1d-SAX : une nouvelle représentation symbolique pour les séries temporelles

Simon Malinowski[*,***], Thomas Guyet[*,***], René Quiniou[**,***], Romain Tavenard[****]

[*]AGROCAMPUS-OUEST
[**]INRIA, Centre de Rennes - Bretagne Atlantique
[***]IRISA - UMR 6074 Campus de Beaulieu, F - 35 042 Rennes Cedex
[****]IDIAP Research Institute, Martigny, Switzerland
simon.malinowski@agrocampus-ouest.fr, thomas.guyet@agrocampus-ouest.fr,
rene.quiniou@inria.fr, romain.tavenard@idiap.ch

Résumé. SAX (Symbolic Aggregate approXimation) est une des techniques majeures de symbolisation des séries temporelles. La non prise en compte des tendances dans la symbolisation est une limitation bien connue de SAX. Cet article présente 1d-SAX, une méthode pour représenter une série temporelle par une séquence de symboles contenant des informations sur la moyenne et la tendance des fenêtres successives de la série segmentée. Nous comparons l'efficacité de 1d-SAX vs SAX dans une tâche de classification de séries temporelles d'images satellites. Les résultats montrent que 1d-SAX améliore les taux de classification pour une quantité d'information identique utilisée.

1 Introduction

La fouille de séries temporelles (FST) a récemment focalisé l'attention des chercheurs en fouille de données en raison de l'augmentation de la disponibilité de données comportant une dimension temporelle. Les algorithmes de FST tels que la classification/regroupement des séries temporelles, l'extraction de motifs ou la recherche de similarités nécessitent une mesure de distance entre séries temporelles. Le calcul de ces distances repose principalement sur la classique distance euclidienne ou l'alignement temporel dynamique (DTW - Dynamic Time Warping) qui peuvent conduire à des temps de calcul trop importants pour s'attaquer à de longues séries ou à des bases de séries volumineuses. Aussi, de nombreuses représentations approchées des séries temporelles ont émergé au cours de la dernière décennie. La représentation symbolique est une technique pour approximer les séries temporelles. L'algorithme SAX proposé par Lin et al. (2003) est un des plus utilisés pour la symbolisation. C'est une technique très simple, permettant de symboliser par segments les séries temporelles sans nécessiter d'information a priori. Lin et al. (2003) ont montré que SAX possède de bonnes performances pour la FST. Elle ne permet néanmoins pas de prendre en compte les informations de tendance dans les segments de séries temporelles. Plusieurs extensions de la représentation SAX ont été proposées pour pallier ce manque (Esmael et al. (2012); Lkhagva et al. (2006); Zalewski et al. (2012)). Cependant l'information sur la pente dans ces travaux est soit très simplifiée,

soit ne tient pas compte du fait que la distribution des valeurs de pente dépend de la taille des segments. De plus, ces méthodes augmentent significativement la taille de la représenation symbolique associée et les résultats n'en tiennent pas toujours compte.

Nous avons proposé dans Malinowski et al. (2013), une nouvelle représentation symbolique 1d-SAX pour les séries temporelles, basée sur la quantification de la régression linéaire des segments de la série. Il est montré que cette nouvelle représentation approche plus précisément les données d'origine que la représentation SAX pour une même quantité de symboles disponibles, et qu'elle permet d'améliorer les performances en termes de recherche efficace de plus proche voisins dans des bases de série. Dans cet article, nous rappelons les grands principes de 1d-SAX, et montrons dans la dernière section que cette technique peut-être utilisée avantageusement (en termes de compromis complexité/stockage/performance) pour de la classification de séries temporelles d'images satellites.

2 La représentation symbolique 1d-SAX

Dans cette section, nous détaillons notre nouvelle représentation symbolique des séries temporelles. Nous décrivons d'abord les principes de la méthode SAX, puis nous présentons l'extension proposée, appelée 1d-SAX.

2.1 SAX (Symbolic Aggregate Approximation)

SAX transforme une série temporelle numérique dans une séquence de symboles prenant leurs valeurs dans un alphabet fini. SAX ne nécessite pas d'information a priori sur les séries temporelles. Elle fait simplement l'hypothèse d'une distribution gaussienne des valeurs de la série. Sans perte de généralité, elle suppose que les séries temporelles sont centrées et réduites. La représentation SAX se calcule en trois étapes :

1. Découper la série en segments de longueur L,
2. Calculer la moyenne de la série temporelle sur chaque segment,
3. Quantifier les valeurs moyennes en un symbole choisi dans un alphabet de taille N.

