(19) RÉPUBLIQUE FRANÇAISE

INSTITUT NATIONAL
DE LA PROPRIÉTÉ INDUSTRIELLE

COURBEVOIE

(11) **N° de publication :** ***3 109 232***
(à n'utiliser que pour les commandes de reproduction)

(21) **N° d'enregistrement national :** **20 03637**

(51) Int Cl[8] : ***G 06 Q 10/04*** *(2019.12), G 06 N 20/20, G 06 F 17/18*

(12) **BREVET D'INVENTION** **B1**

(54) PROCEDE DE PREDICTION INTERPRETABLE PAR APPRENTISSAGE FONCTIONNANT SOUS RESSOURCES MEMOIRES LIMITEES.

(22) **Date de dépôt :** 10.04.20.

(30) **Priorité :**

(43) **Date de mise à la disposition du public de la demande :** 15.10.21 Bulletin 21/41.

(45) **Date de la mise à disposition du public du brevet d'invention :** 16.08.24 Bulletin 24/33.

(56) **Liste des documents cités dans le rapport de recherche :**

Se reporter à la fin du présent fascicule

(60) **Références à d'autres documents nationaux apparentés :**

◯ **Demande(s) d'extension :**

(71) **Demandeur(s) :** *ADVESTIS SAS* — FR.

(72) **Inventeur(s) :** Geissler Christophe et Margot Vincent.

(73) **Titulaire(s) :** ADVESTIS SAS.

(74) **Mandataire(s) :** In Concreto.

FR 3 109 232 - B1

Description

[0001] Titre

[0002] PROCEDE DE PREDICTION INTERPRETABLE PAR APPRENTISSAGE FONCTIONNANT SOUS RESSOURCES MEMOIRES LIMITEES

[0003] L'invention a trait aux algorithmes de prédiction, et notamment les algorithmes de prédiction exploitant des séries temporelles.

[0004] Par « séries temporelles » on désigne ici des données numériques évoluant dans le temps. L'indice temps peut être, selon les cas, par exemple la minute, l'heure, le jour, l'année.

[0005] Les séries temporelles sont des variables dont on dispose d'un échantillon de données $D_n = (X_i, Y_i)_{1<=i<=n}$ où pour tout i désignant le temps, X_i est un ensemble de variables explicatives ou covariables, et Y_i une variable d'intérêt.

[0006] Idéalement, la prédiction des séries temporelles consiste à modéliser le système qui a généré les données de la série, par exemple par un système d'équations mathématiques déterministes. En connaissant les conditions initiales, il serait alors possible de prévoir l'évolution du système.

[0007] Le plus souvent toutefois, les mécanismes ayant généré la série temporelle ne sont pas connus, et les seules informations disponibles sont les données passées. La modélisation se résume alors à imiter les facteurs générateurs de données, à partir des données passées, sans expliciter les mécanismes en action. Cette approche est à l'origine de la théorie statistique de l'apprentissage.

[0008] Les séries temporelles sont omniprésentes et apparaissent par exemples en météorologie, en biologie, dans le domaine médical, ou en économétrie.

[0009] La prédiction effectuée à partir des séries temporelles peut être utile, par exemples, pour la surveillance des patients dans les services médicaux, la détermination d'une charge de consommation d'énergie sur un réseau, la surveillance de l'état des forêts, la communication d'un taux de pollution de l'air, la maintenance prédictive, la prédiction du trafic automobile, l'optimisation de la valeur d'un portefeuille d'actifs.

[0010] La prédiction la plus simple à partir de séries temporelles passe par des approches linéaires et le calcul d'indices, tels que par exemple des indices de tendance centrale (moyenne, médiane), des indices de dispersion (variance), des indices de dépendance (auto-covariance, auto-corrélation). L'on connait ainsi des modèles statistiques anciens de prédiction de séries temporelles univariées, ces modèles étant de type auto-régressifs (*AR Auto-Regressive*), moyenne mobile (*MA Moving Average*), ainsi que leurs combinaisons et variantes (*ARIMA Autoregressive Integrated Moving Average*, *NARMA*) et extension à la prédiction de séries temporelles multivariées, c'est-à-dire celles où plusieurs valeurs évoluent simultanément (*VAR*).

[0011] Les approches linéaires classiques supposent que les séries temporelles sont stationnaires et qu'elles présentent des dépendances linéaires dans le temps.

[0012] Les régressions linéaires postulent une relation de dépendance globale entre la variable à expliquer Y et les variables liées

[0013] $$Y = \hat{Y} + Z = XA + Z$$
$$Y \in R^{N,1}, X \in R^{N,V}, A \in R^{V,1}, Z \in R^{N,1}$$

[0014] où N est la dimension temporelle de l'échantillon d'observation, V est le nombre de covariables, Z est l'erreur d'estimation, A réalise

[0015] $$Min\{\|Y - XA\|^2, A \in R^{V,1}\}$$

[0016] Les régressions linéaires présentent plusieurs inconvénients, en particulier une fragilité aux valeurs manquantes. Par ailleurs, lorsque V est très supérieur à 1, les coefficients de régression sont incontrôlables. Cet inconvénient a conduit à pénaliser les coefficients en norme L1 (Lasso), ou L2 (Ridge), dans lesquels A réalise

[0017] $$Min\{\|Y - XA\|^2 + \alpha\|A\|^2, A \in R^{V,1}\}$$

[0018] α étant un paramètre de contrôle.

[0019] L'on connait également des modèles non linéaires de prédiction de séries temporelles univariées (*ARCH*, *GARCH*).

[0020] Les séries temporelles peuvent également être projetées dans des espaces définis par des descripteurs statiques, par exemple transformation de Fourier, par ondelettes, ou décompositions polynomiales.

[0021] Les algorithmes d'intelligence artificielle pour la prédiction se sont largement développés ces dernières années, notamment pour la prédiction de l'état de santé de patients, même s'il existe des résistances à leur adoption.

[0022] De tels algorithmes peuvent apparaître comme concurrents du personnel professionnel, dans la mesure où ces algorithmes sont construits en vue d'éliminer des biais de jugement et de synthétiser des signaux contradictoires. De tels algorithmes peuvent en outre apparaître comme opaques dans leurs fonctionnements.

[0023] L'invention concerne notamment les algorithmes de prédiction exploitant des séries temporelles issues de systèmes dynamiques qui présentent des irrégularités, par exemple des systèmes déterministes non-linéaires ou chaotiques.

[0024] L'invention concerne également les algorithmes de prédiction exploitant des échantillons de données $D_n = (X_i, Y_i)_{1<=i<=n}$ où pour tout i désignant le temps, X_i est un ensemble de variables explicatives et Y_i une variable d'intérêt, les données D étant modélisées par des variables aléatoires indépendantes, les suites de variables (X_1, Y_1), $(X_2, Y_2)\ldots(X_n, Y_n)$ ne suivant pas une même loi inconnue.

[0025] L'invention trouve en outre des applications avantageuses lorsque l'indépendance des observations n'est pas réaliste ou peu probable, par exemple lorsque la variable

d'intérêt est le taux de pollution de l'air, ou le rythme cardiaque d'un patient, ou encore la valeur d'un actif dans un portefeuille.

[0026] L'invention concerne notamment les algorithmes de prédiction exploitant des séries temporelles et mettant en œuvre un apprentissage, c'est-à-dire la construction de règles pour le traitement automatique des données.

[0027] De tels algorithmes de prédiction ont été proposés dans l'état de la technique, en particulier machines à vecteur support, forêts aléatoires, réseaux de neurones.

[0028] Les machines à vecteur support (*SVMSupport Vector Machine*) travaillent à partir de classes de fonctions hypothèses, consistant en hyperplans d'un espace de caractéristiques, implicitement défini à partir de l'espace original, par une transformation non linéaire, construite via un noyau. L'algorithme SVM comprend une première phase d'apprentissage, consistant à déterminer un modèle de classification à partir des échantillons du jeu d'entraînement. Une deuxième phase consiste ensuite à appliquer le modèle à la totalité de la population à classer.

[0029] Le document CN 106419936 (Shenzhen Oudmon Tech) décrit l'utilisation d'une machine à vecteur support pour l'évaluation de l'état émotionnel d'une personne, par analyse de séries temporelles de photopléthysmographie. Le document EP 3011895 décrit l'utilisation de machines à vecteur support pour la classification de signaux d'électroencéphalogrammes.

[0030] Les forêts aléatoires (*RFRandom Forest*) mettent en œuvre une séparation des classes par un ensemble d'arbres de décision générés aléatoirement. Ces arbres de décision sont appliqués à des sous-ensembles du jeu d'entraînement en phase d'apprentissage. Le modèle final est ensuite appliqué à l'ensemble de la population à classer.

[0031] Le document EP 3564853 décrit l'utilisation de forêts aléatoires pour le traitement des obstacles par un véhicule autonome.

[0032] Les algorithmes de référence pour les arbres de décision sont ID3 (*Iterative Dichotomiser*), C4.5 et CART. Lors de la construction d'un arbre de décision, un critère de pureté comme l'entropie (utilisé dans C4.5) ou Gini (utilisé dans CART) est employé pour transformer une feuille en nœud. Des versions incrémentales des arbres de décision sont proposées (ID4, ID5R, ITI).

