الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي والبحث العلمي
جامعة سعيدة د. مولاي الطاهر
كلية التكنولوجيا
قسم: الإعلام الآلي

Mémoire de Master

Spécialité :Sécurité informatique et Cryptographie

Thème

Application d'une Meta-heuristique pour la détection d'intrusion : L'algorithme à essaim de particule (PSO)

Présenté par :
Mr Boukhobza Ibrahim

Dirigé par :
Mr Lokbani Ahmed Chaouki

Promotion 2021 - 2022

Remerciement

Je tiens à saisir cette occasion et adresser mes profonds remerciements et mes profondes reconnaissances à toutes personnes qui m'ont aidé de près ou de loin dans la réalisation de ce mémoire.

Je remercie Mr Lokbani Ahmed Chaouki pour l'encadrement, l'aide et l'encouragement. Grâce à ses conseils et ses orientations j'ai pu terminer ce travail.

Je remercie vivement l'enseignant Mr Hamou Reda Mohamed pour ses précieux conseils et orientations.

Je joins ces remerciements également aux membres du jury pour leur attention et intérêts portés envers ce travail

Je veut aussi adresser mes sincères remerciements à tous les enseignants de département de l'informatique qui ont contribué à ma formation.

Enfin, et surtout je remercie vivement toute ma famille notamment mes parents qui m'ont toujours encouragé dans la poursuite de mes études, ainsi que pour leur aide, leur compréhension et leur soutien sans oublier de remercier mes amis pour leurs soutiens et leurs bonnes humeurs pendant la préparation de ce mémoire.

Par crainte d'avoir oublié quelqu'un, que tous ceux et toutes celles dont je suis redevable se voie ici vivement remerciés.

Table de matières

table de figure:

Liste des tableaux:

Introduction général:

L' augmentations exigences de calcul augmentent chaque jour et le nombre des attaque qui se multiplien nos besoin en matière de sécurité forte , d'imposition de restrictions ont augmenté et la détection de tout intrusion que se soit de l'extérieur du système ou de l'intérieur.

Ces dernières années, de nombreux chercheurs et développeurs ont découvert des techniques de sécurité afin d'assurer une sécurité de haute qualité et de développer des fonctionnalités qui ne peuvent pas être fournies par les méthodes traditionnelles , et l'une de ces techniques est le système de détection d'intrusion. L'objectif de cette technologie est détecter toute violation de la politique de sécurité et d'avertir et d'alerter les principaux responsables, mais bien que l'IDS présente de nombreux avantages, il présente certains inconvénients notamment le taux élevé de fausses alarmes en plus de le temps de détection relativement élevé.

notre Problématique est comment améliorer les performances de l'IDS avec l'algorithme PSO

Chapitre 1

La sécurité informatique

1.1 Introduction:

À mesure que la technologies de l'information progresse et l'ouverture des systèmes d'information sur Internet, l'évolution de la technologie et des moyens de communication ainsi que la transmission de données à travers les réseaux,il a augmenté des risques d'accès et de manipulation des données par des gents non autorisées d'une façon accidentelle ou bien intentionnelle sont apparus. Donc la mise en place d'une politique de sécurité autour de ces systèmes est devenu une nécessité incontournable.

1.2 Définition de la sécurité informatique:

La sécurité informatique, la cybersécurité ou la sécurité des technologies de l'information (sécurité informatique) est la protection des systèmes et réseaux informatiques contre la divulgation d'informations, le vol ou l'endommagement de leur matériel, de leurs logiciels ou de leurs données électroniques, ainsi que contre la perturbation ou la mauvaise direction des services.[1][2]

1.3 Objectifs de la sécurité informatique:

1.3.1 La Confidentialité:

La confidentialité est la propriété, que l'information n'est pas mise à disposition ou divulguée à des personnes, entités ou processus non autorisé.[3]

