Extraction de sous-groupes exceptionnels de séries temporelles

Josie Signe

Extraction de sous-groupes exceptionnels de séries temporelles

Josie Signe[1]

[1] Inria, Univ Rennes, CNRS, IRISA

josie.signe@irisa.fr

Résumé

La tâche de fouille de modèles exceptionnels extrait des sous-groupes intéressants, selon des attributs cibles, dans des données tabulaires. Nous étendons cette approche aux séries temporelles en les utilisant comme attribut cible pour évaluer les sous-groupes. Un sous-groupe est caractérisé à l'aide d'une description en s'appuyant sur un modèle de séries temporelles calculé à partir de toutes les séries du sous-groupe. L'évaluation de la qualité d'un sous-groupe repose sur la différence entre le modèle du sous-groupe et le modèle général, i.e., de toutes les données.

Mots-clés

Fouille de motifs, Fouille de modèles exceptionnels, Séries temporelles.

Abstract

Exceptionnal model mining task finds interesting subgroups according to several target attributes in tabular data. We extend this approach to time series data, using time series as a target attribute to evaluate subgroups. A subgroup is characterized by a description based on a model of time series computed from all series of the subgroup. Evaluation of subgroup quality is based on the difference between the subgroup model and the model of the whole dataset.

Keywords

Pattern mining, Exceptionnal model mining, Time series.

1 Introduction

L'utilisation croissante de capteurs entraîne une multiplication de données mesurées en temps réel. Ces suites de valeurs numériques, appelées *séries temporelles*, permettent de suivre l'évolution des systèmes dans de nombreux domaines (e.g. le cours de la bourse en finance, la température des animaux en agriculture). Une tâche importante concernant les séries temporelles est la détection de séries dont l'évolution est différente de la norme.

La fouille de modèles exceptionnels (EMM) [2] permet d'extraire et de caractériser des sous-groupes *exceptionnels*, i.e., qui se distinguent par rapport à l'ensemble des données sur des attributs dits *cibles*. Chaque sous-groupe est caractérisé par un ensemble d'*attributs descriptifs*, distincts des attributs cibles. EMM a donné des résultats concluant dans plusieurs domaines [2]. Par exemple, dans le domaine de la bioinformatique, EMM a permis de trouver des sous-groupes de gènes intéressants, i.e., dont les niveaux d'expression des gènes (attributs cibles) se distinguent des autres gènes, et de les caractériser par des données cliniques de patients (attributs descriptifs). Il existe plusieurs instances d'EMM permettant de traiter des attributs numériques, nominaux ou binaires mais aucune permettant de traiter des séries temporelles comme attributs cibles.

Dans cet article nous proposons une nouvelle instance d'EMM pour pouvoir traiter des séries temporelles comme attribut cible. Nous présentons un extrait des résultats obtenus lors de nos expérimentations sur des données d'élevage.

2 Travaux antérieurs

Séries temporelles Une série temporelle est une séquence de l valeurs $t =< (v_1, h_1), ..., (v_l, h_l) >$, avec $v_i \in \mathbb{R}$ la valeur horodatée à h_i.

Pour comparer deux séries t_1 et t_2, leur similarité est mesurée avec une fonction de distance $dist(t_1, t_2) \in \mathbb{R}$, calculée entre les valeurs de t_1 et t_2. Une mesure de distance simple entre deux séries temporelles est la distance euclidienne. Elle calcule la distance entre chaque point de t_1 et t_2 qui ont le même horodatage. Une approche plus flexible appelée déformation temporelle dynamique (DTW) [1] prend en considération les décalages temporels et permet d'aligner de manière optimale chaque point des deux séries.

Fouille de modèles exceptionnels La fouille de modèles exceptionnels (EMM) [2] extrait des sous-groupes *exceptionnels* dans une base de données. Un sous-groupe est dit exceptionnel si les valeurs de ses attributs cibles sont très différentes des valeurs de ces mêmes attributs dans la base de données.

Soit D une base de données de n éléments $\{e_1, .., e_n\}$, un élément e est défini par un ensemble d'attributs descriptifs A et d'attributs cibles C tel que : $e = \{a_1, .., a_m, c_1, ..c_k\}$, $m, k \in \mathbb{N}$, $a_i \in A$, $c_i \in C$.

Un *sous-groupe* S de D est décrit par un ensemble de conditions sur les attributs descriptifs de A, appelé *description*. Une description d *couvre* un élément e de D si les attributs descriptifs de e respectent toutes les conditions de d. Le sous-groupe S décrit par d correspond à l'ensemble des éléments appartenant à D qui sont couverts par d.

Les attributs cibles permettent de déterminer si un sous-groupe S est intéressant. EMM modélise la relation qui existe entre les différents attributs cibles du sous-groupe S. Le modèle utilisé pour représenter cette relation est choisi

en fonction du type des attributs. Par exemple, la relation entre attributs cibles numériques pourra être représentée par une régression linéaire alors que celle entre attributs nominaux par des réseaux bayésiens.
Pour chaque sous-groupe, un modèle est calculé sur l'ensemble de ses attributs cibles. Le but d'EMM est de trouver les sous-groupes les plus exceptionnels. Pour les trouver, une mesure de qualité est utilisée pour attribuer un score à chaque sous-groupe. Cette mesure permet de quantifier la différence entre le modèle d'un sous-groupe et le modèle général. Par exemple, pour une régression linéaire, on peut calculer la différence de pente entre les deux modèles.

