Graph constrained label propagation on water supply networks

Manuel Herrera [a,*], Eva Ramos-Martínez [b], Joaquín Izquierdo [b] and Rafael Pérez-García [b]

[a] *InfraSense Labs, Imperial College London, London, United Kingdom*

[b] *Fluing–IMM, Universitat Politècnica de València, Valencia, Spain*

In many real-world applications we have at our disposal a limited number of inputs in a theoretical database with full information, and another part of experimental data with incomplete knowledge for some of their features. These are cases that can be addressed by a label propagation process. It is a widely studied approach that may acquire complexity if new constraints in the new unlabeled data that should be taken into account are found. This is the case of the membership to a group or community in graphs. The proposal is to add the Laplacian matrix as well as another different similarity measures (may be not found in the original database) in the label propagation. A kernel embedding process together with a simple label propagation algorithm will be the main tools to achieve this approach by the use of all types of available information. In order to test the functionality of this new proposal, this work introduces an experimental study of biofilm development in drinking water pipes. Then, a label propagation through pipes belonging to a complete water supply network is approached. These pipes have their own properties depending on their network location and environmental co-variables. As a result, the proposal is a suitable and efficient way to deal with practical data, based on previous theoretical studies by the constrained label propagation process introduced.

Keywords: Label propagation, semi-supervised learning, kernel methods, water supply networks.

Introduction

In many practical applications of pattern classification and data mining, one often faces lack of sufficient labeled data, since labeling often requires expensive human labor and much time. Semi-supervised learning (SSL) is a class of machine learning techniques that makes use of both labeled and unlabeled data for training their associated algorithms (typically a small amount of labeled data with a large amount of unlabeled data). Thus, it represents a halfway between supervised and unsupervised learning [4], working with unlabeled data, but providing some supervised information [24]. In general, these methods can be categorized into two classes: transductive learning [10] and inductive learning [1]. The goal of transductive learning is only to estimate the labels of the given unlabeled data, whereas inductive learning tries to induce a decision function which has a low error rate on the whole sample space.

The key to semi-supervised learning problems is the prior consistency [24], also called cluster assumption [3]:

1. nearby points are likely to have the same label;
2. points on the same structure (such as a cluster or a sub-manifold) are likely to have the same label.

Note that the first assumption is local, while the second one is global. The cluster assumption compels us to consider both local and global information contained in the dataset during learning. Label propagation has been applied in classification and ranking task in several fields as varied as web page classification [11], genome issues to rank biomarkers [22] or water distribution issues to manage large networks [9] or sectorize [2], among others.

This work introduces a semi-supervised problem in which a database of labeled data is available, but the unlabeled data follow a graph structure. Thus, in the process of label propagation both the usual pairwise similarities and the graph constraints related to these unlabeled data should be taken into account. Then, the proposal is to work with the Laplacian matrix associated with the graph together with the similarity matrices, all of them embedded into a kernel space, [20,21]. This is a high dimensional feature space defined by a

[*] Corresponding Author: Manuel Herrera, InfraSense Labs, Dept. of Civil and Environmental Engineering, South Kensington Campus - Imperial College London - SW7 2BU London, United Kingdom; E-mail: a.herrera-fernandez@imperial.ac.uk

AI Communications
ISSN 0921-7126, IOS Press. All rights reserved

kernel function, i.e. a function returning the inner product between the images of two data points in the feature space. Working with kernels both complex and nonlinear problems usually can be addressed by easier even linear methodologies. Label propagation takes place in this kernel feature space. It uses a simple algorithm, which takes into account different kernel properties. In addition, it is introduced a methodology for tuning both weights and parameters which will appear along this process. Regarding the Laplacian matrix, the proposal is to address its tuning by a local scaling approach [23]. The parameters related to dissimilarities embedded into kernel spaces and the weights to combine these kernels should be tuned by a common process based on new trust-regions achievements [17].

To test the availability of this approach, an experimental study of biofilm development in water supply networks (WSNs) is proposed. In this case, various studies have been approached by independent pipes which have been labeled depending on their membership to a cluster. Nevertheless, in the study of a WSN we have a number of unlabeled pipes but distributed in a graph related to the WSN layout. The above label propagation process allows to take into account the database information but also includes the graph constraints into the WSN.

The outline of the paper is as follows. The following Section proposes the graph constrained label propagation methodology. It also introduces the main tools used for this label propagation process. After that, a method for tuning the corresponding parameters is proposed. Next, an experimental study about biofilm development in WSN is developed. The last section is about conclusions and further challenges for closing the paper.

