Chapter 1

Metric learning in pairwise difference space

Introduction

Constraints on which data points should or should not be grouped together (must-link or cannot-link constraints) can significantly improve the performance of clustering. Most conventional semi-supervised approaches try to minimize must-link distances, maximize cannot-link distances, or to enforce a constant limit on either and optimize the other. They do not make explicit use of the probability distribution of must-link and cannot-link distances.

Here, we propose a framework for semi-supervised clustering based on projecting the data into a distance space in which distances reflect the linkage probabilities of belonging to the same cluster, using simple probabilistic classifiers trained on the available constraints. The framework can be seen as a novel approach to perform non-linear metric learning using weak supervision in the form of pairwise constraints, in order to improve clustering performance, as pioneered by (Xing et al., 2002). Although this problem is very similar to metric learning in general, the criterion of interest is often somewhat different: as opposed to optimizing the performance of some classifier as in e.g. (Xing et al., 2012)), or for a large margin as in e.g. (Weinberger et al., 2005), the goal is ensuring that all instances of a cluster are closer under the learned metric than those of different clusters. Furthermore, semi-supervised methods requiring partially labelled instances are not applicable if only pairwise constraints are available.

We have used this framework to model the structure of spatial representations in human participants in Chapter 5 above, using the information which buildings have

or have not been co-represented as the must-link and cannot-link constraints to train a Gaussian Discriminant Analysis (GDA) model in absolute pairwise difference space (see also Chapter 2.5 for the mathematical formulation, and Chapter 5 for the results).

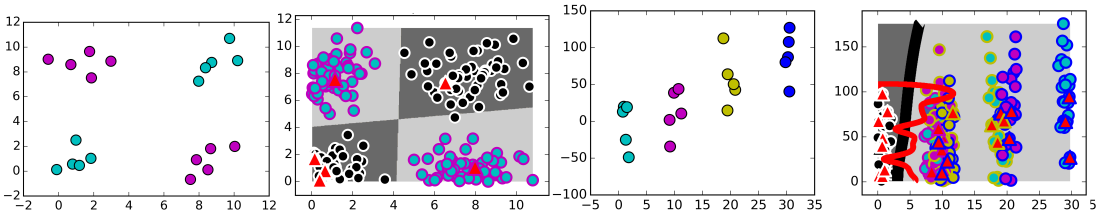
1.1 Motivation

For general applicability, non-linear metrics are vital, because of the problem of *multi-modality*. When data points forming several groups or 'modes' in unweighted feature space actually belong to the same cluster semantically, as indicated by ML constraints, there exists no linear projection able to separate them, and linear methods must inevitably fail. A traditional example is the XOR dataset, consisting of four groups, connected by diagonal ML constraints, such that there exists no linear separating hyperplane - see Figure 1.1a.

Although kernel-based methods can deal with multimodal clustering problems (or any complicated data distribution) according to the Representer Theorem, in theory, given the optimal kernel and suitable parameters, in practice it is often difficult to find such a kernel. Most non-linear metric learning methods able to learn suitable kernels are sensitive to multiple hyperparameters, and, being nonconvex optimization problems, frequently get stuck in local minima.

A further issue with popular isotropic kernels such as the Radial Basis Function (RBF), frequently used in non-linear metric learning (Baghshah & Shouraki, 2010; Chitta et al., 2011), is that they are ill-suited for data with features on very different scales, since the optimal regularization parameters in one dimension can be suboptimal in other dimensions in this case, as pointed out by (Ong et al., 2005) (who propose a solution only in the supervised setting).

Our motivations for proposing a novel non-linear method are threefold. First, we would like to sidestep the difficulty of robustly finding a good non-linear metric for a particular dataset, in a probabilistic framework, without hyperparameter tuning. Second, instead of learning a metric using an objective function based on Lp-distance, which collapses the differences along the individual dimensions into one value, we would like to let the model directly access these individual differences, and thus to learn their importance, allowing it to easily deal with non-isotropic data (Figure 1.1c-d). Third, we would like to make use of prior information regarding what constitutes a 'good' metric for clustering. In particular, unlike using a weighted Lp-metric which



(a) XOR dataset **(b)** Pairwise differences **(c)** Non-isotropic variance **(d)** Pairwise differences

