Influence des Vecteurs Caractéristiques en Stéganalyse par Séparateurs à Vastes Marges

Benoit Roue[1],Patrick Bas[2], Jean-Marc Chassery[1]

[1]Laboratoire des Images et des Signaux INPG/ENSIEG, Grenoble, France

[2]Laboratory of Computer Science, Helsinky University of Technology, Finland

First.Name@lis.inpg.fr

Résumé – Le but de l'analyse stéganographique est de prouver la présence d'une information cachée dans un signal hôte. Cette étude se focalise sur une analyse stéganographique aveugle des images numériques utilisant des Séparateurs à Vastes Marges (SVM).
Dans un premier temps nous décrivons cette méthode de stéganalyse puis nous étudions plusieurs prédicteurs utilisés dans la phase d'extraction de caractéristiques. Enfin nous dégageons des propriétés statistiques et texturelles qui permetteront d'améliorer les performances des classificateurs à vaste marge.

Abstract – The goal of steganalysis is to search for the presence of hidden information in numerical contents.This paper focuses on blind steganalysis using Support Vectors Machine (SVM).
First we describe this algoritm, then we study several predictors used during the caracteristics extraction. Finally some attributes are outlined in order to improve steganalysis accuracy.

1 Introduction

La stéganographie consiste en l'insertion d'une information dans un support hôte sans que celle-ci ne puisse être décelable analytiquement ou visuellement. L'intérêt porté à cette discipline s'est accru depuis que l'on soupçonne son utilisation dans les documents multimédias [2] [5] [4] (images numériques marquées, fichiers audio...).
La stéganalyse (ou analyse stéganographique) a pour objectif de détecter l'éventuelle présence d'un message inséré et d'en estimer, si possible, la taille. Plusieurs familles de techniques d'analyse stéganographique ont été développées dans la littérature, mais nous nous focalisons ici sur des méthodes aveugles (*i.e.* détection de la présence d'un message sans connaître l'algorithme d'insertion) basées sur l'apprentissage par Séparateurs à Vastes Marges (SVM).

2 Méthode de stéganalyse par SVM

L'analyse stéganographique par SVM permet de séparer, à l'aide d'un classifieur, les images "propres" (images non marquées) des stego-images (images contenant un message). La construction du classifieur se fait en deux étapes :

- une phase d'extraction des caractéristiques qui permettront de discriminer les deux classes d'images
- une phase d'entraînement du classifieur par Séparateur à Vaste Marge à l'aide de ces caractéristiques.

2.1 Extraction des caractéristiques

Les caractéristiques extraites de l'image doivent faire ressortir au mieux les propriétés de la marque. Lyu *et al* [6] proposent une méthode basée sur les hautes fréquences de l'image. Dans ce travail cette information est extraite à partir d'une décomposition multirésolution de l'image (*e.g.* décomposition en ondelettes).
Cette décomposition est effectuée via des filtres miroirs en quadrature qui décomposent l'image en sous-bandes d'orientations et de fréquences différentes (Figure 1) : une sous-bande horizontale H, verticale V, diagonale D et basse fréquence L. En itérant le processus sur la sous-bande basse fréquence on obtient une décomposition multirésolution de l'image (Figure 2). Les différentes sous-bandes à l'échelle $i = 1 \dots n$ sont notées H_i, V_i et D_i.
Une première série de caractéristiques est alors extraite de chaque sous-bande aux échelles 1 à $n-1$: les moments centraux normalisés d'ordre 1 à 4 (*i.e.* la *moyenne* μ, la *variance* σ, l'*asymétrie* ζ et le *kurtosis* κ). Pour une sous-bande B cela donne :

$$\begin{aligned} \mu &= \frac{1}{N_x N_y} \sum_{x,y} B(x,y) \\ \sigma &= \frac{1}{N_x N_y} \sum_{x,y} (B(x,y) - \mu)2 \\ \zeta &= \frac{1}{N_x N_y \sigma 3} \sum_{x,y} (B(x,y) - \mu)3 \\ \kappa &= \frac{1}{N_x N_y \sigma 4} \sum_{x,y} (B(x,y) - \mu)4 - 3. \end{aligned} \quad (1)$$

