Classification de séquences basée sur une distance universelle algorithmique

François CAYRE[1], Olivier J.J. MICHEL[1], Nicolas LE BIHAN[1]

[1]Univ. Grenoble Alpes, CNRS, Grenoble INP[†],
GIPSA-Lab, 38000 Grenoble, France.
[†]: Institute of Engineering Univ. Grenoble Alpes
francois.cayre@gipsa-lab.fr,olivier.michel@gipsa-lab.fr, nicolas.le-bihan@gipsa-lab.fr

Résumé – La classification de séquences pour lesquelles un modèle probabiliste n'existe pas nécessite de recourir à la théorie de l'information algorithmique. Classiquement, les mesures d'information utilisées pour construire une distance se basent sur des algorithmes de compression standards (de type gzip). Dans cet article, nous introduisons une nouvelle approche basée sur différents types de conditionnement lors de la factorisation conditionnelle de séquences. Cette approche, appelée SALZA, permet de dériver une semi-distance universelle entre séquences, que nous étendons ici en une distance en vue d'une application de classification.

Abstract – The classification of strings for which a probabilistic model does not exist requires the use of algorithmic information theory. Classically, the information measurements used to construct a distance are based on standard compression algorithms (*e.g.*, gzip). In this article, we introduce a new approach based on different types of conditioning for conditional factorization of sequences. This approach, called SALZA, allowsto derive a universalsemi-distance on sequences, that is here further extended to make a full distance for classification tasks.

La théorie de l'information algorithmique définit un cadre pour l'étude des données pour lesquelles il n'y a pas de modèle probabiliste à disposition [1]. Cette théorie est basée sur l'utilisation de la complexité de Kolmogorov au lieu de l'entropie de Shannon pour dériver une mesure d'information. Parmi ses applications, la plus connue est la définition de la "Normalized Compression Distance" (NCD) [2] qui utilise des compresseurs sans pertes pour estimer une distance entre des séquences de caractères, la complexité de Kolmogorov n'étant pas calculable.

Dans cet article, nous présentons une implémentation de l'information algortihmique mutuelle conditionnelle pour des séquences et illustrons son application dans un problème de classification.

1 Conditionnement et factorisation

Un point crucial en théorie de l'information est la définition d'une information mutuelle conditionelle respectant la règle de chaînage et permettant la définition d'un théorème du traitement de données. Dans le cas des séquences, une telle mesure d'information conduit naturellement à la définition d'une *semi-distance*. La mesure de similarité conditionnelle (devenant une mesure d'information dans un cas particulier) sur laquelle repose la définition de cette *semi-distance* est appelée SALZA (Sequence analysis based on Lempel-Ziv like Algorithms).

Nous considérons des séquences de taille finie et à valeurs sur un alphabet fini $\mathcal{A}$, pour lesquelles nous utilisons l'ordre lexicographique usuel. La séquence nulle (vide) est dénotée $\emptyset$. $\mathcal{A}^+$ est l'ensemble des séquences non-nulles de tailles finies, et on dénote $\mathcal{A}^\star = \mathcal{A}^+ \cup \emptyset$. La taille d'une séquence et le cardinal d'un ensemble ou d'un alphabet sont dénotés $|.|$. Étant données des séquences $x_1, \ldots, x_n$, $x_{\leq k}$ désigne les k premières d'entre elles et $x_{\leq 0} = \emptyset$.

1.1 Types de conditionnement

SALZA analyse une séquence de manière séquentielle, à la manière de LZ77 : le but est de décrire une séquence y à l'aide de son propre passé (potentiellement) et d'un ensemble de séquences x_i ($i = 1, \ldots, n$) contribuant à former l'information *a priori*.

