SVM et Visualisation pour la Fouille de Grands Ensembles de Données

Thanh-Nghi Do*, François Poulet*

*ESIEA Recherche, BP 0339, 53003 Laval-France
(dothanh, poulet)@esiea-ouest.fr

Résumé. Nous présentons un algorithme de SVM et des méthodes graphiques pour le traitement de grands ensembles de données. Pour pouvoir traiter de tels ensembles de données, nous utilisons une représentation des données de plus haut niveau (sous forme symbolique). L'algorithme de séparateur à vaste marge (SVM) est adapté pour pouvoir traiter ce nouveau type de données. Nous construisons un nouveau noyau RBF (Radial Basis Function) que l'algorithme utilise à la fois pour la classification, la régression et la détection d'individus atypiques dans des données de type intervalle. Nous utilisons ensuite des méthodes de visualisation interactive (elles aussi adaptées au cas des variables de type intervalle) pour expliquer les résultats obtenus par les SVM. La méthode est évaluée sur des ensembles de données symboliques existant ou créés artificiellement.

1 Introduction

Le début des années 2000 voit la quantité d'information stockée dans le monde croître de manière très importante. On estime qu'elle augmente de deux exa (10^{18}) octets tous les ans (Lyman et al. 2003). Une telle masse d'information est trop complexe pour pouvoir être appréhendée simplement par un utilisateur. L'extraction de connaissances à partir de données (ECD) s'est développée pour pouvoir découvrir des connaissances à partir de très grandes quantités d'information. Le processus d'ECD (Fayyad et al. 1996) est un processus non trivial permettant d'identifier des structures inconnues, valides et potentiellement exploitables dans les bases de données.

Dans cette problématique, nous nous sommes plus particulièrement intéressés à une méthode récente de fouille de données à l'aide de SVM (Vapnik 1995). Les SVM et les méthodes de noyaux sont reconnues comme une méthodologie efficace pour la résolution de plusieurs problèmes : la classification supervisée, la régression, la détection d'individus atypiques, le clustering. Les SVM donnent de bons résultats dans la pratique en ce qui concerne le taux de précision, mais ils nécessitent la résolution d'un programme quadratique dont la mise en œuvre est coûteuse en temps d'exécution et mémoire. Un autre inconvénient est que les SVM fournissent très peu d'informations en sortie, ils ne retournent que les vecteurs support pour construire la frontière de séparation des données. L'utilisateur peut se servir de cette frontière pour classifier ses données avec de bons taux de précision mais il ne peut pas expliquer le modèle obtenu. Or la compréhensibilité des résultats est elle aussi importante même si elle n'apparaît pratiquement jamais dans l'évaluation des algorithmes de fouille de données. L'interprétation des résultats de SVM est nécessaire pour permettre à l'utilisateur de comprendre les résultats et cela augmente sa confiance dans ces résultats.

Nos principaux apports sont d'une part l'amélioration de la confiance et de la compréhensibilité du modèle obtenu et d'autre part la capacité à traiter de très grandes quantités de données sur du matériel standard. Nous utilisons une représentation de données sous forme symbolique (Bock et Diday 1999). L'utilisation d'attributs symboliques comme des intervalles permet de pouvoir représenter sous une forme résumée l'ensemble de données traité. Il est donc possible dans ce cas de traiter des ensembles de données de très grandes tailles puisque ce ne sont pas les données elles-mêmes qui sont manipulées, mais une représentation plus succincte de celles-ci. Les problèmes à résoudre sont alors d'adapter les méthodes à ce nouveau type de variable. Nous construisons un nouveau noyau de RBF que l'algorithme utilise pour la classification, régression et détection d'individus atypiques dans des données de type intervalle. Nous proposons aussi d'utiliser des méthodes graphiques interactives pour interpréter les résultats de SVM. Les dimensions intéressantes dans le modèle obtenu peuvent être perçues de manière visuelle sur des représentations graphiques des résultats.

Le paragraphe 2 présente le principe de l'algorithme de SVM. Ensuite, nous présentons la fouille de grands ensembles de données avec un SVM sur des données symboliques de type intervalle dans le paragraphe 3. Les résultats numériques sont présentés dans le paragraphe 4. Nous présentons la visualisation des résultats de SVM dans le paragraphe 5 avant de conclure sur nos travaux.