[0033] Un exemple d'utilisation d'arbre de décision CART pour l'aide au diagnostic d'athérosclérose est présenté en février 2020 par *Ghiasi et al, Decision tree-based diagnosis of coronary artery disease : CART model, Computer Methods and Programs in Biomedecine 192*. Les données utilisées concernent 303 patients et 55 paramètres indépendants. Le développement du modèle est effectué sur un ordinateur de bureau (CPU 2,93 GHz, 8 GB RAM).

[0034] Les réseaux de neurones (*ANNArtificial Neural Networks*) consistent en l'association en un graphe de neurones formels, modèles caractérisés par un état interne, des signaux

d'entrée et une fonction d'activation effectuant une transformation d'une combinaison affine des signaux d'entrée. Cette combinaison est déterminée par un vecteur de poids associé à chaque neurone et dont les valeurs sont estimées durant la phase d'apprentissage. Pour obtenir un système totalement non linéaire, le réseau de neurones doit comporter au moins une couche intermédiaire, appelée généralement couche cachée.

[0035] Le document US2018336452 (Sap) décrit un système de prédiction des incendies de forêt utilisant un réseau de neurones.

[0036] Les réseaux de neurones présentent plusieurs inconvénients. En particulier, il est a priori impossible de connaître l'influence effective d'une variable d'entrée sur le système, notamment dès qu'une couche cachée intervient. Ce fonctionnement en boite noire contraint fortement l'interprétation des résultats obtenus. Il semble par ailleurs que les performances des réseaux de neurones chutent lorsque l'on augmente l'horizon de prédiction ou lorsque l'on augmente la dimension des données.

[0037] Les méthodes algorithmiques de l'état de la technique présentent plusieurs inconvénients.

[0038] Les algorithmes existants ne sont que peu voire pas interprétables ou transparents. En d'autres termes, ils ne permettent pas de connaître les variables d'entrée qui ont une influence sur la prédiction. Cet inconvénient est particulièrement présent pour les algorithmes mettant en œuvre des réseaux de neurones avec ou sans apprentissage profond.

[0039] Lorsque les algorithmes de l'état de la technique sont relativement interprétables ou transparents, leur capacité de prédiction est faible lorsque les relations entre variables d'entrée et variables de sortie sont complexes. Cet inconvénient est particulièrement présent pour les algorithmes mettant en œuvre des arbres de décision, ou des machines à vecteur support.

[0040] Les algorithmes existants ont une faible tolérance par rapport aux données manquantes. Or, les données réelles dans les différentes applications scientifiques, industrielles, médicales ou financières de prédiction présentent souvent des plages manquantes ou incomplètes. Un prétraitement des données peut certes être effectué, en inférant les données manquantes à base d'heuristiques. Par exemple, la donnée manquante peut être remplacée par la moyenne des valeurs observées sur une séquence, ou par la dernière valeur observée sur la séquence. Ce prétraitement est toutefois long et couteux.

[0041] L'invention concerne plus particulièrement les algorithmes d'intelligence artificielle, de prédiction, mettant en œuvre une construction de règles par agrégation supervisée.

[0042] De tels algorithmes sont connus dans l'art antérieur, sous différentes formes.

[0043] Le document Patra (*Apprentissage à grande échelle, contribution à l'étude*

d'algorithmes de clustering répartis asynchrones, 2012) propose de mesurer les performances d'une stratégie de prévision quantile à l'aide d'une fonction de perte

[0044] $$L_n(g) = \frac{1}{n}\sum_{t=1}^{n} \rho_\tau\left(Y_t - g_t(Y_1^{t-1})\right)$$

[0045] une stratégie étant d'autant plus précise que la fonction de perte est petite. La stratégie de prévision repose sur une agrégation d'experts, chacun des prédicteurs fondamentaux étant fondé sur la technique des plus proches voisins, les prédictions étant agrégées avec des poids dépendant directement de leurs performances passées. Le poids d'un expert dans l'agrégat évolue ainsi au fur et à mesure du temps et est d'autant plus important que les prévisions passées de l'expert considéré ont été satisfaisantes. Selon Patra, cette stratégie de prévision *g* est universellement convergente. En d'autres termes, pour tout processus stationnaire et ergodique on a la convergence suivante

[0046] $$\lim_{n\to+\infty} L_n(g) = L^*,$$

[0047] où *L** est la plus petite perte asymptotique moyenne possible pour une stratégie de prévision.

[0048] Pour assurer la manipulation de grosses quantités de données, Patra propose un clustering, les séries temporelles étant séparées en sous-groupes présentant des similarités.

[0049] Le document Guedj (*Agrégation d'estimateurs et de classificateurs : théorie et méthodes, 2013*), rappelle que, selon le formalisme issu de la théorie de l'information, le risque associé à un estimateur, dans sa version empirique construite sur l'échantillon *Dn* est noté

[0050] $$R_n(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{f}(\mathbf{X}_i), Y_i).$$

[0051] ou *l* est la fonction de perte, par exemple quadratique. Guedj présente l'implémentation de méthodes d'agrégation à poids exponentiels.

[0052] Une présentation des différentes stratégies d'agrégation d'experts peut être trouvée dans le document *Soltz Aggrégation séquentielle de prédicteurs : méthodologie générale et applications à la prévision de la qualité de l'air et à celle de la consommation électrique, Journal de la société française de statistique, vol 151, n°2, 2010.*

[0053] Le document Margot et al (*Rule Induction Partitioning Estimator, ISSN 0302-9743, pp 288-301*) décrit un algorithme (*RIPE*) sélectionnant, à partir d'un échantillon $(X_i, Y_i)_{1<=i<=n}$, un ensemble de règles de type « Si A alors B », les conditions A étant des évènements du type $\{X \in r\}$, r étant un hyperrectangle. L'ensemble des hyperrectangles est ensuite transformé en une partition de l'espace permettant de construire

un estimateur universellement consistant. L'algorithme sélectionne un sous ensemble de règles S_n dans un ensemble de règles générées sur la base d'une condition de minium de l'erreur moyenne de prédiction

[0054] $$S_n^* := \underset{S \subset \mathcal{R}_n}{\arg\min}\, \mathcal{L}_n(\beta^{(S)})$$

[0055] La demanderesse a constaté que l'algorithme RIPE est bien adapté pour des données statiques identiquement distribuées, et n'exploite pas de manière satisfaisante la structure temporelle des données.

[0056] La demanderesse a constaté que les algorithmes d'apprentissage supervisé de l'état de la technique, notamment ceux disponibles dans les librairies publiques, exigent que les données soient rassemblées dans un fichier ou matrice unique. Or, la taille de cette matrice en mémoire vive peut facilement excéder les possibilités d'une machine ordinaire, notamment lorsque le nombre de variables est important.

[0057] Dès lors que le problème de classification porte sur des entités complexes, comme par exemple des patients dans un centre médical, le nombre total de variables peut être de plusieurs milliers. La totalité des données à prendre en compte croise dont l'ensemble des variables, l'ensemble des entités et l'ensemble des instances d'observations, par exemple à différents instants. Cet ensemble de données occupe donc une taille mémoire proportionnelle au nombre de variables et peut se révéler rapidement impossible à charger en une seule fois. Les modules standards de machine learning ne peuvent donc pas opérer sur ces ensembles de données.

[0058] Pour éviter un temps d'accès aux données sur disque dur, il est connu d'accéder aux données sous forme de flux, à l'aide d'algorithmes en ligne, utilisant des méthodes d'échantillonnage, de résumé de données ou de calcul distribué.

[0059] Pour limiter le temps de calcul et les besoins en mémoire, il a été proposé de construire le modèle au fur et à mesure de l'arrivée des données en utilisant un algorithme d'apprentissage incrémental, capable de mettre à jour son modèle à l'aide des nouvelles données, sans avoir besoin de toutes les revoir.

[0060] De nombreux algorithmes incrémentaux existent, mais leurs besoins en ressource mémoire et processeur ont une croissance non linéaire avec la taille des données.

[0061] On connaît dans l'art antérieur différentes approches pour générer un modèle à partir de données ne pouvant être toutes chargées en mémoire : les données peuvent être découpées en plusieurs ensembles (chunks) et/ou utiliser des techniques de parallélisassions de l'algorithme d'apprentissage.

[0062] L'apprentissage hors lignes correspond à l'apprentissage d'un modèle sur un jeu de données disponible au moment de l'apprentissage. Ce type d'apprentissage est réalisable sur des volumes de taille faible, jusqu'à quelques giga-octets (GO). Au delà, le temps d'accès et de lecture des données devient prohibitif, et il devient difficile de

réaliser un apprentissage qui ne prenne pas des heures ou des jours.

[0063] L'invention vise à pallier les inconvénients des algorithmes connus dans l'état de la technique, en particulier pour la prédiction à partir de séries temporelles.

