Application sur un signal audio

Sommaire

Ayant introduit la théorie des Machines à Vecteurs de Support, nous nous intéressons maintenant à leur mise en œuvre sur le problème spécifique de la classification audio. Après avoir présenté l'architecture globale du système dans la section 6.1, nous verrons en section 6.2 comment le signal audio est traité en entrée pour se conformer au cadre théorique exposé précédemment. La constitution des exemples d'apprentissage pour les SVM se fait par le calcul de descripteurs audio, dont nous présenterons en section 6.5 le panel choisi pour caractériser au mieux les classes mises en jeu. Une courte discussion sur ces derniers (section 6.6) nous permettra de mettre en évidence plusieurs modalités dominantes de description du signal, dont les descripteurs sont fortement corrélés. Ce constat nous conduira donc à nous intéresser dans le chapitre suivant au problème de la sélection automatique de descripteurs.

6.1 Architecture du système de classification

L'architecture du système mis en place est résumée dans la figure 6.1.

Nous avons traité, dans les chapitres 3 et 4, de la question de l'apprentissage des SVM ainsi que de la sélection du noyau optimisant les performances par rapport à un ensemble d'apprentissage. Nous aborderons dans ce chapitre la constitution de l'ensemble d'apprentissage, par l'extraction de descripteurs audio calculés sur le signal audio du corpus d'apprentissage après un découpage en trames.

De par la nature discriminative des SVM, nous avons vu dans le chapitre 5 qu'il est nécessaire de combiner plusieurs discriminateurs dans une situation impliquant plus de deux classes. Le processus hiérarchique de combinaison des classifieurs est synthétisé dans une taxonomie multi-classes, que l'on retrouve en haut à gauche de la figure 6.1. Cette taxonomie implique un ensemble de classifieurs, chacun étant destiné à discriminer une paire de classes donnée. Sur chacune de ces paires on appliquera donc, de manière indépendante, les traitements contenus dans l'encadré jaune de la figure.

Ainsi, nous verrons dans le chapitre 7 qu'à la phase d'extraction générale des descripteurs succède une sélection des descripteurs les plus pertinents, qui est bien sûr propre à la paire de classes considérée. Un modèle SVM est par la suite appris sur les descripteurs sélectionnés.

L'application du système sur le corpus audio d'évaluation (ou de test) suivra la même séparation en processus distincts propres à chaque paire. Ainsi on n'extraira cette fois-ci que les descripteurs sélectionnés, puis on classifiera les exemples inconnus au moyen du modèle SVM appris. Ensuite,

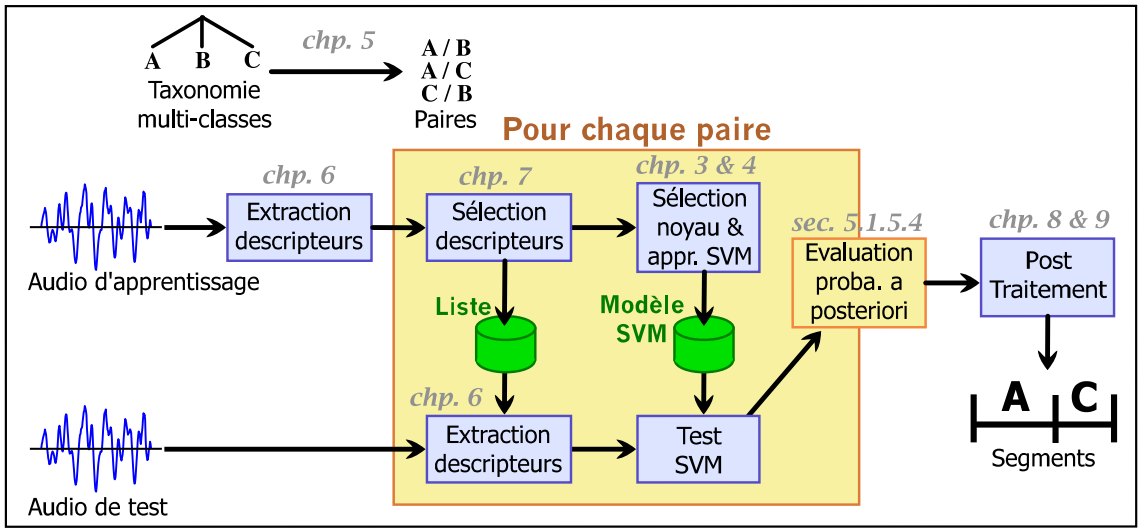


FIGURE 6.1 – Architecture générale du processus de classification audio par combinaison de SVM.

en suivant la méthode présentée dans la section 5.1.5.4, on estime les probabilités a posteriori qui nous permettent d'appliquer l'un des procédés de post-traitement dynamiques qui seront présentés dans les chapitres 8 et 9.

6.2 Analyse du signal en trames

La tâche de classification porte sur un signal audio numérique fini de L échantillons. On supposera par la suite que celui-ci est échantillonné à $f_s = 16$ kHz. Le signal est représenté par un vecteur $\boldsymbol{x}$ de L échantillons : $\boldsymbol{x} = [x_1, \ldots, x_L]^T$. La valeur de chaque échantillon indépendamment des autres n'apporte aucune information concernant les propriétés du signal à un instant donné. On doit donc considérer un ensemble de N échantillons successifs appelé *trame*. On admettra qu'un signal audio est stationnaire sur une durée inférieure à 40 ms. Afin d'optimiser le calcul de la FFT (Transformation de Fourier Rapide), on choisira donc la puissance de 2 la plus élevée satisfaisant cette contrainte, soit $N = 2^{\lfloor \log_2(f_s \times 0.04) \rfloor} = 512$ ce qui correspond à des trames de 32 ms.

On utilise généralement un pas d'avancement de R échantillons entre les trames, qui est inférieur à la taille de la fenêtre, afin d'accroître la précision temporelle de la classification ; on parle alors de trames *chevauchantes*. On choisit ici $R = \frac{N}{2} = 256$.

Le signal sera donc caractérisé par un ensemble de valeurs calculées sur ces trames temporelles. Nous verrons qu'une partie implique le spectre fréquentiel défini par l'analyse de Fourier. On calcule donc les 256 composantes d'amplitude a_k et de phase ϕ_k associées à chaque *bin* fréquentiel d'indice k, après pondération de la trame par une fenêtre de Hamming.

La classification sur des exemples caractérisant des trames temporelles constitue le paradigme de base de l'apprentissage statistique. Cette approche est généralement appelée *sac de trames* (*bag of frames*).

6.3 Intégration temporelle

Nécessaire pour caractériser les propriétés instantanées du signal, le découpage en trames courtes reste cependant lacunaire puisque nombre de phénomènes acoustiques n'ont de sens que sur une portée temporelle plus longue ; par exemple, en musique et en parole, le trémolo et le vibrato sont des grandeurs impliquant des modulations d'amplitude ou de fréquence sur une durée de l'ordre d'une seconde. On peut d'ailleurs montrer que la durée nécessaire à un humain pour la reconnaissance de genre musical est de l'ordre de 0.5 à 3 secondes [184]. On trouvera dans [243] une étude comparative de plusieurs horizons temporels pour la classification de genres musicaux.

La prise en compte d'une échelle temporelle plus étendue se fait généralement au travers de deux moyens [155] : soit par l'inclusion de descripteurs dit « long-terme », directement calculés sur des trames longues de l'ordre d'une seconde, également nommées *fenêtres de texture* par Tzanetakis et Cook [229] et d'autre auteurs [41][164], soit par l'intégration statistique des valeurs des descripteurs court-terme sur les trames longues. Nous présentons ci-dessous ces deux possibilités que nous exploitons conjointement.

Trames longues

Nous définissons, par opposition aux *trames courtes* introduites précédemment, des *trames longues* d'une durée d'une seconde, temps correspondant au consensus global sur la fenêtre de texture citée précédemment. Afin de pouvoir synchroniser ces deux modalités, les trames longues ont pour taille le nombre exact d'échantillons impliqués dans une série de N_{mul} trames courtes, soit la longueur N_l suivante :

$$N_l = (N_{\text{mul}} - 1)R + N.$$

L'avancement R_l entre deux trames longues est également choisi de manière à ce que chaque trame longue soit synchronisée avec le début d'une trame courte, d'où :

$$R_l = R_{\text{mul}}R,$$

où R_{mul} est le nombre de trames courtes d'avancement entre chaque trame longue. Afin d'obtenir des trames longues de l'ordre d'une seconde, on choisit $N_{\text{mul}} = 60$ (soit $N_l/f_s = 0.976$ s). Les trames se chevauchent, comme pour les trames courtes, d'environ 50%, soit $R_{\text{mul}} = 30$. On parlera respectivement, pour les descripteurs calculés sur les trames courtes et longues, de descripteurs court-terme et long-terme.

La figure 6.2 illustre la synchronisation induite par les grandeurs introduites.

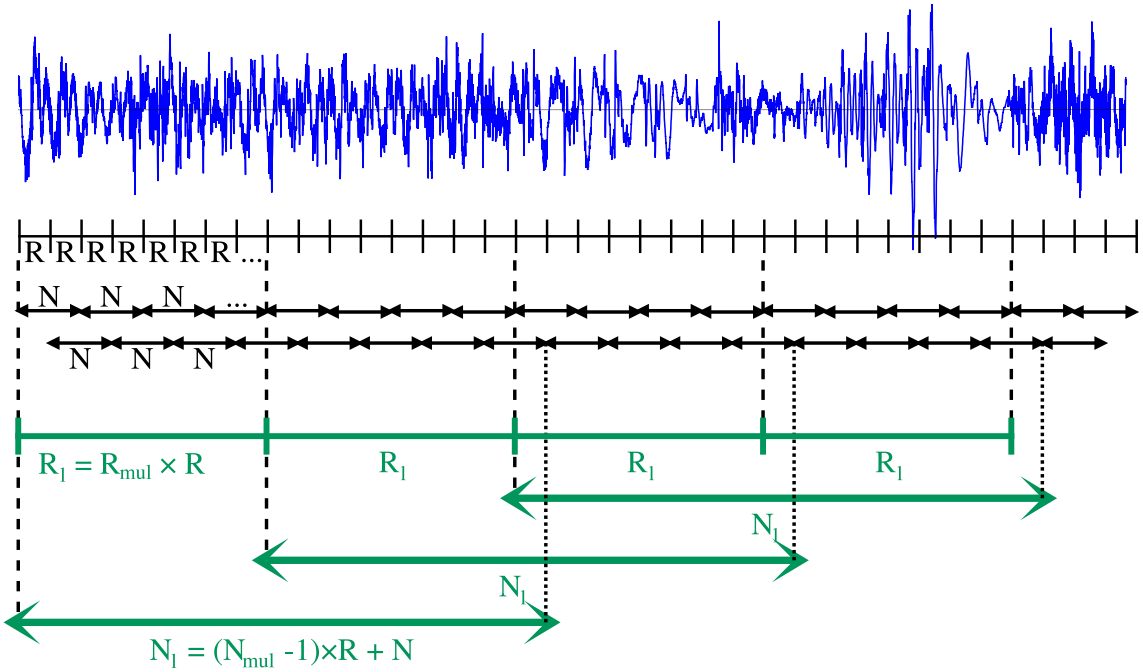


FIGURE 6.2 – Représentation de la synchronisation entre les frontières de trames courtes (segments fléchés en noir) et de trames longues (segments fléchés en vert).

Intégration statistique

Le découpage en trames courtes est nécessaire pour évaluer les propriétés instantanées définies par certains descripteurs. Cependant, la précision induite, de l'ordre de la dizaine de millisecondes,

est largement supérieure aux besoins pratiques. Nous verrons par la suite que la campagne d'évaluation ESTER tolère une erreur de 0.25 s sur les frontières de segments de classification. Plusieurs études [184][7] établissent d'ailleurs que la durée nécessaire à un être humain pour classifier un extrait audio est de cet ordre de grandeur.

De plus, d'un point de vue pratique, la classification sur trames courtes est extrêmement coûteuse en ressources et implique un nombre très élevé d'exemples (de l'ordre de 200000 par heure de signal). Étant donné que l'on se limite à quelques dizaines de milliers d'exemples pour l'apprentissage des SVM [1], on en déduit que le corpus d'apprentissage serait caractérisé de manière très parcellaire sur une base de plusieurs dizaines d'heures.

En pratique on choisit donc non pas de considérer les descripteurs eux-mêmes, mais certaines grandeurs statistiques calculées sur un nombre de trames suffisamment significatif. On exploite ici la synchronisation entre trames courtes et longues en calculant ces mesures statistiques sur les trames courtes couvertes par chaque trame longue. De plus, une comparaison de plusieurs échelles temporelles d'intégration statistique montre [182] que les meilleurs résultats sont obtenus pour 1 s, ce qui correspond à la longueur de nos trames longues. À chaque trame longue est donc associé un vecteur de descripteurs composé de statistiques de descripteurs court-terme et de descripteurs long-terme sans traitement statistique.

Ainsi, si $x(n)$ est la suite des valeurs d'un descripteur court-terme, où n est l'index de trame, on calcule les descripteurs long-terme $X(m)$, indexés par l'indice de trame longue m (on suppose que les indices débutent à 0) :

$$X(m) = f\left(x(mR_{\text{mul}}), x(mR_{\text{mul}}+1), \ldots, x(mR_{\text{mul}}+N_{\text{mul}}-1)\right).$$

f représente ici le traitement statistique appliqué ; on parle également d'intégration de descripteurs, dans un sens plus large. Nous n'exploitons dans cette étude que les grandeurs les plus couramment exploitées [140][30][229] : la moyenne et l'écart type.

Les bornes minimales et maximales sont également exploitées par certains auteurs [48] mais celles-ci sont trop sensibles à d'éventuelles valeurs marginales excentrées. Nous avons d'ailleurs montré dans une étude précédente [191] que leur inclusion dans le processus de classification parole/musique n'améliore pas les performances.

D'autres intégrations ont été explorées dans la littérature. En particulier, Meng a proposé [154] l'exploitation des coefficients d'un modèle auto-régressif appris sur les descripteurs d'une trame longue et fournit par ailleurs une étude comparative impliquant divers procédés d'intégration [155]. Toutefois, les coefficients d'un modèle AR sont connus pour être instables et leur évolution est discontinue par rapport aux variations du signal. On trouvera également dans [116] une étude assez exhaustive des différents procédés d'intégration sur une tâche de reconnaissance automatique des instruments de musique. Les résultats présentés ne montrent cependant pas d'avantage clair et systématique à utiliser des méthodes d'intégration plus complexes et confirment ainsi notre choix de ne pas explorer plus en profondeur ce sujet.

6.4 Normalisation des descripteurs

Les descripteurs obtenus après dérivation et intégration temporelle proviennent de modalités différentes et leur dynamique est très hétérogène. Pourtant, dans tous les noyaux usuels que nous exploitons, les descripteurs sont mis en concurrence au travers de sommes à pondérations uniformes, par exemple $k(\boldsymbol{x}, \boldsymbol{y}) = \sum_{d=1}^{D} x_d y_d$ dans le cas du noyau linéaire, ou $k(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(\sigma^{-2} \sum_{d=1}^{D} (x_d - y_d)^2\right)$ pour le noyau RBF gaussien, où D est le nombre de composantes. Ainsi il est clair qu'un descripteur de moyenne largement supérieure à celle d'un autre descripteur couvrira ce dernier et le rendra presque « muet » dans l'expression de la fonction noyau.