1.3.2 L'Intégrité:

Dans le domaine de la sécurité informatique, l'intégrité des données signifie maintenir et garantir l'exactitude et l'exhaustivité des données tout au long de leur cycle de vie.[4] Cela signifie que les données ne peuvent pas être modifiées de manière non autorisée ou non détectée.[5]

1.3.3 La Disponibilité:

Pour qu'un système d'information remplisse son objectif, l'information doit être disponible lorsqu'elle est nécessaire.[6] Cela signifie que les systèmes informatiques utilisés pour stocker et traiter les informations, les contrôles de sécurité utilisés pour les protéger et les canaux de communication utilisés pour y accéder doivent fonctionner correctement[7]. D'autres aspects peuvent aussi être considérés comme des objectifs de la sécurité des systèmes d'information.

1.3.4 La traçabilité (ou preuve) :

garantie que les accès et tentatives d'accès aux éléments considérés sont tracés et que ces traces sont conservées et exploitables.[8]

1.3.5 La non-répudiation et l'imputation :

En droit, la non-répudiation implique l'intention d'une personne de remplir ses obligations contractuelles. Cela implique également qu'une partie à une transaction ne peut pas nier avoir reçu une transaction, ni que l'autre partie ne peut nier avoir envoyé une transaction.[9]

1.4 Définition de la politique de sécurité:

Une politique de sécurité est un plan d'action défini pour préserver l'intégrité et la pérennité d'un groupe social. Elle reflète la vision stratégique de la direction de l'organisme (PME, PMI, industrie, administration, État, unions d'États. . .)[10]

1.4.1 Les Objectifsde la politique de sécurité:

Une politique de sécurité a pour objectif de définir :

- les grandes orientations et les principes génériques à appliquer, techniques et organisationnels .

- Les responsables.
- L'organisation des différents acteurs.

1.5 Les Soucis de la sécurité informatique:

1.5.1 La vulnérabilité:

une vulnérabilité est une faiblesse dans un système informatique permettant à un attaquant de porter atteinte à l'intégrité de ce système, c'est-à-dire à son fonctionnement normal, à la confidentialité ou à l'intégrité des données qu'il contient.[11]

1.5.2 L' attaque:

Ce sont des opérations dont le but est de pénétrer,des réseaux informatiques,des systèmes d'information informatiques,des infrastructures ou des appareils informatiques personnels.un attaquant est personne ou un système essayant d'accéder à des information confidentielles , à des fonctionnalités ou à d'autres zones sensibles , dont le but peut être malveillant. Selon le contexte, les cyberattaques peuvent faire partie de la cyberterrorisme ou du cyberguerre.

1.5.3 Le menace :

Une menace c'est la possibilité la possibilité de pénétrer le système en exploitant une ou plusieurs vulnérabilités.

1.6 Les programmes malveillants :

Un logiciel malveillant (un portemanteau pour logiciel malveillant) : Ce sont des programmes malveillant dont le but d'espionner de révéler des informations non autorisées ou de désactiver un ordinateur, un serveur, un client ou un réseau informatique, isoler les utilisateurs et de les empêcher d'entrer dans le système et d'utiliser leur comptes à des fins malveillantes.

1.6.1 Logiciels malveillants infectieux:

Les types de logiciels malveillants les plus connus, les virus et les vers, sont connus pour la manière dont ils se propagent, plutôt que pour des types de comportement spécifiques et ont été assimilés à des virus biologiques.[12]

- Le Ver :

Un ver est un logiciel malveillant autonome qui se transmet activement sur un réseau pour infecter d'autres ordinateurs et peut se copier sans infecter les fichiers. Ces définitions conduisent à l'observation qu'un virus nécessite que l'utilisateur exécute un logiciel ou un système d'exploitation infecté pour que le virus se propage, alors qu'un ver se propage. [13]

- Le Virus:

Un virus informatique est un logiciel généralement caché dans un autre programme apparemment inoffensif qui peut produire des copies de lui-même et les insérer dans d'autres programmes ou fichiers, et qui effectue généralement une action nuisible (comme la destruction de données). [14]

