Distance Metric Learning through Minimization of the Free Energy

Dusan Stosic\(^a\)*, Darko Stosic\(^b\), Teresa B. Ludermir\(^a\), Borko Stosic\(^b\)

\(^a\)Centro de Informática, Universidade Federal de Pernambuco, Av. Luiz Freire s/n, 50670-901, Recife, PE, Brazil
\(^b\)Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco, Rua Dom Manoel de Medeiros s/n, Dois Irmãos, 52171-900, Recife, PE, Brazil

Abstract

Distance metric learning has attracted a lot of interest for solving machine learning and pattern recognition problems over the last decades. In this work we present a simple approach based on concepts from statistical physics to learn optimal distance metric for a given problem. We formulate the task as a typical statistical physics problem: distances between patterns represent constituents of a physical system and the objective function corresponds to energy. Then we express the problem as a minimization of the free energy of a complex system, which is equivalent to distance metric learning. Much like for many problems in physics, we propose an approach based on Metropolis Monte Carlo to find the best distance metric. This provides a natural way to learn the distance metric, where the learning process can be intuitively seen as stretching and rotating the metric space until some heuristic is satisfied. Our proposed method can handle a wide variety of constraints including those with spurious local minima. The approach works surprisingly well with stochastic nearest neighbors from neighborhood component analysis (NCA). Experimental results on artificial and real-world data sets reveal a clear superiority over a number of state-of-the-art distance metric learning methods for nearest neighbors classification.

Keywords: Distance metric learning, Statistical physics, Nearest neighbors classification, Free energy

1. Introduction

The success of many machine learning and pattern recognition tasks relies on the choice of an appropriate distance metric, that will be effective in capturing differences between the data. To date, there have been several approaches that attempt to learn distance metrics directly from the input data. Recent works have reported new and more sophisticated methods for learning these distance functions. However, the correct metric choice is problem specific and may dictate the success or failure of the learning algorithm\(^1\). Distance metric learning consists in choosing the distance metric using information contained in the training data. The resulting distance metrics are widely used to improve performance of metric-based methods, such as nearest neighbors classification\(^2\) or k-means clustering\(^3\). Depending on the problem, appropriate distance metrics can achieve substantial improvements over the standard Euclidean distance metric\(^4\)\(^5\). With this in mind distance metric learning plays a crucial role in many pattern recognition tasks, including classification\(^5\)\(^6\), regression\(^7\), clustering\(^4\)\(^7\)\(^8\), feature selection\(^9\)\(^10\), and ranking\(^11\).

Distance metric learning methods can be divided into three categories: supervised, semi-supervised, and unsupervised. Supervised methods for classification aim to ensure that patterns belonging to the same class are close to each other, and those from different classes are farther apart\(^5\)\(^12\). Semi-supervised methods for clustering use the information to form pairwise similarity or dissimilarity constraints\(^7\)\(^13\), and unsupervised methods learn a distance metric that preserves distance relations between most of the patterns for unsupervised dimensionality reduction\(^14\). In the supervised context, a popular class of distance metrics is the Mahalanobis distance (see Section\(^2\)). The distance metric generalizes the standard Euclidean distance by admitting arbitrary linear scalings and rotations. More precisely, learning the distance metric is equivalent to finding an optimal linear transformation that best serves the classification purpose.

While distance metric learning has achieved considerable success in many learning tasks, there remain some challenges to be addressed. For example, the learning process should be flexible enough to support a variety of constraints realized across distinct problem domains and learning paradigms. Distance metric learning approaches based on convex optimizations are formulated under specific constraints, i.e., the matrix must be positive semidefinite, which are unable to generalize well for different tasks. Another challenge is the ability to solve complex constraints, so that the learned distance metric generalizes well to unseen test data. The issue typically arises in distance metric learning methods that take non-convex objective functions with spurious local minima, such as the stochastic neighbors in neighborhood component analysis (NCA)\(^15\) which uses gradient descent. This paper focuses on providing a solution to these obstacles.

For distance metrics learned using complex constraints and

---

*Corresponding author.