2 Algorithme de SVM

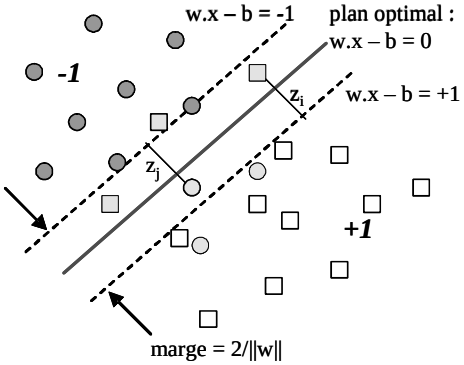

FIG. 1 – *Plan optimal pour la séparation des données en deux classes*

Les algorithmes de SVM sont souvent utilisés pour la classification, régression et détection d'individus atypiques dans la fouille de données.

2.1 Classification à l'aide de SVM

Soit un ensemble de données x_i *(i=1, 2, …, m)* avec deux classes $y_i = \pm 1$. Chaque exemple est représenté dans un espace de dimension *n*. Les SVM recherchent un hyperplan *(w, b)* permettant de séparer les exemples en deux classes comme montré sur la figure 1.

Le meilleur plan est celui qui se trouve le plus loin possible des deux classes. Pour trouver ce plan, il faut simultanément maximiser la marge de séparation ($2/||w||$) en se basant sur les deux plans support des deux classes et minimiser les erreurs (les z_i). L'hyperplan optimal passe au milieu des deux plans support. La recherche de l'hyperplan optimal se ramène à résoudre le programme quadratique (1) :

$$\begin{aligned} &min\ \Psi(w, b, z) = (1/2)\ ||w||^2 + c \sum_{i=1}^{m} z_i \\ &\quad avec : \\ &\qquad y_i(w.x_i - b) + z_i \geq 1 \\ &\qquad z_i \geq 0 \quad (i=1, 2, \ldots, m) \end{aligned} \tag{1}$$

où une constante $c > 0$ est utilisée pour contrôler la marge et les erreurs.

On peut également résoudre le programme quadratique (1) en se basant sur la formulation duale de Lagrange (2) :

$$\begin{aligned} &min\ \Psi(\alpha) = (1/2) \sum_{i=1}^{m} \sum_{j=1}^{m} (y_i\, y_j\, \alpha_i\, \alpha_j\, x_i.x_j) \ - \sum_{i=1}^{m} \alpha_i \\ &\quad avec : \\ &\qquad \sum_{i=1}^{m} y_i\, \alpha_i = 0 \\ &\qquad c \geq \alpha_i \geq 0 \end{aligned} \tag{2}$$

La résolution du programme quadratique (1) donne les α_i. Seuls les α_i correspondant aux exemples les plus proches de l'hyperplan sont supérieurs à zéro et ces exemples sont appelés vecteurs support. Le calcul de l'hyperplan *(w, b)* se base alors sur les vecteurs support (*#SV*) :

$$w = \sum_{i=1}^{\#SV} y_i\, \alpha_i\, x_i \tag{3}$$

Le scalaire *b* est obtenu en utilisant n'importe quelle paire de vecteurs support x_i (de la classe +1) et x_j (de la classe -1) dans l'équation (4) :

$$b = (1/2)\ w.(x_i + x_j) \tag{4}$$

La classification d'un nouvel exemple *x* est basée sur : $f(x) = signe\ (\sum_{i=1}^{\#SV} y_i\, \alpha_i\, x_i.x\ - b)$

En substituant le produit $x_i.x_j$ par une fonction de noyau $K(x_i, x_j)$ dans (2) on peut effectuer une classification non linéaire.