On normalise donc les descripteurs de manière à réduire les disparités statistiques. La méthode la plus classique [225], que nous employons ici, consiste à homogénéiser les statistiques de premier

1. Cette limitation est due à un compromis entre performances et temps de calcul, rendu nécessaire par la complexité quadratique de la phase d'apprentissage.

et de deuxième ordre. Ainsi, si l'on note $x_{i,d}$ la composante d'indice d de l'exemple $\boldsymbol{x}_i$, on estime la moyenne μ_d et la déviation standard σ_d du descripteur d par les estimateurs statistiques classiques :

$$\begin{aligned} \mu_d &= \frac{1}{n}\sum_{i=1}^{n} x_{i,d} \\ \sigma_d^2 &= \frac{1}{n-1}\sum_{i=1}^{n} \left(x_{i,d} - \mu_d\right)^2 . \end{aligned}$$

Les composantes normalisées prennent donc l'expression suivante :

$$\hat{x}_{i,d} = \frac{x_{i,d} - \mu_d}{\sigma_d}.$$

Malgré le consensus autour de cette méthode de normalisation, il est important de rappeler que celle-ci se base sur un postulat de gaussianité des distributions des descripteurs. Or, dans le cas de distributions moins régulière, cette procédure de normalisation ne permet pas d'obtenir, comme souhaité dans l'idéal, des données gaussiennes centrées. Cette hypothèse est confirmée en pratique dans de nombreux cas de descripteurs non-bornées. Toutefois, les descripteurs bornés ou semi-bornés (l'exemple le plus courant est celui de descripteurs positifs, comme c'est le cas pour des mesures d'énergie), ne satisfont généralement pas le modèle gaussien, et s'approchent plutôt d'un modèle de distribution Gamma. L'effet de la normalisation « gaussienne » sur une distribution Gamma n'est pas évident et ne sera pas couvert dans ce document.

Une manière courante [12] pour s'affranchir de la distribution des données consiste à substituer à la valeur $x_{i,d}$ la valeur de fonction de répartition $F(x_{i,d})$, estimée sur l'ensemble des exemples. Ainsi, on a la garantie d'obtenir pour toutes les composantes une distribution quasi-uniforme sur l'intervalle $[0; 1]$. Il est équivalent, à un facteur multiplicatif près, de substituer au descripteur son rang parmi les valeurs de l'ensemble d'apprentissage, triées par ordre croissant, comme le proposent Stolcke et al. [222]. Toutefois, nous n'avons pas constaté un effet notable de ces techniques de normalisations alternatives sur les performances des SVM, nous n'explorons donc pas cette question plus en détail dans ce document. On trouvera dans [12] une étude comparative des différentes méthodes de normalisation de données.

6.5 Liste des descripteurs employés

Nous présentons brièvement dans cette section la collection de descripteurs réunis pour les tâches de classification traitées dans ce document. Ceux-ci sont regroupés selon la modalité de calcul. On distingue ainsi des descripteurs spectraux, calculés sur le spectre estimé par FFT, des descripteurs temporels, calculés directement sur le signal audio, des descripteurs cepstraux et des descripteurs perceptifs, basés sur des modèles de propriétés psychoacoustiques de l'audition humaine. Les descripteurs employés étant pour la plupart bien connus de la communauté, leur présentation détaillée a été reportée dans l'annexe C.

Pour certains des descripteurs proposés, l'évolution temporelle de la valeur peut être aussi significative, voire plus, que la valeur elle-même. Ainsi, on ajoute à la plupart des descripteurs les estimations des dérivées premières et secondes, qui forment elles-mêmes de nouveaux descripteurs.

Descripteurs spectraux

Les descripteurs spectraux sont calculés à partir du spectre estimé par la Transformée de Fourier Discrète (TFD), qui est définie, sur une trame de N échantillons, de la façon suivante :

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-2j\pi k \frac{n}{N}} \quad \forall k \in [0, \ldots, N-1].$$

Le calcul de la TFD est précédé de la pondération du signal de trame par une fenêtre de Hamming, qui limite l'étalement des pics spectraux. En pratique, seuls les amplitudes $a_k = |X(k)|$ sont utilisées dans les descripteurs spectraux présentés ci dessous :

- Les *moments statistiques spectraux* : cette série de descripteurs est basée sur le barycentre et les moments d'ordre i d'un modèle probabiliste du spectre ; elle contient :
  - le *centroïde spectral*,
  - la *largeur spectrale*,
  - l'*asymétrie spectrale* (ou *Skewness*),
  - la *platitude spectrale* (ou *Kurtosis*).
- *Descripteurs MPEG-7* : plusieurs descripteurs liés au standard MPEG-7 [3] sont ici exploités :
  - le *rapport spectral*, qui constitue une alternative à la mesure de platitude spectrale,
  - la *platitude d'amplitude spectrale (ASF)*,
  - le *facteur de crête spectral (SCF)*, proposé par Peeters [178] qui, bien qu'il ne fasse pas partie du standard MPEG-7, reste très proche des deux descripteurs précédents, dans sa définition.
- La *pente spectrale*, qui représente le taux de décroissance spectrale.
- La *décroissance spectrale*, qui mesure la décroissance des amplitudes spectrales.
- La *fréquence de coupure*, liée à une mesure de quantile de l'énergie spectrale.
- Le *flux spectral*, défini par Scheirer et Slaney [207], comme une mesure de variation spectrale entre trames consécutives.
- Les *coefficients LPC (Linear Prediction Coding)*, caractérisant un modèle source-filtre pour le codage audio.
- Les *sous-bandes en octaves*, proposés par Essid pour la reconnaissance d'instruments de musique [72], et destinés à capturer la structure spectrale de sons instrumentaux. Ils se composent des deux sous-groupes suivants :
  - les intensités de signaux de sous-bandes en octaves, nommées OBSI (cf. annexe C)
  - et les rapports d'intensité de signaux de sous-bandes en octaves, nommés OBSIR.
- Plusieurs mesures de *modulation d'amplitude* sont exploités pour caractériser les phénomènes de trémolo et de rugosité, qui se manifestent respectivement sur les bandes de fréquences entre 4 et 8 Hz et entre 10 et 40 Hz. Quatre critères sont définis pour chaque bande :
  - la *fréquence AM* du pic maximal,
  - l'*amplitude AM* du pic maximal,
  - l'*amplitude AM heuristique* du pic maximal, par rapport à la bande de fréquences,
  - et le *produit AM* de la fréquence et de l'amplitude AM.

Descripteurs temporels

Les descripteurs suivants sont basés exclusivement sur la forme d'onde du signal audio dans une trame courte (ou longue si précisée) et ne font pas intervenir le spectre.

- Le *taux de passage par zéro (ZCR)*, proposé par Kedem [118], et dont on peut montrer la corrélation au centroïde spectral.
- Le *taux de passage par zéro long-terme*, calculé sur les trames long-terme.
- Les *moments statistiques temporels*, qui reprennent les mêmes moments que ceux définis sur le spectre. Ils sont calculés :
  - sur les trames court-terme,
  - sur les trames long-terme,
  - sur l'enveloppe des trames longues, estimée par le biais de la transformée de Hilbert.
- Les *coefficients d'autocorrélation*.

Descripteurs cepstraux

Le *cepstre* complexe d'un signal est défini à l'origine en 1963 par Bogert et al. [33], comme la transformée de Fourier du logarithme du spectre, soit :

$$C(\tau) = C(z(t)) = \mathcal{F}\left\{\ln\left(\mathcal{F}\left\{z(t)\right\}\right)\right\}.$$

Originellement proposé pour l'étude de phénomènes d'échos, le cepstre permet l'observation des variations du spectre. Dans sa forme actuelle, le cepstre réel est généralement formulé avec une transformée de Fourier inverse :

$$C(\tau) = C(x(t)) = \left|\mathcal{F}^{-1}\left\{\ln\left(\left|\mathcal{F}\left\{x(t)\right\}\right|^2\right)\right\}\right|^2.$$

Ceci permet d'exprimer le cepstre dans le domaine temporel. Son usage est particulièrement répandu dans le domaine du traitement de la parole, car cette dernière peut être modélisée par un modèle source-filtre $x(n) = (s * h)(n)$. En effet le produit de convolution se traduisant par un produit simple dans le domaine fréquentiel, on montre que le cepstre complexe est un morphisme transformant l'opérateur de convolution en somme :

$$\begin{aligned} C(x) &= \mathcal{F}^{-1}\left\{\ln\left(\mathcal{F}\left\{s * h\right\}\right)\right\} && (6.1)\\ &= \mathcal{F}^{-1}\left\{\ln\left(\mathcal{F}\left\{s\right\}\right) + \ln\left(\mathcal{F}\left\{h\right\}\right)\right\} && (6.2)\\ &= C(s) + C(h). && (6.3) \end{aligned}$$

Ainsi, dans les cas proches de la voix, où la source $s(n)$ est un peigne fréquentiel, tandis que les résonances apportées par le filtre $h(n)$ ont un spectre beaucoup plus lisse, généralement assimilée à une enveloppe spectrale, les cepstres de la source et du filtre sont pratiquement disjoints et donc séparables. Ceci étant, beaucoup de sources sonores musicales ne rentrent pas dans le cadre du modèle source-filtre et ne sont pas aussi aisément interprétables par l'analyse cepstrale.

Nous exploitons pour la tâche de classification audio deux types de descripteurs basés sur la représentation cepstrale :

- Les *coefficients MFCC*, basés sur une échelle mel des fréquences, et où l'on substitue une DCT à la Transformée de Fourier.
- Les *coefficients cepstraux à Q constant*, basés sur une échelle liée à la répartition logarithmique des hauteurs musicales, qui implique en outre l'adaptation des largeurs de bandes par rapport à la fréquence centrale.

Descripteurs perceptifs

- Nous exploitons deux mesures liées au *pitch* (c'est-à-dire la perception de la fréquence fondamentale dominante), calculées par l'algorithme YIN de Cheveigné et al. [60] :
  - la *fréquence fondamentale* F_0,
  - et la *mesure de périodicité*, qui caractérise les spectres harmoniques.
- Enfin, trois descripteurs sont liés à la mesure d'*intensité perceptive* proposée par Moore et al. [162] et appelée *loudness* :
  - la *loudness spécifique relative*, constituée de coefficients d'intensité sur les bandes de fréquences perceptives,
  - l'*acuité perceptive* (ou *sharpness*), proposée par Peeters [178]), qui est l'équivalent du centroïde spectral sur la loudness,
  - et l'*étalement perceptif* (ou *spread*), également proposé par Peeters, défini comme une mesure de contraste sur la loudness.

6.6 Discussion

Les descripteurs présentés ont avant tout été choisis pour respecter deux contraintes essentielles liées à notre cadre particulier.

En premier lieu le système mis en place est conçu pour fonctionner en temps réel ; il ne doit pas impliquer de descripteurs coûteux. Ont donc été rejetés de nombreux descripteurs basés sur une analyse spectrale plus poussée par un modèle psychoacoustique de l'audition [194][157], impliquant une forte charge en temps de calcul ; ou d'autres qui permettent de mieux appréhender le matériau musical éventuel. Ainsi on trouve dans la littérature certaines formes de « transcriptions ébauchées » à travers la recherche de trajectoires de pics spectraux [121]. Ces approches ont par ailleurs généralement l'inconvénient de se baser sur des algorithmes de type Viterbi, impliquant les trames futures dans la prise de décision.

Cette contrainte temporelle est la seconde restriction que nous nous imposons. Une réponse en temps réel implique une décision régulière et instantanée, ou au moins avec retard constant, qui exclue donc l'usage de descripteurs temporellement alignés sur des événements particuliers, comme les attaques. On trouve par exemple dans la littérature plusieurs propositions impliquant une estimation du tempo ou de la structure rythmique [111][41], que nous avons exclues de notre cadre expérimental, puisqu'elles font intervenir une modalité temporelle beaucoup plus large (de

l'ordre de plusieurs secondes). Certains auteurs, comme Lachambre et al. [129], se basent également sur une décomposition temporelle du signal en segments « homogènes » de tailles variables servant chacun de support au calcul d'une valeur.

Beaucoup de ces propositions émanent de domaines annexes, plus directement liés à la musique, comme la reconnaissance de genres ou d'ambiances musicales. Nous supposons ici, sans toutefois l'étayer par un constat psychoacoustique, que les classes impliquées (parole, musique, chant) sont identifiables par un être humain de manière quasi instantanée et surtout hors de tout contexte. Les classes considérées sont plus clairement définies et peu ambigües [2], contrairement à d'autres domaines, comme la reconnaissance de genre musical, qui fait intervenir des notions sémantiques et cognitives.

La plupart des descripteurs réunis ici sont d'usage très courant dans la littérature. On notera en outre que les traitements statistiques décrits dans la section 6.3 font apparaître certaines propriétés qui peuvent sembler absentes de notre liste. Par exemple le taux de trames à basse énergie ou le taux de trames silencieuses, que l'on trouve assez fréquemment [207][138][109] dans la littérature, est équivalent à la moyenne de l'énergie des trames court-terme. De plus, certains descripteurs, difficilement interprétables tels quels, comme les moments temporels, prennent leur sens si l'on considère leur moyenne ou leur variance. Ainsi la variance de la *largeur temporelle* constitue une alternative à l'estimation du taux de trames à basse énergie.

En définitive on remarquera que de nombreux descripteurs, parmi ceux présentés, sont fortement redondants dans leur définition. On peut en effet grouper ceux-ci en 4 modalités principales :

- *Centre spectral* : estimation du centre de gravité du spectre. Par exemple le centroïde spectral, le ZCR, le ZCR long-terme, et l'acuité perceptive.
- *Répartition spectrale* : description de l'allure ou de la répartition spectrale. Par exemple la largeur, l'asymétrie, la platitude, le rapport, la pente, et la décroissance spectraux, la fréquence de coupure, la platitude d'amplitude spectrale, le facteur de crête spectrale, ou encore l'étalement perceptif.
- *Énergies de sous-bandes* : de nombreux descripteurs fournissent une estimation de l'énergie de sous-bandes spectrales, généralement grâce à l'usage de banc de filtres, ou à travers l'estimation d'enveloppes spectrales. On peut ainsi citer les descripteurs OBSI et OBSIR, les coefficients LPC, MFCC, à Q constant, ainsi que la *loudness* spécifique relative.
- *Estimation de pitch* : enfin, une série de descripteurs fournissent une estimation de la fréquence fondamentale dominante, comme le F_0 estimé par YIN, et les coefficients d'autocorrélation.

Les autres descripteurs expriment des caractéristiques plus particulières, comme la mesure d'apériodicité estimée par YIN, les mesures de trémolo et de rugosité par modulation d'amplitude, le flux spectral, décrivant les variations du spectre, ou encore les moments statistiques temporels.

Bien que ces descripteurs soient pertinents pour la discrimination parole/musique, on remarque que la littérature s'attarde très peu sur le problème des classes mixtes. En effet, l'observation comparatée de spectres de parole et de musique met en évidence de nombreuses différences assez claires que beaucoup d'auteurs ont tenté de traduire par des mesures quantifiées. Néanmoins, la plupart de ces observations perdent leur sens lorsque l'on superpose les signaux des deux classes. Or nous verrons par la suite, dans la partie IV sur l'évaluation, que la majeure partie des erreurs du système provient précisément de confusions sur cette situation (la parole accompagnée de musique). On peut bien sûr être tenté de faire appel à des techniques de séparation de sources mais celles-ci sont généralement beaucoup plus complexes et coûteuses que les méthodes impliquées dans la classification audio. De plus, de nombreux algorithmes de séparation de sources font généralement appel à des techniques de classification afin d'identifier les régions pertinentes à séparer, ce qui inverse totalement la méthodologie.

2. Certains problèmes demeurent par exemple sur la détection du chant, comme l'identification du rap, ou de certaines prosodies de chant proches de la parole, comme le Gainsbourg des années 80.

Chapitre 7

Sélection de descripteurs

Sommaire

7.1 Introduction

Nous avons introduit dans le chapitre 6 une large collection de descripteurs destinés à caractériser au mieux les classes mises en jeu pour la segmentation d'un flux audio. Nous avons fait le

constat que beaucoup d'entre eux décrivent des modalités très proches et peuvent donc présenter de fortes redondances. Diversifier les descripteurs reste bien sûr souhaitable, mais peut se révéler contre-productif lors de la phase de classification, en premier lieu à cause de la malédiction de la dimension, que nous avions évoquée dans la section 2.3.4 de l'état de l'art.