[0064] Un premier objet de l'invention est une méthode algorithmique d'exploitation de séries temporelles ne présentant pas les inconvénients des méthodes antérieures et permettant une mise en œuvre sur une machine de bureau, telle qu'un ordinateur personnel, dont les ressources en mémoire vive (RAM) sont limitées.

[0065] Un deuxième objet de l'invention est de fournir une telle méthode algorithmique consommant des ressources de mémoire vive indépendantes du nombre de variables descriptives dans le problème de classification.

[0066] Un troisième objet de l'invention est de fournir une telle méthode algorithmique pouvant opérer sur des données contenues dans des fichiers situés sur des supports séparés.

[0067] Un quatrième objet de l'invention est de fournir une telle méthode algorithmique fournissant des commentaires explicatifs associés à la classification d'observations numériques, les commentaires explicatifs étant exprimés comme des conditions simples portant sur les variables retenues par les utilisateurs pour la classification.

[0068] Un cinquième objet de l'invention est de fournir une telle méthode algorithmique fournissant des commentaires explicatifs ayant la forme de règles d'association du type « si condition 1 et condition 2 et… condition n, alors la variable d'intérêt appartient à la classe K ».

[0069] Un autre objet de l'invention est une telle méthode algorithmique, notamment pour l'exploitation de séries temporelles, ne présentant pas les inconvénients des méthodes antérieures et permettant en particulier une prévision interprétable.

[0070] Un autre objet de l'invention est une telle méthode algorithmique, en particulier d'exploitation de séries temporelles, permettant le traitement de données structurées massives.

[0071] Un autre objet de l'invention est une telle méthode algorithmique, pour l'exploitation de séries temporelles par apprentissage.

[0072] Un autre objet de l'invention est de fournir une telle méthode algorithmique permettant de révéler les variables influentes dans la prévision, et donc dans les décisions prises sur la base de ces prévisions.

[0073] A ces fins, il est proposé, selon un premier aspect, un procédé technique de classification de données apte à être mis en œuvre sur un ordinateur de bureau, le procédé exploitant des données d'entrée d'un ensemble d'apprentissage comprenant :

[0074] - des co-variables X^i explicatives, décrites par un ensemble d'instances indexées par un ensemble d'individus I_k et un ensemble d'occurrence T_l ;

[0075] - les observations d'une variable Y d'intérêt ;

[0076] les données des co-variables X^i explicatives étant contenues dans des fichiers distincts, le procédé comprend les étapes suivantes :

[0077] – définition d'une règle testant si une réalisation de X est dans un hyper-rectangle de l'espace des variables explicatives ;
– définition de la complexité de la règle ;
– discrétisation de l'espace des variables explicatives en M modalités ;
– recherche récursive sur la complexité des règles jusqu'à une complexité maximale fixée ;
– sélection d'un sous ensemble de règles avec prédiction supérieure à zéro et d'un sous ensemble de règles avec prédiction inférieure à zéro, en contrôlant leur chevauchement.

[0078] Les données des co-variables explicatives peuvent se trouver dans différents répertoires, dans différents lecteurs d'un réseau, dans différents périphériques externes.

[0079] Une règle est ainsi un objet de type

[0080] $$\text{"}Si\ X \in r\ Alors\ \hat{Y} = p_r\text{"}$$

[0081] tel que

[0082] – la condition *Si* teste si une réalisation de X est dans un hyperrectangle de l'espace des variables explicatives

[0083] $$r = \prod_{k=1}^{d} I_k$$

[0084] – l'implication est la valeur prédite par la règle sur la condition est vérifiée, avec

[0085] $$p_r = \frac{\sum_{i=1}^{n} y_i \mathbf{1}_{x_i \in r}}{\sum_{i=1}^{n} \mathbf{1}_{x_i \in r}}$$

[0086] – la complexité d'une règle étant définie par

[0087] $$cp(r) = d - \#\{1 \le k \le d ; I_k = \mathcal{X}_k\}$$

[0088] Avantageusement, le procédé comprend une détermination de l'acceptabilité d'une règle, cette détermination comprenant les étapes suivantes :

[0089] - calcul de la couverture de la règle ;

[0090] - calcul de la significativité de la règle ;

[0091] - vérification de ce que la couverture de la règle est comprise entre deux valeurs prédéterminées ;

[0092] - vérification de ce que la significativité de la règle est supérieure à une valeur prédéterminée ;

[0093] - calcul d'un gain pénalisé.

[0094] Avantageusement, une règle est acceptable uniquement si la condition de couverture, la condition de significativité, et la condition sur les gains sont vérifiées.

[0095] La condition de significativité peut ainsi être avantageusement exprimée comme suit :

[0096] $$\frac{\sqrt{n(r,D_n)}|p_r-\bar{y}|}{\hat{\sigma}} \geq a$$

[0097] $n(r,D_n)$ désignant le nombre d'observations de l'ensemble D_n qui satisfont les conditions de la règle r.

[0098] La condition de couverture peut avantageusement être exprimée comme suit :

[0099] $$c_{min} \leq \frac{n(r,D_n)}{n} \leq c_{max}$$

[0100] c_{min} et c_{max} étant deux constantes vérifiant $0 < c_{min} < c_{max} < 1$.

[0101] Avantageusement, le procédé comprend une étape de vérification de ce qu'une condition sur le gain pénalisé est vérifiée.

[0102] Dans certaines mises en œuvre, la condition sur les gains est exprimée comme suit :

[0103] $$\frac{\Delta}{n}\sum_{j=0}^{n/\Delta-1}\left(\sum_{i=j\Delta+1}^{(j+1)\Delta} y_i \times p_r \mathbf{1}_{x_i \in r}\right) - \gamma_r > 0$$

[0104] où Δ est une période fixée et γ_r une pénalisation dépendante de la règle.

[0105] Avantageusement, le procédé est mis en œuvre sur un ordinateur de bureau dont la mémoire vive est d'une capacité inférieure à 20 GO.

[0106] L'invention se rapporte, selon un deuxième aspect, à un procédé d'apprentissage par ordinateur d'une commande d'un système technique, le procédé mettant en œuvre une classification technique de données tel que présenté ci-dessus, le procédé d'apprentissage étant basé sur des séries temporelles sous la forme d'un échantillon de données $D_n = (X_i, Y_i)_{1<=i<=n}$ où pour tout i, X_i est un ensemble de variables explicatives et Y_i une variable d'intérêt.

[0107] L'invention se rapporte, selon un troisième aspect, à un support lisible par ordinateur sur lequel sont stockées des instructions lisibles par machine pour exécuter un procédé tel qu'il vient d'être présenté.

[0108] D'autres objets et avantages de l'invention apparaîtront à la lumière de la description de modes de réalisation, faite ci-après, en référence aux dessins annexés dans lesquels :

[0109] [fig.1] est un schéma illustrant l'élimination de règles similaires, dans la mise en œuvre d'un procédé alternatif d'apprentissage supervisé disponible dans les librairies publiques ;

[0110] [fig.2] est un graphe représentant la profondeur explicative des règles obtenues par un procédé selon l'invention ;

[0111] [Fig.3] est un graphe représentant la profondeur explicative des règles obtenues par un procédé alternatif d'apprentissage supervisé disponible dans les librairies publiques.

[0112] L'invention propose un algorithme de prédiction exploitant des données de séries temporelles, l'algorithme mettant en œuvre un apprentissage, c'est-à-dire la construction de règles de décision et d'inférence pour le traitement automatique des données.

[0113] Les séries temporelles sont des variables dont on dispose d'un échantillon de données $D_n = (X_i, Y_i)_{1<=i<=n}$ où pour tout i, X_i est un ensemble de variables explicatives et Y_i une variable d'intérêt.

[0114] L'on souhaite prédire Y conditionnellement à X.

[0115] Les observations $(X_i, Y_i)_{1<=i<=n}$ sont modélisées par des variables aléatoires.

[0116] On suppose que les variables explicatives et les variables d'intérêts appartiennent à des ensembles mesurables.

[0117] Les observations sont modélisées par des variables aléatoires suivant une même loi ou non, indépendantes ou non.

[0118] L'hypothèse d'indépendance des observations n'est pas retenue lorsque le phénomène observé la rend peu réaliste, comme par exemple dans le cas de la surveillance de la pollution de l'air.

[0119] Pour une application mesurable appelée prédicteur est défini un risque et la prévision consiste à trouver, à l'aide des données D_n uniquement, un prédicteur tel que son risque est minimal.

[0120] La loi suivie par les variables étant inconnue, le risque est celui d'une règle d'apprentissage (ou estimateur) lié à l'échantillon D_n défini par

[0121] $$\mathcal{R}_P(\hat{f}(D_n)) = \mathbb{E}\left[c(\hat{f}(D_n; X), Y) \mid D_n\right]$$

[0122] Dans l'algorithme, la fonction de contraste c est avantageusement la fonction de contraste quadratique.