Figure 1.1: Motivation for the proposed metric learning approach. (a) Example not linearly separable data requiring non-linear metrics. (b) Visualization of the distribution of corresponding absolute pairwise differences (APD), containing the element-wise differences in all dimensions between all possible objects, within (black) and across (coloured) clusters. The background contour shows the probability of a pair with a given distance belonging to the same cluster, learned by Gaussian Discriminant Analysis, and used as the distance pseudometric. Red triangles show the labelled ML and CL constraints. (c) Example data with non-isotropic variance (three orders of magnitude larger in the y-axis direction). (d) Corresponding APD space coloured as in (a). In addition to the modelled probability of belonging to the same cluster, the models decision boundary (black), as well as the decision boundary of a Support Vector Machine with an isotropic RBF kernel (optimal parameters set by grid search), which overfits along the low-variance dimension.

attempts to summarize the structure of within-cluster and across-cluster pairwise distances as scalars, we define *a pseudo-metric based on the vector space of absolute pairwise differences* (APD).

This not only allows straightforward learning of the importance of each feature and thus fitting non-isotropic distributions better than isotropic kernels (Figure 1.1c-d), but also makes explicit the structure in the distribution of constraints. It has been observed before that for data containing clusters, the probability density function of pairwise Lp distances shows two peaks (one for within- and one for across-cluster pairs), e.g. by (Brin, 1995). However, in the case of multiple clusters with different shapes and variances, a bimodal distribution is insufficient to reflect the true distributions of the instance differences within or across clusters. Clearly, within-cluster variances in one cluster do not have to equal those in another cluster, and the same is true for across-cluster variances (as illustrated by the variances of the groups of data in APD space in Figure 1.1b and d). Learning in pairwise difference vector space (instead of collapsing these distances into scalars) allows our model to adapt locally to within- and across-cluster variances of different clusters, and therefore to better approximate the true pairwise distance distribution.

1.2 Supervised learning in absolute pairwise difference space

The formulation of (weakly) supervised learning in absolute pairwise difference space was given in Chapter 2.5 above. We briefly summarize it as follows.

Let $\mathcal{X} = (\boldsymbol{x}_i, ..., \boldsymbol{x}_n)$ be the feature vector representation of n objects which are to be clustered, where $x_i \in \mathbb{R}^D$ are vectors with D dimensions. Let the set of m given pairwise linkage constraints be denoted by $\mathcal{L}$, where $|\mathcal{L}| = m$, and $l_{i,j} \in \mathcal{L}$ is

$$l_{i,j} = \begin{cases} 1, & \text{if } i \text{ and } j \text{ belong to the same cluster (ML constraint)} \\ 0, & \text{if } i \text{ and } j \text{ belong to different clusters (CL constraint)} \end{cases} \tag{1.1}$$

Then, a distance metric d_l between two instances can be defined based on the probability that these instances belong to the same cluster, making use of a generative classifier (constituent model) trained on the given constraints:

$$d_l(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{\theta}) = 1 - p(l = 1 | \Delta\boldsymbol{x}, \boldsymbol{\theta}) = p(l = 0 | \Delta\boldsymbol{x}, \boldsymbol{\theta}) \tag{1.2}$$

There are three advantages of using generative classifiers instead of just the classification output as the pseudometric. First, this makes the model well-suited to learning in the inductive setting (where the examples from the test set are not seen at training time). Many existing approaches are designed for and evaluated in the transductive setting. Second, on more complex, unseparable data, utilizing the models confidence of whether instances should be linked, in addition to a binary prediction, greatly improves the resulting clustering. This choice also allows choosing suitable priors and likelihoods and thus tailoring the model to fit the data at hand.

The learned metric in (1.2) can subsequently be embedded into a new feature space using multi-dimensional scaling (MDS), and then used for clustering or classification.

1.3 Extension of the framework to semi-supervised learning

Given the large number of pairwise differences, $\binom{n}{2} = \frac{n^2-n}{2}$, the following question arises: why only use the tiny set of given pairwise constraints for learning, instead of making use of the entire distribution in APD space?

An extension of a recently proposed framework for semi-supervised learning, called Contrastive Pessimistic Likelihood Estimation (CPLE) (Loog, 2015), can facilitate such an approach for using the entire APD space. Loog (2015) points out that most current semi-supervised approaches are not 'safe' (do not guarantee better performance when including the unlabelled data than without), and often perform sub-optimally due to model misspecification (making assumptions violated by data). Instead, he suggests only using the intrinsic assumptions of a given classifier, and training it in a pessimistic framework: using optimization to find hypothetical labels for the unlabelled data corresponding to the worst case. This pessimism ensures that the inclusion of unlabelled data point cannot make the model accuracy worse than just using the labelled data.

