and also solve the unit difference, scale difference, and integrate MRI biomarkers with clinical data (Fig. 1). The cross-correlation matrix provides a metric for the topological network construction. The similarity of the metric is measured by norm correlation, which is used to define the distance of two data points in a topological space. The norm correlation of two points (X, Y) is calculated by:

$$NormCorr(X,Y) = 1 - r(X',Y')$$

where X and Y represent the data points and X^',Y' are the column-wise, mean-centered, and variance normalized version of X and Y, and the formula of r(X, Y) is :

$$r(X,Y) = \frac{N\sum_{i=1}^{N} X_i Y_i - \sum_{i=1}^{N} X_i \sum_{i=1}^{N} Y_i}{\sqrt{N\sum_i X_i^2 - (\sum_i X_i)^2}\sqrt{N\sum_i Y_i^2 - (\sum_i Y_i)^2}}$$

2) Project data into a three-dimensional cloud space by using the cross-correlation matrix across all selected variables and viewing them though a mathematical "lens" of principal component analysis (see Fig. 2a) [12].

3) Draw topology network by splitting the cloud space into a number of bins which overlapped with each other. We select single-linkage agglomerative as the clustering method, which pairs the clusters that minimally increase a given linkage distance and merge them recursively. The data points were clustered by the clustering method based on the similarity, which is defined by the norm correlation. We connect two clusters if they contain one or more same data points (see Fig. 2b) [13].

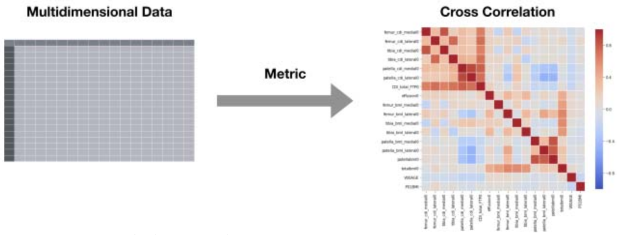

Fig. 1. Cross correlation metric

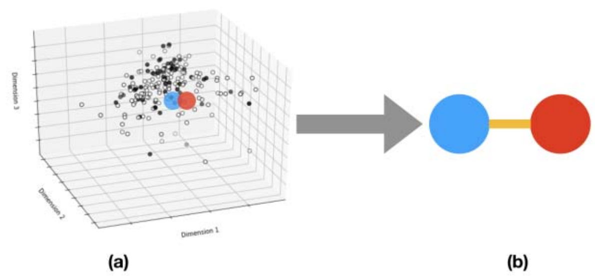

Fig. 2. Multi-dimensional data space

4) Use variables of clinical interests (e.g., smoking, disease severity score, and gender) to color topological network and visualize the hidden information of the data. Fig. 3 shows two topological networks colored by age and the change of KL grade separately. The color bars provide a reference of data representation.

5) Split the generated network into subnetworks at the weakest connection point. The sub-networks of interest were mathematically evaluated by the Kolmogorov-Smirnov (KS)

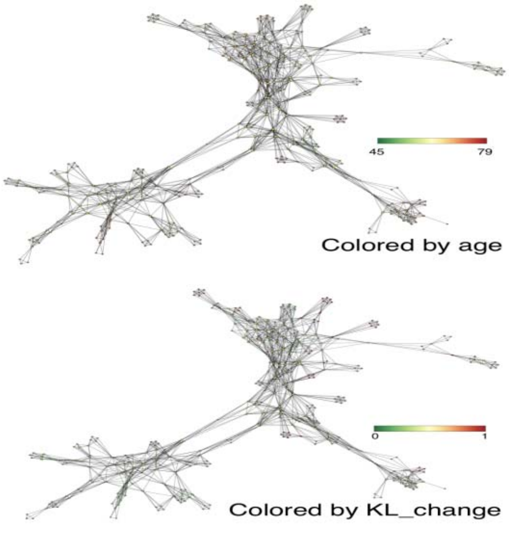

Fig. 3. Colored Topological Network

Authorized licensed use limited to: Pace University. Downloaded on July 08,2020 at 23:57:25 UTC from IEEE Xplore. Restrictions apply.

test [11] to generate hypothesizes about high risk factors. In this paper, we used KS test to rank the features by p-values.

III. Experiment and Results

A data topology network was constructed from the multidimensional data in Fig. 4. We colored the topology network by using OA progression score, i.e., two-year Kellgren-Lawrence (KL) grade change, between baseline year and two-year follow-up. Red nodes represent sample points with OA progression, green nodes represent sample points with no progression, and other colors stand for a mixture of both progression and non-progression. Two subnetworks were separated by the weakest connection points. With similar initial conditions, the two subnetworks show different progression status. We applied Student T-test on the two subnetworks and obtained p-value 0.0082 (< 0.05), which proved that the two subnetworks are statistically significantly different in OA progression. More progressed OA subjects (red nodes) appear in the right subnetwork circled by a blue dashed line, and more non-progressed (green nodes) appear in the left subnetwork circled by a red dashed line. Table I lists the results of the ranking of all the 17 features by their p-values, through running KS test on the two subnetworks. The top 3 key factors for two-year KL change are CDI_total_FTP ($p = 4.57\times10^{-21}$), patella_CDI_medial ($p = 4.75\times10^{-15}$), and femur_tibia_lateral ($p = 8.02\times10^{-14}$). The most important factor identified in this work, CDI_total_FTP, is consistent with domain knowledge about factors that affect OA progression, and it is also consistent with the findings of our previous work [14].