- Le cheval de Troie (trojan) :

Un cheval de Troie est un programme nuisible qui se présente sous un faux jour pour se faire passer pour un programme ou un utilitaire régulier et bénin afin de persuader une victime de l'installer. Un cheval de Troie comporte généralement une fonction destructrice cachée qui est activée au démarrage de l'application. Le terme est dérivé de l'histoire grecque antique du cheval de Troie utilisé pour envahir furtivement la ville de Troie.[15]

- La porte dérobée (backdoor) :

Une porte dérobée est une méthode permettant de contourner les procédures d'authentification normales, généralement via une connexion à un réseau tel qu'Internet. Une fois qu'un système a été compromis, une ou plusieurs portes dérobées peuvent être installées afin de permettre l'accès à l'avenir.[16]

- Le Rootkit :

Une fois qu'un logiciel malveillant est installé sur un système, il est essentiel qu'il reste caché pour éviter d'être détecté. Les progiciels connus sous le nom de rootkits permettent cette dissimulation, en modifiant le système d'exploitation de l'hôte afin que le logiciel malveillant soit caché à l'utilisateur. Les rootkits peuvent empêcher un processus nuisible d'être visible dans la liste des processus du système ou empêcher la lecture de ses fichiers.[17]

1.7 Classification des attaques informatiques :

1.7.1 Selon l'objectif d'attaque:

On trouve deux types d'attaques principaux : passives et actives

- Les attaques actives :

quelqu'un essayant d'influencer de modifier les ressources système ou d'affecter leur fonctionnement.

- Les attaques passives :

quelqu'un essayant d'influencer d'utiliser ou d'apprendre des informations du système mais n'affecte pas les ressources du système (par exemple, l'écoute électronique).Une attaque peut être perpétrée par unl'extérieur ou de interne de l'organisation.

1.7.2 Selon le point d'initiation:

Il y a deux types d'attaques pour ce critère de classification: attaques de l'extérieur et attaques de l'intérieur :

- Attaquede l'intérieur :

est une attaque qui a fait dans l'intérieur ,c'est-à-dire une personne qui est autorisée à accéder aux ressources du système mais qui les utilise d'une manière non approuvée par ceux qui ont accordé l'autorisation.

- Attaque de l'extérieur :

est une attaque qui a fait dans l'extérieur du périmètre,c'est-à-dire une personne qui ne peut pas autorisée à accéder aux ressources du système , les attaquants extérieurs potentiels vont les gouvernements hostiles et les terroristes internationaux , en passant par des farceurs amateurs aux criminels organisés.

1.8 Les motivations des attaques informatiques:

Il existe plusieurs motivations d'un attaquant à vouloir exploiter une vulnérabilité et effectuer une attaque, parmi elles on peut citer les suivantes :

1.8.1 Les motivations financières :

prendre possession des données pour rançonner une personne ou l'entreprise.

1.8.2 Les motivations économique et concurrentielle :

on espionne un concurrent afin de voler des information sensibles ou effetuer des action malveillantes à son encontre afin de le détruire et obtenir un avantage commercial.

1.8.3 Les motivations politique ou idéologique :

C'est une guerre résultent divergnces politiques et destinée à déstabiliser les pays , à détruire leur structure économique et à créer des problèmes de manière indirecte.comme le piratage de l'entreprise Ashley Madison (Mansfield Devine, 2015) ou la cyberattaque Not Petya.