L'approche dans SALZA consiste donc à rechercher la plus longue référence dans l'information à disposition. Un point clé est d'autoriser différentes stratégies pour définir cette information à disposition. Étant données les séquences $y, x_1, \ldots, x_n$ et une stratégie choisie, on y associe $\mathcal{R}$ (un sous-ensemble de l'ensemble $y, x_1, \ldots, x_n$) et SALZA factorise y en cherchant les mots les plus longs dans le sous-ensemble $\mathcal{R}$ correspondant. Ici, nous considérons deux stratégies relatives à la position du pointeur dans la séquence (en rouge sur la figure 1) lors de la factorisation de y :

1. $y|x_1, \ldots, x_n : \mathcal{R}$ est constitué du passé de y (la partie de y déjà factorisée) et de la totalité des chaînes $x_1, \ldots, x_n$;
2. $y|^+x_1, \ldots, x_n : \mathcal{R}$ consiste seulement en la totalité des chaînes $x_1, \ldots, x_n$.

La première est une factorisation de type LZ77 [3]+Ziv-Merhav [6] de y alors que la seconde est une factorisation de type Ziv-Merhav uniquement. Une représentation des deux stratégies de factorisation est donnée sur la figure 1. Dans la suite, les résultats sont présentés avec la notation $y \wr x_1, \ldots, x_n$ qui représente indifféremment l'une des deux stratégies de conditionnement. Les résultats présentés sont donc systématiquement valides pour les deux cas.

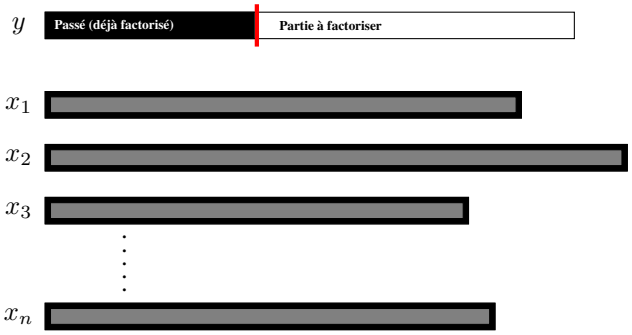

FIGURE 1 – Illustration de l'information à disposition ($\mathcal{R}$) pour les deux stratégies L77+Ziv-Merhav (noir) et Ziv-Merhav uniquement (gris). La position du pointeur d'encodage est donnée par la ligne rouge.

1.2 Factorisation de séquences

La factorisation SALZA est séquentielle. Elle découpe une séquence y, connaissant les séquences $x_1, \ldots, x_n$, en m symboles de la forme $(s_i, l_i, z_i)_{1 \leq i \leq m}$:

$$y \wr x_1, \ldots, x_n = (s_1, l_1, z_1) \ldots (s_m, l_m, z_m).$$

Dans un *symbole* SALZA, s est un pointeur vers une des séquences préalablement rencontrées, l est une longueur et z est un entier positif. Un symbole SALZA (s, l, z) est donc possiblement :

1. un *littéral* : $s = y$, $l = 1$ et z est le littéral de y à insérer dans le tampon de sortie ;
2. une *référence* : $l > 1$ est la taille du plus long *mot* dans $\mathcal{R}$ identique à la nouvelle sous-séquence (ou *mot*) observée. Bien que non utilisés ici, s serait la séquence dans laquelle le nouveau mot a été trouvé et z serait le décallage par rapport à l'origine de s auquel le copier-coller devrait commencer.

Le produit de deux factorisations (avec éventuellement des séquences différentes d'information *a priori*) est définie comme la concaténation des symboles SALZA :

$$y_1 \wr x_{1,1}, \ldots, x_{1,n_1} \times y_2 \wr x_{2,1}, \ldots, x_{2,n_2} = (s_{1,1}, l_{1,1}, z_{1,1}) \ldots (s_{1,m_1}, l_{1,m_1}, z_{1,m_1}) \, (s_{2,1}, l_{2,1}, z_{2,1}) \ldots (s_{2,m_2}, l_{2,m_2}, z_{2,m_2}).$$

On définit alors la *factorisation jointe* SALZA d'un ensemble de séquences $x_1, \ldots, x_n \in \mathcal{A}^\star$. La factorisation jointe SALZA est alors simplement :

$$x_1 \ldots \cdot x_n = \prod_{i=1}^{n} x_i \wr x_{\leq i-1}.$$

Suivant le conditionnement, on a $x_1 = x_1|\emptyset$ la factorisation LZ77 de x_1, et $x_1|^+\emptyset$ consiste en la succession des littéraux qui forment x_1. On remarque également que la factorisation jointe est asymétrique, *i.e.* $x \cdot y \neq y \cdot x$, comme c'est classiquement le cas dans l'analyse de séquences [4].