où la somme s'effectue sur tous les coefficients de la sous-bande B de N_x pixels de largeur et de N_y pixels de hauteur. Cette opération permet de construire un vecteur caractéristique de $4 \times 3 \times (n-1)$ éléments.
Ensuite des caractéristiques tenant compte des propriétés intrinsèques des images sont calculées. En effet dans une

image naturelle les pixels ne varient pas aléatoirement, il est donc alors possible de prédire la valeur d'un pixel grâce à celle de ses voisins (dans le domaine spatial ou multi-résolution).
Cependant, lorsqu'une image a été stéganographiée la corrélation locale est perturbée. L'erreur de prédiction est donc discriminante et peut être utilisée afin de construire le vecteur caractéristique.
Pour prédire les valeurs $H_i(x,y)$, $V_i(x,y)$ et $D_i(x,y)$ $(i = 1 \dots n)$, plusieurs prédicteurs peuvent être utilisés, et la section 3 les étudiera plus en détail.
En combinant les moments issus des sous-bandes elles-mêmes et les moments provenant de l'erreur de prédiction les auteurs obtiennent donc enfin un vecteur caractéristique de $24 \times (n-1)$ composantes qui décrit les statistiques de l'image analysée. C'est ce vecteur qui sera utilisé dans l'étape suivante : la classification par apprentissage.

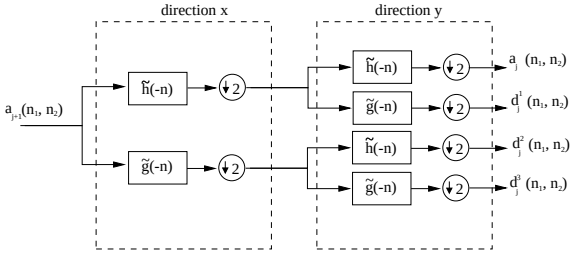

FIG. 1 – Principe de décomposition multirésolution.

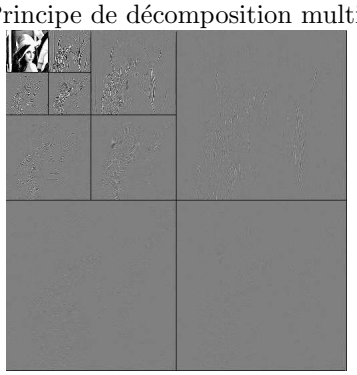
FIG. 2 – L'image Léna et sa décomposition sur 3 échelles.

2.2 Classification par apprentissage

L'étape de classification est faite par apprentissage supervisé et permet de séparer en deux classes (images propres et stégo-images) toutes les images analysées (ou plus formellement leurs vecteurs caractéristiques). Ce dernier point est le coeur de l'analyse stéganographique aveugle.
L'apprentissage supervisé requiert un entraînement du classifieur qui sera construit à l'aide d'échantillons classés au préalable. Dans la littérature, il existe plusieurs méthodes d'apprentissage supervisé (réseaux de neurones, discriminants linéaires de FISCHER, SVM linéaires ou non-linéaires...) mais Farid et Lyu [6] proposent l'utilisation de Séparateurs à Vaste Marge non linéaires qui s'avèrent être, dans le cadre de l'analyse stéganographique, les plus efficaces.
Dans le cas d'une SVM linéaire nous allons chercher un hyperplan de $\mathbb{R}^d$ (de dimension $n-1$) qui séparent les échantillons propres des échantillons marqués (l'hyperplan optimal est obtenu par maximisation de lagrangien), avec, si les échantillons ne sont pas tous séparables, une fonction de coût de mauvaise classification (Figure 3).
En analyse stéganographique il est difficile de séparer les échantillons avec un hyperplan linéaire (Figure 4), il faut donc utiliser une SVM non-linéaire. Les données des vecteurs caractéristiques sont projetées dans un espace euclidien $\mathbb{H}$ de dimension supérieure à d (éventuellement infinie) à l'aide d'une application injective ϕ et on applique une SVM linéaire dans cet espace. Définir ϕ serait coûteux en tant de calcul ou même théoriquement impossible, cependant, comme la résolution de l'hyperplan optimal revient à une maximisation de lagrangien, on peut résoudre le problème sans connaître ϕ. Il suffit de trouver une fonction $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ telle que $K(\overrightarrow{x}, \overrightarrow{y}) = \phi(\overrightarrow{x})\phi(\overrightarrow{y})$. De telles fonctions sont appelées noyaux de MERCER et le noyau utilisé dans cette méthode est le *noyau gaussien* :

$$K(\overrightarrow{x}, \overrightarrow{y}) = exp(-\frac{||\overrightarrow{x} - \overrightarrow{y}||^2}{2\sigma^2}). \qquad (2)$$

où σ est la variance de la gaussienne, valeur à déterminer lors de l'apprentissage.