2.2 Régression à l'aide de SVM

Dans la tâche de régression, on recherche une fonction *f* permettant de prédire des valeurs continues en fonction des dimensions des exemples. La régression de SVM (SVR) utilise une

fonction de perte *ε-insensible* proposée par Vapnik. Avec une dérive ε, la SVR cherche une fonction de prédiction *f* tel que les valeurs prédites s'écartent au maximum de ε des valeurs désirées : $-\varepsilon \leq w.x_i - b - y_i \leq \varepsilon$. La tâche de SVR linéaire est exprimée par le programme quadratique (5) :

$$\begin{aligned} &min\ \Psi(w, b, z^*, z) = (1/2)\ ||w||^2 + c \sum_{i=1}^{m} (z_i^* + z_i) \\ &\quad avec : \\ &\qquad w.x_i - b - y_i - z_i^* \leq \varepsilon \\ &\qquad w.x_i - b - y_i + z_i \geq -\varepsilon \\ &\qquad z_i^*, z_i \geq 0 \qquad (i=1, 2, \ldots, m) \end{aligned} \tag{5}$$

où une constante $c > 0$ est utilisée pour contrôler la marge et les erreurs et z_i^*, z_i sont les variables de ressort.

En substituant le produit scalaire par une fonction de noyau *K* dans la formulation duale du programme quadratique, on obtient également la formule de SVR non linéaire (6) :

$$\begin{aligned} min\ \Psi(\alpha, \alpha^*) &= (1/2) \sum_{i=1}^{m} \sum_{j=1}^{m} ((\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(x_i, x_j)) \\ &\quad - \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) y_i + \varepsilon \sum_{i=1}^{m} (\alpha_i + \alpha_i^*) \\ &avec : \\ &\quad \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) = 0 \\ &\quad c \geq \alpha_i,\ \alpha_i^* \geq 0 \end{aligned} \tag{6}$$

La prédiction d'un nouvel exemple *x* est basée sur : $f(x) = \sum_{i=1}^{\#SV} (\alpha_i - \alpha_i^*) K(x_i, x) - b$

2.3 Détection d'individus atypiques avec SVM

Pour la tâche de détection d'individus atypiques (petit ensemble de données dont le comportement est différent du reste des données), les SVM ont pour objectif de rechercher l'hypersphère de rayon minimal *r* et de centre *o* qui contient la presque totalité des exemples. Un nouvel exemple est atypique s'il est à l'extérieur de l'hypersphère. Le SVM une classe est exprimé par (7) :

$$\begin{aligned} &min\ \Psi(r, o, z) = r^2 + (1/(mv)) \sum_{i=1}^{m} z_i \\ &\quad avec : \\ &\qquad ||x_i - o||^2 \leq r^2 + z_i \\ &\qquad z_i \geq 0 \quad (i=1, 2, \ldots, m) \end{aligned} \tag{7}$$

où une constante $v \in (0..1)$ est utilisée pour contrôler les erreurs et le nombre de vecteurs support.

La formulation duale de (7) avec une fonction de noyau *K* pour le SVM une classe non linéaire est de la forme (8) :

$$min\ \Psi(\alpha) = \sum_{i=1}^{m} \alpha_i \alpha_j K(x_i, x_j) - \sum_{i=1}^{m} \alpha_i K(x_i, x_i)$$

$$avec : \quad \sum_{i=1}^{m} \alpha_i = 1 \qquad (8)$$

$$1/(mv) \geq \alpha_i \geq 0$$

Un nouvel exemple *x* est considéré comme un individu atypique si :

$$f(x) = K(x, x) - 2\sum_{i=1}^{\#SV} \alpha_i K(x_i, x) + \sum_{i,j=1}^{\#SV} \alpha_i \alpha_j K(x_i, x_j) - r^2 \geq 0$$

où r est récupéré à partir d'un vecteur support x_k *(satisfaisant* $0 < \alpha_k < (1/mv)$*) :*

$$r^2 = K(x_k, x_k) - 2\sum_{i=1}^{\#SV} \alpha_i K(x_i, x_k) + \sum_{i,j=1}^{\#SV} \alpha_i \alpha_j K(x_i, x_j)$$

3 SVM et variable de type intervalle pour la fouille de grands ensembles

Fonction de noyau	**Forme fonctionnelle**
Linéaire	$K(u, v) = u.v$
polynomiale de degré *d*	$K(u, v) = (u.v + c)^d$
fonction à base radiale avec l'écart σ^2	$K(u, v) = exp(-\|\|u - v\|\|^2/(2\sigma^2))$
fonction sigmoïde	$K(u, v) = tanh(a(u.v) - b)$

TAB 1 – *Les fonctions de noyau les plus courantes.*

Les algorithmes de SVM sont très flexibles grâce aux méthodes de noyau. Aucun changement d'algorithme n'est nécessaire, par contre le choix des fonctions de noyau est très important pour obtenir de bons résultats. Soit 2 exemples u et v, on a les fonctions de noyau *K(u, v)* couramment utilisées, présentées dans le tableau 1. (Lin 2003) a illustré que le noyau gaussien RBF est intéressant pour les cas de frontières ou modèles linéaire et non linéaire.