Email addresses: ddstosic@bu.edu (Dusan Stosic), ddstosic@cin.ufpe.br (Darko Stosic), tbl@cin.ufpe.br (Teresa B. Ludermir), borkostosic@gmail.com (Borko Stosic)
non-convex objective functions, an important question entails what is the best distance metric that can be achieved for a given data set? The problem is reminiscent of a typical situation in statistical physics, which allows us to apply its concepts for solving the distance metric learning problem. Rather than performing a gradient-driven iterative search, the basic idea involves applying random transformations on the metric space, i.e., scalings and rotations, in search of a distance metric that minimizes the objective function. This provides a natural way to learn the desired distance metric. The learning process can be intuitively seen as stretching the metric space until some heuristic is identified, e.g., similar patterns are closer together and dissimilar patterns are farther apart (see Fig. 1).

In this paper we introduce a novel approach for learning the distance metric. Our approach differs from other distance metric learning methods in the following ways. We formulate this task as a typical problem in statistical physics, which allows us to apply its concepts for solving the distance metric learning problem as finding the ground state of a complex physical system. The heuristic is identified with the energy of the system, which we minimize by Metropolis Monte Carlo. This allows one to learn the distance metric for non-trivial constraints and non-convex objective functions with spurious local minima. As a case study, we consider stochastic nearest neighbors in NCA, but the same approach can be used for other heuristics. Our numerical experiments reveal that the resulting distance metric leads to significant improvements in accuracy for nearest neighbors classification. Several state-of-the-art distance metric learning methods are used for fair comparison.

The paper is organized as follows. Section 2 briefly reviews the existing literature on distance metric learning. Section 3 describes the proposed method. Section 4 demonstrates numerical experiments on nearest neighbors classification. Section 5 discusses important aspects of the proposed distance metric learning method. Section 6 draws the conclusions.

2. Related work

Distance metric learning consists in choosing a suitable distance measure from information in the training data. A popular class of distance metrics is the Mahalanobis distance:

\[ d_M = (u - v)^T M (u - v), \]  

(1)

due to its wide use in many application domains and because it provides a flexible way of learning an appropriate distance metric for complex problems [16, 17]. \( M \) is a symmetric positive semidefinite matrix that can be factorized into \( M = AVA^T \), given a linear transformation \( A \). The distance \( d_M \) between patterns \( u \) and \( v \) is equivalent to the Euclidean distance after performing the linear transformation:

\[ d_A(u, v) = (Au - Av)^T (Au - Av). \]  

(2)

The distance metric generalizes the standard Euclidean distance by admitting arbitrary linear scalings and rotations (see Fig. 1), where cross terms (or product between different components) can appear in distance computations through the off-diagonal elements of \( M \). The current problem is to learn the linear transformation \( u \to Au \), or its quadratic form \( M = A^T A \), that parametrizes the corresponding distance metric. While learning is unconstrained for linear transformations, for the quadratic form we must ensure that \( M \) remains positive semidefinite. As a result, most of distance metric learning methods fall into three broad categories: eigenvalue methods based on second-order statistics, convex optimizations over the space of positive semidefinite matrices, and non-convex optimizations.

2.1. Eigenvalue decomposition

Eigenvalue methods have been widely used to learn the linear transformation since they only need to compute an eigendecomposition. Most of them are based on second-order statistics, such as principal component analysis (PCA) [18], relevant component analysis (RCA) [19], discriminant component analysis (DCA) [20], and local Fisher discriminant analysis (LFDA) [21]. For example, PCA chooses the linear transformation \( A \) that maximizes the variance, \( \max \text{Tr}[A^T CA] \) subject to \( A^T A = I \), where \( C \) and \( I \) are the covariance and identity matrices, respectively. Another important and recent eigenvalue method is DMLMJ [22], which learns the distance metric by maximizing the Jeffrey divergence. Such linear transformations can be used either for dimensionality reduction when the matrix is rectangular, or to rotate and re-order the coordinate system for a better representation of the data. They can also be used to accelerate nearest neighbors calculations in large data sets [5]. However, eigenvalue methods are often viewed as inducing a distance function rather than learning the underlying metric.