[0123] Un expert f_i de poids Π_i est une fonction constante en son premier argument et qui vaut l'espérance empirique de Y sachant X :

[0124] $$f_i(x, D_T) = \frac{\sum_{s=1}^{T} y_s \mathbf{1}_{x_s \in k_i}}{\sum_{s=1}^{T} \mathbf{1}_{x_s \in k_i}}$$

[0125] Au moins un sous ensemble d'expert est identifié par minimum de contraste, soit

[0126] $$\sum_{s=1}^{T} \gamma_q\left(\hat{\theta}_T^{\{\mathbf{f},\mathcal{S}\}}; (x_s, y_s)\right) = \min_{t \subset \mathcal{F}} \frac{1}{t} \sum_{s=1}^{T} \gamma_q\left(\hat{\theta}_T^{\{\mathbf{f},\mathcal{S}\}}; (x_s, y_s)\right)$$

[0127] en prenant le contraste quadratique.

[0128] Le prédicteur est construit sous la forme d'une agrégation d'experts, via une stratégie

S

[0129] $$\hat{g}_T^{\{\mathbf{f}^*,S\}}(x) = \sum_{i=1}^{M} \pi_i(x) f_i(x, D_T)$$

[0130] avec

[0131] $$\pi_i(x) = \frac{\mathbf{1}_{\{x \in k_i\}} \pi_i}{\sum_{j=1}^{M} \mathbf{1}_{\{x \in k_j\}} \pi_j},$$

[0132] Le prédicteur peut ainsi s'écrire aussi sous la forme d'un estimateur linéaire de la fonction de régression :

[0133] $$\hat{g}_T^{\{\mathbf{f}^*,S\}}(x) = \sum_{i=1}^{M} \pi_i(x) \sum_{s=1}^{T} \frac{y_s \mathbf{1}_{x_s \in k_i}}{\sum_{u=1}^{T} \mathbf{1}_{x_u \in k_i}} = \sum_{s=1}^{T} w_s(x) y_s$$

[0134] avec

[0135] $$w_s(x) = \sum_{j=1}^{M} \pi_j(x) \frac{\mathbf{1}_{x_s \in k_j}}{\sum_{u=1}^{T} \mathbf{1}_{x_u \in k_j}}$$

[0136] L'on dispose ainsi d'un estimateur de la fonction de régression et d'un prédicteur ayant des performances comparables à celles de la meilleure combinaison convexe du sous ensemble d'experts identifié.

[0137] Le procédé selon l'invention d'apprentissage par ordinateur permet une commande d'un système technique.

[0138] Le système technique est par exemple un système d'alerte à usage médical, signalant un risque pour un patient, au vu de l'analyse de séries temporelles de rythmes cardiaque.

[0139] Le système technique est, dans un autre exemple, un système de trading.

[0140] Le procédé selon l'invention est basé sur l'analyse de séries temporelles sous la forme d'un échantillon de données $D_n = (X_i, Y_i)_{1<=i<=n}$ où pour tout i, X_i est un ensemble de variables explicatives et Y_i une variable d'intérêt.

[0141] Le procédé comprend les étapes suivantes :

[0142] - définition d'une règle testant si une réalisation de X est dans un hyperrectangle de l'espace des variables explicatives ;

[0143] - définition de la complexité de la règle ;

[0144] - discrétisation de l'espace des variables explicatives en M modalités ;

[0145] - recherche récursive sur la complexité des règles jusqu'à une complexité maximale fixée ;

[0146] - sélection d'un sous ensemble de règles avec prédiction supérieure à zéro et d'un

sous ensemble de règles avec prédiction inférieure à zéro, en contrôlant leur chevauchement.

[0147] Une règle est ainsi un objet de type

[0148] $$\text{"Si } X \in r \text{ Alors } \hat{Y} = p_r\text{"}$$

[0149] tel que

[0150] – la condition *Si* teste si une réalisation de X est dans un hyperrectangle de l'espace des variables explicatives

[0151] $$r = \prod_{k=1}^{d} I_k$$

[0152] – l'implication est la valeur prédite par la règle sur la condition est vérifiée, avec

[0153] $$p_r = \frac{\sum_{i=1}^{n} y_i \mathbf{1}_{x_i \in r}}{\sum_{i=1}^{n} \mathbf{1}_{x_i \in r}}$$

[0154] – la complexité d'une règle étant définie par

[0155] $$cp(r) = d - \#\{1 \leq k \leq d ; I_k = \mathcal{X}_k\}$$

[0156] Le procédé comprend avantageusement une détermination de l'acceptabilité d'une règle, cette détermination comprenant les étapes suivantes :

[0157] - calcul de la couverture de la règle ;

[0158] - calcul de la significativité de la règle ;

[0159] - vérification de ce que la couverture de la règle est comprise entre deux valeurs prédéterminées ;

[0160] - vérification de ce que la significativité de la règle est supérieure à une valeur prédéterminée ;

[0161] - calcul d'un gain pénalisé.

[0162] Une règle est acceptable uniquement si la condition de couverture, la condition de significativité, et la condition sur les gains sont vérifiées.

[0163] La condition de significativité peut ainsi être avantageusement exprimée comme suit

[0164] $$\frac{\sqrt{n(r, D_n)}\,|p_r - \bar{y}|}{\hat{\sigma}} \geq \alpha$$

[0165] La condition de couverture peut avantageusement être exprimée comme suit :

[0166] $$c_{min} \leq \frac{n(r, D_n)}{n} \leq c_{max}$$

[0167] La condition sur les gains peut être exprimée comme suit :

[0168]

$$\frac{\Delta}{n}\sum_{j=0}^{n/\Delta-1}\left(\sum_{i=j\Delta+1}^{(j+1)\Delta} y_i \times \rho_r \mathbf{1}_{x_i \in r}\right) - \gamma_r > 0$$

[0169] où Δ est une période fixée et γ_r une pénalisation dépendante de la règle.

[0170] Les algorithmes selon l'invention exploitent ainsi la structure temporelle des données, au moyen d'une fonction de gain pénalisée.

[0171] La fonction de gain pénalisée combine avantageusement l'espérance conditionnelle des règles, leur fréquence d'occurrence et une mesure de la régularité spectrale des activations des règles.

[0172] L'invention permet ainsi de prendre en compte l'intensité et la fréquence des signaux, pour la recherche d'évènements rares à forte intensité et la recherche de signaux faibles et récurrents.

[0173] Une application de l'invention est l'extraction de signaux prédictifs à partir de données extra financières portant sur des entreprises, par exemple les notations au regard de la politique RSE des organisations.

[0174] Le procédé selon l'invention permet de montrer l'existence d'un lien entre performance financière et critères extra financiers de type ESG.

[0175] Le procédé selon l'invention est ainsi avantageusement utilisé dans l'extraction de signaux prédictifs pour la gestion d'actifs.

[0176] Avantageusement, la mise en œuvre du procédé est effectuée en plusieurs étapes. Dans une première étape, une première preuve de concept est effectuée sur la base d'une simulation de portefeuilles. Dans une deuxième étape, une extension de la preuve de concept est effectuée, à des données de notes d'analystes sur les sociétés, ces données présentant une qualité très supérieure aux données brutes issues du web ou des réseaux sociaux. Dans une troisième étape, un module de visualisation est créé, présentant les indications données par l'algorithme d'apprentissage et s'appuyant sur un algorithme de recherche de configurations proches d'une configuration donnée, dans une base historique.

[0177] Une autre application de l'invention est l'extraction de signaux prédictifs à partir de données portant sur des patients, par exemple en service de réanimation. Les données sont par exemple le rythme cardiaque.

[0178] Lorsque des décisions médicales sont prises sur la base de prévisions fournies par des algorithmes, les exigences des assureurs et les attentes des familles ne sont pas compatibles avec un fonctionnement de type boite noire, dans lequel aucune indication ne peut être trouvée sur les variables ayant participé à une prédiction.

[0179] L'invention fournit avantageusement des commentaires explicatifs associés à la classification des données médicales.

[0180] *Exemple comparatif*

[0181] Les performances du procédé selon l'invention vont être présentées en comparaison avec celles d'un « procédé alternatif » issu de l'état de la technique.

[0182] Plus précisément, un ensemble de données numériques massives a fait l'objet d'un traitement par apprentissage supervisé selon l'invention, et d'un traitement à l'aide d'un procédé d'apprentissage supervisé utilisant des moyens de l'état de la technique (« procédé alternatif »), ces moyens étant présents dans des librairies publiques d'apprentissage supervisé.

[0183] Il va être présenté ci-dessous un procédé technique de classification de données selon l'invention, opérant sur un ordinateur de bureau à ressources limitées en mémoire vive. Le procédé selon l'invention fournit des commentaires explicatifs associés à la classification d'observations numériques.

[0184] *Ensemble de données traitées lors de la mise en œuvre de l'exemple comparatif*

[0185] L'ensemble de données est constitué d'une variable d'intérêt Y à prédire et de covariables X^i ($i \in 1\ldots V$). Y et chacun des X^i sont décrites par un ensemble d'instances indexées par :

[0186]
- un ensemble d'individus I_k, $k \in 1..K$.
- un ensemble d'occurrences T_l, $l \in 1..L$. Les occurrences sont ordonnées selon les relations $T_1 < T_2 < ...T_L$

[0187] Chaque observation d'une variable X^i se note donc

[0188] $X^i_{k,l}$

[0189] et correspond à la mesure de l'attribut X^i effectuée sur l'individu I_k lors du relevé d'occurrence T_l.