De plus, malgré le soin porté dans le choix des descripteurs mis en jeu, il est possible d'introduire une certaine proportion de descripteurs non pertinents qui bruitent l'information considérée par le classifieur. L'introduction d'une composante non corrélée à la tâche considérée peut en effet avoir un impact néfaste sur la mesure des distances entre exemples, qui constitue pourtant l'outil de base de la majorité des méthodes de classification, particulièrement des SVM.

Enfin, d'un point de vue pratique, l'utilisation de descripteurs non pertinents, ou au moins redondants, introduit une complexité inutile (tant calculatoire qu'en termes de mémoire) dans la phase de classification par le calcul coûteux et superflu de ces descripteurs.

La sélection des descripteurs les plus pertinents pour une tâche donnée est un problème à part entière dans le domaine de l'apprentissage statistique, qui a beaucoup occupé la communauté scientifique durant les dernières décennies. Outre la résolution des difficultés exposées précédemment, elle peut apporter une meilleure compréhension d'un problème par l'interprétation des descripteurs les plus pertinents.

On conserve dans ce chapitre les notations introduites dans la section 3.1 en se concentrant sur un problème de discrimination : on travaille donc sur un ensemble d'apprentissage $\mathcal{S} = \{(\boldsymbol{x}_i, y_i)\}_{i=1\ldots n}$, où les exemples $\boldsymbol{x}_i \in \mathbb{R}^D$ sont décrits par D composantes dimensionnelles correspondant chacune à un descripteur donné ($\boldsymbol{x}_i = [x_{i,1}\,, \ldots, x_{i,D}]^T$), et sont associés à un label $y_i \in \{+1, -1\}$. On utilisera également le vecteur des labels $\boldsymbol{y} = [y_1, \ldots, y_n]^T$ et les vecteurs des exemples pour chaque descripteur $\boldsymbol{x}_{\cdot,d} = [x_{1,d}\,, \ldots, x_{n,d}]^T$.

Nous abordons dans la section 7.2 la question de la définition de la pertinence d'un descripteur, qui met en évidence la complexité du problème posé. L'explosion combinatoire qui en découle est généralement contournée par le moyen de stratégies de recherche que nous énumérons par la suite. La relation avec le classifieur est également formalisée par une taxonomie classique qui fait apparaître l'importance de la prise en compte du comportement du classifieur dans le processus de sélection des descripteurs.

Nous détaillerons dans la section 7.3 quelques algorithmes classiques principalement basés sur une mesure de corrélation, pour nous concentrer ensuite, dans la section 7.4, sur des méthodes prenant en compte le comportement des machines à vecteurs de support dans la sélection. L'étude des algorithmes de la littérature nous permet finalement de proposer plusieurs algorithmes basés sur certains des critères de performances abordés dans le chapitre 4.

7.2 Taxonomie des algorithmes de sélection

7.2.1 Classement ou sélection

On peut définir le problème de la sélection de descripteurs comme la recherche d'un sous-groupe des S descripteurs les plus susceptibles d'optimiser la tâche de classification ultérieure, parmi une collection originale de D descripteurs. Plusieurs questions se posent alors :

- Souhaite-t-on déterminer le sous-groupe optimal pour un nombre $S < D$ donné ?
- Souhaite-t-on déterminer le sous-groupe optimal pour *tout* nombre $S < D$ possible ?
- Souhaite-t-on déterminer automatiquement le nombre S de descripteurs conjointement au sous-groupe ?

Ces trois problèmes sont en réalité tout à fait différents d'un point de vue méthodologique, et soulignent l'hétérogénéité de la sélection de descripteurs, qui recouvre en réalité deux problèmes différents :

- Le **classement** (*variable ranking*) vise à ranger les descripteurs par ordre croissant (ou décroissant) de pertinence pour la tâche donnée.
- La **sélection** de sous-ensemble (*subset selection*) vise à extraire de la liste originale un sous-ensemble de descripteurs pertinents, dont la taille est déterminée manuellement ou automatiquement.

Nous verrons que certains des algorithmes présentés sont fondamentalement liés à la notion de classement. D'autres sont souvent présentés dans la littérature sous la forme de problèmes de sélection de sous-ensemble, mais nous les présenterons systématiquement sous forme d'algorithmes de classement, de manière à pouvoir définir un protocole expérimental commun. Cette transformation implique en général d'ignorer les étapes de seuillage introduites par les auteurs.

7.2.2 Notion de pertinence

Les deux problèmes précédents ont pour trait commun de se baser sur la notion centrale de *pertinence*, qui peut être considérée comme une mesure de l'efficacité d'un descripteur pour une tâche donnée. Pourtant, même si elle semble intuitive, celle-ci est difficile à définir analytiquement. De nombreuses propositions ont été formulées dans la littérature, synthétisées par John et al. [117]. Les auteurs montrent que pour un simple problème de OU exclusif (XOR) sur des données corrélées, où l'on définit 5 variables booléennes x_1, x_2, x_3, x_4 et x_5, avec deux couples inversement correllés ($x_4 = \bar{x}_2$ et $x_5 = \bar{x}_3$), associées à un label $y = x_1 \otimes x_2$ (où $\otimes$ représente le OU exclusif), toutes les définitions considérées sont fausses. Ils en déduisent une distinction formelle entre *pertinence forte*, relative à un descripteur indispensable à la tâche (c'est-à-dire dont l'exclusion pénalise celle-ci), et *pertinence faible*, relative à un descripteur utile à la tâche, mais auquel peut cependant se substituer un autre descripteur. La notion de pertinence faible fait apparaître le fait que la pertinence d'un descripteur n'est pas une propriété intrinsèque et ne peut être jugée indépendamment des autres descripteurs mis en jeu.

Guyon et al. [97] montrent par ailleurs, par le biais d'un exemple concret également basé sur le problème du OU exclusif, que deux descripteurs strictement non- pertinents lorsqu'ils sont utilisés chacun seul, peuvent se révéler pertinents lorsqu'ils sont exploités ensemble (dans des cas de séparation non-linéaire). On retrouve ce constat dans une note de Toussaint [226] qui montre que les k pires descripteurs individuels peuvent se révéler meilleurs ensemble que les k meilleurs descripteurs individuels ; nous appellerons par la suite ce phénomène *interpertinence*.

Guyon et al. se penchent en outre sur une autre idée reçue qui consiste à considérer que deux descripteurs corrélés n'apportent pas d'information supplémentaire, en montrant également par un exemple simple que ce raisonnement n'est pas systématiquement vrai.

Nous ne cherchons pas ici à traiter en détail la question de la définition formelle de la pertinence, sur laquelle on pourra trouver plus d'informations dans [160] et [31]. Toutefois, ces considérations montrent les difficultés sous-jacentes à la sélection de descripteurs qui, par une approche « force brute », se heurte au classique problème d'explosion combinatoire : il n'est en général pas possible d'évaluer les $\binom{D}{S} = \frac{D!}{S!(D-S)!}$ possibilités pour sélectionner le sous-ensemble optimal de S descripteurs parmi les D. Ce constat s'aggrave si l'on considère toutes les valeurs S possibles. Il est donc nécessaire d'adopter une stratégie de recherche.

7.2.3 Stratégies de recherches

Webb propose [241] une taxonomie des différentes stratégies généralement employées pour la sélection de descripteurs, reprise par Wang et Chen [239] :

BIN (*Best Individual N*) : C'est la stratégie la plus simple. L'efficacité de chaque descripteur est mesurée indépendamment par un critère donné. La complexité est donc réduite (de l'ordre $O(D)$) et peut par ailleurs largement profiter d'un traitement en parallèle. Les descripteurs sont rangés par ordre décroissant de pertinence, d'après la mesure effectuée. Bien sûr la contrepartie à ce moindre coût computationel est l'absence de prise en compte des dépendances éventuelles entre les descripteurs. On risque ainsi de retrouver de nombreux descripteurs redondants dans les mieux notés, et de passer à côté des phénomènes d'interpertinence présentés plus haut.

SEQ (*SEQuential*) : Afin de prendre en compte les interdépendances entre descripteurs on peut adopter une stratégie séquentielle, qui a pour principe de sélectionner itérativement les descripteurs ou des groupes de descripteurs. On commence ainsi par sélectionner le descripteur le plus pertinent. Par la suite, à chaque itération la sélection prendra en compte (d'une manière non précisée) la liste

des descripteurs déjà choisis pour mesurer la pertinence des descripteurs restants. On évite ainsi la sélection de multiples descripteurs redondants, au prix néanmoins d'un fort accroissement de complexité, de l'ordre $O\left(\frac{D^2}{2}\right)$, puisque l'on doit évaluer à chaque itération tous les descripteurs restants, ou $O\left(SD - \frac{S^2}{2}\right)$ si l'on s'arrête à S descripteurs. On notera toutefois que dans le cas extrême (et bien sûr théorique) du OU exclusif, le problème des descripteurs interpertinents n'est pas résolu par cette stratégie.

À l'approche *forward* décrite ici, on peut substituer une approche *backward*, où les descripteurs les moins pertinents sont itérativement supprimés. Bien que l'approche *backward* soit beaucoup plus coûteuse si $S \ll D$ (on retrouve alors l'ordre $O\left(\frac{D^2}{2}\right)$), Guyon et al. [97] expliquent que cette dernière est moins susceptible d'être biaisée puisque dès la première itération l'effet conjoint de tous les descripteurs est pris en compte, ce qui n'est pas le cas de l'approche *forward*.

PO (*Parameter Optimization*) : La troisième stratégie repose sur une procédure d'optimisation. En pondérant chaque composante (chaque descripteur d) d'un exemple $\boldsymbol{x}$ par un facteur w_d (soit $\boldsymbol{x_w} = \boldsymbol{x} \bullet \boldsymbol{w}$, où $\bullet$ est le produit terme à terme), on minimise un critère donné par mises à jours successives des poids w_d, jusqu'à convergence (nous présenterons dans la suite plusieurs exemples associés à des critères différents). On considère alors les poids w_d comme une mesure de pertinence des descripteurs qui sont ordonnés par ordre décroissant. Cette approche présente l'avantage de faire intervenir la totalité des descripteurs simultanément à chaque itération, ce qui permet de mieux prendre en compte les problèmes d'interdépendances et de redondances. La complexité d'ordre $O(ID)$ est beaucoup plus réduite que pour l'approche séquentielle, si l'on suppose que le nombre d'itérations I est négligeable devant le nombre de descripteurs D. Néanmoins, cette approche suppose l'usage d'un critère dérivable par rapport aux poids w_d. De plus il n'existe pas de preuve théorique que ces poids soient une mesure fiable de pertinence ; en particulier on n'a aucune garantie de la consistance des poids si l'on ôte une grande partie des descripteurs (ce qui est a priori le but souhaité). Certains auteurs se contentent de supprimer les descripteurs dont les poids tendent vers 0, mais cette approche a tendance à supprimer plus de descripteurs qu'il ne faudrait.

7.2.4 Paradigmes

Nous avons jusque-là considéré la sélection de descripteurs en dehors de tout contexte. Pourtant celle-ci précède et détermine l'utilisation d'une méthode de classification, puisqu'elle est dans l'idéal menée pour optimiser cette dernière. On retrouve généralement dans la littérature [97][117][31] une taxonomie distinguant trois paradigmes de sélection de descripteurs, déterminés par la relation avec le classifieur :

Filtres

Les filtres (*filters*) sont considérés comme indépendant du processus de classification. La sélection de descripteurs peut être vue comme une étape de pré-traitement des données qui ne fait pas appel au classifieur. Elle est donc « universelle » de ce point de vue et présente a priori le désavantage de ne pouvoir prendre en compte le biais éventuel introduit par tel ou tel classifieur.

Enveloppeurs

La distinction entre filtres et enveloppeurs (*wrappers*) est introduite par John et al. [117] pour prendre en compte la notion de pertinente faible qu'ils introduisent. Le principe des enveloppeurs est d'inclure le classifieur, comme une boîte noire, dans le processus de sélection. Celle-ci s'opère donc par itérations successives, tirant parti des résultats de classification et prenant ainsi en compte le comportement propre du classifieur. Cependant cette approche est généralement beaucoup plus coûteuse que les filtres, puisqu'elle implique de nombreuses phases d'apprentissage, en plus des calculs directement liés à la sélection de descripteurs.

Sélection embarquée

Les algorithmes de sélection embarquée (*embedded methods*) concernent une classe de classifieurs pour lesquels la sélection de descripteurs fait directement partie du processus de classification. L'exemple des arbres de décision de type CART (*Classification And Regression Trees*) est le plus typique de cette approche. Les sélections embarquées ont le mérite, par rapport aux enveloppeurs qui traitent le classifieur comme une boîte noire, de ne pas nécessiter de corpus de validation distinct du corpus d'apprentissage pour valider les performances du classifieur.

Il est communément admis que les approches filtres sont indépendantes du classifieur mis en jeu. Pourtant, tout critère de mesure de pertinence impliqué dans la sélection induit en réalité une hypothèse sur le processus de classification. Nous présenterons dans la suite de ce chapitre de nouvelles méthodes de sélection de descripteurs, adaptées aux SVM, qui n'incluent aucun apprentissage du classifieur tout en basant la sélection sur les critères de performances introduits dans le chapitre 4, directement liés aux SVM.

7.3 Méthodes filtres classiques

De nombreuses méthodes de sélection de descripteurs ont été proposées sur la base d'une mesure de corrélation entre le descripteur et le label associé à l'exemple. Elles reposent en général sur une approche *filtre* de classement de pertinence. Nous détaillons quelques-unes de ces méthodes dans cette section.

7.3.1 Coefficients de Pearson et critère de Fisher

On exploite en statistiques le coefficient de corrélation de Pearson entre deux variables aléatoires X et Y :

$$\mathcal{R} = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\,\mathrm{var}(Y)}},$$

où cov et var sont respectivement les mesures de covariance et de variance. On peut ainsi estimer empiriquement la corrélation entre le vecteur des labels $\boldsymbol{y}$ et le vecteur $\boldsymbol{x}^d$ des exemples pour le descripteur d :

$$\mathcal{R}(d) = \frac{\sum_{i=1}^{n}(x_{i,d}-\bar{x}_d)(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_{i,d}-\bar{x}_d)^2\sum_{i=1}^{n}(y_i-\bar{y})^2}},$$

où $\bar{x}_d = \frac{1}{n}\sum_{i=1}^{n} x_{i,d}$ est le centre des exemples du descripteur d et $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ le centre des labels.

Dans le cas du problème de discrimination ($y_i \in \{+1;-1\}$), si l'on suppose les deux classes également réparties, on a $\bar{y}=0$, et $\sum_{i=1}^{n}(y_i-\bar{y})^2 = n$, on peut alors montrer que, à un facteur $\sqrt{2}$ près, les coefficients ont pour valeur :

$$R_{corr}(d) = \frac{\mu_{1,d}-\mu_{2,d}}{\sqrt{\sigma_{1,d}^2+\sigma_{2,d}^2}},$$

où $\mu_{c,d}$ et $\sigma_{c,d}^2$ sont respectivement le centre et la variance pour le descripteur d des exemples de la classe c. Cette expression est généralement employée pour définir les coefficients de corrélation de Pearson exploités pour ranger les descripteurs par ordre de pertinence.

On remarquera néanmoins que la corrélation telle quelle est a priori mal définie puisqu'un descripteur inversement corrélé au label se voit attribuer le pire score de pertinence alors qu'il apporte autant d'information que son opposé. On préfère donc généralement employer le carré des coefficients de Pearson, que Bishop [29] nomme critère de Fisher :

$$R_{Fisher}(d) = \frac{|\mu_{1,d}-\mu_{2,d}|^2}{\sigma_{1,d}^2+\sigma_{2,d}^2}.$$

En effet la maximisation de ce critère dans sa forme générale ($R = \frac{|\mu_1 - \mu_2|^2}{\sigma_1^2 + \sigma_2^2}$) est directement liée à l'Analyse Discriminante de Fisher introduite dans le prélude aux Machines à Vecteurs de Support, en section 3.2. On exploite donc ce dernier pour quantifier la séparabilité des classes dans l'espace unidimensionnel défini par chaque descripteur.
On a donc introduit ici une approche filtre couplée à une stratégie de recherche BIN.

