

12 DEMANDE DE BREVET D'INVENTION A1

22 Date de dépôt : 15.11.23.

30 Priorité :

43 Date de mise à la disposition du public de la
demande : 16.05.25 Bulletin 25/20.

56 Liste des documents cités dans le rapport de
recherche préliminaire : *Se reporter à la fin du
présent fascicule*

60 Références à d'autres documents nationaux
apparentés :

○ Demande(s) d'extension :

71 Demandeur(s) : THALES Société anonyme — FR.

72 Inventeur(s) : HÉLIOU Alice, THOUVENOT Vincent,
MORISSE Baptiste, LAMPE Rodolphe et HUYNH Cong
Bang.

73 Titulaire(s) : THALES Société anonyme.

74 Mandataire(s) : Lavoix.

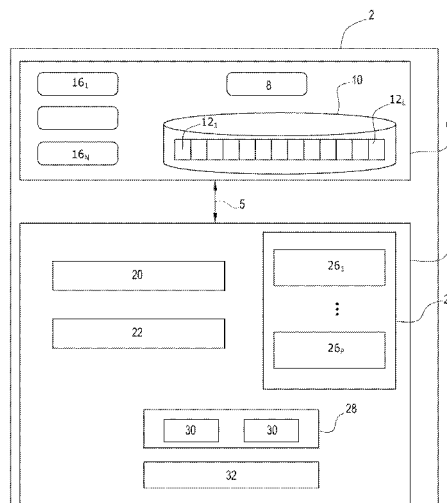
54 Procédé et dispositif de détermination d'appartenance aux données d'apprentissage utilisées pour
l'entraînement d'un modèle de classification de données par apprentissage machine.

57 Procédé et dispositif de détermination d'appartenance
aux données d'apprentissage utilisées pour l'entraînement
d'un modèle de classification de données par apprentissage
machine

Ce dispositif de détermination d'appartenance aux don-
nées d'apprentissage utilisées pour l'entraînement d'un mo-
dèle cible (8) de classification met en œuvre, sur un
ensemble (10) de données à traiter, des modules de :

- génération (20) d'un nombre N supérieur à 2 de parti-
tions distinctes de l'ensemble de données à traiter (10) en
un premier et un deuxième sous-ensembles de données,
- calcul (22) par apprentissage machine de N modèles
témoins (16₁,...,16_N), de même structure que ledit modèle
cible (8);
- apprentissage machine (24) sur un premier sous-en-
semble de modèles témoins d'au moins deux modèles d'in-
férence d'appartenance (26₁,..., 26_P) distincts fournissant
chacun, pour chaque donnée à traiter, une prédiction d'ap-
partenance aux données d'apprentissage du modèle cible,
- apprentissage machine, sur un deuxième sous-en-
semble de modèles témoins, des paramètres d'au moins un
modèle de décision (24) prenant en entrée lesdites prédi-
ctions d'appartenance et fournissant en sortie une prédiction

d'appartenance consolidée.
Figure pour l'abrégé: Figure 1



Description

Titre de l'invention : Procédé et dispositif de détermination d'appartenance aux données d'apprentissage utilisées pour l'entraînement d'un modèle de classification de données par apprentissage machine

- [0001] La présente invention concerne un procédé de détermination d'appartenance aux données d'apprentissage utilisées pour l'entraînement d'un modèle de classification de données par apprentissage machine, dit modèle cible.
- [0002] L'invention concerne également un dispositif associé et un programme d'ordinateur associé.
- [0003] L'invention se situe dans le domaine de la sécurité pour des applications utilisant l'intelligence artificielle.
- [0004] De nombreux systèmes appliqués, par exemple dans le domaine industriel, médical ou militaire, utilisent des modèles de classification automatique des données, par exemple des modèles basés sur des réseaux de neurones artificiels, qui sont entraînés par apprentissage machine.
- [0005] Ce type de modèle de classification comporte un très large nombre de paramètres, dont les valeurs sont calculées et mises à jour de manière dynamique. Il est nécessaire d'apprendre les valeurs des paramètres définissant un modèle de classification, l'apprentissage (également appelée entraînement) des paramètres s'effectuant dans une phase d'apprentissage sur des quantités de données très importantes. De manière classique, de tels modèles de classification prennent en entrée des données d'entrée, par exemple sous forme vectorielle ou matricielle, de type connu, et fournissent en sortie un label de classification (ou étiquette de classification) des données.
- [0006] Par exemple, les données d'entrée sont des images, de dimensions prédéterminées, et les labels de classification représentent des classes prédéterminées, par exemple représentatives de types d'objet présent dans la scène représentée par l'image, selon l'application envisagée.
- [0007] Les données d'entrée utilisées pour l'apprentissage, appelées données d'apprentissage, sont issues soit de bases de données publiques, soit de bases de données privées. Pour certaines applications nécessitant un certain niveau de sécurité, par exemple des applications médicales, ou militaires, il est critique de protéger les données d'apprentissage, et en particulier de s'assurer qu'il n'est pas possible d'obtenir des informations relatives aux données d'apprentissage à partir d'un modèle de classification entraîné.

- [0008] Des publications récentes décrivent des modèles d'attaque par inférence d'appartenance (de l'anglais « Membership inference attacks »). Un modèle d'attaque par inférence d'appartenance permet de prédire, à partir d'un modèle de classification cible, si une donnée à traiter appartient ou n'appartient pas aux données d'apprentissage utilisées pour l'entraînement du modèle cible. Ce type d'attaque peut alors être mis en œuvre par un propriétaire légitime de données qui ont par exemple été piratées par un tiers malveillant sur un serveur, de déterminer ultérieurement si les données piratées ont été utilisées pour entraîner un modèle de classification. De plus, la connaissance de ce type d'attaque permet par la suite de renforcer la sécurité des données effectivement utilisées dans une phase d'apprentissage machine des paramètres d'un modèle de classification par un acteur légitime.
- [0009] L'article de Di Wu, Saiyu Qi, Yong Qi, Qian Li, Bowen Cai, Qi Guo, Jingxian Cheng, "Understanding and defending against White-box membership inference attack in deep learning, Knowledge-Based Systems", publié en 2023, décrit l'utilisation d'attaques d'inférence d'appartenance dans un objectif de défense.
- [0010] On connaît plusieurs méthodes d'attaque par inférence d'appartenance, mettant en œuvre des modèles d'inférence d'appartenance.
- [0011] En particulier, on connaît des modèles d'inférence d'appartenance nécessitant la connaissance de la structure du modèle cible, par exemple, lorsqu'il s'agit d'un réseau de neurones, la structure comprend le nombre de couches, le nombre de neurones par couches, les liaisons entre les couches, les fonctions d'activation et les valeurs des poids. En d'autres termes, tous les paramètres définissant le modèle cible sont connus. Ces modèles d'inférence d'appartenance font partie des attaques dites « white box », le modèle cible étant considéré transparent.
- [0012] Par exemple, l'article « Membership Inference Attack Using Self Influence Functions » de G. Cohen et R. Giryès publié en 2022 décrit une attaque par inférence d'appartenance de type « white box » utilisant des fonctions d'auto-influence.
- [0013] On connaît également des attaques par inférence d'appartenance sur un modèle cible utilisable pour obtenir une classification à partir d'une donnée à traiter, mais dont la structure, et plus généralement les paramètres ne sont pas connus. Ces attaques sont également appelées attaques « black box », le modèle cible étant considéré opaque, tel une boîte noire.
- [0014] Par exemple, l'article « Membership inference attacks from first principles » de N. Carlini et al, publié dans IEEE Symposium on Security and Privacy, 2022, décrit une attaque de ce type.
- [0015] Cependant, chaque méthode d'attaque par inférence connue a des avantages et des inconvénients, et leurs performances varient en fonction de l'ensemble de données à traiter.