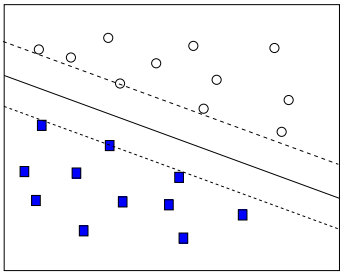
FIG. 3 – SVM linéaire en dimension 2.

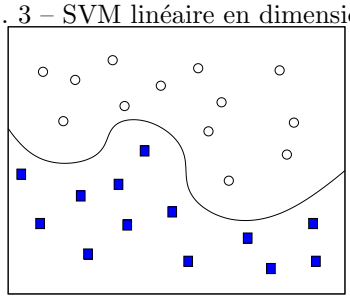
FIG. 4 – SVM non-linéaire en dimension 2

3 Etape de prédiction

Le choix des caractéristiques (Section 2.1) qui composent les vecteurs d'entrée du système d'apprentissage est une étape cruciale qui déterminera la qualité du classifieur. Le prédicteur utilisé lors de l'acquisition des statistiques doit donc être le plus efficace possible.

3.1 Prédicteur linéaire

Pour prédire les valeurs $H_i(x,y)$, $V_i(x,y)$ et $D_i(x,y)$ $(i = 1 \dots n)$, FARID *et al* [3] proposent un prédicteur linéaire [1] :

[1] * désigne la valeur prédite

$$
\begin{aligned}
H_i^*(x,y) = \; & w_1 H_i(x-1,y) + w_2 H_i(x+1,y) \\
& + w_3 H_i(x,y-1) + w_4 H_i(x,y+1) \\
& + w_5 H_{i+1}(x/2,y/2) + w_6 D_i(x,y) \\
& + w_7 D_{i+1}(x/2,y/2) \\
V_i^*(x,y) = \; & w_1 V_i(x-1,y) + w_2 V_i(x+1,y) \\
& + w_3 V_i(x,y-1) + w_4 V_i(x,y+1) \\
& + w_5 V_{i+1}(x/2,y/2) + w_6 D_i(x,y) \\
& + w_7 D_{i+1}(x/2,y/2) \\
D_i^*(x,y) = \; & w_1 D_i(x-1,y) + w_2 D_i(x+1,y) \\
& + w_3 D_i(x,y-1) + w_4 D_i(x,y+1) \\
& + w_5 D_{i+1}(x/2,y/2) + w_6 H_i(x,y) \\
& + w_7 V_{i+1}(x/2,y/2)
\end{aligned}
\tag{3}
$$

Les w_i sont les paramètres du prédicteur qu'il faut ajuster pour minimiser l'erreur de prédiction. On montre que l'erreur sur chaque sous bande est donnée, pour le prédicteur optimal, par[2] :

$$\overrightarrow{E} = \mathtt{log}_2(\overrightarrow{B}) - \mathtt{log}_2(|Q\overrightarrow{w}|) \tag{4}$$

où $\overrightarrow{B}$ contient les coefficients des sous-bandes, $\overrightarrow{w}$ contient les paramètres du prédicteur, et Q les amplitudes des coefficients voisins. A partir de $\overrightarrow{E}$ sont calculées sur chaque sous- bande les mêmes statistiques que dans le paragraphe 2.1 (*moyenne* μ, *variance* σ, *asymétrie* ζ et *kurtosis* κ). Dans l'optique d'améliorer la détection, nous proposons d'utiliser d'autres prédicteurs.

