

Manifold Coarse Graining for Online Semi-supervised Learning

Mehrdad Farajtabar, Amirreza Shaban,
Hamid Reza Rabiee, and Mohammad Hossein Rohban

Digital Media Lab, AICTC Research Center,
Department of Computer Engineering,
Sharif University of Technology, Tehran, Iran
{farajtabar, shaban, rahban}@ce.sharif.edu,
rabiee@sharif.edu

Abstract. When the number of labeled data is not sufficient, Semi-Supervised Learning (SSL) methods utilize unlabeled data to enhance classification. Recently, many SSL methods have been developed based on the manifold assumption in a batch mode. However, when data arrive sequentially and in large quantities, both computation and storage limitations become a bottleneck. In this paper, we present a new semi-supervised coarse graining (CG) algorithm to reduce the required number of data points for preserving the manifold structure. First, an equivalent formulation of Label Propagation (LP) is derived. Then a novel spectral view of the Harmonic Solution (HS) is proposed. Finally an algorithm to reduce the number of data points while preserving the manifold structure is provided and a theoretical analysis on preservation of the LP properties is presented. Experimental results on real world datasets show that the proposed method outperforms the state of the art coarse graining algorithm in different settings.

Keywords: Semi-Supervised Learning, Manifold Assumption, Harmonic Solution, Label Propagation, Spectral Coarse Graining, Online Classification.

1 Introduction

Semi-supervised learning is a topic of recent research that effectively addresses the problem of limited data [1]. In order to use unlabeled data in the learning process efficiently, certain assumptions on the relation between the possible labeling functions and the underlying geometry should hold [2]. In many real world classification problems, data points lie on a low dimensional manifold. The manifold assumption states that the labeling function varies smoothly with respect to underlying manifold [3]. Manifold structure is modeled by the neighborhood graph of the data points. SSL methods with manifold assumption prove to be effective in many applications including image segmentation[4], handwritten digit recognition and text classification [5].

Online classification of data is required in common applications such as object tracking [6], face recognition in surveillance systems [11], and image retrieval [7].

Usually unlabeled data is easily available in such classification problems. However, most of the classic SSL algorithms are not efficient in an online classification setting. This is due to repeated invocations of a computationally demanding label inference algorithm which takes a time of $O(n^3)$ in standard implementations. Moreover, when the number of arrived data grows large, space complexity becomes an important issue. Consequently, designing efficient label prediction algorithms for the online setting is essential.

Recently online manifold classification algorithms have been proposed to address these challenges. The manifold regularized passive-aggressive online algorithm [8] uses a smoothness regularization term on the τ most recent data in order to reduce the number of samples needed to be stored and processed. This method fails when the windowed data are not representative of the true underlying manifold. This case may happen when data arrive in a biased order. Authors in [9] use RPtree [10] to partition the graph into clusters which grow incrementally in size and cover the manifold structure. Despite promising experimental results, no theoretical guaranty is provided on the error bound for this method.

An state of the art method aimed at reducing the size of data and coarsening the graph is proposed in [11]. Coarsening is done by replacing neighboring points in Euclidean space with fixed number of centroids. Experiments show that considering geodesic distances on manifold results in more accurate data reduction. Authors in [12] propose a data reduction method based on a mathematical framework with an interesting upper bound on the eigenvector distortion after every coarsening of data. However, minimizing this upper bound is hard to tackle. They use a variation of k -means to minimize this bound which is prone to local minima. In addition, a drawback of their method is its prior determination of the number of new nodes, while it is better to concede this decision to the manifold structure itself. To the best of our knowledge none of the previous works in this area take advantage of labeled data to reduce the size of the graph, i.e. coarsening of the graph is performed independent of the given classification task.

Our method like some recent works on online manifold learning [7,9,11], rely on data reduction to overcome memory and time limitations. In this paper we propose a semi-supervised data reduction method that not only captures the geometric structure of data, but also considers the labeled data as a cue to better preserve the classification accuracy. Spectral decomposition is used to find similar nodes on the manifold in order to be merged. Assuming a maximum buffer length of k , we do data reduction whenever this limit is reached, So the time to predict the label of each newly arrived data will not exceed $O(k^3)$. Moreover, the complexity of our CG method is equivalent to the complexity of eigenvector decomposition which similarly takes a time of $O(k^3)$ and will be done just when the buffer limit is reached. As a result overall time complexity is constant with time.

The rest of the paper is organized as follows. In Section 2 basics of HS and LP are briefly introduced. A new formulation of LP and its spectral counterpart is derived in Section 3. Section 4 is the core of this paper where we introduce

coarse graining in exact and approximate modes and explain how it helps us to preserve LP and manifold structure while reducing the number of data points. In Section 5 experimental results are provided, after which the paper is concluded in Section 6.

2 Basics and Notations

Let $\mathcal{X}_u = \{x_1, \dots, x_u\}$ and $\mathcal{X}_l = \{x_{u+1}, \dots, x_{u+l}\}$ be sets of unlabeled and labeled data points respectively, where $n = u + l$ is the total number of data points. Also let $y = (y(u+1), \dots, y(u+l))^\top$ be the vector labels on \mathcal{X}_l . Our goal is to predict labels of $\mathcal{X} = \mathcal{X}_u \cup \mathcal{X}_l$ as $f = (f_u^\top; f_l^\top) = (f(1), \dots, f(u), f(u+1), \dots, f(u+l))^\top$, where $f(i)$ is the label associated to x_i for $i = 1, \dots, n$.

