SONY CSL-Paris

Technical Report CSLP-TR-06-01

Organisation et catégorisation de la musique populaire par apprentissage statistique

Roman Campana

September 2006

Romain CAMPANA
Rapport de stage dans le cadre de la validation du Master2 Mathématiques, Vision et Apprentissage de l'ENS Cachan

Organisation et catégorisation de la musique populaire par apprentissage statistique

Stage réalisé au Sony CSL sous la direction de François PACHET

Introduction

L'ensemble de la musique digitalisée répandue à travers le Monde représente, même s'il est difficile de l'évaluer exactement, une masse colossale d'information. Cette information est le sujet d'étude de la communauté MIR (Music Information Retrieval), à laquelle j'ai tenté d'apporter ma contribution le temps de mon stage au Music Sony CSL. La difficulté dans ce domaine est de réussir l'intégration des méthodes automatiques issues de l'informatique et du caractère subjectif inhérent à la musique.

Avant d'aller plus loin, je ne résiste pas à l'envie d'évoquer des réflexions d'ordre esthétique qui nécessairement surgissent lorsque l'on tente de faire cohabiter science et musique. Ces réflexions je préfère les esquisser en amont afin qu'elles n'interfèrent pas avec les sections purement techniques. Lorsque l'on s'intéresse aux problématiques de la communauté MIR, il faut bien souvent abandonner ses propres jugements musicaux, car dans ce qui suit une improvisation de John Coltrane, un quatuor de Beethoven ou le dernier tube de Madonna seront mis sur un pied d'égalité et seront tous appréhendés

comme des fichiers mp3 accompagnés d'un ensemble d'adjectifs attribués par des auditeurs aux techniques d'écoute codifiées.

Le travail que j'ai effectué au Sony CSL s'inscrit dans la volonté d'appliquer les techniques d'apprentissage, supervisée ou non, à la musique digitalisée. Le problème majeur de l'apprentissage statistique est la constitution de bases d'entraînement suffisamment représentatives de l'univers étudié, dans le but d'obtenir des modèles robustes. A cela s'ajoute le fait que pour la musique, les tables de vérités obtenues par le jugement humain peuvent très aisément revêtir un caractère subjectif. Or Sony CSL dispose de bases de données fiables qui permettent d'envisager un travail véritablement innovant. Ces données évoquées à l'instant sont purement confidentielles, et à ce titre je ne donnerai que les informations nécessaires à une bonne compréhension scientifique. J'attire donc l'attention sur le caractère exclusif des données dont dispose le laboratoire CSL et sur la quasi absence de travaux scientifiques similaires à ceux menés par le Sony CSL.

Données disponibles et perspectives de recherche

Sony CSL dispose d'une importante base de titres mp3, chacun des titres est décrit par un milliers d'attributs booléens, par exemple « Style Rock » = true. Ces attributs sont groupés en 17 catégories : Aera/Epoch, Affiliate, Character, Country, Dynamics, Genre, Language, Main Instrument, Metric, Mood, Musical Setup, Rythmics, Style, Tempo, Text Category, Variant.

Dans un premier temps j'ai opté pour une approche naïve de ces données, en appliquant des méthodes d'apprentissage non supervisées. Le but , ambitieux s'il en est, était de dégager une structure d'ensemble de la musique populaire du vingtième siècle. Peut-on mettre en évidence des clusters de titres que la sensibilité humaine n'aurait pas imaginé ? Dans un premier temps j'ai étudié les résultats d'une analyse en composantes principales, avant de partir à la recherche de procédés plus originaux reliés à la théorie de la compression. J'ai notamment implémenté un algorithme basé sur la Minimum Description Length.

Un autre visage du stage a été un retour sur la récente thèse de J.J Aucouturier, ancien thésard du laboratoire Sony CSL. L'objet de sa thèse a été la construction d'une distance timbrale. Je suis parti de son travail pour construire un noyau timbral que j'ai ensuite utilisé dans la suite de mon travail.

Enfin, pour l'essentiel de mon stage, je me suis intéressé à l'inférence d'attributs par des méthodes de classification supervisées (les Support Vector Machines uniquement). Peut-on prédire un attribut connaissant les autres ? Peut-on inférer un attribut à partir de l'analyse timbrale ? Dans quelle mesure la connaissance des attributs booléens, que nous appellerons aussi tags, permet d'améliorer la classification par le timbre seule ?

Table des matières

I. Mise en évidence de profils de titres (approche non supervisée)

Les fichiers mp3 dont je dispose pour ce stage ont été taggés à la main, c'est-à-dire parmi une liste d'un millier d'attributs (ou tags), ceux qui correspondent ont été cochés. Ce processus est long, et fastidieux sous certains aspects. Si ce processus était assisté (ou entièrement réalisé) par ordinateur, on augmenterait plus rapidement la taille des données disponibles. C'est dans ce souci de vitesse que je me suis intéressé à dégager des structures au sein de la base.

1. Analyse en composantes principales

On a recours à l'analyse en composantes principales lorsque l'on désire réduire la dimensionnalité de nos données d'entrée. Elle est par exemple utile en reconnaissance de digits lorsque ceux-ci sont représentés par une matrice binaire. On procède ainsi : à partir d'un ensemble de digits, on détermine les directions de variations qui sont elles-mêmes des images. L'information est en général contenue dans les premières directions de variations, ainsi un digit de la base d'entraînement peut s'approximer de manière

satisfaisante comme la combinaison linéaire de quelques directions seulement. Dans le cas de digit on peut ainsi réduire la dimensionnalité d'un facteur 10 de manières à par la suite appliquer des algorithmes d'apprentissage de manière plus viable.

Les titres de notre base d'étude étant originellement décrits par approximativement un millier d'attributs booléens, j'ai naturellement été tenté d'expérimenter l'ACP afin de faciliter les tâches d'apprentissage ultérieures. Je propose un petit rafraîchissement théorique avant d'aller plus loin.

a. Rappels théoriques

Un titre est ramené à un vecteur x à N dimensions (N voisin de 1000) dans la base triviale (en continuant l'analogie avec l'image, chaque pixel est ici un attribut booléen). Le but de l'ACP est de construire une nouvelle base, appropriée à la description des metadata(les attributs ici) des fichiers mp3 disponibles. Ceci est réalisé en étudiant à partir d'un ensemble de 9000 titres les directions de grandes variations des données $(x_i)_{i=1..9000}$ .Les N directions fourniront une nouvelle base. L'intérêt de l'ACP vient du fait que l'information est concentrée sur un petit nombre de vecteurs de cette base constituée des directions principales.

La projection orthogonale sur une direction w appartenant à $\Re^N$ est la fonction $h_w : \Re^N \to \Re$ définie par :

$$h_w(x) = x^T \frac{w}{\|w\|}$$

La variance empirique de h_w est :

$$\hat{\text{var}}(h_w) = \frac{1}{n}\sum_{i=1}^{n} h_w(x_i)^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{(x_i{}^T w)^2}{\|w\|^2}$$ où n est le nombre de données

La i-ème direction principale (i=1..N) est définie par :

$$w_i = \underset{w \perp \{w_1,..,w_{i-1}\}}{\arg\max}\ \hat{\text{var}}(h_w)$$

Soit X la matrice 9000xN contenant les chiffres de la base d'entraînement sous la forme de vecteurs ligne, on a alors :

$$\hat{\text{var}}(h_w) = \frac{1}{n}\frac{w^T X^T X w}{w^T w}$$

$$w_i = \underset{w \perp \{w_1,\ldots,w_{i-1}\}}{\arg\max} \frac{1}{n}\frac{w^T X^T X w}{w^T w}$$

Les solutions de cette dernière équation sont les vecteurs propres de $X^T X$, rangés par valeurs décroissantes.

Une fois que l'on dispose des directions principales, ainsi que des valeurs propres associées qui rendent compte des poids informatifs de ces directions, on peut se restreindre à une description selon les *nb* directions prépondérantes. Ainsi tout titre T peut être approximé de la manière suivante :

$$T = \sum_{i=1}^{nb} \left(x^T \frac{w_i}{\|w_i\|} \right) w_i$$

L'ensemble des $c_i = x^T \frac{w_i}{\|w_i\|}$ pour i=1..nb serait alors la nouvelle manière de décrire les metadata des titres de notre base.

b. Résultats pratiques

Les calculs précédents ont été réalisés avec Matlab. Lorsque l'on étudie des images binaires, les directions principales sont elles en niveau de gris. Cela n'est guère gênant pour la reconnaissance de caractère puisque la couleur n'intervient pas. En revanche cela est plus problématique pour nos attributs booléens. Une direction ne peut en effet pas contenir l'attribut Style Rock à 0.8 et l'attribut Style Reggae à -0.05, il faut imposer un seuil au sein des directions principales et ramener les attributs qui le dépasse à la valeur 1. Je me rends bien compte que mes expériences prennent ici un tour artisanal et qu'il faudra songer à des méthodes spécialement dédiées aux attributs booléens. Néanmoins voici le résultat de ma version de l'ACP. Le seuil adopté pour décider de la présence ou non d'un attribut dans une direction principale a été fixé à la moitié de la plus grande valeur (en valeur absolue) prise au sein de cette direction.
Les résultats qui vont suivre, outre leur valeur instructive, permettent à mes lecteurs de se faire une idée des attributs booléens utilisés pour décrire les titres de la base.

