様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月 ２５ 日現在

機関番号：１４３０１
研究種目：挑戦的萌芽研究
研究期間：2013～2014
課題番号：２５５４０１００

研究課題名（和文）Empirical Bayes Kernels: Unsupervised Kernel Learning

研究課題名（英文）Empirical Bayes Kernels: Unsupervised Kernel Learning

研究代表者
Ｃｕｔｕｒｉ　Ｍａｒｃｏ（Cuturi, Marco）

京都大学・情報学研究科・准教授

研究者番号：８０５９７３４４

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：この研究に置ける目標は、ラベルなしデータの大規模なデータセットを活用するために、教師なしの方法でカーネル/距離学習を行うことが原則的アプローチを提供することでした。私たちは、ヒストグラムデータに主に焦点を当てて、この研究の道を調査しました。 AitchisonとLebanonとHintonによって3既知のアプローチの組み合わせを使用して、我々は最先端のレベルで実行するか、または直接競合するアプローチをアウトパフォーム異なるアルゴリズムを提案することができました。

研究成果の概要（英文）：Our goal in this work was to provide a principled approach to carry out kernel/metric learning in an unsupervised way, to take advantage of large datasets of unlabeled data. We investigated this research avenue by focusing mostly on histogram data (bags-of-features). Using a combination of 3 known approaches by Aitchison, Lebanon and Hinton, we were able to propose different algorithms which perform at state-of-the art level or directly outperform competing approaches.

研究分野：機械学習

キーワード：機械学習　距離学習　ヒストグラムデータ

１．研究開始当初の背景

Data analysts are now confronted to gigantic datasets. These datasets contain for the most part unlabeled data, such as transactional data, text or images that can be harvested at a very little cost on the internet but do not specify labels.

Many machine learning tasks, such as classification or regression, are inherently supervised, meaning that they need labeled data to work. A recent trend of data analysis algorithms collectively known under the name of semi-supervised algorithms, among which deep networks, have proposed to leverage these vast resources of knowledge to improve the performance of classifiers. In this context, the motivation of this research was to use these vast databases to improve the performance of inference algorithms by learning a novel geometry for data, using the framework of kernel methods / metric learning.

２．研究の目的

Our project proposed to study algorithms that could exploit unlabeled data to construct metrics and kernel functions (similarity functions) that can be efficiently applied in machine learning, using for instance nearest-neighbor methods or kernel support vector machines.

After a few experimentations, we felt that we would be able to make the most salient contributions in the field of *metric learning for histograms*. The rationale for this choice was as follows: most supervised and unsupervised approaches to learn a metric for arbitrary data have been confined to standard vectors, despite the fact that a large share of these datasets contain in fact *histograms* or *bags-of-features*.

３．研究の方法

Our contributions have relied on different tools, among which (1) the body of work by Aitchison [a] to define Riemannian metrics on the space of discrete probability measures (2) a geometric formulation by Lebanon [b] which inspired the definition of a criterion to learn a metric in an unsupervised way (3) recent advances in the field of maximum likelihood methods (contrastive divergences [c], pseudo-contrastive divergences) that can exploit samples randomly generated from the candidate model p_ϑ to maximize approximately the log-likelihood of the data given a certain probabilistic model.

４．研究成果

We have proposed three contributions which all have in common a set of mappings proposed by Aitchison [a] to define Riemannian metrics on the space of discrete probability measures.

The first two contributions were published in the ACML conference, with an extended version published in the Machine Learning Journal. The third publication has recently appeared in the International Conference on Machine Learning.

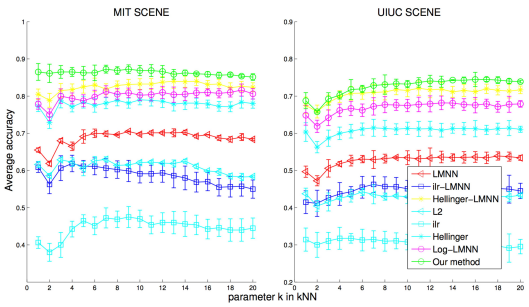
Learning Aitchison Mappings

We proposed in these two papers to address the problem of learning a metric in the probability simplex (a metric for histograms) by generalizing a family of embeddings proposed by Aitchison (1982) to map the probability simplex onto a suitable Euclidean space. The family of maps that we considered was of the following form:

$$\mathfrak{a}(\mathbf{x}) \stackrel{\text{def}}{=} \mathbf{P} \log(\mathbf{x} + \mathbf{b}) \in \mathbb{R}^m.$$

In that formula, **x** is a probability vector (nonnegative components that sum to 1), whereas P (a map) and **b** (a vector) are parameters of the embedding.

We provided in these papers various gradient descent type algorithms to estimate the parameters of such maps. We showed that these algorithms led to representations that outperformed alternative approaches to compare histograms, as can be seen in the results displayed below for two datasets (MIT and UIUC) of scene classification.