L'apprentissage des SVM se ramène à résoudre un programme quadratique, la mise en œuvre d'un algorithme de SVM est donc coûteuse en temps. Pour pouvoir traiter des ensembles de données ayant un très grand nombre d'individus, nous réduisons la taille des ensembles de données en utilisant le concept de données symboliques de type intervalle.

Soit un très grand ensemble de données, nous utilisons un algorithme de clustering comme les K-moyennes (MacQueen 1967) ou les cartes de Kohonen (Kohonen 1995) pour créer des clusters pour chaque classe. Un cluster est considéré comme un vecteur de type intervalle dont une composant de type intervalle [min, max] représentant les valeurs d'une dimension d'un groupe d'individus dans ce cluster. Donc, un grand ensemble de données est représenté par des vecteurs de type intervalle (les clusters) avec une très petite taille. Nous

construisons un noyau gaussien que l'algorithme de SVM utilise pour la classification, régression et détection d'individus atypiques dans des données de type intervalle.

Soit une fonction de noyau RBF de deux individus *p* et *q* de type numérique classique en dimension *n* :

$$K(p, q) = \exp(-\gamma\|p - q\|^2) \quad (9)$$

Avec des données de type intervalle, nous avons substitué la distance euclidienne entre les deux individus p et q ($d_E = \|p - q\|$) classique (9) par la distance de Hausdorff (1868-1942) pour deux individus de type intervalle. Nous avons choisi la distance de Hausdorff dans ce cadre d'utilisation parce que cette distance est facile à calculer et vérifie les trois axiomes d'un espace métrique (axiome d'identité, de symétrie et triangulaire).

Soit deux intervalles $u = [u_{min}, u_{max}]$ et $v = [v_{min}, v_{max}]$, la distance de Hausdorff entre ces deux intervalles u, v est définie par (10) :

$$d_H(u, v) = \max\{| u_{min} - v_{min} |, | u_{max} - v_{max} |\} \quad (10)$$

La distance de Hausdorff entre les deux individus p et q en n dimensions de type intervalle est définie par (11) :

$$d_H(p, q) = \sqrt{\sum_{i=1}^{n} \max\left(\left| p_{i,\min} - q_{i,\min} \right|, \left| p_{i,\max} - q_{i,\max} \right|\right)^2} \quad (11)$$

En substituant $d_E(p, q) = \|p - q\|$ par $d_H(p, q)$ dans la fonction de noyau RBF (9), nous obtenons un nouveau noyau RBF pour les données de type intervalle.

Cette construction du noyau non linéaire gaussien est utilisée par les SVM pour la classification, régression et détection d'individus atypiques dans des données intervalle.

4 Résultats

Nous avons rajouté le code de la construction du noyau gaussien pour des données de type intervalle dans le programme LibSVM (Chang et Lin 2003) permettant de faire la classification, régression et détection d'individus atypiques.

Nous avons utilisé les ensembles de données de Statlog, de l'UCI (Blake et Merz 1998), de Delve (Delve 1996) et de Régression (Torgo 2003) pour évaluer des performances de notre approche. Nous avons fait le choix de créer des données de type intervalle à partir de ces ensembles de données de type continu. Nous avons utilisé l'algorithme des K-moyennes (MacQueen 1967) pour résumer les données sous la forme de clusters. Les valeurs minimum et maximum sur chaque variable continue des individus d'un même cluster servent ensuite de bornes pour créer la variable de type intervalle. Nous avons obtenu les ensembles de données de type intervalle présentés dans le tableau 2. D'autres méthodes (Bock et Diday 1999) proposent de créer des objets symboliques à partir de variables de type continu.