Let W be the weight matrix of the k -NN graph of \mathcal{X} ,

$$W(i, j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

where σ is the bandwidth parameter. Define the diagonal matrix D with nonzero entries $D(i, i) = \sum_{j=1}^n W(i, j)$. Thus the laplacian matrix $L_{un} = D - W$. Manifold regularization algorithms minimize the smoothness functional

$$\mathcal{S}(f) = \frac{1}{2} \sum_{i,j} W(i, j)(f(i) - f(j))^2 = f^T L_{un} f \quad (1)$$

under some appropriate criteria[3,13,14]. Minimizing $f^T L_{un} f$ such that $f_l = y$ is called Harmonic Solution for manifold regularization problem[14].

Label Propagation [15] is a way for computing HS. In this algorithm labels are propagated from labeled to unlabeled nodes through edges in an iterative manner. Edges with larger weights propagate labels easier. In each step a new label is computed for each node as a weighted average of its neighboring labels. The stochastic matrix P is defined such that

$$P(i, j) = \frac{W(i, j)}{\sum_{k=1}^n W(i, k)}. \quad (2)$$

$P(i, j)$ can be interpreted as the effect of $f(j)$ on $f(i)$. The algorithm is stated as follows:

1. Propagation: $f^{(t+1)} \leftarrow Pf^{(t)}$
2. Clamping: $f_l = y$

Where $f^{(t)}$ is the estimated label at step t . If we decompose W and P according to labeled and unlabeled parts,

$$W = \begin{bmatrix} W_{uu} & W_{ul} \\ W_{lu} & W_{ll} \end{bmatrix} \quad P = \begin{bmatrix} P_{uu} & P_{ul} \\ P_{lu} & P_{ll} \end{bmatrix}, \quad (3)$$

then under appropriate conditions [15], the solution of LP converges to the HS and is independent of the initial value (i.e. $f^{(0)}$) and may be written as

$$f_u = (I - P_{uu})^{-1} P_{uly} \quad f_l = y. \quad (4)$$

3 Spectral View of Label Propagation

In this section the LP solution is derived in terms of the spectral decomposition of a variation of the stochastic matrix, P . This helps us find a spectral property of the stochastic matrix, the invariance of which will guarantee that the solution of LP remains approximately constant throughout CG.

Consider the process of propagating labels. Each new label is computed as the weighted average of its neighboring labels. However, for a labeled node the process is undone by clamping its label to the true initial value.

These two steps for labeled nodes may be integrated in one step. For a labeled node i , we remove $P(i, j)$ for all j s and set $P(i, i) = 1$. This causes LP to have an update rule like $f_l^{(t+1)}(i) = f_l^{(t)}(i)$ for labeled nodes. Using this updated stochastic matrix, we can remove the clamping procedure and state the entire process in a coherent fashion.

We mimic the effect of this new process using a new stochastic matrix denoted by Q , which we call the *absorbing stochastic matrix*:

$$Q \triangleq \begin{bmatrix} P_{uu} & P_{ul} \\ 0 & I \end{bmatrix} \quad (5)$$

With the absorbing stochastic matrix the entire process of LP may be rewritten as

$$f^{t+1} = Qf^t, \quad (6)$$

where the initial value is $f^{(0)} = (f_u^{(0)}; y)$, and $f_u^{(0)}$ may be arbitrary. In this new formulation estimated labels are computed as $\lim_{n \rightarrow \infty} Q^n f^{(0)}$. Defining Q^∞ as

$$Q^\infty \triangleq \lim_{n \rightarrow \infty} Q^n,$$

we can write $(f_u; f_l) = Q^\infty f^{(0)}$. Since the result is independent of initial states of unlabeled data, $f_u(j)$ can be rewritten as

$$f_u(j) = \sum_{k=u+1}^{l+u} Q^\infty(j, k)y(k). \quad (7)$$

We wish to relate $Q^\infty(j, k)$ to the right eigenvectors of Q ; to this end we need the following two lemmas.

Lemma 1. *The matrix Q defined in (5) has following properties:*

- Every eigenvalue λ is real and $|\lambda| \leq 1$
- Dimension of the eigenspace corresponding to $\lambda = 1$ is equal to the number of labeled data l .
- Rows of

$$[\mathbf{0}_{l \times u} \ \mathbf{I}_{l \times l}]$$

are the left eigenvectors of Q corresponding to $\lambda = 1$.

Proof. The eigenvalues of Q are roots of the characteristic polynomial i.e. $p(\lambda) = \det(Q - \lambda I) = 0$. Considering the special form of Q ,

$$p(\lambda) = (1 - \lambda)^l \det(P_{uu} - \lambda I)$$

the magnitude of all eigenvalues of P_{uu} is less than one, due to the fact that $P_{uu}^n \rightarrow 0$ as $n \rightarrow \infty$ [15]. Therefore, $\lambda = 1$ has multiplicity l and the magnitude of all other eigenvalues of Q is less than one and real. It is straightforward to show that eigenvalues of a stochastic matrix and the new variation are all real.