7.3.2 Inertia Ratio Maximization using Feature Space Projection (IRMFSP)

Peeters propose [181] une méthode séquentielle de classement de descripteurs, dont le principe est proche du critère de Fisher. Il y ajoute une phase qui, à chaque itération, permet de prendre en compte les descripteurs déjà sélectionnés. Il se base en premier lieu sur le critère de séparabilité que nous avons introduit dans la section 4.4.2. On rappelle que ce dernier est le rapport des *dispersions inter-classes* et *intra-classes* (équations 4.14 et 4.15, page 58) qui, sur des données uni-dimensionnelles, prennent les expressions suivantes :

$$S_b = \frac{1}{n} \sum_{c=1,2} n_c \left(\mu_{c,d} - \mu_d\right)^2 \tag{7.1}$$

$$S_w = \sum_{c=1,2} \sum_{\boldsymbol{x}_i \in \mathcal{S}_c} \left(x_{i,d} - \mu_{c,d}\right)^2 . \tag{7.2}$$

On peut en outre introduire la mesure de *dispersion globale* (ou totale), qui mesure la dispersion de tous les exemples par rapport au centre global μ_d :

$$S_T = \frac{1}{n} \sum_{i=1}^{n} \left(x_{i,d} - \mu_d\right)^2 . \tag{7.3}$$

Peeters propose donc, comme critère de pertinence pour les descripteurs, le rapport entre les dispersions inter-classes et globale (qu'il nomme *inerties*, d'où le terme *Inertia Ratio*) :

$$R_{IR}(d) = \frac{\sum_{c=1,2} n_c \left(\mu_{c,d} - \mu_d\right)^2}{\sum_{i=1}^{n} \left(x_{i,d} - \mu_d\right)^2}.$$

L'apport principal de sa contribution consiste à choisir le descripteur maximisant le critère à chaque itération, puis d'appliquer une procédure d'orthogonalisation par rapport aux descripteurs sélectionnés sur les descripteurs restants afin de réduire la corrélation entre ces derniers et le descripteur sélectionné.

Ainsi, si l'on suppose qu'à l'itération k, le descripteur d'indice d_k maximise le critère R_{IR}, alors pour tout descripteur restant d'indice e, on applique la mise à jour suivante par rapport au vecteur normalisé $\tilde{\boldsymbol{x}}_{\cdot,d_k} = \frac{\boldsymbol{x}_{\cdot,d_k}}{\|\boldsymbol{x}_{\cdot,d_k}\|}$:

$$\boldsymbol{x}_{\cdot,e} \longleftarrow \boldsymbol{x}_{\cdot,e} - \left(\boldsymbol{x}_{\cdot,e}^T \tilde{\boldsymbol{x}}_{\cdot,d_k}\right) \tilde{\boldsymbol{x}}_{\cdot,d_k}.$$

La méthode IRMFSP apporte ainsi une stratégie de recherche séquentielle sur un critère proche du critère de Fisher. Celle-ci reste très peu coûteuse de par la simplicité du critère exploité. Cependant, la phase d'orthogonalisation introduite nécessite un certain nombre d'exemples pour être statistiquement fiable. De plus cette fiabilité décroît fortement à mesure que les effets des orthogonalisations successives se cumulent. On peut donc finir par travailler sur des descripteurs totalement bruités après un certain nombre d'itérations.

7.3.3 Test de Kolmogorov-Smirnov

Il est également possible de construire une mesure de pertinence sur la base du test de Kolmogorov-Smirnov. Ce dernier est un test d'hypothèse utilisé en statistiques pour déterminer si un échantillon suit une loi définie par sa fonction de répartition $F_X(x) = \mathcal{P}(X \leq x)$, où X est une variable aléatoire modélisant ici le comportement d'un descripteur donné. Il est défini comme le maximum de la différence absolue entre deux fonctions de répartitions, l'une supposant la classe positive, l'autre non ; soit :

$$KS(d) = \sqrt{n} \max_{1 \leq i \leq n} \left|F_{X_d}(x_{i,d}) - F_{X_d}(x_{i,d} | \, y = +1)\right| .$$

On estime la fonction de répartition F_{X_d} à partir des échantillons de $\boldsymbol{x}_{\cdot,d}$ par :

$$F_{X_d}(x) = \frac{1}{n} \operatorname{Card} \{x_{i,d} \mid x_{i,d} < x\}_{1 \leq i \leq n} .$$

Le critère ainsi défini ne fait donc aucune supposition sur la répartition des observations (contrairement à la modélisation gaussienne implicite dans les deux approches précédentes), et permet ainsi la constitution d'une approche filtre avec stratégie BIN pour la sélection de descripteurs. On notera toutefois qu'un algorithme a été proposé [28] pour adapter ce critère à une stratégie séquentielle et ainsi prendre en compte les redondances entre descripteurs.

Il existe de nombreuses autres approches de type filtres dans la littérature, par exemple basées sur l'Information Mutuelle [253][255], que nous ne détaillerons pas ici. Notre propos est avant tout de montrer la pertinence des méthodes prenant en compte la classification par SVM, par contraste avec les méthodes filtres classiques uniquement basées sur des mesures de corrélation ou d'information entre les descripteurs ou entre les descripteurs et les labels de classe.

7.4 Méthodes à noyaux

7.4.1 Feature Selection concaVe (FSV)

Parallèlement au développement des Machines à Vecteurs de Support, Mangasarian a dirigé de nombreux travaux sur la séparation linéaire par programmation linéaire [147]. Il formule ainsi l'algorithme de Programmation Linéaire Robuste [23] (RLP *Robust Linear Programming*) pour la détermination d'un hyperplan linéaire optimal, même dans le cas de données non-séparables linéairement :

$$\begin{array}{ll} \min\limits_{\boldsymbol{w},\gamma,\boldsymbol{y},\boldsymbol{z}} & \dfrac{\mathbf{1}^T\boldsymbol{y}}{n_1} + \dfrac{\mathbf{1}^T\boldsymbol{z}}{n_2} \\ & \\ \text{sous les contraintes} & \begin{array}{c} -\boldsymbol{A}\boldsymbol{w} + \gamma\mathbf{1} + \mathbf{1} \leq \boldsymbol{y} \\ \boldsymbol{B}\boldsymbol{w} - \gamma\mathbf{1} + \mathbf{1} \leq \boldsymbol{z} \\ \boldsymbol{y} \geq \mathbf{0}, \boldsymbol{z} \geq \mathbf{0}, \end{array} \end{array} \tag{7.4}$$

où $\mathbf{1}$ est un vecteur dont les composantes sont égales à 1 (et permet d'exprimer la norme $L1$ $\|\boldsymbol{x}\|_1 = \sum x_i = \mathbf{1}^T\boldsymbol{x}$), n_c le nombre d'exemples de la classe c, et $\boldsymbol{A} \in \mathbb{R}^{n_1 \times D}$ et $\boldsymbol{B} \in \mathbb{R}^{n_2 \times D}$ les matrices dont les lignes contiennent respectivement les exemples des classes 1 et 2, et $\boldsymbol{w}$ le vecteur normal de l'hyperplan de séparation.

Mangasarian et Bradley proposent par la suite [36][35] de résoudre de manière conjointe le problème de la sélection de descripteurs par l'ajout d'un terme contraignant la minimisation du nombre de composantes non-nulles du vecteur $\boldsymbol{w}$. L'expression à minimiser devient donc, sous les mêmes contraintes :

$$\min_{\boldsymbol{w},\gamma,\boldsymbol{y},\boldsymbol{z}} \quad (1-\lambda)\left(\frac{\mathbf{1}^T\boldsymbol{y}}{n_1} + \frac{\mathbf{1}^T\boldsymbol{z}}{n_2}\right) + \lambda \|\boldsymbol{w}\|_0 , \tag{7.5}$$

où $\|\cdot\|_0$ est la « *norme zéro* » [1] et est égale au nombre de composantes non nulles. Le paramètre additionnel λ fixe le compromis entre les deux termes à minimiser.

Toutefois la dite norme zéro pose problème parce qu'elle n'est ni continue, ni dérivable. Les auteurs résolvent ce problème en approchant cette dernière par une exponentielle inverse :

$$\|\boldsymbol{w}\|_0 = \mathbf{1}^T\left(\mathbf{1} - e^{-\alpha\boldsymbol{v}}\right),$$

où $\boldsymbol{v}$ est un vecteur aux termes positifs bornant à l'excès les composantes de $\boldsymbol{w}$ (il définit donc une nouvelle contrainte $-\boldsymbol{v} \leq \boldsymbol{w} \leq \boldsymbol{v}$). L'exponentielle inverse est choisie par les auteurs pour sa simplicité et sa concavité qui garantit de bonnes propriétés de convergence.

L'algorithme FSV (*Feature Selection concaVe*) [35] ainsi défini est un exemple de méthode *embarquée* impliquant dans un même processus d'optimisation l'apprentissage du classifieur et

1. l'appellation courante de norme est ici abusive puisqu'elle ne vérifie pas la propriété d'homogénéité, à savoir $\|\lambda\boldsymbol{x}\| = |\lambda| \cdot \|\boldsymbol{x}\|$.

la sélection de descripteurs. Les auteurs font par ailleurs apparaître un problème d'optimisation de SVM en remplaçant la norme zéro par une classique norme $L2$ (terme additionnel $\frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w}$). On retrouve alors le principe de maximisation de la marge, mais l'algorithme se révèle incapable d'annuler des composantes du vecteur $\boldsymbol{w}$ et n'opère donc plus de sélection.

7.4.2 Approximation of the zeRO-norm Minimization (AROM)

L'utilisation des composantes du vecteur normal $\boldsymbol{w}$ pour le classement de descripteurs a été largement exploitée dans d'autres publications sur la sélection de descripteurs liée aux SVM. Comme nous l'avons expliqué, l'optimisation de Machines à Vecteurs de Support consiste en la détermination d'un vecteur $\boldsymbol{w}$ optimal, exprimé à partir des exemples $(\boldsymbol{x}_i, y_i)$ et des multiplicateurs de Lagrange α_i :

$$\boldsymbol{w} = \sum_i \alpha_i y_i \Phi(\boldsymbol{x}_i). \tag{7.6}$$

Le problème majeur lié à l'usage de $\boldsymbol{w}$ provient du fait que la fonction Φ n'est pas explicite pour la plupart des noyaux, on ne peut donc exprimer numériquement les composantes du vecteur. Plusieurs propositions contournent cette difficulté en se restreignant au noyau linéaire, parmi lesquelles la méthode d'Approximation de Minimisation de la norme zéro (AROM *Approximation of the zeRO-norm Minimization*) [244].

Se concentrant sur le problème de la minimisation de la norme zéro, les auteurs font le constat que l'approche de Bradley et Mangasarian souffre d'une grande complexité lorsque le nombre de descripteurs est élevé, et laisse de plus ouverte la question de la détermination du paramètre α. Ils proposent ainsi de substituer à ce problème la minimisation de la grandeur suivante :

$$\min_{\boldsymbol{w}} \quad \sum_{d=1}^{D} \ln |w_d|,$$

et montrent que le minimum atteint est pratiquement égal au minimum de $\|\boldsymbol{w}\|_0$. On peut constater intuitivement que la fonction ln favorise les composantes proches de zéro, forçant ainsi la minimisation du nombre de composantes non nulles. En appliquant la méthode de Franke et Wolke de descente de gradient, ils montrent que le problème converge vers un minimum local. Celui-ci est atteint par mises à jour successives du vecteur $\boldsymbol{w}$ (initialisé par exemple à $\boldsymbol{w} = \mathbf{1}$) en résolvant le problème suivant :

$$\begin{array}{ll} \min\limits_{\hat{\boldsymbol{w}}} & \sum_{d=1}^{D} |\hat{w}_d| \\ \text{sous les contraintes} & y_i \left(\hat{\boldsymbol{w}}^T (\boldsymbol{x}_i \bullet \boldsymbol{w}) + b\right) \geq 1, \end{array} \tag{7.7}$$

où $\bullet$ est le produit terme à terme (dit de Hadamard). La mise à jour du vecteur $\boldsymbol{w}$ se fait simplement en multipliant à chaque itération les composantes terme à terme avec le vecteur $\hat{\boldsymbol{w}}$ évalué : $\boldsymbol{w} \leftarrow \boldsymbol{w} \bullet \hat{\boldsymbol{w}}$, jusqu'à la convergence de $\boldsymbol{w}$. Nous désignerons par la suite cette méthode par AROM $L1$, du fait de l'usage de la norme correspondante.

Les auteurs proposent ensuite une approximation très efficace de l'algorithme en substituant à $L1$ la norme $L2$, ce qui mène à la formulation duale (on peut également exploiter la formulation primale) d'un problème d'apprentissage de SVM :

$$\begin{array}{ll} \min\limits_{\alpha_i} & \sum_{i=}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \left(\boldsymbol{w} \bullet \boldsymbol{x}_i\right)^T \left(\boldsymbol{w} \bullet \boldsymbol{x}_j\right) \\ \text{sous les contraintes} & \sum_{i=1}^{n} \alpha_i y_i = 0 \\ & \alpha_i \geq 0 \end{array} \tag{7.8}$$

La mise à jour du vecteur $\boldsymbol{w}$ se fait de la même manière à partir du vecteur $\hat{\boldsymbol{w}}$, que l'on déduit des α_i estimés par la relation 7.6. Cette méthode, nommée AROM $L2$, est très avantageuse par rapport à la précédente puisqu'elle bénéficie de toutes les techniques de décomposition proposées dans la littérature pour l'apprentissage des SVM.

Les formulations 7.7 et 7.8 permettent en outre d'interpréter la méthode AROM comme une mise à jour de coefficients de pondération (les composantes w_d) sur les descripteurs. Ainsi un descripteur qui intervient peu dans le processus de classification sera pénalisé jusqu'à ne plus être pris en compte par ce dernier (lorsque le poids w_d est nul). La méthode AROM peut donc être considérée comme une méthode de sélection embarquée avec stratégie de recherche par optimisation de paramètres (PO).

7.4.3 Recursive Feature Extraction (RFE)

Par un développement en série de Taylor de la fonction objectif des SVM ($J = \frac{1}{2}\|\boldsymbol{w}\|^2$) au voisinage de son optimum, Guyon et al. [99] montrent que l'estimation de la dérivée $\frac{\partial J}{\partial w_d}$ justifie l'usage de la grandeur $R(d) = w_d^2$ comme critère de classement des descripteurs.

Néanmoins, se basant sur les constats énoncés en section 7.2.2, ils estiment qu'une simple stratégie BIN basée sur un tel classement peut se révéler largement sous-optimale, puisqu'elle équivaut à exclure un grand nombre de descripteurs en même temps, ce qui biaise la pertinence du critère énoncé. Il proposent donc une méthode à stratégie séquentielle *backward* (SEQ) d'éliminations successives des descripteurs.

L'algorithme RFE (*Recursive Feature Extraction*) consiste donc à éliminer à chaque itération le descripteur minimisant le critère $R(d) = w_d^2$ après apprentissage d'un SVM sur les descripteurs restants. La structure itérative permet de mettre à jour le classement des critères de pertinence après chaque élimination de descripteur.

Bien qu'efficace, la méthode se révèle ainsi beaucoup plus coûteuse que la méthode AROM, sans pourtant que ce coût soit réellement justifié par un écart notable de performances. Les auteurs proposent notamment d'éliminer plusieurs descripteurs à chaque itération pour en réduire la complexité, mais sans apporter de réponse théorique au nombre optimal de descripteurs à éliminer.