	#intervalles (taille originale)		#dim	Protocole de test
Wave (3 classes)	30	(300)	21	Leave-1-out
Iris (3 classes)	30	(150)	4	Leave-1-out
Wine (3 classes)	36	(178)	13	Leave-1-out
Pima (2 classes)	77	(768)	8	Leave-1-out
Segment (7 classes)	319	(2310)	19	10-fold
Shuttle (7 classes)	594	(58000)	9	10-fold
RingNorm (2 classes)	10000	(1000000)	20	10-fold
Bank8FM (1 classes)	450	(4499)	8	

TAB 2 – *Description des données de type intervalle.*

Les SVM ne traitent que les ensembles de données de haut niveau dont la taille est plus petite que les ensembles de données originaux. Par exemple, le plus grand ensemble de données RingNorm avec un million d'individus est représenté par dix mille individus de type intervalle (soit 1 % de l'ensemble total). Ensuite, les SVM peuvent traiter sans difficulté cet ensemble de données de type intervalle et conservent également de bons résultats. Les résultats obtenus pour la classification et la régression sur les 7 premiers ensembles de données de type intervalle sont présentés dans le tableau 3.

	Classification : précision % sur données de type intervalle (originaux)		Régression : squared-error sur données de type intervalle
Wave	80,00 %	(86,33 %)	0,462389
Iris	100,00 %	(97,35 %)	0,078389
Wine	97,22 %	(98,88 %)	0,075182
Pima	79,22 %	(77,34 %)	0,212736
Segment	91,22 %	(97,10 %)	1,696050
Shuttle	94,78 %	(99,80 %)	1,096640
RingNorm	100 %	(97,89 %)	0,037762

TAB 3 – *Résultats de classification et régression.*

Le tableau 4 présente les résultats de la détection d'individus atypiques dans 2 ensembles de données de type intervalle Shuttle et Bank8FM. Nous présentons le nombre d'individus atypiques et le nombre d'individus les plus significativement atypiques.

	#ind. atypiques	#ind. significativement atypiques
Shuttle	31	9
Bank8FM	12	6

TAB 4 – *Résultats de détection d'individus atypiques.*

Ces résultats sont potentiellement de traiter des fichiers de données de très grandes tailles sous la préservation de la qualité des modèles obtenus. Nous n'avons pas pu effectuer de comparaison par rapport à d'autres algorithmes parce qu'aucun résultat obtenu par un modèle non linéaire sur des données de type intervalle n'est disponible.

5 Visualisation des résultats de SVM

Les SVM ont illustré leur efficacité en fouille de données et donnent de bons résultats en classification, régression et détection d'individus atypiques. L'utilisateur n'obtient la plupart du temps que les vecteurs support ou les coefficients de l'hyperplan sans aucune autre indication, il est très difficile d'expliquer les résultats. Nous proposons une approche graphique interactive pour interpréter les résultats de SVM en classification, régression et détection d'individus atypiques. La visualisation des résultats de SVM se base sur la visualisation interactive multi-vue pour expliquer les modèles obtenus par les SVM.

5.1 Visualisation des résultats de classification

Dans la classification à l'aide d'algorithmes de SVM, la compréhension de la marge est très importante car elle représente la frontière (la plus large possible) entre les deux classes. La robustesse des modèles est mesurée par la taille de la marge et les erreurs commises par le modèle. Donc, il est intéressant de voir les individus les plus proches de la marge. Ces individus représentent naturellement la frontière de séparation des données en deux classes. L'utilisateur a aussi des informations sur la largeur de la marge et le nombre d'erreurs.

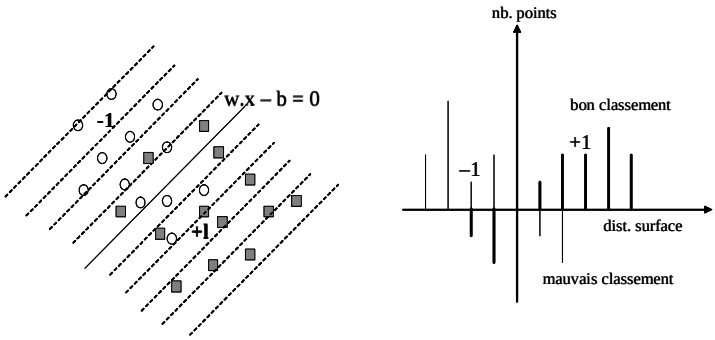

FIG. 2 – *Distribution de données en fonction de la distance à l'hyperplan*

Pour pouvoir visualiser les individus les plus proches de la marge, nous utilisons la visualisation multi-vue en se basant sur la distribution des individus en fonction de leur distance à la surface de séparation. Nous commençons par calculer la distribution des individus pendant la classification des données, avec en positif, les individus bien classés et en négatif les individus mal classés, la couleur représentant la classe. La figure 2 est un exemple du calcul de la distribution des données en fonction de la distance à l'hyperplan de séparation. Cette distribution est ensuite affichée sous la forme d'un histogramme.

