A GRAPH CONSTRUCTION STUDY FOR GRAPH BASED SEMI SUPERVISED LEARNING

Sumedh Yadav , Gautam Kumar, Shivam Kumar*[1]
Gstech Tech. Pvt. Ltd., Bengaluru, India

INTRODUCTION

Graph-based SSL algorithms are traditionally popular among graph-structured datasets, such as online social networks et cetera (Subramanya et al. 2014), however, researchers and practitioners are beginning to realize that graph can be used to represent data organically (Wu et al. 2018). Although extensive research has been done in the classification sub-problem of the graph-based SSL techniques (Peel 2016, Talukdar et al. 2009), creation of knowledge graph sub-problem is the area of current interest in the machine learning community. Distance metric such as cosine similarity (CS) is shown to be effective for quantifying text-similarity in the case of unstructured data, which can be used to construct a graph. Particularly, state-of-the-art metric improved sqrt-cosine (ISC) similarity is shown to be effective (Sohangir et al. 2017) for text-based data. We present a study on how the distance/similarity metrics impact the graph construction and the subsequent classification task in cases of unstructured text data.

***Keywords** - graph construction, similarity metrics, unstructured text data, graph-based SSL, graph spectrum, transductive learning*

THE MATHEMATICS

The classification based on the graph based SSL has two major components :

- Graph construction
- Label propagation

In Graph construction following distance metrics are used for finding approximate nearest neighhbors (Using FLANN) :

- Euclidean L2 norm,

$$\text{Euclidean}(x, y) = \sqrt{\sum_{i}^{n} (x_i - y_i)^2}$$

- Cosine similarity,

$$\text{CS}(x, y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}$$

- Improved sqrt-cosine similarity,

$$\text{ISC}(x, y) = \frac{\sum_{i=1}^{n} \sqrt{x_i y_i}}{\sqrt{\sum_{i=1}^{n} x_i} \sqrt{\sum_{i=1}^{n} y_i}}$$

where x and y are datapoints, n is the number of features.

The similarity metric differs from euclidean in the sense that the similarity metric uses a pivot to find the distance between two points.

In graph construction approximate nearest neighbor (ANN) search is used to find the nearest neighbors and distance metrics were used to weigh it as shown in Algorithm 1. Fast library for ANN (FLANN) was used, and the metrics CS and ISC were implemented in it.

Algorithm 1: Graph construction

Input: $G(V, \phi, \phi)$, nn - number of neighbors, $metric$ - L2/CS/ISC
Output: $G(V, E, W)$

```
1  neigh_list[] ← 0
2  indices[][] ← 0
3  dists[][] ← 0.0
4  flann() ← V                                   ▷ FLANN instance with all nodes
5  flann.index()
6  indices, dists ← flann.ann_search(nn, metric)                ▷ ANN search
7  for i ∈ V do
8      for j ∈ indices[i] do
9          if j ∉ neigh_list[i] then
10             neigh_list[i] += j                ▷ add j to neighbor list of i
11             W_{i,j} = dists[i][j]
12         end
13         if i ∉ neigh_list[j] then
14             neigh_list[j] += i                            ▷ undirected graph
15             W_{j,i} = dists[i][j]
16         end
17     end
18 end
```

The output weighted graph is used as an input to the LP_ZGL inference algorithm (Label propagation algorithm) (Xiaojin Z et al. 2002) for which the Junto label propagation toolkit (Talukdar et al. 2010) is used.

DATASET

The first dataset, Dataset I, is a E-com. data in which various sources of product information are available, as shown in Figure 1.

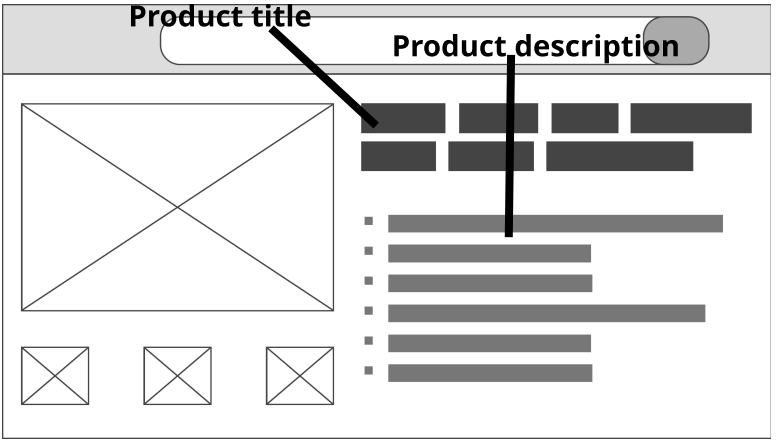


Figure 1 A typical product page on a E-commerce platform has plenty of text data.