7.4.4 Sélection par minimisation de la borne Rayon-Marge (R2W2)

Il est également possible d'appliquer une stratégie PO (beaucoup plus économique que la stratégie SEQ déployée dans l'algorithme RFE) basé sur les facteurs de pondération introduits dans l'algorithme AROM, sans avoir à évaluer le vecteur normal de l'hyperplan, que nous noterons $\boldsymbol{w}_h$ dans cette section. Weston et al. ont proposé [47][245] un algorithme de sélection de descripteurs basé sur les critères d'évaluation de SVM introduits dans les sections 4.3.6 et 4.3.7. Nous nous concentrons ici sur la borne Rayon-Marge, la borne sur l'étendue étant beaucoup trop coûteuse quoique plus resserrée. La borne Rayon-Marge, dont on rappelle l'expression :

$$\mathcal{P}_{RM} = \frac{1}{n}\frac{R^2}{M^2} = \frac{1}{n}R^2\|\boldsymbol{w}_h\|^2,$$

a pour avantage de ne faire intervenir que la norme quadratique du vecteur $\boldsymbol{w}_h$, qui ne nécessite pas de connaître la fonction Φ puisque $\|\boldsymbol{w}_h\|_2^2 = k(\boldsymbol{w}_h, \boldsymbol{w}_h)$, où k est le noyau impliqué dans le classifieur SVM.

Les auteurs utilisent, de manière similaire à la méthode AROM, un vecteur de pondération $\boldsymbol{w}$ (qui ici n'est pas lié au vecteur normal $\boldsymbol{w}_h$) dont on peut résumer l'effet en introduisant le noyau pondéré $k_{\boldsymbol{w}}$:

$$k_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{y}) = k(\boldsymbol{w} \bullet \boldsymbol{x}, \boldsymbol{w} \bullet \boldsymbol{y}).$$

On peut ainsi exprimer la dérivée de la borne Rayon-Marge par rapport aux composantes w_d du vecteur $\boldsymbol{w}$:

$$\frac{\partial R^2\|\boldsymbol{w}_h\|^2}{\partial w_d} = R^2\frac{\partial\|\boldsymbol{w}_h\|^2}{\partial w_d} + \|\boldsymbol{w}_h\|^2\frac{\partial R^2}{\partial w_d},$$

avec :

$$\begin{aligned}\frac{\partial\|\boldsymbol{w}_h\|^2}{\partial w_d} &= -\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j\frac{\partial k_{\boldsymbol{w}}(\boldsymbol{x}_i, \boldsymbol{x}_j)}{\partial w_d}\\ \frac{\partial R^2}{\partial w_d} &= \sum_{i=1}^{n}\beta_i\frac{\partial k_{\boldsymbol{w}}(\boldsymbol{x}_i, \boldsymbol{x}_i)}{\partial w_d} - \sum_{i,j=1}^{n}\beta_i\beta_j\frac{\partial k_{\boldsymbol{w}}(\boldsymbol{x}_i, \boldsymbol{x}_j)}{\partial w_d},\end{aligned}$$

où l'on utilise l'expression du rayon R introduite dans l'équation 4.3. La dérivée $\frac{\partial k_{\boldsymbol{w}}(\boldsymbol{x}_i, \boldsymbol{x}_j)}{\partial w_d}$ est généralement évidente pour les noyaux usuels et ne pose donc pas de problème.

La sélection se fait donc par minimisation du critère Borne-Marge, en suivant une descente de gradient sur les composantes w_d. Les auteurs accélèrent l'algorithme en fixant à chaque itération les pires poids à zéro, arrêtant celui-ci lorsque seuls S poids non-nuls subsistent. On peut cependant

déduire de l'algorithme un classement de tous les descripteurs, sans sélection de sous-groupe, après convergence de la borne Rayon-Marge. Nous désignerons par la suite cette méthode par l'acronyme R2W2.

Comme nous le verrons dans la section expérimentale 7.7, la méthode R2W2 forme un très bon compromis entre performances et complexité. En effet, la stratégie de recherche PO garantie une complexité proportionnelle à la dimension, et se révèle donc, si l'algorithme converge en peu d'itérations, beaucoup plus rapide que la méthode RFE. De plus, contrairement aux autres méthodes liées aux SVM que nous avons présentées, celle-ci peut prendre en compte tout type de noyaux (pour peu que ce dernier soit dérivable par rapport aux w_d, ce qui est le cas de tous les noyaux usuels), et donc adapter la sélection de descripteurs à des cas non séparables linéairement.

La méthode est de type enveloppeur parce qu'elle implique l'apprentissage de SVM dans le processus de sélection (nécessaire pour évaluer le rayon R et le vecteur $\boldsymbol{w}_h$). Nous proposons dans la suite de ce chapitre plusieurs nouvelles méthodes de type filtre dont le principe est proche de R2W2. Nous verrons qu'elle réduisent la complexité en se passant de l'apprentissage de SVM.

7.5 Propositions d'algorithmes efficaces de sélection

7.5.1 Sélection pondérée basée sur le critère d'Alignement (SAS)

Toutes les méthodes de sélection présentées dans la section précédente impliquent l'apprentissage d'un SVM afin d'évaluerle critère utilisé. Nous avons présenté dans la section 4.4.1 le critère d'Alignement (ou KTA) qui permet d'évaluer les performances d'un noyau pour une tâche de classification donnée. Nous proposons ici un algorithme de sélection de descripteurs, basé sur la maximisation de l'Alignement par la mise à jour des composantes du vecteur de pondération $\boldsymbol{w}$. On rappelle que le KTA a pour expression :

$$\mathcal{A}(\boldsymbol{K}, \boldsymbol{K^*}) = \frac{\langle \boldsymbol{K}, \boldsymbol{K^*} \rangle_F}{\|\boldsymbol{K^*}\|_F \|\boldsymbol{K}\|_F}, \tag{7.9}$$

où $\boldsymbol{K}^* = \boldsymbol{y}\boldsymbol{y}^T$ est la matrice cible, et $\langle \cdot, \cdot \rangle_F$ le produit de Frobenius défini par $\langle \boldsymbol{A}, \boldsymbol{B} \rangle_F = \sum_{i,j} a_{ij} b_{ij}$. On définit ici la matrice de Gram pondérée $\boldsymbol{K}_{\boldsymbol{w}}$ du noyau pondéré $k_{\boldsymbol{w}}$. La dérivée de l'Alignement, pour la matrice de Gram pondérée, par rapport à la composante w_d du vecteur $\boldsymbol{w}$, a pour expression (équation 4.12, appliquée sur $\boldsymbol{K}_{\boldsymbol{w}}$) :

$$\frac{\partial}{\partial w_d} \mathcal{A}(\boldsymbol{K}_{\boldsymbol{w}}, \boldsymbol{K^*}) = \frac{\langle \partial_{w_d} \boldsymbol{K}_{\boldsymbol{w}}, \boldsymbol{K^*} \rangle_F}{\|\boldsymbol{K}_{\boldsymbol{w}}\|_F \|\boldsymbol{K^*}\|_F} - \frac{\langle \boldsymbol{K}_{\boldsymbol{w}}, \boldsymbol{K^*} \rangle_F \langle \boldsymbol{K}_{\boldsymbol{w}}, \partial_{w_d} \boldsymbol{K}_{\boldsymbol{w}} \rangle_F}{\|\boldsymbol{K}_{\boldsymbol{w}}\|_F^3 \|\boldsymbol{K^*}\|_F}. \tag{7.10}$$

Seule la matrice $\partial_{w_d} \boldsymbol{K}_{\boldsymbol{w}} = [\partial_{w_d} k_{\boldsymbol{w}}(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{ij}$ nous fait défaut pour évaluer la dérivée de l'Alignement. Nous la noterons par la suite $\partial_d \boldsymbol{K}_{\boldsymbol{w}}$ pour simplifier les notations, de même pour la dérivée $\partial_d k_{\boldsymbol{w}}$ du noyau pondéré par rapport à la composante w_d.

Tout comme pour l'algorithme R2W2, on doit évaluer la dérivée du noyau par rapport au poids w_d. L'introduction d'une décomposition des noyaux en *noyaux dimensionnels* κ^d permet de simplifier l'expression des dérivées. On détaille ici les décompositions naturelles pour les noyaux usuels :

- **Linéaire :**
  $k_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{w} \bullet \boldsymbol{x})^T (\boldsymbol{w} \bullet \boldsymbol{y}) = \sum_d w_d^2 \kappa^d(\boldsymbol{x}, \boldsymbol{y})$
  $\kappa^d(\boldsymbol{x}, \boldsymbol{y}) = x_d \cdot y_d$
  $\partial_d k_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{y}) = 2\, w_d\, \kappa^d(\boldsymbol{x}, \boldsymbol{y})$

- **RBF Gaussien :**
  $k_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{||\boldsymbol{w} \bullet (\boldsymbol{x} - \boldsymbol{y})||^2}{2\sigma^2}\right) = \exp\left(-\sum_d w_d^2 \kappa^d(\boldsymbol{x}, \boldsymbol{y})\right)$
  $\kappa^d(\boldsymbol{x}, \boldsymbol{y}) = \frac{(x_d - y_d)^2}{2\sigma^2}$
  $\partial_d k_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{y}) = -2\, w_d\, \kappa^d(\boldsymbol{x}, \boldsymbol{y})\, k_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{y})$

- **Polynomial :**
  $k_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{y}) = \chi_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{y})^\delta$
  $\chi_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{y}) = 1 + c\, (\boldsymbol{w} \bullet \boldsymbol{x})^T (\boldsymbol{w} \bullet \boldsymbol{y}) = 1 + c \sum_d w_d^2 \kappa^d(\boldsymbol{x}, \boldsymbol{y})$
  $\kappa^d(\boldsymbol{x}, \boldsymbol{y}) = x_d \cdot y_d$
  $\partial_d k_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{y}) = 2\, \delta\, c\, w_d\, \kappa^d(\boldsymbol{x}, \boldsymbol{y})\, \chi_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{y})^{\delta - 1}$

Il ressort de ces décompositions, particulièrement pour les noyaux linéaires et RBF gaussien, que le calcul des dérivées se déduit du noyau et des noyaux dimensionnels avec un coût additionnel très modéré.

L'algorithme SAS (*Scaled Alignment Selection*) de sélection que nous proposons consiste donc à déduire de la maximisation de l'Alignement le classement des descripteurs par ordre décroissant des poids w_d. On utilise pour la procédure d'optimisation un simple algorithme de montée de gradient avec initialisation de tous les poids à 1. Il s'agit d'une approche *filtre* (puisque aucun apprentissage de SVM n'est impliqué dans la sélection) liée à une stratégie PO.

On remarquera que dans le cas du noyau RBF gaussien pondéré, le paramètre σ est implicitement fixé par la détermination des facteurs de poids. En effet, si l'on définit le vecteur de paramètres $\boldsymbol{\Theta} = (\sigma, \boldsymbol{w})$, soit une valeur arbitraire $\tilde{\sigma}$, et $\tilde{\boldsymbol{\Theta}} = \left(\tilde{\sigma}, \frac{\tilde{\sigma}}{\sigma}\boldsymbol{w}\right)$, alors

$$k_{\boldsymbol{\Theta}}(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{\sum_i w_i^2 (x_i - y_i)^2}{2\sigma^2}\right) = \exp\left(-\frac{\sum_i \left(\frac{\tilde{\sigma}}{\sigma} w_i\right)^2 (x_i - y_i)^2}{2\tilde{\sigma}^2}\right) = k_{\tilde{\boldsymbol{\Theta}}}(\boldsymbol{x}, \boldsymbol{y}).$$

Notre proposition n'est pas la première à faire intervenir le critère d'Alignement pour la sélection de descripteurs. La principale contribution sur le sujet [166] est basée sur une minimisation conjointe de l'opposé de l'Alignement et de la norme zéro du vecteur normal de l'hyperplan $\boldsymbol{w}_h$, s'inspirant de la stratégie proposée par Bradley et Mangasarian pour l'algorithme FSV (section 7.4.1). Néanmoins, la fonction objectif n'étant pas convexe, elle est décomposée comme une différence de deux fonctions convexes, qui permet l'usage d'une technique de minimisation spécifique (DCA, *Difference of Convex functions minimization Algorithm*, proposé dans [224]). La minimisation s'opère par une double boucle ; chaque étape de la boucle intérieure implique donc le calcul de l'Alignement, ce qui rend l'algorithme beaucoup plus complexe et coûteux que l'algorithme SAS proposé ici. De plus, afin de faire apparaître la décomposition comme différence de deux fonctions convexes, les auteurs suppriment le dénominateur de normalisation de l'Alignement, se justifiant par le fait que dans le cas du noyau RBF gaussien, la matrice de Gram est déjà bornée. Nous montrons dans la partie expérimentale l'importance de ce dénominateur en comparant l'algorithme SAS au SFS proposé ci-dessous.

7.5.2 Sélection Pondérée basée sur le produit de Frobenius (SFS)

L'algorithme SFS (*Scaled Frobenius Selection*) suit exactement le même principe que l'algorithme SAS, en substituant au critère d'Alignement le dit *critère de Frobenius*, qui se résume au produit de Frobenius entre la matrice de Gram et la matrice cible, c'est-à-dire un critère d'Alignement non normalisé :

$$\mathcal{F}(\boldsymbol{K}, \boldsymbol{K}^*) = \langle \boldsymbol{K}, \boldsymbol{K}^* \rangle_F . \tag{7.11}$$

Cette algorithme n'est défini ici que pour mettre en évidence sa moindre efficacité par rapport au critère d'alignement complet, comme le montrera la partie expérimentale.

L'absence de normalisation se révèle d'ailleurs immédiatement préjudiciable pour l'usage du noyau linéaire puisque dans de nombreux cas, le critère de Frobenius diverge vers l'infini lorsque les poids w_d augmentent arbitrairement. La maximisation étant impossible en pratique, on se limitera donc pour la méthode présente aux noyaux non-linéaires.

7.5.3 Sélection Forward basée sur le critère d'Alignement (FAS)

Les approches par noyau pondéré, de même que la méthode R2W2, offrent un très bon compromis entre performances et complexité. Toutefois, comme nous l'avons précisé, la stratégie PO ne permet pas de lier d'un point de vue théorique la pertinence des S meilleurs descripteurs au classement de leurs poids parmi les D descripteurs originaux.

La décomposition des noyaux linéaires et RBF gaussien en noyaux dimensionnels, présentée plus haut dans la section 7.5.1, offre un grand avantage puisqu'elle permet l'évaluation de l'effet individuel d'un descripteur additionnel sur une matrice de Gram. Les noyaux sont ici dans leur forme originale non pondérée. On définit les *matrices de Gram dimensionnelles* $\boldsymbol{\kappa}^d = \left[\kappa^d(\boldsymbol{x}_i, \boldsymbol{x}_j)\right]_{ij}$, afin de transposer les relations énoncées sous forme matricielle.

Dans le cas du noyau linéaire, la contribution individuelle se fait par sommation des matrices de Gram dimensionnelles, avec $\kappa_d(\boldsymbol{x}, \boldsymbol{y}) = x_d \cdot y_d$:

$$\boldsymbol{K} = \sum_{d=1}^{D} \boldsymbol{\kappa}^d.$$

Dans le cas du noyau RBF gaussien, la contribution individuelle se fait par le produit des matrices de Gram dimensionnelles ($\bigotimes$ représente le produit terme à terme de différentes matrices), avec $\kappa^d(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{(x_d - y_d)^2}{2\sigma^2}\right)$:

$$\boldsymbol{K} = \bigotimes_{d=1}^{D} \boldsymbol{\kappa}^d.$$

Ce constat nous permet de définir une stratégie de recherche séquentielle (SEQ) *forward* de sélections successives de descripteurs, basée sur le critère d'Alignement. Nous proposons ainsi la méthode de sélection FAS (*Forward Alignement Selection*), détaillée dans l'algorithme 3.