In practice, his procedure pessimistically assign soft labels to the unlabelled data, such that the improvement over the supervised version is minimal; at the same time maximize log likelihood over labelled data. That is, the parameters θ_{semi} of a semi-supervised model in the CPLE framework are given by

$$\theta_{semi} = \arg\max_{\theta} \arg\min_{q} L(\theta|X,U,q) - L(\theta_{sup}|X,U,q), \tag{1.3}$$

where $X = (x_i, y_i)_{i=1}^{N}$ denotes the labelled data, U the unlabelled data, and q the hypothetical soft labels. (Loog, 2015) only provides a solution for Linear Discriminant Analysis (LDA), and his framework requires an explicit generative likelihood $L(\theta|X,U,q)$.

However, his framework can be modified to use discriminative likelihoods instead. This allows using both generative and discriminative classifiers in this framework, provided that they can output a prediction probability (such probability estimates can also be provided by Platt scaling (Platt et al., 1999) for classifiers which don't support them).

Figure 1.2 shows the objective function this extended CPLE model, supporting pessimistic semi-supervised learning with any constituent model. Instead of optimizing the generative likelihood L, the negative log loss $J(y,r) = logp(y|r) = \frac{1}{N}\sum_{i=1}^{N}(y_i \log(r_i) + (1-y_i)\log(1-r_i))$ is used in the optimization objective, where y_i is the predicted class of the i'th instance, and r_i is the probability associated with this prediction.

The modified CPLE objective function in Figure 1.2 can be used with any global optimization approach to train a model to fit the unlabelled data distribution in a pessimistic fashion, while maximizing performance over the labelled data. We used the locally biased variant (Gablonsky & Kelley, 2001) of DIRECT (DIviding RECTangles)

Algorithm 1.3.1: CPLEOBJECTIVEFUNCTION($model, \theta, q, X, y, U$)

$$1: \text{set } z_i \leftarrow \begin{cases} 1 & \text{if } q_i \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \text{ for all } i \in [1, |U|]$$
$$2: X' \leftarrow X \cup U$$
$$3: y' \leftarrow y \cup z$$
$$4: \text{set } w_i \leftarrow \begin{cases} 1 & \text{if } X'_i \in X \\ q_i & \text{otherwise} \end{cases} \text{ for all } i \in [1, |X'|]$$
$$5: model \leftarrow train(model, X', y', w)$$
$$6: r \leftarrow predictionprobability(model, X)$$
$$7: r' \leftarrow predictionprobability(model, U)$$
$$8: J_{labelled} \leftarrow \frac{1}{|X|} \sum_{i=1}^{|X|} (y_i \log(r_i) + (1 - y_i) \log(1 - r_i))$$
$$9: J_{unlabelled} \leftarrow \frac{1}{|U|} \sum_{i=1}^{|U|} (z_i \log(r'_i) + (1 - z_i) \log(1 - r'_i))$$
$$10: return(J_{unlabelled} - J_{labelled})$$

Figure 1.2: A general pessimistic semi-supervised learning framework, based on a generalization of (Loog, 2015) to use discriminative likelihoods and to be usable with any classifier supporting prediction probabilities. It requires a constituent model, current hypothetical labels q of the unlabelled data points and model parameters θ which are to be optimized, the unlabelled data U, labelled data X, labels y of X, and the negative log loss $J_{labelled}$ of the model trained only on X. Minimizing this objective increases accuracy over the labelled data while assuming worst-case labels for the unlabelled data

(Jones et al., 1993), a global, deterministic, derivative-free optimization method based on Lipschitzian optimization, which can handle the kinds of non-linear and non-convex functions constituted by Figure 1.2.

The resulting semi-supervised learning framework is highly computationally expensive, but has the advantages of being a generally applicable framework, needing low memory, and making no additional assumptions except for the ones made by the chosen classifier.

1.4 Constituent models

In principle, model able to produce probability estimates could be used as the constituent model in (1.2) to model the linkage probability distribution. Here, we focus on three models of this class: Gaussian Discriminant Analysis (GDA), Gaussian Process

(GP), and Random Forest (RF). The first (GDA) has a closed form solution, thus constituting the - to our knowledge - first probabilistically motivated approach to learning a non-linear distance metric in closed form.

