Nonnegative Sparse and KNN graph for semi-supervised learning

Yunbin ZHANG[1], Chunmei ZHANG[2, a], Qianqi ZHOU[3]

[1] College of Computer science and engineering, Beifang University of Nationalities, Yinchuan, 750021, China

[2] College of Computer science and engineering, Beifang University of Nationalities, Yinchuan, 750021, China

[3] College of Computer science and engineering, Beifang University of Nationalities, Yinchuan, 750021, China

[a]email: chunmei66@hotmail.com

Keywords: Sparse graph; KNN graph; NSKNN-graph; semi-supervised learning

Abstract. For the graph-based semi-supervised learning, the performance of a classifier is very sensitive to the structure of the graph. So constructing a good graph to represent data, a proper structure for the graph is quite critical. This paper proposes a novel model to construct the graph structure for semi-supervised learning. In this new structure, the weights of edges in the graph are obtained by the linear combination of a Nonnegative Sparse graph and K Nearest Neighbour graph (NSKNN-graph). The NSKNN-graph can capture both the global structure (by global sparse graph) and the local structure (by the KNN graph). We demonstrate the effectiveness of NSKNN-graph on the UCI dataset. Experiments show that the NSKNN-graph has advantages over graphs constructed by conventional methods.

Introduction

The semi-supervised learning problem has attracted an increasing amount of interest recently. Many new mathematical methods have been introduced into this research field, such as minimum cut[1] 、 Gaussian fields[2] and spectral graph theory[3]. Conventionally, the graph-based semi-supervised learning [4], directly or indirectly making use of manifold hypothesis, first constructs a graph structure with all the training instances (marked or unmarked) as the nodes of the graph and the similarity of the nodes as (weighted) edges of the graph. Then, the objective functions are defined and optimized. Eventually the decision functions as the regularization are smoothed on the graph to obtain the optimal model parameters.

The graph-based semi-supervised learning has been applied in many fields, but how to construct effective structures for the graph with semi-supervised learning is still an open topic. Conceptually, a good graph should reveal the intrinsic complexity of the data (through local linear relationships), and also capture the whole structures of the data globally. Traditional methods (such as K nearest neighbors) mainly rely on pair-wise Euclidean distances and construct the graph by a family of overlapped local patches. Such kinds of graphs only capture local structures without figuring the whole structures of data globally.

Liu et al. [5] constructed a low rank graph. Although the graph can capture the global structure of data, still ignores the local details of the data structure. In addition, because of its coefficients can be negative, such will make the data offset each other by subtracting. Yan et al. [6] proposed to construct a sparse graph by solving a l^1 norm optimization problem. Such kinds of sparse graphs can capture the global structures of data by choosing the sparsest representation for each sample from all the linear combinations of other samples.

To improve the structure of the graph, we propose to construct a composite graph, in which the weights of edges are obtained by the linear combination of a Nonnegative Sparse graph and a K nearest neighbor graph (NSKNN-graph). The NSKNN-graphs can reveal the intrinsic complexity of the data with capturing certain global structures by the sparse representation and capturing local

structures by the KNN. We select two popular graph-based semi-supervised learning frameworks, Gaussian Harmonic Function (GHF) and Local and Global Consistency (LGC), to compare the effectiveness of different graphs, which include NSKNN-graph, low rank graph, KNN graph, sparse graph and exp-weight graph. We conducted experiments on UCI dataset. Experiments show that the NSKNN-graph can significantly improve the performance of semi-supervised learning. These results clearly demonstrate that NSKNN-graph can reveal the true intrinsic complexity of the data.

NSKNN-graph Construction

Let $X = [x_1, x_2, ..., x_n] \in \mathrm{R}^{m \times n}$ be a matrix whose columns are n data samples. Then each column can be represented by a linear combination of a basis A:

$$x_i = As_i \tag{1}$$

where is the overcomplete dictionary with atoms of a_i, $S = [s_1, s_2, ..., s_n] \in \mathrm{R}^{n \times n}$ is the coefficient matrix for decomposing $X = [x_1, x_2, ...x_n]$ with A. We can get the sparest solution to each $x_i = As_i$ by solving the following optimization problem:

$$\min_{s_i} \|s_i\|_1, \ s.t. \ x = As_i, s_i \geq 0 \tag{2}$$

where $\|\cdot\|_1$ denotes l^1 norm.