- [0016] Il n'existe pas à l'heure actuelle d'outil pour comparer les modèles d'inférence d'appartenance ou pour déterminer quel modèle produira de meilleures performances sur un ensemble de données à traiter.
- [0017] Le but de l'invention est alors de proposer un procédé de détermination d'appartenance aux données d'apprentissage d'un modèle de classification permettant d'obtenir des performances améliorées, en particulier dans la détermination d'appartenance effective ou détection de vrais positifs.
- [0018] A cet effet, l'invention a pour objet un procédé de détermination d'appartenance aux données d'apprentissage utilisées pour l'entraînement d'un modèle de classification de données par apprentissage machine, dit modèle cible, le procédé étant mis en œuvre par un processeur de calcul, et comportant, dans une phase d'apprentissage, mise en œuvre sur un ensemble de données à traiter, des étapes de :
- [0019] - génération d'un nombre N supérieur à 2 de partitions distinctes de l'ensemble de données à traiter en un premier et un deuxième sous-ensembles de données,
- [0020] - calcul par apprentissage machine de N modèles témoins, chaque modèle témoin ayant une même structure que ledit modèle cible, le calcul comportant l'apprentissage des paramètres de chaque modèle témoin à partir d'un premier sous-ensemble de données d'une des partitions, pour fournir en sortie une classification sensiblement identique à celle fournie par le modèle cible;
- [0021] - apprentissage machine sur un premier sous-ensemble de modèles témoins d'au moins deux modèles d'inférence d'appartenance distincts fournissant chacun, pour chaque donnée à traiter, une prédiction d'appartenance ou de non-appartenance aux données d'apprentissage du modèle cible,
- [0022] - apprentissage machine, sur un deuxième sous-ensemble de modèles témoins, des paramètres d'au moins un modèle de décision prenant en entrée lesdites prédictions d'appartenance ou de non-appartenance et fournissant en sortie une prédiction d'appartenance ou de non-appartenance consolidée.
- [0023] Avantageusement, le procédé de détermination d'appartenance aux données d'apprentissage proposé combine au moins deux modèles d'inférence d'appartenance distincts, et met en œuvre un modèle de décision. De préférence, le modèle de décision prend en considération également des scores de confiance de chaque modèle d'inférence d'appartenance.
- [0024] Le procédé de détermination d'appartenance aux données d'apprentissage selon l'invention peut également présenter une ou plusieurs des caractéristiques ci-dessous, prises indépendamment ou selon toutes les combinaisons techniquement envisageables.
- [0025] Au moins un des modèles d'inférence d'appartenance fournit, pour chaque donnée à traiter, un score de confiance associé à la prédiction d'appartenance ou de non-

appartenance aux données d'apprentissage du modèle cible, et l'apprentissage machine dudit au moins un modèle de décision prend en entrée lesdits scores de confiance.

- [0026] Chaque modèle d'inférence d'appartenance distinct fournit pour chaque donnée à traiter, un score de confiance associé à la prédiction d'appartenance ou de non-appartenance aux données d'apprentissage du modèle cible, et l'apprentissage machine, sur un deuxième sous-ensemble de modèles témoins, des paramètres d'au moins un modèle de décision prend en entrée lesdites prédictions d'appartenance ou de non-appartenance et les scores de confiance associés fournis par chaque modèle d'inférence d'appartenance distinct.
- [0027] Le procédé comporte en outre une phase opérationnelle comportant, pour au moins une donnée à traiter, des étapes de :
- mise en œuvre de chacun desdits modèles d'inférence d'appartenance sur ladite donnée à traiter pour obtenir des prédictions d'appartenance ou de non-appartenance et les scores de confiance associés ;
 - détermination d'une prédiction consolidée d'appartenance ou de non-appartenance de ladite donnée à traiter aux données d'apprentissage par application dudit modèle de décision sur les prédictions d'appartenance ou de non-appartenance et les scores de confiance associés.
- [0028] Le procédé met en œuvre un premier modèle d'inférence d'appartenance utilisant une connaissance des paramètres du modèle cible, et un deuxième modèle d'inférence d'appartenance sans utilisation des paramètres du modèle cible.
- [0029] Une pluralité de modèles de décision sont mis en œuvre parmi des modèles de : régression logistique, forêt aléatoire, amplification adaptative, amplification de gradient, modèle bayésien naïf.
- [0030] Le procédé comporte en outre une étape de détermination d'appartenance ou non-appartenance pour obtenir une prédiction finale d'appartenance ou non-appartenance par combinaison des résultats des modèles de décision.
- [0031] La combinaison des résultats est effectuée par vote majoritaire ou vote majoritaire pondéré.
- [0032] Selon un autre aspect, l'invention concerne un dispositif de détermination d'appartenance aux données d'apprentissage de détermination d'appartenance aux données d'apprentissage utilisées pour l'entraînement d'un modèle de classification de données par apprentissage machine, dit modèle cible, le dispositif comportant un processeur de calcul, configuré, dans une phase d'apprentissage, pour mettre en œuvre sur un ensemble de données à traiter :
- [0033] - un module de génération d'un nombre N supérieur à 2 de partitions distinctes de l'ensemble de données à traiter en un premier et un deuxième sous-ensembles de données,

- [0034] - un module de calcul par apprentissage machine de N modèles témoins, chaque modèle témoin ayant une même structure que ledit modèle cible, le calcul comportant l'apprentissage des paramètres de chaque modèle témoin à partir du premier sous-ensemble de données d'une des partitions, pour fournir en sortie une classification sensiblement identique à celle fournie par le modèle cible;
- [0035] - un module d'apprentissage machine sur un premier sous-ensemble de modèles témoins d'au moins deux modèles d'inférence d'appartenance distincts fournissant chacun, pour chaque donnée à traiter, une prédiction d'appartenance ou de non-appartenance aux données d'apprentissage du modèle cible,
- [0036] - un module d'apprentissage machine, sur un deuxième sous-ensemble de modèles témoins, des paramètres d'au moins un modèle de décision prenant en entrée lesdites prédictions d'appartenance ou de non-appartenance et fournissant en sortie une prédiction d'appartenance ou de non-appartenance consolidé.
- [0037] Avantageusement, ce dispositif est configuré pour mettre en œuvre un procédé de détermination d'appartenance aux données d'apprentissage tel que brièvement décrit ci-dessus.
- [0038] Selon un autre aspect, l'invention concerne un support d'enregistrement d'informations, sur lequel sont mémorisées des instructions logicielles pour l'exécution d'un procédé de détermination d'appartenance aux données d'apprentissage tel que brièvement décrit ci-dessus, lorsque ces instructions sont exécutées par un dispositif électronique programmable.
- [0039] Selon un autre aspect, l'invention concerne un programme d'ordinateur comportant des instructions logicielles qui, lorsqu'elles sont mises en œuvre par un dispositif électronique programmable, mettent en œuvre un procédé de détermination d'appartenance aux données d'apprentissage tel que brièvement décrit ci-dessus.
- [0040] L'invention apparaîtra plus clairement à la lecture de la description qui va suivre, donnée uniquement à titre d'exemple non limitatif, et faite en référence aux dessins dans lesquels :
- [0041] [Fig.1] la [Fig.1] est une représentation schématique des principaux modules d'un dispositif de détermination d'appartenance aux données d'apprentissage utilisées pour l'entraînement d'un modèle de classification de données par apprentissage machine selon un mode de réalisation ;
- [0042] [Fig.2] la [Fig.2] est un synoptique des principales étapes d'un mode de réalisation du procédé de détermination d'appartenance dans la phase d'apprentissage ;
- [0043] [Fig.3] la [Fig.3] est une illustration schématique d'une pluralité de partitions de données ;
- [0044] [Fig.4] la [Fig.4] est un synoptique des principales étapes d'un mode de réalisation du procédé de détermination d'appartenance dans la phase opérationnelle.