Lorsque l'on sélectionne les barres de l'histogramme dans la vue de la distribution (les points les plus proches de la frontière de séparation), ces points sont alors automatiquement sélectionnés dans les matrices de scatter-plot 2D où une croix représente la projection d'un individu selon deux dimensions de type intervalle, la couleur correspondant à la classe (Poulet 2003). Cette approche donne à l'utilisateur des informations sur la marge. Il peut

trouver quelles sont les dimensions intéressantes dans le modèle obtenu (si ces dimensions montrent une frontière claire de séparation des données).

Sur la figure 3, nous avons utilisé le résultat obtenu par la classification de la classe *7* contre le reste des données Segment. Dans la visualisation de la distribution des individus, on sélectionne les individus les plus proches de la frontière de séparation (les barres de l'histogramme les plus proches de l'origine), ces individus sont automatiquement sélectionnés (les points noirs) dans la vue scatter-plot 2D. Ils représentent la marge de séparation de données. La projection correspondant aux dimensions *2* et *16* présente une frontière claire de séparation des données, ces deux dimensions sont intéressantes dans le modèle pour déterminer l'appartenance ou non à la classe *7*.

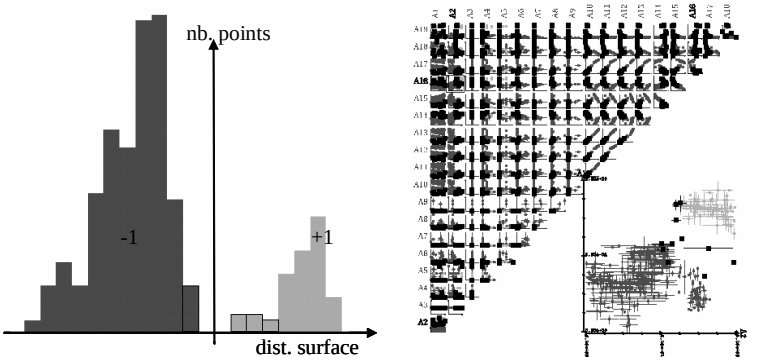

FIG. 3 – *Visualisation du résultat de SVM sur les données Segment (classe 7 contre le reste) avec l'histogramme et les matrices de scatter-plot en 2D*

5.2 Visualisation des résultats de régression

Nous proposons aussi de visualiser les résultats de la régression à l'aide de SVM pour permettre à l'utilisateur d'évaluer la qualité des résultats. Nous fournissons à l'utilisateur la visualisation des individus les plus éloignés de la fonction de régression. L'utilisateur peut avoir les informations intéressantes sur l'allure de la fonction de régression. Il sait comment la fonction de régression suit ses données.

Nous calculons d'abord la distribution des données en fonction de la distance entre les individus et la fonction de la régression. Ensuite, l'histogramme de la distribution de ces distances est lié aux matrices de scatter-plot 2D pour interpréter les résultats de régression. L'utilisateur obtient des informations sur la qualité de la régression et sur la fonction de régression, il peut trouver les dimensions importantes dans le modèle obtenu.

Nous avons fait une régression non linéaire sur l'ensemble de données de type intervalle Shuttle. La visualisation de ce résultat est présentée sur la figure 4 avec l'histogramme et les matrices de scatter-plot en 2D. Nous sélectionnons les barres les plus éloignées de la fonction de régression (les barres noires) dans l'histogramme, les individus correspondants (les primitives 2D noires) sont sélectionnés dans les matrices 2D. Si ces individus les plus éloignés de la fonction de régression sont atypiques, alors la fonction de régression suit bien les données. L'utilisateur peut avoir des informations sur la qualité de la régression. Avec les

matrices 2D, on aperçoit que les dimensions 1 et 8 sont importantes dans le modèle de régression parce qu'elles présentent clairement les individus les plus éloignés de la fonction de régression. L'utilisateur a une information intéressante sur la fonction de régression.