Selon ces règles la première direction principale, dont la valeur propre associée normalisée est de 4.89%, est constituée des attributs suivants:

- Dynamics steady
- Character well-balanced
- Situation Car Ride
- Aera/Epoch 1990-2000
- Genre Mainstream
- Mood strong
- Character powerfull
- Character pulsating
- Popularity Popularity low
- Mood pulsating
- Situation Party/Dance
- Musical Setup Vocals / Accompany
- Main Instruments male
- Main Instruments Vocals
- Language English
- Metric 4/4
- Genre Dancemusic

La deuxième direction principale dotée d'un poids de 2.17% est ainsi constituée :

- Mood dreamy
- Rhythmics solid
- Main Instruments Guitar (acoustic)
- Situation Night
- Variant synthetic
- Rhythmics groovy
- Situation Landscape/Panning Shot
- Style Rock/Pop
- Character calm
- Mood harmonious
- Style Hip Hop
- Mood marching
- Main Instruments Beatbox / electr. Drums
- Style Urban
- Style Rap
- Genre Mainstream
- Main Instruments Vocals (Spoken; Rap)
- Style Pop
- Situation Evening Mood
- Character pulsating
- Mood sentimental
- Genre Adult Contemporary (AC)
- Variant soft
- Character melodical
- Mood pulsating
- Situation City Life
- Musical Setup Synthetic
- Main Instruments Vocals
- Mood strong
- Situation Love
- Genre Ballad
- Character powerfull
- Character well-balanced

La première direction s'interprète comme les attributs apparaissant les plus souvent dans notre base, ce qui est également vrai pour la deuxième avec des attributs moins fréquents. L'interprétation des autres directions est plus hasardeuse.

c. Conclusion

Une expérience discréditant la notion même de direction (si on considère un titre sous la forme d'un vecteur de valeurs binaires) est la suivante : (expérience vraie pour tout n) on ne prend pas en considération les n attributs les plus importants de la première direction. On relance une analyse en composante principale : la première direction principale est constituée des mêmes attributs que dans nos premières expériences moins les n attributs retirés.

Cette première tentative malheureuse dans le sens d'une représentation concise des éléments de la base, m'a encouragé à prendre impérativement en compte le caractère binaire.

2. Clustering utilisant la Minimum Description Length

J'ai eu l'intuition de considérer l'ensemble des réalisations des attributs booléens comme un code. J'ai ensuite trouvé un papier de Plumbley [5] qui détaille une méthode basée sur la MDL pour le Clustering de données binaires creuses. J'en ai réalisé une implémentation Matlab et l'ai testée sur la base du Sony CSL.
La méthode à laquelle je me suis intéressée est utile pour des études de marketing ou pour la recherche d'information dans des textes. Typiquement dans le cadre d'une étude de marché pour laquelle on s'intéresse aux pratiques d'achat d'un consommateur, on dispose pour chaque individu d'une liste d'attributs booléens correspondant aux objets achetés ou non. Si la liste d'attributs est l'ensemble des produits d'une grande surface la représentation peut être très creuse. Le but recherché est l'extraction de profils de clients type afin de savoir qui consomme quoi.
Dans l'optique d'une application au texte, il s'agit à partir d'une base de documents écrits d'origines diverses, de réaliser un Clustering par contenu lexical, chaque mot étant ici un attribut booléen. Un texte de botanique aura un lexique spécialisé très différent de celui d'un rapport d'apprentissage statistique à moins que ce dernier ne comporte les charmants corymbes, pneumatophore ou autre spadice…

Je vais présenter l'algorithme de Clustering dans le cas qui nous concerne, à savoir l'extraction de profils musicaux. J'espère ainsi faire des regroupements statistiques entre attributs auxquels les préjugés, humains, seraient restés imperméables.

a. Formalisme

Le kième titre de la base est décrit par un vecteur $(x_{1k}, x_{2k}, ..., x_{mk}) = x_k$, les coordonnées étant 0 ou 1 en fonction de la présence ou non de ce tag.
Ainsi la matrice X est la matrice m x p, nombre d'attributs x nombre de titres considérés. X est maintenant considéré comme une information à transmettre. Il s'agit de trouver un encodage de X qui assure le code le plus court. Cette approche est connue comme l'approche MDL.
Supposons que X puisse être approximé par $\hat{X} = A \circ S$, où A est une matrice m x n et S une matrice n x p, toute deux matrice binaire.

$$\hat{x}_{ik} = \vee_{j=1}^{n} a_{ij} s_{jk} \quad \text{expression qui vaut 1 si } \sum_j a_{ij} s_{jk} \neq 0 \text{ et 0 sinon}$$

A représente n clusters caractérisés par leurs attributs et S l'appartenance ou non des titres à chacun des clusters.
Au lieu d'encoder les éléments de X seuls, nous encodons maintenant les trois matrices A, S et $\left(X \middle| \hat{X}\right)$. On a ainsi la longueur totale du code :

$$L = L(A, S, X) = L(A) + L(S) + L\left(X \middle| \hat{X}\right)$$

A partir de là on recherche le couple (A,S) qui minimise l'encodage de la longueur totale. Le pseudo code qui suit, appuyée par une démonstration dans [5], permet de déterminer le Clustering optimal au sens de la MDL.

b. Algorithme implémenté

L'algorithme qui suit est un algorithme de Clustering descendant. A chaque étape descendante une nouvelle matrice A est calculée ainsi :

$$A = X \circ S^T$$

$$\text{Soit explicitement } a_{ij} = 1 \text{ si } \exists k, (x_{ik} = 1) \wedge (s_{jk} = 1) \text{ et O sinon}$$

Ce calcul (je ne présente pas la démonstration) correspond au choix de A permettant d'optimiser la MDL.
Dans l'algorithme décrit, un titre ne peut appartenir qu'à un seul cluster.

1. Initialisation :
Il y a un cluster pour un attribut, la matrice S est diagonale de taille p

2. Fusion (opération logique OU) des 2 clusters les plus proches au sens de la mesure

$$D(1,2) = n_3|a_3| - \{n_1|a_1| + n_2|a_2|\}$$

L'indice 3 est celui de la fusion potentielle, les n représente le nombre de titres appartenant à un cluster donné et les |a| les cardinaux des clusters en terme d'attributs. On obtient ainsi une nouvelle matrice S (je rappelle qu'elle contient l'information d'appartenance des titres aux clusters).

3. Partant de la nouvelle matrice S, on calcule la nouvelle matrice A et ainsi la nouvelle approximation de X.

4. On répète 2 et 3 jusqu'à la fin de la récurrence (on a envisagé tous les Clustering possibles : d'un cluster par titre jusqu'à un cluster pour tous les titres).