3 Contribution

Le but de notre approche consiste à trouver des sous-groupes dont les séries temporelles se distinguent dans leur évolution par rapport aux séries temporelles de l'ensemble des données D.
Nous proposons donc une nouvelle instance d'EMM pour laquelle l'attribut cible est une série temporelle. Soit D une base de données de n éléments $\{e_1, .., e_n\}$, un élément e est défini par un ensemble de n attributs descriptifs de A et d'un attribut cible t tel que : $e = \{a_1, ..., a_m, t\}$, $m \in \mathbb{N}$, $a_i \in A$, et t une série temporelle.
Nous avons du définir un modèle représentant un sous-groupe de séries temporelles ainsi qu'une mesure permettant de comparer deux de ces modèles. Le modèle pour représenter un sous-groupe de séries temporelles est une série temporelle elle-même. Dans notre approche, elle est calculée avec la méthode DBA (DTW Barycenter Averaging) [3]. DBA calcule une série temporelle moyenne s d'un ensemble de séries temporelles $S = \{t_1, ..., t_x\}$ qui minimise :

$$\min \sum_{i=1}^{x} dist_{DTW}(s, t_i)^2$$

Cette approche est appliquée à l'ensemble des données D pour générer le modèle général. Puis elle est utilisée pour chaque sous-groupe S afin de calculer leur modèle propre. Lors de l'évaluation d'un sous-groupe S, son modèle est comparé au modèle général. Nous utilisons une mesure de qualité q qui repose sur la distance DTW entre ces deux modèles pour attribuer un score d'exceptionnalité aux sous-groupes :

$$q(S) = dist_{DTW}(DBA_D, DBA_S)$$

avec DBA_D le modèle issu du DBA de l'ensemble des séries temporelles de D, et DBA_S le modèle issu du DBA des séries temporelles du sous-groupe S.

4 Expérimentations et perspectives

Nous avons expérimenté notre approche dans le cadre d'une étude sur le bien-être animal. Le but est de trouver des sous-groupe de vaches qui supportent plus ou moins bien les périodes de forte chaleur en analysant l'évolution de leur température corporelle. Dans cette expérience les éléments de

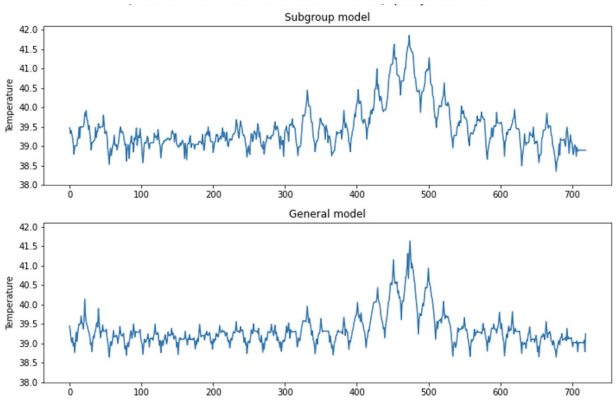

FIGURE 1 – Modèle du sous-groupe {Poids Vif <= 685.9 AND Ration Totale >= 21.48} et modèle général.

D sont des vaches. Elles sont décrites par des attributs descriptifs numériques tel que leur poids ou leur production de lait, et leur attribut cible correspond à leur série temporelle de température corporelle mesurée en période de forte chaleur. Nous avons extrait tous les sous-groupe possible et à la figure 1, nous présentons deux modèles extraits de nos premières expérimentations. Celui du bas est le modèle général calculé sur toutes les données. Celui du haut est le modèle représentant un sous-groupe de vaches dont la description associée est "des vaches dont le poids n'est pas élevé et qui mangent beaucoup". On voit que la température entre les heures 400 et 525 présente des pics plus élevés et qui redescendent moins par rapport au modèle général. Ce sous-groupe décrit donc des vaches sensibles aux fortes chaleurs. Dans nos expérimentations, beaucoup de sous-groupes sont extraits dont certains redondants et de petite taille. Une de nos perspectives est d'intégrer directement la taille des sous-groupes dans la mesure de qualité. De plus, nous envisageons d'explorer des approches de fouille de données s'appuyant sur la théorie de l'information pour sélectionner les sous-groupes les plus intéressants.

5 Remerciements

Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme d'Investissements d'avenir portant la référence ANR-16-CONV-0004

Références

[1] Eamonn Keogh. Exact indexing of dynamic time warping. In *Proceedings of the 28th International Conference on Very Large Data Bases*, page 406–417, 2002.

[2] Dennis Leman, Ad Feelders, and Arno Knobbe. Exceptional model mining. In *Machine Learning and Knowledge Discovery in Databases*, pages 1–16, 2008.

[3] François Petitjean, Alain Ketterlin, and Pierre Gançarski. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3) :678–693, 2011.