In this way, a non-negative sparse coefficient matrix $S = [s_1, s_2, ..., s_n]^T$ is constructed.

For any two nodes x_i and x_j, let $K_j(i)(i = 1, 2, .., k, i \neq j)$ denote a k neighbor set of the node x_j, KNN method usually chooses the distance as its measurement constraint. Firstly we calculate the distance between a node x_i and all other nodes, then select k nearest nodes x_i which belong to the k neighbor set $K_j(i)$, finally measure whether a node x_i is a k nearest neighbor of the node x_j or not, if x_i belongs to $K_j(i)$, the corresponding weight of the edge x_j and x_j is 1, otherwise 0. The specific processing is as follows.

$$K(i, j) = \begin{cases} 1 & if \ x_i \in K_j(i), i \neq j \\ 0 & otherwise \end{cases} \tag{3}$$

In this way, a matrix $K \in \mathrm{R}^{n \times n}$ is constructed, which is composed of 1 and 0.

A good graph should reveal the intrinsic complexity of the data. To improve the classification accuracy, we propose to construct a composite graph, in which the weights of edges are obtained by the linear combination of a Nonnegative Sparse graph and K Nearest Neighbor graph (NSKNN-graph), as well as a nonnegative constraint. That is, we construct a composite matrix W by solving the following linear problem:

$$W = S + \beta K, \quad \beta \in [0, 1] \tag{4}$$

In this way, a matrix $W \in \mathrm{R}^{n \times n}$ is constructed.

The algorithm of constructing the NSKNN-graph

Given a matrix of data $X = [X_l, X_u] = [x_1, x_2, x_l, x_{l+1}, x_{l+u}] \in \mathrm{R}^{m \times n}$, we may construct an undirected graph $G = (V, E)$, where l is the number of labeled samples and u is the number of unlabeled samples. $V = \{v_{ij}\}$ is the vertex set, each node v_i corresponding to a data point x_i, and $E = \{e_{ij}\}$ is the edge set, each edge e_{ij} associate nodes v_i and v_j. When we assign different weights to edge e_{ij}, we can get different matrixes of weight $W = \{w_{ij}\}$, e.g. weight matrix $S = \{s_{ij}\}$ and weight matrix $K = \{k_{ij}\}$ represent the weights of non-negative sparse coefficient matrix and KNN matrix respectively. Then we add the two matrixes to get a new weight matrix $W = S + \beta K$, with this composited matrix, we can derive the adjacency structure of the graph and its weight matrix. The

steps of constructing an NSKNN-graph are as follows:

Constructing NSKNN-graph algorithm

Input: The sample dataset set denoted as the matrix $X=[X_l, X_u]=[x_1, x_2, x_l, x_{l+1}, x_{l+u}]$, where $x_i \in \mathrm{R}^m$, X_l is the labeled samples set, X_u is the unlabeled samples set.
Output: Graph G.
According to formula (2), calculate the non-negative sparse coefficient matrix S.
According to formula (3), calculate the KNN matrix K.
According to formula (4), calculate the composite matrix W.
Denote $G=(X,W)$

Semi-supervised classification based on NSKNN-graph

In this subsection, we select two popular methods, Gaussian Harmonic Function (GHF) [7] and Local and Global Consistency (LGC) [8], to compare the effectiveness of different graphs. Let $y=\left[y_l, y_u\right]^T \in R^{|V|\times c}$ be a label matrix, where $y_{mn}=1$, if a sample x_i is associated with label m for $n \in \{1,2,\ldots,c\}$ and $y_{mn}=0$ otherwise. The GHF realize the label propagation by learning a classification function $F=\left[F_l, F_u\right]^T \in R^{|V|\times c}$. It utilizes the graph and the known labels to recover the continuous classification function by optimizing the predefined energy functions. GHF combines Gaussian random fields and Harmonic Function for optimizing the following cost on a weighted graph to recover the classifiction function F:

