Semi-Supervised Learning via Geodesic Weighted Sparse Representation

Jianqiao WANG†(a), Nonmember, Yuehua LI†(b), Member, Jianfei CHEN†, and Yuanjiang LI††, Nonmembers

SUMMARY The label estimation technique provides a new way to design semi-supervised learning algorithms. If the labels of the unlabeled data can be estimated correctly, the semi-supervised methods can be replaced by the corresponding supervised versions. In this paper, we propose a novel semi-supervised learning algorithm, called Geodesic Weighted Sparse Representation (GWSR), to estimate the labels of the unlabeled data. First, the geodesic distance and geodesic weight are calculated. The geodesic weight is utilized to reconstruct the labeled samples. The Euclidean distance between the reconstructed labeled sample and the unlabeled sample equals the geodesic distance between the original labeled sample and the unlabeled sample. Then, the unlabeled samples are sparsely reconstructed and the sparse reconstruction weight is obtained by minimizing the L1-norm. Finally, the sparse reconstruction weight is utilized to estimate the labels of the unlabeled samples. Experiments on synthetic data and USPS handwritten digit database demonstrate the effectiveness of our method.

key words: label estimate, geodesic distance, sparse reconstruction, semi-supervised learning

1. Introduction

Semi-supervised learning methods use a small number of labels and a large number of samples to perform learning tasks. The key to semi-supervised learning problems is the cluster assumption: (1) nearby samples are likely to have the same label; (2) samples on the same structure are likely to have the same label [1]. According to the cluster assumption, we can estimate the labels of the unlabeled samples directly, rather than learning a classifier. In recent years, several label estimation algorithms have been proposed. According to the implementation, these methods can be classified into iterative methods and non-iterative methods.

Linear Neighborhood Propagation (LNP) [2],[3] is an iterative method. Based on a linear neighborhood model, LNP propagates the labels from the labeled samples to the whole sample set. The propagation is an iterative process. In each iteration, each sample “absorbs” a fraction of the label information from its neighborhood and retains some label information of its initial state. A parameter is defined to indicate how much label information is absorbed. Thus, LNP is a parameter-sensitive method. The selection of the number of nearest neighbors and the absorption fraction affect the convergence speed and the recognition rate.

Nearest Neighbor (NN) can be thought as the simplest non-iterative label estimation method. For each unlabeled sample, NN calculates the Euclidean distance between the unlabeled sample and each labeled sample. The labeled sample that has the shortest distance and the corresponding unlabeled sample belong to the same class. In recent years, with the development of compressed sensing, the sparse reconstruction and sparse coding technique have been studied widely. Sparse Representation (SR) [4],[5] is a sparse coding technique. It represents each unlabeled sample as a linear reconstruction of the labeled samples. The reconstruction coefficient is utilized to reconstruct the labels of the unlabeled samples from the labels of the labeled samples. Such representations are usually sparse and robust to noises.

From the implementations of LNP and SR, we see that both LNP and SR utilize the first cluster assumption: in LNP, neighbor samples are likely to have the same label; while in SR, similar samples in the sense of L1-norm minimization are likely to have the same label. The similarity is approximately measured by the Euclidean distance, rather than the sample structure. In this paper, we propose a novel non-iterative label estimation method, called Geodesic Weighted Sparse Representation (GWSR), to perform semi-supervised learning tasks. Our method utilizes both the first and the second cluster assumption. First, the geodesic distance is calculated for all samples. The geodesic distance contains the structure information, but it cannot be used directly. We need to reconstruct the labeled samples by using the geodesic weight. Then, the unlabeled samples are sparsely reconstructed by the geodesic weighted labeled samples. The sparse reconstruction weight is obtained by minimizing the L1-norm. Finally, the labels of the unlabeled samples are estimated by the sparse reconstruction weight.

2. Geodesic Weighted Sparse Representation

2.1 The Label Estimation Problem in Nonlinear Space

As shown in Fig. 1, we have 200 samples. These samples are sampled from the two-moon data set, which contains two classes: the top moon and the bottom moon. The two red dots in the top moon and the two blue dots in the bottom moon are labeled. We want to estimate the label of sample E from sample A, B, C, and D. Most of the existing methods, such as NN and SR, utilize the Euclidean distance to search the nearby samples. However, it does not work well in the
nonlinear space. For convenience, we define the Euclidean distance between sample \( x_i \) and \( x_j \) as \( d_E(x_i, x_j) \). It is easy to see that \( d_E(A, A) < d_E(E, C) < d_E(E, B) < d_E(E, D) \). That means the contribution of sample C is larger than that of sample B when we estimate the label of sample E. Obviously, this is not reasonable. Moreover, the contribution of sample B is overestimated. Since sample A is close to sample E, the similarity between A and E can be characterized by \( d_E(E, A) \). While sample B is far away from sample E, the distance that is used to measure the similarity is larger than \( d_E(E, B) \). Thus, the similarity between B and E can not be characterized by \( d_E(E, B) \) any more.