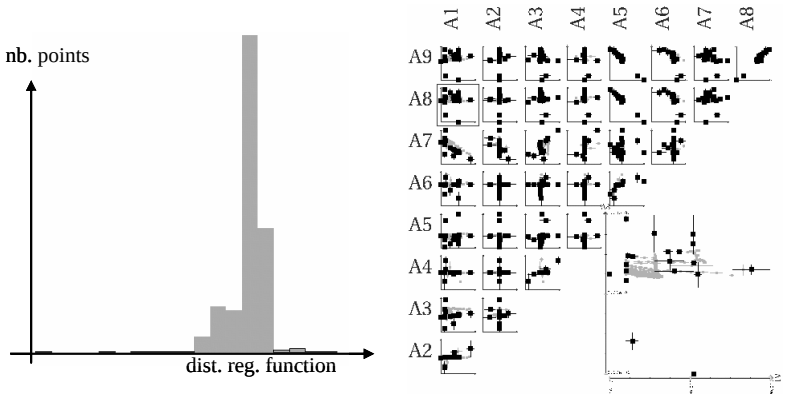

FIG. 4 – *Visualisation du résultat de régression de SVM sur les données Shuttle avec l'histogramme et les matrices 2D*

5.3 Visualisation des résultats de détection d'individus atypiques

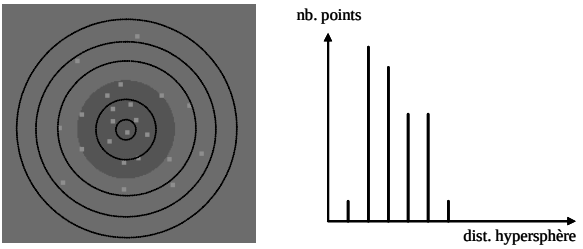

FIG. 5 – *Distribution des données en fonction de la distance à l'hypersphère*

Pour aider l'utilisateur à pouvoir voir et interpréter ou qualifier les individus atypiques. Nous utilisons aussi la visualisation multi-vue pour visualiser les individus en fonction de la distance à l'hypersphère obtenue par l'algorithme de SVM.

Nous calculons d'abord la distribution des données en fonction de la distance à l'hypersphère obtenue par l'algorithme. Ensuite, nous affichons cette distribution sous la forme d'un histogramme comme sur la figure 5. Lorsque l'on sélectionne les barres de l'histogramme (les points les plus éloignés de l'hypersphère), ces points sont alors automatiquement sélectionnés dans les matrices de scatter-plot 2D. L'utilisateur peut valider les individus atypiques.

La figure 6 est un exemple de la visualisation des résultats obtenus par la détection d'individus atypiques par SVM sur les données de type intervalle Bank8FM. Dans la visualisation de la distribution des individus en fonction de la distance à l'hypersphère, on

sélectionne les individus les plus éloignés de l'hypersphère, ces individus sont automatiquement sélectionnés (les points noirs) dans la vue scatter-plot 2D. Ils sont vraiment atypiques sur la projection des deux dimensions 5 et 7. Donc, ces deux dimensions sont celles qui ont un rôle important pour la détermination d'individus atypiques.

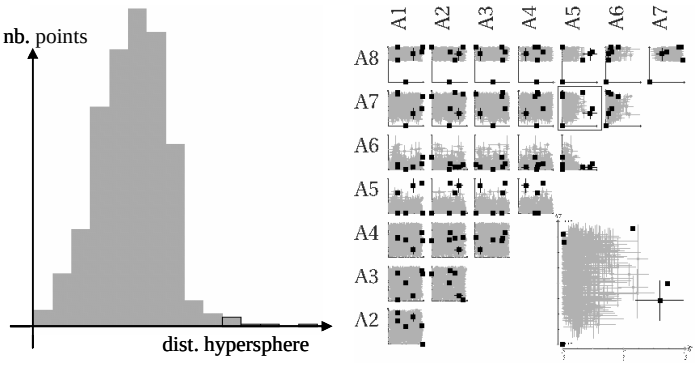

FIG. 6 – *Visualisation des résultats de détection d'individus atypiques sur les données de Bank8FM avec l'histogramme et les matrices 2D*