c. Exploitation de l'algorithme

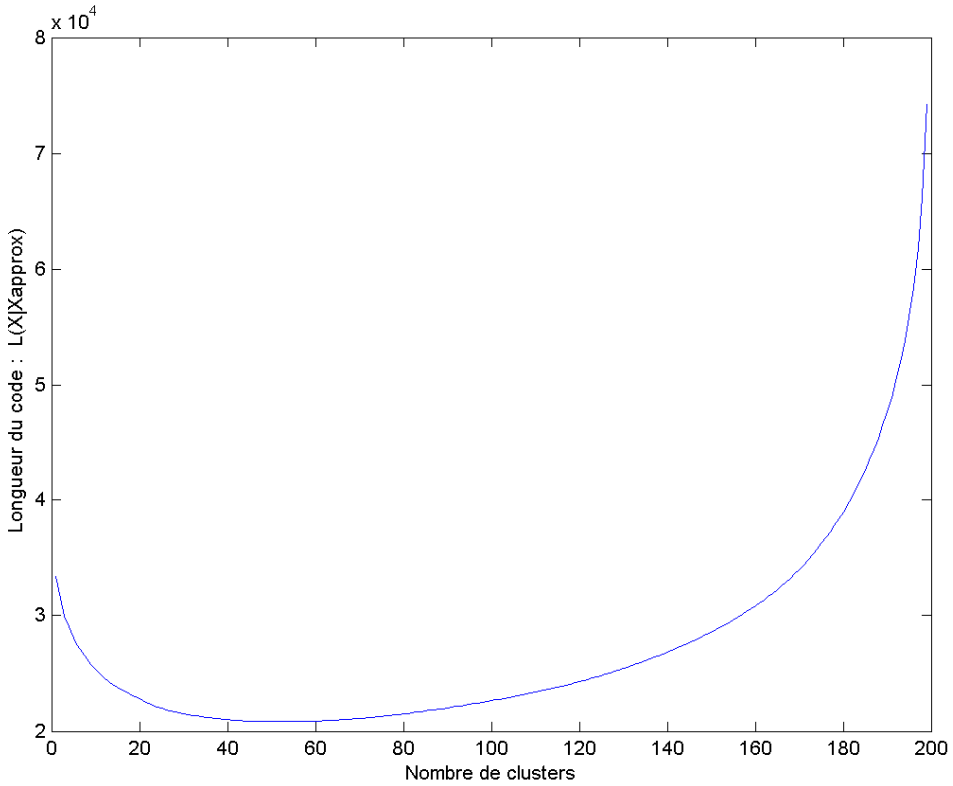
A chaque étape, on calcule et on garde en mémoire la longueur de code résultant de l'approximation de X. Ceci utilise la formule :

$$L\left(X \middle| \hat{X}\right) = -N\left(\hat{X}\right)\log(q_{01}) - N(X)\log(q_{11})$$

$N(X)$ et $N\left(\hat{X}\right)$ représentent respectivement le nombre de 1 dans X et $\hat{X}$, et $q_{11} = \dfrac{N(X)}{N\left(\hat{X}\right)}$ avec $q_{01} = 1 - q_{01}$.

d. Expérimentations sur des titres musicaux

Je considère 200 titres pris au hasard dans la base, représentés à l'aide de 760 attributs : X est donc une matrice 760 x 200. Le taux de remplissage de X est de 4.7 %.
Je « lance » l'algorithme ci-dessus sur ces données, j'obtiens 53 pour le nombre optimal de clusters. Ce chiffre se vérifie sur le graphe de L(X|Approximation de X) en fonction du nombre de clusters.



La longueur de code minimale pour X sachant son approximation est 2.10^E4, chiffre à comparer avec les 15.10^E5 (pour 760 x 200) éléments de X. On met en évidence la redondance de notre description des titres musicaux.
Néanmoins, les résultats que nous fournit cet algorithme ne sont pas très probants du point de vue du jugement humain. Je parle du résultat supposément optimal. En effet, j'ai remarqué qu'en général le n optimal est de l'ordre du tiers du nombre de titres en entrée. Dans le cas de 200 titres en entraînement nous avons 53 clusters constitués en moyenne de 56 attributs, soit une valeur près de 2 fois supérieure au nombre d'attributs en général nécessaire pour décrire un titre. Cet accroc en quelques sortes est lié à la présence redondante de certains attributs, exactement comme le mot « le » ou « et » le sont pour le Clustering de textes.

e. Singularités

A l'intérieur de chacun des clusters obtenus par l'algorithme basé sur la MDL, j'extrais les attributs idiosyncrasiques.
Parmi les 53 clusters, 20 ne présentent aucun attribut susceptible de les caractériser. En moyenne, les clusters ont 1.16 attribut spécifique. Voici quelques exemples de clusters présentant de « nombreuses » singularités :

Language Vocalize Mood comic Situation Slapstick Style Elektro Variant pointed Variant staccato	Main Instruments Recorder Main Instruments Tin Whistle Mood solemn/festive Musical Setup Orchestra (Studio-)	Main Instruments Woodwinds Mood heavenly Situation Church Situation Silent Movie	Genre Live Classic Genre Nightclub Music Genre Singer/Songwriter Musical Setup Duo	Metric 12/8 Situation Spring Style Folk Style Folklore Style Folkpop

S'ils n'ont rien d'extraordinaires, ces résultats permettent de vérifier le fonctionnement du processus de fusion. On peut véritablement extraire les singularités, mais cette étape d'extraction est standard et ne fait pas intervenir le critère d'optimalité de l'encodage.

f. Conclusion

La recherche du nombre optimal de clusters n'a pas véritablement abouti sur notre ensemble de titres. Ceci peut s'expliquer par le fait que l'algorithme utilisé est plus efficace pour des données plus creuses (soit une matrice X avec un taux de remplissage inférieur à 1%).
D'autre part la complexité cubique de l'algorithme est assez dissuasive pour une application sur une grande base de données.

II. Le Timbre

Jean Julien Aucouturier, ancien thésard au Sony CSL, a présenté en Avril 2006 une thèse [6] sur la modélisation du timbre polyphonique. Je vais en rappeler les éléments les plus importants puisqu'au cours de mon stage j'ai utilisé bons nombres de ses résultats, stockés dans les bases de données du laboratoire. Sans ce travail préalable et sans l'environnement objet développé sur plusieurs années, je n'aurais pu réaliser mes expériences d'apprentissage statistique.

1. Définition psycho acoustique

Le timbre est défini par l'ASA (American Standards Association) comme l'attribut sensitif qui permet à un auditeur de juger de la dissimilarité de deux sons de même hauteur et de même intensité.

2. Vers une définition du timbre pour le traitement du signal

Le manque de modèles psycho acoustiques pour la perception du timbre de textures polyphoniques pousse les chercheurs à utiliser une approche pragmatique. Je précise au passage que la mention de musique polyphonique est à prendre au sens perceptif et non musicologique du terme ! L'approche pragmatique, donc, est en fait celle utilisée en reconnaissance de formes : on extrait des descripteurs informatifs à partir desquels on pourra réaliser une tâche de classification. Pour le timbre, on étend la technique la plus simple de reconnaissance dans le cas monophonique. Le signal audio est découpé en frames avec un certain recouvrement, typiquement des frames de 50ms avec un recouvrement de 50%. Pour chaque frame on extrait un vecteur de description constitué des MFCCs : Mel Frequency Cepstrum Coefficients.
Aucouturier met cependant l'accent sur un point : son but est de modéliser le timbre et non d'extraire des descripteurs haut niveau afin de réaliser des classifieurs. Il veut définir une métrique pour comparer les titres entre eux (j'utilise le mot titre comme équivalent de l'anglais « song »).
Pour chaque frame on conserve un certain nombre de coefficients du cepstre (les MFCCs).
Le cepstre est la transformée de Fourier inverse du logarithme du spectre [1], [4] :

$$c_n = \frac{1}{2\pi} \int_{\omega=-\pi}^{\omega=+\pi} \log S(\omega) \exp(j\omega n) d\omega$$

On appelle mel-cepstre, le cepstre après avoir opéré une transformation non linéaire des fréquences vers une échelle perceptive qui reproduit la non-linéarité de la résolution fréquentielle de l'oreille humaine.
Les MFCCs sont « traditionnellement » utilisés en reconnaissance de la parole et Eronen/Kalpuri les ont utilisé couplés à des descripteurs temporels pour réaliser de la reconnaissance d'instruments de musique.

3. Le timbre comme « GMM de MFCCs »

L' « information timbre » n'est pas contenue dans une frame mais dans plusieurs frames. Le nombre de frames à considérer pose problème, combien de temps faut-il à l'oreille humaine pour apprécier le timbre ? C'est un problème de psycho acoustique intéressant mais qui ne s'inscrit aucunement dans la perspective de recherche de mon stage. Toutefois je mentionne que l'oreille humaine peut ne pas réussir à distinguer une

trompette d'un violon si elle n'a pas entendu l'attaque de la note. Cette anecdote, je trouve, met en évidence toute la difficulté de construire une mesure du timbre.
Aucouturier à l'aide d'un algorithme EM entraîne un modèle de mélange de gaussiennes (GMM) à partir de l'ensemble des frames d'un titre, soit :

$$p(\Im_t) = \sum_{m=1}^{M} \pi_m \mathrm{N}(\Im_t, \mu_m, \Sigma_m)$$

où $\Im_t$ est le vecteur de MFCCs observé à la frame t

Le timbre d'un titre est donc modélisé par un GMM de MFCCs, on considère ainsi une approche du timbre à grande échelle, celle d'un titre dans son intégralité, nous n'avons ainsi plus affaire à la notion perceptive du timbre. Le modèle de gaussienne donne une information plus globale pouvant répondre à la question « Comment est-ce que cela sonne ? »
En soi un tel modèle n'a presque aucune signification, il est seulement intéressant si l'on peut comparer le timbre de titres entre eux. Il faut pour cela disposer d'une distance entre deux modèles de mélange de gaussiennes.
Une approche de type Monte Carlo est utilisée pour fournir une distance entre deux GMM A et B :

$$D(A,B) = \sum_{i=1}^{DSR} \log p\left(S_i^A | A\right) + \sum_{i=1}^{DSR} \log p\left(S_i^B | B\right) - \sum_{i=1}^{DSR} \log p\left(S_i^A | B\right) - \sum_{i=1}^{DSR} \log p\left(S_i^B | A\right)$$

où DSR est le Distance Sample Rate et les S les vecteurs échantillons. L'expression, à base de vraisemblances conditionnelles, est symétrisée afin d'obtenir une distance.