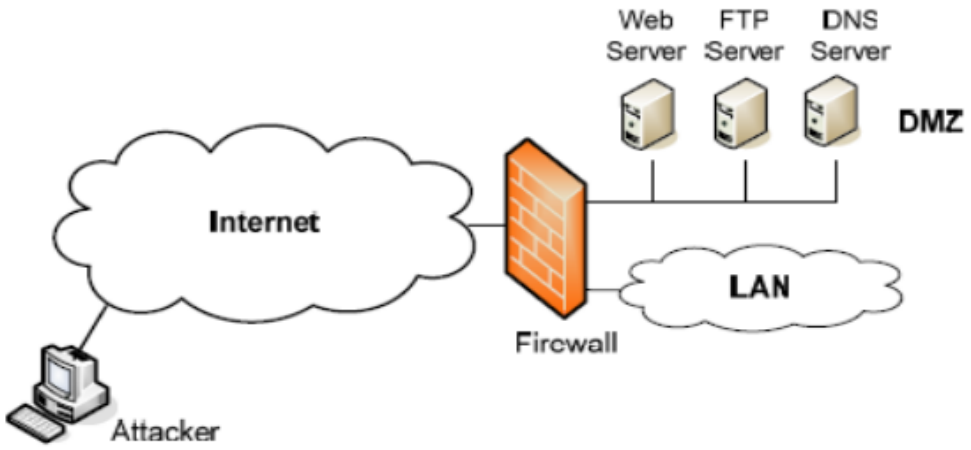
1.9 L'antivirus:

Un logiciel antivirus, également connu sous le nom d'anti-malware, est un programme informatique utilisé pour prévenir, détecter et supprimer les logiciels malveillants.

Le logiciel antivirus a été développé à l'origine pour détecter et supprimer les virus informatiques, d'où son nom. Cependant, avec la prolifération d'autres logiciels malveillants, les logiciels antivirus ont commencé à protéger contre d'autres menaces informatiques. En particulier, les logiciels antivirus modernes peuvent protéger les utilisateurs contre les objets d'assistance de navigateur malveillants (BHO), les pirates de navigateur, les ransomwares, les enregistreurs de frappe, les portes dérobées, les rootkits, les chevaux de Troie, les vers, les LSP malveillants, les numéroteurs, les outils de fraude, les logiciels publicitaires et les logiciels espions.[18]

1.10 Le pare-feu ou Firewall :

En informatique, un pare-feu est un système de sécurité réseau qui surveille et contrôle le trafic réseau entrant et sortant en fonction de règles de sécurité prédéterminées. Un pare-feu établit généralement une barrière entre un réseau approuvé et un réseau non approuvé, comme Internet.[19]



figuere 01 : Déploiement trihébergé d'un parefeu de réseau d'entreprise[19]

1.10.1 Le Filtre de paquets:

Le premier type de pare-feu réseau signalé est appelé filtre de paquets, qui inspecte les paquets transférés entre ordinateurs. Le pare-feu maintient une liste de contrôle d'accès qui dicte quels paquets seront examinés et quelle action doit être appliquée, le cas échéant, avec l'action par défaut définie sur rejet silencieux. Trois actions de base concernant le paquet consistent en une suppression silencieuse, une suppression avec Internet Control Message Protocol ou une réponse de réinitialisation TCP à l'expéditeur et une transmission au saut suivant.[20]

1.10.2 Le Suivi de connexion:

Flux de paquets réseau via Netfilter, un module du noyau Linux De 1989 à 1990, trois collègues d'ATT Bell Laboratories, Dave Presotto, Janardan Sharma et Kshitij Nigam, ont développé la deuxième génération de pare-feu, les appelants des passerelles au niveau du circuit. Les pare-feu de deuxième génération effectuent le travail de leurs prédécesseurs de première génération,

mais maintiennent également la connaissance des conversations spécifiques entre les points de terminaison en se souvenant du numéro de port que les deux adresses IP utilisent au niveau de la couche 4 (couche de transport) du modèle OSI pour leur conversation, permettant l'examen de l'échange global entre les nœuds.[21]

1.10.3 Couche d'application:

Marcus Ranum, Wei Xu et Peter Churchyard ont publié un pare-feu d'application connu sous le nom de Firewall Toolkit (FWTK) en octobre 1993. Cela est devenu la base du pare-feu Gauntlet chez Trusted Information Systems.