[0190] De même, l'observation correspondante de la variable Y est notée $Y_{k,l}$.

[0191] Les variables prennent des valeurs dans

[0192] $\mathbb{R} \bigcup \{\mathrm{nan}\}$

[0193] nan étant une valeur non numérique attribuées aux valeurs non renseignées.

[0194] *Cahier des charges de l'essai comparatif*

[0195] Afin de mener la comparaison à parité, les deux procédés sont soumis au même cahier des charges.

[0196] Les données d'entrée sont :

[0197]
- un ensemble de covariables X^i ($i \in 1\ldots V$) et une variable d'intérêt Y, classifiable en M classes. La donnée des covariables X^i ($i \in 1\ldots V$) et de Y constitue l'ensemble d'apprentissage ;
- une consigne de complexité maximale. La valeur de 2 a été retenue pour cet exemple. Une complexité K implique de vérifier 2xk conditions sur les variables ;

- une condition de significativité statistique minimum ;
- une condition de couverture relative minimum. La valeur de 5% a été retenue dans cet exemple ;
- une condition de taux d'intersection maximum entre deux règles. La valeur de 80% a été retenue pour cet exemple.

[0198] Les données de sortie : fournir une liste de règles *Rj* satisfaisant aux quatre conditions présentées ci-dessus (consigne de complexité maximale, condition de significativité statistique minimum, condition de couverture relative minimum, condition de taux d'intersection maximum entre deux règles). Ces règles doivent s'appliquer à l'ensemble des données, et n'être spécifiques ni à un individu particulier, ni à une occurrence particulière.

[0199] Les objectifs sont :

[0200]
- consommer en cours d'exécution une quantité de mémoire vive maximale indépendante du nombre de variables ;
- maximiser la moyenne de la profondeur explicative sur l'ensemble des exemples d'apprentissage.

[0201] *Définitions utilisées*

[0202] Une règle R_j est définie par

[0203] – des conditions portant sur k co-variables

[0204] $$X^{i_1} \in [a_1, b_1] \wedge ... \wedge X^{i_v} \in [a_v, b_v]$$

[0205] – une affectation de classification

[0206] $$Y \in \hat{Y}^i$$

[0207] associée aux conditions.

[0208] La complexité $c(R_j)$ est le nombre v de co-variables présentes dans les conditions de la règle.

[0209] L'ensemble d'activation Act (R) d'une règle R est l'ensemble des paires

[0210] $$(k, l) \in K \times L$$

[0211] vérifiant la règle

[0212] $$Act(R) = \{(k,l) \in K \times L | X_{k,l}^{i_1} \in [a_1, b_1] \wedge ... \wedge X_{k,l}^{i_v} \in [a_v, b_v]\}$$

[0213] Par convention, on a

[0214] $$\mathbf{nan} \notin [a, b]$$

[0215] ce qui implique qu'une observation manquante d'une variable ne peut jamais appartenir à l'ensemble d'activation d'une règle contenant cette variable.

[0216] La couverture cov(R) d'une règle est définie par

[0217] $$cov(R) = card(Act(R))$$

[0218] La couverture relative rcov(R) d'une règle est définie par

[0219] $$rcov(R) = cov(R)/card(K \times L)$$

[0220] Le taux d'intersection entre deux règles R et R' est défini par

[0221] $$int\_rate(R, R') = card(Act(R) \cap Act(R'))/min(cov(R), cov(R'))$$

[0222] Le contexte explicatif *Expl* (*i*,*k*) d'une observation indexée

[0223] $$(k, l) \in K \times L$$

[0224] est l'ensemble des observations défini par

[0225] $$Expl(i, k) = \{j | (i, k) \in Act(R_j)\}$$

[0226] Il s'agit donc de l'ensemble des règles qui englobent une observation donnée dans leur ensemble d'activation.

[0227] La profondeur explicative *expl_d*(*i*,*k*) d'une observation indexée est définie par la taille du contexte explicatif de ce point, soit

[0228] $$expl\_d(i, k) = card(Expl(i, k))$$

[0229] *Le jeu de données utilisé lors de la mise en œuvre de l'exemple comparatif*

[0230] Les données correspondent à V=1009 attributs numériques X^i concernant K=657 individus selon L=1850 occurrences, et un vecteur de résultats Y également renseigné pour les K individus et les L occurrences.

[0231] Le nombre total de cellules est donc de $(V+1)KL$=1.23 10^9.

[0232] L'occupation en mémoire vive d'un tel ensemble dépend de la représentation des nombres flottants dans le langage utilisé. Dans le cas du langage Python, utilisé dans cet exemple comparatif, les nombres flottants sont codés sur 8 octets (64 bits). La taille de la matrice totale en mémoire est donc de 9,8 GO.

[0233] Afin que les règles conservent un caractère interprétable, les attributs numériques

[0234] $$X^i_{k,l}$$

[0235] sont discrétisés en cinq modalités.

[0236] *Imputation des valeurs manquantes*

[0237] Les modules d'arbres de décision de Scikit-Learn ne permettent pas de traiter les valeurs manquantes dans les variables. Si, pour un indice

[0238] $$T_l, l \in 1..L$$

[0239] et un individu I_k, on a

[0240] $$X^i_{k,l} = \mathrm{nan}$$

[0241] c'est à dire si l'attribut X^i n'est pas renseigné pour l'occurrence T_l de l'individu I_k, alors aucune comparaison ne pourra être faite sur cet attribut et l'ensemble des occurrences de cet attribut pour l'individu I_k sera ignoré.

[0242] Le procédé selon l'invention rejette uniquement les occurrences non renseignées d'un attribut lorsqu'elles interviennent dans une règle.

[0243] Pour permettre une comparaison, il est donc nécessaire de définir une stratégie d'imputation des valeurs manquantes, dans le procédé alternatif.

[0244] La stratégie d'imputation des valeurs manquantes définie dans le procédé alternatif est la suivante. Pour un individu I_k et un attribut X^i, si l'occurrence T_l est manquante,

[0245] $$X^i_{k,l} = \mathbf{nan}$$

[0246] alors on effectue l'imputation selon la dernière occurrence connue:

[0247] $$X^i_{k,l} \longleftarrow X^i_{k,l*} \text{ avec } l* = Max\{m | 0 \leq m < l \wedge X^i_{k,m} \neq \mathbf{nan}\}$$

[0248] Si aucune occurrence précédente n'est renseignée, alors l'imputation est effectuée selon la valeur moyenne pour l'occurrence 0, des observations prises pour tous les individus possédant un attribut

[0249] $$X^i_{k,0}$$

[0250] différent de nan. Par convention, la moyenne d'une observation sur un ensemble vide d'individus est fixée à zéro.

[0251] *Mise en œuvre du procédé selon l'invention*

[0252] Le procédé selon l'invention est mis en œuvre sur un ordinateur de bureau.

[0253] Les données d'entrées comprennent :

[0254]
- les fichiers contenant les variables X^i. Ces fichiers peuvent se trouver dans des répertoires, des lecteurs réseau ou des périphériques séparés;
- le fichier contenant les observations de la variable Y;
- un fichier de consignes d'entrées.

[0255] Les consignes d'entrée comprennent:

[0256]
- l'information relative à la localisation physique des fichiers,
- le nombre maximum de variables utilisées dans une règle (complexité), fixée à deux pour cet exemple;
- le taux d'intersection maximum entre deux règles, fixé à 0,8 pour cet exemple;
- le nombre de modalités selon lequel les variables X^i sont discrétisées, fixé à cinq pour cet exemple;
- des seuils de significativité statistique pour la rétention des règles.

[0257] Les données de sortie sont un fichier contenant la description des règles retenues par l'algorithme.

[0258] Chaque rège est décrite par des bornes portant sur deux co-variables:

[0259] $$X^{i_1} \in [a_1, b_1], X^{i_2} \in [a_2, b_2]$$

[0260] Un exemple de transcription de règle en langage naturel est donné dans le tableau ci-

dessous. Chaque règle comporte deux variables. Les bornes respectives des variables sont Bmin et Bmax. La transcription en langue naturelle de la règle est donnée dans la colonne “description”.

[0261] [Tableaux1]

Rule Id	Features_Names	BMin	BMax	Description
R_0(2)+	['FIN_corr260d_GavSec_Consumer Durables', 'Esg_NBySector_HR1.1_SCORE']	[3.0, 3.0]	[4.0, 4.0]	WHEN FIN_corr260d_GavSec_Consumer Durables is very high AND Esg_NBySector_HR1.1_SCORE is very high
R_1(2)+	['Esg_NByUniver_ENV_I-SCORE', 'FIN_DevCorr260dFinancials_ByMc']	[4.0, 2.0]	[4.0, 4.0]	WHEN Esg_NByUniver_ENV_I-SCORE is at the maximum AND FIN_DevCorr260dFinancials_ByMc is not low

[0262] *Mise en œuvre du procédé alternatif*

[0263] Le procédé alternatif de comparaison utilise les arbres de décision.

