

Rapport sur la thèse de Saeid SOHEILY-KHAH

Generalized k-means based clustering for temporal data

Pour obtenir le grade de docteur de l'université Grenoble Alpes
par

Mohamed Nadif, Professeur à l'Université Paris Descartes

Comme la taille des bases de données à traiter sont de plus en plus volumineuses, les synthétiser, y découvrir des informations cachées, les visualiser voire s'en servir pour faire de la prévision sont des objectifs importants dans divers domaines. Ces données peuvent être de différents types et lorsque la fonction temporelle est présente tels que dans le domaine médical, industriel, télécommunication et réseaux de capteurs, des méthodes appropriées sont nécessaires. C'est dans cette problématique que s'inscrit le travail de Mr Saied Soheily-Khah pour traiter le problème de la classification ou *clustering* des séries temporelles. Le manuscrit de thèse est divisé en 3 principaux chapitres, en sus d'un chapitre introductif et un chapitre conclusion et perspectives.

Chapitre 1, présente l'état de l'art sur les similarités, dissimilarités et distances dans le contexte statique et évoque leurs limites dans le contexte temporel notamment pour capturer la similarité entre deux séries temporelles. Pour ce faire, l'auteur commence par rappeler la mesure de **déformation temporelle dynamique (Dynamic Time Warping : DTW)** qui permet de mesurer la similarité entre deux suites qui peuvent varier au cours du temps. A noter que DTW permet d'obtenir un **alignement optimal** entre deux séquences. Plusieurs algorithmes s'appuyant sur la mesure DTW ont été proposés dans la littérature et ont donné des résultats intéressants dans plusieurs domaines dont celui de la reconnaissance vocale. Cependant le DTW conventionnel est trop lent pour la recherche d'un chemin d'alignement pour les grands ensembles de données. Pour y remédier trois variantes ont été décrites utilisant 1) les contraintes (par exemple, bande sakoe-chiba), 2) l'indexation (par exemple, piecewise DTW), et 3) l'abstraction de données (par exemple, multiscale DTW). Ce chapitre comporte également des indices temporels s'inspirant de la corrélation (CORT, DACO). Une présentation sur les méthodes à noyaux est faite pour clôturer cette introduction. On retrouve les mesures communément utilisées par ce type de méthodes : DTAK (Dynamic Temporel Alignment Kernel), GDTW (Gaussian DTW) et GA (Global Alignment). Ce chapitre de 21 pages est très bien écrit, contient l'essentiel sur les méthodes utilisées qui seront pour la plupart évaluées sur des données réelles (Chapitre 4).

Chapitre 2 se focalise sur l'estimation du centroïde d'un ensemble de séries temporelles ; un processus incontournable dans le clustering. L'auteur souligne la difficulté d'un alignement multiple au lieu d'un alignement par paires. Trois approches sont discutées. La première basée sur la programmation dynamique pour déterminer directement un alignement multiple empêche son utilisation à cause de complexité en $O(T^N)$ où N le nombre de séries et T la longueur de chaque série. La deuxième dite progressive consiste à estimer le centroïde en combinant des centroïdes des paires suivant différentes stratégies NLAFF (Non linear Alignment and Averaging Filters), PSA (Prioritized Shape Averaging) ou CWRT (Cross-Words Reference Template). La troisième dite itérative quant à elle propose une méthode globale dont l'objectif est de minimiser une inertie s'appuyant sur DTW. Celle-ci, appelée DBA (Dtw Barycenter Averaging) est dans l'esprit du clustering de type k-means.

Chapitre 3 constitue la contribution principale de cette thèse. Après un rappel bref mais efficace sur les algorithmes des k-means, des k-medoides et à noyaux, l'auteur s'attaque au problème de la classification des séries temporelles via un k-means généralisé s'appuyant sur DTW. Un cadre général est proposé définissant un DTW pondéré (W_{DTW} : Weighted DTW) entre deux séries permettant de généraliser au cas où chaque instant est pondéré. Le calcul des centroïdes se base l'optimisation d'une fonction de coût qui dépend d'une dissimilarité ϕ entre une série et un centroïde, celle-ci est pondérée par une fonction des poids f . Ces deux fonctions conduisent facilement à diverses extensions des mesures classiques utilisées dans les méthodes à noyaux et donnent naissance à WK_{DTAK} , WK_{GDTW} et WK_{GA} . Le problème étant bien posé, il reste à définir l'étape cruciale de calcul des centroïdes qui dépendra des différentes mesures employées. Pour chaque mesure l'optimisation à la fois des centres et des poids est clairement argumentée. Ce chapitre est original et comporte une innovation dans le domaine. Tous les aspects mathématiques sont rigoureusement présentés dans le manuscrit.