Algorithme 3 Forward Aligment Selection

D le nombre de descripteurs
S le nombre de descripteurs à sélectionner
Initialisation : $\boldsymbol{K}_0 \leftarrow 0$, liste de descripteurs restants $I_1 \leftarrow 1, \ldots, D$.
Calcul des matrices de Gram dimensionnelles $\boldsymbol{\kappa}^i \ \ \forall i \in I_1$.
pour $d = 1$ à S **faire**
  pour $i \in I_d$ **faire**
    Noyau linéaire : $\boldsymbol{K}_d^i = \boldsymbol{K}_{d-1} + \boldsymbol{\kappa}^i$
    ou RBF gaussien : $\boldsymbol{K}_d^i = \boldsymbol{K}_{d-1} \bullet \boldsymbol{\kappa}^i$
    Calcul de l'Alignement : $A_d^i = \mathcal{A}(\boldsymbol{K}_d^i, \boldsymbol{K}^*)$.
  fin pour
  Sélection du descripteur de rang d : $i_d = \arg\max_{i \in I_d} A_d^i$
  $\boldsymbol{K}_d = \boldsymbol{K}_d^{i_d}$
  $I_{d+1} = I_d \setminus \{i_d\}$
fin pour
résultat : Liste ordonnée des descripteurs sélectionnés $i_1, \ldots, i_S$.

La stratégie séquentielle favorise ainsi la sélection des premiers descripteurs en considérant toutes les possibilités parmi les D descripteurs disponibles. Le calcul préalable des matrices de Gram dimensionnelles offre de plus une importante réduction du coût de calcul, l'essentiel des boucles de l'algorithme consistant donc en sommes ou produits terme à terme (donc parallélisables) de matrices. Cependant la complexité reste beaucoup plus élevée que l'algorithme pondéré SAS, en particulier lorsque le nombre de descripteurs D est élevé, comme nous l'avons précisé dans la présentation de la stratégie de recherche séquentielle (section 7.2.3). De plus l'algorithme tend à accumuler les erreurs au fil des itérations, et devient donc de moins en moins fiable à mesure que le nombre de descripteurs sélectionnés augmente.

Cependant pour une sélection restreinte ($S \ll D$), l'algorithme FAS obtient de meilleurs résultats que l'approche SAS, tout en impliquant une complexité raisonnable (en $O(SD)$).

7.5.4 Sélection Pondérée sur le critère de Séparabilité (SCSS)

Sur la base des critiques formulées à l'égard du critère d'Alignement (section 4.4.1.4), nous avons introduit le critère de Séparabilité de Classes « Kernelisé » (KCS), défini comme le rapport des dispersions inter-classes et intra-classes dans l'espace transformé. On rappelle l'expression du critère KCS, défini par rapport aux matrices $\boldsymbol{B}$ et $\boldsymbol{W}$, exprimées dans les équations 4.17 et 4.18 :

$$\mathcal{J} = \frac{\mathbf{1}_n^T \boldsymbol{B} \mathbf{1}_n}{\mathbf{1}_n^T \boldsymbol{W} \mathbf{1}_n} = \frac{\Sigma(\boldsymbol{B})}{\Sigma(\boldsymbol{W})},$$

où l'opérateur Σ est égal à la somme des composantes d'une matrice ($\Sigma(\boldsymbol{A}) = \sum_{ij} a_{ij}$).

Dans le cas du noyau pondéré $k_{\boldsymbol{w}}$, on peut donc en déduire l'expression de la dérivée par rapport au poids w_d :

$$\frac{\partial \mathcal{J}}{\partial w_d} = \frac{\Sigma(\partial_{w_d}\boldsymbol{B})\Sigma(\boldsymbol{W}) - \Sigma(\boldsymbol{B})\Sigma(\partial_{w_d}\boldsymbol{W})}{\Sigma(\boldsymbol{W})^2},$$

avec :

$$\Sigma(\partial_{w_d}\boldsymbol{B}) = \Sigma(\partial_{w_d}\boldsymbol{K_{11}}) + \Sigma(\partial_{w_d}\boldsymbol{K_{12}}) - \Sigma(\partial_{w_d}\boldsymbol{K}), \tag{7.12}$$

$$\Sigma(\partial_{w_d}\boldsymbol{W}) = \sum_{i=1}^{n} \partial_{w_d} k_{ii} - \Sigma(\partial_{w_d}\boldsymbol{K_{11}}) - \Sigma(\partial_{w_d}\boldsymbol{K_{12}}). \tag{7.13}$$

On a ôté l'indice $\boldsymbol{w}$ de la matrice de Gram $\boldsymbol{K_w}$ et de ses sous-blocs pour clarifier les notations, mais ici le critère est bien calculé sur la matrice de Gram du noyau pondéré.

On propose donc l'algorithme SCSS (*Scaled Class Separability Selection*) basé sur le même principe que l'algorithme SAS, où l'on substitue le critère de Séparabilité dans l'espace transformé ($\mathcal{J}$) au critère d'Alignement.

Cependant, bien que le critère de séparabilité soit a priori plus fiable que le critère d'Alignement (on a vu que le premier prend en compte la dispersion intra-classes, tandis que le second ne fait intervenir que la mesure de dispersion inter-classe), son expression introduit une instabilité numérique dans la maximisation de $\mathcal{J}$ que nous avons évoquée dans la section 4.4.2. On appliquera donc ici à nouveau la procédure de régularisation proposée dans cette précédente section pour éviter ce problème. Nous verrons cependant dans la partie expérimentale que le critère KCS n'apporte pas de gain en pratique par rapport au critère d'Alignement pour l'approche par optimisation sur noyau pondéré.

Notons qu'un algorithme de sélection de descripteur basé sur le critère KCS a également été proposé par Wang [239]. Néanmoins, afin de contourner le problème de régularisation, l'auteur avance que le critère KCS est borné inférieurement par la grandeur $\operatorname{tr} S_b$ (qui se trouve en fait être son numérateur), et base en conséquence toutes ses expériences sur ce critère plus simple, qui se trouve être, comme nous l'avons montré, le produit de Frobenius entre la matrice de Gram et la matrice cible (soit le critère de Frobenius, comme nous l'avons appelé dans la section 7.5.2). La dispersion intra-classes est alors absente du critère employé. On peut donc considérer que le critère KCS n'est pas réellement exploité par l'auteur.

7.5.5 Sélection sur le Discriminant de Fisher Kernelisé (KFDS)

La dernière méthode que nous proposons s'appuie sur un modèle très proche du critère de séparation des classes, basé sur les matrices de dispersion dans l'Analyse Discriminante de Fisher Kernelisée (*Kernel Fisher Discriminant Analysis* KFDA).

Le problème de l'Analyse Discriminante de Fisher a été introduit dans la section 3.2 et consiste en la détermination d'un hyperplan (de vecteur normal $\boldsymbol{w}_h$) de séparation entre deux classes, autrement dit d'un axe $\boldsymbol{w}_h$ de projection des données maximisant le critère de l'équation 4.13 ($J = \frac{\operatorname{tr} \boldsymbol{S}_b}{\operatorname{tr} \boldsymbol{S}_w}$). Le problème peut ainsi être formulé sous la forme de la maximisation du critère $J(\boldsymbol{w}_h)$ suivant :

$$J(\boldsymbol{w}_h) = \frac{\boldsymbol{w}_h^T \boldsymbol{S}_b \boldsymbol{w}_h}{\boldsymbol{w}_h^T \boldsymbol{S}_w \boldsymbol{w}_h}.$$

où l'on retrouve les matrices de dispersion inter-classes et intra-classes (équations 4.14 et 4.15, l'expression suivante de $\boldsymbol{S}_b$ est égale à un facteur près à celle de l'équation 4.14) :

$$\begin{aligned} \boldsymbol{S}_b &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T, \\ \boldsymbol{S}_w &= \sum_{c=1,2} \sum_{\boldsymbol{x}_i \in \mathcal{S}_c} (\boldsymbol{x}_i - \boldsymbol{\mu}_c)(\boldsymbol{x}_i - \boldsymbol{\mu}_c)^T. \end{aligned}$$

Mika et al. ont montré [158] qu'il est possible de formuler ces dernières exclusivement en termes de produits scalaires, ce qui permet, de manière similaire aux SVM, d'étendre le champ des surfaces

de séparation à des surfaces plus complexes, en leur substituant une fonction noyau k. Soit Φ la fonction de transformation relative au noyau k, on « kernelise » les matrices de dispersion de la manière suivante :

$$\begin{aligned} \boldsymbol{S}_b^\Phi &= (\boldsymbol{\mu}_1^\Phi - \boldsymbol{\mu}_2^\Phi)(\boldsymbol{\mu}_1^\Phi - \boldsymbol{\mu}_2^\Phi)^T, \\ \boldsymbol{S}_w^\Phi &= \sum_{c=1,2} \sum_{\boldsymbol{x}_i \in \mathcal{S}_c} (\Phi(\boldsymbol{x}_i) - \boldsymbol{\mu}_c^\Phi)(\Phi(\boldsymbol{x}_i) - \boldsymbol{\mu}_c^\Phi)^T, \end{aligned}$$

où l'on a introduit les centres des classes dans l'espace transformé $\boldsymbol{\mu}_c^\Phi = \frac{1}{n_c} \sum_{\boldsymbol{x}_i \in \mathcal{S}_c} \Phi(\boldsymbol{x}_i)$. Les résultats de la Théorie des Noyaux Reproduisants nous permettent d'affirmer que le vecteur $\boldsymbol{w}_h$ se trouve dans l'espace engendré par les exemples de l'ensemble d'apprentissage, soit : $\boldsymbol{w}_h = \sum_{i=1}^n \alpha_i \Phi(\boldsymbol{x}_i)$, ce qui permet [244] de reformuler les termes du critère $J(\boldsymbol{w}_h)$:

$$\begin{aligned} \boldsymbol{w}_h^T \boldsymbol{S}_b^\Phi \boldsymbol{w}_h &= \boldsymbol{\alpha}^T \boldsymbol{M} \boldsymbol{\alpha}, \\ \boldsymbol{w}_h^T \boldsymbol{S}_w^\Phi \boldsymbol{w}_h &= \boldsymbol{\alpha}^T \boldsymbol{N} \boldsymbol{\alpha}, \end{aligned}$$

où l'on a introduit les deux matrices $\boldsymbol{M}$ et $\boldsymbol{N}$. On décompose la matrice de Gram $\boldsymbol{K}$ en deux blocs verticaux, chacun relatif à une classe, soit $\boldsymbol{K} = [\boldsymbol{K}_1 \boldsymbol{K}_2]$, avec $[\boldsymbol{K}_c]_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ pour $i = [1, \ldots, n]$ et $\boldsymbol{x}_j \in \mathcal{S}_c$. De plus, on définit les vecteurs $\boldsymbol{M}_c$ comme les moyennes de colonnes des matrices $\boldsymbol{K}_c$, soit $[\boldsymbol{M}_c]_i = \frac{1}{n_c} \sum_{\boldsymbol{x}_j \in \mathcal{S}_c} k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Cette décomposition nous permet d'exprimer les matrices introduites :

$$\begin{aligned} \boldsymbol{M} &= (\boldsymbol{M}_1 - \boldsymbol{M}_2)(\boldsymbol{M}_1 - \boldsymbol{M}_2)^T \\ \boldsymbol{N} &= \sum_{c=1,2} \boldsymbol{K}_c \left(\boldsymbol{I} - \frac{1}{n_c} \boldsymbol{1}_{n_c} \right) \boldsymbol{K}_c^T. \end{aligned}$$

La matrice $\boldsymbol{1}_{n_c}$ étant une matrice de taille $n_c \times n_c$ dont les termes sont égaux à 1.

L'Analyse du Discriminant de Fisher Kernelisé (KFDA) consiste donc en la maximisation du critère J suivant, par rapport aux coefficients α_i :

$$J(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^T \boldsymbol{M} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \boldsymbol{N} \boldsymbol{\alpha}}.$$

Le problème est résolu de manière analogue à l'algorithme LDA non-kernelisé, par la recherche des vecteurs propres de la matrice $\boldsymbol{N}^{-1}\boldsymbol{M}$.

L'extraction d'un vecteur $\boldsymbol{w}_h$ de séparation nous mène à appliquer une transposition de la méthode AROM (basée sur les SVM), sur le résultat précédent. On propose l'algorithme KFDS (*Kernel Fisher Discriminant Selection*) qui consiste à mettre à jour itérativement les facteurs de pondération de descripteurs du vecteur $\boldsymbol{w}$ par un produit terme à terme avec le vecteur $\boldsymbol{w}_h$ ($\boldsymbol{w} \leftarrow \boldsymbol{w} \bullet \boldsymbol{w}_h$), puis à appliquer à l'itération suivante l'algorithme d'analyse de Fisher sur les descripteurs pondérés (on utilise donc la matrice de Gram pondérée $\boldsymbol{K}_{\boldsymbol{w}}$).

Cependant, afin de pouvoir exprimer le vecteur $\boldsymbol{w}_h$, l'algorithme proposé est soumis aux mêmes contraintes qu'AROM (transformée Φ explicite), et l'on limite donc pour l'instant son usage au noyau linéaire.

7.5.6 Avantages des méthodes proposées

Volume en mémoire

L'un des principaux avantages des méthodes proposées est leur adaptabilité en matière d'espace mémoire. En effet, la plupart des méthodes de sélection de descripteurs ne peuvent être appliquées sur un nombre trop élevé (plusieurs milliers) d'exemples puisqu'elles impliquent des structures numériques trop grandes pour être contenues dans des volumes classiques de mémoire vive.

Les deux critères proposés, de même que leurs dérivées par rapport aux facteurs de pondération, diffèrent des autres critères, du fait qu'ils sont exclusivement basés sur des termes additifs (des traces et des sommes). Ils peuvent donc être calculés itérativement au moyen d'une décomposition

en blocs des matrices de Gram ; le calcul de la matrice cible n'est d'ailleurs pas nécessaire puisqu'il équivaut, si le bloc est homogène en termes de classes, qu'à un éventuel changement de signe. Les besoins en mémoire peuvent donc être arbitrairement faibles, selon le compromis choisi avec la complexité (par exemple le pré-calcul des matrices de Gram dimensionnelles est largement profitable en termes de complexité, mais suppose un coût en stockage conséquent).

De plus, les matrices de Gram étant symétriques, seule la moitié des blocs non diagonaux est nécessaire pour l'évaluation des critères, ce qui apporte un gain en complexité supplémentaire.

Ces remarques ne s'appliquent pas cependant au dernier algorithme (KFDS), qui implique une inversion de matrice.

Complexité

L'approche pondérée basée sur le critère d'Alignement (SAS, section 7.5.1) est proposée ici comme une alternative à la méthode R2W2 présentée plus haut (section 7.4.4). La méthode R2W2 est basée sur une double boucle d'optimisation, chaque itération externe implique donc :

1. L'apprentissage d'un SVM pour évaluer les facteurs α_i et la norme $\|\boldsymbol{w}\|^2$.
2. L'évaluation de la matrice de Gram $\boldsymbol{K}$.
3. Une boucle d'optimisation par programmation quadratique pour évaluer le rayon R.
4. Le calcul des matrices dérivées $\partial_{w_d}\boldsymbol{K}$ pour chaque facteur de pondération w_d.

Les méthodes basées sur l'Alignement (KTA) et la Séparabilité des Classes (KCS) ne nécessitent que les étapes 2 et 4 de la boucle R2W2, sans impliquer l'apprentissage de SVM ou la résolution de problèmes de programmation quadratique, qui sont tous les deux des phases particulièrement coûteuses de l'algorithme. Les mesures de temps de calcul prodiguées dans la section expérimentale 7.7 confirmeront que les méthodes proposées sont plus rapides que R2W2, et présentent des performances comparables, voire meilleures dans certains cas.

De la même manière, l'algorithme KFDS (section 7.5.5), proposé comme alternative à la méthode AROM (section 7.4.2), se révèle beaucoup plus rapide que cette dernière, la résolution d'un système matriciel $\boldsymbol{Ax} = \boldsymbol{B}$ étant moins coûteuse que l'apprentissage d'un SVM, nécessaire dans les itérations de la méthode AROM.