[0264] Dans cet exemple comparatif, le même jeu de données est utilisé.

[0265] Les modules CART (*Classification and regression Tree*) de la librairie publique Scikit-Learn sont utilisés.

[0266] Comme pour les autres modules d’apprentissage supervisé de l’état de la technique, le module CART s’exécute en chargeant en mémoire l’ensemble des données. Cette exigence rendrait la mise en œuvre potentiellement inopérante pour un nombre de variable élevé.

[0267] Pour contourner cette difficulté, un arbre de décision est construit pour chaque individu, soit 657 arbres pour le jeu de données utilisé.

[0268] Le nombre de cellules chargées simultanément est ainsi limité à l’ensemble des observations relatives à cet individu, dont la taille est VL=1866650.

[0269] La boucle suivante est mise en œuvre.

[0270] Pour chaque individu X^i :

[0271]
- la construction d’un arbre à partir des variables de l’individu X^i est effectuée par appel au module CART;
- les caractéristiques de cet arbre sont sauvegardées sur disque;
- la mémoire est vidée pour le prochain arbre.

[0272] La profondeur d’un arbre est fixée à 4. Cette profondeur est à parité avec la complexité maximale de 2, fixée comme consigne pour le procédé selon l’invention.

[0273] *Transformation des arbres en règles*

[0274] Un arbre de décision correspond à une procédure dichotomique de classification d’individus, base sur la comparaison de certaines co-variables à des seuils.

[0275] L’ensemble des résultats possibles de comparaisons entre les variables et les seuils correspondants constitue une branche de l’arbre.

[0276] L’ensemble des individus identifiés par les conditions d’une branche constitue une feuille de l’arbre.

[0277] Ces conditions sont directement exprimables sous forme d’une règle, dont les variables ne sont autres que les variables intervenant le long de la branche.

[0278] La complexité de la règle obtenue est égale au nombre de variables.

[0279] Les feuilles de l'arbre fournissent une partition de l'ensemble des observations au moyen de règles.

[0280] La transformation en règles des arbres construits à l'étape précédente donne environ 13000 règles.

[0281] *Vérification des règles*

[0282] Les règles obtenues à l'étape précédente sont chacune spécifiques à un individu, alors que le cahier des charges impose de fournir des règles générales valables pour l'ensemble des individus.

[0283] Il est donc nécessaire de filtrer les règles obtenues et de ne conserver que celles qui satisfont le critère de significativité statistique pour l'ensemble des individus.

[0284] Cette étape peut être complétée au moyen d'une boucle sur les individus, en ne chargeant que les variables intervenant dans la règle.

[0285] La vérification de la significativité des règles se fait dans l'ensemble d'apprentissage.

[0286] Seules les règles passant le test de significativité sur l'ensemble des individus sont conservées. Un total de 2725 règles et alors obtenu, contre environ 13000 avant filtrage. Cette étape de vérification prend environ neuf heures.

[0287] *Elimination des règles redondantes*

[0288] Après la vérification, les règles sont filtrées. En construisant un arbre par individu, il se peut qu'une même règle apparaisse dans différents arbres. L'objectif étant de produire un ensemble de règles s'appliquant uniformément à l'ensemble des individus, le procédé alternatif élimine les règles syntaxiquement identiques, définies par des conditions identiques sur les mêmes variables.

[0289] Dans une étape suivante, afin de réduire le nombre de règles, un second filtrage est effectué.

[0290] Ce second filtrage vise à éliminer les règles ayant des ensembles d'activation trop proches. Cette étape élargit l'élimination des règles identiques à celles de règles simplement similaires.

[0291] La figure 1 illustre la stratégie appliquée.

[0292] Dans une première étape, les règles sont triées selon un critère de qualité métier variant en fonction de l'application. La règle ayant le meilleur critère est automatiquement conservée.

[0293] Dans une deuxième étape, le nombre de points qui activent la règle (r_1) est calculé, ainsi que le nombre de points qui activent la seconde règle (r_2) et le nombre de points activant les deux règles (r_1&r_2). Ce nombre de points en commun ne doit pas dépasser 80% de r_1 ou r_2, pour que la seconde règle soit conservée. Cette condition correspond à la première ligne de la figure 1.

[0294] Les règles suivantes sont traitées de la même manière, la différence étant que la règle

r_1 est remplacée par le nombre de points activant au moins une des règles déjà sélectionnées. Ceci correspond à la deuxième ligne de la figure 1.

[0295] Cette étape de filtrage retient finalement 9 règles sur les 2725 de l'étape précédente.

[0296] *Comparaison des résultats obtenus par le procédé selon l'invention et le procédé alternatif*

[0297] Le procédé selon l'invention produit 27 règles de complexité 2 sur l'ensemble de données. La consommation maximum de mémoire au cours de l'exécution est de 15,2 GO, quel que soit le nombre de variables.

[0298] Le procédé alternatif produit 9 règles de complexité 4 sur l'ensemble de données. La consommation maximum de mémoire au cours de l'exécution est de 9,6 GO, proportionnelle au nombre de variables.

[0299] Les performances comparées du procédé selon l'invention et selon le procédé alternatif peuvent être détaillées, en répétant l'essai comparatif avec différentes valeurs du nombre de variables V, du nombre d'individus X et du nombre d'occurrences.

[0300] Le tableau ci-dessous présente la consommation maximale de mémoire vive (RAM) durant l'exécution du procédé selon l'invention et du procédé alternatif.

[0301] [Tableaux2]

Consommation maximale de mémoire vive en cours de calcul	X=10	X=100	X=657	X=657	X
	V=1009	V=1009	V=1009	V=4000	V
	L=100	L=1850	L=1850	L=1850	L
Procédé selon l'invention	2.1 GO	2.3 GO	15.2 GO	15.2GO	$C_1 + C_2 \times L \times X$
Procédé alternatif	2.4 GO	9.6 GO	9.6 GO	38.4 GO	$C_3 + C_4 \times V \times L \times X$

[0302] Dans ce tableau, C_1, C_2, C_3 et C_4 sont des constantes dont les valeurs sont estimées à

[0303] $C_1=C_3=2,0$GO (quantités de mémoire minimales pour mettre en œuvre les procédés)

[0304] $C_2=8,2\ 10^{-7}$ et $C_4=5,4\ 10^{-7}$ (constantes de proportionnalité respectives des deux procédés, par rapport à la taille du problème posé). `

[0305] Comme le montre le tableau ci-dessus, le procédé selon l'invention présente l'avantage d'une consommation maximale de mémoire vive indépendante du nombre de variables.

[0306] Le procédé alternatif, comme tout procédé issu de librairies standards qui exige le chargement de l'ensemble des données simultanément, présente une consommation de mémoire vive linéairement croissante en fonction du nombre de variables.

[0307] Dans les deux procédés, il existe une dépendance linéaire par rapport au nombre d'individus de la base (nombre d'individus X et nombre d'occurrences L). Cette dépendance est conventionnellement contournée, pour un nombre d'occurrences très élevé, par la mise en œuvre de techniques de type « map reduce » opérant sur des données distribuées.

[0308] Les figures 2 et 3 permettent de comparer la profondeur explicative entre les règles produites par le procédé selon l'invention (figure 2) et selon le procédé alternatif (figure 3). La profondeur explicative correspond au nombre moyen de règles activées par individu.

[0309] Sur les figures 2 et 3, l'axe des abscisses représente la date d'occurrence du relevé des attributs des individus X, groupés par année. L'axe des ordonnées représente les individus regroupés selon quatre familles. La nuance de gris, sur l'échelle de droite du graphe, représente la profondeur explicative moyenne de chaque groupe, c'est à dire le nombre moyen de règles actives pour chaque occurrence.

[0310] La profondeur explicative du procédé selon l'invention apparaît deux fois plus élevée que celle du procédé alternatif.

[0311] Ceci s'explique notamment par le fait que le procédé selon l'invention recherche directement des règles statistiquement significatives sur l'ensemble des individus. Cette caractéristique est rendue possible par la capacité de l'algorithme de parcourir l'espace de recherche sans charger l'ensemble des variables.

[0312] Le procédé alternatif, comme tout procédé construit à partir d'une librairie d'apprentissage telle que Scikit-Learn, contraint à rechercher initialement des règles localement valables pour un individu seulement. La généralisation à l'ensemble des individus n'intervient que dans un second temps. Cette généralisation fournit in fine moins de règles que le procédé selon l'invention, car une règle optimisée pour un individu subit en moyenne une perte de significativité élevée lorsqu'on temps de la généraliser à l'ensemble des individus, ce mécanisme expliquant le taux de perte de 13000 à 9 règles dans le procédé alternatif.

[0313] Avantages des algorithmes selon l'invention

[0314] Comme montré dans l'exemple comparatif, l'invention permet de fournir des contextes explicatifs à partir d'un ensemble arbitraire de variables structurées, non nécessairement situées dans un unique fichier. Ces variables peuvent contenir des données manquantes.