The geodesic distance indicates the shortest path between two samples on the manifold. It can be used to measure the distance in the nonlinear space if the samples are sufficiently sampled. A successful application of the geodesic distance is Isomap [6], which is a nonlinear dimensionality reduction algorithm. Similar with \( d_E(x_i, x_j) \), we define the geodesic distance between sample \( x_i \) and \( x_j \) as \( d_G(x_i, x_j) \). Given a data set \( X = \{x_1, \ldots, x_n\} \), the geodesic distance can be obtained by the following two steps.

1) Construct the neighborhood graph: Define the graph \( G \) over all samples by connecting samples \( x_i \) and \( x_j \) if \( x_i \) is one of the \( k \) nearest neighbors of \( x_j \). Set edge lengths equal to \( d_E(x_i, x_j) \);

2) Compute the geodesic distance: Initialize \( d_G(x_i, x_j) = d_E(x_i, x_j) \) if \( x_i \) and \( x_j \) are linked by an edge. \( d_G(x_i, x_j) = \infty \) otherwise. Then for each value of \( t = 1, \ldots, n \) in turn, replace all entries \( d_G(x_i, x_j) \) by \( \min(\{d_G(x_i, x_k), d_G(x_k, x_i) + d_G(x_i, x_j)\}) \).

In Fig. 1, we see that \( d_G \) describes the similarity between sample B and E better than \( d_E \). Thus, we should utilize \( d_G \) to estimate the label of sample E. However, we can not utilize the geodesic directly. When we reconstruct the unlabeled sample with the labeled samples, what we utilize is the feature vector in each sample, while the geodesic is not a vector. Figure 1 shows a method to overcome this shortcoming. Take sample B for example, we construct another sample \( B' \) which satisfies \( d_E(E, B') = d_G(E, B) \). We name \( B' \) the geodesic weighted sample. In order to calculate conveniently, we locate \( B' \) on the extension cord of EB. The features of \( B' \) can be obtained by a simple geometric relationship. Assume that there are \( m \) features in each sample. The \( t \)th feature of \( B' \) can be expressed as:

\[
B'_t = \frac{d_G(E, B)}{d_E(E, B)} B_t + (1 - \frac{d_G(E, B)}{d_E(E, B)}) E_t
\]

For sample C, which locates on the different manifold of sample E, the geodesic distance \( d_G(E, C) \) is set to \( \infty \). For the stabilization of the numerical, we set

\[
d_G(E, C) = \alpha \max(\{d_G(E, P)\})
\]

where sample \( P \) and E are on the same manifold, \( \alpha \) is a parameter that adjust the distance between the manifolds. Similar with Eq. (1), the geodesic weighted sample \( C' \) and \( D' \) can be obtained by the same way.

2.2 Geodesic Weighted Sparse Representation

Assume that we have \( n \) labeled samples and \( m \) unlabeled samples. These samples belong to \( c \) classes and the class label of the \( i \)th sample is denoted by \( l_{ti}(t_i = 1, \ldots, c) \). The goal of GWSR is to estimate the \( m \) unknown labels from the \( n \) given labels. We utilize \( x^L = [x^L_1, \ldots, x^L_n] \) and \( X^U = [x^U_1, \ldots, x^U_m] \) to indicate the labeled sample matrix and unlabeled sample matrix, respectively. First, we calculate the geodesic distance \( d_G \) for all samples and construct the geodesic weighted labeled matrix. For the \( j \)th unlabeled sample \( x_j^U \), we define the geodesic weighted labeled matrix as \( X^GL = [x^{GL}_1, \ldots, x^{GL}_n] \). The superscript \( GL \) means the sample is geodesic weighted and the corresponding label is known. The subscript \( ij \) (\( i = 1, \ldots, n \)) means the sample has the same label with \( x^L_i \) and the geodesic weight depends on \( x^L_i \) and \( x_j^U \). The \( p \)th coordinate of \( x^{GL}_{ij} \) is calculated as follow:

\[
x^{GL}_{ij,p} = \frac{d_G(x^L_i, x_j^U)}{d_E(x^L_i, x_j^U)} x^{L}_{ji,p} + (1 - \frac{d_G(x^L_i, x_j^U)}{d_E(x^L_i, x_j^U)}) x^U_{ij,p}
\]

