Fast and Regularized Local Metric for Query–based Operations

Karim Abou–Moustafa and Frank Ferrie
The Artificial Perception Laboratory
Centre for Intelligent Machines, McGill University
3480 University street, Montreal, QC, Canada H3A 2A7
{karimt,ferrie}@cim.mcgill.ca

Abstract

To learn a metric for query–based operations, we combine the concept underlying manifold learning algorithms and the minimum volume ellipsoid metric in a unified algorithm to find the nearest neighbouring points on the manifold on which the query point is lying. Extensive experiments on standard benchmark data sets in the context of classification showed promising and interesting results with regard to our proposed algorithm.

1 Introduction

Query–based operations are used in a plethora of algorithms in the literature of pattern recognition, machine learning and computer vision. A typical scenario is to have a set $\mathcal{X}$ of high dimensional vectors (images, image patches, feature vectors, etc.) where it is required to find a subset of nearest neighbors or matches to a point that is either within the same set or a new incoming one. The Euclidean distance is usually the measure of choice for assessing the similarity between points. From a statistical perspective, a sober look at the Euclidean distance can raise questions about its full validity when used with high dimensional data arising from real life applications. Usually, such data are (1) High dimensional, highly structured (images, proteins, etc.) and nonlinear; (2) Measurements from various sources at different scales and with various degrees of variability and correlation; (3) Prone to various sources of noise that may largely deviate measurements and give rise to outliers in the data. These combined characteristics will be referred to as "*data complexity issues*".

By expanding the squared Euclidean norm $\|\mathbf{x}-\mathbf{y}\|_2^2$ to $(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T\mathbf{I}(\mathbf{x}-\mathbf{y})$, where $\mathbf{I}$ is the identity matrix, one directly obtains an instance of the general family of Mahalanobis distances between points $\mathbf{x}$ and $\mathbf{y}$: $D_\Sigma(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{y})$, where $\mathbf{\Sigma}$ is a symmetric and positive definite matrix. Replacing $\mathbf{\Sigma}$ by $\mathbf{I}$ implies that the Euclidean distance takes for granted that all variables are independent, the variance across all dimensions is one and that covariances among all variables are zero, a situation that is hardly attained in real life data. Therefore, the Euclidean distance, by definition, ignores the structure, scale, variance and correlations in the data and consequently it is wise to say that *"in the absence of clear evidence of Euclidean geometry, the metric structure should be inferred from the data"* [9].

In our previous work [1], we introduced the Minimum Volume Ellipsoid Metric (MVEM) and the Minimum Volume Ellipsoid of Nearest Neighbors (MiniVenn) algorithm to learn the MVEM. The MVEM is a similarity measure that tries to mitigate the effects of data complexity by using a parametrized Mahalanobis distance instead of the Euclidean distance. The MVEM is defined independently for each point in a data set $\mathcal{X}$ based on the information in a small neighborhood around it. Hence, it is adaptively and locally defined for each point (referred to as a query point) whether it is within the original data set or a new incoming one.

However, MiniVenn and consequently the MVEM suffer from two drawbacks. The first is that MiniVenn does not consider the notion of intrinsic dimensionality [5] in defining the similarity measure between points. That is, the literature on manifold learning algorithms assumes that despite the high dimensionality d of the input space, most of the data variability can be captured by far fewer dimensions d_0 than the dimensionality of the input space ($d_0 \ll d$). Accordingly, it is assumed that the data actually lies on, or near (due to noise), a lower dimensional nonlinear manifold that captures most of the data variability, and is embedded in the high dimensional input space. The dimensionality of this lower dimensional manifold is the intrinsic dimensionality of the data. The second drawback is due

to the computational requirements of MiniVenn. That is, in order to compute the MVEM, MiniVenn has to solve a determinant minimization problem under linear inequality constraints which is a difficult convex optimization problem. Directly solving this formulation of the problem without further manipulation and using a general purpose library for convex optimization [7] that does not consider the special problem structure resulted in a very slow algorithm for computing the MVEM. This, in turn, hinders the usage of the MVEM in practical situations that require fast query–based operations.