The graph constrained label propagation algorithm

The proposed algorithm has two different parts. The first one address the kernel embedding process of the data. The second one approach a label propagation algorithm that takes advantage of the kernel properties of the space where it works.

Kernel embedding process

To propagate labels on a graph we should take into account both graph structure and pairwise similarities. Thus, building the Laplacian matrix, L, associated with the graph is one of the main processes to achieve. This matrix is defined as the diffence between the degree diagonal matrix, D, of the graph and the affinity matrix, A. Nevertheless, this basic definition may be slighty modified in order to exploit the good propierties in their representation of graphs [15]. Then, the basic expression of Laplacian matrix, $L = D - A$, evolves to the so-called normalized Laplacian $L = I - D^{-1/2}AD^{-1/2}$ (where I is the identity matrix) which we use in this work because its symmetry and good properties of its spectrum [14].

Through the kernel embedding process we can work in a high dimensionality space where linear methods work well to solve complex problems. In addition, the main processes take place in the feature space and the learning algorithms can be expressed so that the data points only appear inside dot products with other points. This is often referred to as the “kernel trick” [20]. Besides of this, it is possible to add as much information as we have for this label propagation. This allows to include the necessary similarities obtained from vector data to the structure of the graph. Thus, the process starts embedding the Laplacian of the graph together with the different similarity matrices into a kernel matrix for the unlabeled data K_U. We can directly transform similarity matrices into kernel matrices, which should be symmetric and positive semidefinite matrices that enconde the relative positions of all points [12].

Given a set of unlabeled data $X_U = \{x_1, ..., x_{n_U}\}$ in $\mathbb{R}^u$, the process steps are as follows:

1. Build the affinity matrix $A \in \mathbb{R}^{n \times n}$ defined by $A_{ij} = exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2}\right)$ if $i \neq j$ and $A_{ii} = 0$.
2. Define D to be the degree diagonal matrix whose (i,i)-element is the sum of the entries in A’s ith row.
3. Build the matrix $L = I - D^{-1/2}AD^{-1/2}$.
4. Embed into a kernel space the Laplacian and dissimilarity matrices associated with the problem.
   (a) Scale data between 0 and 1;
   (b) Plug a diagonal of 1’s into the diagonal of each matrix.
   (c) Next, matrices are mirrored through their diagonals to make them symmetric.
5. Follow Equation 1 to combine individual kernels.

$$K_U = w_{Lap}K_{Lap} + \sum_{i \in I} \omega_i K_i \tag{1}$$

The scaling parameter of step 1, σ^2, controls how rapidly the affinity A_{ij} falls within the distance between

x_i and x_j [13]. $D = diag(d_1, \dots d_{n_U})$ is the degree matrix, which is the diagonal matrix formed from the vertex degrees and A is the adjacency matrix. Each ω_i (step 5) allows to give different importance to each dissimilarity matrix, K_i, involved in the performance of K_U. In step 5 and Equation 1 expression ω_{Lap} and K_{Lap} are the weight and kernel matrix, respectively, associated with the Laplacian. These weights should only meet the condition of being positive; allowing a conic combination for Equation 1.

The K_i kernels associated with dissimilarity matrices can directly be these kind of matrices embedded into a kernel space or their expression can be chosen between already known kernel expressions (see Table 1). As a suggestion, given the categorical nature of labels, we attempt to check firstly kernels independent on any specific metric (such as Linear or Polynomial kernels, Table 1), because they usually offer better response working with categorical variables.

Table 1

Short list of some common kernel functions

Name	Expression
Gaussian	$K(x,x') = \exp\left(-\frac{\|\|x-x'\|\|^2}{2\sigma^2}\right)$
ANOVA	$K(x,x') = \sum \exp\left(-\sigma(x^k - x'^k)^2\right)^d$
Linear	$K(x,x') = x^T x' + c$
Polynomial	$K(x,x') = (\alpha x^T x' + c)^d$
Rational Quadratic	$K(x,x') = 1 - \frac{\|\|x-x'\|\|^2}{\|\|x-x'\|\|^2+c}$

There are a number of reasons to justify that the process of kernelization of this Laplacian matrix is correct [8]. There are two key properties that a kernel function must meet [21]. Firstly, it should capture the measure of similarity approximate to the particular task and domain, and, secondly, its evaluation should require significantly less computation than it would be needed in an explicit evaluation of a corresponding feature mapping. Furthermore, as the sum of kernel matrices is another kernel matrix, we propose to build an accumulative matrix, which is the weighted sum of the normalized dissimilarities in the different characteristics of the data.

