Iterative Semi Supervised Data Denoising with Procrustes Analysis

Abhishek RSI2016006@IIITA.AC.IN
Priyanshu Goyal PRIYANSHU.GOYAL92@GMAIL.COM
Shekhar Verma SVERMA@IIITA.AC.IN
Indian Institute of Information Technology, Deoghat, Jhalwa, Allahabad, 211012.

Abstract

A wireless sensor network localization with only few location aware nodes is difficult due to noisy medium and other environmental effects. This situation is similar to semi supervised learning where in the given data set, a small portion is labeled while majority remains unlabeled and the aim is to find unknown labels based on available information. This is achieved by exploiting the underlying intrinsic geometrical manifold structure between data including unlabeled data. In this paper we propose iterative graph Laplacian least square regression with Procrustes analysis which ensures error minimization in localization through iterative manifold structure learning. The results indicate that high localization accuracy is achieved as compared to other similar technique.

1. Introduction

In a wireless sensor network, nodes need to localize themselves using a few location aware nodes before commencement of the sensing operation. The localization process can be modeled as follows. Given n data with $\{x_i, y_i\}_{i=1}^{n}$ where x_i is independent variable and y_i is its respective label, noise removal is easy. A data set with few m labeled and $\{x_j\}_{j=m+1}^{n}$, $\{n-m \gg m\}$ unlabeled data, noise identification and removal is difficult. Semi supervised manifold learning (Chapelle et al., 2006) for error minimization has been previously proposed in (Liu et al., 2014; Pan & Yang, 2007; Biswas et al., 2006; Chen et al., 2011). The basic idea behind semi supervised learning is to efficiently utilize those unlabeled data by exploiting the intrinsic geometrical manifold structure. In the underlying algorithm, a prediction function based on labeled data cost function (such as square loss) along with intrinsic graph structure learning is derived which is further used to find missing labels. By using the labeled data only to compute prediction function leads to underutilization, moreover with single regression the error boundary remains equal to original noisy data error margin.

In (Liu et al., 2014) authors solved the problem of impulse noise in image by exploiting the natural image property of local smoothness and global structural similarity. A graph is created with adjoining pixels as vertex and edge weight containing similarity value between them. MDS-MAP as described in (Shang et al., 2003) illustrates localization which uses multidimensional scaling on the available nodes' distance within the range. However MDS-MAP is centralized algorithm which leads to scaling problem and performance degrades in irregular node deployment. MDS-MAP(P) was proposed in (Shang & Ruml, 2004), which divides the original network graph into several local graphs called patches. It solves the scaling problem but depends on the correctness of node distances. For higher utilization of unlabeled samples efficiently, a family of learning algorithms exploiting geometry of marginal distribution has been proposed in (Belkin et al., 2006). A regularization is performed using cost function along with manifold learning. A graph is created with data points as vertex, edge weight calculated through kernel function. By calculating graph Laplacian over the weight and the kernel matrix with appropriate kernel used, the QP coefficient is calculated which is finally combined with kernel to solve the convex function. The method performs well with irregular data models.

In this paper, we propose an error minimization algorithm Iterative Semi Supervised Data Denoising with Procrustes analysis (IS^2D^2P) which not only utilizes labeled data more than once for noise reduction but unlabeled data learning is also done iteratively.The pa-

Preliminary work. Under review by the Indian Workshop on Machine Learning (iWML). Do not distribute.

per has been organized with section 1 containing introduction to the problem and existing solutions. Section 2 states the theory behind proposed method along with algorithm. In section 3 we verify our proposed algorithm on wireless sensor node localization which is followed by section 4 concluding the findings of this paper.

2. Proposed Technique

In the proposed **IS^{2}D^{2}P** algorithm, we minimize the noise and find labels in the given data by iterative graph Laplacian regularization with lease squares regression. To undo the transformation from previous step, we employ Procrustes analysis to obtain denoised data.