Beaucoup de paramètres sont intervenus successivement, il nous faut pouvoir évaluer leur influence sur la performance de la distance. Je définis maintenant cette performance, qui est une « précision ».

A partir des données du AMG(All Music Guide) légèrement modifiées, on forme des clusters à partir de 350 titres de la base de données Cuidado. Ces clusters sont consistants du point de vue du timbre, par exemple on trouvera dans un même cluster les pistes contenant du piano solo indépendamment du genre musical. Les titres intervenants dans les clusters sont choisis pour leur homogénéité de timbre.
Considérons un titre S_i appartenant à un cluster C_{S_i} de taille N_i, on calcule à partir de la méthode de mélange de gaussiennes exposée plus haut, les N_i plus proches voisins de S_i. Parmi ces plus proches voisins, N appartiennent effectivement à C_{S_i}. Le rapport N/ N_i est appelé R-précision (méthode issue de la classification de textes). La précision générale est la moyenne sur tous les titres de la R-précision. C'est la quantité à optimiser par un bon choix de paramètres : fréquence d'échantillonnage du signal(SR),

nombre de MFCCs(N), nombre de gaussiennes pour le mélange(M), choix de DSR, choix de la distance entre deux GMM, durée d'une frame.

Une R-précision maximale de 0.63 est atteinte pour SR=44kHz, N=20, M=50, DSR=2000(Monte Carlo) et une frame autour de 30ms. Ce score peut être augmenté de deux points avec quelques travaux de pré processing sur le signal audio. Il est à noter le GMM de MFCCs ne prend pas en compte l'organisation temporelle dans sa structure contrairement à un modèle de Markov caché (HMM). Mais les résultats du HMM ne sont pas meilleurs que ceux du GMM.

Cette précision de 65%, Aucouturier la considère selon ces termes comme un glass-ceiling relativement admis dans la communauté ISMIR.

4. Reconnaissance audio

Aucouturier a construit une distance non euclidienne basée sur les MFCCs. Il s'est le plus possible, tenu à l'écart de l'approche reconnaissance de formes, celle que je vais à présent considérer.

La reconnaissance audio suit le même schéma que la reconnaissance de formes : extraction de descripteurs informatifs utilisés en entrée d'algorithmes de classification. Des descripteurs usuels sont par exemple ceux inclus dans la norme MPEG 7. Ces descripteurs peuvent être de nature temporelle ou spectrale. Ces descripteurs sont généralement perçus comme des « descripteurs du commerce ». Ils sont trop génériques pour résoudre efficacement des problèmes spécifiques. Pour ces derniers il faut élaborer minutieusement des descripteurs adaptés. L'inconvénient est que dès qu'un nouveau problème de reconnaissance apparaît, il faut refaire une étude signal isolée. Pour éviter cet écueil, le Sony CSL a développé un algorithme génétique (EDS pour Exctractor Discovery System) qui permet de trouver des descripteurs spécifiques à un problème donné (exemples : aboiements de chiens, bruits urbains…). Frédéric Roskam, élève du M2 ATIAM ayant effectué son stage en même temps que moi, a eu pour objectif d'évaluer précisément les performances d'EDS.

5. Noyau timbral/Table de similarité

Mon approche est différente. Je pars de l'espace de timbre d'Aucouturier et je vais proposer une mesure de similarité qui pourra être utilisée dans les méthodes dites « à noyaux » dont le SVM est le plus illustre représentant.

Aucouturier a pré calculé les distances timbrales sur une base de 5000 titres. A partir de celles-ci j'obtiens des similarités de la manière suivante :

$$s = \left(1 + \frac{1}{d}\right)^p$$

Dans cette expression d est la distance pré calculée dans une table, p sera un paramètre à optimiser.

Je dispose donc d'une matrice de similarité. Cette matrice est symétrique, définie et positive. Elle définit donc un noyau de Mercer. Je pourrai ainsi licitement entraîner un SVM à partir de celui-ci.

III. Catégorisation automatique de la musique

Lorsque j'emploie le terme catégorisation, je fais référence aux 17 catégories de la base de données dont je dispose, à savoir :

-Aera/Epoch
-Affiliate
-Character
-Country
-Dynamics
-Genre
-Language
-Main Instrument
-Metric
-Mood
-Musical Setup
-Rythmics
-Style
-Tempo
-Text Category
-Variant

Chacune de ces catégories est divisée en attributs booléens qui correspondent aux différents choix possibles. Pour la catégorie Aera/Epoch, une seule réponse est possible ; en revanche un fichier mp3 peut être catégorisé simultanément Style Rock et Style Funk. Pour cette dernière raison je n'ai pas considéré le problème de catégorisation comme 17 problèmes de classification multi-classes mais comme 780 problèmes de classification binaire.

Les classificateurs envisagés pour réaliser la tâche de catégorisation sont de deux types selon la nature des données d'entrée : le signal numérique ou des attributs booléens. Comme je dispose d'une base fiable de titres annotés à la main ainsi que des fichiers mp3 associés, j'ai développé un algorithme faisant intervenir les deux types d'information.

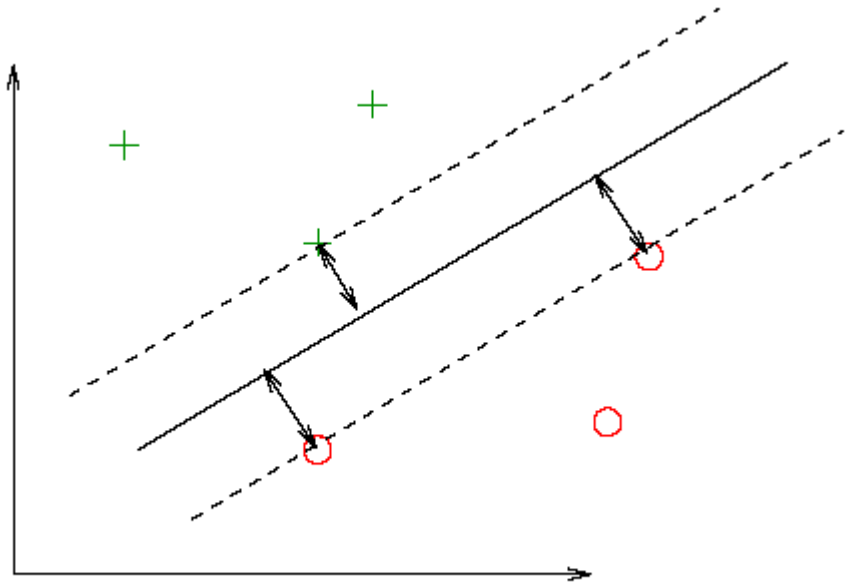
1. Catégorisation à partir d'attributs booléens : Peut-on prédire la popularité à partir des autres attributs ?

