ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

Metric Learning with Submodular Functions

Jiajun Pan and Hoel Le Capitaine

University of Nantes, LS2N UMR CNRS 6004, Nantes, France

Abstract. Metric learning mainly focuses on learning distances (or similarities) that use single feature weights with L_p norms, or pair of features with Mahalanobis distances. In this paper, we consider higher order interactions in the feature space, with the help of submodular set-functions. We propose to define a metric for continuous features based on Lovasz extension of submodular functions, and then present a dedicated metric learning approach. This is naturally at the price of higher complexity so that we use k-additive fuzzy measures to decrease this complexity, by reducing the order of interactions that are taken into account. This approach finally gives a computationally feasible problem. Experiments on various datasets show the effectiveness of the approach.

1 Introduction

Since the seminal paper of Xing et al. [1], metric learning has attracted a lot of interest in the machine learning community. It is now widely known that using a convenient metric in machine learning algorithms is fundamental [2, 3]. A common practice consists in considering the Mahalanobis metric defined by

$$D_M^2(x, y) = (x - y)^T M(x - y), \tag{1}$$

where x and y are d-dimensional vectors, M is a positive semi-definite matrix that can be learned.

In practice, however, the distribution of the data has been often complex, so that non-linear approaches have been proposed. In [4], they directly model non-linear metrics with a discriminative objective, while the authors of [5] propose a kernelization of a linear metric. In [6], they extend the linear metric learning approach LMNN to χ^2- distances specialized for histogram data and give a gradient boosted LMNN for non-linear mapping combined with a traditional Euclidean distance. Having a closer look at the Mahalanobis metric shows that it consists in giving weight to all possible feature pairs. The use of the inverse of the covariance matrix for M, i.e. the historical Mahalanobis distance, implies that the weight of a feature pair is proportional to the cofactor of the features. Although the cofactor of a pair of features depends on all other pairwise covariances, the actual distance definition only considers the pair-wise combination of features, whereas d-tuple-wise combinations bring a lot more information.

We are investigating the possibility of giving (and learning) weights to coalitions of features whose cardinal can be greater than two. This clearly allows high order interactions between features, at the price of higher complexity than regular Mahalanobis based approaches. We will show that we can decrease this

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

complexity thanks to constraints on the optimization problem, making the problem computationally feasible, even for the moderately large dimensional problem. Moreover, the complexity of the problem may become independent of the volume of the data.

We consider a class of set-functions $f(S) : 2^V \rightarrow [0, 1]$, that maps subsets S of a ground set V to unit interval values. Note that, in general, the codomain of f is not restricted to be the unit interval, but belongs to $\mathbb{R}$. In the sequel, we will denote $d = |V|$ the dimension of the ground set. This definition allows associating weights to subsets, in our case subsets of features.

We propose to use set-functions, and in particular submodular set-functions, in order to weight coalition of features. By a minimum norm point algorithm used on submodular functions, we propose a linear programming approach for learning the new metric considering high order interactions between features.

2 Preliminaries

The general formulation of supervised metric learning using the Mahalanobis distance D^2_M (see Eq. 1) is to find M such that it minimizes $\mathcal{L}(M) = \ell(M, \mathcal{C}) + \lambda R(M)$, where ℓ is a loss function penalizing unsatisfied constraints, with $\mathcal{C}$ the set of constraints. Constraints are selected by splitting the samples into a similar set $\mathcal{S}$ containing pairs with same target label, and a dissimilar set $\mathcal{D}$ with different labels. λ is a trade-off parameter between the regularization term and the loss, and $R(M)$ is a regularization function. If feasible, this model is generally casted as a constrained optimization problem $\min R(M)$ while s.t. $\ell(M, i) \leq 0, \forall i \in \mathcal{C}$

The objective of this paper is not to present state of the art metric learning algorithms, and we refer the reader to recent surveys [7] for more details about historical and new methods dedicated to metric learning using D_M.

Following usual metric learning formulation, we use the set-function f with a newly defined metric $D^2_f(x, y) = L_f((x - y)^2)$ for the ability to weight the d-tuple-wise combination of features, and using relative constraints $\mathcal{R}$ for the following optimization problem.

$$min_f \sum_{(i,j,k)\in\mathcal{R}} \ell(i, j, k) + \lambda R(f), \tag{2}$$

where R is the regularizer on f, and ℓ is the hinge loss defined as $\ell(i, j, k) = [\gamma + D^2_f(x_i, x_j) - D^2_f(x_i, x_k)]_+$. In the sequel, and following earlier works, the margin γ is set to 1.