Now it is possible to expand K_U plugging together both labeled and unlabeled data into just one matrix $\hat{K}$ (Equation 2). X_L is the set of labeled data with labels Y_L. Its kernel mapping, $K(X_L, X_L)$ of Equation 2, can be understood as a distance or dissimilarity measure based on data that engages well with the rest of the process.

$$\hat{K} = \begin{pmatrix} K_U & K(X_U, X_L) \\ K(X_L, X_U) & K(X_L, X_L) \end{pmatrix} = \begin{pmatrix} K_U & K_{UL} \\ K_{LU} & K_L \end{pmatrix} \tag{2}$$

Label propagation

After the previous kernel embedding of the problem, a slight modification of one of the more classical label propagation algorithms [25] will be enough to approach the labels for the unlabeled data structured by a graph. Figure 1 summarizes the process that we will follow.

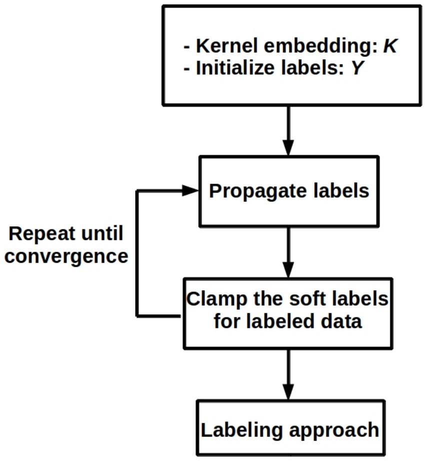

Fig. 1. Graph constrained label propagation

Given the number of classes, C, for labeling, an initial $(n_L + n_U) \times C$ probability label matrix is defined by an iterative process that starts with Equation 3.

$$Y^0 = \begin{pmatrix} Y_U \\ Y_L \end{pmatrix} \tag{3}$$

Y_L corresponds to the probability distribution of labeled data and Y_U to the unlabeled ones. The Y_U probability values can be initialized arbitrary by assigning values of the possible probabilities of the labels in a random way. An iterative propagation process will update this matrix: $Y^0 \to Y^1 \to \dots \to Y^m$.

The main propagation idea is based on Equation 4.

$$Y_U \leftarrow K_U Y_U + K_{UL} Y_L \tag{4}$$

Then, the iterative process is:

$$Y_U^1 = K_U Y_U^0 + K_{UL} Y_L \tag{5}$$

$$Y_U^2 = K_U (K_U Y_U^0 + K_{UL} Y_L) + K_{UL} Y_L \tag{6}$$

and so on. Assuming we iterate infinite times then:

$$Y_U = \lim_{n \to \infty} \left(K_U^n Y^0\right) + \left[\sum_{i=1}^{n} K_U^{(i-1)}\right] K_{UL} Y_L \tag{7}$$

The process converges since $\hat{K}$ is a normalized matrix and K_U is a submatrix of $\hat{K}$. Thus, we can find Y_U by solving the assignment of Equation 4 and we finally have Equation 8.

$$Y_U = (I - K_U)^{-1} K_{UL} Y_L \tag{8}$$

Tuning parameters in kernel label propagation

Once the scaling parameter for the Laplacian matrix has been tuned by well known methods: automatically [16] or by local scaling [23], the tuning process of the hyper-parameters coresponding to the kernel label propagation continues. The subjacent label propagation idea is related to the membership to one label, similarly to the common membership concept used in clustering. This makes useful to be label propagation based on similar approaches than we use in clustering to measure the goodness of each label distribution. The method of silhouette, introduced by [19], is a cluster validation and interpretation process that combines both cohesion and separation criteria. The value of the silhouette coefficient can vary between -1 and 1. A negative value corresponds to a case in which the average distance to points in the same label is greater than the minimum average distance to points with a different label. The ideal case is that the silhouette coefficient to be positive, and as close as possible to 1. The goodness of the final configuration of labels will be related to the weighted average of the silhouette width (ASW).

Once the criterion of searching best combination of parameters is fixed, a methodology based on multi-start trust region will be launched. Multi-start algorithms are an option if global extrema are searched, since these algorithms can explore more than a single basin of attraction of the objective function. The starting points should be enough and well distributed onto the design space [17].