Notre base de données contient trois attributs de popularité : Low, Medium et High Popularity. Le problème posé est le suivant : connaissant les valeurs prises par chacun des autres attributs que la popularité, avec quelle exactitude peut-on prédire cette dernière ? Nous allons répondre à cette question en utilisant des Support Vector Machines comme classificateurs.

a. Support Vector Machines

Je n'ai pas l'intention de recopier un des nombreux cours disponibles sur les Support Vector Machines ; je me contenterai seulement d'en rappeler les grandes lignes. Un SVM est un classificateur, il est obtenu à partir de données d'entrée qui usuellement se

présentent sous la forme de vecteurs de descripteurs numériques obtenus à partir d'objets dont on connaît la classe d'appartenance. Nous sommes dans le cadre de l'apprentissage supervisé.
Schématiquement, dans l'espace vectoriel des données, le SVM est défini à partir d'un hyperplan optimal. Cet hyperplan sépare en maximisant la marge les données « true » des données « false » dans le cas d'un problème binaire.
En pratique cet hyperplan optimal est calculé par un algorithme d'optimisation quadratique. Ayant travaillé en Java, j'ai utilisé l'implémentation SMO (Sequential Minimum Optimization) disponible dans le package WEKA.
Voici une illustration simple d'un SVM, les données d'entraînement sur la marge sont appelées les vecteurs de support :



On note l'ensemble d'entraînement $(x_i, y_i)_{i=1..N}$ où les x_i sont les vecteurs de descripteurs et les y_i la valeur de leur classe d'appartenance. Dans le cas où ces données ne sont pas linéairement séparables, ce qui est effectivement le cas, le problème d'optimisation primale SVM s'écrit sous la forme :

$$\min_{w,b,\varepsilon} \frac{\|w\|^2}{2} + M\sum_{i=1}^{N}\varepsilon_i$$

Sous les contraintes : $y_i(w^T x_i + b) \geq 1 - \varepsilon_i$ et $\varepsilon_i \geq 0$ pour i=1..N

La marge, distance des vecteurs de support à l'hyperplan séparateur, vaut $\frac{2}{\|w\|}$ et $\sum_{i=1}^{N}\varepsilon_i$ représente l'erreur de classification en terme de distance. M est appelé dans la littérature anglo-saxonne Soft Margin Parameter, on voit qu'une faible valeur de M pénalisant peu l'erreur va conduire à un SVM qui ne 'colle' pas trop aux données, alors qu'une grande valeur de M va conduire à un SVM donnant de bons résultats sur la base d'entraînement mais dont la généralisation n'est pas assurée. Il faudra trouver une valeur de M adaptée.

En utilisant la méthode des multiplicateurs de Lagrange, on obtient le problème dual suivant :

$$\max_{\alpha_i} \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j x_i x_j$$

Sous les contraintes $\forall i, 0 \leq \alpha_i \leq M$ et $\sum_{i=1}^{N}\alpha_i y_i = 0$

Dans le cadre général des méthodes à noyaux, on montre via le théorème du représentant, que la solution de notre problème prend la forme :

$$f_{SVM}(x) = sign\left(\sum_{i=1}^{SV}\alpha_i y_i K(s_i, x) + b\right)$$

Les multiplicateurs de Lagrange $(\alpha_i)_{i=1..N}$ sont les solutions du problème d'optimisation dual, les $(s_i)_{i=1..SV}$ étant les SV vecteurs de supports. K est un noyau défini positif que l'on utilise en remplacement du produit scalaire dans le problème dual. Cette opération est connue sous le nom « astuce noyau », elle sous-tend un mapping (dont la formulation analytique n'a pas à être déterminée) des données dans un espace de dimension supérieur, espace dans lequel est effectué la recherche de l'hyperplan optimal par l'algorithme SMO.
Un noyau standard réputé pour ses bonnes performances est le noyau RBF, qui sous-tend un mapping dans un espace de Hilbert. Pour deux vecteurs de descripteurs x et z, son expression est la suivante :

$K(x,z) = \exp\left(-\frac{(x-z)^2}{2\sigma^2}\right)$ où σ est à paramétrer empiriquement

C'est ce noyau que j'ai utilisé pour tenter d'inférer la popularité d'un morceau à partir de ses autres attributs.

b. Complexité du problème et recherche des paramètres optimaux

Je pense, à tort peut-être, que la popularité des titres n'est régie par aucune loi de probabilité. C'est la raison pour laquelle je ne propose pas de borne théorique sur l'erreur de généralisation des modèles que nous allons apprendre.

Ma démarche est complètement empirique : à partir de mon ensemble de 10000 titres, taggé chacun avec approximativement 800 attributs booléens, je cherche de manière gloutonne les paramètres du SVM ainsi que la taille des données d'entraînement qui me donneront les meilleurs résultats.
Il a d'abord fallu tâtonner pour trouver la gamme viable de variations des paramètres. En effet, en dehors de celle-ci, toutes les mauvaises situations sont possibles : du total sur apprentissage au non apprentissage ! Cette gamme trouvée on peut commencer à travailler plus systématiquement.
Une bonne manière pour décider de l'optimalité de paramètres d'un SVM est de savoir si en les modifiant on fait chuter l'erreur de classification sur l'ensemble de test (on pratique la cross validation autant que possible) tout en parvenant à diminuer le nombre de vecteurs de support.

c. Résultats

Je présente ici les meilleurs résultats obtenus avec des bases d'entraînement de 1000 titres constituées à égale part d'exemples positifs et négatifs, et ceci pour chacune des trois classes. Les résultats présentés sont des taux d'erreurs (nombres d'erreurs de classification divisé par la taille de la base d'entraînement). Les paramètres que l'on fait varier sont le coefficient de pénalisation M et le paramètre du noyau gaussien σ.

Low Popularity

		σ		
		0.005	0.01	0.05
M	0.5	35%	33.5%	38.5%
	1	33%	31.5%	32%
	2	32%	34%	34%

	5	29.5%	36%	32.5%

Medium Popularity

		σ		
		0.005	0.01	0.05
M	0.5	25%	30%	31.5%
	1	34.5%	30.5%	26%
	2	33%	34.5%	32.5%
	5	34%	26%	26%

High Popularity

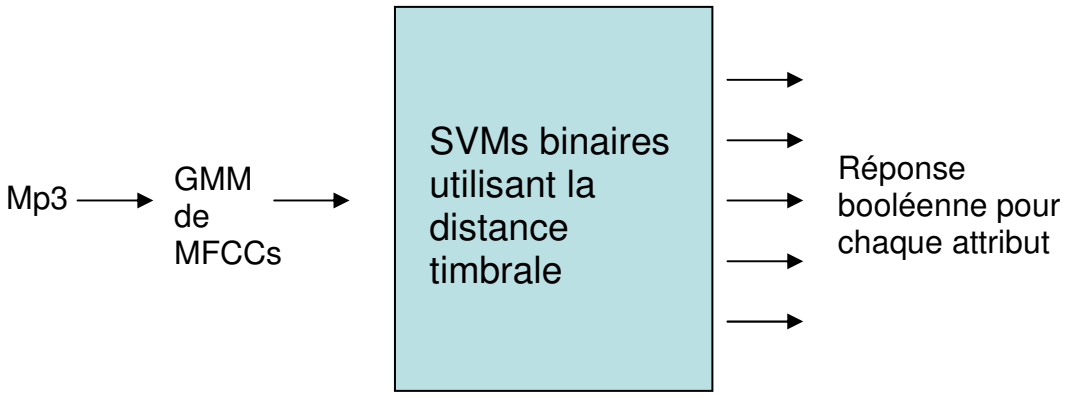
		σ		
		0.005	0.01	0.05
M	0.5	23.5%	28.5%	25%
	1	25.5%	26.5%	24%
	2	31.5%	21.5%	27.5%
	5	29.5%	23%	29%

Pour chacun des SVMs obtenu je tiens à préciser qu'avec des valeurs de soft margin supérieures ou égales à 5, l'erreur est nulle sur l'ensemble d'apprentissage, autrement dit notre classificateur apprend par cœur les données.
Même s'il semble difficile d'obtenir des taux très bons, on constate quand même que des trois attributs relatifs à la popularité, l'impopularité est la plus difficile à prédire. En choisissant les paramètres M=2, σ=0.05 on peut prédire à 78.5% si un titre sera très populaire. Ce qui n'est déjà pas si mal !

J'ai réalisé mon étude sur les attributs « Popularity », mais pareille étude peut naturellement être menée pour tous les autres attributs. Néanmoins pour obtenir les trois tableaux de résultats ci-dessus il faut trois jours de calculs. Il n'a donc pas été envisageable de mener l'étude pour tous les attributs. On sent pourtant qu'un jeu de paramètre n'est pas optimal pour tous les attributs.

2. Catégorisation à partir du Timbre

Le but recherché est le suivant : construire à partir du timbre des classificateurs pour chacun des attributs disponibles. Ainsi nous disposerons du système suivant :



a. Construction des classificateurs

J'ai choisi d'utiliser les SVMs pour réaliser cette tâche, parce que je dispose de données stockées dans une table de distance.

Je tiens à faire une courte remarque technique. La plupart des algorithmes libres qui réalisent l'entraînement des machines à vecteurs de support prennent en entrée des vecteurs de descripteurs. Le programme calcule ensuite par lui-même les valeurs des noyaux et gère la mémoire par une habile gestion du cache pour ne pas avoir à stocker une volumineuse matrice de similarité.
La librairie SMO de Weka, implémentant l'algorithme [3], ne déroge pas à la règle et il faut modifier laborieusement le code source pour parvenir à court-circuiter l'étape de calcul du noyau et le remplacer par un appel à la base de données.