- [0045] La [Fig.1] représente schématiquement les principaux modules d'un dispositif 2 de détermination d'appartenance aux données d'apprentissage utilisées préalablement pour l'entraînement d'un modèle de classification cible MC.
- [0046] Le dispositif 2 est un dispositif électronique programmable, e.g un ordinateur.
En variante non représenté, le dispositif 2 est formé d'une pluralité de dispositifs électroniques programmables connectés entre eux.
- [0047] Le dispositif 2 comporte au moins un processeur de calcul 4, une unité de mémoire électronique 6, adaptées à communiquer via un bus de communication 5.
- [0048] En complément, le dispositif 2 comporte également une interface de communication avec des dispositifs distants, par un protocole de communication choisi, par exemple un protocole filaire et/ou un protocole de communication radio, ainsi qu'une interface homme machine ; ces interfaces sont réalisées de manière classique et ne sont pas représentées dans la [Fig.1].
- [0049] L'unité de mémoire électronique mémorise le modèle cible (MC) 8, qui est un modèle de classification de données d'entrée préalablement entraîné par apprentissage machine sur un ensemble de données d'apprentissage.
- [0050] L'ensemble de données d'apprentissage préalablement utilisé n'est pas connu.
- [0051] Le modèle cible est entraîné pour classifier des données d'entrée représentées sous forme vectorielle ou matricielle de dimension(s) prédéterminées.
- [0052] Par exemple, dans une application, les données d'entrée sont des images numériques d'objets, à classifier dans un ensemble prédéterminé de classes.
- [0053] La classification consiste par exemple à associer un label de classification (ou étiquette de classification) à chaque donnée d'entrée.
- [0054] De nombreuses applications utilisent ce type de modèle de classification pour la reconnaissance d'objets à partir d'images numériques.
- [0055] Dans un mode de réalisation, le modèle cible est réalisé sous forme d'un réseau de neurones.
- [0056] De manière connue, un réseau de neurones comporte une succession ordonnée de couches de neurones dont chacune prend ses entrées sur les sorties de la couche précédente.
- [0057] Plus précisément, chaque couche comprend des neurones prenant leurs entrées sur les sorties des neurones de la couche précédente, ou bien sur les variables d'entrée pour la première couche.
- [0058] En variante, des structures de réseaux de neurones plus complexes peuvent être envisagées avec une couche qui peut être reliée à une couche plus lointaine que la couche immédiatement précédente.

- [0059] À chaque neurone est également associée une opération, c'est-à-dire un type de traitement, à effectuer par ledit neurone au sein de la couche de traitement correspondante.
- [0060] Chaque couche est reliée aux autres couches par une pluralité de synapses. Un poids synaptique est associé à chaque synapse, et chaque synapse forme une liaison entre deux neurones. C'est souvent un nombre réel, qui prend des valeurs positives comme négatives. Dans certains cas, le poids synaptique est un nombre complexe.
- [0061] Chaque neurone est propre à effectuer une somme pondérée des valeur(s) reçue(s) de la part des neurones de la couche précédente, chaque valeur étant alors multipliée par le poids synaptique respectif de chaque synapse, ou liaison, entre ledit neurone et les neurones de la couche précédente, puis à appliquer une fonction d'activation, typiquement une fonction non-linéaire, à ladite somme pondérée, et à délivrer en sortie dudit neurone, en particulier aux neurones de la couche suivante qui lui sont connectés, la valeur résultant de l'application de la fonction d'activation. La fonction d'activation permet d'introduire une non-linéarité dans le traitement effectué par chaque neurone. La fonction sigmoïde, la fonction tangente hyperbolique, la fonction de Heaviside sont des exemples de fonction d'activation.
- [0062] En complément facultatif, chaque neurone est également apte à appliquer, en outre, un facteur multiplicatif, et un biais additif, à la sortie de la fonction d'activation, et la valeur délivrée en sortie dudit neurone est alors le produit de la valeur facteur multiplicatif et de la valeur issue de la fonction d'activation, additionnée du biais.
- [0063] Un réseau de neurones à convolution est aussi parfois dénommé réseau de neurones convolutif ou par le sigle CNN qui renvoie à la dénomination anglaise de « *Convolutional Neural Networks* ».
- [0064] Dans un réseau de neurones à convolution, chaque neurone d'une même couche présente exactement le même schéma de connexion que ses neurones voisins, mais à différentes positions d'entrée. Le motif de connexion est appelé noyau de convolution ou, plus souvent, « *kernel* » en référence à la dénomination anglaise correspondante.
- [0065] Une couche de neurones entièrement connectée est une couche dans laquelle les neurones de ladite couche sont chacun connectés à tous les neurones de la couche précédente.
- [0066] Un tel type de couche est plus souvent dénommé selon le terme anglais de « *fully connected* », et parfois désigné sous la dénomination « couche dense ».
- [0067] Les valeurs des poids, facteurs multiplicateurs et biais le cas échéant sont apprises lors d'une phase d'apprentissage machine pour réaliser la tâche de classification.
- [0068] L'invention s'applique avec tous les types de réseau de neurones.

- [0069] De manière générale, il est considéré que les paramètres du modèle cible, et en particulier du réseau de neurones formant le modèle cible, sont connus, les paramètres incluant les structures et les valeurs des poids appris par apprentissage machine.
- [0070] Outre le modèle cible, l'unité de mémoire 6 mémorise également un ensemble 10 de données 12_i à traiter.
- [0071] Chaque donnée à traiter 12_i est une donnée numérique, représentée sous la même forme vectorielle ou matricielle que les données d'entrée du modèle cible 8, et de même dimension(s) prédéterminée(s).
- [0072] Les données à traiter sont des données susceptibles d'avoir été utilisées lors de l'entraînement du modèle cible 8. Chaque donnée à traiter est une instance (ou un exemple) d'une des classes de sortie du modèle cible.
- [0073] L'ensemble de données à traiter a un cardinal L , L étant un nombre entier positif, de préférence supérieur à 2.
- [0074] En complément facultatif, l'ensemble 8 de données à traiter peut être augmenté en appliquant une méthode d'augmentation de données, par exemple par traitement d'images lorsque les données à traiter sont des images numériques (e.g. rotation, translation, retournement (ou flipping en anglais) vertical ou horizontal). Cela permet notamment d'assurer qu'un nombre suffisant d'instances des classes est disponible pour le calcul de modèles témoins, décrit ci-après.
- [0075] Le processeur 4 est configuré pour mettre en œuvre un module 20 de génération d'un nombre N supérieur à 2 de partitions distinctes de l'ensemble de données à traiter en un premier et un deuxième sous-ensembles de données.
- [0076] Il comporte également un module 22 de calcul par apprentissage machine de N modèles témoins (MT) $16_1 \dots 16_N$, qui sont mémorisés dans l'unité de mémoire 6.
- [0077] Chaque modèle témoin 16_j est un modèle de classification de même structure que le modèle cible 8, et les paramètres du modèle témoin sont appris par apprentissage machine sur le premier sous-ensemble de données d'une des partitions générées, avec l'objectif d'obtenir une classification sensiblement identique à celle du modèle cible, et en particulier avec l'objectif d'obtenir une distribution statistique en sortie similaire à la distribution statistique en sortie du modèle cible.
- [0078] En d'autres termes, chaque modèle témoin « imite » le modèle cible, le modèle témoin étant entraîné sur une base de données d'apprentissage (e.g. le premier-sous-ensemble d'une partition choisie) connue.
- [0079] Le processeur 4 comporte en outre un module 24 d'apprentissage machine sur un premier sous-ensemble de modèles témoins d'au moins deux modèles d'inférence d'appartenance $26_1, \dots, 26_P$ distincts fournissant chacun, pour chaque donnée à traiter,

une prédiction d'appartenance ou de non-appartenance aux données d'apprentissage du modèle de classification cible et de préférence un score de confiance associé.