Consider a one dimensional regression problem which is to learn a function $f : x \to y$ where x denotes the input space and y denotes the label. Here we are concerned with planar localization, therefore $x, y \subseteq \mathbb{R}^2$ and our aim is to find function f which can reestimate the non-anchor position. Given n nodes with few $\{x_i, y_i\}_{i=1}^m$ anchor nodes and $\{x_j\}_{j=m+1}^n$ non-anchor nodes where $m \ll (n-m)$, the prediction function can be trained using those m samples by (Belkin et al., 2006)

$$f^* = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} \| y_i - f(x_i) \|^2 + \lambda \| f \|^2 \quad (1)$$

By creating prediction function based on only m anchors will not lead to noise minimization but we also have to use those $n-m$ unlabeled data. In order to efficiently include them for prediction function, we will apply the well known manifold assumption on the graph structure. Here all $\{x_{i=1}^n\}$ become vertex and distance between them is calculated to draw edges. Graph Laplacian on previously created graph gives us the intrinsic geometrical structure and we will include this in our prediction function for assumed manifold learning. The manifold assumption is given by (Belkin et al., 2006)

$$R(f) = \frac{1}{2} \sum_{i=j=1}^{n} (f(x_i) - f(x_j))^2 \mathcal{W}_{ij} \quad (2)$$

where (Liu et al., 2014),

$$\mathcal{W}_{ij} = \frac{1}{C} exp\left\{-\frac{\|x_i - x_j\|^2}{\varepsilon^2}\right\} exp\left\{-\frac{\|b_i - b_j\|^2}{\varepsilon^2}\right\}$$

$\| x_i - x_j \|^2$ data distance, $\| b_i - b_j \|^2$ average distance of immediate neighbors, and $\varepsilon > 0$

Laplacian is given by $\mathcal{L} = \mathcal{D} - \mathcal{W}$ where $\mathcal{D}_{ii} = \sum_{j=1}^n \mathcal{W}_{ij}$, expanding eqn 2 and substituting with $\mathcal{L}$ gives

$$\begin{aligned} R(f) &= \sum_{i=1}^{n} f(x_i^2) \sum_{j=1}^{n} \mathcal{W}_{ij} - \sum_{i=j=1}^{n} \mathcal{W}_{ij} f(x_i) f(x_j) \\ &= f^T \mathcal{D} f - f^T \mathcal{W} f \\ &= f^T \mathcal{L} f \end{aligned} \quad (3)$$

On including eqn 3 in eqn 1, it becomes

$$f^* = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} \| y_i - f(x_i) \|^2 + \lambda \| f \|^2 + \gamma \| f^T \mathcal{L} f \| \quad (4)$$

The extended Representer Theorem states that optimal f exists in $\mathcal{H}_K$ and is given by

$$\begin{aligned} f(x) &= \sum_{i=1}^{n} \alpha_i \kappa(x_i, x) \\ \Rightarrow f &= [f(x_1), f(x_2), \ldots, f(x_n)]^T \\ &= \left[\sum_{i=1}^{n} \alpha_i \kappa(x_i, x_1), \ldots, \sum_{i=1}^{n} \alpha_i \kappa(x_i, x_n) \right]^T \\ f &= \kappa \alpha \end{aligned} \quad (5)$$

Here κ denotes $n \times n$ Kernel gram matrix $\kappa_{ij} = \kappa(x_i, x_j)$ and α is representation coefficient matrix. Replacing f from eqn 5 in eqn 4 we get

$$f^* = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \| y_m - \kappa_m \alpha \|^2 + \lambda \alpha^T \kappa \alpha + \gamma \alpha^T \kappa \mathcal{L} \kappa \alpha \quad (6)$$

To obtain optimal solution, set partial derivative with α on eqn 6

$$\begin{aligned} \frac{\partial f}{\partial \alpha} &= 0 \\ \Rightarrow \alpha^* &= \left(\kappa_m \kappa_m^T + \lambda \kappa + \gamma \kappa \mathcal{L} \kappa\right)^{-1} \kappa_m y_m \end{aligned} \quad (7)$$

On substituting eqn 7 in eqn 5, it gives the final prediction output. κ is calculated using geodesic similarity distance matrix (Pan & Yang, 2007) $\mathcal{G}$ which is obtained by

$$\begin{aligned} d_{ij} &= \sqrt{\| x_i - x_j \|^2} \\ \mathcal{G}_{ij} &= \min\left\{d_{ij}, d_{ik} + d_{kj}\right\} \\ \kappa(x_i, x_j) &= exp\left\{-\frac{\mathcal{G}^2(x_i, x_j)}{\sigma^2}\right\} \end{aligned} \quad (8)$$

Wireless sensor node communication can be viewed as spherical shape which helps us forming the needed

Algorithm 1 IS^{2}D^{2}P

Input: n data with $(x_i, y_i)_{i=1}^{m}$ labeled and $(x_j)_{j=m+1}^{n}$ unlabeled

Calculate $adj(x_i, x_j) = \begin{cases} 1, \, if \, i, \, j \, are \, neighbor \\ 0, \, else \end{cases}$