For Step 1 of Figure 2, title and description are combined to form a document. Table 1 shows rest of the parameters.

Table 1 Dataset paramters.

Parameter	Value	
Dataset	E-commerce	20 newsgroups
# classes	2	4
# instances	19 224	3 905
Variance	0.004	0.09

A number of feature engineering steps, as shown in Figure 2, were involved in preparing the feature vectors. For example, for Dataset II, number of features were reduced by almost 36% after the application of the cleaning tool.

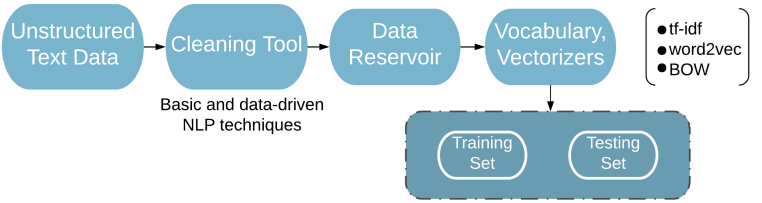


Figure 2 Feature engineering pipeline includes data cleaning and the vectorisation before graph construction.

The second dataset, Dataset II, is the 20 newsgroups data, which is popular for experiments in text applications of machine learning techniques, such as text classification and text clustering.

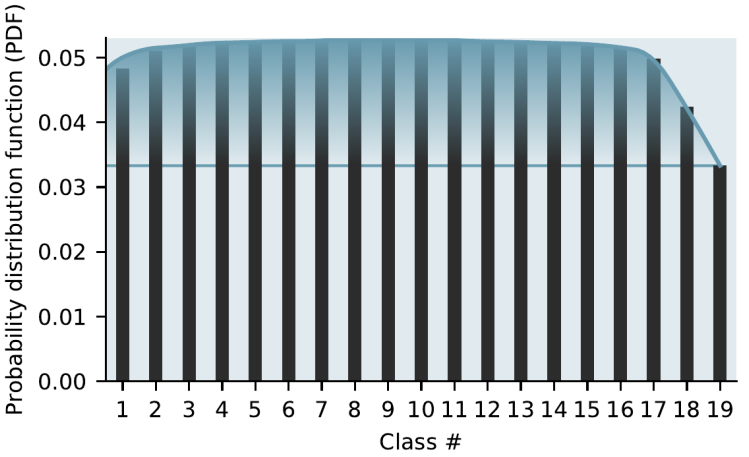


Figure 3 PDF of the classes for 20 newspaper group.

Figure 3 shows the probability distribution function (PDF) of the 20 classes, which is very uniform, and 4 target classes are chosen. Table 1 shows rest of the parameters.

TESTS

Distance metrics are compared in two major ways,

- Quality of graph construction. That is, in Algorithm 1, the number of neighbors were tracked, which was used to report the two variables,
  - Average node degree (AND) of the graph,
  - Standard deviation of the node degree (SDND) of the graph.
- Prediction accuracy (PA) on the testing set which measures the quality of label propagation.

Testing parameters are further summarized in Table 2, in which it is observed that testing/training ratio is kept high. This can be done because of a spectral transformation of the graph using a function such as gaussian field kernel (GFK), which leads to smoothening of the edge weights across nodes, as shown in Figure 4.

Prediction accuracies were averaged over 7 runs of the inference algorithm. Lastly, the partial truth table, Table 3, was used to choose the four most confused classes in Dataset II.