L'hypothèse de gaussiannité conduit à quantifier les valeurs moyennes calculées sur chaque segment (étape 2) selon les quantiles de la loi gaussienne. Ces quantiles sont déterminés au moyen de points de coupure (β_k) disponibles dans des tables.

2.2 La représentation 1d-SAX pour les séries temporelles

L'inconvénient majeur de SAX est qu'elle repose uniquement sur les valeurs moyennes de chaque segment de la série. Ainsi, deux segments de valeur moyenne proche mais de comportement différent seront quantifiés selon la même valeur. Par exemple, un segment d'une série ayant une tendance croissante, peut être classé dans la même zone qu'un segment de tendance décroissante mais de valeur moyenne proche.
Nous proposons d'intégrer dans la représentation SAX une information sur la tendance de la série sur chaque segment. Cette nouvelle représentation, notée 1d-SAX dans la suite, est calculée en trois grandes étapes similaires à celles de SAX :

1. Découper la série en segments de longueur L,

2. Calculer la régression linéaire de la série temporelle sur chaque segment,
3. Quantifier les couples $(moyenne, pente)$ de la régression linéaire en un symbole choisi dans un alphabet de taille N.

Dans l'étape 2, l'algorithme calcule la régression linéaire de chaque segment produit dans l'étape 1, puis cette régression est quantifiée dans un alphabet fini (étape 3). La régression linéaire est calculée selon l'estimation des moindres carrés : soit $V_1, \ldots, V_L$, les valeurs d'un segment V sur les instants $T = [t_1, \ldots, t_L]$. La régression linéaire de V sur T est la fonction $l(x) = sx + b$ qui minimise la distance euclidienne entre l et V sur T. Elle est entièrement déterminée par les deux valeurs s et b. s représente la pente de l et b la valeur de l pour $x = 0$:

$$s = \frac{\sum_{i=1}^{L}(t_i - \overline{T})(V_i - \overline{V})}{\sum_{i=1}^{L}(t_i - \overline{T})^2}, \text{ et } b = \overline{V} - s \times \overline{T}, \tag{1}$$

où $\overline{T}$ et $\overline{V}$ représentent respectivement la moyenne des valeurs de V et de T.
Dans ce qui suit, nous avons choisi de représenter une régression linéaire de V sur les instants T par la valeur de la pente s, et la valeur moyenne du segment a. a est défini par l'équation $a = s \times (t_1 + t_L)/2 + b$. À l'issue de l'étape 2, la série temporelle est représentée par des couples (s, a) pour chaque segment résultant de la segmentation de la série. Il faut ensuite quantifier ces couples dans un alphabet de N symboles. À cette fin, les deux valeurs d'un couple sont quantifiées séparément puis combinées en un symbole. Avec la même hypothèse de gaussiannité de la série temporelle, les propriétés statistiques de la régression linéaire garantissent que les distributions des valeurs de moyenne et des valeurs de pente sont gaussiennes de moyenne 0, de variance 1 pour les valeurs de moyenne et de variance σ_L^2, une fonction décroissante de L, pour les valeurs de pente. Selon ces propriétés, la quantification de la moyenne et de la pente peut se faire de la même manière que pour SAX. Les valeurs des moyennes sont quantifiées sur N_a niveaux selon les $(N_a - 1)$ quantiles de la distribution gaussienne $\mathcal{N}(0, 1)$, tandis que les valeurs des pentes sont quantifiées sur N_s niveaux selon les $(N_s - 1)$ quantiles de la distribution gaussienne $\mathcal{N}\left(0, \sigma_L^2\right)$. Le choix du paramètre σ_L^2 est important. À partir de l'analyse de l'impact de σ_L sur de nombreuses séries temporelles de distribution gaussienne, $\sigma_L^2 = 0.03/L$ semble un bon compromis. Pour une même nombre de niveaux $N = N_a \times N_s$, la représentation 1d-SAX permet différentes configurations, suivant le nombre de niveaux affectés respectivement à la moyenne et à la pente. Par exemple, une représentation symbolique sur 64 niveaux peut être répartie en 32 pour la moyenne et 2 pour la pente ou à 16 pour la moyenne et 4 pour la pente, etc.