The core part of the new metric D_f is the Lovasz extension L_f defined by the set-function f. The Lovasz extension [8] (also known as the Choquet integral), allows extending a set-function defined on the vertices of the unit hypercube to the full unit hypercube $[0, 1]^d$. Another appealing property of the Lovasz extension is its ability to draw a link between set-functions and convex functions.

The Lovasz extension L_f of x with respect to a set-function f is defined as:

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

$$L_f(x) = \sum_{p=1}^{d} x_{(p)} \left[f(\{q|x_{(q)} \geq x_{(p)}\}) - f(\{q|x_{(q)} \geq x_{(p+1)}\}) \right] \tag{3}$$

where $(\cdot)$ denotes a nondecreasing permutation of the input vector x such that $x_{(d)} \geq \cdots \geq x_{(1)}$ and $x_{(d+1)} = \infty$ by convention.

The Lovasz extension allows to set weights to subsets, and we now turn to its use for defining a metric. It is well known that if N is a norm, then $D(x, y) = N(x - y)$ is a metric. Consequently, defining a metric with the Lovasz extension reduces to prove that the Lovasz extension defines a norm, given some conditions on f.

A norm is a function $N : \mathbb{V} \to \mathbb{R}^+$ on a vector space $\mathbb{V}$ satisfying the following conditions: $N(x) = 0 \Leftrightarrow x = 0$ for separates points, $N(ax) = |a|N(x) \forall x \in \mathbb{V}$ $\forall a \in \mathbb{R}$ for absolute homogeneity, and triangular inequality $N(x) + N(y) \geq N(x + y)$ $\forall x, y \in \mathbb{V}$.

According to [9, 8], we know that a set-function f is submodular, if and only if $L_f(x)$ is convex. As in [10], the convexity of $L_f(x)$ implies convexity of $L_f(|x|)$ (by composition of convex non-decreasing functions), and by convexity of $L_f(|x|)$ we have that $L_f(|x|)$ is a norm.

We now specifically define the Lovasz Extension Metric D_f using the squared Lovasz Extension norm as follow:

$$D_f^2(x, y) = L_f\left((x - y)^2\right) \tag{4}$$

Note that Equation (4) can be easily generalized on p-Lovasz Extension norms, exactly the same way as L_p norms in Euclidean spaces.

3 Learning Lovasz Extension Metric

As can be seen in Equation (2), we use the general metric learning formulation of relative constraints with the new define Lovasz extension metric. The condition of Lovasz extension metric is the set function f should be submodular. Written as a constrained optimization problem, we obtain

$$\begin{aligned} &\min R(f) \\ &\text{s.t. } \ell(i, j, k) \leq 0, \forall (i, j, k) \in \mathcal{R} \\ &\quad\ f \text{ is submodular} \end{aligned} \tag{5}$$

Although we are aware that one can consider sparse LP solutions [11] to tackle this problem, we do not consider this family of approaches in this paper. Naturally, it can be used to further improve our proposition. Let us use the following vector notation for the set-function

$$\mathbf{f} = (f(\{1\}), f(\{2\}), \cdots, f(\{1, 2\}), \cdots, f(\{1, \cdots, d\}))^T$$

.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

Straightforward manipulation of the Lovasz extension w.r.t. the set-function $\mathbf{f}$ leads to the following expression, $D_f^2((x_i, x_j)) = \mathbf{a}_{ij}^T \mathbf{f}$, where:

$$\mathbf{a}_{ij} = \begin{pmatrix} (x_{i_{(1)}} - x_{j_{(1)}})^2 \\ (x_{i_{(2)}} - x_{j_{(2)}})^2 \\ \cdots \\ (x_{i_{(2^d-1)}} - x_{j_{(2^d-1)}})^2 \end{pmatrix}, \tag{6}$$

where (.) is the permutation defined within (3).

Therefore, the first constraint in (5) can be written as the inequality $M^T \mathbf{f} + \mathbf{b} \leq 0$, where $\mathbf{b}$ is the constant margin vector γ, and $A = \left(\mathbf{a}_{ij}^1 - \mathbf{a}_{ik}^1, \cdots, \mathbf{a}_{ij}^m - \mathbf{a}_{ik}^m\right)$, corresponding to the m constraints of $\mathcal{R}$. The submodularity of f can also be written as an inequality. In particular, by using a matrix of $\{-1, 0, 1\}$ values, one can write each of the $\frac{1}{2}2^d(2^d + 1)$ submodular constraints.