2.2. Convex optimization

Distance metric learning can also be formulated as a convex optimization over the cone of positive semidefinite matrices \( M \). These approaches are more complicated to devise as they must enforce the constraint that \( M \) is positive definite. However, the
Monte Carlo Step

Initial matrix, A

| 1 | 0 | 0 |
|---|---|---|
| 0 | 1 | 0 |
| 0 | 0 | 1 |

Candidate matrix, B

| a11 | a12 | a13 |
|-----|-----|-----|
| a21 | 1   | 0   |
| a31 | 0   | 0   |

aik = Aik + U(-1,1)

Figure 2: Illustration of Metropolis Monte Carlo for distance metric learning. Given the current distance matrix A (e.g., starts with the identity matrix), each Monte Carlo step generates a new candidate matrix B which is accepted or rejected based on the Metropolis choice, W(A → B). Candidate matrices are formed by uniformly random deviations U(-1, 1) to all elements (background) in A. The Monte Carlo step is repeated until some convergence criterion is met, such as when the change in energy, ΔE = E_B − E_A, falls below a given threshold ε.

Monte Carlo Step

ΔE < ε

Final matrix

| a | b | c |
|---|---|---|
| d | e | f |
| g | h | i |

The classes of problems in pattern recognition have a close mathematical connection to those in statistical physics [29]. In relation with pattern recognition, statistical physics has paved the road for much of the research done in academia and industry, ranging from the revival of (79) to modern day deep learning [31]. Moreover, a parallel can be drawn between statistical physics - the behavior of systems with many degrees of freedom in thermal equilibrium at a finite temperature, and distance metric learning - the search of an appropriate metric by which to measure distances between patterns. As a result, we can employ concepts from statistical physics to learn the distance metric.

The focus of this work is on distance metric learning methods that contain complex constraints and non-convex objective functions, which make them susceptible to getting stuck in spurious local minima rather than the global optimum. For example, the gradient descent used in NCA often converges to local solutions representing suboptimal distance metrics. To solve this issue, we formulate the task as a minimization of the free energy of a statistical system which is equivalent to the distance
metric learning problem, under constraints on the distance metric. By analogy with statistical physics, we represent the set of distances between patterns as the constituents of a physical system, where the corresponding energy is given by the constraints or objective function. Since ground states and other low energy states are extremely rare among all possible states of a macroscopic system, we can expect that the methods from statistical physics used to find them are more than capable to search for the global optimum distance metric.

3.1. Problem formulation

We treat the distance metric learning problem as a search for the ground state of a complex system. The problem is reminiscent of simulations of the thermal motion of particles when the system is in equilibrium, which was solved by random walks in the phase space realized from Metropolis Monte Carlo simulations [32]. Here we propose a similar approach to learn the distance metric from the system is in equilibrium, which was solved by random walks. The problem is reminiscent of simulations of the thermal motion of particles when the system is in equilibrium.

Metropolis Monte Carlo can be used to learn the distance metric (see Fig. 3 for an illustration) as follows: First an index $i$ is randomly chosen and the matrix element, $A_i$, is displaced randomly to a new candidate value, $B_i$. The difference in energy, $\Delta E = E(B_i) - E(A_i)$, is then evaluated and the move is accepted with probability $W$ given by the Metropolis choice [32]:

$$W = \min \left[ 1, \exp \left( -\frac{\Delta E}{T} \right) \right],$$

where $T$ is the effective temperature, a disorder parameter in the same units as the cost function [34]. The second argument corresponds to the ratio $P(B_i)/P(A_i)$ of Boltzmann probabilities:

$$P(X) = \frac{e^{-E(X)/T}}{\sum_S e^{-E(S)/T}},$$

where the normalization factor (or partition function) guarantees that the probabilities sum to one. Only calculations of the energy difference, $\Delta E$, are necessary, while knowledge of the individual state probabilities and partition function (which would incur impossible computational costs) is not necessary. Transitions that lead to a decrease in energy are always accepted, while transitions to higher energies are accepted with a probability that decays exponentially with $\Delta E$ on the scale defined by the choice of $T$.