Set $pre\_error = \infty$, $now\_error = rms(x)$

while $now\_error < pre\_error$ **do**

  Set $\mathcal{W} = distance(x)$

  Set $\mathcal{G} = geodesic(\mathcal{W})$

  Set $\kappa = rbf(\mathcal{G})$

  Set $\mathcal{D}_{ii} = \sum_{j=1}^{n} \mathcal{W}_{ij}$, $\mathcal{L} = \mathcal{D} - \mathcal{W}$

  Calculate α from eqn 7

  Calculate $f_{n\times 2}$ from eqn 5

  Set $pre\_error = now\_error$, $now\_error = rms(f)$

end while

Set $\mathcal{Z}_{n\times 2} = procrustes(f)$

Output: $\mathcal{Z}$

manifold structure, hence in order to learn it efficiently we use an explicit RBF kernel which is isotropic in nature and has a spherical symmetry. Above prediction function is executed in multiple iterations as shown in eqn 9 for l^{th} instance by calculating distance on previous approximate position which proportionally reduces error on each data.

$$d_{ij}^{(l)} = \sqrt{\| f_i^{(l-1)} - f_j^{(l-1)} \|^2}$$

$$f^{(l)} = \kappa^{(l)} \alpha^{(l)} \tag{9}$$

We know the true and estimated labels of labeled data. Procrustes analysis is done to determine the affine transformations required to map estimated labels to true labels. The same transformation is applied to estimated labels of unlabeled data as neighboring unlabeled data are affected by similar noise and undergo same transformation. This further reduces the error.

3. Simulation and Results

Table 1 contains all the simulation parameters we have used, apart from 36 anchor nodes we have added gaussian noise with $\sigma = 0.7$ to all other 364 nodes' position representing the error during RSSI based localization due to wireless medium and environmental conditions. Among the four figures, fig. 1 shows 400 original node positions deployed in $10\times10 squnits$ area and the localized erroneous positions. Difference between fig. 1 and fig. 2 shows that the error boundary gets reduced from outside the deployment area to inside. This happens because we learn an approximate position with first regression and it becomes maximum area of freedom for that node in next regression, this itself minimizes a

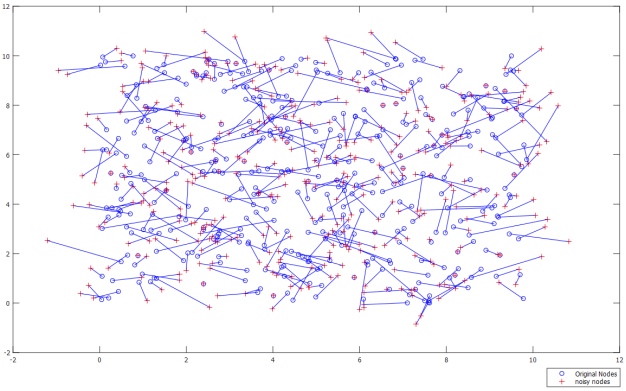

Figure 1. Nodes with original and noisy location

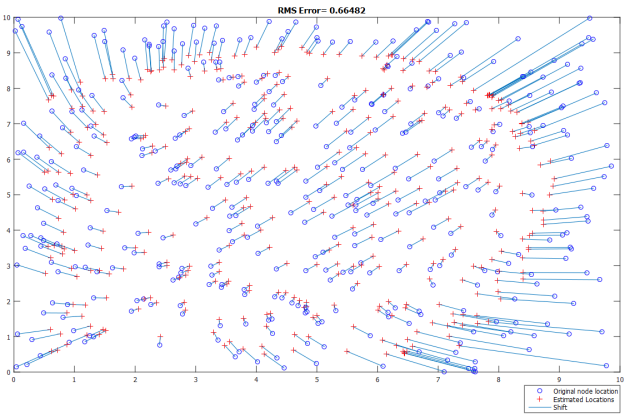

Figure 2. Node location- post first iteration

major error fraction on boundary nodes with low reference nodes. On computing the second iteration as shown in fig. 3 to optimize the defined convex function based on previous predicted location, error is further reduced as verified from the rms error value calculated over the nodes. To efficiently utilize the anchor nodes as well as exploit natural property of uniform error distribution within local subregions, we perform Procrustes analysis on anchor nodes with respect to the prediction we got from second regression. By using this analysis data we can undo the previous transformation. When we apply the obtained anchor node Procrustes among their neighbors, the error is further minimized as seen in fig. 4.

