Classification de courbes, application aux économies d'énergie

Francisco de Carvalho, Yves Lechevallier, Guillaume Pilot, Brigitte Trousse

Classification de courbes, application aux économies d'énergie

Francisco de A.T. de Carvalho*, Yves Lechevallier**, Guillaume Pilot***, Brigitte Trousse***

*Centro de Informatica - CIn/UFPE, Av. Jornalista Anibal Fernandes, s/n
Cidade Universitária, 50.740-560, Recife - PE, Brésil
fatc@cin.ufpe.br
**INRIA, Paris-Rocquencourt - 78153 Le Chesnay cedex, France
*** INRIA Sophia Antipolis-Méditerranée - 06902 Sophia Antipolis cedex, France
{Brigitte.Trousse,Guillaume.Pilot,Yves.Lechevallier}@inria.fr

Résumé. L'objectif de cet article est de montrer les intérêts et les inconvénients de deux approches classificatoires de courbes. La première est basée sur une représentation sous forme d'intervalles, la seconde utilise une distance basée sur les propriétés mathématiques des courbes (dérivée première et seconde).

Les courbes sont issues de capteurs de température mis dans 40 bureaux durant un an. Cette période a été divisée en périodes de pré et post challenge et la période du challenge. La période de pré challenge contenant une période de chauffage et une période sans chauffage. Durant la période du challenge les occupants avaient une information sous forme de bonus/malus de l'impact de leurs comportements par rapport à la consommation d'énergie.

1 Introduction

En France, le Grenelle de l'Environnement a défini, comme priorité, de parvenir à une réduction importante de la consommation d'énergie, dans tous les domaines, notamment dans le secteur du bâtiment.

Le challenge ECOFFICES tente de répondre à cette problématique d'économies d'énergie.

L'idée du challenge est de mettre en compétition au sein d'une même entreprise des équipes afin de les inciter à améliorer les usages qu'ils font des équipements de chauffage ou de climatisation (via un système de bonus/malus) en développant leurs comportements éco-responsables et ainsi réduire leurs consommations énergétiques sur leur lieu de travail. Le challenge Ecoffices s'est déroulé dans les bâtiments du CSTB (Centre Scientifique et Technique du Bâtiment) de Sophia-Antipolis. La compétition repose, d'une part, sur les informations concernant les consommations énergétiques réelles, et, d'autre part, sur la connaissance des usages des différents équipements de chauffage ou de climatisation. Les bureaux des 3 équipes participantes ont été équipés de plus de 300 capteurs.

L'objectif de cette étude est de réaliser plusieurs approches classificatoires sur diverses représentations des ces courbes de températures.

2 Les données

Toutes les données, statiques et dynamiques, ont été stockées dans la base de données Ecoffices qui est composée de 15 tables contenant environ 9 000 000 de lignes. Nous avons pris uniquement les données issues des capteurs de températures (ambiante, radiateur, ventilo-convecteur) fournissant toutes les 15 minutes une température moyenne.

Ces mesures de températures ont été prises entre le 01/02/2011 et le 31/10/2011 (273 jours) qui ont été divisées en 4 périodes : chauffage et sans chauffage avant le challenge, challenge et après le challenge.

3 Les distances

Sur ces courbes de températures et pour chaque période nous proposons de calculer deux distances ; la première est la distance de D'Urso et Vichi qui se calcule directement sur les points de discrétisation de ces courbes ; la seconde utilise la distance de Hausdorff sur les minima et maxima journaliers sur chacune de ces quatre périodes.

3.1 La distance de D'Urso et Vichi : une mesure qui compare deux courbes

Ces auteurs (**?**) proposent une mesure de dissimilarité calculée sur les différents points de discrétisation de la courbe. Cette mesure est basée sur trois dissimilarités, l'une compare les valeurs de cette courbe aux positions prédéfinies, la seconde mesure la vélocité (la dérivée discrète entre deux positions) et la troisième mesure l'accélération (dérivée seconde discrète).

Soit $\mathbf{x}_i = (x_i(t_1), \ldots, x_i(t_p))$ la i-ème trajectoire sur p points de discrétisation. La vélocité de la i-ème trajectoire est définie par $\mathbf{v}_i = (v_i(t_2), \ldots, v_i(t_p))$, où $v_i(t_j) = \frac{x_i(t_j) - x_i(t_{j-1})}{t_j - t_{j-1}}$. L'accélération de la i-ème trajectoire est définie par $\mathbf{a}_i = (a_i(t_3), \ldots, a_i(t_p))$, où $a_i(t_j) = \frac{v_i(t_j) - v_i(t_{j-1})}{t_j - t_{j-2}}$

La distance de D'Urso et Vichi est une combinaison linéaire pondérée des trois distances euclidiennes calculées sur la position, la vélocité et l'accélération soit $d^2(s_i, s_l) = \alpha_1 \|\mathbf{x}_i - \mathbf{x}_l\|^2 + \alpha_2 \|\mathbf{v}_i - \mathbf{v}_l\|^2 + \alpha_3 \|\mathbf{a}_i - \mathbf{a}_l\|^2$.