[0315] Le procédé selon l'invention consomme une quantité de mémoire vive indépendante du nombre de variables. L'exemple comparatif montre que le procédé de l'invention peut être mis en œuvre sur un ordinateur de bureau doté d'une mémoire vive de 16 GO.

[0316] Le procédé selon l'invention peut fonctionner sur une machine de bureau en

consommant des ressources de mémoire vive indépendantes du nombre de variables descriptives présentes dans le problème de classification.

[0317] Le procédé selon l'invention peut opérer sur des données contenues dans des fichiers situés sur des supports séparés.

[0318] Si l'on cherche à obtenir des résultats analogues en utilisant des algorithmes d'apprentissage supervisés disponibles dans les librairies publiques, un obstacle technique se présente, les librairies publiques exigeant que les données soient rassemblées dans un fichier ou matrice unique. La taille de cette matrice en mémoire vive peut facilement excéder les possibilités d'une machine ordinaire, en particulier si le nombre de variables est important.

[0319] Le procédé de classification de données selon l'invention est capable de fournir des éléments d'explication sur la façon dont chaque donnée est classifiée. Les commentaires explicatifs sont exprimés comme des conditions simples portant sur les variables retenues par les utilisateurs pour la classification.

[0320] Ces commentaires explicatifs ont la forme de règles d'association du type « Si condition 1 et condition 2 et … conditio n, Alors la variable d'intérêt appartient à la classe K ».

[0321] Les algorithmes d'apprentissage supervisé qui viennent d'être décrits permettent d'expliquer une performance, à partir d'un échantillon de co-variables.

[0322] Les algorithmes selon l'invention présentent de nombreux avantages :

[0323]
- capacité à traiter un grand nombre de co-variables ;
- tolérance vis-à-vis de données manquantes, contrairement aux machines à vecteur support (SVM), et aux régressions linéaires ;
- rendre compte d'effets de seuils sur des variables, contrairement aux machines à vecteur support (SVM), aux régressions linéaires, et aux approches topologiques de type plus proche voisin ;
- rendre compte de dépendances non linéaires ;
- traçabilité et parcimonie du modèle prédictif ;
- pas d'hypothèse sur la distribution statistique des variables ;
- pas d'à priori sur une hiérarchie inter variables, contrairement aux arbres de décision ;
- évolutivité en fonction de nouvelles données ;
- fournir des prédicteurs concurrents et partiellement corrélés.

[0324] Les algorithmes d'apprentissage selon l'invention sont déterministes et interprétables par tous, contrairement aux machines à vecteur support, aux forêts aléatoires et aux réseaux de neurones.

[0325] Par interprétable, on souligne ici qu'une personne peut comprendre la logique ayant conduit à la prédiction fournie par l'algorithme.

[0326] Les algorithmes selon l'invention sont adaptés à des données qualitatives et quantitatives.

[0327] L'agrégation d'experts fournit un prédicteur dont les performances sont comparables à celle de la meilleure combinaison convexe.

[0328] La construction du prédicteur permet de l'exprimer comme un estimateur de la fonction de régression.

[0329] Les algorithmes selon l'invention permettent d'éviter les biais de jugement, et permettent une synthèse de signaux contradictoires.

[0330] Les algorithmes selon l'invention permettent d'extraire les bons signaux d'une masse de données pour enrichir la variété des données, incorporer des indicateurs propriétaires.

[0331] Les algorithmes selon l'invention ne fonctionnent pas en boite noire, et leur fonctionnement peut être expliqué par les utilisateurs. Ils permettent une représentation des tendances et de leurs intermittences.

[0332] L'utilisation des algorithmes selon l'invention permet de fournir des prévisions interprétables.

[0333] Cette performance est avantageuse dans de nombreux secteurs.

[0334] En effet, par exemple, lorsque des décisions d'investissement sont prises sur la base de prévisions fournies par des algorithmes, les exigences réglementaires de traçabilité des décisions et de suivi des risques ne sont pas compatibles avec un fonctionnement de type boite noire, dans lequel aucune indication ne peut être trouvée sur les variables ayant participé à une prédiction.

[0335] Dans la plupart des problèmes de prédiction rencontrés, en particulier dans des situations industrielles, médicales ou environnementales, le nombre d'individus et d'occurrences est une donnée fixe (taille d'une banque de données d'images, de parcours clients, ou de caractéristiques biologiques de patients). La différence en matière de pouvoir prédictif entre plusieurs algorithmes se fait sur la capacité à créer de nouvelles variables X^i adaptées au problème posé. Cette étape de « features engineering » doit pouvoir être menée avec aussi peu de contraintes que possible sur le nombre de variables présentées à l'algorithme. Dans ce contexte, la capacité technique de l'invention à fonctionner en utilisant une quantité de mémoire indépendante du nombre de variables est particulièrement avantageuse.

Revendications

[Revendication 1] Procédé technique de classification de données apte à être mis en œuvre sur un ordinateur de bureau, le procédé exploitant des données d'entrée d'un ensemble d'apprentissage comprenant :
- des co-variables X^i explicatives, décrites par un ensemble d'instances indexées par un ensemble d'individus I_k et un ensemble d'occurrence T_l ;
- les observations d'une variable Y d'intérêt ;
caractérisé en ce que les données des co-variables X^i explicatives sont contenues dans des fichiers distincts, le procédé comprend les étapes suivantes :

– définition d'une règle testant si une réalisation de X est dans un hyperrectangle de l'espace des variables explicatives ;
– définition de la complexité de la règle ;
– discrétisation de l'espace des variables explicatives en M modalités ;
– recherche récursive sur la complexité des règles jusqu'à une complexité maximale fixée ;
– sélection d'un sous ensemble de règles avec prédiction supérieure à zéro et d'un sous ensemble de règles avec prédiction inférieure à zéro, en contrôlant leur chevauchement.

[Revendication 2] Procédé selon la revendication 1, caractérisé en ce qu'il comprend une détermination de l'acceptabilité d'une règle, cette détermination comprenant les étapes suivantes :
- calcul de la couverture de la règle ;
- calcul de la significativité de la règle ;
- vérification de ce que la couverture de la règle est comprise entre deux valeurs prédéterminées ;
- vérification de ce que la significativité de la règle est supérieure à une valeur prédéterminée ;
- calcul d'un gain pénalisé.

[Revendication 3] Procédé selon la revendication 2, caractérisé en ce qu'il comprend une étape de vérification de ce qu'une condition sur le gain pénalisé est vérifiée.

[Revendication 4] Procédé selon la revendication 3, caractérisé en ce que la condition sur

les gains est exprimée comme suit :

$$\frac{\Delta}{n}\sum_{j=0}^{n/\Delta-1}\left(\sum_{i=j\Delta+1}^{(j+1)\Delta} y_i \times \rho_r 1_{x_i \in r}\right) - \gamma_r > 0$$

où Δ est une période fixée et γ_r une pénalisation dépendante de la règle.

[Revendication 5] Procédé selon l'une quelconque des revendications précédentes, caractérisé en ce qu'il est mis en œuvre sur un ordinateur de bureau dont la mémoire vive est d'une capacité inférieure à 20 GO.

[Revendication 6] Procédé d'apprentissage par ordinateur d'une commande d'un système technique, le procédé mettant en œuvre une classification technique de données selon l'une des revendications 1 à 5, le procédé d'apprentissage étant basé sur des séries temporelles sous la forme d'un échantillon de données $D_n=(X_i, Y_i)_{1<=i<=n}$ où pour tout i, X_i est un ensemble de variables explicatives et Y_i une variable d'intérêt.

[Revendication 7] Support lisible par ordinateur sur lequel sont stockées des instructions lisibles par machine pour exécuter un procédé selon l'une quelconque des revendications précédentes.