2 Mesure de similarité SALZA

Il est possible de baser une théorie algorithmique de l'information à l'aide de SALZA. Nous introduisons ici la notion de similarité et quelques exemples utiles pour la tâche de classification de la Section 3.

2.1 Similarité conditionnelle SALZA

Un paramètre central dans l'approche SALZA est l'ensemble des longueurs d'une factorisation :

$$\mathcal{L}_{y \wr x_1, \ldots, x_n} = \{l_i\}_{1 \leq i \leq m}.$$

Afin de tenir compte de ce paramètre, nous introduisons une *fonction admissible* et un *niveau de bruit*. Ainsi, pour une séquence x, $f : \mathbb{N}^\star \to [0, 1]$ est appelée *fonction admissible* ssi :

1. f est monotone croissante,
2. $\exists\, 0 < T < |x| \,, \forall\, l \geq T \,,\ f(l) = 1.$

Dans cette définition, le paramètre T agit comme un seuil au-dessus duquel les *références* sont considérées comme d'importance équivalente. Le *niveau de bruit*, noté $l^\star$, est défini comme la valeur centrale de f, *i.e.* $f(l^\star) = 1/2$. Plus de détails sur les fonctions admissibles sont donnés dans [5].

Définition 2.1. *Étant données une fonction admissible f et les séquences $y, x_1, \ldots, x_n \in \mathcal{A}^*$, la similarité conditionnelle SALZA de y sachant $x_1, \ldots, x_n$, notée $S_f(y \wr x_1 \ldots, x_n)$, est définie comme :*

$$S_f(y \wr x_1, \ldots, x_n) = |y| - \sum_{l \in \mathcal{L}_{y \wr x_1, \ldots, x_n}} (l - 1) f(l). \qquad (1)$$

La similarité conditionnelle SALZA est bornée, *i.e.* : $1 \leq S_f(y \wr x_1, \ldots, x_n) \leq |y|$. Elle atteint son minimum quand y est une sous-chaîne d'une des x_i, et correspond aux notions de complexités *classiques* quand $f = \mathbb{1}$.

2.2 Théorie de l'information algorithmique basée sur SALZA

À l'aide de la mesure de similarité conditionnelle SALZA, il est possible de proposer une implémentation complète de la

théorie de l'information algorithmique. Afin de proposer par la suite des outils de classification (Section 3), nous introduisons la notion de similarité jointe SALZA. Étant données une fonction admissible f et les séquences $x_1, \ldots, x_n \in \mathcal{A}^*$, la similarité jointe SALZA est de la forme :

$$S_f(x_1 \ldots \cdot x_n) = \sum_{i=1}^{n} S_f(x_i \wr x_{\leq i-1}),$$

en prenant soin à l'ordre dans $x_1, \ldots, x_n$. Cette similarité devient une complexité de type LZ77 dans le cas $S_{\mathbb{1}}(x_1) = |x_1|$. À present, étant données une fonction admissible f et les séquences $x, y, z \in \mathcal{A}^*$, la *similarité mutuelle conditionnelle* [1] SALZA de x et y sachant z, notée $I_f(x : y \wr z)$, est définie comme suit :

$$I_f(x : y \wr z) = S_f(z \cdot x) + S_f(z \cdot y) - S_f(z \cdot x \cdot y) - S_f(z),$$