- [0080] Par exemple et comme décrit plus en détail ci-après, deux modèles d'inférence d'appartenance distincts et complémentaires sont mis au point, un premier modèle de type « white box », le modèle cible étant considéré connu, et un deuxième modèle de type « black box », le modèle cible étant considéré inconnu.
- [0081] De manière plus générale, le nombre P de modèles d'inférence d'appartenance est quelconque, supérieur ou égal à 2.
- [0082] Enfin, le processeur 4 comporte un module 28 d'apprentissage machine, sur un deuxième sous-ensemble de modèles témoins, des paramètres d'un ou plusieurs modèle(s) de décision 30 prenant en entrée lesdites prédictions d'appartenance ou de non-appartenance et les scores de confiance associés fournis par les P modèles d'inférence d'appartenance distincts. Le module de décision fournit en sortie une prédiction d'appartenance ou non-appartenance consolidée.
- [0083] Dans un mode de réalisation, les modèles de décision 30 comprennent un ou plusieurs classifieurs choisis parmi des modèles de : régression logistique, forêt aléatoire, amplification adaptative, amplification de gradient, modèle bayésien naïf. Bien entendu, cette liste n'est pas exhaustive, toute méthode de classification statistique est applicable.
- [0084] Lorsque plusieurs modèles de décision sont mis en œuvre, un module de détermination 32 par combinaison des résultats des modèles de décision 30 est également mis en œuvre. Par exemple, la combinaison est réalisée par vote majoritaire pour obtenir une prédiction finale, également appelée décision finale. Ainsi, une décision finale est obtenue pour chaque donnée à traiter quant à l'appartenance ou la non-appartenance de la donnée à traiter aux données d'apprentissage du modèle cible.
- [0085] Bien entendu, après l'apprentissage machine des modèles d'inférence d'appartenance 26₁..26_p, et du ou des modèles de décision 30, chacun de ces modèles devient un modèle entraîné par apprentissage machine exécutable pour mettre en œuvre la tâche pour laquelle il a été entraîné.
- [0086] Ainsi, dans une phase opérationnelle postérieure à la phase d'apprentissage, le modèle cible 8, les modèles d'inférence d'appartenance 26₁..26_p, le ou les modèles de décision 30 et le module de détermination 32 sont exécutés sur une donnée à traiter pour déterminer l'appartenance ou la non-appartenance de cette donnée aux données d'apprentissage du modèle cible 6.
- [0087] Dans un mode de réalisation, les modules 20, 22, 26, 24, 28, 30, 32 sont réalisés sous forme d'instructions logicielles formant un programme d'ordinateur, qui, lorsqu'il est exécuté par un dispositif électronique programmable, met en œuvre un procédé de détermination d'appartenance aux données d'apprentissage selon l'invention.

- [0088] En variante non représentée, les modules 22, 24, 26, 28,30, 32 sont réalisés chacun sous forme de composants logiques programmables, tels que des FPGA (de l'anglais *Field Programmable Gate Array*), des microprocesseurs, des composants GPGPU (de l'anglais *General-purpose processing on graphics processing*), ou encore de circuits intégrés dédiés, tels que des ASIC (de l'anglais *Application Specific Integrated Circuit*).
- [0089] Le programme d'ordinateur comportant des instructions logicielles est en outre apte à être enregistré sur un support d'enregistrement d'informations, non-transitoire et lisible par ordinateur. Ce support lisible par ordinateur est par exemple, un médium apte à mémoriser les instructions électroniques et à être couplé à un bus d'un système informatique. A titre d'exemple, ce support est un disque optique, un disque magnéto-optique, une mémoire ROM, une mémoire RAM, tout type de mémoire non-volatile (par exemple EPROM, EEPROM, FLASH, NVRAM), une carte magnétique ou une carte optique.
- [0090] La [Fig.2] est un synoptique des principales étapes d'un procédé de détermination d'appartenance dans la phase d'apprentissage, selon un mode de réalisation dans lequel le nombre P de modèles d'inférence d'appartenance est égal à 2.
- [0091] En entrée sont fournis le modèle cible et l'ensemble de données à traiter.
- [0092] Le procédé comprend une étape 40 de génération de N partitions de l'ensemble des données à traiter.
- [0093] Comme montré plus explicitement dans la [Fig.3], pour chaque partition $Part_i$ parmi les N partitions $Part_1, Part_2, \dots, Part_N$, un premier sous-ensemble de données S1 comprend des instances destinées à l'apprentissage, notamment à l'apprentissage des modèles témoins, et un deuxième sous-ensemble de données S2 comprend des données destinées aux tests et validations.
- [0094] Les données de validation sont utilisées pour choisir des hyperparamètres du modèle témoin et vérifier la convergence de ce modèle.
- [0095] Les données de test ne sont pas utilisées dans la phase d'apprentissage et servent à évaluer la performance du modèle.
- [0096] Dans un mode de réalisation, le premier sous-ensemble de données S1, utilisées pour l'apprentissage ; comporte plus de données que le deuxième sous-ensemble de données S2, utilisées pour le test ou la validation, de préférence le premier sous-ensemble de données S1 comporte au moins 60% des données.
- [0097] Dans un autre mode de réalisation, les sous-ensembles de données comportent le même nombre de données.
- [0098] Par exemple pour la génération d'une partition donnée, chaque donnée d'entrée est attribuée, par exemple aléatoirement, au premier sous-ensemble de données S1 ou au deuxième sous-ensemble de données S2.

- [0099] Dans un mode de réalisation, $N/2$ partitions sont générées aléatoirement, et les autres $N/2$ sont générées par complémentarité, en inversant l'attribution de chaque donnée, de manière à assurer que chaque donnée est attribuée $N/2$ fois au premier sous-ensemble de données $S1$ (i.e. pour l'apprentissage) et $N/2$ fois au deuxième sous-ensemble de données $S2$.
- [0100] Dans un mode de réalisation préféré, pour chaque donnée, on choisit aléatoirement $N/2$ partitions dans lesquelles la donnée d'entrée est attribuée au premier sous-ensemble de données $S1$ (i.e. pour l'apprentissage), et par complémentarité, la donnée est attribuée au deuxième sous-ensemble de données $S2$ dans les $N/2$ partitions restantes. Dans cette variante, l'attribution à l'un ou l'autre des sous-ensembles dans chaque partition se fait par donnée.
- [0101] Un ensemble de N partitions telles que chaque donnée attribuée $N/2$ fois au premier sous-ensemble $S1$ et $N/2$ fois au deuxième sous-ensemble $S2$ est dit ensemble équilibré.
- [0102] Le procédé comporte ensuite une étape 42 de calcul par apprentissage machine de N modèles témoins du modèle cible, i.e. qui fournissent des résultats sensiblement identiques à ceux fournis par le modèle cible sur un sous-ensemble de données d'apprentissage.
- [0103] Chaque modèle témoin est entraîné sur une partition distincte de l'ensemble de données à traiter, et plus spécifiquement sur le premier sous-ensemble de données de la partition.
- [0104] A l'issue de l'étape 42, N modèles témoins distincts sont obtenus et mémorisés.
- [0105] Dans un mode de réalisation, l'étape 42 effectue un calcul par apprentissage machine d'un premier sous-ensemble de modèles témoins, sur un ensemble équilibré de $N1$ partitions, et un deuxième sous-ensemble de modèles témoins, sur un ensemble équilibré de $N2$ partitions, avec $N=N1+N2$.
- [0106] Le procédé comporte alors, dans un mode de réalisation, un apprentissage machine 44 d'un premier modèle d'inférence d'appartenance, en connaissance des paramètres du modèle cible, et un apprentissage machine 46 d'un deuxième modèle d'inférence d'appartenance, sans connaissance du modèle cible.
- [0107] Chacune des étapes d'apprentissage 44 et 46 met en œuvre un premier sous-ensemble de modèles témoins.
- [0108] Par exemple, pour $N=288$ modèles témoins calculés, le premier-sous ensemble comporte $N1=256$ modèles témoins, et un deuxième sous-ensemble de $N2=32$ modèles témoins, chacun des premier et deuxième sous-ensembles de modèles témoins étant appris sur un sous-ensemble équilibré de données d'apprentissage.
- [0109] Le premier sous-ensemble de modèles témoins est utilisé pour l'apprentissage des modèles d'inférence d'appartenance.