3 Interrogation asymétrique d'une base de séries

Nous avons appliqué cette nouvelle représentation des séries temporelles au problème de la recherche du plus proche voisin (1-PPV). L'objectif de cette application est le suivant. Étant donné une base de données D contenant $\#D$ séries temporelles et une série requête q, nous souhaitons trouver les séries temporelles de D les plus proches de q. Nous supposons dans la suite de cet article que la série requête q et toutes les séries temporelles de la base de données sont de même longueur. La méthode naïve consiste à calculer la distance entre la requête et toutes les séries de la base de données et à retourner la série la plus similaire à q. Le nombre

de distances à calculer est donc $\#D$. Nous pouvons tirer parti de la représentation symbolique afin d'accélérer la recherche dans de grandes bases de données de séries temporelles. La représentation SAX a par exemple été utilisée pour indexer et interroger des bases contenant des téraoctets de séries temporelles (Shieh et Keogh (2008)).
Nous définissons dans cette section un schéma d'interrogation asymétrique pour la recherche approchée de séries temporelles dans une base de données. Le terme asymétrique signifie que les requêtes ne sont pas quantifiées pour éviter d'avoir une double erreur de quantification (quand les requêtes et les séries de la base de données sont quantifiées). Jégou et al. (2011) ont montré que l'interrogation asymétrique améliore la précision de l'approximation pour la recherche de vecteurs. Nous proposons une méthode basée sur cette même idée pour la recherche de séries temporelles dans une base de données.
Nous supposons que D contient les séries temporelles, ainsi que leur représentation symbolique (1d-SAX) pour un ensemble de paramètres donnés (L, N_a, N_s). Cet ensemble de paramètres définit complètement les $N = N_a \times N_s$ symboles $s_1, \ldots, s_N$ utilisés pour quantifier la série. L'algorithme de recherche du 1-PPV de la requête q est le suivant :

1. Découper q en segments de longueur L : $q = q_1, \ldots, q_w$
2. Calculer les distances euclidiennes entre chaque segment de q et les symboles $s_j, 1 \leq j \leq N$. Ces valeurs sont stockées dans une table $A = (a_{i,j})$ de dimension $w \times N$, où $a_{i,j} = ED(q_i, s_j)^2$. ED représente la distance euclidienne.
3. Pour toute série d de D, la version quantifiée de d, $\hat{d} = \hat{d_1}, \ldots, \hat{d_w}$ est disponible. La distance approchée $Dist_{asym}$ entre q et $\hat{d}$ donnée par

$$Dist_{asym}(q, \hat{d}) = \sum_{i=1}^{w} ED(q_i, \hat{d_j})^2 = \sum_{i=1}^{w} a_{i, s_{\hat{d}_j}}, \tag{2}$$

   obtenue par simple accès à la table A et sommation sur w éléments.

Une fois ces étapes effectuées, les distances entre q et toutes les séries temporelles de D sont disponibles pour identifier les plus proches voisins de q. Le nombre ν_q d'opérations arithmétiques élémentaires nécessaires pour satisfaire une requête est :$\nu_q = (3L - 1) \times w \times N + (w - 1) \times \#D$, où la partie gauche est le coût de l'étape 2 ci-dessus et la partie droite celui de l'étape 3. Dans le cas de la méthode naïve le nombre d'opérations élémentaires est $(3Lw - 1) \times \#D$. Le coût du calcul dans le schéma de recherche approchée est plus faible, en particulier pour de grandes bases de données où $N \ll \#D$.

4 Expérimentations et évaluations

Dans cette section, nous évaluons les performances de notre représentation symbolique des séries temporelles dans une tâche de classification de séries temporelles d'images satellites. Les images ont été acquises sur une période de 6 ans à raison d'une image tous les 16 jours et recouvrent le Sénégal avec une résolution spatiale de 250m. Chaque pixel d'une image contient une valeur normalisée qui quantifie la végétation au sol (indice NDVI). La succession dans le temps des indices NDVI d'un pixel constitue une série temporelle représentant l'évolution de cet indice sur toute la durée de l'acquisition. On cherche à caractériser les types de végétation au sol par l'évolution de l'indice NDVI. Les différents types d'occupation du sol sont, par

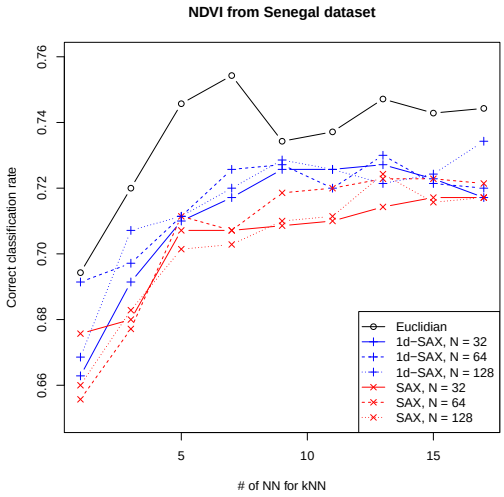

FIG. 1 – *Taux de classification correcte en fonction du nombre de plus proches voisins gardés. Comparaison entre les méthodes se basant sur SAX, 1d-SAX et sur la distance euclidienne.*

exemple : forêt, arbuste, zone urbaine, fleuve, ... Les types d'occupation du sol de chaque pixel ont été déterminés par photointeprétation [1] et donnent les classes d'appartenance des pixels. Nous avons utilisé ces données dans un schéma de classification par plus proches voisins en appliquant une technique de validation croisée. Chaque série de l'ensemble d'apprentissage est représentée par sa version symbolique (par SAX et 1d-SAX pour comparaison). Ensuite, pour chaque série de l'ensemble de test, la classe est estimée en regardant la classe majoritaire des k plus proches voisins dans l'ensemble d'apprentissage (méthode de la section 3). Un taux de classification correcte est calculé sur tout l'ensemble de test.