7.6 Synthèse

Le tableau 7.1 synthétise la taxonomie des différents algorithmes de sélection de descripteurs présentés dans ce chapitre. La troisième colonne précise, dans le cas des méthodes adaptées aux SVM, quels types de noyaux sont pris en compte, la quatrième indique si l'algorithme est de type filtre, enveloppeur ou embarqué (selon la taxonomie présentée dans la section 7.2.4), enfin la cinquième précise la stratégie de recherche associée (parmi les stratégies énumérées dans la section 7.2.3).

On peut voir que les algorithmes proposés sont exclusivement des méthodes de type filtre, ce qui explique leur moindre complexité, mais prenant toutefois en compte les spécificités des SVM, et adoptant en majorité une stratégie d'optimisation, également largement exploitée dans les méthodes existantes basées sur les SVM.

7.7 Expériences comparatives

Nous présentons dans cette section un protocole expérimental, à la fois sur des données synthétiques et réelles, destiné à comparer les diverses approches pour la sélection de descripteurs et à valider les algorithmes proposés dans ce document. Nous exploiterons toutes les méthodes énumérées dans la synthèse de la section 7.6 précédente, à l'exception de IRMFSP, Kolomogorov-Smirnov et FSV, dont les résultats sont globalement décevants par rapport aux méthodes plus

	Section	Noyaux	Paradigme	Stratégie de recherche
Méthodes existantes				
Fisher	7.3.1	·	Filtre	BIN
IRMFSP	7.3.2	·	Filtre	SEQ
Kolmogorov-Smirnov	7.3.3	·	Filtre	BIN
FSV	7.4.1	Linéaire	Embarquée	PO
AROM	7.4.2	Linéaire	Embarquée	PO
RFE	7.4.3	Linéaire	Enveloppeur	SEQ
R2W2	7.4.4	Tous	Enveloppeur	PO
Méthodes proposées				
SAS	7.5.1	Tous	Filtre	PO
SFS	7.5.2	Non-linéaire	Filtre	PO
FAS	7.5.3	Linéaire & RBF	Filtre	SEQ
SCSS	7.5.4	Tous	Filtre	PO
KFDS	7.5.5	Linéaire	Filtre	PO

TABLE 7.1 – Tableau récapitutif des méthodes de sélection de descripteurs présentées dans ce chapitre, avec renvoi aux sections correspondantes. Y sont précisés les types de noyaux pris en compte, le type d'algorithme et la stratégie de recherche.

récentes. Nous conservons néanmoins le critère de Fisher, qui sert de référence, en raison de son usage très courant dans la littérature, et de son coût extrêmement réduit en temps de calcul.

Les commentaires porteront sur les courbes d'erreur, plus lisibles que les tableaux de résultats. On remarquera que sur toutes les figures, les méthodes proposées sont représentées par des pointillés et les méthodes de l'état de l'art sont en traits plein.

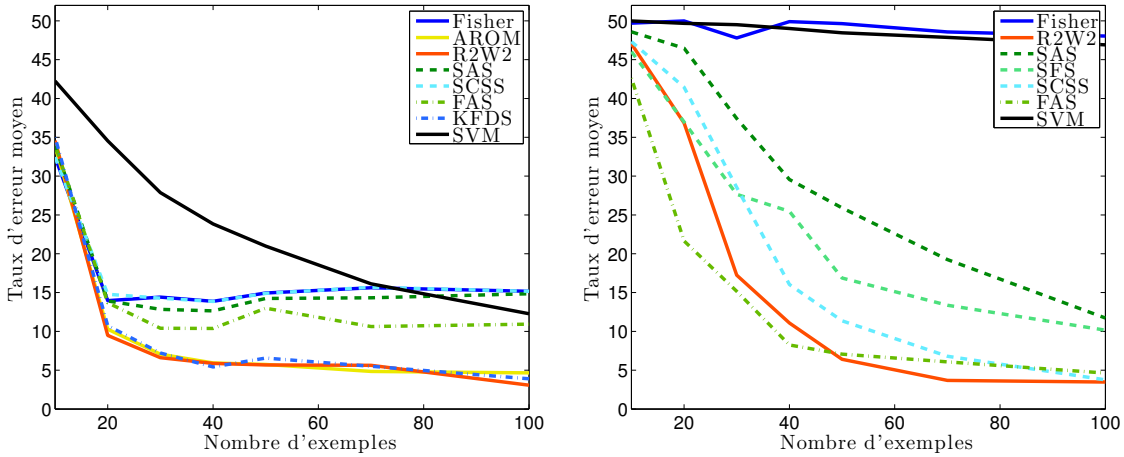
7.7.1 Données artificielles

Nous avons reproduit ici l'expérience décrite dans [245] et [47], incluant les protocoles de synthèse des données et d'évaluation des résultats. L'expérience compare les performances des différents algorithmes sur un problème linéairement séparable (que nous désignerons par « problème linéaire ») et un problème non-séparable linéairement (que nous désignerons par « problème non-linéaire »), par l'évaluation des SVM sur deux descripteurs sélectionnés parmi un ensemble conséquent de descripteurs redondants ou non-pertinents. Les approches basées sur des noyaux emploient respectivement un noyau linéaire et un noyau RBF gaussien pour chacun des cas. Nous précisons que sur toutes les expériences impliquant un noyau linéaire, l'approche par noyau pondéré sur le critère de Frobenius (SFS) n'est pas considérée car la définition du critère implique une divergence des facteurs de pondération vers l'infini lors de la phase d'optimisation.

Le problème linéaire réunit 202 descripteurs synthétiques dont seulement 6 ne sont pas du bruit, et sont corrélés entre eux par groupes de 3. Le problème non-linéaire regroupe 52 descripteurs dont seulement 2 ne sont pas du bruit, mais définissent des distributions multi-gaussiennes non-séparables linéairement. Nous invitons le lecteur à consulter l'article original [245] pour une description plus détaillée de la synthèse des données.

L'expérience consiste à observer l'évolution des performances lorsque le nombre n d'exemples d'apprentissage (générés aléatoirement) varie entre 10 et 100. A chaque itération les deux meilleurs descripteurs sont sélectionnés, un classifieur SVM est appris sur ces deux composantes, à partir des mêmes exemples d'apprentissage, et le taux d'erreur est calculé sur un ensemble de 500 exemples de test (générés selon la même distribution). La procédure est répétée de manière à évaluer l'erreur moyenne sur 40 itérations. La figure 7.1 montre les performances des différentes méthodes sur les deux problèmes. La courbe noire pleine indique les résultats obtenus en conservant tous les descripteurs pour l'apprentissage des SVM, afin d'évaluer l'amélioration apportée par la sélection de descripteurs.

Le problème linéaire, dont les résultats apparaissent sur la figure 7.1(a), illustre le principal défaut des méthodes proposées basées sur l'alignement (SAS, SFS et FAS) et le critère de séparabilité des classes (SCSS) dans leur incapacité à trouver la meilleure solution en présence de descripteurs



(a) noyau linéaire (b) noyau non-linéaire

FIGURE 7.1 – Comparaison des performances sur un problème séparable linéairement (a) et un problème non-séparable linéairement (b), impliquant un grand nombre de descripteurs bruités.

fortement redondants. Avec ces méthodes, de même qu'avec le critère de Fisher, le taux d'erreur tend vers 15% lorsque n augmente, parce que deux descripteurs pertinents mais redondants sont sélectionnés, au lieu de deux descripteurs indépendants apportant plus d'information. L'approche *forward* sur l'alignement (FAS) apporte cependant une amélioration sensible des performances par rapport à l'approche pondérée (SAS). Par contre les approches R2W2, AROM et la méthode proposée sur le discriminant de Fisher kernelisé (KFDS) se montrent plus aptes à prendre en compte ces redondances et présentent des résultats comparables.

Le problème non-linéaire (figure 7.1(b)) n'implique pas de redondance et est centré sur la capacité à détecter la complémentarité de deux descripteurs indépendants pour la séparation des classes. Sans surprise, le critère de Fisher est ici totalement incapable de distinguer les descripteurs pertinents, de même que les autres méthodes basées sur un noyau linéaire (AROM et le Fisher kernelisé KFDS) dont les résultats ne sont pas affichés pour une meilleure lisibilité, mais similaires à ceux du critère de Fisher. L'approche *forward* sur l'alignement (FAS) obtient ici les meilleurs résultats, avec l'approche R2W2. Ce constat montre la pertinence de l'approche *forward*, comparée aux approches par noyaux pondérés (SAS, SFS et SCSS), pour la sélection d'un nombre réduit de descripteurs. On remarque pour finir que le critère de séparabilité de classes (SCSS) apparaît ici plus efficace que le critère d'alignement (SAS), mais ce constat ne se retrouve malheureusement pas sur les données réelles. On remarque également que l'usage du critère de Frobenius (SFS) par rapport à l'Alignement complet (SAS) réduit les performances du système. Ce résultat sera confirmé par les expériences suivantes.

7.7.2 Données réelles

Nous avons également testé les méthodes proposées sur des données réelles. Nous réemployons ici les bases de données introduites dans la section 4.6 et présentées en détail dans l'annexe B, dont trois sont des bases disponibles sur le dépôt public UCI [16], et la dernière une base construite à partir de nos données sur le problème de classification *parole/musique*. Les bases ont été choisies pour couvrir un éventail assez diversifié de configurations, en termes de nombre de descripteurs originaux et de nombre d'exemples d'apprentissage. Ainsi, les bases *Ionosphere* et *Spambase* offrent un ensemble modéré d'exemples et de descripteurs (de l'ordre de la centaine) tandis que la base *Lymphoma*, tirée d'une expérience décrite dans [244] et [99], est caractérisée par une très vaste collection de descripteurs (plusieurs milliers) et un nombre réduit d'exemples (quelques dizaines). La base *parole/musique* se distingue par un nombre très conséquent d'exemples (20000).

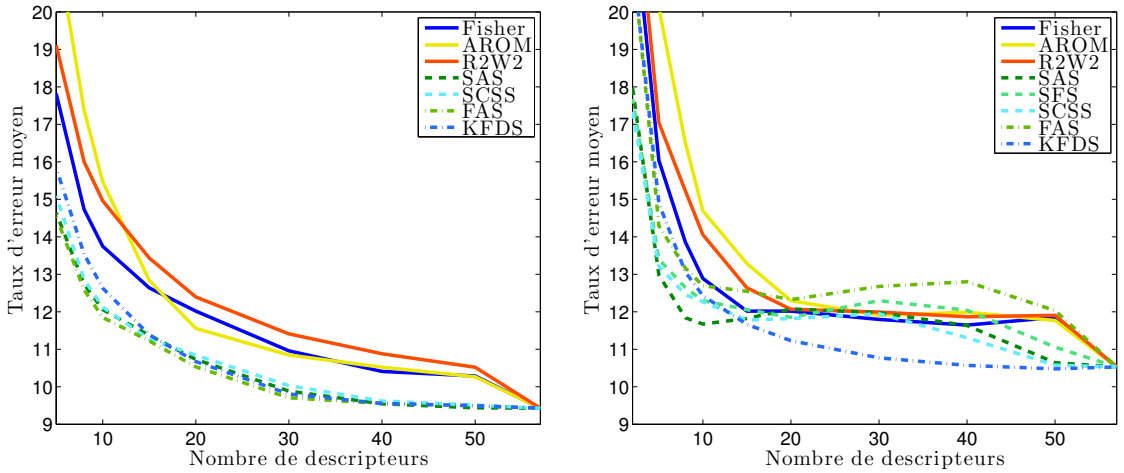
Le protocole d'évaluation est proche de celui déployé sur les données artificielles. La sélection de descripteurs est appliquée sur un ensemble d'apprentissage de n_{appr} exemples tirés aléatoirement parmi les n exemples de la base (qui contient n_1 et n_2 exemples pour chacune des deux classes). Un

classifieur SVM est ensuite appris sur le même ensemble d'apprentissage dont on a sélectionné les D descripteurs les plus pertinents, D étant ici le paramètre variable de l'évaluation, alors que l'on faisait varier n_{appr} pour les problèmes sur données synthétiques. Le taux d'erreur moyen est calculé sur 30 itérations de ce processus, et l'on fournira généralement les résultats avec un noyau linéaire et un noyau non-linéaire (RBF gaussien). Le tableau 7.2 résume les principales caractéristiques des bases exploitées ici.

Base	Nb d'exemples n (n_1/n_2)	n_{appr}	n_{test}	Nb descripteurs
Artificiel linéaire	synthétique	10 à 100	500	202
Artificiel non-linéaire	synthétique	10 à 100	500	52
Lymphoma	96 (34 / 62)	60	36	4026
Ionosphere	351 (126 / 225)	250	101	34
Spambase	4601 (2788 / 1813)	500	1000	57
Parole/musique	20000 (10000 / 10000)	500	500	321

TABLE 7.2 – Caractéristiques comparées des bases employées pour l'évaluation.

7.7.2.1 Spambase



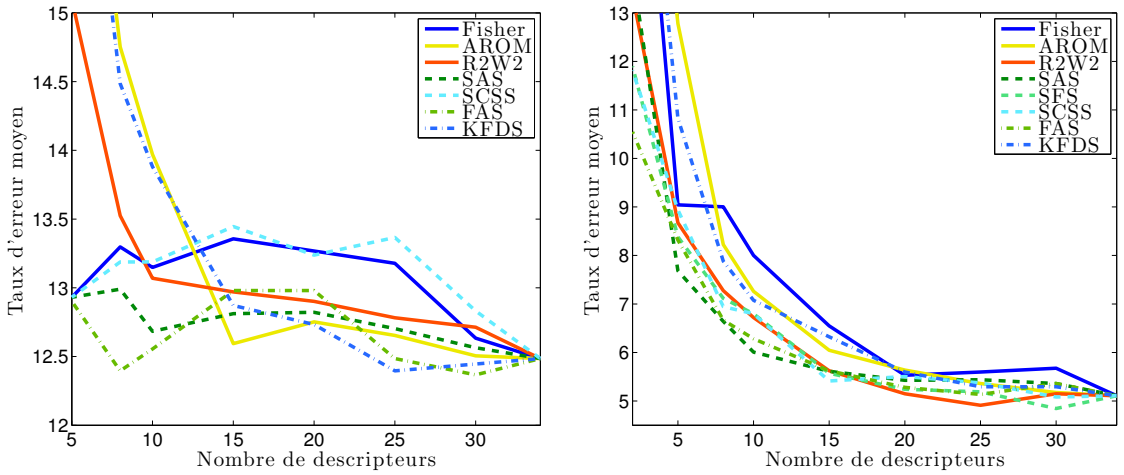
(a) noyau linéaire (b) noyau RBF gaussien

FIGURE 7.2 – Comparaison des performances entre les différentes méthodes sur la base *Spambase*.

Les résultats observés sur la base *Spambase*, illustrés par la figure 7.2, montrent en premier lieu la séparabilité linéaire du problème, étant donné que le noyau RBF gaussien n'apporte aucune amélioration par rapport au noyau linéaire. Cet exemple ne semble pas impliquer de descripteurs non pertinents (ou bruités) pénalisant la classification, puisque le taux d'erreur décroît de manière monotone lorsque le nombre de descripteurs sélectionnés augmente. Pour les deux noyaux, on constate que les algorithmes proposés apportent un net avantage par rapport aux approches existantes. Ainsi, dans le cas linéaire, on observe un réduction de 3% du taux d'erreur pour $d = 10$ sur les 15% d'erreur mesurés avec l'approche R2W2 (soit une réduction relative du taux d'erreur de l'ordre de 20%). Les très bonnes performances de l'approche basée sur le discriminant de Fisher kernelisé (KDFS) sont en outre surprenantes lorsqu'elle est suivie d'une classification avec noyau non-linéaire (figure 7.2 b), bien que l'algorithme KDFS soit basé exclusivement sur un noyau linéaire, ce qui tend à confirmer la séparabilité linéaire du problème. Le handicap de l'approche *forward* FAS sur le noyau RBF gaussien (qui sera confirmé par la suite), comparé aux approches par noyaux pondérés, montre que la complexité de cet algorithme n'apporte pas d'amélioration notable sur de petits ensembles d'apprentissage, à cause de la modélisation trop parcellaire des distributions de classes. L'écart de performances entre les approches pondérées SAS et SFS (où

le dénominateur de l'alignement est omis) sur le noyau RBF gaussien confirme par ailleurs la pertinence du terme de normalisation dans l'expression de l'alignement.

