Graph Segmentation in Scientific Datasets

Rajat Sahay[1] and Savannah Thais[2]
[1]Vellore Institute of Technology [2]Princeton University

ABSTRACT

Deep learning tools are being used extensively in a range of scientific domains; in particular, there has been a steady increase in the number of geometric deep learning solutions proposed to a variety of problems involving structured or relational scientific data. In this work, we report on the performance of graph segmentation methods for two scientific datasets from different fields. Based on observations, we were able to discern the individual impact each type of graph segmentation methods has on the dataset and how they can be used as a precursors to deep learning pipelines.

DATASETS

TrackML

- Comprises multiple events.
- Each event contains simulated measurements of particles generated in a collision between proton bunches.
- All events are statistically independent and contain directional and unique particle information.

sc-PDB

Protein	N	*UniProt ID*	n
Cationic Trypsin	315	P00918	968
Bromodomain-Containing Protein 4 (BRD4)	93	O60885	192
Cyclin-Dependent Kinase 2 (CDK2)	148	P24941	490
Estrogen Receptor (ER)	52	P03372	241
Human Immunodeficiency Virus-1 (HIV-1) Protease	335	N/A*	481
Prothrombin	142	P00734	336

GRAPH CONSTRUCTION

TrackML

- Track hits are mapped into *eta-phi* space

$$\phi = \arctan 2(y, x)$$
$$\eta = \operatorname{arctanh}\left(\frac{z}{\sqrt{x^2+y^2+z^2}}\right)$$

- Hits are filtered based on p_t^{min}(>2GeV) and the number of hits in the track belong to (>2 hits/track).

sc-PDB

- Binding sites for proteins were generated using the *SiteHopper* create tool.
- Each binding cavity is described by *VolSite* using a set of pharmacological properties laid out in a 3D grid.
- Data was split into 10 groups, based on the UniProt ID of the proteins.

GRAPH SEGMENTATION

DBSCAN

- DBSCAN finds core samples of high density and expands clusters from them.
- It requires the Eps, the neighbourhood radius and the *MinPts*, which is the minimum number of points required to seed a cluster.

Spectral Clustering

- Leverage non-negative and symmetric similarity function to measure pairwise similarities and construct similarity matrix *S.*
- We leverage the ***Eigengap Heuristic*** to determine the number of partitioning clusters.
- In EH, the goal is to choose *k* such that all eigenvalues $\lambda_1,\ldots,\lambda_k$ are small but λ_{k+1} is relatively larger.

Dynamic kNN

- DkNN works on the basis of the principle of choosing the best *k* value to perform a kNN segmentation.
- It recursively uses an individual observation from the original sample for validation and the rest as training data.

Gaussian Mixture Models

- Probabilistic models that use a soft-clustering approach to distribute data points into different clusters.
- The objective of a GMM is to maximise the likelihood value of data *X* which can be formulated as a marginalised property summed up over *G* clusters.
- Mathematically: $p(X_i) = \sum_{g=1}^{G} p(X_i|c_g)p(c_g)$

PRELIMINARY RESULTS ON *SC-PDB*

Method	Protein	Cationic Trypsin	BRD-4	CDK2	Estrogen Receptor	HIV-1 Protease	Prothrombin
DBSCAN	Cationic Trypsin	-	0.04	0.07	0.03	0.03	0.08
	BRD-4	0.03	-	0.08	0.11	0.05	0.13
	CDK2	0.09	0.13	-	0.06	0.11	0.18
	Estrogen Receptor	0.08	0.18	0.04	-	0.12	0.08
	HIV-1 Protease	0.17	0.01	0.19	0.18	-	0.31
	Prothrombin	0.16	0.05	0.24	0.11	0.09	-
Spectral Clustering	Cationic Trypsin	-	0.06	0.17	0.13	0.08	0.20
	BRD-4	0.15	-	0.18	0.09	0.12	0.06
	CDK2	0.14	0.15	-	0.07	0.14	0.05
	Estrogen Receptor	0.02	0.18	0.16	-	0.03	0.06
	HIV-1 Protease	0.15	0.02	0.08	0.07	-	0.16
	Prothrombin	0.11	0.11	0.13	0.22	0.05	-
Dynamic kNN	Cationic Trypsin	-	0.04	0.15	0.11	0.27	0.19
	BRD-4	0.19	-	0.35	0.14	0.08	0.13
	CDK2	0.16	0.19	-	0.15	0.19	0.22
	Estrogen Receptor	0.10	0.18	0.22	-	0.18	0.09
	HIV-1 Protease	0.15	0.15	0.21	0.17	-	0.24
	Prothrombin	0.19	0.28	0.31	0.23	0.17	-
GMM	Cationic Trypsin	-	0.04	0.10	0.08	0.16	0.09
	BRD-4	0.07	-	0.17	0.14	0.11	0.03
	CDK2	0.01	0.08	-	0.04	0.13	0.19
	Estrogen Receptor	0.11	0.18	0.09	-	0.11	0.07
	HIV-1 Protease	0.05	0.11	0.17	0.06	-	0.02
	Prothrombin	0.18	0.24	0.07	0.04	0.09	-

- DkNNs display higher values compared to other methods and can be considered an appropriate method for mitigating noise introduced by fpocket cavity detection.