Contribution : We are interested in learning a metric for query–based operations that (1) considers the intrinsic dimensionality of the data and (2) is computed using a fast and efficient algorithm that allows it to be used in fast query–based operations. To this end, we combine the concepts underlying manifold learning algorithms and the minimum volume ellipsoid metric (MVEM) [1] in a unified, fast and efficient algorithm that tries to overcome data complexity issues.

The proposed algorithm is different from previous metric learning algorithms that focus on learning a metric specifically for k–NN (nearest neighbors) classification, exemplified by [11, 8], in that the *proposed algorithm* is unsupervised, self–adaptive for each new query point, and defines the metric on the lower dimensional manifold on which the query point is lying. *The proposed algorithm* is also different from algorithms that learn a global metric using similarity constraints (or side–information) [14, 2], or fully labeled data for k–NN classification such as [13]. These algorithms [14, 2, 13] learn a *global* metric through the general family of Mahalanobis distances $\mathrm{D}_A(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T\mathbf{A}(\mathbf{x}-\mathbf{y})$ and the differences between these algorithms are due to the constraints defining each metric.

2 Modifying the MiniVenn algorithm

In [1] we introduced the MiniVenn algorithm to learn the MVEM. This work strongly builds and extends our previous work on the MVEM and MiniVenn specifically, therefore the interested reader is encouraged to review [1] for more details and explanations. Due to space limitations, we directly proceed with the details of our contribution.

The modified MiniVenn algorithm, shown in Algorithm (1), overcomes the computational bottleneck of the original algorithm, and takes into consideration the intrinsic dimensionality of the data via two added steps to the algorithm. The algorithm proceeds as follows. In Step 1, similar to the original algorithm, MiniVenn defines a local neighborhood $\mathcal{N}_{\mathbf{x}_q}$ for $\mathbf{x}_q$ using the Euclidean distance as a similarity measure. In Step 2, also similar to the original MiniVenn, the algorithm computes the robust estimate of the covariance matrix $\mathbf{S}_q$ using the Minimum Volume Covering Ellipsoid (MVCE) estimator of the set $\mathcal{N}_{\mathbf{x}_q}$. The difference in the modified algorithm is in the new formulation and algorithm used to compute the MVCE of $\mathcal{N}_{\mathbf{x}_q}$. Steps 3 an 4 are the new steps in the algorithm and they are concerned with manifold detection, estimation of the local intrinsic dimensionality at $\mathbf{x}_q$, and MVEM regularization. In the following, each modification and addition will be further explained in detail.

Algorithm 1 Regularized Minimum Volume Ellipsoid of Nearest Neighbors : *Learns a local metric for query point $\mathbf{x}_q$ on the manifold on which $\mathbf{x}_q$ is lying.*

Require: $\mathcal{X}_{n\times d}$, $\mathbf{x}_q$, m, τ and ρ where $\mathcal{X}_{n\times d}$ is the training set with n d-dimensional samples, $\mathbf{x}_q$ is the query point, $m \geq d+1$ is a user input that controls the size of the neighborhood, $\tau > 0$ is the threshold to select the leading (tangent) directions with large eigenvalues along the manifold and $\rho \in [0,1]$ is the MVEM regularization parameter.

1: Find the set $\mathcal{N}_{\mathbf{x}_q}$ that has the m nearest neighbors to $\mathbf{x}_q$ using the Euclidean distance.

2: Compute the robust estimate of the covariance matrix $\mathbf{S}_q$ defined by the MVCE estimator for the set $\mathcal{N}_{\mathbf{x}_q}$ and centre $\mathbf{x}_q$ using Titterington algorithm [12].

3: Compute the eigen decomposition of $\mathbf{S}_q = \mathbf{V}\mathbf{L}\mathbf{V}^T$ where $\mathbf{V} = [V_1\ \ldots\ V_d]$, $\mathbf{L} = \mathrm{diag}(\lambda_1,\ldots,\lambda_d)$ are the matrices of eigenvectors and eigenvalues respectively and $\lambda_1 > \lambda_2 > \cdots > \lambda_d$.