In the case of the **GDA** (Bensmail & Celeux, 1996), the likelihoods of a pair of instances belonging to the same cluster $p(\Delta\mathbf{x}|l=1;\mu_1,\Sigma_1)$ or to different clusters $p(\Delta\mathbf{x}|l=0;\mu_0,\Sigma_0)$, are modelled using multivariate Gaussians:

$$p(\Delta\mathbf{x}|l=i;\mu_i,\Sigma_i) = (2\pi)^{-\frac{D}{2}}|\mathbf{\Sigma}_i|^{-\frac{1}{2}}\, e^{-\frac{1}{2}(\Delta\mathbf{x}-\boldsymbol{\mu}_i)^{\intercal}\mathbf{\Sigma}_i^{-1}(\Delta\mathbf{x}-\boldsymbol{\mu}_i)}, \tag{1.4}$$

where $i \in \{0,1\}$. (μ_1,Σ_1) are the means and covariances of the APD distances of pairs in the same cluster, and (μ_0,Σ_0) those in different clusters. These parameters can be easily estimated from the instances corresponding to the must-link and cannot-link constraints, respectively, by calculating their means and covariances.

In the case of **GP**, we model the linkage probability distribution using a Gaussian Process with mean function $m(\cdot)$ and covariance function $k(\cdot)$:

$$p(\Delta\mathbf{x}|l=i;M,K) = \mathcal{GP}(m(\cdot),k(\cdot)) = \mathcal{N}\left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_m) \end{bmatrix}, \begin{bmatrix} k(x_1,x_1),\ldots,k(x_1,x_m) \\ \vdots \quad \ddots \quad \vdots \\ k(x_m,x_1),\ldots,k(x_m,x_m) \end{bmatrix}\right) \tag{1.5}$$

In the experiments below, we have used a zero mean function and the square exponential kernel as the covariance function. See (Rasmussen & Williams, 2005) for an extensive book on Gaussian Processes.

Finally, in the case of **RF**, an ensemble of B decision trees is learned. Each decision tree is trained on a bootstrap resample of data (drawn with replacement), and only considering a randomly selected subset of the available features. See (Breiman, 2001) for details. There are several ways to obtain class probabilities from trained random forests (Boström, 2007). We used the relative class frequency, i.e. the averaged fractions of samples falling on that same class in the leaves of individual trees:

$$p(\Delta\mathbf{x}|l=i;M,K) = \frac{1}{B}\sum_{j=1}^{B}\frac{l(t_j,\Delta\mathbf{x},i)}{l(t_j,\Delta\mathbf{x},i)+l(t_j,\Delta\mathbf{x},1-i)}, \tag{1.6}$$

where t_j denotes the individual trained trees, and $l(t_j,\Delta\mathbf{x},i)$ is a function returning the number of training examples falling into the same leaf as $\Delta\mathbf{x}$ in tree t_j.