PRELIMINARY RESULTS ON *TRACKML*

Method	p_t^{min}	Truth Efficiency	Edge Efficiency	Number of Nodes	Number of Edges
DBSCAN	0.5	0.972 ± 0.03	0.046 ± 0.00	1.41×10^5	2.35×10^6
	0.6	0.974 ± 0.05	0.051 ± 0.00	1.02×10^5	1.80×10^6
	0.75	0.979 ± 0.09	0.091 ± 0.01	8.68×10^4	1.17×10^6
	1.0	0.981 ± 0.10	0.14 ± 0.03	5.93×10^4	7.99×10^5
	1.5	0.983 ± 0.14	0.21 ± 0.05	2.19×10^4	9.27×10^4
	2.0	0.982 ± 0.15	0.25 ± 0.09	1.02×10^4	5.09×10^4
Spectral Clustering	0.5	0.968 ± 0.003	0.042 ± 0.01	1.59×10^5	2.382×10^5
	0.6	0.972 ± 0.004	0.108 ± 0.03	8.268×10^4	9.358×10^4
	0.75	0.979 ± 0.004	0.108 ± 0.03	7.97×10^4	9.27×10^4
	1.0	0.981 ± 0.005	0.180 ± 0.13	6.351×10^3	7.591×10^4
	1.5	0.981 ± 0.006	0.399 ± 0.17	3.742×10^4	4.260×10^4
	2.0	0.981 ± 0.008	0.719 ± 0.15	1.924×10^4	1.834×10^4
Dynamic kNN	0.5	0.972 ± 0.003	0.028 ± 0.00	1.621×10^5	3.209×10^5
	0.6	0.974 ± 0.003	0.074 ± 0.00	8.491×10^4	5.372×10^5
	0.75	0.977 ± 0.005	0.081 ± 0.00	8.699×10^4	9.973×10^4
	1.0	0.979 ± 0.005	0.119 ± 0.01	7.372×10^4	6.138×10^4
	1.5	0.983 ± 0.008	0.253 ± 0.03	4.206×10^4	4.528×10^4
	2.0	0.984 ± 0.010	0.375 ± 0.04	2.519×10^4	1.972×10^4
GMM	0.5	0.966 ± 0.002	0.115 ± 0.00	8.491×10^4	5.372×10^5
	0.6	0.974 ± 0.003	0.151 ± 0.01	7.916×10^4	4.684×10^5
	0.75	0.979 ± 0.006	0.207 ± 0.02	6.582×10^4	2.461×10^5
	1.0	0.982 ± 0.006	0.207 ± 0.03	5.218×10^4	9.207×10^4
	1.5	0.983 ± 0.008	0.361 ± 0.05	2.625×10^4	6.948×10^4
	2.0	0.982 ± 0.010	0.452 ± 0.05	9.350×10^3	9.958×10^3

- GMMs generally achieve high combined Truth Efficiency and Edge Efficiency across the full range of p_t^{min}.
- The raw input of TrackML can be broken down into distinct subgraphs to simplify the tasks of downstream track-finding.
- This will also be helpful in accelerating the training of graph-based deep learning architectures on distributed systems.

FURTHER RESULTS

We validate previous results by looking at the labels of individual particles within each cluster.

Dataset	Method	$e_{TrackML} \uparrow$	$e_{sc-PDB} \uparrow$	$\chi_{TrackML} \uparrow$	$\chi_{sc-PDB} \uparrow$
DBSCAN	TrackML	0.579	-	0.7424	-
	sc-PDB	-	0.481	-	0.2863
Spectral Clustering	TrackML	0.602	-	0.5968	-
	sc-PDB	-	0.517	-	0.4262
Dynamic kNN	TrackML	0.513	-	0.5079	-
	sc-PDB	-	0.594	-	0.5038
GMM	TrackML	0.735	-	0.8194	-
	sc-PDB	-	0.408	-	0.3920

CONCLUSION

We take a look at how different types of graph segmentation approaches work on scientific datasets and how they could be used as a precursor for deep learning pipelines with graph-based data. We conduct comprehensive evaluations over two scientific datasets used in separate fields and show how graph segmentation would be able to point towards factors that would inevitably help speed-up or improve the accuracy of the overall pipeline it is fitted into.

Contact:
rajat.sahay2018@vitstudent.ac.in

This work is supported by IRIS-HEP through the U.S. National Science Foundation (NSF) under Cooperative Agreement OAC-1836650.