Dans le cas où $I_f(x : y \wr z) = 0$, x et y sont dits *dissimilaires* connaissant z. Dans le cas particulier $f = \mathbb{1}$, on montre que $S_{\mathbb{1}}$ sastifait les conditions [4, Section 7.1] et est donc bien une mesure d'information. Il est à noter que pour des séquences $y, x_1, \ldots, x_n \in \mathcal{A}^*$, la complexité relative SALZA de y connaissant $x_1, \ldots, x_n$, est définie comme $S_{\mathbb{1}}(y \wr x_1, \ldots, x_n)$. Une conséquence directe est que : Si $I_{\mathbb{1}}(x : y \wr z) = 0$, alors x et y sont dits indépendants connaissant z. Enfin, la complexité relative SALZA est également non-croissante par conditionnement, tel que pour trois séquences $x, y, z \in \mathcal{A}^\star$, il vient que : $S_{\mathbb{1}}(y \wr x, z) \leq S_{\mathbb{1}}(y \wr x)$. Les propriétés de *mesure d'information* de la similarité SALZA peuvent se résumer ainsi :

1. Normalisation : $S_f(\emptyset) = S_f^+(\emptyset) = 0$;
2. Monotonicité : $x \leq y \implies S_f(x) \leq S_f(y)$;
3. Sous-modularité approchée :
   $S_{\mathbb{1}}(z \cdot x) + S_{\mathbb{1}}(z \cdot y) \geq S_{\mathbb{1}}(z \cdot x \cdot y) + S_{\mathbb{1}}(z)$.

À titre d'exemple, nous proposons de montrer l'application de SALZA à un problème de classification.

3 Classification

Dans de nombreuses applications, les données disponibles ne comportent souvent qu'une seule réalisation, n'autorisant pas une modélisation statistique pertinente. Nous proposons ici un algorithme de classification dans ce cas de figure pour des séquences à valeurs sur un alphabet fini. Nous dérivons une distance à partir de la semi-distance basée sur SALZA et montrons sa pertinence pour classifier des jeux de séquences.

3.1 NSD : une semi-distance avec SALZA

Il est possible de définir une semi-distance universelle entre séquences gâce à SALZA en se basant sur le codeur relatif de type Ziv-Merhav [6]. Cela correspond pour nous au cas du conditionnement $|^+$. Étant données une fonction admissible f, et deux séquences $x, y \in \mathcal{A}^+$ telles que $|x|, |y| > 1$, la semi-distance SALZA normalisée, notée NSD_f, est définie comme :

$$\mathrm{NSD}_f(x, y) = \max\left\{ \frac{S_f(x|^+y) - 1}{|x|}, \frac{S_f(y|^+x) - 1}{|y|} \right\}.$$

Cette semi-distance a été utilisée en classification [5], et comparée à la NCD (Normalized Compression Distance) [2] pour de la classification de séquences. La NSD s'est d'ailleurs montrée plus rapide et plus discriminante que la NCD [5]. Ici, nous nous focalisons sur le problème qui est que la NSD n'est qu'une semi-distance : nous proposons une procédure permettant d'en déduire une distance et nous l'appliquons à un problème de classification. Dans la suite, f est une fonction admissible utilisant l'exponentielle et $l^\star$ est la moyenne arithmétique des longueurs des symboles de l'ensemble des factorisations. Ainsi, nous filtrons automatiquement les références courtes, réputées être moins pertinentes.

