

Title	Kernelized Evolutionary Distance Metric Learning
Author(s)	Kalintha, Wasin
Citation	大阪大学, 2019, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/72578
rights	
Note	

Osaka University Knowledge Archive : OUKA

<https://ir.library.osaka-u.ac.jp/>

Osaka University

論文内容の要旨

氏名 (KALINTHA WASIN)

論文題名

Kernelized Evolutionary Distance Metric Learning
(カーネル法と進化計算による距離計量学習)

論文内容の要旨

Recently, clustering has played an important role in data mining and machine learning. Semi-supervised clustering is an extension of conventional clustering technique by integrating background information in the clustering, e.g., pairwise constraints or class labels. The conventional way to do semi-supervised clustering is Mahalanobis-based distance metric learning which penalizes objective function using the constraints satisfactory in order to find a suitable metric. Although, state-of-the-art semi-supervised clustering has a rich performance to improve the clustering accuracy by utilizing the class information from human intervention; however, it is reported that hard pairwise constraints, i.e., instance-level constraints, sometimes destroy the clustering quality, depending on relationship between the constraints and the data distribution and there is no monotonicity to the number of constraints, that is the improvement of cluster quality is not guaranteed by adding constraints. These drawbacks are critical issues in practice. Evolutionary distance metric learning (EDML) has been proposed to address the problem of instance-level constraints by directly improve cluster validity index, however, it is categorized as a linear distance metric learning, which yields a small benefit when the data is not linearly separable, like many other distance metric learning techniques. Even though many researchers proposed non-linear distance metric learning, it could not get away from the problem of instance-level constraints.

This study proposes a distance metric learning method which addresses the problem of non-linearly separable data and the problem of instance-level constraints simultaneously. Hence, this research provides an integration of kernelization technique with evolutionary distance metric learning called kernelized evolutionary distance metric learning (K-EDML). The proposed methods are able to handle either class labels or pairwise constraints and directly improve any clustering index as an objective function and can also perform a non-linear distance metric simultaneously. It can be viewed as utilizing cluster-level soft constraints, unlike other instance-level hard constraints which sometimes collapse the clustering. This research demonstrates the performance of the proposed method on UCI dataset compare with other well-known clustering and distance metric learning technique. As a result, the proposed method empirically overcomes other methods in many datasets and secure the highest average ranking in all dataset both in training and test sample. Moreover, the results demonstrate the benefit of kernelization in distance metric learning on the real-world dataset. The advantage of directly optimize the cluster validity index is illustrates by the improvement of cluster quality in EDML and K-EDML from baseline and also state-of-the-art distance metric learning technique. In addition, the proposed method demonstrates generalize performance over the evaluation environment which different from training scheme.

Finally, the proposed method maintains neighbor relation of clusters and can lead to a better visualization of the clustering result. Thus, it can be used as a novel cluster analysis technique that analyzes both class label and features sample simultaneously as a human-centered computing. This method is applied to the real-world problem of facial images and recipes data. The analysis provided promising insights, i.e., more intelligible cluster structure with neighbor relations can be obtained, and a particular cluster structure can be obtained according to the purpose of analysis.

論文審査の結果の要旨及び担当者

氏 名 (KALINTHA WASIN)			
	(職)		氏 名
論文審査担当者	主 査	教授	沼尾 正行
	副 査	教授	藤崎 泰正
	副 査	教授	谷田 純
	副 査	准教授	福井 健一

論文審査の結果の要旨

近年、クラスタリングはデータマイニングや機械学習において重要な役割を果たしている。クラスタリングは観測データ点間の距離に応じた類似データの集合を獲得する技術であるが、そのクラスタリング性能は距離計量の定義に大きく左右される。そこで、部分的に得られる教師情報を活用して、対象毎に適切な距離計量をデータから学習する距離計量学習に関する研究が盛んに行われている。従来の多くの距離計量学習法は、インスタンス（データ点）レベルの制約に基づいて評価関数を設計し、距離計量を最適化するものであった。しかし、インスタンスレベルの制約は、選ばれる制約によってはクラスタリング精度を悪化させる場合があることが知られている。そこで、本論文ではクラスタレベルのソフト制約（クラスタリング妥当性指標）に基づく距離計量学習に着目した。しかし、クラスタリング妥当性指標を評価関数とする場合、超多峰性の関数となるため最適化が困難である。そこで、進化計算の一種である差分進化法によって任意のクラスタリング妥当性指標を最適化する進化的距離計量学習（Evolutionary Distance Metric Learning）がこれまでに提案されている。しかしながら、EDMLは他の距離計量学習と同様にマハラノビス距離に基づくため、距離計量は線形変換に限定されていた。

それに対して本論文第4章では、距離計量学習において線形変換の限界とインスタンスレベル制約の問題を同時に解決する距離計量学習法として、カーネル法と進化計算による距離計量学習（Kernelized Evolutionary Distance Metric Learning : K-EDML）を提案している。サポートベクタマシンで知られるカーネルトリックを用いることで、距離計量の非線形変換を可能とし、同時にEDMLによるクラスタレベルのソフト制約を継承することで両問題に対処している。実験では、データマイニング・機械学習の領域で広く使用されている複数のベンチマークデータセットに対して、提案法（K-EDML）、線形EDML、いくつかの近年の距離計量学習法と比較を行い、提案法は平均して良好なクラスタリング性能が得られることを示した。複数のデータ、異なる条件での広範な比較実験を丁寧に行っており、提案法の有効性を実験的に示している。

続く第5章では、K-EDMLによるクラスタ構造の可視化の例として、顔画像およびレシピデータのクラスタ分析への適用例を示している。K-EDMLは、クラスタリングと同時にクラスタ間の隣接関係をグラフ構造として可視化できる機能を持っている。また、分析の目的に従って制約として与える教師情報を変えることで、属性やカテゴリに応じた特定のクラスタ構造を得ることができる。顔画像およびレシピデータのクラスタ分析において、教師なしのクラスタ分析では発見できなかった、入力データのクラスタ構造とクラスタラベル間の関係性を提案法は視覚的に把握することができることを示した。

以上を要するに、本論文は半教師ありクラスタリングにおける距離計量学習に関する研究を行い、その有効性を確かめたもので、データマイニング・機械学習の分野において情報科学技術の果たす役割の進展に大きく貢献するものである。よって、博士(情報科学)の学位論文として価値のあるものと認める。