4. Conclusion

In this work we proposed iterative regression feedback algorithm which reduces the error margin in each iteration which in turn minimizes the error in data. Man-

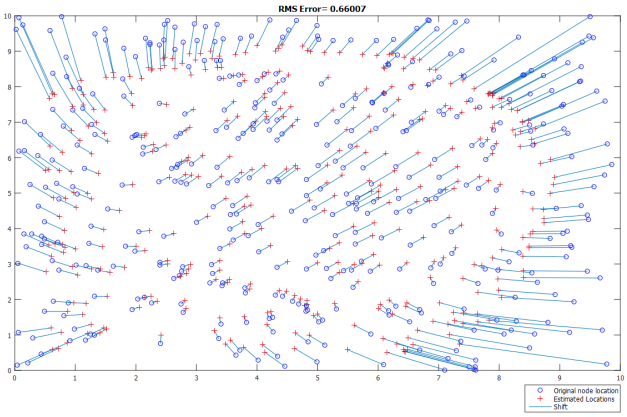

Figure 3. Node location- post second iteration

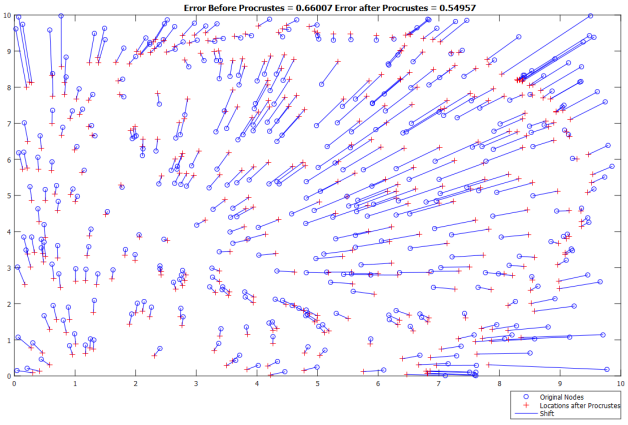

Figure 4. Node location- post Procrustes analysis

ifold regularization was able to exploit the inherent geometrical structure to enhance noise free labeling. In case of 2-D localization, only a few iterations were sufficient to increase position accuracy. The choice of parameters plays a vital role in proposed algorithm's performance. In comparison to existing algorithms where labeled data was used once to learn prediction function, our method was able to use the geometrical structure and achieve high accuracy by canceling out the errors introduced by the method itself.

Table 1. Simulation parameters

Parameter	**Value**
AREA	$10 \times 10\, sq\, units$
n	400
m	36
RANGE	$3\, units$
ε^2 IN $\mathcal{W}$	0.1
C IN $\mathcal{W}$	2
λ IN α	0.5
γ IN α	0.01
σ^2 IN κ	0.5

References

Belkin, Mikhail, Niyogi, Partha, and Sindhwani, Vikas. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.*, 7:2399–2434, 2006.

Biswas, P., Liang, T. C., Toh, K. C., Ye, Y., and Wang, T. C. Semidefinite programming approaches for sensor network localization with noisy distance measurements. *IEEE Transactions on Automation Science and Engineering*, 3:360–371, 2006.

Chapelle, Olivier, Schlkopf, Bernhard, and Zien, Alexander. *Semi-supervised learning*. Adaptive computation and machine learning. 2006.

Chen, Jiming, Wang, Chengqun, Sun, Youxian, and (Sherman) Shen, Xuemin. Semi-supervised laplacian regularized least squares algorithm for localization in wireless sensor networks. *Computer Networks*, 55:2481–2491, 2011.

Liu, Xianming, Zhai, Deming, Zhao, Debin, Zhai, Guangtao, and Gao, Wen. Progressive image denoising through hybrid graph laplacian regularization: A unified framework. *IEEE Trans. Image Processing*, 23:1491–1503, 2014.

Pan, Jeffrey Junfeng and Yang, Qiang. Co-localization from labeled and unlabeled data using graph laplacian. In *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*, pp. 2166–2171, 2007.

Shang, Yi and Ruml, W. Improved mds-based localization. volume 4, pp. 2640–2651, 2004.

Shang, Yi, Ruml, Wheeler, Zhang, Ying, and Fromherz, Markus P. J. Localization from mere connectivity. In *Proceedings of the 4th ACM International Symposium on Mobile Ad Hoc Networking &Amp; Computing*, MobiHoc '03, pp. 201–212, 2003.