Le noyau utilisé est le noyau timbral présenté dans la section précédente. Il me faut néanmoins trouver les paramètres optimaux du noyau « pseudo polynomial » que j'ai construit. Je dis « pseudo » car le « vrai » noyau polynomial a pour expression, dans le cas de données d'entrées sous forme de vecteurs de descripteurs :

$$K(x, y) = \left(1 + \langle x, y \rangle\right)^p$$

Dans mon cas les descripteurs sont les paramètres des modèles de gaussiennes, et donc mon noyau est polynomial mais fait intervenir une distance non euclidienne.

Certains attributs de ma base sont peu représentés, c'est pourquoi pour alléger les expériences je me suis concentré sur ceux qui apparaissaient un grand nombre de fois. C'est sur ceux-là que je vais préciser le choix des paramètres de mon noyau, et plus particulièrement sur ceux relatifs à la guitare. Dans la base les « attributs guitare » qui apparaissent au moins 250 fois sont : guitare électrique, guitare distorsion et guitare acoustique. Je vais entraîner des classificateurs timbraux pour ces 3 attributs en faisant varier l'exposant de mon noyau polynomial.

Je considère des ensembles d'entraînement de 250 titres constitués à égale proportion d'exemples positifs et négatifs pour l'attribut considéré. Je teste ensuite sur un ensemble de test de même taille contenant des éléments strictement différents de ceux utilisés pour l'entraînement.

Les résultats furent décevants et je les présenterai plus tard. Pour me faire une meilleure idée des performances je me suis intéressé aux éléments les plus représentatifs de chaque classe (les éléments les plus représentatifs de la classe guitare distorsion par exemple). J'expose ici ce que j'entends par plus représentatif.

b. Confiance d'un SVM

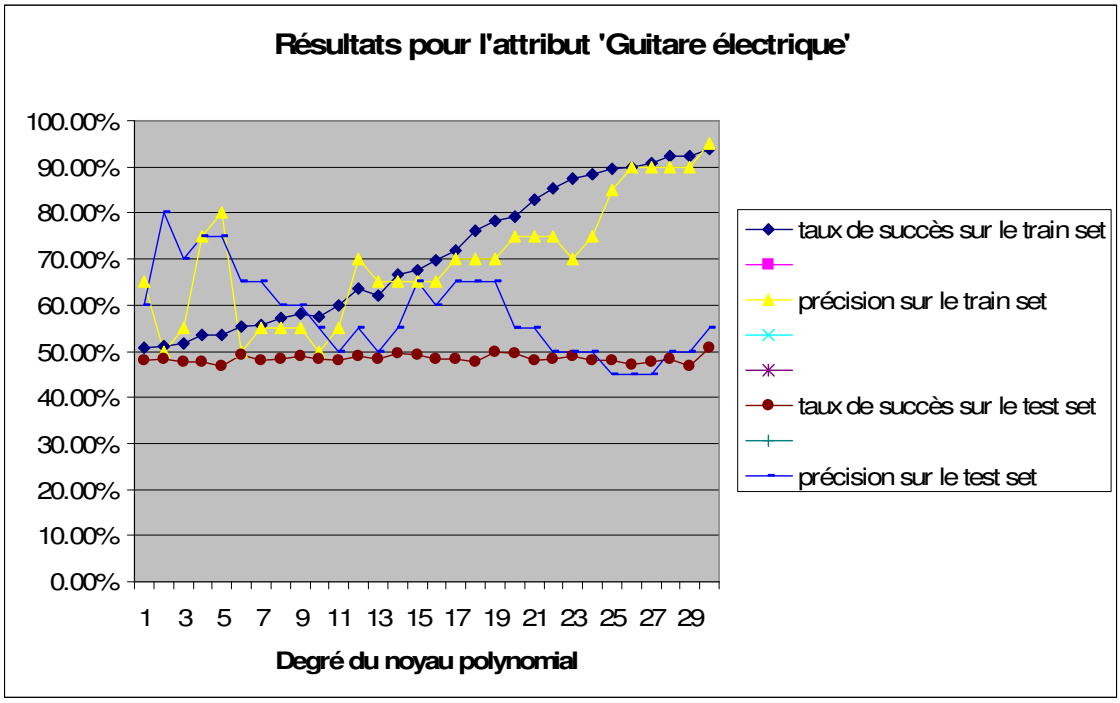
Considérons que nous disposons d'un SVM déjà entraîné, soit en fait son ensemble de vecteurs de support et les multiplicateurs de Lagrange associés.
Lorsqu'on teste un titre (une instance de test), pour savoir s'il est, ou non guitare distorsion, on calcule :

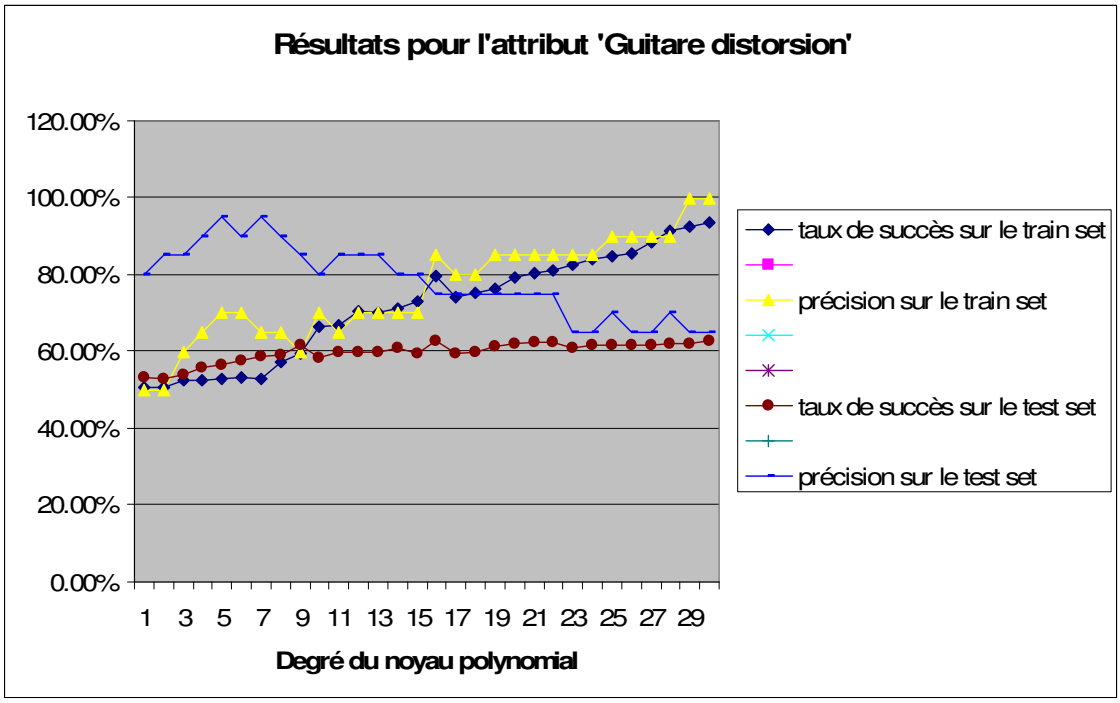
$$\sum_{i=1}^{SV} \alpha_i y_i K(s_i, x) + b$$