Because L_f is linear in f, the problem (5) can be finally written as a linear inequality program:

$$\min \mathbf{f}^T \mathbf{r} \tag{7}$$

$$\text{s.t. } C^T \mathbf{f} + \mathbf{t} \leq 0 \text{ and } 0 \leq \mathbf{f} \leq 1$$

$$\text{where } C = \begin{pmatrix} A \\ S^T \end{pmatrix}, \mathbf{t} = \begin{pmatrix} \mathbf{b} \\ \mathbf{0} \end{pmatrix}, \tag{8}$$

and $\mathbf{r}$ is the unit $2^d - 2$ dimensional vector. In practice, all constraints cannot be satisfied with real data, so that we introduce non-negative slack variables ξ_i for each of these constraints. We subsequently add the penalty term $\alpha \sum_{i=1}^m \xi_i$ to $\mathbf{f}^T \mathbf{r}$, where α is a trade-off parameter (set to 1 in our experiments). Solving the revised program gives the solution denoted as L_f hereafter. Clearly, this problem does not scale with the dimension of the data. The number of values to be learned, for a d-dimensional dataset is $2^d - 2$. Furthermore, as indicated earlier, the number of constraints for verifying submodularity is $\frac{1}{2}2^d(2^d + 1)$.

In order to deal with this problem, we use k-additive fuzzy measure, see [12], to set-functions to simplify the optimization problem. A fuzzy measure f is called k-additive if its Mobius transform θ verifies $\theta(S_A) = 0$ for any subset S_A with more than k elements $|S_A| > k$, and there exists a subset S_B with k elements such that $\theta(S_B) \neq 0$. If a set-function is k-additive, it implies that there are no interactions between subsets of more than k elements. According to [12], using 2-additive fuzzy measure formulation as a limited submodularity representation, the number of constraints decreases to $\frac{1}{8}2^d(d^2 - d)$ for d-dimensional dataset, which is much more reasonable for practical problems. In this paper, we show more result with the k-additive fuzzy measure with k greater than 2. Using this proposition gives the solution denoted as L_f^k hereafter.

4 Experiments and Results

Now, we conduct experiments which demonstrate the performance, and in particular the classification generalization performance, of the proposed method of

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

Data	Euc.	LMNN	ITML	LSML	LFDA	GMML	GBNN	L_f	L_f^k
bala.	72.66	78.86 (16.95)	77.17 (7.35)	73.82 (0.01)	80.22 (0.01)	80.34(0.42)	68.90(0.23)	81.02 (0.01)	**81.12** (0.01)
digits	93.77	94.33 (254.0)	90.94 (0.71)	92.83 (0.01)	94.10 (0.07)	94.73(1.76)	**94.87**(4.60)	93.89 (126.0)	94.05 (2.60)
glass	61.02	65.21 (4.19)	57.24 (7.53)	64.27 (3.95)	58.17 (0.01)	64.86(0.32)	66.79(0.39)	**68.70** (0.94)	68.17 (0.09)
iono	85.75	87.17 (6.58)	85.18 (6.12)	86.04 (0.09)	77.18 (0.09)	87.56(0.16)	**94.32**(2.32)	86.61 (150.9)	88.60 (2.54)
liver	66.48	63.87 (36.11)	62.26 (7.55)	65.70 (5.43)	66.14 (0.02)	64.72(0.85)	66.48(3.52)	66.48 (0.01)	**66.59** (0.01)
seeds	82.57	88.52 (2.45)	87.62 (14.30)	88.10 (2.71)	89.48 (0.01)	88.67(0.61)	88.10(2.42)	**90.95** (0.02)	90.49 (0.01)
sonar	50.87	55.69 (2.80)	48.92 (3.11)	51.53 (0.02)	52.86 (0.01)	55.83(1.04)	**66.76**(2.07)	56.63 (124.9)	59.02 (2.17)
segm.	78.10	82.52 (76.75)	80.29 (1.26)	**85.67** (0.05)	83.61 (0.01)	83.34(1.03)	83.42(2.12)	84.38 (168.8)	83.81 (2.76)

Table 1: Accuracy of KNN with different metrics learning algorithm and their running time in seconds.

metric learning on some real-world datasets. We use 8 data sets from the UC Irvine Machine Learning Repository, which are Seeds, Sonar, Balance, Glass, Digits, Liver Segment, and Ionosphere.