Le principal avantage du filtrage de la couche application est qu'il peut comprendre certaines applications et certains protocoles tels que le protocole de transfert de fichiers (FTP), le système de noms de domaine (DNS) ou le protocole de transfert hypertexte (HTTP). Cela lui permet d'identifier les applications ou services indésirables utilisant un port non standard, ou de détecter si un protocole autorisé est abusé.[22]

Chapitre 2

Le Système de détection d'intrusions

2.1 Introduction :

L'avancée phénoménale qui touche le domaine des technologies d'informations et de communications les rend de plus en plus faciles d'accès et d'usage, et la connectivité des systèmes d'informations de plus en plus grande, notamment à internet.

Cette connectivité en constante croissance, n'est pas sans conséquences. Les tentatives d'intrusions aux systèmes d'informations sont rendues plus facile grâce notamment aux failles, en constante croissance, découvertes dans les systèmes informatiques, augmentant de ce fait le risque d'attaques distantes.

Les techniques de préventions actuelles telles que les pare-feu et les antivirus permettent de diminuer le risque partiellement, mais il n'existe toujours pas un mécanisme de sécurité optimal.

De plus, des études ont montré que la plupart du temps, les attaques sont dues à des individus internes, ou bien le fait de personnes de leurs entourages. ce raisons ont conduit a l'émergence de Système de détection d'intrusion, dans ce chapitre Nous présenterons un concept pour détection d'intrusion.

2.2 Définition de Le Système de détection d'intrusions:

Un système de détection d'intrusion (IDS ; également système de prévention d'intrusion ou IPS) est un appareil ou une application logicielle qui surveille un réseau ou des systèmes pour détecter toute activité malveillante ou vio-

lation de politique. Toute activité d'intrusion ou violation est généralement signalée à un administrateur ou collectée de manière centralisée à l'aide d'un système de gestion des informations et des événements de sécurité (SIEM). Un système SIEM combine les sorties de plusieurs sources et utilise des techniques de filtrage des alarmes pour distinguer les activités malveillantes des fausses alarmes.[23]

2.2.1 Faux positif :

Cela signifie une tentative d'effraction, et c'est à ce moment paquet anodin est détecté (ne consistuant aucune menace réelle) .

2.2.2 Faux négatif :

C'est à dire lorsque la tentative d'intrusion n'est pas détectée.

2.2.3 Les tâches effectuées par un IDS:

1.Analyse et Surveillance des utilisateurs et activités du système.
2.Audit de la structure du système et erreur.
3.Mise en correspondance du modèle d'activité de reconnaissance connue attaques et alerte.
4.Analyse statistique des anomalies modèle de comportement.
5.Évaluer l'intégrité des systèmes et fichiers de données.

2.3 Anatomie d'une intrusion : [24]

2.3.1 Probe(Analyser) :

une personne mal intentionnée ,il veut trouver des failles et les exploiter pour pénétrer le réseau.

2.3.2 Penetrate (Pénétrer) :

Une fois trouvé une ou plusieurs failles identifiées, le pirate profite des failles pour pénétrer dans le system.

2.3.3 Persist (Peréniser) :

Le réseau infiltré, le pirate cherchera à y revenir facilement. Pour cela, il installera par exemple des back doors. Cependant, en général, il corrigera

la faille par laquelle il s'est introduit afin de s'assurer qu'aucun autre pirate n'exploitera sa cible.

2.3.4 Propagate (Propager) :

Le réseau est infiltré, l'accès est périnée.un pirate peut explorer le réseau et accéder à des informations sensible et gêner le fonctionnement du système.

2.3.5 Paralyze(Paralyser) :

Les objective sont fixés et le pirate nuira au système.