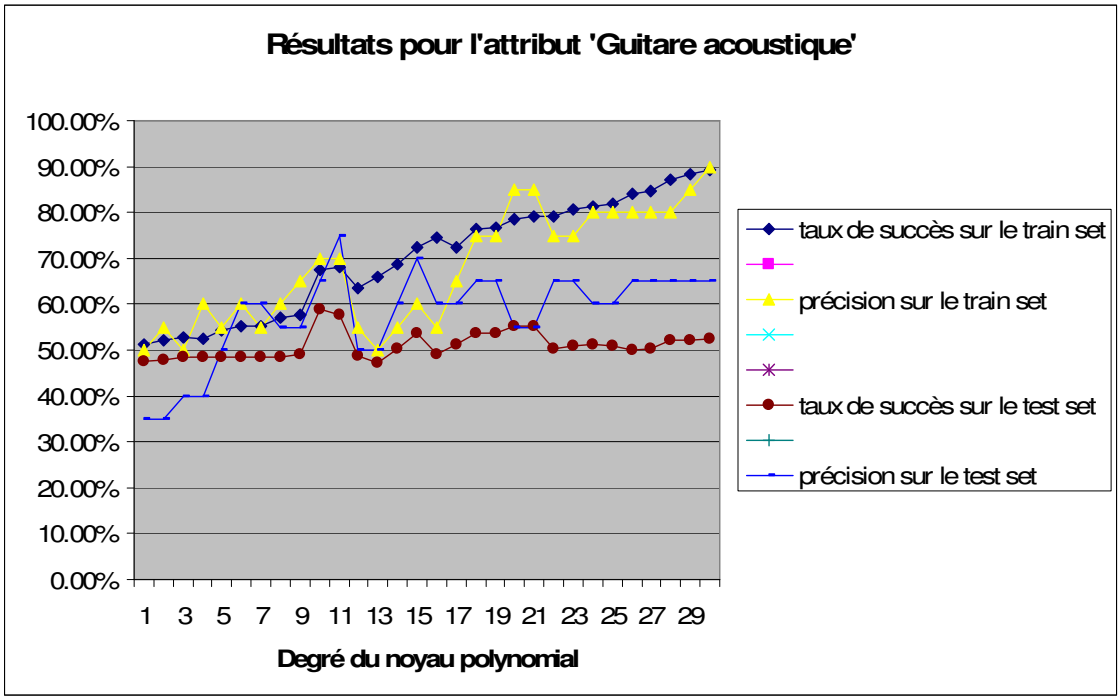
Si cette quantité est positive, alors le classificateur répondra oui, et si cette quantité est très grande (relativement aux résultats obtenus avec d'autres instances) on pourra considérer que notre SVM est sûr de sa prédiction.
Considérons maintenant un ensemble de test de 250 titres, je récupère à chaque fois la quantité ci-dessus et garde en mémoire les 10 titres qui ont les valeurs positives les plus grandes ainsi que les 10 titres qui ont les valeurs négatives les plus basses. Pour le classificateur dont j'évalue les performances, ces 20 instances sont celles pour lesquelles il est le plus sûr de sa prédiction. Sur celles-ci je calcule un taux de succès. On peut l'appréhender comme une précision.

c. Détermination des paramètres du noyau

Je présente les performances des 3 classificateurs relatifs à la guitare. Sont sur le même graphe l'erreur brute (taux de succès : nombre de titres correctement classé divisé par la taille de la base) sur les ensembles d'entraînement et de test ainsi que la précision timbrale (taux de succès pour les titres les plus distants en principe de la frontière de décision).


Résultats pour l'attribut 'Guitare électrique'
100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29
Degré du noyau polynomial
taux de succès sur le train set
précision sur le train set
taux de succès sur le test set
précision sur le test set


Résultats pour l'attribut 'Guitare distorsion'
120.00%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29
Degré du noyau polynomial
taux de succès sur le train set
précision sur le train set
taux de succès sur le test set
précision sur le test set



Présenter le taux de succès sur les instances les plus typiques m'a été inspiré par D. Ellis [2]. Le système qu'il a conçu retourne des titres musicaux à la manière de Google, en renvoyant d'abord les plus significatifs d'une classe donnée (Jazz par exemple). Il considère les catégories Style, Artist et Mood. La moyenne des performances sur tous les éléments de style de AMG (All Music Guide) est de 69% au mieux. Bien qu'il ne dispose pas de la même base de titres, ni du même système d'annotation, nos résultats sont du même ordre. Les miens sont meilleurs sans doute parce que mes bases d'entraînement sont 4 fois plus grosses. Présenter les performances d'un SVM en terme de précision et de confiance sur les instances les plus sûres permet de parler de taux de succès plus élevés et donc de « gonfler » les chiffres. Dans le cas d'Ellis ce mode de présentation des résultats est justifié par la nature de son système.

Le dernier exemple est enrichissant sur un point : lorsque le degré du polynôme est faible, le classificateur n'apprend rien, pas même l'ensemble d'entraînement. Lorsque la complexité du modèle augmente avec l'élévation de p, le SVM devient capable d'apprendre la base d'entraînement mais sa performance brute sur le test stagne « autour du hasard ». En revanche les deux précisions augmentent. C'est assez étonnant: on améliore la frontière de décision pour les exemples les plus typiques alors que la performance générale ne bouge pas.
Ces paradoxes de test sont d'autant moins appréhendables que l'on ne peut pas représenter graphiquement notre distance (qui est je le rappelle obtenue après un tirage de Monte Carlo).
Je me fie plus particulièrement à la précision sur l'ensemble de test, puisque l'erreur brute n'y est pas exploitable. Et là encore un problème se pose puisque des faibles degrés polynomiaux donnent de mauvais résultats pour la guitare acoustique, alors qu'ils sont bons sur les sonorités plus agressives.
L'attribut guitare distorsion est intéressant puisque les performances brutes sont « bonnes », relativement aux autres. L'augmentation du degré fait chuter la précision.

Comme par la suite je ne pourrai me permettre, pour des raisons de temps de calcul, de réaliser une recherche gloutonne sur le noyau optimal, je fais le choix du compromis. Le « noyau timbral standard » sera de degré 13 pour porter chance. La mesure de similarité timbrale standard entre 2 titres sera :

$$K(titre1, titre2) = \left(1 + \frac{1}{d_{12}}\right)^{13}$$

d. Performances des classificateurs timbraux

J'ai entraîné des SVMs utilisant le noyau timbral standard pour des attributs apparus au moins 200 fois. Les bases d'entraînement sont constituées de titres pris au hasard dans la base totale. Comme il a été toujours le cas dans mes expériences, les bases d'entraînement contiennent autant d'instances positives que négatives. Il en résulte que 249 attributs sont candidats à servir de cobaye pour la construction d'un classificateur timbral.
On calcule un taux de succès sur les ensembles d'entraînement et de test pour chacun des attributs, les moyennes respectives de ces taux sont 66.6% et 53.7%. Ces résultats ne sont vraiment pas bons.
Il n'y rien de très surprenant à ce que l'attribut 'Genre Well-Known Music' ait un taux de succès de 44% sur l'ensemble de test. Nul de besoin de faire un calcul d'entropie conditionnelle pour conclure que le timbre ne contient pas d'information sur la popularité relative d'un titre.
En revanche les bonnes performances de certains attributs sont moins étonnantes :

-Style Hip Hop (73%)
-Style Rap (71.5%)
-Variant aggressive (71%)

Si on ne considère que les 50 meilleurs attributs, les taux de succès sur train et test sont 71.3% et 61.6%.
On ne va bientôt plus considérer que ces 50 meilleurs classificateurs. Peut-on améliorer leur performance individuelle en les faisant collaborer ? Par exemple si on s'intéresse à construire un classificateur pour l'attribut 'Textcategory Anger'. Imaginons que sur un titre de test il prédise 'faux' et qu'il se trompe. Imaginons maintenant que pour ce même titre en cours d'analyse nous disposions maintenant des classificateurs pour les attributs 'Variant aggressive', 'Main Instruments Vocals (Spoken; Rap)', 'Mood belligerent', et que ces classificateurs prédisent 'true' alors il y a de fortes chances pour que nous puissions corriger notre classificateur timbrale seul 'Textcategory Anger'.

Je vais maintenant présenter un algorithme qui réalise la catégorisation automatique des 50 attributs qui fournissent les classificateurs timbraux les plus performants.

3. Catégorisation automatique pour 50 attributs

Je tiens à faire une petite remarque d'ordre lexicale. Les mots catégorisation et classification sont ici synonymes, à la nuance près que catégorisation peut sous-entendre l'ensemble des résultats de classification.

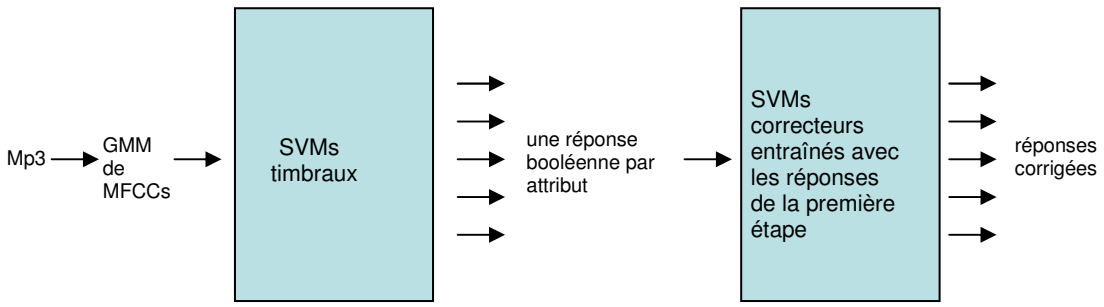
a. Architecture du classificateur

Une fois entraîné mon modèle fonctionne pour un titre de cette manière :