4: Select the d_0 leading eigenvalues such that $\lambda_{[1:d_0]} > \tau$ and form the matrix $\tilde{\mathbf{L}} = \mathrm{diag}(\rho,\ldots,\rho,\frac{1}{\lambda_{d_0+1}},\ldots,\frac{1}{\lambda_d})$

5: **return** $\tilde{\mathbf{S}}_q^{-1} = \mathbf{V}\tilde{\mathbf{L}}\mathbf{V}^T$

Fast computation of the MVCE Let $\mathcal{N}_{\mathbf{x}_q} = \{\mathbf{x}_j \mid 1 \leq j \leq m,\ \mathbf{x}_j \in \mathcal{X}\}$ be the set of nearest neighbors to the point $\mathbf{x}_q$. The MVCE of $\mathcal{N}_{\mathbf{x}_q}$ with centre $\mathbf{x}_q$ is denoted by $\mathcal{E}$ and is parameterized by a symmetric and positive definite matrix $\mathbf{S}_q \in \mathbb{R}^{d\times d}$ as follows [3]:

$$\mathcal{E} \quad = \quad \{\mathbf{x}_j \mid \|\mathbf{S}_q^{-\frac{1}{2}}\mathbf{x}_j - \mathbf{b}\|_2^2 \leq 1,\ \forall j\}, \qquad (1)$$

where $\mathbf{b} = \mathbf{S}_q^{-\frac{1}{2}}\mathbf{x}_q$. Since $V(\mathcal{E}) \propto \det(\mathbf{S}_q^{-1})$, where $V(\mathcal{E})$ is the ellipsoid's volume, minimizing this volume can be formulated as follows:

$$\min_{\mathbf{S}_q}\ \log\det\mathbf{S}_q,\ \text{ s.t. }\ \|\mathbf{S}_q^{-\frac{1}{2}}\mathbf{x}_j - \mathbf{b}\|_2^2 \leq 1,\ \forall j. \qquad (2)$$

The objective and the constraints in (2) are convex in $\mathbf{S}_q^{-1}$, therefore this optimization problem has a unique global optimal solution. However, as in [1], directly solving this optimization problem using standard convex optimization libraries such as CVX [7] showed to

be computationally expensive and not efficient for practical situations. Alternatively, the dual of this optimization problem, thanks to Titterington [12], is easier to optimize and has a very fast and efficient algorithm to compute it (see [12] for algorithm details) :

$$\max_{\mathbf{S}_q,\Phi} \ \log\det(\mathbf{S}_q) \ \text{ s.t. } \ \mathbf{\Phi} \in \mathbb{R}^m, \ \mathbf{\Phi} \geq 0, \ \mathbf{\Phi}^T\mathbf{e} = 1 \quad (3)$$

$$\mathbf{S}_q = \sum_{j=1}^{m} \phi_j (\mathbf{x}_j - \mathbf{x}_q)(\mathbf{x}_j - \mathbf{x}_q)^T + \gamma\mathbf{I}$$

where Φ is the vector of dual variables ϕ_j, $\gamma \geq 0$ and $\gamma\mathbf{I}$ is an extra constraint that guarantees a minimal diameter of the ellipsoid in all directions. This would prevent the ellipsoid from collapsing to zero volume especially in high dimensional spaces [4].

Manifold detection To detect the manifold on which the query point $\mathbf{x}_q$ is lying, MiniVenn performs an eigen–decomposition and a regularization step for the robust estimate $\mathbf{S}_q$. The benefit of the eigen–decomposition is twofold: (1) It can estimate the intrinsic dimensionality of the data using Fukunaga's algorithm [6] (which is the role of parameter τ), and (2) the orthogonal eigenvectors of $\mathbf{S}_q$ decide which vectors are tangent or normal to the underlying manifold. That is, the eigenvector associated with the smallest eigenvalue (or lowest variance in $\mathcal{N}_{\mathbf{x}_q}$) is normal to the manifold, while the eigenvector associated with the largest eigenvalue (or highest variance in $\mathcal{N}_{\mathbf{x}_q}$) is tangent to the manifold. The latter is the main direction of interest since it is the direction that goes along the manifold and contributes the most to the similarity measure defined for $\mathbf{x}_q$. Note that in a d–dimensional space and for a d_0–dimensional manifold with $d_0 \ll d$, there will be approximately d_0 tangent vectors associated with the d_0 largest eigenvalues.