Table 2 Testing parameter

Parameter	Value
Distance metric	ISC, CS, L2
# neighbors or nn	4
Junto inference algorithm	lp_zgl
Testing/training ratio	10:1 to 40:1
datapoints fraction	0.01 to 1.0

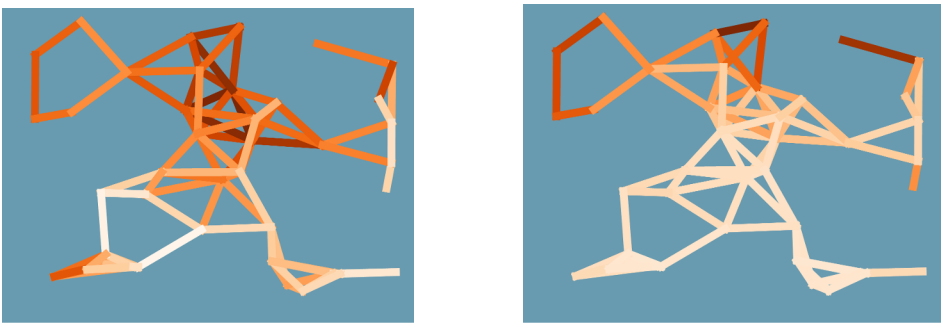
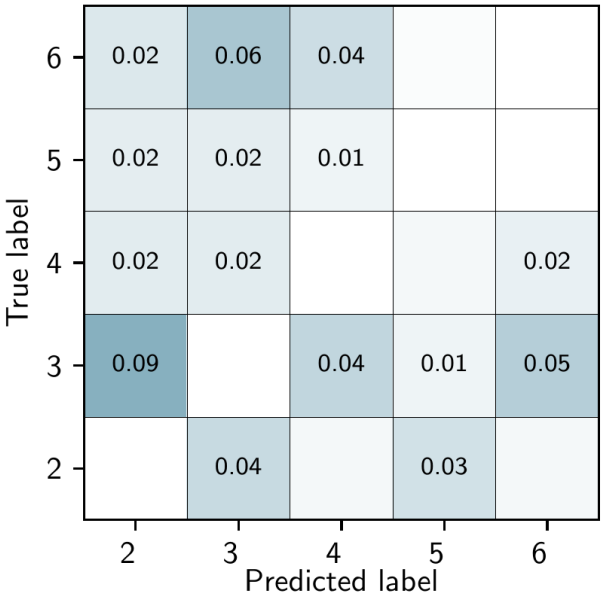


Figure 4 Subfigure on the left is the original eigen matrix (similar to W) of the graph, and on the right is after applying the spectral transformation of GFK.

Table 3 Partial truth table for Dataset II.



FINDINGS

First, for Dataset I, Table 4 presents the results. It can be seen that the inference algorithm crashes for the fractions 0.50 and 0.75 with L2. For fraction 0.25, Figure 5 emphasizes the findings, particularly AND and SDND are seen to be significantly lower for CS and ISC compared to L2. Furthermore, ISC performs decisively better than CS with AND = 6.09 being closest to $nn = 4$, and half the deviation.

Table 4 Results on Dataset I, vectorizer was tf-idf.

Parameter	Value											
Distance metric	L2				CS				ISC			
data points fraction	0.10	0.25	0.50	0.75	0.10	0.25	0.50	0.75	0.10	0.25	0.50	0.75
Average degree of graph	7.60	7.43	-	-	6.24	6.21	6.13	6.11	6.12	6.09	6.10	6.10
Standard deviation	98.95	121.21	-	-	18.14	24.51	17.13	15.33	9.05	9.49	8.58	8.36
Accuracy	92.93	97.91	-	-	95.48	96.30	97.22	94.63	96.94	97.61	98.09	98.21

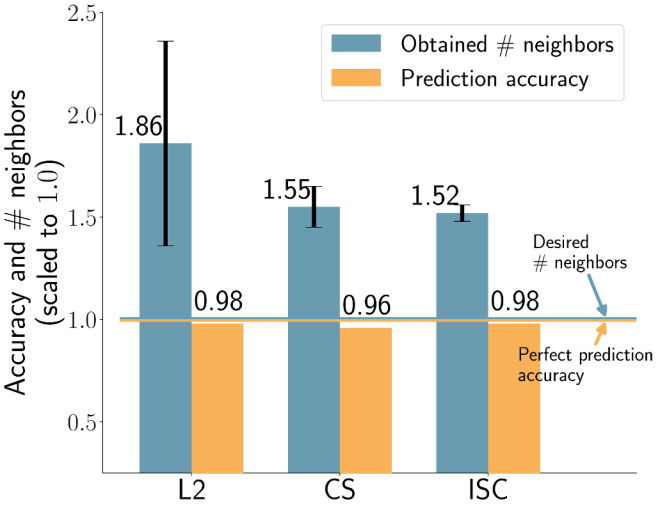


Figure 5 Test was conducted on quarter of Dataset I and, ISC performs best in graph construction.