For the last part, it can be verified that

$$\begin{bmatrix} \mathbf{0}_{l \times u} & \mathbf{I}_{l \times l} \end{bmatrix} \times \begin{bmatrix} P_{uu} & P_{ul} \\ 0 & I \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{l \times u} & \mathbf{I}_{l \times l} \end{bmatrix}.$$

Therefore, rows of $\begin{bmatrix} \mathbf{0}_{l \times u} & \mathbf{I}_{l \times l} \end{bmatrix}$ are the left eigenvectors of Q associated to $\lambda = 1$.

Definition 1. From now we refer to eigenvectors corresponding to eigenvalues equal to one as unitary eigenvectors, which is different from unit eigenvectors that have unit norm.

Lemma 2. (Spectral decomposition)[16] Every squared matrix A of dimension n with n independent eigenvectors could be decomposed as

$$A = V_R D V_L^T$$

and

$$V_L^T V_R = I,$$

where D is the diagonal matrix of eigenvalues, columns of V_R and V_L are the right and left eigenvectors of A , respectively.

Corollary 1. By unfolding above decomposition we get another expression for spectral decomposition as

$$A = \sum_{i=1}^n \lambda_i p_i u_i^T,$$

where λ_i , p_i and u_i are the i^{th} eigenvalue, right eigenvector and left eigenvector respectively.

Now we are ready to prove the main result of this part.

Theorem 1. $Q^\infty(j, k) = p_k(j)$, for $u + 1 \leq k \leq l + u$ and $1 < j < n$, where $p_k(j)$ denotes element j of the k^{th} right eigenvector which is unitary.

Proof. By Lemma 2 we can write $Q = V_R D V_L^T$. Since $V_L^T V_R = I$, It's easily seen that $Q^n = V_R D^n V_L^T$ or equivalently $Q^n = \sum_{i=1}^n \lambda_i^n p_i u_i^T$. So as $n \rightarrow \infty$ all eigenvectors with eigenvalue less than one disappear and the unitary eigenvalues and eigenvectors remain:

$$Q^\infty = \sum_{i=u+1}^{l+u} p_i u_i^T.$$

By Lemma 1 the left eigenvector u_i can be represented as a vector of zeros with the exception of the i^{th} element equal to one for $u+1 \leq i \leq l+u$. Therefore $Q^\infty(.,k)$ is constructed with u_k and all other u_i s have zero elements in the corresponding places. Consequently $Q^\infty(j,k) = p_k(j)$ for $u+1 \leq k \leq l+u$.

Applying Theorem 1 in equation (7), the final solution of LP is stated as:

$$f_u(j) = \sum_{k=u+1}^{l+u} p_k(j) y(k). \quad (8)$$

Therefore, f_u can be expressed in terms of the right unitary eigenvectors of Q . As a result, f_u remains unchanged if these eigenvectors are preserved in a CG process. This fact will become clear in the next section.

4 Manifold Coarse Graining

In some cases amount of data is so large that storing and manipulating them consumes large memory and imposes high processing cost. We will show in the next subsections that some graph nodes can be merged without seriously affecting LP on the remaining and oncoming data.

4.1 Exact Coarse Graining

Consider the graph in Figure 1 constructed from data. Nodes 1 and 2 have the same neighbors and are both unlabeled. Suppose rows of the absorbing stochastic matrix Q , corresponding to these nodes are the same i.e.

$$\frac{w_{13}}{w_{13} + w_{14}} = q_{13} = q_{23} = \frac{w_{23}}{w_{23} + w_{24}}, \quad \frac{w_{14}}{w_{13} + w_{14}} = q_{14} = q_{24} = \frac{w_{24}}{w_{23} + w_{24}}.$$

Then these two nodes take the same effect from their neighbors in label propagation. Intuitively merging these two nodes should not disturb the process of propagating the labels. After this step, weights should be summed up. This process is illustrated in Figure 1. Node 0 is formed by summing the weights of nodes 1 and 2.

This intuition can be verified analytically. If f and f' are the estimated label functions before and after this merge respectively, then:

Before merge	After merge
$f(1) = q_{13}f(3) + q_{14}f(4)$	$f'(0) = q_{03}f'(3) + q_{04}f'(4)$
$f(2) = q_{23}f(3) + q_{24}f(4)$	$f'(3) = q_{30}f'(0) + \dots$
$f(3) = q_{31}f(1) + q_{32}f(2) + \dots$	$f'(4) = q_{40}f'(0) + \dots$
$f(4) = q_{41}f(1) + q_{42}f(2) + \dots$	

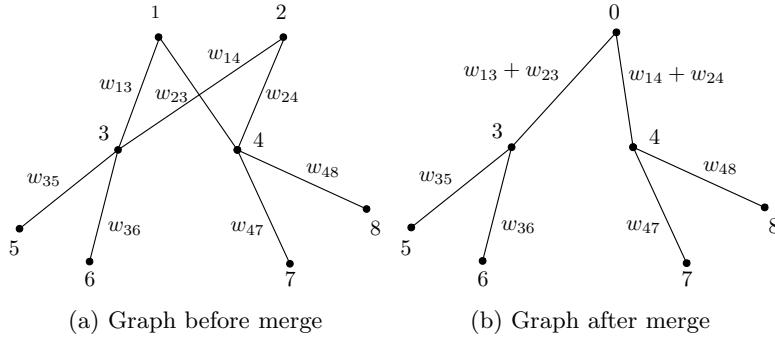


Fig. 1. Merging two vertices 1 and 2 would not disturb label propagation

It is straightforward to see that $q_{03} = q_{13} = q_{23}$ and $q_{04} = q_{14} = q_{24}$, so columns in the first two rows of the above equations are equivalent. Also since after merging we have $q_{31} + q_{32} = q_{30}$ and $q_{41} + q_{42} = q_{40}$ columns of the last two rows impose the same effect on nodes 3 and 4. Thus if nodes 1 and 2 are unlabeled, $f^{(t)}(1) = f^{(t)}(2) = f'^{(t)}(0)$ and $f^{(t)}(3) = f'^{(t)}(3)$ and $f^{(t)}(4) = f'^{(t)}(4)$ in all steps of LP in the original and reduced graph.