3.2 Prédicteur DCT

L'idée est d'utiliser un prédicteur plus performant utilisé dans la norme MPEG4, qui travaille sur les coefficients DCT de l'image. Il s'agit de prédire les coefficients DCT qui ont une forte probabilité d'être modifiés par un schéma de stéganograhie : les coefficients supérieurs gauche de chaque bloc DCT (*e.g.* les 10 premiers coefficients lors d'une lecture en *zig-zag*). Ensuite les mêmes moments centraux (Section 2.1) sont extraits de l'image reconstruite et ces statistiques s'ajoutent à celles déduites de l'image originale pour former un nouveau vecteur caractéristique à utiliser lors de l'entraînement du classifieur.

3.3 Prédicteur basé B-splines

Ici on utilise des B-splines pour prédire les coefficients des sous bandes de la décomposition multirésolution c'est à dire que les coefficients à prédire sont interpolés grâce aux B-splines construites par les coefficient voisins (ligne ou colonne). Ensuite les statistiques sont extraites comme dans la méthode de FARID (paragraphe 3.1) et s'ajoutent aux caractéristiques de la section 2.1 pour former un autre vecteur caractéristique.

4 Résultats

4.1 Mise en œuvre

Les méthodes décrites dans cette étude ont été testées sur plusieurs algorithmes de stéganographie existant dans la littérature :

- Outguess, développé par N.Provos en 1999. Cet algorithme modifie les bits de poids faibles de certains coefficients DCT, puis modifie les coefficients DCT afin d'obtenir un histogramme identique à celui de l'image originale.
- Jphide, développé par A.Latham en 1998. Cet algorithme modifie les bits de poids faibles de certains coefficients DCT choisis aléatoirement.
- F5, développé par A.Westfeld et A.Pfitzmann. Comme les deux autres cet algorithme travaille sur les coefficients DCT mais décrémente leur valeur absolue et utilise une matrice de codage afin de minimiser le nombre de coefficients à modifier.

Ensuite le classifieur a été entraîné sur une base d'images en niveaux de gris (2700 images) contenant le même nombre d'images marquées et non marquées. Les images ont toutes la même taille (512×512) et les messages insérés via les algorithmes décrits sont de taille 48×48. Pendant cette phase, les paramètres σ et C (respectivement la largeur de gaussienne et la fonction de coût pour mauvaise classification) sont déterminés par validation croisée.
Enfin une même quantité d'images marquées par les algorithmes ci-dessus et d'images propres sont utilisées pour tester les performance du classifieur SVM obtenu.

4.2 Performances

Les résultats obtenus sur des bases de 1500 images par algoritme d'insertion sont présentés sur les Figures 5 (pour le taux de bonne classifications des images propres) et 6 (pour le taux de bonne classification des images contenant un message).
La méthode de Farid a une très bonne précision en ce qui concerne les vrais positifs (*images marquées et reconnues comme telles*) et nos résultats avec le prédicteur linéaire (Figure 6) sont équivalents (seulement équivalents car la qualité d'un classifieur dépend de la base d'image utilisée lors de la phase d'entraînement, on ne peut donc raisonnablement s'attendre à avoir exactement les mêmes résultats). Les résultats obtenus avec les prédicteurs DCT et B-spline en ce qui concerne les vrais positifs sont bons aussi, voire très bons pour Outguess car les coefficient DCT sont modifiés pour restaurer l'histogramme original. On peut donc imaginer un classifieur SVM ayant des caractéristiques extraites à l'aide de ce prédicteur dans le cadre d'une stéganalyse non aveugle (où on connait l'algorithme d'insertion).
Cependant nos erreurs de classification, comme celles de Farid (Figure 5), sont très importantes pour les taux de faux positifs (30% des images non marquées sont mal classées) et surtout pour des images texturées (hautes fréquences, beaucoup de contours). Dans ce cas précis le SVM va considérer les images texturées comme marquées, en effet les erreurs de prédiction, avec tous les prédicteurs, sont importantes dans ce type d'image. Pour ces images il est donc préférable de ne pas utiliser de prédicteur.

[2] le `log` est calculé pixel par pixel (ou coefficient par coefficient).