The Mahalanobis distance, however, measures the similarity using $\mathbf{S}_q^{-1}$, i.e. by taking the inverse of the eigenvalues. Thus it assigns by that small weights to high variance components (tangent eigenvectors) and large weights to low variance components (normal eigenvectors). It is at this point that the regularization parameter ρ is needed to emphasize the contribution of the main tangent vectors over the contribution of both normal and also less significant tangent vectors. More specifically, ρ influences the notion of similarity of the MVEM, however this influence is task and data dependent since it can tune the MVEM according to the objective of the task under consideration.

3 Generalization of the MVEM

Generalization of the MVEM is controlled by the MiniVenn's four parameters: m, τ, ρ and implicitly γ in Equation (3). While m and τ reflect the topological properties of the data, ρ influences the notion of similarity of the obtained metric. Using [6], τ can be fixed for a data set since it is a threshold on the normalized eigenvalues. Similarly, γ can be fixed for each data set separately although it was fixed to either 0 or 0.1 in all our experiments. More attention however, is required to select m and ρ. A large value of m will over–smooth the main tangent directions of the patch on which $\mathbf{x}_q$ is lying, while a very small value will lead to crude and rather fragile estimates of these directions. An intuitive approach is to select m and ρ via an optimization procedure. This can be achieved by linking the two parameters to an objective function that can be optimized. The optimal objective function in this case would be the objective function of the task under consideration. Implicitly, this means that the metric (or the MVEM) will be tuned to maximize or minimize this objective function. For instance, in the case of our experiments on query-based learning, m and ρ were optimized by a grid search to minimize the expected zero–one loss $E[L(Y, f(X))] = E[1-\delta(Y, f(X))]$ (or missclassification rate) on the available training set, where Y is the true label of the input X, $f(X)$ is the decision obtained from the classifier, and the $\delta(.,.)$ is the Kronecker delta function. Accordingly, since there is a training phase to optimize m and ρ directly on the task's objective function, the MVEM is expected to generalize well on unseen data sets.

It is worth noting that when MiniVenn forms a local neighborhood for the query point, it does not depend on labels or side-information [14] from the data, but rather on the similarity measure and the parameter m. This is unlike other metric learning algorithms that rely on the availability of *a priori* information in the form of fully/partially labeled data or side–information. The importance and contribution of any *a priori* knowledge only appears when optimizing m and ρ as mentioned earlier. Therefore, MiniVenn can be considered an unsupervised metric learning algorithm in that regard.

4 Experimental results

The experimental setting for query–based operations consisted of thirteen data sets, shown in Table 1, from the UCI Machine Learning Repository [10] and a k–Nearest Neighbours (k–NN) classifier with three different values for $k = (1, 3, 5)$. Since there are no explicit training and test sets for the used UCI data sets, 10 Folds Double Cross Validations (FDCV) were used to report the error rate. In terms of comparisons, the k–NN classifier using the MVEM was compared to k–NN classifier using the Euclidean metric and a more dedicated

Table 1. The thirteen UCI [10] data sets used in our experiments.

DataSet	classes	size	dim.
Balance (bal)	3	625	4
Liver Disorders (bup)	2	345	6
Glass (gla)	7	214	9
Housevotes (hou)	2	341	16
Ionosphere (ion)	2	350	33
Iris (iri)	3	150	4
Lymphography (lym)	4	148	18
New–Thyroid (new)	3	215	5
Pima–Diabetes (pim)	2	768	8
TicTacToe (tic)	2	958	9
WDBC (wdb)	2	569	30
Wine (win)	3	168	12
Yeast (yes)	10	1484	6