This process can be modeled by the transformation $Q' = LQR$ where

$$L = \begin{bmatrix} \frac{d_1}{d_1+d_2} & \frac{d_2}{d_1+d_2} & 0 & \cdots & 0 \\ 0 & 0 & & & \\ \vdots & \vdots & I_{n-2} & & \\ 0 & 0 & & & \end{bmatrix} \quad R = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & I_{n-2} & \\ 0 & & & \end{bmatrix} \quad (9)$$

and $d_i = \sum_j W(i, j)$. One can see that the transformation simply merges rows and columns of Q corresponding to nodes 1 and 2 such that all rows are still normalized. For an undirected graph its stochastic matrix has the property that its first left eigenvector is proportional to $\text{diag}(d_1, \dots, d_n)$. It's easy to see that this is also true for the unlabeled part of the absorbing stochastic matrix Q_{uu} , which can be viewed as the scaled stochastic matrix of an undirected graph. Since the first u elements of the eigenvectors of Q are equal to the eigenvectors of Q_{uu} , this is true for unlabeled nodes. For unlabeled nodes $d_i = u_1(i)$; and only unlabeled nodes are coarsened, so alternatively $u_1(i)$ may be used in (9).

This transformation has interesting properties and is well studied in [17] which presents a similar algorithm based on random walk in social networks. In the general case, R and L can be defined such that the transform merges all the nodes with the same neighbors. There may be more than two nodes that have similar neighbors and can thus be merged.

We will proceed using spectral analysis which will help us in later sections where we introduce non-exact merging of nodes. The next lemma relates spectral analysis to CG.

Lemma 3. *Rows i and j of Q are equal if and only if $p_k(i) = p_k(j)$ for all k with $\lambda_k \neq 0$, where p_k is a right eigenvector of Q .*

Proof. Proof is immediate from the definition of eigenvectors and Corollary 1 of spectral decomposition.

Lemma 3 states that nodes to be merged can be selected via eigenvectors of the absorbing stochastic matrix Q , instead of comparing the rows of Q itself. We decide to merge nodes if their corresponding elements are equal along all the eigenvectors with nonzero eigenvalue. We will see how this spectral view of merging helps us develop and analyze the non-exact case where we merge nodes even if they aren't exactly identical along eigenvectors.

We should fix some notations before we proceed. Superscript “’’ will be used to indicate objects after CG. i.e. Q' , p' , u' , n' are the stochastic matrix, right and left eigenvectors, and number of nodes after CG. Let $S_1, \dots, S_{n'}$ be the n' clusters of nodes found after CG. Also let S be the set of all nodes that are merged into some cluster.

We wish to use ideas from section 3 to provide a spectral view of coarsening stated so far in this section. We need the following lemma.

Lemma 4. [17] *If conditions of Lemma 3 hold Lp is the right unitary eigenvector of Q' with the same eigenvalue as p , where p is the right unitary eigenvector of Q .*

First note that Lp simply retains elements of p which are not merged and removes repetition of the same elements for nodes that are merged. So after CG, the right eigenvectors of Q and associated eigenvalues are preserved. Recall from the previous section that the right eigenvectors are directly related to the result of LP. We are now ready to prove the following theorem.

Theorem 2. *LP solution is preserved for nodes or cluster of nodes in exact CG, i.e. when we merge nodes if their corresponding elements are the same along all right eigenvectors with nonzero eigenvalues.*

Proof. Consider equation (8) from previous section for computing labels based upon right eigenvectors,

$$f_u(j) = \sum_{k=u+1}^{u+l} p_k(j)y(k). \quad (10)$$

We know from Lemma 4, Lp_k is also a right unitary eigenvector of Q' . Suppose j' is the new index of the node or cluster of nodes that node j will reside after CG, similarly

$$f'_u(j') = \sum_{k=u'+1}^{u'+l} (Lp_k)(j')y(k), \quad (11)$$

Considering $(Lp_k)(j') = p_k(j)$ we get the result, $f_u(j) = f'_u(j')$. This means that labels of unlabeled nodes are preserved under CG.

This kind of data reduction will preserve LP results in the manifold of data and as a consequence manifold structure in the reduced graph. This is elaborated upon in the next subsections. Equality along all eigenvectors with nonzero eigenvalues is a restrictive constraint for CG. In the next section we will see how this criterion may be relaxed.

4.2 Approximate Coarse Graining

In real problems the case where neighbors of two or more nodes are exactly the same rarely occurs. Thus the motivation for an approximate coarse graining, i.e. merging nodes when their corresponding elements in eigenvectors are close enough. For example along i^{th} eigenvector we consider two elements approximately the same if their difference is no more than η_i , then merge nodes if they are pairwise approximately the same.