where \( x^{L}_{ji,p} \) indicates the \( p \)th coordinate of \( x^L_i \) and \( x^{U}_{ij,p} \) indicates the \( p \)th coordinate of \( x_j^U \). Then, we utilize \( X^{GL}_U \) as a dictionary [7] to reconstruct \( X^U \). For the unlabeled sample \( x^U_j \), we reconstruct it by solving the following L1-norm minimization problem:

\[
\begin{align}
\min & \|w_j\|_1 \\
\text{s.t.} & \|X^{GL}_j w_j - x_j\|_2 \leq \varepsilon \\
& \sum_p w_{jp} = 1
\end{align}
\]

where \( w_j \) is the reconstruction coefficient, \( w_{jp} \) is the \( p \)th coordinate of \( w_j \), \( \varepsilon \) is the reconstruction error which is usually set very small. In this paper, we set \( \varepsilon = 0.01 \). Since \( x^U_j \) is linearly reconstructed from the geodesic weighted labeled samples with weight \( w_j \), the label of \( x^U_j \) can also be reconstructed by the given labels with the same reconstruction weight. We utilize \( f_j = [f_{j1}, \ldots, f_{jc}] \) to indicate the class information of \( x^U_j \), where \( f_{jq} \) indicates the probability of \( x_j \) belonging to the \( q \)th class. \( f_{jq} \) can be calculated as follow:

\[
f_{jq} = \sum_p w_{jp} \delta(l_j - q)
\]

Fig. 1 The label estimation problem in two-moon data set. Sample A, B, C and D are labeled. The label of sample E is what we want to obtain. B’ is the reconstructed sample of B and it satisfies \( d_E(E, B') = d_G(E, B) \).
Table 1  GWSR algorithm.

| Input:  | the labeled sample matrix $X^L$ and the labels, the unlabeled sample matrix $X^U$, the number of nearest neighbor $k$, the parameter $\alpha = 100$ and $\varepsilon = 0.01$. |
| Output: | the labels of unlabeled samples. |

Algorithm:

1: Calculate the geodesic distance $d_G$ for all samples;
2: For each unlabeled sample $x^U_j$, calculate the geodesic weighted labeled matrix $X^{GL}_j$;
3: Calculate the reconstruction weight $w_j$ by using Eq. (4);
4: Calculate the class probability $f_j$ by using Eq. (5);
5: Search the largest $f_j q$, then we set the label of unlabeled sample $x^U_j$ as $q$.

where $\delta()$ is the discrete unit impulse function, and defined as:

$$
\delta(l_j - q) = \begin{cases} 
1, & l_j = q \\
0, & l_j \neq q 
\end{cases}
$$

We set $\max\{f_j 1, \ldots, f_j k\} = 1$ and the rest to 0. If $f_j q = 1$, $x^U_j$ belongs to the $q$th class. We summarize the GWSR algorithm in Table 1.

3. Experimental Results

We perform experiments on two-moon data set and USPS hand-written digit database. To demonstrate the effectiveness of our proposed algorithm, we compare our GWSR with NN, SR and USSL [8].

3.1 Two-moon data set

Most of the real world data sets are high dimensional and not suitable for visualization. To give an intuition and visible result, we perform experiment on two-moon data set. Each moon has 100 samples. One sample in each moon is labeled and labels of the rest samples will be estimated. The labeled samples are indicated by circular rings. The number of nearest neighbor $k$ is set to 9. The results of the label estimation by NN, SR, USSL and GWSR are shown in Fig. 2 (a), (b), (c) and (d), respectively. From the results, we see that our method recognizes the two moons correctly, while NN, SR and USSL only recognize a part of the moons.

3.2 Digit Recognition

USPS handwritten digits data set contains 8-bit gray-scale images of “0” through “9”. The size of each image is $16 \times 16$ pixels. Thus, each digit image is represented as a 256-dimensional vector. We use three subsets to test the algorithms. These three subsets are defined as USPS1 = {1, 3, 7}, USPS2 = {0, 6, 9}, USPS3 = {4, 6, 8, 9}, respectively. For each digit, 200 samples are selected randomly. 100 tests are performed and the average recognition rates are compared. Since our method is parameter-sensitive, different parameters (the number of nearest neighbor $k$) lead to different results. First, we examine the impact of $k$. The recognition rates on different $k$ are shown in Fig. 3.