[Fig. 1]

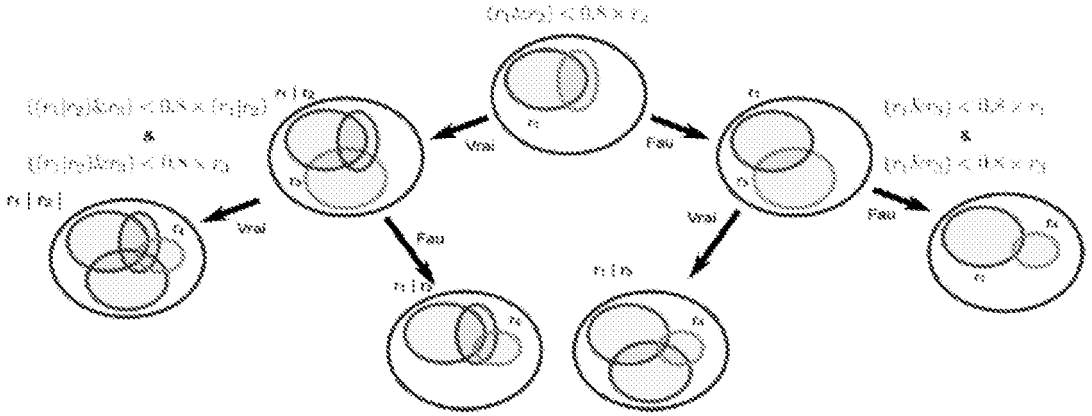

Fig. 1

[Fig. 1]

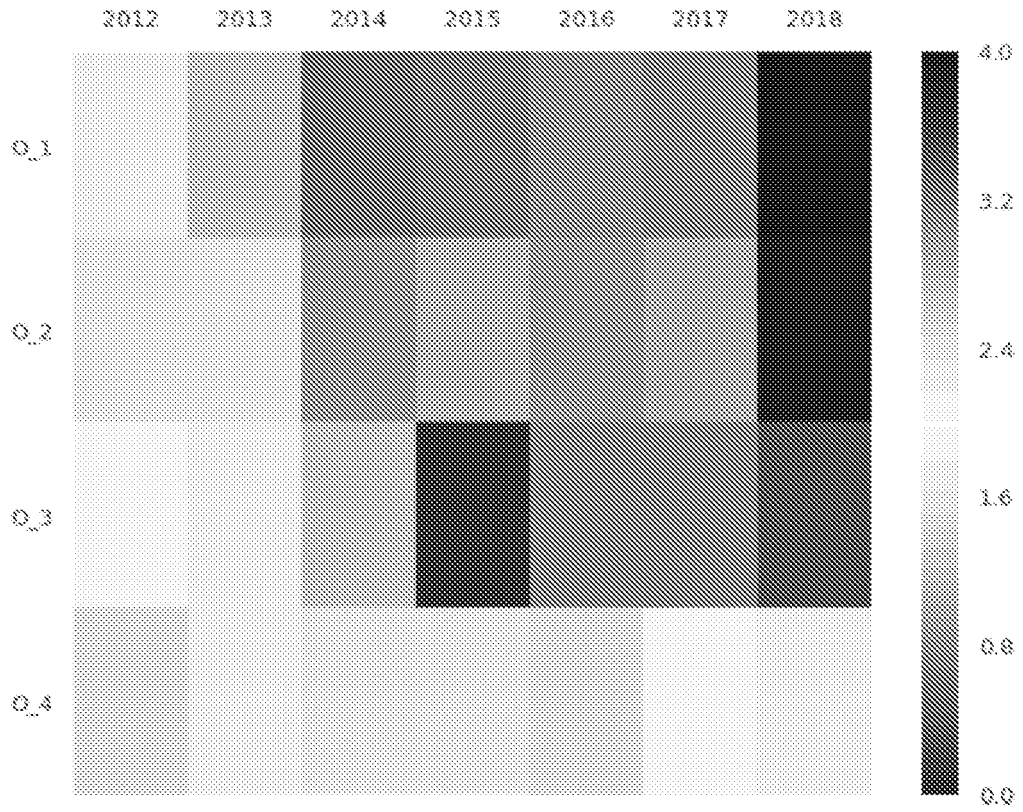

Fig.2

[Fig. 2]

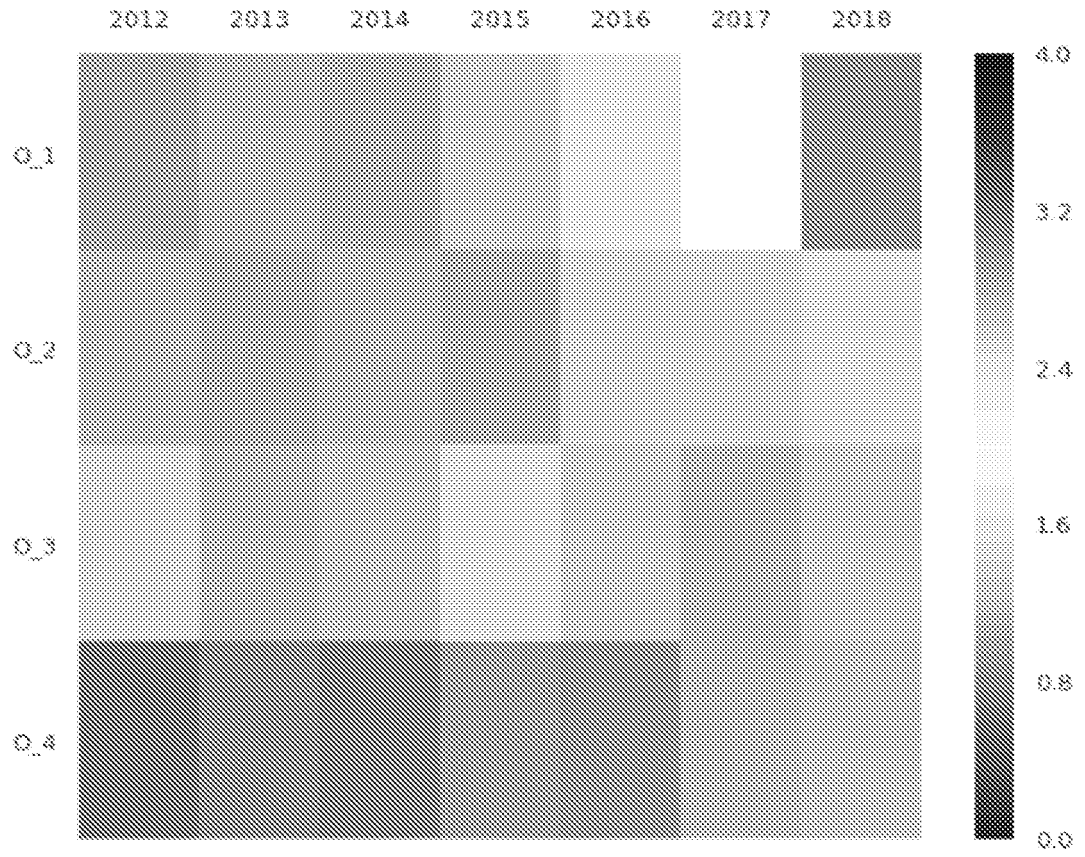

Fig. 3

RAPPORT DE RECHERCHE

articles L.612-14, L.612-53 à 69 du code de la propriété intellectuelle

OBJET DU RAPPORT DE RECHERCHE

L'I.N.P.I. annexe à chaque brevet un "RAPPORT DE RECHERCHE" citant les éléments de l'état de la technique qui peuvent être pris en considération pour apprécier la brevetabilité de l'invention, au sens des articles L. 611-11 (nouveauté) et L. 611-14 (activité inventive) du code de la propriété intellectuelle. Ce rapport porte sur les revendications du brevet qui définissent l'objet de l'invention et délimitent l'étendue de la protection.
Après délivrance, l'I.N.P.I. peut, à la requête de toute personne intéressée, formuler un "AVIS DOCUMENTAIRE" sur la base des documents cités dans ce rapport de recherche et de tout autre document que le requérant souhaite voir prendre en considération.

CONDITIONS D'ETABLISSEMENT DU PRESENT RAPPORT DE RECHERCHE

[x] Le demandeur a présenté des observations en réponse au rapport de recherche préliminaire.

[x] Le demandeur a maintenu les revendications.

☐ Le demandeur a modifié les revendications.

☐ Le demandeur a modifié la description pour en éliminer les éléments qui n'étaient plus en concordance avec les nouvelles revendications.

☐ Les tiers ont présenté des observations après publication du rapport de recherche préliminaire.

☐ Un rapport de recherche préliminaire complémentaire a été établi.

DOCUMENTS CITES DANS LE PRESENT RAPPORT DE RECHERCHE

La répartition des documents entre les rubriques 1, 2 et 3 tient compte, le cas échéant, des revendications déposées en dernier lieu et/ou des observations présentées.

[x] Les documents énumérés à la rubrique 1 ci-après sont susceptibles d'être pris en considération pour apprécier la brevetabilité de l'invention.

[x] Les documents énumérés à la rubrique 2 ci-après illustrent l'arrière-plan technologique général.

☐ Les documents énumérés à la rubrique 3 ci-après ont été cités en cours de procédure, mais leur pertinence dépend de la validité des priorités revendiquées.

☐ Aucun document n'a été cité en cours de procédure.

1. ELEMENTS DE L'ETAT DE LA TECHNIQUE SUSCEPTIBLES D'ETRE PRIS EN CONSIDERATION POUR APPRECIER LA BREVETABILITE DE L'INVENTION

US 2019/379589 A1 (RYAN SID [CA] ET AL)
12 décembre 2019 (2019-12-12)

US 2016/210552 A1 (KASABOV NIKOLA KIRILOV [NZ] ET AL) 21 juillet 2016 (2016-07-21)

2. ELEMENTS DE L'ETAT DE LA TECHNIQUE ILLUSTRANT L'ARRIERE-PLAN TECHNOLOGIQUE GENERAL

FR 3 069 357 A1 (WORLDLINE [FR])
25 janvier 2019 (2019-01-25)

3. ELEMENTS DE L'ETAT DE LA TECHNIQUE DONT LA PERTINENCE DEPEND DE LA VALIDITE DES PRIORITES

NEANT