Before we proceed it is beneficial to consider the term $p_i - RLp_i$. Lp_i is the approximate right eigenvector of Q' . Multiplying by R unfolds the clusters. Defining $\varepsilon_i = p_i - RLp_i$, we would like to find an upper bound of ε_i for nodes to be merged. The smaller $\|\varepsilon_i\|$ is, the more similar Lp_i is to p_i . So minimizing ε_i better preserves the i^{th} eigenvector. On the other hand LP results depend on unitary eigenvectors, so a good practice is to do CG on unitary eigenvectors only. In spite of approximately preserving LP this allows more reduction. This approximation will be clearer when error bounds are analyzed.

For simplicity consider node 1 that is placed in a cluster $S_1 = \{1, \dots, m\}$, ($r \leq m$). Using

$$(RLp_i)(1) = \frac{u_1(1)}{\sum_{j=1}^m u_1(j)} p_i(1) + \dots + \frac{u_1(m)}{\sum_{j=1}^m u_1(j)} p_i(m) \quad (12)$$

we may write

$$\begin{aligned} \varepsilon_i(1) &= p_i(1) - (RLp_i)(1) = \\ &= \frac{\sum_{j=2}^m u_1(j)}{\sum_{j=1}^m u_1(j)} p_i(1) - \frac{u_1(2)}{\sum_{j=1}^m u_1(j)} p_i(2) - \dots - \frac{u_1(m)}{\sum_{j=1}^m u_1(j)} p_i(m) = \\ &= \frac{u_1(2)}{\sum_{j=1}^m u_1(j)} (p_i(1) - p_i(2)) + \dots + \frac{u_1(m)}{\sum_{j=1}^m u_1(j)} (p_i(1) - p_i(m)) \leq \quad (13) \\ &= \eta_i \frac{\sum_{j=2}^m u_1(j)}{\sum_{j=1}^m u_1(j)} = \eta_i \left(1 - \frac{u_1(1)}{\sum_{j=1}^m u_1(j)}\right). \end{aligned}$$

The last inequality is due to the fact that in each cluster along the i^{th} eigenvector, differences between elements are no more than η_i . Inequality (13) bounds the difference between elements of eigenvectors corresponding to a node before CG and the desired value after CG. Note that ε_i is zero if CG is exact or for a node that is not merged.

Suppose p' is the true right eigenvector of Q' . However we would like to have Lp as its right eigenvector so as to better preserve the manifold structure and

LP. It is thus interesting to see whether Lp can be approximately considered as an eigenvector of Q' with approximately the same eigenvalue as p . Considering

$$Q'(Lp) = \lambda(Lp) + e, \quad (14)$$

we would like to minimize $\|e\|$. Following [12] we have

$$\sum_{i=1}^{n'-l} \frac{e(i)^2}{u'_1(i)} \leq 2\mathcal{D}, \quad (15)$$

where

$$\mathcal{D} = \sum_{i=1}^k \widehat{\lambda}_i^2 \sum_{j=1}^n u_1(j) \widehat{\varepsilon}_i(j)^2 \quad (16)$$

and $\widehat{\varepsilon}_i = \widehat{p}_i - RL\widehat{p}_i$ where \widehat{p}_i s and $\widehat{\lambda}_i$ s are right eigenvectors and associated eigenvalues that CG is performed along (As this bound hints CG need not be performed along all eigenvectors. We will explain this point shortly). It's noticeable that the bound (15) is a general bound for any coarsening algorithm, It's also originally stated for stochastic matrix of undirected graphs such as P , However as stated in 4.1 the unlabeled part of Q can be considered as such a matrix.

Considering (14) and (15), if $\lambda \gg \sqrt{\mathcal{D}}$ then Lp is a good approximation of p . Given the eigenvectors that must be preserved we can determine how to choose η_i for a good approximation. The inequality $\widehat{\lambda}_l \geq \sqrt{\mathcal{D}}/\omega(1)$ should be satisfied. For example we may seek for sufficient conditions to satisfy

$$\widehat{\lambda}_l \geq \sqrt{\mathcal{D}}/n \quad (17)$$

for every eigenvector \widehat{p}_l that we wish to preserve. Using equation (16) we want to find η_i for all i such that (17) holds.

For simplicity consider cluster $S_1 = \{1, \dots, m\}$. By using inequality (13),

$$\begin{aligned} \sum_{j \in S_1} u_1(j) \widehat{\varepsilon}_i(j)^2 &= \sum_{j \in S_1} u_1(j) \widehat{\eta}_i^2 \left(1 - \frac{u_1(j)}{\sum_{r \in S_1} u_1(r)}\right)^2 = \\ &= \widehat{\eta}_i^2 \sum_{j \in S_1} \left(u_1(j) - 2 \frac{u_1(j)^2}{\sum_{r \in S_1} u_1(r)} + \frac{u_1(j)^3}{(\sum_{r \in S_1} u_1(r))^2}\right) = \\ &= \widehat{\eta}_i^2 \left(\sum_{j \in S_1} u_1(j) - 2 \sum_{j \in S_1} \frac{u_1(j)^2}{\sum_{r \in S_1} u_1(r)} + \sum_{j \in S_1} \frac{u_1(j)^3}{(\sum_{r \in S_1} u_1(r))^2}\right) = \\ &= \widehat{\eta}_i^2 \left(\sum_{j \in S_1} u_1(j) - 2 \sum_{j \in S_1} \frac{u_1(j)^2}{\sum_{r \in S_1} u_1(r)} + \frac{(\sum_{r \in S_1} u_1(r))^3}{(\sum_{r \in S_1} u_1(r))^2} - C\right) \leq \\ &\leq 2\widehat{\eta}_i^2 \sum_{j \in S_1} u_1(j) \end{aligned} \quad (18)$$