Second, for Dataset II, Table 5 presents the results. Observations follow the previous case, and a significant trend in PA is observed, where ISC > CS > L2, as can be seen in Figure 6.

Table 5 Results on Dataset II, vectorizer was tf-idf.

Parameter	Value											
Distance metric	L2				CS				ISC			
datapoints fraction	0.25	0.50	0.75	1.00	0.25	0.50	0.75	1.00	0.25	0.50	0.75	1.00
Avg. degree	7.73	7.82	7.81	7.83	6.55	6.34	6.22	6.10	6.40	6.25	6.08	5.99
SD	102.57	114.257	102.74	126.96	34.23	57.51	61.88	33.85	37.52	32.60	28.56	22.11
Accuracy	44.83	48.24	53.28	53.20	53.75	62.61	63.67	65.73	59.93	69.06	71.53	73.98

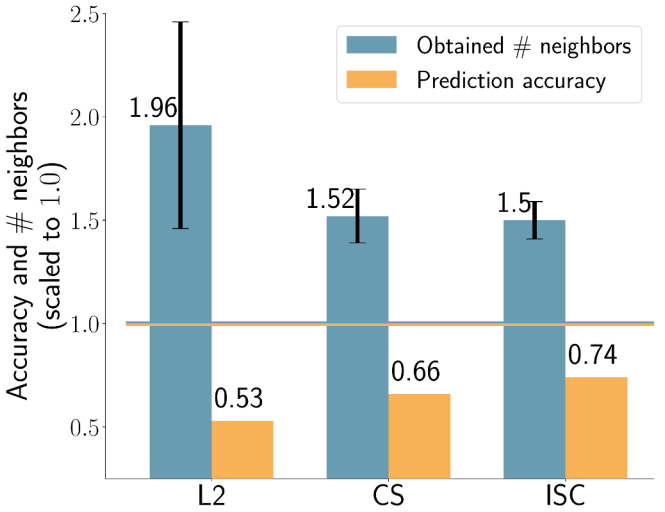


Figure 6 On full Dataset II, a clear performance trend of ISC > CS> L2 is observed in graph construction and prediction accuracy.

Overall, ISC performs the best in the quality of graph construction, and the subsequent classification/inference, followed by CS, however, L2 is evidently not comparable to the other two similarity metrics for the graph construction on text data.

REFERENCES

- Subramanya A, Talukdar PP (2014) Graph-based semi-supervised learning. Morgan & Claypool Publishers. url:https://books.google.co.in/books?id=fzKNBQAAQBAJ.
- Wu X, Zhao L, Akoglu L (2018) A quest for structure: Jointly learning the graph structure and semi-supervised classification. Proceedings of the 27th ACM international conference on information and knowledge management. doi:10.1145/3269206.3271692.
- Peel L (2016) Graph-based semi-supervised learning for relational networks. arXiv. url:https://arxiv.org/abs/1612.05001.
- Sohangir S, Wang D (2017) Improved sqrt-cosine similarity measurement. Journal of Big Data. doi:0.1186/s40537-017-0083-6.
- Muja M, Lowe DG (2009) Fast approximate nearest neighbors with automatic algorithm configuration. url:https://www.cs.ubc.ca/research/flann/.
- Xiaojin Z, Zoubin G (2002) Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University.
- Talukdar PP, Pereira F (2010) Experiments in graph-based semi-supervised learning methods for class-instance acquisition. ACL. https://github.com/parthatalukdar/junto. Accessed 15 August 2019.

DATASET AND MATERIALS

Dataset I is available at doi:10.5281/zenodo.3355823, and Dataset II at url:http://qwone.com/~jason/20Newsgroups/.

Tests were performed with the following computing specifications,

- Operating system: Ubuntu 16.04 LTS
- Programming Language: C/C++, Java, Python
- External Libraries: FLANN (1.8.4), Junto, Keras (2.2.4)

Resources including codes are available at the project webpage.