3.2 Distance universelle algorithmique

Soit une matrice SALZA, notée $\mathbf{SD}$, dont le terme général est $[\mathbf{SD}]_{ij} = \mathrm{NSD}_f(y_i, y_j)$. L'ensemble $\{y_i, i = 1, \ldots, N\}$ représente un ensemble de séquences (génétiques sur l'exemple présenté dans la suite) à segmenter (au sens 'clustering'). $\mathbf{SD}$ est considérée comme une matrice de poids de connections d'un graphe complet $\mathcal{G}(\mathcal{V}, \mathcal{SD})$ [2]. Les sommets $v_i \in \mathcal{V}$ sont associés aux y_i. Un graphe de recouvrement minimal (Minimum Spanning Tree) unique et calculable en $O(N \log(N))$ opérations peut être construit, *e.g.* par l'algorithme de Prim. Dans [7], il est proposé d'"enraciner" deux MSTs en deux sommets différents, et à faire croître les MSTs ainsi enracinés de manière compétitive (un seul sommet est connecté à l'un des deux arbres à chaque itération de l'algorithme de Prim, et doit être de poids minimal), jusqu'à ce que les deux arbres se rencontrent. La réunion des deux arbres ainsi construits est un MST pour l'ensemble des sommets qu'ils connectent et la longueur totale (somme des poids associés aux segments) est un estimateur de l'entropie de la distribution des sommets connectés : les propriétés de minimalité des MSTs permettent d'établir que l'algorithme de construction à partir de ces deux racines tend à connecter entre eux les sommets dans les voisinages où ils sont densément distribués. Notons $\mathrm{NSD}_f(y_k^{(i)}, y_l^{(j)})$ le poids du dernier segment construit, joignant les $v_k^{(i)}$ et $v_l^{(i)}$ connectés respectivement aux arbres enracinés en v_i et v_j. Définissons $\delta(y_i, y_j) = \delta(v_i, v_j) = \mathrm{NSD}_f(y_k^{(i)}, y_l^{(j)})$. **La mesure $\delta(.,.)$ est positive, symétrique, et satisfait l'inégalité triangulaire**. Les deux premières propriétés sont triviales. La troisième est établie dans [7]. Une propriété importante des mesures induites par des MSTs (par des graphes quasi-additifs en général) est leur relation dans la limite d'un grand nombre de sommets, avec la notion de distance géodésique sur la variété (généralement inconnue) sur laquelle évoluent les sommets du graphe [8].

1. La notation $I_f(.)$ est utilisée car dans le cas où $f = \mathbb{1}$ on retrouve la notion standard d'information mutuelle algorithmique [1].

2. $\mathcal{SD}$ est l'ensemble des liens de poids décrits dans $\mathbf{SD}$

3.3 Résultats de classification

Nous présentons un exemple de "clustering" réalisé sur un ensemble de 14 séquences génétiques $\{y_i, i = 1, \ldots, 14\}$, pour lesquelles les mesures $\mathrm{NSD}_f(y_i, y_j)$ ont été réalisées. Les matrices $\mathbf{SD}$ et $\mathbf{\Delta}$ [3] sont représentées en Figure 2. Les indices des séquences (ou des sommets associés) ont été ré-ordonnés *a posteriori* pour faire apparaître les clusters sur ces matrices. Indépendamment de la propriété de distance satisfaite par $\mathbf{\Delta}$, un contraste supérieur est observé : cette propriété se comprend aisément en notant que la distance entre deux séquences est toujours la longueur du plus grand segment (poids maximum) dans un parcours de longueur minimale reliant les sommets associés à ces séquences [7]. Sur la Figure 3 (en haut) est représenté le dendrogramme associé à la matrice $\mathbf{\Delta}$. Il apparaît que ce classifieur hiérarchique est satisfaisant au sens ou les groupes définis semblent logiques au sens des classes biologiques qu'il découvre. Les valeurs prises par le dendrogramme montrent qu'il reste cependant difficile de segmenter automatiquement les classes, voire d'en définir le nombre. Un algorithme de clustering spectral permet un changement de représentation des points abstraits associés aux sommets du graphe, par projection dans un sous-espace propre principal du Laplacien du graphe. Le dendrogramme de la Figure 3 (en bas) représente le classifieur hiérarchique obtenu pour la matrice de distances Euclidiennes des sommets projetés sur l'espace propre de Laplacien normalisé ([9]) du graphe de similarité $\mathbf{W}_{i,j} = exp(-\frac{\alpha.\mathbf{\Delta}_{i,j}}{2\sigma^2})$. α est un paramètre ajustable (ici α = nb de clusters cherchés = 4). σ^2 est la variance empirique de la distribution des valeurs $[\mathbf{\Delta}]_{i,j}$. Le dendrogramme obtenu fait apparaitre des clusters très concentrés et bien séparés, facilement identifiables par un simple algorithme de "K-Means" (Lloyd-Max).