The procedure is repeated until $N$ trial moves have been attempted, where $N$ is the number of patterns in the system. Each set of $N$ moves comprises a single Monte Carlo step. Repeating this procedure many times allows us to simulate the system in equilibrium, which corresponds to the minimum of the “free energy”:

$$\mathcal{F} = -T \log \sum_S e^{-E(S)/T},$$

and approaches the ground state (or global minimum) as $T \to 0$. Distance metrics with higher energies (objective function values) are realizable at higher temperatures. For the choice of temperature, we can fix $T = 0$ which corresponds to an extremely rapid quenching of the system. The transition probability, $W$, then acquires a particularly simple form:

$$W = \begin{cases} 1, & E(A_i) \geq E(B_i) \\ 0, & E(A_i) < E(B_i) \end{cases},$$

yielding a greedy algorithm which only accepts changes to the matrix that reduce the objective function. This makes the approach susceptible to getting stuck in spurious local minima, like gradient descent, but there is no dependence on derivatives. Another possible choice is simulated annealing [34] which
Algorithm 1 DMLFE

1: Initialization:
2: \( \alpha \leftarrow \) annealing parameter
3: \( T_0 \leftarrow \) initial temperature
4: \( N \leftarrow \) number of patterns
5: \( M \leftarrow \) maximum number of iterations
6: \( \epsilon \leftarrow \) tolerance for convergence criteria
7: \( A = U(0, 1) \) \( \triangleright \) Initialize transformation matrix \( A \)
8: Monte Carlo procedure:
9: for \( k = 1 \ldots M \) do
10: \( T \leftarrow a^k \times T_0 \) \( \triangleright \) Temperature annealing
11: \( B \leftarrow A + U(-1, 1) \) \( \triangleright \) Update candidate matrix \( B \)
12: \( p_{ij} \leftarrow \frac{e^{-\frac{\Delta(E_B - E_{ij})}{T}}}{\sum_{l=1}^{N} e^{-\frac{\Delta(E_B - E_{lj})}{T}}} \) \( \triangleright \) Stochastic neighbors
13: \( E_B \leftarrow 1 - \sum_{i=1}^{N} \sum_{j=1}^{N} p_{ij}Y_{ij} \) \( \triangleright \) Objective func. Eq. (4)
14: \( \Delta W \leftarrow \min \left(1, \exp \left(\frac{\Delta(E_B)}{\epsilon} \right)\right) \) \( \triangleright \) Transition probability
15: if \( \Delta W \geq U(0, 1) \) then
16: \( A \leftarrow B \) \( \triangleright \) Accept update move
17: if \( \Delta E < \epsilon \) then \( \triangleright \) Convergence criteria
18: break
19: return \( A \) \( \triangleright \) Output learned distance matrix \( A \)

draws an analogy with annealing in solids, such as crystals. The process consists of first “melting” the system at high temperatures, and then a very slow cooling occurs where the temperature is lowered in stages (that must proceed long enough for the system to reach equilibrium) until the system “freezes” and no further changes occur. While the annealing procedure is computationally slower than quenching, convergence to the ground state is guaranteed for sufficiently long times by ergodicity of the dynamical process. The effect of different temperature choices is further discussed in Section 5.

We refer to the proposed method as Distance Metric Learning through minimization of the Free Energy (DMLFE). A simplified pseudo-code implementation of DMLFE is given in Algorithm 1.

3.2. Intuition

Monte Carlo provides a natural way to learn the distance metric. The method involves applying random transformations on the metric space in search of a distance metric that minimizes the objective function. For example, a candidate matrix \( B \) in two dimensions corresponds to the following combination of scaling and rotation transformations of the patterns:

\[
\begin{bmatrix}
  x' \\
  y'
\end{bmatrix} = \begin{bmatrix}
  x \\
  y
\end{bmatrix} \begin{bmatrix}
  t_x \cos \theta & -t_y \sin \theta \\
  t_x \sin \theta & t_y \cos \theta
\end{bmatrix}
\]

(10)

where \( t_x \) and \( t_y \) are the scaling factors, \( \theta \) is the angle of rotation. The learning process can then be intuitively seen as stretching and rotating the metric space until some heuristic is satisfied, e.g., similar patterns are closer together and dissimilar patterns are farther apart, as illustrated in Fig. 3.