We compare the results obtained with the proposed method against several state-of-the-art linear and non-linear metric learning algorithms: LMNN [3], ITML [2], LSML [13], LFDA [14], GMML[15] and GBNN [6] using 10 fold cross validation on the task of K-nearest neighbors classification, with $K = 5$ (other values for K were tested, without significant modification)

Finally, we also give the results obtained without metric learning, i.e. the Euclidean distance for which $M = Id$. In the first part of the experiments, we are using the first model L_f, that is using all possible orders of feature interactions. In particular, the only constraints are related to submodularity and relative distance constraints. As mentioned earlier, due to the complexity of the model, our first proposition L_f is not able to process datasets for which the dimension is (even moderately) large. Consequently, we first use a PCA on the data whose dimension is greater than 10: sonar, ionosphere, digits, and segment, for which the lost variance is 12.02, 21.97, 26.26 and 0.008, respectively. The other datasets remained unchanged.

Accuracy (and running times) obtained on the 8 datasets for each method are given in Table 1. As can be seen, the proposed L_f generally performs better than all the other metric learning algorithms (with the notable exception of Ionosphere and Segment datasets). More precisely, given the rank of averaging accuracy of each method, we obtain the following ranking $L_f \succ GBNN \succ GMML \succ LFDA \sim LMNN \succ LSML \succ ITML \succ Id$. Note that for low dimensional datasets, the running time of the proposed method is low, and quickly increases with the dimension of the data.

The second part of the experiments uses the modified L_f^k. We also use the k-additive constraint on f in order to decrease the complexity. Increasing k adds orders of interaction, and finally reaches the order of interaction tackled by the first L_f approach. It can be noted that each time we decrease k, the number of free parameters of f is divided by 2, so that running time of the method is now very reasonable, even for quite large dimensional data. Table 1 also gives the results obtained through a grid search of k (last column). Interestingly, we can see that L_f^k often gives better results than L_f, showing that using all the d-tuple-wise combinations are not always useful, and may even penalize the

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

performances (e.g. balance, ionosphere, liver, and sonar).

5 Conclusions and Future Works

In this paper, we present a new metric distance based on the Lovasz extension of a submodular set-function and give the necessary conditions for defining a proper metric. Then, we present a linear program allowing to learn this metric and some variations around the constraints imposed on the set-function. Experiments show the efficiency of the proposition on rather low dimensional datasets, by outperforming state-of-the-art metric learning approaches in terms of accuracy. Potential future work will consist of improving the complexity of the algorithm.

References

[1] Eric P Xing, Andrew Y Ng, Michael I Jordan, and Stuart Russell. Distance metric learning with application to clustering with side-information. In *NIPS*, volume 15, page 12, 2002.

[2] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *ICML*, pages 209–216. ACM, 2007.

[3] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 10(Feb):207–244, 2009.

[4] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, volume 1, pages 539–546. IEEE, 2005.

[5] Ratthachat Chatpatanasiri, Teesid Korsrilabutr, Pasakorn Tangchanachaianan, and Boonserm Kijsirikul. A new kernelization framework for mahalanobis distance learning algorithms. *Neurocomputing*, 73(10):1570–1579, 2010.

[6] Dor Kedem, Stephen Tyree, Fei Sha, Gert R Lanckriet, and Kilian Q Weinberger. Non-linear metric learning. In *Advances in Neural Information Processing Systems*, pages 2573–2581, 2012.

[7] Brian Kulis et al. Metric learning: A survey. *Foundations and Trends® in Machine Learning*, 5(4):287–364, 2013.

[8] László Lovász. Submodular functions and convexity. In *Mathematical Programming The State of the Art*, pages 235–257. Springer, 1983.

[9] Francis Bach. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends® in Machine Learning*, 6(2-3):145–373, 2013.

[10] Mikhail Timonin. Robust optimization of the choquet integral. *Fuzzy sets and systems*, 213:27–46, 2013.

[11] Ian En-Hsu Yen, Kai Zhong, Cho-Jui Hsieh, Pradeep K Ravikumar, and Inderjit S Dhillon. Sparse linear programming via primal and dual augmented coordinate descent. In *NIPS*, pages 2368–2376, 2015.

[12] Michel Grabisch. *Set Functions, Games and Capacities in Decision Making*. Springer, 2016.

[13] Eric Yi Liu, Zhishan Guo, Xiang Zhang, Vladimir Jojic, and Wei Wang. Metric learning from relative comparisons by minimizing squared residual. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 978–983. IEEE, 2012.

[14] Masashi Sugiyama. Local fisher discriminant analysis for supervised dimensionality reduction. In *Proceedings of the 23rd international conference on Machine learning*, pages 905–912. ACM, 2006.

[15] Pourya Zadeh, Reshad Hosseini, and Suvrit Sra. Geometric mean metric learning. In *International Conference on Machine Learning*, pages 2464–2471, 2016.