Le chapitre 4 traite de l'évaluation des différentes méthodes décrites dans la thèse. L'auteur s'appuie sur 24 benchmarks avec des classes connues. Ces classes sont de différentes formes, différents proportions et différents degré de mélange. La description des données est claire, les méthodes comparées ont toutes été décrites auparavant et les objectifs des expériences menées sont : clustering, temps d'exécution, classification supervisée et enfin un examen approfondi des centroïdes obtenus. En termes de clustering l'indice de Rand est utilisé pour comparer les partitions obtenues par les différents algorithmes. Les méthodes proposées s'avèrent intéressantes et en particulier celle qui s'appuie sur W_{DTW} . Dans le contexte de classification supervisée, un algorithme de type k-plus proches voisin est proposé en s'appuyant sur le calcul des centroïdes sur la base de W_{DTW} , WK_{DTAK} et WK_{GDTW} . La version W_{DTW} ($k=1$) apparaît la plus intéressante. Dans les deux études (clustering et classification supervisée), le choix de l'analyse des correspondances multiples pour synthétiser les différentes informations issues des méthodes et des 24 bases, a été pertinent et pourrait même suggérer une investigation sur la structure des données étant donné la présence de l'effet Guttman. Enfin et sur une étude sur des données simulées, l'auteur montre par examen des différents centroïdes obtenus par les différentes méthodes que les méthodes proposées W_{DTW} , WK_{DTAK} et WK_{GDTW} ont la capacité de surmonter le problème du bruit et de révéler le partage local des séries d'une même classe.

Une conclusion générale reprend l'essentiel des chapitres et donne quelques perspectives à la thèse dont le choix de nouvelles métriques et fonctions de déformation.

Cette thèse est claire et agréable à lire dans l'ensemble, la présence d'une conclusion par chapitre est très appréciable pour le lecteur. L'étude expérimentale est riche et s'appuie sur la statistique inférentielle et l'analyse des données, un élément que j'ai apprécié particulièrement. L'évaluation sur des données présentant différents challenges est intéressante. Cependant, une discussion sur le nombre de classes dans le contexte clustering des séries temporelles et également sur le choix de l'indice de Rand aurait été bénéfique ; pourquoi pas l'indice de Rand ajusté ou la NMI ? D'autre part, une étude sur les proportions des classes, quand le degré de séparabilité des classes est faible, aurait été utile pour mesurer la robustesse des méthodes proposées. Ces critiques n'enlèvent rien à l'intérêt du travail réalisé par Mr Saied Soheily-Khah. Au-delà de l'aspect théorique rigoureusement présenté, le choix des données jusqu'au traitement et l'analyse ont été bien menés et bien commentés. Une diffusion des programmes réalisés serait appréciée dans la communauté.

En conclusion, Mr Saied Soheily-Khah propose une contribution originale et très utile. Le manuscrit qui est rédigé en anglais, est très bien écrit, facile à lire et également très bien structuré. Pour toutes ces raisons je donne un avis très favorable à la soutenance par Mr Saied Soheily-Khah de ses travaux de recherche pour obtenir le grade de docteur de l'université Grenoble Alpes.

Paris, le 20 septembre 2016
Professeur Mohamed Nadif



Laurent BESACIER
Directeur
Ecole Doctorale M.S.T.I.I

Rapport d'évaluation du mémoire de thèse / Evaluation report of the PhD thesis

THES_FOR_10

Doctorant	Nom prénom / Full name	SOHEILY-KHAH, Saeid
PhD student	École doctorale / Doctoral School	ED MSTII
	Titre thèse / PhD Title	Generalized k-means Based Clustering for Temporal Data

Rapporteur	Nom prénom / Full name	HONEINE, Paul
Reviewer	Établissement / Institution	Université de Rouen Normandie
	Statut, fonction / Status, position	Professeur des Universités

Qualité du mémoire, rédaction & illustrations / Thesis quality, style & illustrations

 Satisfaisant / Satisfactory [] Bon / Good [] Très bon / Very good [] Excellent []

Commentaires/comments :

Contexte, état de l'art, collaborations / Background, state of the art, collaborations :

Commentaires/comments :

Qualité scientifique, méthodologie, expérimentations, validation

Scientific quality, methodology, experiments, validation

 Satisfaisant / Satisfactory [] Bon / Good [] Très bon / Very good [] Excellent []

Commentaires/comments :

Apports personnels, originalité, valorisation, perspectives

Personal contributions, originality, valorization, prospects

Commentaires/comments :

Conclusions du rapporteur / Reviewer's conclusions

Commentaires/comments :

Avis du rapporteur / Reviewer's opinion :

Défavorable à la soutenance / Unfavorable to the defence []

Favorable []

Date 14/09/2016

Signature



Laurent BESACIER
Directeur

Ecole Doctorale M.S.T.I.I

Commentaires libres, questionnements, correction demandées
Free comments, questions, requested corrections

Review Report on the Ph.D. Thesis Submitted by
Mr. Saeid Soheily-Khah to the Université de Grenoble
“Generalized k-means Based Clustering for Temporal Data”

Reviewer

Paul Honeine, Professeur des Universités at Université de Rouen Normandie, France