Table 1: Brief description of methods used in the experiments

| Method         | Description                                      |
|----------------|--------------------------------------------------|
| ITML           | Information-theoretic metric learning            |
| LMNN           | Large margin nearest neighbor classification     |
| DML-eig        | Distance metric learning with eigenvalue optimiza-|
|                | tion                                            |
| SCML           | Sparse compositional metric learning             |
| NCA            | Neighborhood component analysis                  |
| LMNN           | Distance metric learning through maximization of  |
|                | the Jeffrey divergence                           |
| DMLMJ          | Distance metric learning through maximization of  |
|                | the Jeffreys divergence                          |
| DMLFE          | Distance metric learning through minimization of  |
|                | the free energy                                  |
| RF             | Random forest                                    |
| RF             | Random Forest                                    |

4. Experiments

In this section, we describe some experiments to evaluate the performance of distance metric learning methods. As a case study, we consider the stochastic nearest neighbors in NCA (see Eq. (4)) for our method, but other non-convex heuristics can be used. The proposed method is compared to the standard Euclidean distance and state-of-the-art distance metric learning methods listed in Table 1, which includes those based on eigenvalue decomposition, convex optimization, and non-convex optimization. Comparisons are also made with Random Forest (RF) [35]. For our experiments we consider supervised learning for nearest neighbors classification.

To make fair comparisons, we consider the following configurations for the experiments. The numerical experiments are empirically tested in the context of nearest neighbors classification, where the number of neighbors is varied between 1 and 40, and the best classification accuracy is chosen. RF is trained with 100 decision trees. We use the source codes of the distance metric learning methods supplied by the authors, and tune their parameters to get the best results. DMLFE uses an annealing procedure which starts with \( T = 0.1 \) and successively reduces the temperature by a multiplicative factor of \( \alpha = 0.9 \). Monte Carlo steps are repeated until convergence or a maximum number of steps is reached.

We compare the distance metric learning methods on 36 data sets from UCI [36, 37] and LIBSVM [38] machine learning repositories. The information of these data sets is summarized in Table 2. The classification accuracies are obtained by averaging over five runs of two-fold cross-validation. Divisions of the data sets for training and testing are made randomly. The features of the patterns are normalized into the interval [0, 1], which significantly improves performance of many standard distance metric learning methods.

Table 3 shows the average classification accuracies obtained by the competing methods. Fig. 4 further illustrates side by side comparisons of the results in Table 3. Since many points lie below the diagonal line, we can conclude that DMLFE presents a clear advantage over other approaches to nearest neighbors classification. Particularly, we find that our methods beats the standard Euclidean distance and is competitive with RF. The approach also seems to widely outperform distance metric learning methods based on eigenvalue decomposition and convex
optimization. Lastly, we observe that the Monte Carlo annealing procedure obtains better distance metrics than gradient descent for stochastic neighbors in NCA. The data sets where our method performs poorly can be explained by one of two causes: 1) convergence to a local minimum due to fast annealing, or 2) the free energy landscape is not represented by a complex surface with numerous local minima, which occurs in many situations in statistical physics such as spin glasses, but rather by a well-behaved surface with pronounced global minimum, where even gradient-based searches can find the optimal solution.

We rank the methods based on their classification accuracy.
on each data set, e.g., a rank of one is assigned to the method that has the highest accuracy, a rank of two to the method obtaining the second higher accuracy, and so on. The average ranks and other statistics of the competing methods are listed in the last rows of Table 4. We find that our approach achieves better results than NCA because the annealing procedure can overcome spurious local minima in which the gradient descent gets stuck. Our approach also significantly outperforms all other distance metric learning methods and has comparable classification performance to random forest (RF), which is often regarded as one of the best classifiers [39].