mp3 --> 50 réponses des classificateurs timbraux -> réponses corrigées

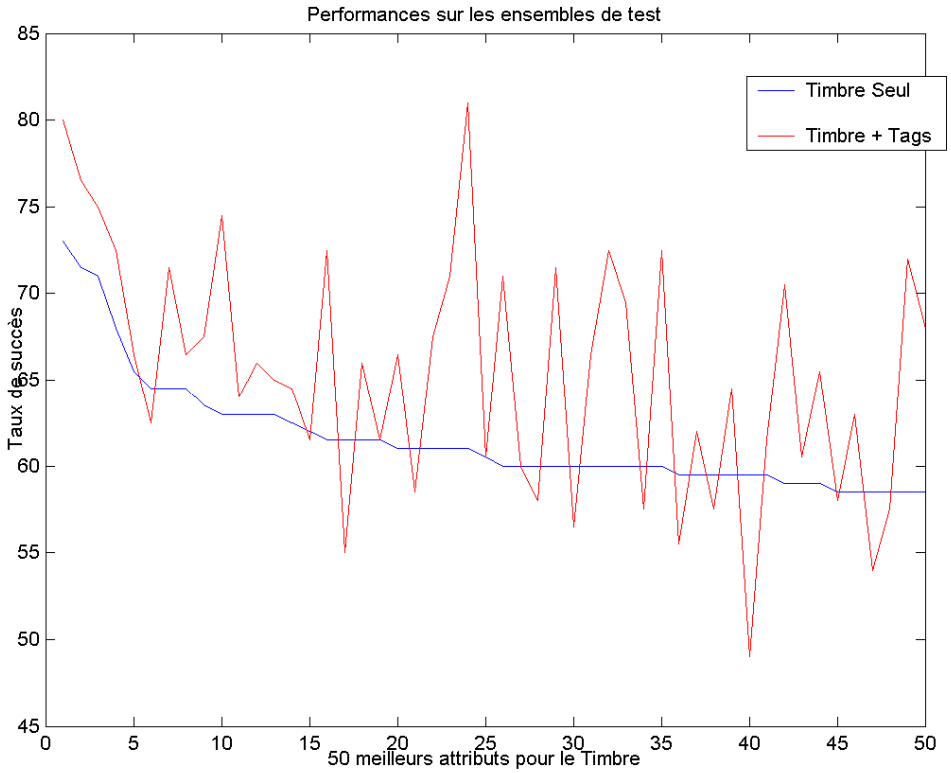
La première phase de catégorisation aboutit aux réponses de 50 classificateurs timbraux. Chacun de ces classificateurs a été entraîné sur une base propre à l'attribut considéré. La deuxième phase est la réponse de 50 classificateurs SVM prenant en entrée les 50 réponses booléennes de la première étape. C'est une étape de correction. Ces classificateurs correcteurs sont indépendants entre eux.

Pour bien faire comprendre la deuxième étape je dirai qu'elle est similaire à la catégorisation par attributs booléens, à deux différences prêt : 1.la classe pour laquelle on construit le classificateur est en entrée, 2.toutes les entrées comme résultats de classificateurs timbraux fournissent des réponses fiables en moyenne à 61.6%.
Les SVMs de la deuxième étape ont un noyau gaussien avec les paramètres « optimaux », M=1 et gamma=0.01. (déterminés en I.)



b. Résultats et discussion

Mon algorithme de catégorisation est entraîné sur des bases équilibrées de 200 titres. Je le teste sur les ensembles utilisés pour la catégorisation timbrale, ceci afin de pouvoir apprécier l'amélioration en performances.



Pour ces 50 attributs les résultats sont bons. En moyenne le taux de succès sur l'ensemble de test est maintenant de 65.3%. Nous avons également amélioré les taux des 3 meilleurs attributs de l'étape Timbre :

-Style Hip Hop (80%)
-Style Rap (76.5%)
-Variant aggressive (75%)

Les attributs qui ont le plus bénéficié de l'étape de correction sont : 'Vocals spoken (Rap)', 'Mood intimate', 'Musical Setup DJ' et 'Situation Dream'. Ce dernier attribut a été correctement classé à 72%, au lieu de 58.5% avec le timbre seul.

Je présente en fin de paragraphe l'ensemble des résultats. Pour réaliser cette même tâche de catégorisation faisant intervenir timbre et tags, les chercheurs du SONY CSL avaient procédé d'une manière différente [7]. Leur première étape sélectionne les meilleurs

attributs pour le timbre en utilisant un critère de kNN. La deuxième étape construit des arbres de décision à partir des attributs sélectionnés à la première étape. Leurs expériences ont porté sur l'amélioration de 45 attributs qui présentaient une bonne corrélation au timbre.

Je n'ai pas eu le temps de comparer mes résultats avec les précédents travaux du SONY CSL sur le même sujet. Néanmoins mes résultats sont prometteurs et des expériences complémentaires vont être lancées. Notamment il serait intéressant d'augmenter le nombre de SVMs utilisés à la seconde étape et de voir si les performances augmentent, malgré la moins grande fiabilité des SVMs supplémentaires.

Attributs	Taux de succès sans correction	Taux de succès après correction
Style Hip Hop	73%	80%
Style Rap	71.5%	76.5%
Variant aggressive	71%	75%
Mood aggressive	68%	72.5%
Popularity Popularity low	65.5%	66.5%
Mood melancholic/longing	64.5%	62.5%
Situation Fight	64.5%	71.5%
Situation Country Life	64.5%	66.5%
Aera/Epoch 1960-1970	63.5%	67.5%
Main Instruments Vocals (Spoken; Rap)	63%	74.5%
Variant natural / acoustic	63%	64%
Main Instruments Synth Bass	63%	66%
Mood mighty	63%	65%
Genre Ballad	62.5%	64.5%
Rhythmics with Swing	62%	61.5%
Variant Distortion	61.5%	72.5%
Aera/Epoch 70s	61.5%	55%
Mood romantic	61.5%	66%
Main Instruments SFX (Sound Effects)	61.5%	61.5%
Character calm	61%	66.5%
Situation Danger	61%	58.5%
Mood belligerent	61%	67.5%
Mood intimate	61%	71%
Musical Setup DJ	61%	81%
Character repeating	60%	60.5%
Musical Setup Synthetic	60%	71%
Style Rock/Pop	60%	60%
Rhythmics laidback	60%	58%
Variant metalic	60%	71.5%
Mood nostalgic	60%	56.5%
Character creamy	60%	66.5%
Style Dance	60%	72.5%
Style Alternative Rock	60%	69.5%
Country England (United Kingdom)	60%	57.5%
Situation Computer Animation	60%	72.5%

Affiliate Affiliate International	59.5%	55.5%
Musical Setup Pop Band	59.5%	62%
Character friendly	59.5%	57.5%
Mood sensual	59.5%	64.5%
Musical Setup Choir	59.5%	49%
Textcategory Anger	59.5%	61.5%
Character distorted	59%	70.5%
Main Instruments Strings	59%	60.5%
Mood technical	59%	65.5%
Rhythmics rhythmic	58.5%	58%
Variant synthetic	58.5%	63%
Mood dreamy	58.5%	54%
Character narrative	58.5%	57.5%
Situation Dream	58.5%	72%

Bibliographie

[1] J-J. Aucouturier, F. Pachet « Improving Timbre Similarity: How high's the sky? » (2004) SONY CSL
[2] D. Ellis, G. Poliner, M. Mandel "Support Vector Machine Active Learning for Music Retrieval" (2005) ACM Multimedia Systems Journal
[3] J. Platt "Fast Training of Support Vector Machines using Sequential minimal Optimization" (1999) Advances in Kernel Methods
[4] J-J. Aucouturier, F. Pachet "Music similarity measures: What's the use?"(2002) SONY CSL
[5] M. Plumbey "Clustering of Sparse Binary Data using a Minimum Description Length Approach" (2002) Queen Mary, University of London
[6] J-J. Aucouturier "Dix expériences sur la modélisation du timbre polyphonique » Thèse soutenue à Paris 6 (Avril 2006)
[7] J-J. Aucouturier, F. Pachet, P. Roy, A. Beurivé « Towards Automatic Music Categorization »(2005) Internal Report-Confidential

Remerciements

Antony Beurivé m’a été d’un grand secours à chaque fois que je bloquais sur des problèmes d’implémentation. N’étant pas expert en programmation JAVA, il m’a permis d’optimiser le code. Sans son aide je n’aurais pu mener à bien mon stage. Je le remercie donc tout particulièrement.
Je remercie tous les chercheurs du SONY CSL ainsi que Sophie Boucher pour leur accueil agréable. Je remercie également Frédéric Roskam, étudiant du M2 ATIAM, qui m’a fait partager ses connaissances en informatique musicale.
Enfin je remercie François Pachet pour ses justes critiques tout au long du stage.