7.7.2.2 Ionosphere



(a) noyau linéaire (b) noyau RBF gaussien

FIGURE 7.3 – Comparaison des performances entre les différentes méthodes sur la base *Ionosphere*. On remarquera la différence d'échelle en ordonnées entre les deux figures (a) et (b).

Contrairement au cas précédent, le base *Ionosphere*, dont les résultats sont reportés sur la figure 7.3, constitue clairement un problème non-séparable linéairement. On constate en effet que le taux d'erreur reste supérieur à 12.5% avec le noyau linéaire, tandis que le noyau RBF gaussien permet de réduire celui-ci jusqu'à environ 5%. Les résultats mitigés des approches linéaires (Fisher, Fisher kernelisé et AROM) par rapport aux autres approches, viennent confirmer cet constat. Nous préférons donc ici nous concentrer sur les résultats mesurés avec le noyau RBF gaussien (figure 7.3(b)). À nouveau la pente décroissante quasi-monotone du taux d'erreur, que l'on retrouve sur toutes les méthodes, exclue l'hypothèse de la présence de descripteurs bruités pénalisant la classification. Les résultats confirment globalement les observations portées sur la base *Spambase*, à savoir une légère baisse des performances lorsque l'on substitue à l'alignement standard sur noyau pondéré (SAS) le critère allégé de Frobenius (SFS), ainsi que l'absence d'amélioration lorsque l'on emploie l'algorithme *forward* (FAS) très coûteux, par rapport à l'approche par noyau pondéré (SAS). De même, on n'observe pas d'amélioration lorsque l'on emploie le critère de séparabilité de classes (SCSS), pourtant théoriquement mieux justifié que le critère d'alignement. Toutefois, tous les algorithmes proposés ici, à l'exception du déterminant de Fisher kernelisé (KFDS) qui ne peut prendre en compte qu'un noyau linéaire, permettent d'obtenir des performances comparables à celles mesurées avec l'approche R2W2, tout en impliquant une charge de calcul plus réduite, comme nous le montrerons dans la section 7.7.3.

7.7.2.3 Lymphoma

La base de données *Lymphoma* est constituée d'un nombre très réduit d'exemples caractérisés par plusieurs milliers de composantes, cas typique des données génétiques traitées en bioinformatique. Ce problème particulier nous permet d'évaluer le comportement des algorithmes proposés sur de très larges collections de descripteurs, généralement fortement redondantes. L'approche *forward* sur critère d'alignement (FAS) n'a pas été appliquée ici du fait de sa complexité quadratique par rapport au nombre de descripteurs. Nous reproduisons ici l'expérience décrite dans [244] pour l'évaluation de la méthode AROM, qui se restreignait donc à l'usage du seul noyau linéaire. La méthode SFS n'est donc pas testée ici. La figure 7.4 montre les résultats obtenus pour les différentes méthodes, avec un noyau linéaire.

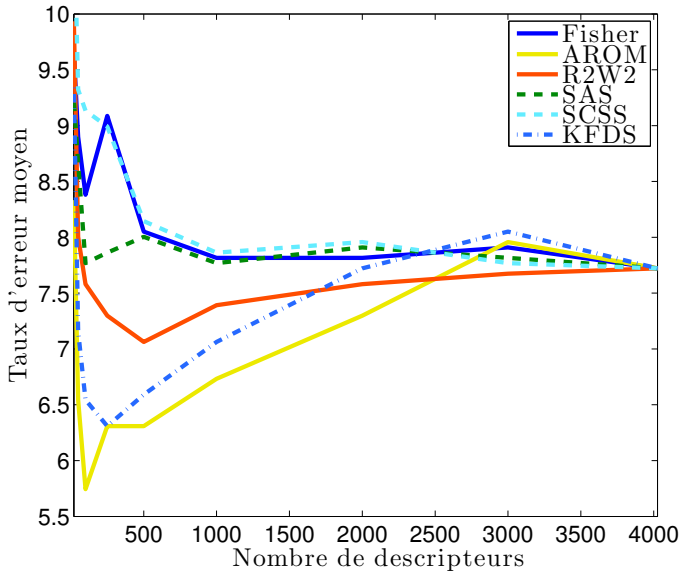


FIGURE 7.4 – Comparaison des performances entre les différentes méthodes sur la base *Lymphoma*.

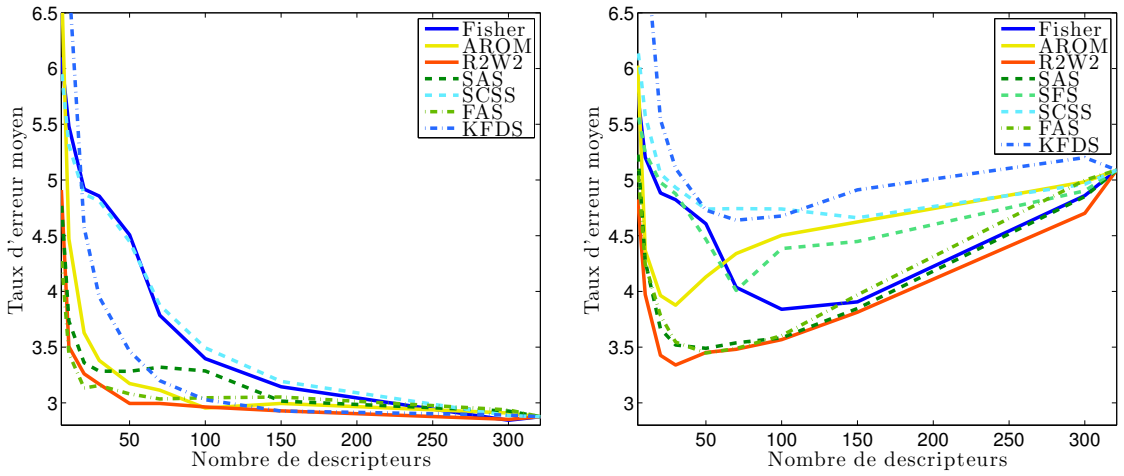
Les minima observés sur les courbes d'erreur des méthodes R2W2, AROM et Fisher kernelisé (KFDS) confirment le fait que la base contient un grand nombre de descripteurs non-pertinents, dont l'écrémage améliore les performances de classification. On constate d'ailleurs que le processus de sélection se révèle largement bénéfique en restreignant fortement le nombre de descripteurs. Ceci s'explique par le fait que l'information apportée par les descripteurs pertinents se trouve « diluée » dans la part dominante de bruit portée par le reste des descripteurs. De plus, les mauvais résultats obtenus avec le critère de Fisher montrent la forte interdépendance des descripteurs pertinents dans l'optimisation du problème, qui ne peut être mesurée indépendamment sur chacun d'entre eux.

L'expérience montre les limites des approches par noyau pondéré présentées ici (SAS, SFS, SCSS) sur de trop larges ensembles de descripteurs. En effet, l'optimisation conjointe sur l'ensemble des facteurs de pondération se révèle ici suboptimale et inefficace face à ce genre de configurations. Cependant la méthode basée sur le discriminant de Fisher kernelisé (KFDS), bien que légèrement inférieure à l'approche AROM, réussit dans ce cas à améliorer les performances tout en supprimant des descripteurs (on constate une réduction absolue de l'erreur de l'ordre de 1.5% sur les 7.7% d'erreur mesurés sans sélection), contrairement aux approches par noyau pondéré, et surpasse en outre les performances de l'approche R2W2.

7.7.2.4 Classification parole/musique

Nous terminons cette évaluation par une expérience similaire sur le problème central de cette thèse : la classification parole/musique. Ici, comme indiqué dans l'annexe B.3, la base ne contient que des exemples calculés sur des trames de parole ou de musique pure. Les descripteurs employés dans la base sont ceux que nous avons décrits dans le chapitre précédent.

Les résultats obtenus avec un noyau linéaire, représentés sur la figure 7.5(a), n'indiquent pas la présence de descripteurs non-pertinents pénalisant la classification à haute dimension. La pente très faible de l'erreur au-delà de 100 descripteurs constitue néanmoins une preuve de la forte redondance entre ces derniers ; nous avons vu en effet dans la section 6.6 que beaucoup d'entre eux traduisent des propriétés très similaires. On peut ainsi interpréter les descripteurs redondants comme une unique variable fortement pondérée qui, à travers l'amplification exponentielle implicite du noyau RBF gaussien, devient un fort handicap dans le cas non-linéaire. Les résultats, dans ce second cas, illustrés par la figure 7.5(b), confirment cette interprétation puisque que l'on constate que les méthodes les plus efficaces présentent un minimum très marqué autour de $D = 30$. Cette forte redondance explique ainsi les faibles performances du critère de Fisher avec le noyau linéaire,



(a) noyau linéaire (b) noyau RBF gaussien

FIGURE 7.5 – Comparaison des performances entre les différentes méthodes sur la base *parole/musique*.

puisque les descripteurs sont alors évalués indépendamment.

Bien que l'approche R2W2 se révèle la plus efficace, les méthodes proposées basées sur le critère d'alignement (SAS et FAS) présentent des résultats tout à fait comparables, l'écart étant compensé par le gain en temps de calcul. À nouveau on constate, en particulier sur le noyau RBF gaussien, que le critère de Frobenius (SFS) est clairement pénalisé par rapport à l'alignement normalisé (SAS). On observe enfin (figure 7.5 a) que l'approche *forward* sur l'alignement (FAS) est plus efficace à basse dimension que l'approche par noyau pondéré, ce qui confirme sa pertinence pour la sélection d'un ensemble très restreint de descripteurs.

7.7.3 Coût en temps de calcul

Le tableau 7.3 indique les temps de calcul en secondes moyennés sur l'ensemble des itérations, pour chaque méthode impliquée dans les expériences décrites. Toutes les durées que figurent ici proviennent des expériences sur noyau RBF gaussien, à l'exception de la base *Lymphoma*. Les méthodes proposées figurent en italique et la valeur soulignée pour l'approche FAS sur la base *Lymphoma* n'a été calculée que sur une seule itération. Les calculs ont été effectués sur un MacBook Core 2 Duo à 2.16 GHz avec 2 Giga-Octets de mémoire, sur un seul cœur du processeur.

Ces mesures sont bien sûr largement dépendantes de l'implémentation de chaque méthode. Aussi, afin d'homogénéiser au mieux cette comparaison, tous les SVM impliqués dans les diverses méthodes sont basés sur l'outil *SVMlight* de Thorten Joachims [114]. Les temps de calculs sont également directement liés au nombre d'itérations des processus d'optimisation (sauf pour l'approche *forward* FAS).

On constate que le discriminant de Fisher kernelisé (KFDS) constitue un compromis intéressant entre complexité et performances par rapport à l'approche AROM. Les deux méthodes sont construites sur le même principe et leur comportement est comparable sur les différentes expériences détaillées plus haut. Le pas de gradient et la condition d'arrêt sont strictement identiques dans les implémentations des deux algorithmes. On peut donc considérer l'algorithme KFDS, proposé ici, comme une alternative plus rapide à l'approche AROM, en particulier sur de larges ensembles de descripteurs (par exemple KFDS ne prend en moyenne que 12.5 s sur la base *Lymphoma*, tandis qu'AROM nécessite 130 s), au prix d'une légère baisse des performances.

Les expériences nous ont également montré que la méthode à noyau pondéré sur l'alignement (SAS) est globalement comparable en performances à l'approche R2W2, et même meilleure dans certains cas. De plus, les temps de calcul mesurés nous permettent de constater que la méthode proposée est plus rapide, en particulier en présence d'un grand nombre d'exemples d'apprentissage, comme sur la base *parole/musique*, où la méthode R2W2 nécessite 281 s de calcul tandis que la sélection par SAS s'exécute en seulement 54.6 s. Le coût en temps de calcul est à peu près le même

Base	**Lymphoma**	**Ionosphere**	**Spambase**	**Par/mus**
n_{appr}	60	250	500	500
Nb de descripteurs D	4026	34	57	321
Noyau	Linéaire	RBF	RBF	RBF
Fisher	19ms	2ms	3ms	10ms
AROM	130.0	2.6	4.8	12.3
Fisher kernelisé (KFDS)	*12.5*	*0.9*	*3.6*	*5.0*
R2W2	387.9	24.1	113.4	281.3
Frobenius pondéré (SFS)	.	*2.4*	*10.8*	*68.0*
Alignment pondéré (SAS)	*103.8*	*3.5*	*13.0*	*54.6*
Sép classes (SCSS)	*103.7*	*2.7*	*12.6*	*108.5*
Alignment forward (FAS)	*7432*	*3.9*	*28.2*	*1122*

TABLE 7.3 – Comparaison des temps de calcul moyens (en secondes) des méthodes impliquées dans l'évaluation (Par/mus désigne ici la base Parole/musique).

pour toutes les approches par noyau pondéré (SAS, SFS, SCSS), mais le tableau nous confirme le coût très élevé de la méthode *forward* (FAS) lorsqu'elle est appliquée sur un grand nombre de descripteurs (plus de 2 heures de temps d'exécution sur la base *Lymphoma*), ce qui proscrit son usage pour ce genre de situations.

7.8 Commentaires

Nous avons proposé des alternatives fiables aux méthodes de l'état de l'art pour la sélection de descripteurs, adaptées à la discrimination par Machines à Vecteurs de Support. Alors que la plupart des méthodes existantes sont de type enveloppeur (*wrapper*) et se basent donc sur des phases d'apprentissage par SVM pour l'évaluation des descripteurs pertinents, le récent critère d'alignement du noyau nous permet ici d'évaluer directement les performances d'un noyau par rapport à une base d'apprentissage donnée. Ce critère, proposé à l'origine pour la sélection de noyau, est pour l'instant peu exploité dans le domaine de la sélection de descripteurs. La méthode SAS, basée sur une stratégie de recherche par optimisation sur les facteurs de pondération du noyau, se révèle très efficace et comparable en performances aux méthodes les plus récentes tirant parti des noyaux, pour un coût moindre en temps de calcul. De plus, l'expression additive du critère, et l'absence de techniques complexes de programmation mathématique (comme la minimisation par programmation quadratique) en permettent une implémentation rapide et scalable, qui peut s'appliquer sur des ensembles d'apprentissage arbitrairement grands.

L'inclusion de la mesure de dispersion intra-classes a également été étudiée, à travers l'usage du critère de séparabilité de classes. Mais ce critère n'apporte pas les résultats escomptés avec l'approche par noyau pondéré. La tentative de régularisation pour éviter la convergence vers zéro des facteurs de pondération n'est malheureusement pas suffisante ici pour constituer une alternative fiable au critère d'alignement. Toutefois la pertinence de la mesure de dispersion complète (c'est-à-dire incluant les termes intra et inter-classes) est validée en suivant une autre approche basée sur le discriminant de Fisher kernelisé, similaire à la méthode AROM. La méthode KFDS, proposée ici, se révèle comparable en performances à AROM, pour un coût en temps de calcul largement réduit. Son usage reste toutefois limité au noyau linéaire, et son application à un ensemble plus large de noyaux constitue une perspective intéressante pour ces travaux.

Nous avons évoqué l'équivalence, démontrée par Shashua [216], entre la solution d'un SVM et la solution du discriminant linéaire de Fisher sur l'ensemble des vecteurs de support dans l'espace transformé. Ce constat ouvre une perspective très attirante qui consisterait à restreindre aux seuls vecteurs de support obtenus après un apprentissage SVM les méthodes de sélection de descripteurs basées sur la séparabilité de classes kernelisée ou le discriminant de Fisher kernelisé. Cependant, l'objectif étant de réduire le nombre de descripteurs, il nous faut pouvoir garantir le fait que le sous-ensemble des vecteurs de support demeuré inchangé sur un sous-ensemble de descripteurs. Des travaux dans ce sens permettraient ainsi, nous l'espérons, de réduire encore la complexité des