To detect whether there are significant differences among the results, we follow Demsar [40] for statistical comparisons of classifiers over many data sets. In each case we test the null-hypothesis that the distance metric learning methods obtain the same results on average. We start with the Wilcoxon

d signed rank [41] and Friedman [42] tests. The p-value for the Wilcoxon test is less than the confidence level of $\alpha = 0.05$ for all competing methods except RF, so we can reject the null-hypothesis as shown in Table 4. Similar conclusions can be made about the Friedman test where the p-value is $10^{-23}$. Then we proceed to the Bonferroni-Dunn [43] and Nemenyi [44] post-hoc tests to determine which distance metric learning methods perform the same or significantly different from our proposed method. The post-hoc tests can identify significant differences between the control method, which we take to be DMLFE, and other competing methods by computing a critical difference. Two distance metric learning methods are deemed significantly different if their average ranks differ by at least the critical difference [45]:

$$CD = q_0 \sqrt{\frac{n_c(n_c + 1)}{6n_t}} \quad (11)$$

where $q_0$ is the critical value, $n_c = 13$ and $n_t = 36$ are the number of competing methods and number of data sets, respectively. We consider the Nemenyi and Bonferroni-Dunn tests at a confidence level of $\alpha = 0.05$, with critical differences of 3.04 and 2.64, respectively. Fig. 4 graphically represents the significant differences among the performances of the different distance metric learning methods. Any distance metric learning method outside the marked area is deemed significantly different from the control method, or DMLFE. We find that the post-hoc tests reject the null-hypothesis for all competing methods except DMLMJ, LMNN, and RF as shown in Table 4. The statistical results allow us to conclude that our method significantly outperforms competing distance metric learning methods, and its performance is on par with RF.

5. Discussion

In this section we outline some important factors which lead to the preferred utilization of DMLFE. We start by considering the choice of temperature, which plays a crucial role in the performance of our proposed distance metric learning method. An annealing procedure will usually guarantee proper convergence to the global minimum, while quenching can get stuck in spurious local minima. However, there are many situations where quenching might provide “good enough” results and avoid the
larger computational costs of annealing. For example, some problems in physics contain many degenerate ground states which are easy to obtain, such as spin glasses, while others are composed of one unique minimum that can be found even with gradient-based searches. One can then imagine that similar situations might occur when learning the distance metric.

To analyze the temperature effects we conduct experiments on several of the considered data sets. Fig. 6 plots evolution of the energy (or objective function) when using the annealing procedure or multiple quenching with random initialization of the distance matrix. It follows that annealing often results in better solutions (see Figs. 6b–e)), achieving states that are closer to the ground state in energy, but at the cost of many more Monte Carlo Steps. However, there are situations where multiple quenching can achieve similar or even better results than annealing but with much fewer steps (see Fig. 6a)). This resembles the case of a frustrated system with nearly degenerate ground states, as mentioned above. While the choice of an annealing procedure with decaying temperature ensures finding the optimal solution, one may still opt to perform several runs with the simpler and faster quenching for different initializations, and then retain only the result with the lowest obtained energy.

We also analyze the performance of competing distance metric learning methods with the number of nearest neighbors. Fig. 7 illustrates the classification accuracy as we vary the methods from one to forty nearest neighbors in three of the data sets. We find that our method significantly outperforms NCA for all neighborhoods, which reaffirms our previous conclusion that the Monte Carlo annealing procedure achieves better solutions than gradient descent for stochastic nearest neighbors. In Fig. 7a), DMLFE achieves better performance than state-of-the-art methods for all neighborhoods, while only for the best performing number of nearest neighbors in Fig. 7b). Interestingly, in Fig. 7c) we find that our method achieves better classification accuracy using much fewer nearest neighbors than competing methods based on eigenvalue decomposition and convex optimization. Table 5 reveals similar findings based on statistics over all data sets of the number of nearest neighbors.

In Fig. 7a), DMLFE achieves better performance than state-of-the-art methods for all neighborhoods, while only for the best performing number of nearest neighbors in Fig. 7b). Interestingly, in Fig. 7c) we find that our method achieves better classification accuracy using much fewer nearest neighbors than competing methods based on eigenvalue decomposition and convex optimization. Table 5 reveals similar findings based on statistics over all data sets of the number of nearest neighbors that results in best classification. This suggests that the proposed method with stochastic nearest neighbors can achieve better classification with smaller computational cost than competing distance metric learning methods.