Since we are interested in tuning weights and parameters within other optimization process (ASW on the label propagation, in this case), and we are not interested in the resulting model, we propose to apply derivative-free optimization algorithms [17]. Thus, only the objective function value is required. Our proposal is to adopt a linear second order model to interpolate (zones of) the parametric space by a simple surface where we can easily search its extreme. The use of a surrogate model, instead of the computation of the real objective function, reduces the computational time and extends the possible solutions from a set of points (Grid Search) to an entire surface. Thus, a so-called Design Of Experiments (DOE) is required at the begin of the algorithm by a double use: firstly, it searches different zones where locate our trust regions; next, DOE samples a number of starting locations into each of these regions.

The basic algorithm is outlined in the following steps, where the process starts selecting m regions of the parametric space (trust-regions). These trust-regions are specified with a center point and a radius, r. Once initialized, the trust region radius is dynamically adjusted by checking the quality of the clustering configuration for parameters at a certain distance from the search point.

1. Compute a number, m, of initial areas from the whole parametric space (DOE).
2. Sample points (parameters) in the selected regions (DOE).
3. Compute the labels for the sampled weights and parameters.
4. Create m surrogates based on the sampled points: weights, parameters, and labels.
5. Search the maximum values of ASW on the m surrogate surfaces.
6. Compute the label propagation process with the new selected parameters.
7. Redefine each trust-region and resize it by the factor h.
8. Repeat steps 4-7 until reach a stop condition.

The rule for redefining these trust-regions depends on the ASW. Since the ASW is in the [-1,1] interval, it is possible to calculate its increase, for each combination of weights and parameters during the iterations. The trust-region size is updated by the following rule (taking into account that the value of h will directly affect each trust-region radius):

- If the value of r is $<$ - 0.1, we set the value of the growth parameter, $h = h_{shrink}$.
- If the value of r is $>$ 0.1, we set the value of the growth parameter, $h = h_{grow}$.
- If the value of r is between -0.1 and 0.1, then the region size is not changed: $h = 1$.

By default the value of $h_{shrink} = 0.5$ and $h_{grow} = 2$ to contract and expand the regions, respectively. These criteria may change depending on the problem to be solved and its desirable convergence speed.

The trust-region process is stopped if one of the following criteria is met: the new increment is lower than

0.1 times the initial one; the new solution improves the last one by a number lower than 1.e-06; the number of iterations is greater than 30.

Experimental study

Biofilm is formed by a complex mixture of microbes, organic and inorganic material accumulated amidst a microbially-produced organic polymer matrix attached to the pipes inner surface. Once developed, biofilm is very resistant to disinfectant and can lead to various undesirable problems such as deterioration of bacterial water quality, generation of water bad tastes and odors, biocorrosion, and disinfectant decay, among others. It is known that, since the physical and hydraulic characteristics of the WSNs vary within the distribution systems creating heterogeneous habitats over time and space, biofilm exist at different levels within a WSN. That is why this work focuses on identifying the WSN's areas that are more or less prone to biofilm development according to the physical and hydraulic characteristics of the distribution system.

For the experimental study we use the biofilm database of 210 pipes belonging to different WSNs, proposed and studied by the authors in [18]. In that paper, a clustering of the biofilm database was approached, among other analysis. For our new interests, we keep on using the clustering medoids obtained in [18], which Table 2 shows. The corresponding clustering membership will be the labels for the elements of this study.

Table 2
Medoids of the clusters (labels) from the theoretical database

Medoids	Cluster 0	Cluster 1	Cluster 2
H. Regime	Turbulent	Turbulent	Turbulent
Velocity	High	Medium	Medium
P. Material	Plastic	Cement	Metal
P. Age	Young	Old	Medium
Biofilm	Low	Medium	High

Now the interest focuses on propagating these labels obtained from independent pipes to a set of pipes that integrate a complete WSN. In order to illustrate how our label propagation proposal works, we have chosen the "Example-3" of Epanet [7] (Figure 2). There are two raw water sources – one that is used continuously from high quality river water and another used for a portion of the day that comes from lower quality lake water. We could be interested in the possible biofilm development in a distribution system like that, following the natural constraints of its layout (graph architecture).