In that work, our focus was still on learning metrics using supervised knowledge. The methods proposed in that paper inspired us to pursue additional work, focused this time on learning metrics in an unsupervised setting. This contribution exploits further Aitchison's [a] geometry and is also inspired by deep learning optimization methods [c].

Unsupervised Riemannian Metric Learning

We considered in this latest paper the problem of learning a Riemannian metric on the simplex using unlabeled histogram data. We followed the approach of Lebanon [b], who proposed to estimate such a metric, within a parametric family of metrics, by maximizing the inverse volume of the Riemannian metric computed at each data point in the training set. The intuition behind this reasoning is that distances should move more slowly as they go through areas in the probability simplex in which sampled points are very dense, in order to be the most informative.

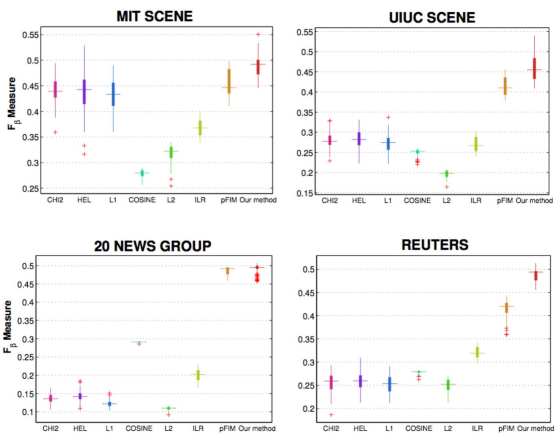
The metrics we consider on the simplex are pull-back metrics of the Fisher information parameterized by operations within the simplex, inspired again by Aitchison's (1982) transformations. Our algorithmic approach to maximize inverse volumes used sampling and contrastive divergences.

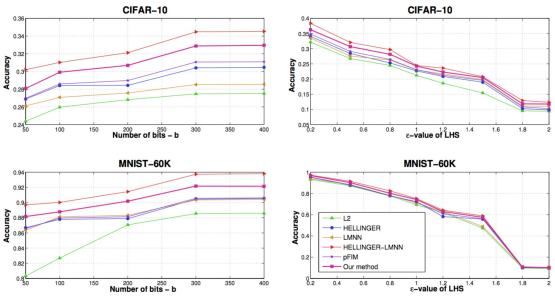
To be more precise, we learned parameters α, λ to learn a map of the following form

$$F(\mathbf{x}) = H \circ G(\mathbf{x}) = \left[\sqrt{\frac{\mathbf{x}_i^{\boldsymbol{\alpha}_i} \boldsymbol{\lambda}_i}{\sum_{j=1}^{n+1} \mathbf{x}_j^{\boldsymbol{\alpha}_j} \boldsymbol{\lambda}_j}} \right]_{1 \leq i \leq n+1} \in \mathbb{S}_n^+.$$

This map associates to a histogram x another histogram $F(x)$ whose weights have been rescaled geometrically and multiplicatively.

Experimental evidence shows that the metric obtained under our proposal outperforms alternative approaches, as can be seen in the figure below, where we used a F measure to compare several metrics, including ours, to measure clustering performance. Our method, plotted in the rightmost column, displays the best performance on these two datasets.





Additional results in the paper illustrate the favorable empirical behavior of our method compared to all other baselines, even when such baselines make actual use of labels, as can be seen in the classification results provided in the figure below.

To summarize, we were able to propose an *unsupervised* metric learning approach that was able to perform at the same level as comparable *supervised* metric learning approachs.

We expect in future research to continue on this trend, and test these approaches on larger datasets, to test some of our hypothesis in a very large scale setting.

[a] Aitchison, J. The statistical analysis of compositional data. Journal of the Royal Statistical Society, 44:139–177, 1982

[b] Lebanon, G. Metric learning for text documents. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(4):497–508, 2006

[c] Hinton, G.E. Training products of experts by minimizing contrastive divergence. Neural computation, 14(8): 1771–1800, 2002.

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　1　件）

T. Le, M. Cuturi, Adaptive Euclidean maps for histograms: generalized Aitchison embeddings, *Machine Learning*, May 2015, Volume 99, Issue 2, pp 169-187.

〔学会発表〕（計　2　件）

T. Le, M. Cuturi, Unsupervised Riemannian metric learning for histograms using Aitchison transformations, *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, JMLR W&CP (to appear), 2015. Acceptance: 270/1037=26%.

T. Le and M. Cuturi. Generalized Aitchison embeddings for histograms. *In Proceedings of the Asian Conference in Machine Learning (ACML), 2013*. Acceptance: 11/100=11% (long oral).

〔図書〕（計　　件）

〔産業財産権〕
○出願状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
https://sites.google.com/site/lttamvn/research/generalized-aitchison-embeddings

６．研究組織
(1)研究代表者
クトゥリ マルコ（Cuturi Marco）

研究者番号：23700172