Active Metric Learning for Supervised Classification

**Krishnan Kumaran, Dimitri Papageorgiou Laurens Lueg, Nicolas Sahinidis**

ExxonMobil Corporate Strategic Research Department of Chemical Engineering

Clinton, NJ 08869 Carnegie Mellon University

*krishnan.kumaran@exxonmobil.com* Pittsburgh, PA 15213

dimitri.j.papageorgiou@exxonmobil.com sahinidis@cmu.edu, llueg@andrew.cmu.edu

Abstract

Clustering and classification critically rely on distance metrics that provide meaningful comparisons between data points. While extensive past work exists on optimal metric selection (usually restricted to the Mahalanobis metric – generalized Euclidean with orthonormal axis rotation and scaling), these approaches are challenged by typical traits of realistic datasets which could contain multiple “islands” belonging to the same class, non-convex class boundaries and the presence of outliers/misclassifications. Further, they do not provide experiment design procedures for selective acquisition of new data to improve classification performance, or online methods to systematically build a minimal but complete dataset.

In this work, we present a novel set of mathematical formulations and solution methods aimed at overcoming the above-described challenges. In particular, we propose methods to construct optimal metrics – first using a given fixed dataset - followed by experimental design procedures for new data selection that manage both metric complexity and data requirement. Starting with a mathematical form for a generalized metric, the Mahalanobis metric being a special case that retains only the first term in a metric expansion, we present a mixed integer optimization (MILP) approach to solve for the optimal metric to minimize misclassification. Specializing to the Mahalanobis case, we generalize and improve upon leading methods by removing reliance on pre-designated “target neighbors,” “triplets,” and “similarity pairs”. We then propose methods to identify “boundary points” of classes, which can be ranked in priority order to guide new data selection in combination with outliers, which are also identified as part of the metric optimization. This form of targeted data acquisition can significantly reduce computational burden by ensuring training data completeness, representativeness, and economy.

Finally, we demonstrate classification and computational performance of the algorithms through simple and intuitive synthetic examples, to be followed in the full final version of this report by results on real datasets from image processing, medical data and chemical engineering applications.

Keywords: Clustering, supervised classification, metric learning, active learning.

Introduction

Selecting an appropriate distance metric is fundamental to many learning algorithms such as clustering, nearest neighbor searches, SVM and others, as observed by (Davis et al., 2007) and other researchers in this field. Further, they observe that choosing such a measure is highly problem-specific and ultimately dictates the success - or failure - of the learning algorithm. To deal with this issue, most of these methods introduce a user-selected *kernel*, typically a non-linear transformation of the space that makes the classification easier by segregating points of the same class into the same region of the transformed space, as shown in the picture below.



However, this type of transformation often requires substantial distortion of the original space, and as we will show in our approach, is unnecessary for good classification performance. Approaches such as Artificial Neural Networks and Random Forests can deal with non-linear and non-convex classes with multiple domains while providing good classification when “sufficient” quantities of training data are available, however they remain “black-boxes” with an unpredictable set of internal parameters with unknown significance, and often do not provide guidance on economic data selection to achieve high classification performance.

In this work, we show that it is possible for the metric selection algorithm to be more general while preserving transparency and interpretability. Past approaches to metric selection, while transparent, have not demonstrated the performance and flexibility of the highly parametrized black-box approaches like Neural Networks. Further, they have been challenged by prohibitive computational requirements for large datasets in high dimensional spaces. To address these issues, we propose a new mathematical framework for choosing metrics that improves and extends previous formulations in some important ways. In particular, we attempt to couple the metric learning problem with that of recommending targeted data acquisition that manages data size, which has not been sufficiently addressed in past work which has mostly assumed that the data is a given, static collection of N-dimensional vectors. However, in many real-world settings, one does not have the luxury of learning a once-and-for-all distance metric. Rather, an iterative approach is required whereby an initial distance metric is learned, new data is selected based on the metric and the metric is continuously refined. A main goal of this work is to present a systematic framework for optimizing this iterative procedure so that an optimal and interpretable metric is learned and is, in turn, used to recommend precise regions to sample in order to acquire new data to be used to further refine the metric and improve classification performance.

**Nearest-Neighbor Classification Criterion**

As described in (Xiang et al., 2008), there are two prominent batch distance metric learning settings, both of which assume that we are given a set of $N$ points $x\_{i}\in R^{D}$. In the first setting, a class (or label) $C\_{i}$ is explicitly given for each point $i\in N=\{1,…,N\}$. In the second setting, classes are implicitly furnished through pairwise constraints in the form of must-links and cannot-links. Must-links are given as $\{\left(i,j\right):i and j are in the same class\}$, whereas cannot-links are specified as $\{\left(i,k\right):i and k are not in the same$ $ class\}$. For both settings, we let $C\_{i}$ and $\overbar{C}\_{i}$ denote the co-class and non-class neighbors of $i$, respectively.

We ask the question: Is there a metric $D(x,y) $that enforces the condition that the nearest neighbor of each point is a co-class point, i.e.