- [0110] Le deuxième sous-ensemble de modèles témoins est utilisé ultérieurement pour l'apprentissage d'un modèle de décision.
- [0111] A titre d'exemple non limitatif, l'étape 44 met en œuvre l'apprentissage d'un premier modèle d'inférence d'appartenance, qui met en œuvre un calcul de valeurs caractéristiques d'influence de chaque donnée à traiter, en utilisant les modèles témoins préalablement calculés.
- [0112] Le calcul de fonctions d'influence est décrit dans l'article « Membership Inference Attack Using Self Influence Functions » de G. Cohen et R. Giryès.
- [0113] Dans un mode de réalisation lors de l'étape 44 :
- [0114] -pour chaque modèle témoin du premier sous-ensemble de modèles témoins et chaque instance des données à traiter, l'influence de l'instance considérée par rapport aux instances de sa classe étant donné le modèle cible est calculée ;
- [0115] -pour chaque modèle témoin et chaque instance, des valeurs caractéristiques sont calculées ;
- [0116] -un apprentissage machine d'un modèle de décision, par exemple par régression logistique, d'appartenance ou de non-appartenance aux données d'apprentissage, la décision étant effectuée sur les valeurs caractéristiques calculées, en connaissance de l'appartenance effective de chaque instance au sous-ensemble de données d'apprentissage de chaque modèle témoin.
- [0117] Par exemple, dans un mode de réalisation, les valeurs caractéristiques calculées comprennent, pour chaque instance et chaque modèle témoin, une ou plusieurs des valeurs caractéristiques suivantes : la valeur d'auto-influence, la moyenne de l'influence de l'instance sur les données de sa classe, la moyenne de l'influence des données de sa classe sur l'instance considérée, l'opposé de la valeur de sortie du modèle témoin pour la classe de l'instance, la perte de classification (en anglais « hinge loss »).
- [0118] De manière classique dans le domaine de l'intelligence artificielle, le terme « logit » désigne un vecteur de sortie d'un modèle de classification entraîné par apprentissage machine, ce vecteur ayant une taille égale au nombre de classes, chaque composante de ce vecteur correspondant à une des classes et ayant une valeur, appelée valeur de sortie ci-dessus, pour ladite classe.
- [0119] De manière classique dans le domaine de l'intelligence artificielle, la perte de classification, ou « hinge loss » en anglais, pour une instance d'une classe, est égale à la valeur de sortie pour sa classe moins la plus grande valeur parmi les autres valeurs de sortie.
- [0120] Le premier modèle d'inférence d'appartenance ainsi entraîné fournit en sortie, pour chaque donnée à traiter E_i , une prédiction $\text{Préd1}(E_i)$ d'appartenance ou de non-

appartenance de la donnée E_i aux données d'apprentissage du modèle de classification cible et un score de confiance associé Score_{1i} .

- [0121] A titre d'exemple non limitatif, l'étape 46 met en œuvre un deuxième modèle d'inférence d'appartenance met en œuvre un ratio de vraisemblance de la fonction de perte de classification (en anglais « hinge loss »).
- [0122] Un exemple de modèle d'inférence d'appartenance de ce type est décrit dans l'article « Membership inference attacks from first principles » de N. Carlini et al, publié dans IEEE Symposium on Security and Privacy, 2022. Dans un mode de réalisation, l'étape 46 comporte des sous-étapes de :
 - [0123] -calcul de la valeur de perte de classification (« hinge loss ») pour chaque modèle témoin et chaque instance ;
 - [0124] - détermination, pour chaque instance, des paramètres de deux gaussiennes, les paramètres étant respectivement la médiane et l'écart-type, et les gaussiennes étant respectivement :
 - [0125] une première gaussienne représentative de la distribution de la valeur de perte de classification lorsque l'instance a été utilisée à l'apprentissage du modèle témoin ;
 - [0126] une deuxième gaussienne représentative de la distribution de la valeur de perte de classification lorsque l'instance n'a pas été utilisée à l'apprentissage du modèle témoin.
 - [0127] -pour chaque instance, et chaque modèle témoin, détermination d'appartenance ou de non-appartenance en fonction des gaussiennes, comprenant un calcul du rapport entre le score de vraisemblance d'appartenance en utilisant la première gaussienne et du score de vraisemblance d'appartenance en utilisant la deuxième gaussienne ;
 - [0128] -et calcul d'un score moyen de l'attaque sur les modèles témoins, ce score étant analogue à une valeur de fidélité de l'apprentissage, permettant d'obtenir un score de confiance du deuxième modèle d'inférence d'appartenance pour l'instance considérée.
- [0129] Ce deuxième modèle d'inférence d'appartenance ainsi entraîné fournit en sortie, pour chaque donnée à traiter E_i , une prédiction $\text{Préd}_2(E_i)$ d'appartenance ou de non-appartenance de la donnée E_i aux données d'apprentissage du modèle de classification cible, un score de confiance associé Score_2 .
- [0130] De plus, dans ce mode de réalisation, le deuxième modèle d'inférence d'appartenance fournit aussi une valeur de fidélité de l'apprentissage (en anglais « train accuracy »), qui est également considérée comme une autre valeur de score de confiance associée à l'application de ce deuxième modèle d'inférence.
- [0131] Le procédé comporte ensuite une étape 48 d'apprentissage d'un modèle de décision.
- [0132] Cet apprentissage prend en entrée, pour chaque donnée à traiter, et pour chaque modèle témoin du deuxième sous-ensemble de modèles témoins, la prédiction d'appartenance obtenue par le modèle témoin. Comme déjà indiqué ci-dessus, les modèles témoins du deuxième sous-ensemble de modèles témoins sont de préférence

des modèles témoins qui n'ont pas été utilisés pour l'apprentissage des modèles d'inférence d'appartenance.

- [0133] Dans le mode de réalisation décrit en référence à la [Fig.2], l'apprentissage prend en entrée également le score d'appartenance fourni par chacun des modèles d'inférence d'appartenance.
- [0134] Par ailleurs, pour chaque modèle témoin, l'appartenance ou non-appartenance de la donnée à traiter au premier ensemble de données (i.e. données d'apprentissage) utilisées pour l'entraînement du modèle témoin est connue, par construction du modèle témoin.
- [0135] Il est alors possible d'entraîner un modèle de décision, par exemple un classifieur binaire, qui fournit en sortie une prédiction d'appartenance ou de non-appartenance qui est appelée prédiction consolidée, et optionnellement, un score de confiance associé.
- [0136] La décision ainsi obtenue est plus fidèle que les prédictions respectives de chacun des modèles d'inférence d'appartenance, a fortiori lorsque les prédictions sont combinées en tenant compte de leurs scores de confiance respectifs.
- [0137] Dans le cas où un seul classifieur binaire est entraîné à l'étape 48, la prédiction consolidée d'appartenance ou de non-appartenance fourni par ce classifieur constitue une décision finale.
- [0138] Dans le cas où plusieurs classifieurs binaires distincts sont entraînés à l'étape 48, le procédé comporte également une étape 50 de détermination d'appartenance ou non appartenance (i.e décision finale) par combinaison des prédictions consolidées des classifieurs binaires.
- [0139] Par exemple, lorsque le nombre de classifieurs binaires est impair, l'étape 50 met en œuvre un vote majoritaire : la prédiction consolidée majoritaire est retenue comme décision finale d'appartenance ou de non-appartenance aux données d'apprentissage du modèle cible.
- [0140] Dans un autre mode de réalisation, l'étape 50 met en œuvre un apprentissage de pondérations à attribuer aux classifieurs utilisés.
- [0141] Bien entendu, d'autres méthodes de combinaisons différentes sont envisageables pour un homme du métier.
- [0142] La [Fig.4] est un synoptique des principales étapes d'un procédé de détermination d'appartenance dans la phase opérationnelle, après mise en œuvre de l'apprentissage décrit ci-dessus en référence à la [Fig.2], selon un mode de réalisation.
- [0143] Pour une donnée à traiter E_T , par exemple une image, le procédé comprend l'application 60 du modèle cible, l'application 62, 64 des premier et deuxièmes modèles d'inférence d'appartenance, puis l'application 66 des modèles de décision préalablement entraînés et l'application, le cas échéant, de l'étape de détermination 68

pour obtenir une prédiction consolidée, et une décision finale d'appartenance ou non appartenance de la donnée à traiter E_T aux données d'apprentissage du modèle cible.