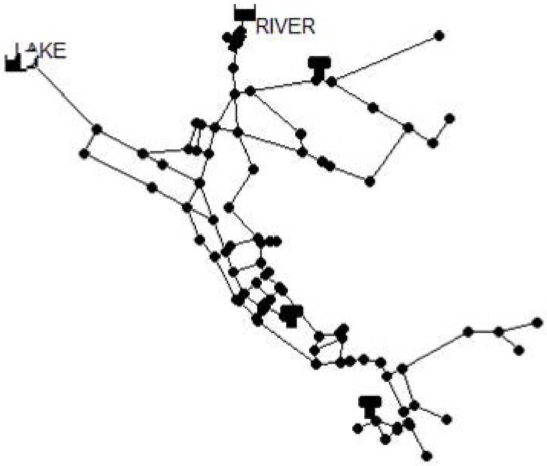

Fig. 2. Layout of the case-study WSN.

This WSN is composed of 91 consumption nodes, 117 pipes, and 5 sources of water (3 tanks, 1 river, and 1 lake); the total pipeline length is 117 Km. The average elevation and demand of the nodes are 25 m and 31.8 l/s, respectively.

To apply the above described label propagation process, we follow the steps to embed in a kernel space the unlabeled 117 data of the WSN (K_U) and the 210 labeled pipes K_L. In both cases, dissimilarities related to pipe variables such as material and age are taken into account. Other hydraulic variables also are considered (water velocity and age). Once the main kernel matrix is provided (Eq. 2), it is possible to calculate the labels for the WSN under study by Equation 8.

We have different kernels (and parameters to be tuned) depending on the nature of the variables. As a summary, we should tune the weight originated by the Laplacian matrix (w_{Lap}) and its parameter α related to the scale. It is also necessary to take into account the rest of the information by both, tuning the parameters of the matrices and tuning the weights to combine all the information. In this case, the study is focused on the following 3 variables: pipe material, pipe age, and water velocity. Thus, there will be necessary tuning their related weights: w_1, w_2, w_3; together with w_{Lap}. All of these inputs are categorical, such as the information of the labels in Table 2. We use Linear kernel for embedding these dissimilarities in the kernel space, then we are also interested on tuning the parameters associated with these kernels, such as c_1, c_2, c_3; together with α. These 4 weights are positive numbers and we work without constraints respect to its sum (conic combination). All these 4 weights and 3 parameters start from 7 different points in their parametric space running the trust-region process above proposed. The best configuration for label propagation is selected by the criterion of maximum ASW, which reaches to 0.37 in the final approach.

As a result, we have 21 pipes labeled by the group associated with a more likely high biofilm formation (see Figure 3). A majority of them (in orange) are situated in the North-East area of the WSN. This area is composed of pipes of 45 years on average and their material are mainly asbestos cement and cast iron. There are another area in the center of the WSN layout marked by this label. It is composed of short pipes which are also made of asbestos cement and cast iron. Finding metallic and cement old pipes in the area prone to high biofilm development agrees with what was expected from the bibliography. It is known that metallic and cement pipes tend to support more biofilm development than plastic pipes. This is because pipes with rough inner surface have greater potential for biofilm growth [5]. Rough surfaces provide more area for biofilm growth and protect biofilm from water shear forces. The accumulation of corrosion and dissolved substances in older pipes increase their roughness [6], thus, old pipes tend also to have greater biofilm development.

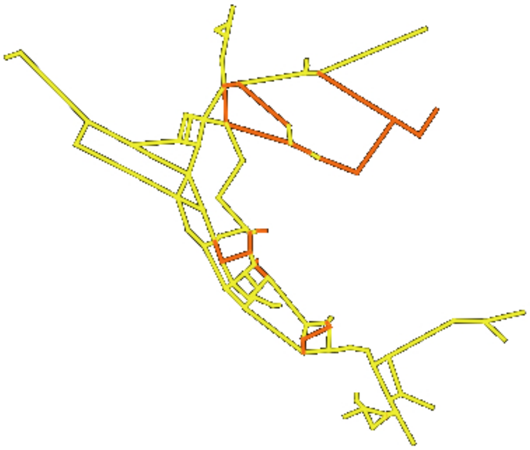

Fig. 3. Results of the label propagation process

Conclusions

A kernel approach to graph constrained label propagation has been studied. The main advantage of this methodology is its flexibility to take into account all the information available in the process of label propagation. In the case of biofilm development in WSNs, this kernel could gather the graph structure of the unlabeled data together with other similarities. Thus, the proposed label propagation is not only an inheritance process, but it is guided by graph constraints on the unlabeled data set together with its own similarities.