 (1)

This formulation can be easily generalized to *k*-nearest neighbors by modifying the condition to require that the majority of *k* nearest points be co-class. Note that the above condition does not a priori define target neighbors, as required by most previous work. This is an important distinction because the closest neighbors of a point are not determined unless the metric is specified. Our formulation incorporates variables that compare distances between true neighbors contingent on the distance metric. This property avoids the pitfalls of pre-specified target neighbors as shown below, while preserving the desirable characteristics of agglomerative and K-nearest neighbor clustering methods such as permitting multiple disjoint islands of the same class and non-convex class regions while maintaining simplicity and interpretability of the metric.

**Form of Distance Metric**. In general, we allow the metric to be a power series of the form

𝔻(𝒙,𝒚) = 𝐚∙|𝒙−𝒚| + (𝒙−𝒚)∙𝑩∙(𝒙−𝒚) +…

higher order terms

Higher order terms involve polynomial products of coordinate differences, e.g. third order term would introduce a triple-index symmetric tensor as



Strictly speaking, the distance measure $D$ may not satisfy the four properties - non-negativity, symmetry, triangle inequality, distinguishability - that are required to be a “metric”, nevertheless we use the term metric for brevity. Restricting to the first term loosely corresponds to SVM/discriminant analysis, while the second term is commonly known in the literature as the Mahalanobis metric if the matrix B is symmetric and positive definite. Higher order terms introduce additional parameters at a power law rate, e.g., the fully symmetric tensor $C\_{ijk}$ has $O(D^{3})$ parameters. In principle, almost any dataset can eventually be fitted using a metric with a potentially infinite number of these parameters. However, in practice, a misclassification trade-off curve and knee-point of diminishing returns can be used to prevent overfitting. Further, the misclassified points could point to possible outliers, errors in input data, or class boundaries. In all of these cases, the algorithm points to regions in the space where further data acquisition/quality testing would be of most value. This aspect of our method is unique, and provides significant value in selecting the most effective training for supervised classification in general, even applied to SVM/Deep Neural Networks or other algorithms. Further, we will show in our experimental results that the ratio of closest co-class to closest non-class point, defined as

$$R\_{i}=\frac{\min\_{j\in Class\_{i}}D\_{ij}}{\min\_{k\notin Class\_{i}}D\_{ik}}$$

is a useful parameter to separate “interior” points from “boundary points” of classes. Even in the case where we consider only the second-order term, the Mahalanobis distance, our approach differs from past approaches due to our fundamental classification criterion (1) which we show empirically results in better solutions. The figure below summarizes the results of our approach on synthetic data for the Mahalanobis metric, and also demonstrates how selecting boundary points only results in similar optimal metric as full data, providing significant data economy.



**Figure 1**: Left column shows 2-d synthetic datasets. Middle column represents Mahalanobis metric as ellipse – principal axes show coordinate scaling to maximize class separation. R-ratio histogram reveal (top right) uses knee point to identify class boundaries (bottom left). Optimal metrics (middle column) similar in both cases. Squished figures on top of each ellipse shows data represented in scaled and rotated coordinates corresponding to optimal Mahalanobis metric.



**Figure 2**: Synthetic example of a 6-D dataset with 3 classes. Top figures show pictorial representation of class points – x-axis is dimension number and y-axis stacks all points within each class. The pattern of the data within classes, as well as the differences between classes, can be visualized much more easily on the scaled coordinates from our optimized metric (middle row). Vertical stripes in these figures shows each class has relatively uniform coordinate values in the transformed metric space.



**Figure 3**: Application to handwriting recognition shows digit examples belonging to interior, boundary and outliers of classes as extracted from optimal metric. Results conform to intuition.

Acknowledgments

We acknowledge contributions from Prof. Martin Takac of Lehigh University and graduate students Bolun Zhang, Yutong Chang and Minhan Li in early stages of this effort.

References

1. **Davis, Jason V, Kulis, Brian, Jain, Prateek, Sra, Suvrit, and Dhillon, Inderjit S**. *Information-theoretic metric learning*.In Proceedings of the 24th international conference onMachine learning, pp. 209–216. ACM, 2007.
2. **Rosales, R´omer and Fung, Glenn**. *Learning sparse metrics via linear programming*. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 367–373. ACM, 2006.
3. **Wang, Fei and Sun, Jimeng**. *Survey on distance metric learning and dimensionality reduction in data mining*. Data Mining and Knowledge Discovery, 29(2):

534–564, Mar 2015. ISSN 1573-756X. doi: 10.1007/ s10618-014-0356-z. URL https://doi.org/10.

1007/s10618-014-0356-z.

1. **Weinberger, Kilian Q and Saul, Lawrence K**. *Distance metric learning for large margin nearest neighbor classification*. Journal of Machine Learning Research, 10(Feb): 207–244, 2009.
2. **Xiang, Shiming, Nie, Feiping, and Zhang**, **Changshu**i. *Learning a mahalanobis distance metric for data clustering and classification*. Pattern Recognition, 41(12):3600 – 3612, 2008. ISSN 0031-

3203. doi: https://doi.org/10.1016/j.patcog.2008.05.

018. URL http://www.sciencedirect.com/

science/article/pii/S0031320308002057.

1. **Xing, Eric P, Ng, Andrew Y, Jordan, Michael I, and Russell**, **Stuart**. *Distance metric learning with application to clustering with side-information*. Advances in neural information processing systems, 15:505–512, 2003.
2. **Ying, Yiming and Li, Peng**. *Distance metric learning with eigenvalue optimization*. Journal of Machine Learning Research, 13(Jan):1–26, 2012.