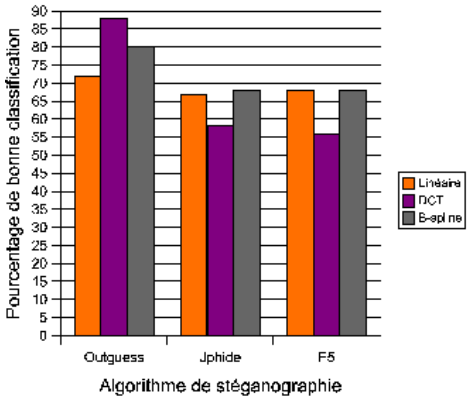

FIG. 5 – Taux de bonne classification des images non marquées.

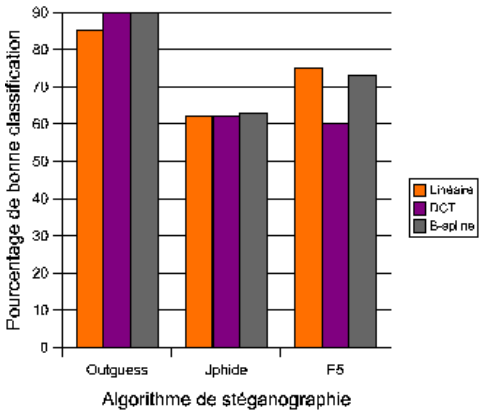

FIG. 6 – Taux de bonne classification des images marquées.

4.3 Analyse

Les performances obtenus avec les trois classifieurs sont bonnes dans l'ensemble, mais nous pouvons d'ors et déjà tirer des enseignements à partir des erreurs de prédictions obtenues. Deux causes principales des ces imprécisions se dégagent de notre analyse :

- Les erreurs sur les images texturées nous poussent à développer des solutions pour réduire l'erreur de classification par l'utilisation d'outils d'analyse de texture comme les matrices de cooccurrenceen s'inspirant de la stéganalyse LSB [1].
- Nous avons, pour entrainer les classifieurs, des vecteurs de dimension 72. c'est beaucoup trop pour une base de 2700 images. Effectivement en très grande dimension il faut un nombre quasi-infini d'échantillons pour définir un hyperplan. Des méthodes de sélection des de variable peuvent nous permettre de réduire significativement ce nombre de variables.

5 Conclusion et Perspectives

L'extraction des caractéristiques est une étape primordiale dans l'entraînement du classifieur. Nous avons voulu améliorer les performances du classifieur en tenant compte, via de nouveaux prédicteurs, des propriétés des images analysées et des algorithmes d'insertion.

Nous pouvons utiliser d'autres prédicteur optimaux (Filtre de Kalman...) et il sera également important de réduire l'erreur de classification relative aux images texturées par l'utilisation d'outils de texture comme les matrices de cooccurrence, comme nous l'avons fait pour la stéganalyse LSB [1].

De plus, plutôt que d'obtenir une base d'images très grande afin d'obtenir un hyperplan séparateur plus fiable, nous pouvons réduire le nombre de coefficients des vecteurs caractéristiques, en effet ces coefficients tels qu'ils sont décrits dans cette étude sont extrémement corrélés. Une étude en cours vise à obtenir des vecteurs caractéristiques de dimension 8 en utilisant les mêmes procédés d'extractions de caractéristiques.

Références

[1] B.Roue, P.Bas, and J-M.Chassery. Improving lsb steganalysis using marginal and joint probabilistic distributions. In *Multimedia and Security Workshop*, Magdeburg, 2004.

[2] H. Farid. Detecting hidden messages using higher-order statistical models. In *International Conference on Image Processing*, NY, 2002.

[3] H.Farid and S.Lyu. Higher-order wavelet statistics and their applications to digital forensic. In *IEEE Workshop on Statistical Analysis in Computer Vision*, Wisconsin, 2003.

[4] J.Fridrich, M.Goljan, and R.Du. Detecting LSB Steganogeraphy in color and grayscale images. In *Magazine of IEEE Multimedia Special Issue on Security*, pages 22–28, October 2001.

[5] S.Dumitrescu, X.Wu, and Z.Wang. Detection of LSB steganography via sample pair analysis. In *IEEE transactions on Signal Processing*, pages 1995–2007, 2003.

[6] S.Lyu and H.Farid. Detecting hidden message using higher-order statistics and support vector machine. In 5^{th} *International Workshop on Information Hiding*, Pays-Bas, 2002.