Where the last inequality is due to fact that

$$\sum_{j \in S_1} \frac{u_1(j)^2}{\sum_{r \in S_1} u_1(r)} > 0 \quad , \quad C > 0.$$

$\widehat{\varepsilon}_i = 0$ for nodes which are not merged, thus

$$\sum_{j=1}^n u_1(j) \widehat{\varepsilon}_i(j)^2 \leq 2\widehat{\eta}_i^2 \sum_{j \in \mathcal{X}_u} u_1(j) \quad (19)$$

Now we are ready to find an appropriate value for $\widehat{\eta}_i$ to satisfy (17):

$$\mathcal{D} = \sum_{i=1}^k \widehat{\lambda}_i^2 \sum_{j=1}^n u_1(j) \widehat{\varepsilon}_i(j)^2 \leq 2 \sum_{i=1}^k \widehat{\lambda}_i^2 \widehat{\eta}_i^2 \sum_{j \in \mathcal{U}} u_1(j) \quad (20)$$

Let $M = \sum_{j \in \mathcal{U}} u_1(j)$. For $\widehat{\lambda}_l \geq \sqrt{\mathcal{D}}/n$ to be satisfied for every l :

$$2 \sum_{i=1}^k \widehat{\lambda}_i^2 \widehat{\eta}_i^2 M \leq \sqrt{\frac{\widehat{\lambda}_l}{n}} \quad (21)$$

It's easy to verify that choosing η_i such that

$$\widehat{\eta}_i^2 \leq \frac{1}{2kM\widehat{\lambda}_i^2} \sqrt{\frac{\widehat{\lambda}_l}{n}} \quad (22)$$

is true for every l , is sufficient condition that will ensure Lp_l is almost surely preserved, i.e., if $\widehat{\lambda}_{min}$ is the minimum eigenvalue among the eigenvectors that must be preserved, then

$$\widehat{\eta}_i^2 \leq \frac{1}{2kM\widehat{\lambda}_i^2} \sqrt{\frac{\widehat{\lambda}_{min}}{n}}. \quad (23)$$

The bound derived in (23) shows how η_i should be chosen to ensure that Lp_i is similar to a right eigenvector of Q' .

4.3 Preserving Manifold Structure

We have seen how the size of data may be reduced while preserving LP properties. Theorem 1 shows that the LP solution is directly related to unitary eigenvectors of the absorbing stochastic matrix. Thus by CG along these eigenvectors we could retain labels while reducing the amount of data. This process is sufficient to preserve LP but may disturb the true underlying manifold structure. To overcome this we can do CG not only along unitary eigenvectors, but we also along eigenvectors with larger eigenvalues.

To elaborate, note that manifold structure is closely related to the evolution of LP in its early steps, and not just the limiting case where the steps tend to infinity. Consider the one step process of propagating labels, $f^{t+1} = Qf^t$. The more properties of Q is preserved in Q' , the more the underlying structure is retained. Also after k steps of propagations we have $f^{t+k} = Q^k f^t$. Using Corollary 1, we can write $Q^k = \sum_{i=1}^n \lambda_i^k p_i u_i^T$, so as k becomes larger the effect of large eigenvalues and their associated eigenvectors become more important. To preserve LP in early steps it is reasonable to choose eigenvectors with larger eigenvalues and do CG along them. In this manner in addition to LP, the general structure of the manifold is preserved. Figure 2 illustrates the process of CG on a toy dataset with one labeled node from each class. In this figure the general structure of the manifold and its preservation under CG is shown. Also note that sparse section of green nodes is preserved which is essential to capture the manifold structure.

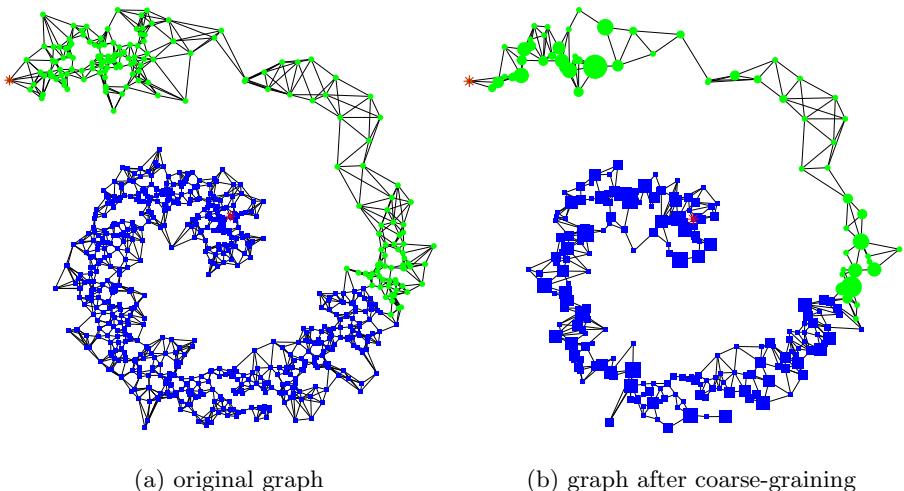


Fig. 2. Process of CG on a toy dataset with 800 nodes. Two labeled nodes are provided on head and tail of the spiral and are red asterisks. Green circle and blue square nodes represent different classes. The area of each circle is proportional to the number of nodes that reside in the corresponding cluster. After CG 255 nodes remain which means a reduction of 68%.