6. Conclusions

In this paper we developed a novel distance metric learning method. We showed that learning the distance metric can be formulated as a typical problem in statistical physics: distances represent constituents of a physical system and the corresponding energy is given by the leave-one-out classification error in Eq. 4. Then we demonstrated that this problem is equivalent to minimizing the free energy of a complex physical system. We proposed an approach to find the ground state using Metropolis Monte Carlo simulations. This allows one to learn the distance metric for non-trivial constraints and non-convex objective functions with spurious local minima. As a case study, we considered stochastic neighbors in NCA, but the same approach can be used for other heuristics. The experimental results on synthetic and real-world data sets demonstrate that the proposed method significantly outperforms other state-of-the-art distance metric learning methods for nearest neighbors classification. Future works should be directed towards the construction of more complex objective functions targeted for specific applications.

References

[1] J. V. Davis, B. Kulis, P. Jain, S. Sra, I. S. Dhillon, Information-theoretic metric learning, in: Proceedings of the 24th International Conference on Machine Learning, 2007, pp. 209–216.
[2] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Transactions on Information Theory 13 (1) (1967) 21–27.
[3] S. Lloyd, Least squares quantization in pcm, IEEE Transactions on Information Theory 28 (2) (1982) 129–137.
[4] S. Xiang, F. Nie, C. Zhang, Learning a mahalanobis distance metric for data clustering and classification, Pattern Recogn. 41 (12) (2008) 3600–3612.
[5] K. Weinberger, L. Saul, Distance metric learning for large margin nearest neighbor classifications, J. Mach. Learn. Res. 10 (2009) 207–244.
[6] B. Nguyen, C. Morell, B. D. Baets, Large-scale distance metric learning for nearest neighbors regression, Neurocomputing 214 (2016) 805–814.
[7] E. Xing, A. Ng, M. Jordan, S. Russell, Distance metric learning with application to clustering with side-information, in: Advances in Neural Information Processing Systems, Vol. 14, 2002, pp. 505–512.
[8] Y. Zhang, C. Zhang, D. Zhang, Distance metric learning by knowledge embedding, Pattern Recognition 37 (2004) 161–163.
[9] B. Liu, B. Fang, X. Liu, J. Chen, Z. Huang, X. He, Large margin subspace learning for feature selection, Pattern Recognition 46 (10) (2013) 2798–2806.
[10] C.-C. Chang, Generalized iterative relief for supervised distance metric learning, Pattern Recognition 43 (8) (2010) 2971–2981.
[11] B. McFee, G. Lanckriet, Metric learning to rank, in: Proceedings of the 27th International Conference on International Conference on Machine Learning, 2010, pp. 775–782.
[12] C. Shen, J. Kim, L. Wang, A. V. D. Hengel, Positive semidefinite metric learning using boosting-like algorithms, J. Mach. Learn. Res. 13 (2012) 1007–036.
[13] Q. Wang, P. C. Yuen, G. Feng, Semi-supervised metric learning via topology preserving multiple semi-supervised assumptions, Pattern Recogn. 46 (9) (2013) 2576–2587.
[14] J. Tenenbaum, D. Vin, J. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.
[15] J. Goldberger, G. Hinton, S. Roweis, R. Salakhutdinov, Neighborhood component analysis, in: L. K. Saul, Y. Weiss, L. Bottou (Eds.), Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 2005, pp. 513–520.
[16] B. Kulis, Metric learning: a survey, Found. Trends Mach. Learn. 5 (4) (2012) 287–364.
[17] A. Bellet, A. Habrard, M. Sebba, Metric learning, Synth. Lect. Artif. Intell. Mach. Learn. 9 (1) (2015) 1–151.
[18] I. T. Jolliffe, Principal Component Analysis, Springer, New York, 2002.
[19] N. Shental, T. Hertz, D. Weinshall, M. Pavel, Adjustment learning and relevant component analysis, in: Proceedings of the 7th European Conference on Computer Vision, ECCV ’02, London, UK, 2002, pp. 776–792.
[20] W. Zhao, Discriminant component analysis for face recognition, in: Proceedings 15th International Conference on Pattern Recognition., Vol. 2, 2000, pp. 818–821.