Summary

Mr. Saeid Soheily-Khah studies in his Ph.D. thesis the issue of temporal data mining, with a particular interest on the clustering task in unsupervised learning. When dealing with a set of temporal data in general, the time series may have varying time delays and may be of different lengths. Therefore, clustering temporal data is a challenging problem in machine learning since, prior to any clustering task, it requires aligning simultaneously all the studied time series. State-of-the-art techniques roughly investigate costly time warping with either (kernel) k-means or k-medoids algorithms. The main contribution of this Ph.D. thesis is a novel framework of temporal data clustering methods relying on a weighted formulation that uses kernel-based time warping for extracting global and local features. The derived algorithms to solve the resulting non-convex optimization problems are computationally efficient and their relevance is demonstrated on several well-known challenging benchmarks including non-isotropic and not well-isolated time series datasets.

Overview

The document is well written and the contributions of the Ph.D. candidate are well highlighted. The document consists of four main chapters, as well as an introduction and a concluding chapter. In the following, a summary of the major contributions of the four chapters is given.

The first chapter is essentially a bibliographical survey of the crucial issue of comparing temporal data, with an emphasize on the comparison between pairs of time series. This issue is a particularly complex one due to the dynamic character of the series, since the samples in the times series may have different time delays and thus need to be optimally aligned prior to quantifying any similarity or dissimilarity. Mr. Saeid Soheily-Khah gives a survey of the state-of-the-art comparing metrics using the alignment of the times series, by describing in detail the well-known dynamic time warping method and some of its variants. The review categorizes and explains the conventional proximity measures and the more elaborated kernel-based warping techniques, including the dynamic time warping, the Gaussian dynamic time warping, the kernel global alignment, and the dynamic temporal alignment kernel.

The second chapter, of tutorial nature, is devoted to the problem of estimating the centroid of a set of time series. It addresses the problem of aligning more than two time series, which is a more challenging issue due to the simultaneous alignment of multiple time series. Mr. Saeid Soheily-Khah describes state-of-the-art multiple temporal alignment techniques, grouped into three classes. The first one searches, by using the computationally expensive dynamic programming, the optimal path within a multi-dimensional grid that crosses the times series. The second one determines the global centroid by combining progressively pairs of time series centroids, the price to pay being the early error propagation issue. The third class overcomes this obstacle with a repeatedly refined estimation at each iteration by realigning it to a reference time series. Most of these methods have been limited to the dynamic time warping metric. This chapter provides a fine analysis of the centroid estimation problem and the state-of-the-art solutions. We think that it would have benefited from a concluding table that summaries the state-of-the-art methods with the positive and negative attributes of arguments (*e.g.*, length of centroid, accuracy, computational complexity and memory usage).


The third chapter provides the main contribution of this Ph.D. thesis, which is a novel framework of temporal data clustering that allows to capture global and local temporal features. To this end, Mr. Saeid Soheily-Khah first revisit the formulation of the centroids with time warp alignments, by associating to each centroid a weighting time series that

provides a measure of the relevance of the time stamps of the centroid. The resulting formulation, investigated with the so-called generalized k-means clustering algorithm, extends the standard time warp measures by using a weighted warping function that guides the learned alignment in order to capture local temporal features. The proposed framework is well described, by revisiting the aforementioned warping techniques with several weighting functions, and by providing some theoretical analysis on the pseudo-convexity of the resulting problems. Solutions of the resulting non-convex quadratic constrained optimization problem are provided using alternating optimization, for several pairs of warping techniques and weighting functions.

The fourth chapter provides an extensive experimental analysis on well-known benchmark time series, including non-isotropic (*i.e.*, non-spherical) and not well-isolated (non linearly separable). Mr. Saeid Soheily-Khah examines the performance of the proposed methods on 24 datasets of temporal data available from various fields, with up to 37 clusters per dataset. Some datasets include time series of distinct global behaviors within clusters, making the clustering a very challenging issue. Mr. Saeid Soheily-Khah shows that the proposed weighted centroid formulation allows to capture the characteristics shared locally within the clusters, and outperforms state-of-the-art clustering methods. While the extensive experiments provide qualitative and quantitative analysis, I would have appreciated the analysis of the proposed methods a dataset with time series of variable lengths.

Final remarks

To conclude, the Ph.D. thesis provides very rich contributions in a highly challenging problem in computer science. This work has resulted in papers published in a highly selective international journal (Pattern Recognition Letters), a book chapter and a peer-reviewed international workshop with proceedings. For all the aforementioned reasons, I think that the work of Mr. Saeid Soheily-Khah is of great quality and of wide interest to researchers. Therefore, I give a very positive recommendation for an oral presentation towards the fulfillment of the requirements for the degree of Docteur de l'Université de Grenoble.


Laurent BESACIER
Directeur
Ecole Doctorale M.S.T.I.I

Rouen (France), 10th of August 2016


Paul Honéine

Professeur à l'Université de Rouen Normandie