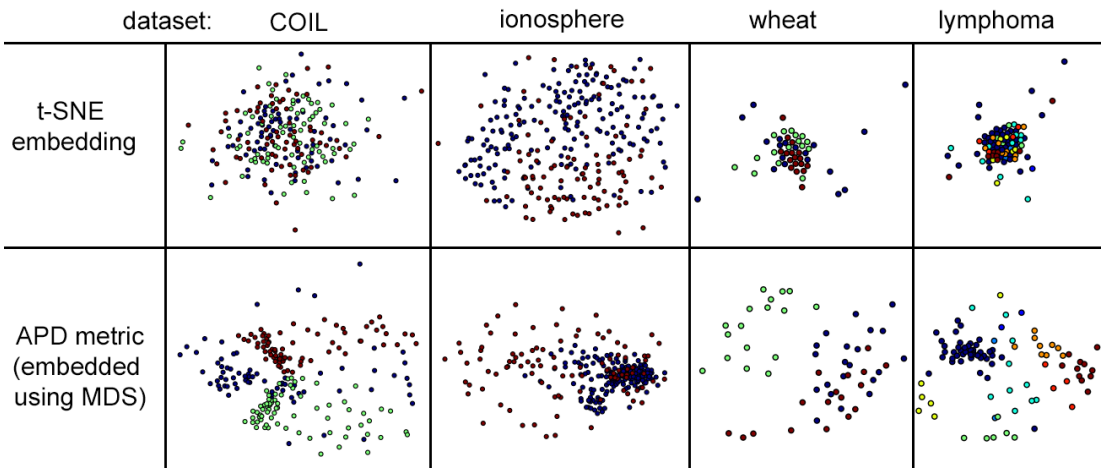


Figure 1.3: Embedding of pairwise distances as suggested by our metric learning approach (top) and by t-SNE (bottom) on the COIL-3 dataset (128x128px images of 3 similar-looking toy cars) used by (Zeng & Cheung, 2012), the ionosphere and wheat datasets from the UCI machine learning repository, and the lymphoma microarray dataset obtained from the Broad Institute. The APD plot was created using supervised GDA as the constituent model.

1.5 Preliminary results

We will first assume a purely supervised case - plugging in a GDA model (Equation (1.4)) into the APD metric framework (Equation (1.2)), as used to model human spatial representation structure in Chapters 2.5 and 5. We have claimed in Chapter 2.5 that our metric is well-suited for clustering, in the sense that for most within-cluster differences $\Delta \boldsymbol{x}_r$ and across-cluster differences $\Delta \boldsymbol{x}_n$ it holds that $d_r(\boldsymbol{x}_{r,1}, \mathbf{x}_{r,2}; \boldsymbol{\theta}) < d_n(\boldsymbol{x}_{n,1}, \mathbf{x}_{n,2}; \boldsymbol{\theta})$.

Figure 1.3 shows distances under our metric in three example datasets, embedded into two dimensions for visualization using MDS. It contrasts these distances with embeddings by t-SNE (Van der Maaten & Hinton, 2008), a state of the art method for visualization, designed to represent high-dimensional data points in such a way that similar objects are modelled by nearby points, and dissimilar objects by distant points. Figure 1.3 shows that the APD metric 'pulls together' different clusters (shown using different colours) more effectively than t-SNE.

The Figure is intended to show that datasets with no clear intrinsic cluster structure can be made separable by learning in APD space given only a few pairwise constraints (20 in these examples, which is less than 0.1% of all pairwise constraints), even with a fully supervised constituent model. It is not intended as a fair comparison, as t-SNE is an unsupervised method.

Clustering results on benchmark datasets

We have compared the described metric learning framework in APD space against several state of the art semi-supervised clustering models on multiple datasets. The constituent models in this section were trained in a semi-supervised fashion. In the case of GDA, we applied the extended CPLE learning principle shown in Figure 1.2 in APD space. In the case of the LS (label spreading) constituent model, we applied the learning algorithm by (Zhou et al., 2004), which can be described as a spreading activation model, labelling unlabelled data points based on their closest neighbours.

Tables 1.1 and 1.2 summarize the performance of semi-supervised clustering with APD metrics using these constituent models, compared to other recent approaches. In all cases, our model learned the metric in APD space, inferred and embedded all pairwise distances based on this metric, and then performed K-Means clustering.

dataset	cons-traints	DCA+ KMeans	Kmeans	MPC KMeans	SS-MMC	**APD-GDA**	**APD-LS**
pendigit	40	0.328	0.545	0.556	0.743	0.554	**0.874**
	60	0.553	0.545	0.535	0.831	0.717	**0.855**
	80	0.677	0.545	0.58	0.833	0.769	**0.901**
	100	0.721	0.545	0.537	0.861	0.8	**0.878**
letterIJL	40	0.207	0.266	0.223	**0.546**	0.194	0.467
	60	0.352	0.266	0.226	**0.691**	0.3	0.431
	80	0.4	0.266	0.167	**0.568**	0.417	0.329
	100	0.582	0.266	0.211	**0.655**	0.517	0.571
vehicle	100	0.233	0.186	0.122	**0.277**	0.161	0.208
	150	0.331	0.186	0.122	**0.365**	0.248	0.215
	200	0.418	0.186	0.102	**0.411**	0.355	0.288
	250	0.459	0.186	0.113	**0.458**	0.351	0.26
ionosphere	40	0.205	0.135	0.101	**0.242**	0.181	0.174
	60	0.167	0.135	0.094	**0.304**	0.293	0.148
	80	0.197	0.135	0.078	0.297	**0.332**	0.216
	100	0.199	0.135	0.094	0.313	**0.419**	0.282

Table 1.1: Semi-supervised clustering comparison - many constraints ($10^1 - 10^3$)**.** The table shows results of the APD-space metric learning framework applied to clustering, with two constituent models, one using GDA and one using label spreading (LS) (Zhou et al., 2004). Compared methods: Discriminative Component Analysis (DCA)+KMeans (Hoi et al., 2006), KMeans, MPCKmeans (Bilenko et al., 2004), and Semi-Supervised Maximum Margin Clustering (Zeng & Cheung, 2012). Non-APD model performances were taken from (Zeng & Cheung, 2012).