This paper works also tuning both weights of the kernel combination and label propagation parameters by proposals such as the more known local scaling of the Laplacian matrix or the novel process in which the tuning is approached by multi-start trust regions. Further works in label propagation could be addressed on how the kernel spectral matrix works in order to use the more representative eigenvectors to lead the propagation by diminishing computational costs. Also, other alternatives to this work based on, for example, multi-agent systems are worth exploring to complete the approach.

References

[1] M. Belkin, Niyogi, P., Sindhwani, V. On Manifold Regularization. Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics, 2005.

[2] E. Campbell, Ayala-Cabrera, D., Izquierdo J., Pérez-García, R., Tavera, M. Label propagation algorithm based methodology for water supply networks sectorization. *International Journal of Complex Systems in Science* 4(1), 2014, pp. 35–39.

[3] O. Chapelle, Weston, J., Schölkopf, B., Cluster Kernels for Semi-Supervised Learning. *Advances in Neural Information Processing Systems*, 15, 2003, pp. 585-592

[4] O. Chapelle, Schölkopf, B., Zien, A. Semi-supervised learning. Cambridge, Mass., MIT Press, 2006.

[5] S. Chowdhury, Heterotrophic bacteria in drinking water distribution system: a review, *Environmental Monitoring and Assesment*, 2011, pp. 2407–2415.

[6] R.T. Christensen, Age Effects on Iron-Based Pipes in Water Distribution Systems, Tech. Report Utah State University, 2009.

[7] EPA, Effects of water age on distribution system water quality. International Water Association, 2002.

[8] M. Herrera, Improving water network management by efficient division into supply clusters. PhD dissertation, Universitat Politècnica de València, Spain, 2011.

[9] M. Herrera, Izquierdo, J., Pérez-García, R., Montalvo, I. Multi-agent adaptive boosting on semi-supervised water supply clusters, *Advances in Engineering Software* 50, 2012, pp. 131–136.

[10] T. Joachims, Transductive Inference for Text Classification using Support Vector Machines, Proceedings of the 16th International Conference on Machine Learning, 1999.

[11] S. M. Kim, Pantel, P., Gaffney, S. Improving web page classification by label-propagation over click graphs, Proceedings of Conference on Information and Knowledge Management, 2009.

[12] G. Lanckriet, Cristianini, N., Bartlett, P., El Ghaoui, L., Jordan, M. I. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research* 5, 2004, pp. 27–72.

[13] X. Y. Li, Guo, L. Constructing affinity matrix in spectral clustering based on neighbor propagation. *Neurocomputing* 97, 2012, pp. 125–130.

[14] U. von Luxburg, A tutorial on spectral clustering. Statistics and Computing, 17, 2007, pp. 395–416.

[15] M. Newman. Networks: An Introduction. Oxford University Press, 2010.

[16] A. Ng, Jordan, M., Weiss, Y. On spectral clustering: Analysis and an algorithm, Proceedings of the Neural Information Processing Systems (NIPS) 2001, 14, pp. 849–856.

[17] D. Peri, Tinti, F. A multistart gradient-based algorithm with surrogate model for global optimization, *Communications in Applied and Industrial Mathematics* 3(1), 2012.

[18] E. Ramos-Martínez, Herrera, M., Izquierdo, J., Pérez-García, R. Ensemble of naive Bayesian approaches for the study of biofilm development in drinking water distribution systems, *International Journal of Computer Mathematics*, 91(1), 2014, pp. 135–146

[19] P. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Computational and Applied Mathematics*, 1987, 20, pp. 53–65.

[20] B. Schölkopf, Smola, A. J. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA, 2001.

[21] J. Shawe-Taylor, Cristianini, N. Kernel Methods for Pattern Analysis. Cambridge University Press, 2006.

[22] M. Stokes, Barmada, M., Kamboh, I., Visweswaran, S. The application of network label propagation to rank biomarkers in genome-wide alzheimer's data, *BMC Genomics*, in press, 2014.

[23] L. Zelnik-Manor, Perona, P. Self-Tuning Spectral Clustering, Proceedings of the Neural Information Processing Systems (NIPS) 2004, 17, pp. 1601–1608.

[24] X. Zhu, Semi-Supervised learning literature survey. Computer Sciences Tech. Report 1530, University of Wisconsin-Madison, 2005.

[25] X. Zhu, Ghahramani, Z. Learning from labeled and unlabeled data with label propagation, Computer Sciences Tech. Report, Carnegie Melon University, 2002.