$$\min_{F\in R^{|V|\times c}} tr(F^T LF), \quad LF_u=0, F_l=y_l \tag{5}$$

where $L=D-W$ is the graph Laplacian, in which D is a diagonal matrix with $D_{mn}=\sum_n W_{mn}$. Instead of clamping the classification function on labeled nodes by setting hard constraints $F_l=y_l$, LGC introduces an elastic fitness term as follows:

$$\min_{F\in R^{|V|\times c}} tr\{F^T SF+\mu(F\text{-}Y)^{\mathrm{T}}(F\text{-}Y)\} \tag{6}$$

where $\mu \in [0,1]$ is a parameter which balances the tradeoff between the local fitting and global smoothness of function F, S is the normalized Laplacian of the graph, $S=D^{-1/2}WD^{-1/2}$. In our experiment, we simply fix $\mu=0.99$. The steps of semi-supervised classification based on NSKNN-graph are as follows:

NSKNN-graph algorithm based on GHF

Input: the composite matrix W.
Output: unlabeled samples label matrix y_u.
Calculate diagonal matrix D, where $D_{mn}=\sum_n W_{mn}$, and then calculate the non-standardized graph Laplacian $L=D-W$.
According to formula (5), calculate the classification function $F=[F_l, F_u]$.
Calculate the unlabeled samples label matrix $y_i=\arg\max_{j\le c} F_u(l+1\le i\le l+u)$.

NSKNN-graph algorithm based on LGC

Input: the composite matrix W.
Output: unlabeled samples label matrix y_u.
Calculate diagonal matrix D, where $D_{mn}=\sum_n W_{mn}$, and then calculate the normalized graph Laplacian $S=D^{-1/2}WD^{-1/2}$.
According to formula (6), calculate the classification function $F=[F_l, F_u]$.
Calculate the unlabeled samples label matrix $y_i=\arg\max_{j\le c} F_u(l+1\le i\le l+u)$.

Experiment

We carry out the classification experiments on the three types of UCI dataset and compare the performance of NSKNNG , low rank graph (LRG), KNN graph (KNNG), sparse graph (SG) and exp-weight graph (EWG). Table1 is the basic attributes of the three types of UCI datasets used in our experiments.

Table1 The basic attributes of the three types of UCI dataset

dataset	categories	dimensions	samples
vote	2	16	435
wdbc	2	30	569
iris	3	4	150

As can be seen from the experimental results of Table 2, the classification accuracy is highest when the value of the parameter β is 0.5. So we finally fix $\beta = 0.5$ in all experiments.

Table2 The influence of β value changing on the classification accuracy under different percentages of labeled samples

β	0	0.3	0.5	0.7	1
vote(5%)	70.72	83.09	87.85	84.56	85.50
vote(10%)	72.17	87.72	90.84	88.62	88.90
vote(15%)	78.64	90.27	90.84	89.19	89.54
vote(20%)	86.41	90.23	92.25	90.25	90.23
wdbc(5%)	72.27	80.50	84.25	81.12	83.35
wdbc(10%)	73.71	85.34	87.16	84.33	85.20
wdbc(15%)	75.59	88.56	90.51	88.63	87.35
wdbc(20%)	76.97	89.87	91.97	89.98	90.00
iris(5%)	70..72	88.01	93.62	80.00	88.65
iris(10%)	75.70	95.04	95.26	93.78	93.41
iris(15%)	78.62	94.68	95.40	80.16	92.70
iris(20%)	80.83	94.50	96.08	91.17	95.58

For the three types of UCI dataset, we randomly selects 5%, 10%, 15% and 20% of the total samples as the train samples with the remaining as the test samples. Classification accuracy takes the average accuracy of 10 independent experiments. The percentage of labeled samples ranges from 5% to 20%, instead of ranging from 25% to 40% or 45% to 60%. This is because the goal of semi-supervised learning is to reduce the number of labeled samples. So we are more interested in the performance of semi-supervised learning methods with low labeling percentages. The classification results are reported in Table3 and Table4. From these results, we can see that NSKNN-graph consistently achieves the lowest classification error rate compared to the other graphs, even with low labeling percentages, and still have the validity with different label propagation method. This suggests that NSKNN-graph is good at representing the intrinsic structure of the data and more suitable than other graphs for graph-based semi-supervised learning.