Dataset	*MPC*-Kmeans	CP-Kmeans	CNP-Kmeans	PCA-Kmeans	GP-Kmeans	LDA-Kmeans	APD+QDA	APD+LS
pima	0.6021±0.0028	0.6406±0.0018	0.6602±0.0014	0.6602±0.0000	0.6602±0.0026	**0.6848**±0.0041	0.6391±0.0088	OOM
iris	0.9000±0.0014	0.9133±0.0022	0.9200±0.0026	0.9133±0.0008	0.9000±0.0042	**0.9400**±0.0026	0.6133±0.0430	0.8687
wdbc	**0.8541**±0.0000	**0.8541**±0.0000	**0.8541**±0.0000	**0.8541**±0.0000	**0.8541**±0.0000	0.7326±0.0012	0.7976±0.0300	OOM
wine	**0.9438**±0.0027	0.6651±0.0012	0.7047±0.0019	0.7022±0.0008	0.9123±0.0021	0.9123±0.0028	0.6413±0.0156	0.5729
glass	0.5286±0.0202	0.5467±0.0010	**0.5561**±0.0122	0.5187±0.0091	0.5324±0.0216	0.5066±0.0082	0.4610±0.0127	**0.9109**
bupa	0.5449±0.0032	**0.5739**±0.0022	0.5536±0.0041	0.5536±0.0021	0.5521±0.0025	0.5159±0.0126	**0.5540**±0.0045	0.3785
balance	0.5136±0.0046	0.6160±0.0076	**0.6192**±0.0080	0.5088±0.0094	0.5246±0.0062	0.5376±0.0026	0.4956±0.0129	0.5884
spectheart	**0.6142**±0.0180	0.6067±0.0032	0.6060±0.0026	0.5543±0.0049	0.6043±0.0061	0.5032±0.0082	**0.6250**±0.0165	**0.7378**
Hepatitis	0.5935±0.0021	0.6129±0.0042	**0.6226**±0.0061	0.5935±0.0082	0.6210±0.0077	0.6181±0.0024	**0.6219**±0.0160	0.651

Table 1.2: Semi-supervised clustering comparison - few constraints ($10^0 - 10^1$)**.** The table shows results of the APD-space metric learning framework applied to clustering, with two constituent models, one using GDA and one using label spreading (LS) (Zhou et al., 2004). Compared methods: MPC-Kmeans (Bilenko et al., 2004), Constraint Projection (CP)-Kmeans (Tang et al., 2007), Constraint Neighbourhood Projection (CNP)-Kmeans (Wang et al., 2014), Principal Component Analysis (PCA)-Kmeans, GP-Kmeans (Eaton, 2005) and linear discriminant analysis (LDA)-Kmeans (Ding & Li, 2007). Non-APD model performances were taken from (Wang et al., 2014). 'OOM' stands for a model running out of memory.

Clustering results on cancer microarray data

Finally, we tested the other described constituent models on a bioinformatics dataset, to show its utility on non-benchmark problems with real-world relevance. It was recently proposed that semi-supervised consensus clustering (SSCC) outperforms the state of the art on cancer microarray data (Wang & Pan, 2014). We compared our APD-space based clustering with this ensemble clustering method, using the GP and RF constituent models.

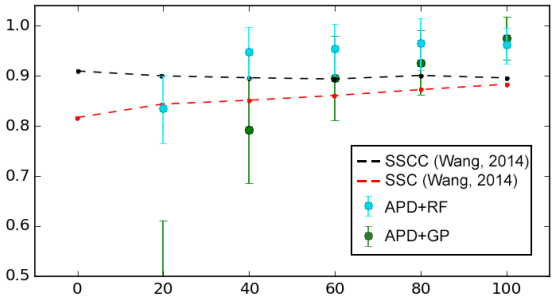


Figure 1.4: Semi-supervised clustering results on cancer microarray data from (Golub et al., 1999), compared to the state of the art semi-supervised consensus clustering method by (Wang & Pan, 2014).