[0144] Le mode de réalisation décrit en référence aux figures 2 et 4 comporte la mise en œuvre de deux modèles d'inférence d'appartenance distincts.

[0145] Bien entendu, le procédé décrit est généralisable sans difficulté à un nombre quelconque de modèles d'inférence d'appartenance distincts, par exemple trois, quatre ou plus, chaque modèle d'inférence d'appartenance étant appris sur des modèles témoins et fournissant une prédiction d'appartenance ou de non-appartenance aux données d'apprentissage et optionnellement, un ou plusieurs scores de confiance associés. L'apprentissage machine sur un deuxième sous-ensemble de modèles témoins, des paramètres d'au moins un modèle de décision prenant en entrée les prédictions d'appartenance ou de non-appartenance est effectué alors sur les prédictions obtenues par la pluralité de modèles d'inférence d'appartenance, et le cas échéant, les scores de confiance associés.

Revendications

[Revendication 1]

Procédé de détermination d'appartenance aux données d'apprentissage utilisées pour l'entraînement d'un modèle de classification de données par apprentissage machine, dit modèle cible (8), le procédé étant mis en œuvre par un processeur de calcul, et comportant, dans une phase d'apprentissage, mise en œuvre sur un ensemble (10) de données à traiter, des étapes de :

- génération (40) d'un nombre N supérieur à 2 de partitions distinctes ($Part_1, \dots, Part_N$) de l'ensemble de données à traiter en un premier ($S1_1, \dots, S1_N$) et un deuxième ($S2_1, \dots, S2_N$) sous-ensembles de données,
- calcul (42) par apprentissage machine de N modèles témoins ($16_1, \dots, 16_N$), chaque modèle témoin ($16_1, \dots, 16_N$) ayant une même structure que ledit modèle cible (8), le calcul (42) comportant l'apprentissage des paramètres de chaque modèle témoin à partir d'un premier sous-ensemble ($S1_1, \dots, S1_N$) de données d'une des partitions, pour fournir en sortie une classification sensiblement identique à celle fournie par le modèle cible (8);
- apprentissage machine (44, 46) sur un premier sous-ensemble de modèles témoins d'au moins deux modèles d'inférence d'appartenance ($26_1, \dots, 26_P$) distincts fournissant chacun, pour chaque donnée à traiter, une prédiction d'appartenance ou de non-appartenance aux données d'apprentissage du modèle cible,
- apprentissage machine, sur un deuxième sous-ensemble de modèles témoins, des paramètres d'au moins un modèle de décision (48) prenant en entrée lesdites prédictions d'appartenance ou de non-appartenance et fournissant en sortie une prédiction d'appartenance ou de non-appartenance consolidée.

[Revendication 2]

Procédé selon la revendication 1, dans lequel au moins un des modèles d'inférence d'appartenance fournit, pour chaque donnée à traiter, un score de confiance associé à la prédiction d'appartenance ou de non-appartenance aux données d'apprentissage du modèle cible, et dans lequel l'apprentissage machine dudit au moins un modèle de décision (48) prend en entrée lesdits scores de confiance.

[Revendication 3]

Procédé selon la revendication 2, dans lequel chaque modèle d'inférence d'appartenance distinct fournit pour chaque donnée à traiter, un score de confiance associé à la prédiction d'appartenance ou de non-appartenance aux données d'apprentissage du modèle

cible, et dans lequel l'apprentissage machine, sur un deuxième sous-ensemble de modèles témoins, des paramètres d'au moins un modèle de décision (48) prend en entrée lesdites prédictions d'appartenance ou de non-appartenance et les scores de confiance associés fournis par chaque modèle d'inférence d'appartenance distinct.

[Revendication 4]

Procédé selon l'une des revendications 1 à 3, comportant en outre une phase opérationnelle comportant, pour au moins une donnée à traiter, des étapes de :

- mise en œuvre de chacun desdits modèles d'inférence d'appartenance (62, 64) sur ladite donnée à traiter pour obtenir des prédictions d'appartenance ou de non-appartenance et les scores de confiance associés ;
- détermination (66, 68) d'une prédiction consolidée d'appartenance ou de non-appartenance de ladite donnée à traiter aux données d'apprentissage par application dudit modèle de décision sur les prédictions d'appartenance ou de non-appartenance et les scores de confiance associés.

[Revendication 5]

Procédé selon l'une quelconque des revendications 1 à 4, dans lequel sont mis en œuvre un premier modèle d'inférence d'appartenance utilisant une connaissance des paramètres du modèle cible, et un deuxième modèle d'inférence d'appartenance sans utilisation des paramètres du modèle cible.

[Revendication 6]

Procédé selon l'une quelconque des revendications 1 à 5, dans lequel une pluralité de modèles de décision sont mis en œuvre parmi des modèles de : régression logistique, forêt aléatoire, amplification adaptative, amplification de gradient, modèle bayésien naïf.

[Revendication 7]

Procédé selon la revendication 6, comportant en outre une étape de détermination d'appartenance ou non-appartenance pour obtenir une prédiction finale d'appartenance ou non-appartenance par combinaison (50) des résultats des modèles de décision.

[Revendication 8]

Procédé selon la revendication 7, dans lequel la combinaison (50) des résultats est effectuée par vote majoritaire ou vote majoritaire pondéré.

[Revendication 9]

Programme d'ordinateur comportant des instructions logicielles qui, lorsqu'elles sont exécutées par un dispositif électronique programmable, mettent en œuvre un procédé de détermination