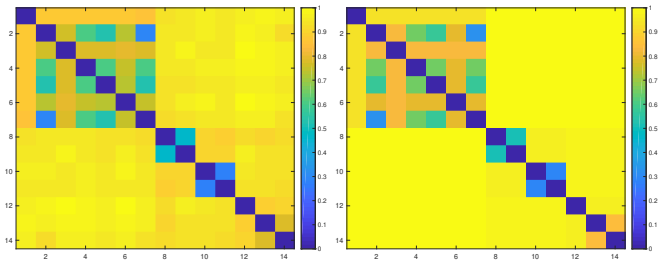

FIGURE 2 – Matrices de distances $\mathbf{SD}$ et $\mathbf{\Delta}$.

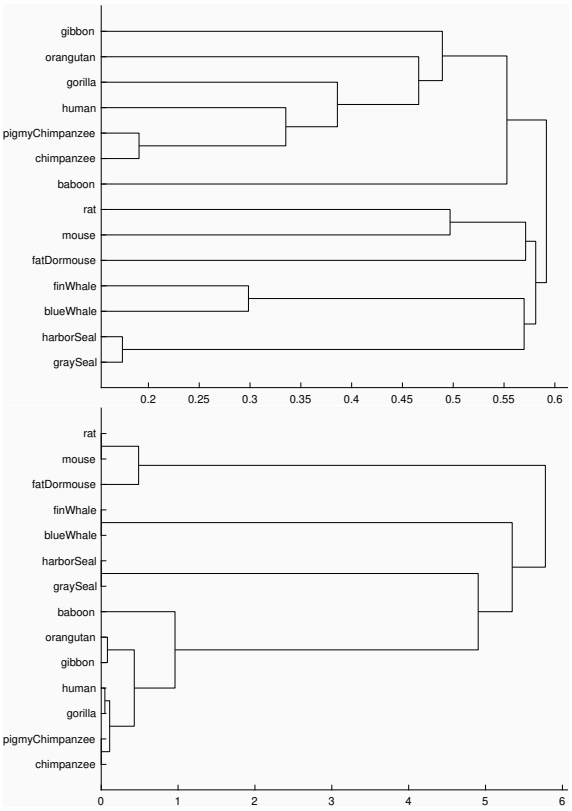

FIGURE 3 – Dendrogrammes pour $\mathbf{\Delta}$ (haut) et pour l'analyse par clustering spectral (bas).

Références

[1] Ming Li and Paul M.B. Vitániy. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer-Verlag New-York, 2008.

[2] Rudi Cilibrasi and Paul M.B. Vitányi. Clustering by Compression. *IEEE Transactions on Information Theory*, 51 :1523–1545, Apr. 2005.

[3] Jacob Ziv and Abraham Lempel. A Universal Algorithm for Sequential Data Compression. *IEEE Transactions on Information Theory*, 23 :337–343, May 1977.

[4] Bastian Steudel, Dominik Janzing, and Bernhard Schölkopf. Causal Markov Condition for Submodular Information Measures. In *Proceedings of the 23rd Annual Conference on Learning Theory*, USA, 2010.

[5] Marion Revolle, François Cayre, and Nicolas Le Bihan. Clustering and causality inference using algorithmic complexity. EUSIPCO, Kos, Greece, August 2017.

[6] Jacob Ziv and Neri Merhav. A Measure of Relative Entropy Between Individual Sequences with Application to Universal Classification. *IEEE Transactions on Information Theory*, 39 :1270–1279, Jul. 1993.

[7] L. Galluccio, O.Michel, P. Comon, M. Kliger, A.O.Hero. Clustering with a New Distance Measure based on a Dual-Rooted Tree. *Information Sciences*, 251 : 96-113, 2013.

[8] J. Costa and A. O. Hero. Manifold Learning with Geodesic Minimal Spanning Trees. ArXiv manuscript cs.CV/0307038 (arXiv), July 2003.

[9] J. Shi and J. Malik Normalized Cuts and Image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8) : 888-905, 2000.

3. $[\mathbf{\Delta}]_{i,j} = \delta(v_i, v_j)$