La figure 1 compare les taux de classifications correctes obtenus en gardant les séries temporelles telles quelles (et en utilisant la distance euclidienne pour les comparer), et en les symbolisant par les techniques SAX et 1d-SAX. L'axe des abscisses représente le paramètre k de la classification par k plus proches voisins. On peut constater que les taux de classification correctes obtenus par la méthode 1d-SAX sont globalement supérieurs à ceux obtenus avec SAX. De plus, ils sont assez proches de ceux obtenus avec la distance euclidienne (méthode beaucoup plus coûteuse en temps et en espace de stockage des images).

5 Conclusion

Dans cet article, nous avons proposé une nouvelle représentation symbolique des séries temporelles. Cette représentation est basée sur la quantification de la régression linéaire des segments constituant la série. Les symboles prennent en compte des informations sur la moyenne et la pente des segments de la série temporelle. Dans toutes nos expérimentations, nous

1. Données issues du programme *Global Land Cover Network* de la FAO (`http://www.glcn.org/`).

avons utilisé la même quantité d'information pour représenter une série temporelle par SAX ou 1d-SAX – *i.e.* un même nombre de symboles N – là où les autres méthodes utilisent une quantité d'information supérieure à celle dévolue à SAX. Nous avons utilisé cette représentation pour la classification de séries temporelles d'images satellites et montré qu'on obtient de meilleures performances en terme de taux de classification qu'en utilisant SAX. Notre méthode nécessite le réglage d'un seul paramètre supplémentaire : le rapport entre le nombre de niveaux pour les valeurs des moyennes (N_a) et celui des valeurs de pente (N_s). Dans un proche avenir, nous étudierons l'apprentissage automatique de la configuration optimale (N_a vs N_s).

Références

Esmael, B., A. Arnaout, R. K. Fruhwirth, et G. Thonhauser (2012). Multivariate time series classification by combining trend-based and value-based approximations. In *Proc. of the 12th Int. Conf. on Computational Science and Its Applications (ICCSA)*, Volume 7336 of *Lecture Notes in Computer Science*, pp. 392–403.

Jégou, H., M. Douze, et C. Schmid (2011). Product quantization for nearest neighbor search. IEEE *Transactions on Pattern Analysis and Machine Intelligence 33*(1), 117–128.

Lin, J., E. Keogh, S. Lonardi, et B. Chiu (2003). A symbolic representation of time series, with implications for streaming algorithms. In *Proc. of the 8th ACM SIGMOD workshop on Research issues in Data Mining and Knowledge Discovery*, pp. 2–11.

Lkhagva, B., Y. Suzuki, et K. Kawagoe (2006). New time series data representation esax for financial applications. In *Proc. of the 22nd Int. Conf. on Data Engineering Workshops*, pp. 17–22.

Malinowski, S., T. Guyet, R. Quiniou, et R. Tavenard (2013). 1d-SAX : a novel symbolic representation for time series. In *Proc. of The Twelfth International Symposium on Intelligent Data Analysis (IDA)*, Volume 8207 of *Lecture Notes in Computer Science*, pp. 273–284.

Shieh, J. et E. Keogh (2008). iSAX : Indexing and mining terabyte sized time series. In *Proc. of the ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pp. 623–631.

Zalewski, W., F. Silva, H. D. Lee, A. G. Maletzke, et F. C. Wu (2012). Time series discretization based on the approximation of the local slope information. In *Proc. of the 13th Ibero-American Conference on AI (IBERAMIA)*, Volume 7637 of *Lecture Notes in Computer Science*, pp. 91–100.

Summary

SAX (Symbolic Aggregate approXimation) is one of the main symbolization techniques for time series. A well-known limitation of SAX is that trends are not taken into account in the symbolization. This paper proposes 1d-SAX a method to represent a time series as a sequence of symbols that each contain information about the average and the trend of the series on a segment. We compare the efficiency of SAX and 1d-SAX in a satellite image time series classification scheme. Results show that 1d-SAX improves performance using equal quantity of information.