Performing CG along all the eigenvectors should better preserve manifold structure. For merging two nodes this requires that they be close along all the eigenvectors, resulting in less reduction contradicting our initial goal, i.e., data reduction. So in practice generally a few eigenvectors are chosen to be preserved and as we have seen the best choices are the eigenvectors associated to larger eigenvalues. The importance of preserving manifold structure becomes evident when labels are to be predicted for unseen data, e.g., in online learning.

5 Experiments

We evaluate our method empirically on 3 real world datasets: digit, letter and image classification. The first is UCI letter recognition dataset [18]. The next is USPS digit recognition. We reduce the dimension of each data to 64 with PCA. Caltech dataset [19] is used for image classification. Features are extracted using CEDD [20]. Adjacency matrices are constructed using 5-NN with the bandwidth size set to mean of standard deviation of data. 20 data points are labeled. In addition to these 20 unitary eigenvectors 5 other top eigenvectors are selected for CG. η_i is set to divide values along i^{th} eigenvector into I groups, where I is the final parameter that varies to get different reduction sizes. In all experiments on digits and letters the average accuracy among 10 pairwise problems are reported. On Caltech we use 2 specific classes. Four experiments are designed to evaluate our method.

5.1 Eigenvector Preservation

Our CG method captures manifold structure based on eigenvector preservation. To show how well eigenvectors are preserved we compare Lp_i and p'_i for top ten eigenvectors that are to be preserved in USPS dataset. We reduce the number of nodes from 1000 to 92. Table 1 shows eigenvalues and cosine similarity of eigenvectors before and after CG. It is easily seen that eigenvalues and eigenvectors are well preserved. This guarantees a good accuracy of classification after reduction as demonstrated in the next subsections.

Table 1. Eigenvalue and eigenvector preservation in CG for top ten eigenvectors which CG is performed along them

i	1	2	3	4	5	6	7	8	9	10
λ_i	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	-0.9999	0.9999	0.9999	0.9997
λ'_i	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	-0.9999	0.9999	0.9998	0.9997
$\frac{(Lp_i)^\top p'_i}{\ Lp_i\ \ p'_i\ }$	0.9967	0.9925	0.9971	0.9910	0.9982	0.9964	0.9999	0.9909	0.8429	0.9982

5.2 Online Classification

In this experiment, we design a real online scenario where the data buffer size is at most 200 and CG is done when maximum buffer limit is reached. Data arrive sequentially and the label of new data is predicted. Classification result in time t is reported for all data up to this time. We compare our result with the graph quantization method [11] and a baseline method which performs classification without reducing the size. As Figure 3 shows our method is quite effective with a performance comparable to the baseline. This efficiency is due to the fact that manifold structure and label information is considered in the process of data reduction. Note the inefficiency of graph quantization method which performs data reduction regarding to Euclidean space which is not the case when data lie on a manifold.

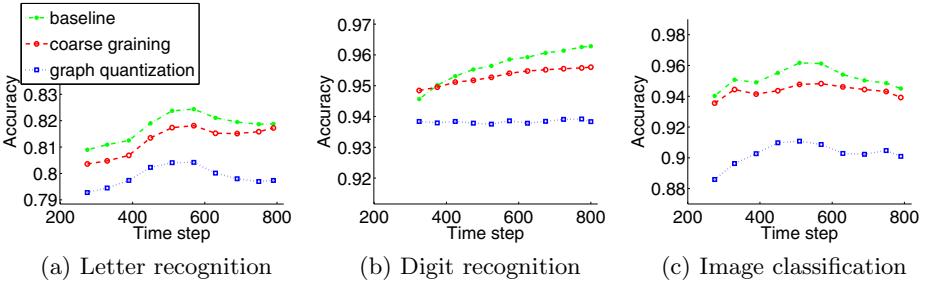


Fig. 3. Online classification. Data is arrived sequentially and maximum buffer size is 200.

5.3 Manifold Structure Preservation

In this experiment CG is done for 500 data points to reduce the data size to 100. One test data point is added and its label is predicted. Accuracy is averaged over 500 new data points added separately. We do in this manner intentionally to prevent new data points recover the manifold structure. So the result is an indication of how well the manifold structure is preserved in CG. Figure 4 shows the effectiveness of our CG method compared to graph quantization method [11] on USPS, UCI letters. Again we think this is due to the "manifoldwise" nature of our method.