Table3 The classification accuracy of different methods on UCI dataset with the GHF label propagation method, the bold numbers are the highest accuracy .

dataset	NSKNNG	KNNG	SG	EWG	LRRG
vote(5%)	**85.54**	84.97	70.72	80.60	83.09
vote(10%)	**90.84**	87.94	72.17	85.93	85.68
vote(15%)	**90.84**	87.43	78.64	86.16	86.75
vote(20%)	**92.25**	89.5	86.41	87.97	87.33
wdbc(5%)	**84.25**	42.93	72.27	37.22	83.33
wdbc(10%)	**87.16**	46.71	73.71	37.27	86.54
wdbc(15%)	**90.51**	49.40	75.59	37.30	87.66
wdbc(20%)	**91.97**	51.78	76.97	37.28	90.75
iris(5%)	**83.38**	68.50	70.72	66.31	79.79
iris(10%)	**91.20**	76.88	75.70	75.56	80.74
iris(15%)	**94.8**	84.72	78.62	80.16	83.73
iris(20%)	**95.43**	87.72	80.83	82.83	84.01

Table4 The classification accuracy of different methods on UCI dataset with the LGC label propagation method, the bold numbers are the highest accuracy .

dataset	NSKNNG	KNNG	SG	EWG	LRRG
vote(5%)	**86.43**	84.12	73.20	75.12	80.09
vote(10%)	**87.75**	88.02	75.33	80.33	84.59
vote(15%)	**88.65**	88.05	77.98	82.50	87.00
vote(20%)	**90.12**	89.38	85.22	85.44	88.00
wdbc(5%)	**82.54**	47.47	72.27	37.22	83.33
wdbc(10%)	**84.02**	57.49	75.00	53.81	81.44
wdbc(15%)	**90.02**	62.39	76.32	37.27	84.44
wdbc(20%)	**90.85**	64.73	77.05	37.28	86.36
iris(5%)	**90.09**	71.93	75.54	75.93	79.98
iris(10%)	**94.52**	77.68	77.32	77.30	80.52
iris(15%)	**94.52**	86.51	79.05	78.57	85.50
iris(20%)	**95.35**	89.07	81.42	80.55	86.04

Conclusion

This paper proposes a new framework to capture the intrinsic structure of data and construct a novel graph, called the Nonnegative Sparse and KNN graph (NSKNN-graph) for graph-based semi-supervised learning. The weights of edges in the graph are obtained by the linear combination of a Nonnegative Sparse graph and KNN graph. The classification results show that the NSKNN-graph is good at representing the true structure of data and thus is more suitable than other graphs for graph-based semi-supervised learning.

Acknowledgement

In this paper, the research was sponsored by the National Natural Science Foundation of China (No.61461002) and the postgraduate innovation project of Beifang University of Nationalities (ycx1556).

References

[1] Zhu, J. Lafferty. Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning [J]. In The 22nd International Conference on Machine Learning, 2005:1052-1059.

[2] X. Zhu, J. Lafferty etc. Semi-supervised learning: From Gaussian fields to Gaussian Processes(D).Carnegie Mellon University (2003)

[3] T. Joachims. Transductive learning via spectral graph partitioning [C]. In The 20th International Conference on Machine Learning,pp.290-297 (2003)

[4] X. Zhu. Semi-supervised learning with graphs [D]. Carnegie Mellon University (2005).

[5] Liu G, L Z, Yu Y. Robust subspace segmentation by low-rank representation[C]//Proceedings of the 27th international conference on machine learning (ICML-10), 2010:663-70.

[6] B. Cheng, J. Yang etc. Learning with graph for image analysis [J]. IEEE Trans. on ImageProcessing, 2010:858-866.

[7] X. Zhu, Z. Ghahramani etc. Semi-supervised learning using Gaussian Fields and Harmonic Functions [C]. In The 20th International Conference on Machine Learning. pp. 912-919 (2003)

[8] Zhou D, Bousquet O, Lal T N, et al. Learning with local and global consistency[J]. Advances in neural information processing systems, 2004, 16(16): 321-328.