[21] M. Sugiyama, Local fisher discriminant analysis for supervised dimensionality reduction, in: Proceedings of the 23rd International Conference on Machine Learning, ACM, New York, NY, 2006, pp. 905–912.
[22] B. Nguyen, C. Morell, B. D. Baets, Supervised distance metric learning through maximization of the jeffrey divergence, Pattern Recognition 64 (2017) 215–225.
[23] S. Boyd and L. Vandenberghe, Cambridge University Press, New York, USA, 2004.
[24] Y. Ying, P. Li, Distance metric learning with eigenvalue optimization, J. Mach. Learn. Res. 13 (2012) 1–26.
[25] Y. Shi, A. Bellet, F. Sha, Sparse compositional metric learning, in: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014, pp. 2078–2084.
[26] D. Wang, X. Tan, Bayesian neighborhood component analysis, IEEE Transactions on Neural Networks and Learning Systems (99) (2018) 1–12.
[27] L. Torresani, K.-C. Lee, Large margin component analysis, in: Advances in Neural Information Processing Systems, Vol. 19, 2007.
[28] P. Yang, K. Huang, C.-L. Liu, Multi-task low-rank metric learning based on common subspace, in: B.-L. Lu, L. Zhang, J. Kwok (Eds.), Neural Information Processing, Springer, 2011, pp. 151–159.
[29] J. M. Richardson, Pattern recognition and statistical mechanics, Journal of Statistical Physics 1 (1969) 71–88.
[30] J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, Proceedings of the National Academy of Sciences 79 (1982) 2553–2558.
[31] G. E. Hinton, S. Osindero, Y. W. Teh, A fast learning algorithm for deep belief nets,” neural computation, Neural computationSciences 18 (2006) 1527–1554.
[32] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, Equation of state calculations by fast computing machines, The Journal of Chemical Physics 21 (6) (1953) 1087–1092.
[33] J. S. Thomsen, Logical relations among the principles of statistical mechanics and thermodynamics, Phys. Rev. 91 (1953) 1263–1266.
[34] S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, Optimization by simulated annealing, Science 220 (4598) (1983) 671–680.
[35] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.
[36] M. Lichman, UCI machine learning repository (2013). URL http://archive.ics.uci.edu/ml
[37] A. R. Group, Learning and Artificial Neural Networks(2016). URL http://www.uco.es/grupos/ayrna/index.php?lang=en
[38] F. Rong-En, LIBSVM data: Classification, regression, and multi-label (2011). URL https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/
[39] M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim, Do we need hundreds of classifiers to solve real world classification problems?, Journal of Machine Learning Research 15 (2014) 3133–3181.
[40] J. Demsar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.
[41] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics Bulletin 1 (6) (1945) 80–83.
[42] M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, Ann. Math. Stat. 11 (1940) 86–92.
[43] O. Dunn, Multiple comparisons among means, J. Am. Stat. Assoc. 56 (293) (1961) 52–64.
[44] W. Hollander, D. A., Nonparametric Statistical Methods, 1999.
[45] D. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, 2007.
Figure 6: Evolution of the free energy for the annealing procedure (red) and for multiple quenching (shades of blue). The quenching procedure is repeated 20 times with different initializations of the distance matrix.

Figure 7: Classification error of the competing methods as a function of the number of nearest neighbors.

Table 5: Statistics of the number of nearest neighbors that results in best classification for competing methods.

| Measure | Euclidean | PCA | RCA | DCA | LFDA | DMLMJ | LMMN | ITML | NCA | DMLFE |
|---------|-----------|-----|-----|-----|------|-------|------|------|-----|-------|
| Mean    | 9.77      | 11.97 | 10.93 | 16.63 | 12.73 | 13.4 | 11.6 | 10.5 | 14.3 | 9.3   |
| Median  | 6         | 7.5  | 8.5 | 14 | 8.5 | 8 | 8 | 5.5 | 14.15 | 8     |
| Rank    | 2.77      | 4.33 | 3.97 | 6.87 | 3.93 | 5.6 | 5 | 3.87 | 5.53 | 3.67  |