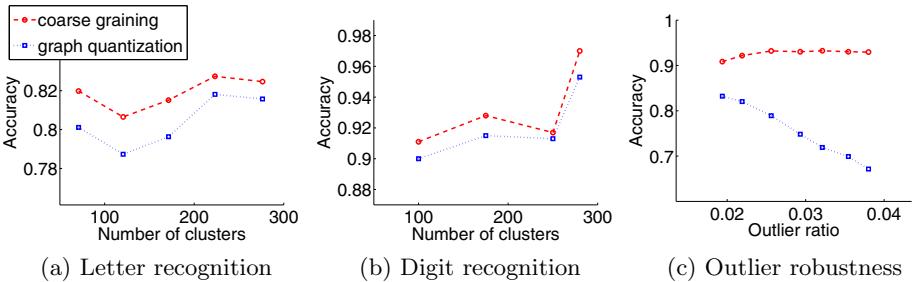


Fig. 4. (a,b): Capability of methods to preserve manifold structure. 500 nodes are coarse grained and the classification accuracy is averaged for separately added 500 new data. (c): Comparison of robustness to outliers in USPS.

5.4 Outlier Robustness

In this experiment we evaluate robustness of our method to outliers in data from USPS. Noise is added manually and classification accuracy is calculated. Outliers are generated by adding multiples of the data variance. Figure 4-c shows robustness of our method compared to the graph quantization method. In our method outliers are merged and their effect is reduced while in the graph quantization method separate clusters are devoted to outliers.

6 Conclusion

In this paper, a novel semi-supervised CG algorithm is proposed to reduce the number of data points while preserving the manifold structure. To this end a new formulation of LP is used to derive a new spectral view of the HS. We show that the manifold structure is closely related to the eigenvectors of a variation of the stochastic matrix. This structure is well preserved by any algorithm which guarantees small distortions in the corresponding eigenvectors. Exact and approximate coarse graining algorithms are provided alongside a theoretical analysis of how well the LP properties are preserved. The proposed method is evaluated on three real world datasets and outperforms the state of the art CG in the following scenarios, namely online classification, manifold preservation and robustness against outliers. The performance of our method is comparable to that of an algorithm that utilizes all the data in a simulated online scenario.

A theoretical analysis of robustness against noise, extending the spectral view point to other manifold learning methods, and deriving tighter error bounds on CG, to name a few, are interesting problems that remain as future work.

Acknowledgments. We would like to thank M. Valko and B. Kveton for providing us with experimental details of quantization method, A. Soltani-Farani for reviewing the manuscript, and anonymous reviewers for their helpful comments. This work was supported by National Elite Foundation of Iran.

References

1. Zhu, X.: Semi-Supervised Learning Literature Survey. Technical Report 1530, Department of Computer Sciences, University of Wisconsin Madison (2005)
2. Chapelle, O., Scholkopf, B., Zien, A.: Semi-supervised Learning. MIT Press, Cambridge (2006)
3. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold Regularization: a Geometric Framework for Learning from Labeled and Unlabeled Examples. *Journal of Machine Learning Research* 7, 2399–2434 (2006)
4. Duchenne, O., Audibert, J., Keriven, R., Ponce, J., Segonne, F.: Segmentation by Transduction. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8 (2008)
5. Belkin, M., Niyogi, P.: Using Manifold Structure for Partially Labeled Classification. *Advances in Neural Information Processing Systems* 15, 929–936 (2003)
6. Grabner, H., Leistner, C., Bischof, H.: Semi-supervised On-Line Boosting for Robust Tracking. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008*, Part I. LNCS, vol. 5302, pp. 234–247. Springer, Heidelberg (2008)
7. He, X.: Incremental Semi-Supervised Subspace Learning for Image Retrieval. In: Proceedings of the ACM Conference on Multimedia (2004)
8. Moh, Y., Buhmann, J.M.: Manifold Regularization for Semi-Supervised Sequential Learning. In: ICASSP (2009)
9. Goldberg, A., Li, M., Zhu, X.: Online Manifold Regularization: A New Learning Setting and Empirical Study. In: Proceeding of ECML (2008)
10. Dasgupta, S., Freund, Y.: Random Projection Trees and Low Dimensional Manifolds. Technical Report CS2007-0890, University of California, San Diego (2007)

11. Valko, M., Kveton, B., Ting, D., Huang, L.: Online Semi-Supervised Learning on Quantized Graphs. In: Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence, UAI (2010)
12. Lafon, S., Lee, A.B.: Diffusion Maps and Coarse-Graining: A Unified Framework for Dimensionality Reduction, Graph Partitioning, and Data Set Parameterization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(9), 1393–1403 (2006)
13. Zhou, D., Bousquet, O., Lal, T., Weston, J., Scholkopf, B.: Learning with local and global consistency. *Neural Information Processing Systems* (2004)
14. Zhu, X., Ghahramani, Z., Lafferty, J.: Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. In: ICML (2003)
15. Zhu, X., Ghahramani, Z.: Learning from Labeled and Unlabeled Data with Label Propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University (2002)
16. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes, The Art of Scientific Computing, 3rd edn. Cambridge University Press, Cambridge (2007)
17. Gfeller, D., De Los Rios, P.: Spectral Coarse Graining of Complex Networks. *Physical Review Letters* 99, 3 (2007)
18. Frank, A., Asuncion, A.: UCI Machine Learning Repository (2010)
19. Fei, L., Fergus, R., Perona, P.: Learning Generative Visual Models From Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories. In: IEEE CVPR 2004, Workshop on Generative Model Based Vision (2004)
20. Chatzichristofis, S.A., Boutsalis, Y.S.: CEDD: Color and Edge Directivity Descriptor: A Compact Descriptor for Image Indexing and Retrieval. In: ICVS, pp. 312–322 (2008)