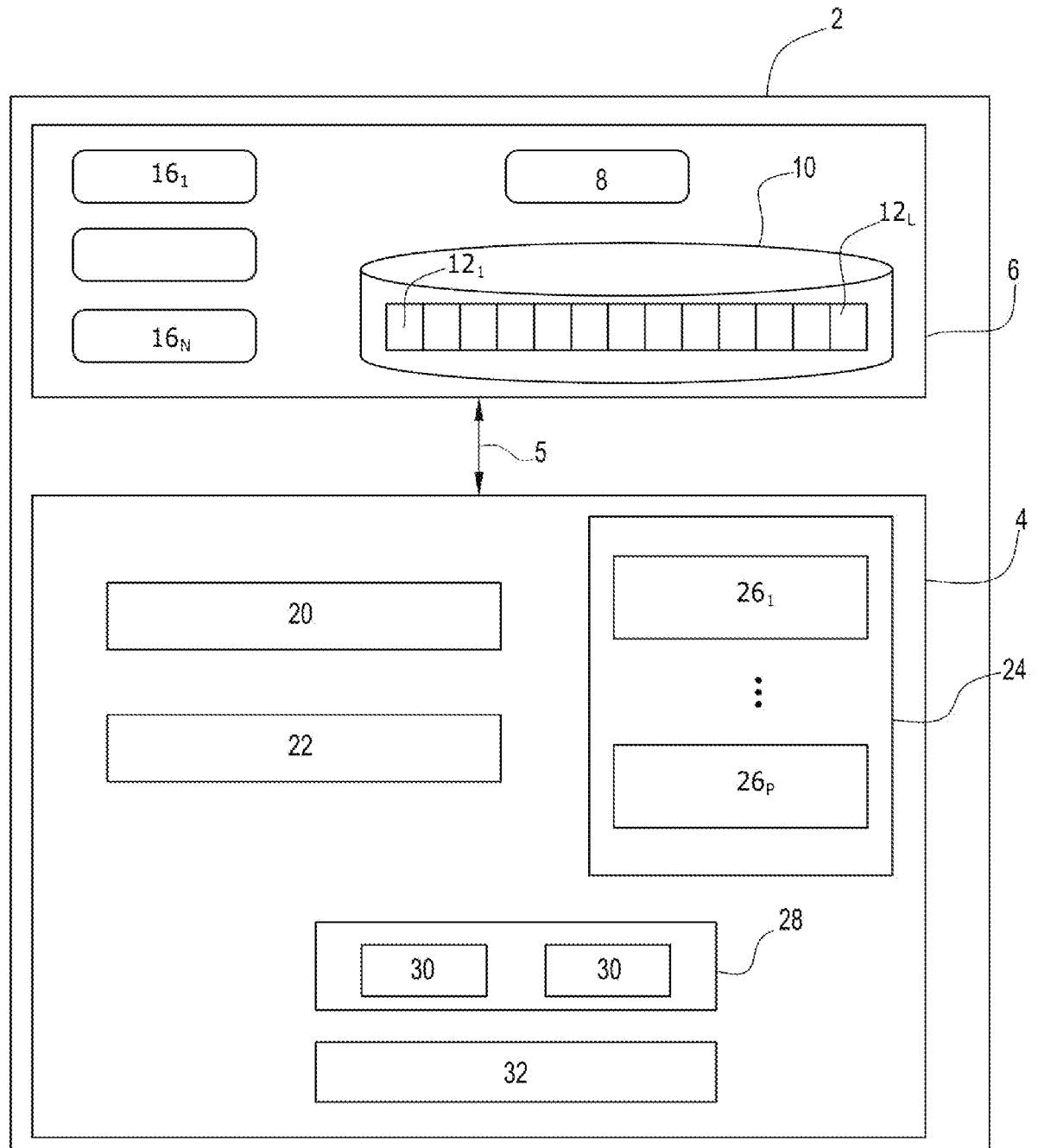
d'appartenance aux données d'apprentissage conforme aux revendications 1 à 8.

[Revendication 10]

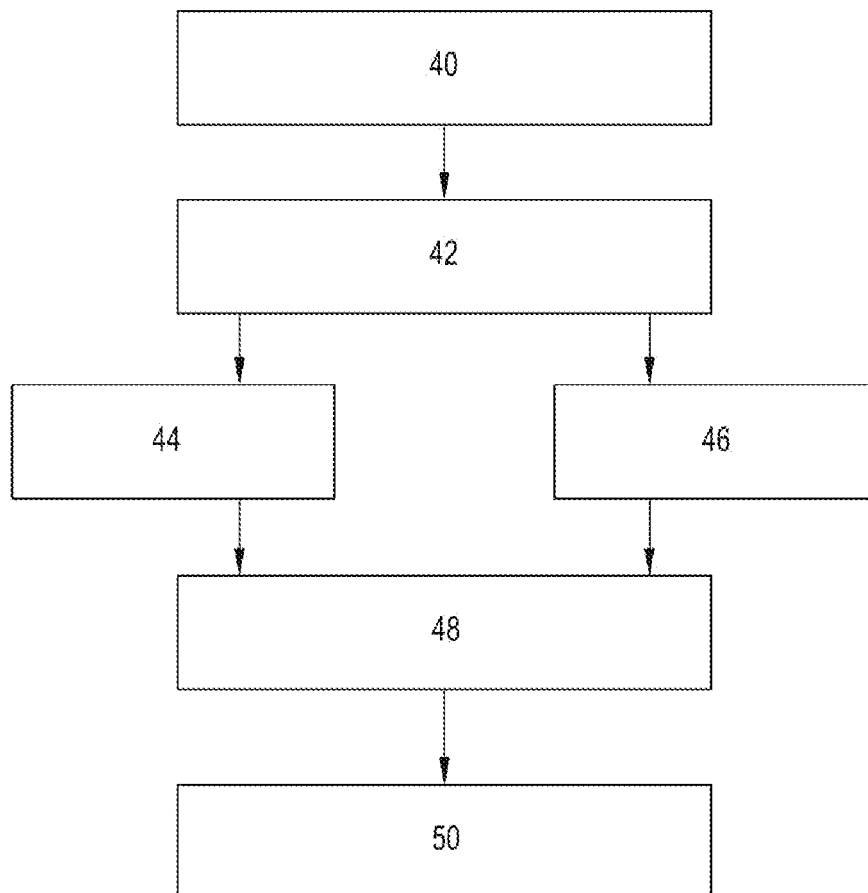
Dispositif de détermination d'appartenance aux données d'apprentissage utilisées pour l'entraînement d'un modèle de classification de données par apprentissage machine, dit modèle cible (8), le dispositif (2) comportant un processeur de calcul (4), configuré, dans une phase d'apprentissage, pour mettre en œuvre sur un ensemble (10) de données à traiter :

- un module (20) de génération d'un nombre N supérieur à 2 de partitions distinctes ($Part_1, \dots, Part_N$) de l'ensemble de données à traiter en un premier ($S1_1, \dots, S1_N$) et un deuxième ($S2_1, \dots, S2_N$) sous-ensembles de données,
- un module (22) de calcul par apprentissage machine de N modèles témoins ($16_1, \dots, 16_N$), chaque modèle témoin ($16_1, \dots, 16_N$) ayant une même structure que ledit modèle cible (8), le calcul comportant l'apprentissage des paramètres de chaque modèle témoin à partir du premier sous-ensemble ($16_1, \dots, 16_N$) de données d'une des partitions, pour fournir en sortie une classification sensiblement identique à celle fournie par le modèle cible (8);
- un module (24) d'apprentissage machine sur un premier sous-ensemble de modèles témoins d'au moins deux modèles d'inférence d'appartenance distincts fournissant chacun, pour chaque donnée à traiter, une prédiction d'appartenance ou de non-appartenance aux données d'apprentissage du modèle cible,
- un module (30) d'apprentissage machine, sur un deuxième sous-ensemble de modèles témoins, des paramètres d'au moins un modèle de décision prenant en entrée lesdites prédictions d'appartenance ou de non-appartenance et fournissant en sortie une prédiction d'appartenance ou de non-appartenance consolidé.