Figure 1.4 shows the results on the Leukemia dataset, published and made available by (Golub et al., 1999). Despite the SSCC being an ensemble method, the APD-space methods significantly outperform it when given a sufficient number of constraints.

Bibliography

Baghshah, M. S., & Shouraki, S. B. (2010). Kernel-based metric learning for semi-supervised clustering. *Neurocomputing*, *73*, 1352–1361.

Bensmail, H., & Celeux, G. (1996). Regularized gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American statistical Association*, *91*, 1743–1748.

Bilenko, M., Basu, S., & Mooney, R. J. (2004). Integrating constraints and metric learning in semi-supervised clustering. In *Proceedings of the twenty-first international conference on Machine learning* (p. 11). ACM.

Boström, H. (2007). Estimating class probabilities in random forests. In *Machine Learning and Applications, 2007. ICMLA 2007. Sixth International Conference on* (pp. 211–216). IEEE.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Brin, S. (1995). Near neighbor search in large metric spaces. In U. Dayal, P. M. D. Gray, & S. Nishio (Eds.), *VLDB '95: proceedings of the 21st International Conference on Very Large Data Bases, Zurich, Switzerland, Sept. 11–15, 1995* (pp. 574–584). Los Altos, CA 94022, USA: Morgan Kaufmann Publishers.

Chitta, R., Jin, R., Havens, T. C., & Jain, A. K. (2011). Approximate kernel k-means: Solution to large scale kernel clustering. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 895–903). ACM.

Ding, C., & Li, T. (2007). Adaptive dimension reduction using discriminant analysis and k-means clustering. In *Proceedings of the 24th international conference on Machine learning* (pp. 521–528). ACM.

Eaton, E. R. (2005). *Clustering with Propagated Constraints*. Ph.D. thesis Citeseer.

Gablonsky, J. M., & Kelley, C. T. (2001). A locally-biased form of the direct algorithm. *Journal of Global Optimization*, *21*, 27–37.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A. et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, *286*, 531–537.

Hoi, S. C., Liu, W., Lyu, M. R., & Ma, W.-Y. (2006). Learning distance metrics with contextual constraints for image retrieval. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* (pp. 2072–2078). IEEE volume 2.

Jones, D. R., Perttunen, C. D., & Stuckman, B. E. (1993). Lipschitzian optimization without the lipschitz constant. *Journal of Optimization Theory and Applications*, *79*, 157–181.

Loog, M. (2015). Contrastive pessimistic likelihood estimation for semi-supervised classification. *arXiv preprint arXiv:1503.00269*, .

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*, 85.

Ong, C. S., Williamson, R. C., & Smola, A. J. (2005). Learning the kernel with hyperkernels. In *Journal of Machine Learning Research* (pp. 1043–1071).

Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, *10*, 61–74.

Rasmussen, C. E., & Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning*. The MIT Press.

Tang, W., Xiong, H., Zhong, S., & Wu, J. (2007). Enhancing semi-supervised clustering: a feature projection perspective. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 707–716). ACM.

Wang, H., Li, T., Li, T., & Yang, Y. (2014). Constraint neighborhood projections for semi-supervised clustering. *Cybernetics, IEEE Transactions on*, *44*, 636–643.

Wang, Y., & Pan, Y. (2014). Semi-supervised consensus clustering for gene expression data analysis. *BioData mining*, *7*, 1–13.

Weinberger, K. Q., Blitzer, J., & Saul, L. K. (2005). Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems* (pp. 1473–1480).

Xing, E. P., Jordan, M. I., Russell, S., & Ng, A. Y. (2002). Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems* (pp. 505–512).

Xing, E. P., Jordan, M. I., Russell, S., & Ng, A. Y. (2012). Similarity learning for provably accurate sparse linear classication. In *Proceedings of the 29th International Conference on Machine Learning (ICML)* (pp. 1871–1878).

Zeng, H., & Cheung, Y.-m. (2012). Semi-supervised maximum margin clustering with pairwise constraints. *Knowledge and Data Engineering, IEEE Transactions on*, *24*, 926–939.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., & Schölkopf, B. (2004). Learning with local and global consistency. *Advances in neural information processing systems*, *16*, 321–328.