From Fig. 3, we see that not every value of $k$ is suitable. When $k$ is smaller than 8, the performance is unstable. Especially when $k = 4$, the performance is the worst. When $k$ is larger than 8, the performance remains stable and recognition rates are relatively high. Thus, a suitable $k$ for digit recognition can be selected from 8 to 20.

Next, we compare our method with NN, SR and USSL on the three USPS subsets. We set $k = 17$ in our method. The recognition rates on different number of training samples are shown in Fig. 4. The average recognition rates and the stand deviations are shown in Table 2. From the results, we find that our method always outperforms NN, SR and USSL. As the number of training samples increases, the recognition rates of the three algorithms increase and become stable finally. A significant advantage of our method is that our method has a relatively high recognition rate when there are only a small number of training samples.

3.3 Discussion

The above experiments show that our method outperforms NN, SR and USSL. NN and SR only utilize the neighborhood information but ignore the structure information. USSL is a semi-supervised method and it utilizes the unlabeled sample to estimate the labels. However, it is a linear
Fig. 4 Performance comparison of recognition rate on (a) USPS1, (b) USPS2 and (c) USPS3.

Table 2 Average recognition rate and standard deviations (percent) on USPS data set.

| Data   | nn   | SR   | USSL  | GWSR  |
|--------|------|------|-------|-------|
| USPS1  | 95.93±0.91 | 97.25±0.85 | 96.91±1.86 | 98.25±2.32 |
| USPS2  | 94.81±1.41 | 95.78±1.57 | 95.75±1.93 | 97.26±1.55 |
| USPS3  | 86.54±1.96 | 86.10±2.44 | 86.03±2.81 | 92.94±2.53 |

approach in essence. As the number of class increases, the performance of USSL decreases. Our method utilizes the geodesic distance to represent the nonlinear sample structure and the neighborhood information is contained in the geodesic distance. Moreover, we reconstruct the labeled samples by the geodesic weight to utilize the Euclidean distance directly. This is equivalent to transform the nonlinear structure to the linear structure. However, in order to transform the nonlinear structure to the linear structure, we need a large number of samples for each class. This is because the transformation is based on the geodesic distance, while the geodesic distance can be accurately calculated only when a large number samples are given. In addition, our method is not suitable for data set which describes the attributes of an object, such as UCI data set. This is because the distance between these attributes cannot represent the distance between the objects themselves. This result has been demonstrated by experiment. However, owning to the limitation of the space, we do not present here.

4. Conclusion

In this paper, we proposed a novel semi-supervised learning method, called GWSR. It utilizes both local property and global structure information to estimate the labels of the unlabeled samples. Since the geodesic distance represents the similarity between samples better than the Euclidean distance in the nonlinear space, we first use the geodesic distance to reconstruct the labeled samples. Then, we use the geodesic weighted labeled samples to linearly and sparsely reconstruct the unlabeled samples. The reconstructed weight is used to estimate the labels. Experiments on two-moon data set and USPS data set show that our algorithm yields better performance than NN, SR and USSL.

Acknowledgements

This work is supported by National (Jiangsu) Natural Science Foundation of China under Grants 60901008 and 61001010 (BK2010490), National Ministry Foundation of China under Grants 51305050102.

References

[1] O. Chapelle, J. Weston, and B. Schölkopf, “Cluster kernels for semi-supervised learning,” Advances in NIPS 15, 2002.
[2] F. Wang and C. Zhang, “Label propagation through linear neighborhoods,” IEEE Trans. Knowl. Data Eng., vol.20, no.1, pp.55–67, 2008.
[3] J. Wang, F. Wang, C. Zhang, H.C. Shen, and L. Quan, “Linear neighborhood propagation and its applications,” IEEE Trans. Pattern Anal. Mach. Intell., vol.31, no.9, pp.1600–1615, 2009.
[4] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” IEEE Trans. Pattern Anal. Mach. Intell., vol.31, no.2, pp.210–227, 2009.
[5] J. Lu, X. Zhou, Y.P. Tan, Y. Shang, and J. Zhou, “Cost-sensitive semi-supervised discriminant analysis for face recognition,” IEEE Trans. Inf. Forensics Security, vol.7, no.3, pp.944–953, 2012.
[6] J.B. Tenenbaum, V. Silva, and J.C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” Science, vol.290, no.5500, pp.2319–2323, 2000.
[7] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” IEEE Trans. Pattern Anal. Mach. Intell., vol.35, no.8, pp.1798–1828, 2013.
[8] K. Huang, Z. Xu, I. King, and M.R. Lyu, “Semi-supervised learning form general unlabeled data,” Proc. ICDM, pp.273–282, 2008.