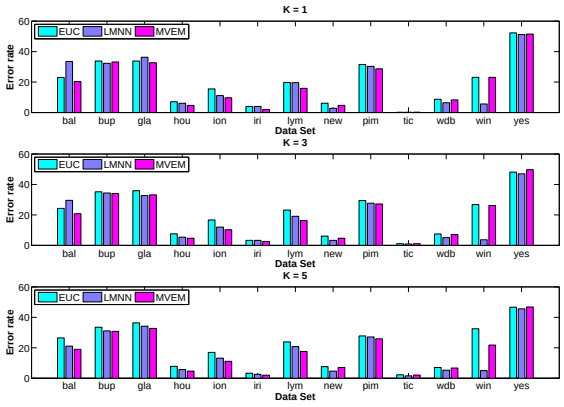

Figure 1. Comparing the error rates for the k–NN classifiers ($k = 1, 3, 5$) using the three different metrics: Euclidean, LMNN [13] and MVEM on the thirteen UCI data sets.

metric learning algorithm, LMNN [13] [1] that learns the metric specifically for k–NN classification. Our hypothesis is that k–NN classifier using MVEM will be very competitive with LMNN and consistently better than a k–NN classifier using the Euclidean metric.

Figure (1) shows the error rates for the three classifiers using the three different metrics. As a quantitative measure for the overall performance of the three metrics, statistical significance tests with $\alpha-$level of 5% show that, on average, the MVEM error rate is statistically significant than the Euclidean error rate while not statistically significant than LMNN. This confirms our hypothesis that the MVEM will be consistently better than the Euclidean metric while very competitive with a more dedicated algorithm like LMNN. This results is very relevant since MiniVenn had less *a priori* information during training and yet it has the same performance as LMNN. These results motivate us to extend the proposed algorithm and metric to the domain of clustering and unsupervised learning with complete absence of side–information and labels.

[1] The source code was downloaded from the author's website

Conclusion We have introduced an algorithm for learning an adaptive metric for query–based operations. The algorithm combines ideas from the minimum volume ellipsoid metric and from manifold learning algorithms to define the metric on the lower dimensional manifold of the query point. In the context of classification using a k–NN classifier, the metric showes very promising results in this regard and is competitive with other metric learning algorithms in the literature.

References

[1] K. Abou-Moustafa and F. Ferrie. The minimum volume ellipsoid metric. In *LNCS 4713, 29th Symposium of the German Association of Pattern Recognition, Heidelberg*, pages 335–344. Springer, 2007.

[2] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning a mahalanobis metric from equivalence constraints. *JMLR*, 6:937–965, 2005.

[3] S. Boyd and L. Vandenberghe, editors. *Convex Optimization*. Cambridge Univ. Press, 2004.

[4] A. Dolia, T. D. Bie, C. Harris, J. Shawe-Taylor, and D. Titterington. The minimum volume covering ellipsoid estimation in kernel-defined feature spaces. In *Proc. of the 17th ECML, Berlin*. Springer, 2006.

[5] K. Fukunaga, editor. *Introduction to Statistical Pattern Recognition*. Academic Press, 1972.

[6] K. Fukunaga and R. Olsen. An algorithm for finding intrinsic dimensionality of data. *IEEE Trans. on Computers*, 20(2):176–183, 1971.

[7] M. Grant, S. Boyd, and Y. Yinyu. Matlab software for disciplined convex programming, 2005. `http://www.stanford.edu/~boyd/cvx`.

[8] T. Hastie and R. Tibshirani. Discriminant adaptive nearest neighbour classification. *IEEE Trans. PAMI*, 18(6):607–615, 1996.

[9] G. Lebanon. Metric learning for text documents. *IEEE Trans. PAMI*, 28(4):497–508, 2006.

[10] D. Newman, S. Hettich, C. Blake, and C. Merz. UCI repository of machine learning databases, 1998.

[11] R. Short and K. Fukunaga. The optimal distance measure for nearest neighbour classification. *IEEE Trans. on Information Theory*, 27(5):622–627, 1981.

[12] D. Titterington. Estimation of correleation coefficients by ellipsoidal trimming. *J. of Royal Statistical Society*, 27(3):227–234, 1978.

[13] K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. In NIPS *18*, pages 1473–1480. 2006.

[14] E. Xing, A. Ng, M. Jordan, and S. Russell. Distance metric learning with application to clustering with side-information. In NIPS *15*, pages 505–512. 2003.