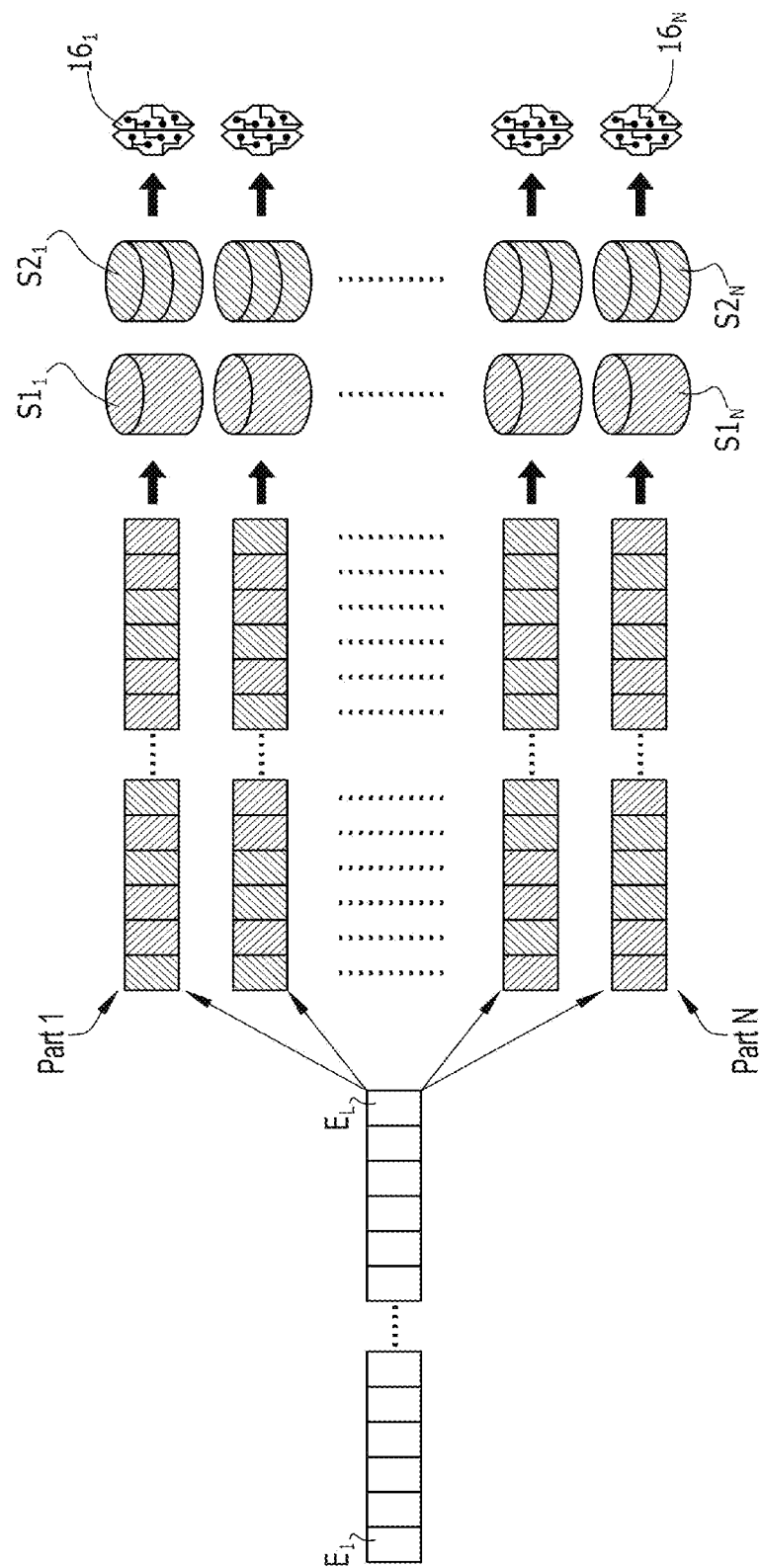
[Fig. 1]



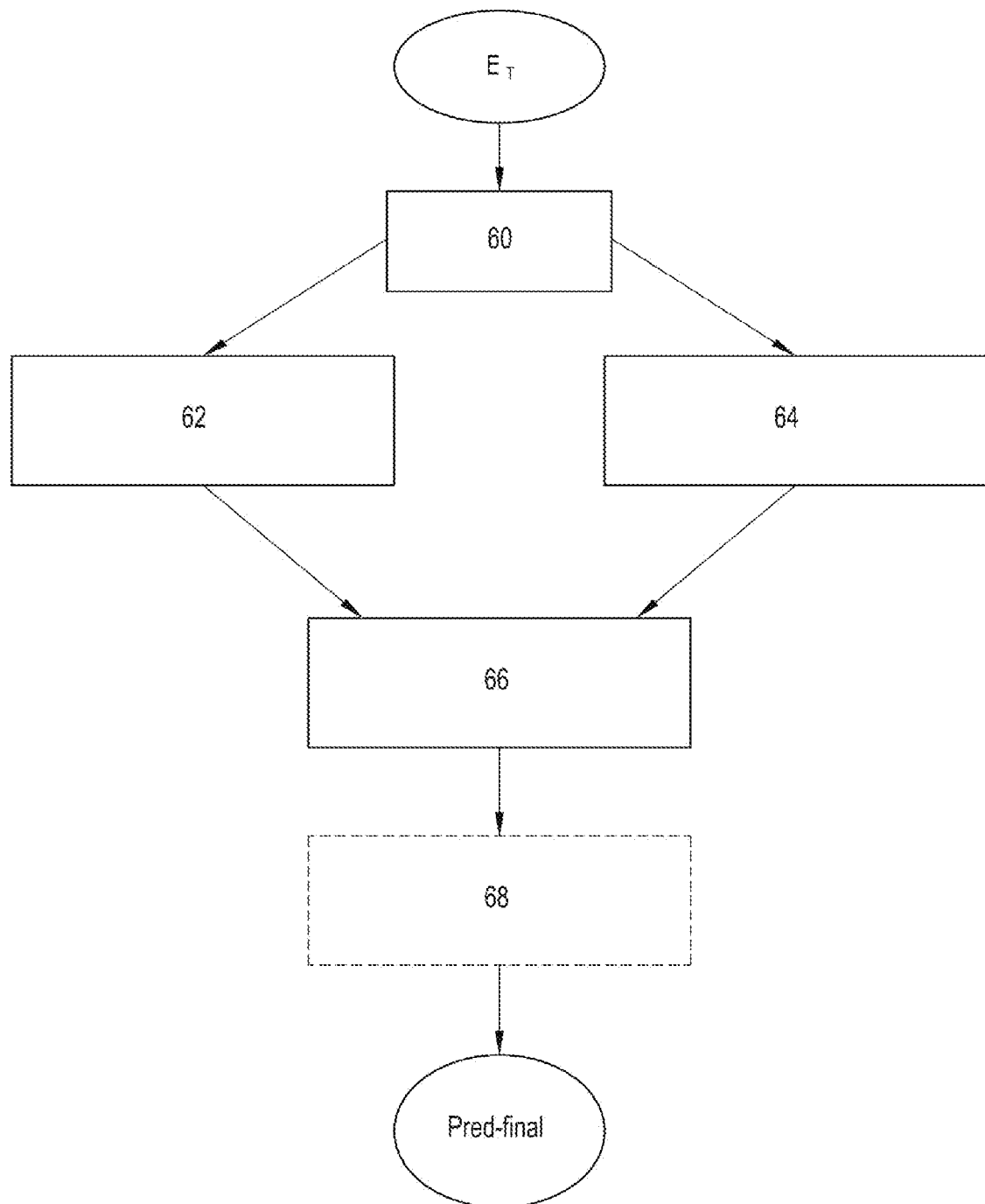
[Fig. 2]



[Fig. 3]



[Fig. 4]



**RAPPORT DE RECHERCHE
PRÉLIMINAIRE**

établi sur la base des dernières revendications
déposées avant le commencement de la recherche

N° d'enregistrement
national

FA 928528
FR 2312499

DOCUMENTS CONSIDÉRÉS COMME PERTINENTS		Revendication(s) concernée(s)	Classement attribué à l'invention par l'INPI
Catégorie	Citation du document avec indication, en cas de besoin, des parties pertinentes		
X	SHOKRI REZA ET AL: "Membership Inference Attacks Against Machine Learning Models", 2017 IEEE SYMPOSIUM ON SECURITY AND PRIVACY (SP), IEEE, 22 mai 2017 (2017-05-22), pages 3-18, XP033108084, DOI: 10.1109/SP.2017.41 [extrait le 2017-06-23] * page 1 - page 12; figures 1-3 * -----	1-10	G06N 20/20 G06N 3/08
			DOMAINES TECHNIQUES RECHERCHÉS (IPC)
			G06N
Date d'achèvement de la recherche		Examineur	
15 juillet 2024		Nourestani, S	
CATÉGORIE DES DOCUMENTS CITÉS		T : théorie ou principe à la base de l'invention E : document de brevet bénéficiant d'une date antérieure à la date de dépôt et qui n'a été publié qu'à cette date de dépôt ou qu'à une date postérieure. D : cité dans la demande L : cité pour d'autres raisons & : membre de la même famille, document correspondant	
X : particulièrement pertinent à lui seul Y : particulièrement pertinent en combinaison avec un autre document de la même catégorie A : arrière-plan technologique O : divulgation non-écrite P : document intercalaire			