6 Conclusion et perspectives

Nous avons présenté dans cet article un nouvel algorithme de SVM pour traiter de grands ensembles de données. L'utilisation de données symboliques de type intervalle permet de pouvoir représenter sous une forme condensée l'ensemble de données traité. Ensuite, nous avons adapté l'algorithme de SVM à ce nouveau type de variables. Ces algorithmes de SVM sont capables de construire un modèle non linéaire pour les cas de classification supervisée, régression et détection d'individus atypiques avec des individus de type intervalle. Les résultats obtenus montrent que l'on peut potentiellement traiter des fichiers de données de très grandes tailles en préservant la qualité des modèles obtenus. De plus, comme le nombre d'individus est relativement peu élevé, les méthodes de visualisation peuvent ici être utilisées efficacement (alors que leurs limites en ce qui concerne le nombre d'individus sont bien connues). Les méthodes de visualisation des résultats permettent à l'utilisateur d'évaluer leur qualité et améliorent la compréhension des données et du modèle construit.

Les méthodes présentées sont des méthodes automatiques, il serait intéressant d'étudier les possibilités de mixage de ces approches automatiques avec les méthodes interactives que nous avons développées pour les variables de type intervalle et ou continues.

Références

Bennett K. et Campbell C. (2000), Support vector machines: hype or hallelujah ?, SIGKDD Explorations, 2(2), pp 1-13, 2000.

Blake C. et Merz C. (1998), UCI repository of machine learning databases, 1998. http://www.ics.uci.edu/~mlearn/MLRepository.html.

Bock H.H. et Diday E. (1999), Analysis of Symbolic Data, Springer-Verlag, 1999.

Chang C.C. et Lin C.J. (2003), LIBSVM -- A library for support vector machines, 2003. http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Cristianini N. et Shawe-Taylor J. (2000), An introduction to support vector machines and other kernel-based learning methods, Cambridge University Press, 2000.

Delve (1996), Data for evaluating learning in valid experiments, 1996. http://www.cs.toronto.edu/~delve.

Do T.N. et Poulet F. (2004), Enhancing SVM with visualization, in Discovery Science 2004, Suzuki E. et Arikawa S. Eds., Lecture Notes in Artificial Intelligence 3245, Springer-Verlag, 2004, pp 183-194.

Fayyad U., Piatetsky-Shapiro G. et Smyth P. (1996), From data mining to knowledge discovery in databases. AI Magazine, 17(3), pp 37-54, 1996.

Kohonen T. (1995), Self-Organizing Maps, Springer-Berlin, 1995.

Lin C.J. (2003), A practical guide to support vector classification. Talk at Freiburg University, Germany, 2003.

Lyman P, Varian H.R., Swearingen K., Charles P., Good N., Jordan L. et Pal J., How much information, 2003. http://www.sims.berkeley.edu/research/projects/how-much-info-2003/.

MacQueen J. (1967), Some methods for classification and analysis of multivariate observations, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, Vol. 1, pp 281-297, 1967.

Poulet F. (2003), Interactive decision tree construction for interval and taxonomical data, Proceedings of VDM@ICDM'03, 3nd Workshop on Visual Data Mining, Melbourne, USA, 2003, pp 183-194.

Poulet F. (2004), SVM and graphical algorithms: a cooperative approach, Proceedings of ICDM'04, 4th IEEE Int. Conf. on Data Mining, UK, 2004, pp 499-502.

Poulet F. et Do T.N. (2004), Mining very large datasets with support vector machine algorithms, in Enterprise Information Systems V, Camp O., Filipe J., Hammoudi S. et Piattini M. Eds., Kluwer Academic Publishers, 2004, pp 177-184.

Torgo L. (2003), Regression data sets, 2003. http://www.liacc.up.pt/~ltorgo/Regression-/DataSets.html.

Vapnik V. (1995), The nature of statistical learning theory, Springer-Verlag, 1995.

Summary

We present a new SVM algorithm and graphical methods for mining very large datasets. We summarize the massive datasets into the interval data. We adapt the SVM algorithm to deal with this interval data. We construct a new RBF kernel of interval data used for classification, regression and novelty detection tasks. We present interactive graphical methods (be also adapted to the interval data) for trying to explain the SVM results. The numerical test results are obtained on real interval and artificial datasets.