3.2 La distance de Hausdorff : une mesure qui compare deux intervalles

La *distance Hausdorff* est usuellement utilisée pour mesurer la proximité entre deux ensembles A et B et elle est définie à partir d'une distance δ définie sur les éléments de A ou B par :

$$\delta_H(A, B) = max\{sup_{a \in A}\ inf_{b \in B}\ \delta(a, b),\ sup_{b \in B}\ inf_{a \in A}\ \delta(a, b)\}.$$

Si A et B sont deux intervalles $[\underline{a}, \overline{a}]$, $[\underline{b}, \overline{b}]$ alors la distance de Hausdroff (voir Chavent (2004)) est égale à : $\delta_H(A, B) = max\{|\underline{a} - \underline{b}|, |\overline{a} - \overline{b}|\}$.

Une variable intervalle Y est une correspondance de l'ensemble E des objets dans $\Re$ qui vérifie la propriété suivante sur son graphe : pour tout individu s le sous-ensemble $Y(s) = [a, b]$ est un intervalle fermé de $\Re$. On notera $\Im$ l'ensemble des intervalles fermés de $\Re$.

Chaque individu s de E est représenté par un vecteur de p intervalles $x_s = (x_s^1, \ldots, x_s^p)$, où $x_s^j, = [a_s^j, b_s^j] \in \Im$. Nous proposons d'utiliser comme distance la somme des p distances de Hausdorff calculées sur chaque intervalle.

4 Problème de classification

Le critère d'adéquation J optimisé par notre algorithme de classification est égal à :

$$J = \sum_{k=1}^{K} \sum_{s_i \in C_k} d_{\lambda_k}(s_i, \mathbf{g}_k) = \sum_{k=1}^{K} \sum_{s_i \in C_k} \sum_{j=1}^{q} \lambda_{kj} \sum_{h \in V_j} \xi_j d_h(x_i^h, g_k^h)) \quad (1)$$

où $d_{\lambda_k}(e_i, \mathbf{g}_k)$ est la dissimilarité globale entre l'objet $s_i \in C_k$ et le prototype $\mathbf{g}_k$ de la classe C_k.

Dans notre application cette dissimilarité globale se décompose en fonction des q périodes. Dans ce cas λ_{kj} est la pondération de la classe C_k pour la période j. V_j et ξ_j sont définis en fonction de la représentation utilisée.

Si les individus sont modélisés par une courbe nous utilisons la distance de D'Urso et Vichi. V_j est l'ensemble des positions, des vélocités et des accélérations avec ξ_j comme pondération. La méthode de classification utilisée est celle décrite dans **?** où le prototype associé à chaque classe est une courbe.

Si les individus sont modélisés sous forme d'un vecteur d'intervalles la distance de Hausdorff est utilisée. V_j est l'ensemble des variables intervalles avec ξ_j comme facteur de normalisation. La méthode de classification est celle qui a été proposée par **?**. Le prototype associé à chaque classe est un vecteur d'intervalles.

La normalisation choisie est celle proposée par **?**. Le bon nombre de classes est déterminé par la procédure décrite dans **?**.

4.1 Algorithme de classification

Au départ les k prototypes sont choisis aléatoirement. L'algorithme alterne, jusqu'à la convergence, les trois étapes décrites dans **?** ou **?** . La valeur stationnaire du critère d'adéquation est un minimum local.

- ***Etape 1 :*** *Recherche de la meilleure partition* $(C_1, \ldots, C_K)$;
- ***Etape 2 :*** *Calcul des pondérations* λ_{kj} ;
  - avec l'approche locale nous avons une **matrice** de pondération ;
  - avec l'approche globale nous avons un **vecteur** de pondération (les pondérations sont indépendantes des classes).
- ***Etape 3 :*** *Construction des prototypes* $\boldsymbol{g}_k$.

5 Application au challenge ECOFFICES

Une interprétation des résultats de ces deux approches et une étude comparative seront présentées. Les premières analyses de cette étude sont disponibles sur le site :

http ://www-sop.inria.fr/axis/pages/ecoffices.html

Remerciements : Le projet ECOFFICES a été co-financé par la Région PACA et FEDER dans le cadre du programme régional PACALABS et a bénéficié de l'aide de la plateforme Focuslab (PACA CPER Telius) du living Lab ICT usage lab (Label ENoLL depuis 2006). Les auteurs remercient chaleureusement B. Senach et C. Goffart d'Inria (AxIS) pour les aspects expérimentation, les challengers du CSTB ainsi que les partenaires (Osmose, CSTB, CASA, Inria) du projet.

Références

Chavent, M. (2004). An hausdorff distance between hyper-rectangles for clustering interval data. In D. Banks, L. House, F. McMorris, P. Arabie, et W. Gaul (Eds.), *Classification, Clustering, and Data Mining applications*, pp. 333–339. Springer Verlag.

Chavent, M. et Y. Lechevallier (2002). Dynamical clustering of interval data. optimization of an adequacy criterion based on hausdorff distance. In K. Jajuga, A. Sokolowski, et H.-H. Bock (Eds.), *Classification, Clustering, and Data Analysis*, Berlin, pp. 53–60. Springer Verlag.

Summary

The aim of this paper is to show interest and disadvantages of both approaches to classifying curves. The first is based on the representation of interval, the second uses a distance based on the mathematical properties of curves (first derivative and second).

The curves are issued from temperature sensors placed in 40 offices during one year. This period was divided into the periods before and after challenge and the challenge period. During the challenge period the occupants had information by bonus / malus messages on energy consumption.