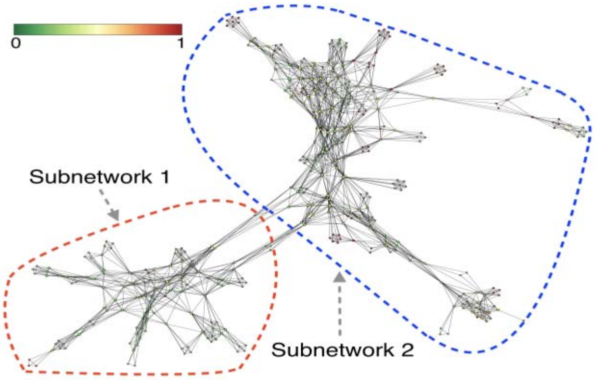

Fig. 4. Subnetworks extracted from the network colored by KL

TABLE I. Topological Data Analysis on multidimensional data and KL grade

Feature	KS score	p-value
CDI_total_FTP	0.9620	4.57E-21
patella_CDI_medial	0.8098	4.75E-15
femur_tibia_lateral	0.7750	8.02E-14
femur_tibia_medial	0.7653	1.74E-13
patella_CDI_lateral	0.7497	5.82E-13
femur_CDI_medial	0.7338	1.96E-12
femur_CDI_lateral	0.7244	3.96E-12
tibia_CDI_lateral	0.7244	3.96E-12
tibia_CDI_medial	0.6705	1.88E-10
patella_BML	0.2966	0.0219
tibia_BML_medial	0.2904	0.0263
age	0.2383	0.1082
patella_BML_lateral	0.2369	0.1121
patella_BML_medial	0.2271	0.1414
effusion volume	0.2257	0.1462
total_BML	0.1696	0.4510
tibia_bml_lateral	0.1476	0.6310
BMI	0.1233	0.8297
femur_BML_medial	0.1219	0.8402
femur_BML_lateral	0.0911	0.9833

Note: FTP = femur, tibia, patella

IV. Conclusions

TDA provides a better understanding of the clinical relations existent in multidimensional data, due to its unique advantage of data integration and visualization that traditional data analysis methods do not have. Through the analysis we found several related features with OA progression, and among them, the total CDI ranked as the top key factor. This result is consistent with the domain knowledge that cartilage is an important indicator for OA disease, which proves TDA as a reliable data analysis tool to fuse multidimensional data.

References

[1] C. H. Lee and H. J. Yoon, "Medical big data: promise and challenges," Kidney Res Clin Pract, vol. 36, no. 1, pp. 3-11, Mar 2017.

[2] I. D. Dinov, "Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data," Gigascience, vol. 5, p. 12, 2016.

[3] C. Epstein, G. Carlsson, and H. E. (2011-12-01), "Topological Data analysis," Inverse Problems, vol. 27, 2011.

[4] N. I. o. Health, "Osteoarthritis Initiative Releases First Data," in News Releases, N. I. o. Health, Ed., ed: U.S. Department of Health & Human Services, August 1, 2006.

[5] D. T. Felson et al., "Bone marrow lesions in knee osteoarthritis change in 6-12 weeks," Osteoarthritis Cartilage, vol. 20, no. 12, pp. 1514-8, Dec 2012.

[6] F. Eckstein et al., "Double echo steady state magnetic resonance imaging of knee articular cartilage at 3 Tesla: a pilot study for the Osteoarthritis Initiative," Ann Rheum Dis, vol. 65, no. 4, pp. 433-41, Apr 2006.

[7] F. Eckstein and W. Wirth, "Quantitative cartilage imaging in knee osteoarthritis," Arthritis, vol. 2011, p. 475684, 2011.

[8] M. Zhang et al., "Development of a rapid knee cartilage damage quantification method using magnetic resonance images," BMC Musculoskelet Disord, vol. 15, p. 264, 2014.

[9] M. Zhang, J. B. Driban, L. L. Price, G. H. Lo, E. Miller, and T. E. McAlindon, "Development of a Rapid Cartilage Damage Quantification Method for the Lateral Tibiofemoral Compartment Using Magnetic Resonance Images: Data from the Osteoarthritis Initiative," Biomed Res Int, vol. 2015, p. 634275, 2015.

[10] M. Zhang et al., "Cartilage Loss Primarily Occurs in the Most Affected Tibiofemoral Compartment with No Evidence of a Ceiling Effect Among Advanced-Stage Disease: A Two-Year Longitudinal Study of Data from the Osteoarthritis [abstract]," Arthritis Rheumatol, vol. 68, 2016.

[11] J. Shlens, "A Tutorial on Principal Component Analysis," Educational, vol. 51, 04/03 2014.

[12] V. Pedoia *et al.*, "MRI and biomechanics multidimensional data analysis reveals R2 -R1rho as an early predictor of cartilage lesion progression in knee osteoarthritis," *J Magn Reson Imaging,* vol. 47, no. 1, pp. 78-90, Jan 2018.

[13] P. Y. Lum *et al.*, "Extracting insights from the shape of complex data using topology," *Scientific Reports,* vol. 3, pp. 1236 EP -, 2013.

[14] Y. Du, J. Shan, R. Almajalid, T. Alon, and M. Zhang, "Using Whole Knee Cartilage Damage Index to Predict Knee Osteoarthritis: A Two-year Longitudinal Study," in 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018, pp. 623-628.

Authorized licensed use limited to: Pace University. Downloaded on July 08,2020 at 23:57:25 UTC from IEEE Xplore. Restrictions apply.