2.4 Les Attaques :

Une attaque est l'exploitation d'une faille d'un système informatique connecté à un réseau. Pour réussir leur exploit, les attaquants tentent d'appliquer un plan d'attaque bien précis pour aboutir à des objectifs distincts.[25]

2.4.1 Les différentes étapes d'une attaque :

La plupart des attaques, de la plus simple à la plus complexe fonctionnent suivant le même schéma :

- L'Identification de la cible :

cette étape est indispensable à toute attaque organisée, elle permet de récolter un maximum de renseignements sur la cible en utilisant des informations publiques et sans engager d'actions hostiles. On peut citer par exemple l'interrogation des serveurs DNS...

- Le Scanning :

l'objectif est de compléter les informations réunies sur une cible visées. Il est ainsi possible d'obtenir les adresses IP utilisées, les services accessibles de même qu'un grand nombre d'informations de topologie détaillée (OS, versions des services, règles de pare-feu...).

- L'Exploitation :

cette étape permet à partir des informations recueillies d'exploiter les failles identifiées sur les éléments de la cible, que ce soit au niveau protocolaire, des services et applications ou des systèmes d'exploitation présents sur le réseau.

- Progression :

il est temps pour l'attaquant de réaliser ce pourquoi il a franchit les précédentes étapes. Le but ultime étant d'obtenir les droits de l'utilisateur root (ou system) sur un système afin de pouvoir y faire tout ce qu'il souhaite (inspection de la machine, récupération d'informations, nettoyage des traces...).

2.4.2 Les différents types d'attaques :

Il existe un grand nombre de type d'attaques, que nous pouvons classifier endifférentes catégories selon plusieurs critères:

- Le sniffing :

grâce à un logiciel appelé "sniffer", il est possible d'intercepter toutes les trames que notre carte réseau reçoit et qui ne nous sont pas destinées. Si quelqu'un se connecte par Telnet par exemple à ce moment-là, son mot de passe transitant en clair sur le net, il sera donc aisé de le lire. De même, il est facile de savoir à tout moment quelles pages web sont consultées par les personnes connectées au réseau, les sessions ftp en cours, les mails en envoi ou réception. Un inconvénient de cette technique est de se situer sur le même réseau que la machine ciblée.

- L'IP spoofing :

cette attaque est difficile à mettre en œuvre et nécessite une bonne connaissance du protocole TCP. Elle consiste, le plus souvent, à se faire passer pour une autre machine en falsifiant son adresse IP de manière à accéder à un serveur ayant une "relation de confiance" avec la machine "spoofée". Cette attaque n'est intéressante que dans la mesure où la machine de confiance dont l'attaquant à pris l'identité peut accéder au serveur cible en tant que root.

- Les programmes cachés ou virus :

il existe une grande variété de virus. On ne classe cependant pas les virus d'après leurs dégâts mais selon leur mode de propagation et de multiplication. On recense donc les vers (capables de se propager dans le réseau), les troyens (créant des failles dans un système), Les bombes logiques (se lançant suite à un événement du système (appel d'une primitive, date spéciale)).

- Les scanners :

un scanner est un programme qui permet de savoir quels ports sont ouverts sur une machine donnée. Les Hackers utilisent les scanners pour savoir comment ils vont procéder pour attaquer une machine. Leur utilisation n'est heureusement pas seulement malsaine, car les scanners peuvent aussi permettre de prévenir une attaque. Le plus connu des scanners réseau est $\mathrm{WS}_{P}ingProPack$.

- Le SYN Flooding :

une connexion TCP s'établie en trois phases. Le SYN Flooding exploite ce mécanisme d'établissement en trois phases. Les trois étapes sont l'envoi d'un SYN, la réception d'un SYN-ACK et l'envoi d'un ACK. Le principe est de laisser sur la machine cible un nombre important de connexions TCP en attentes. Pour cela, l'attaquant envoie un très grand nombre de demandes de connexion, la machine cible renvoie les SYN-ACK en réponse au SYN reçus. L'attaquant ne répondra jamais avec un ACK, et donc pour chaque SYN reçu la cible aura une connexion TCP en attente.

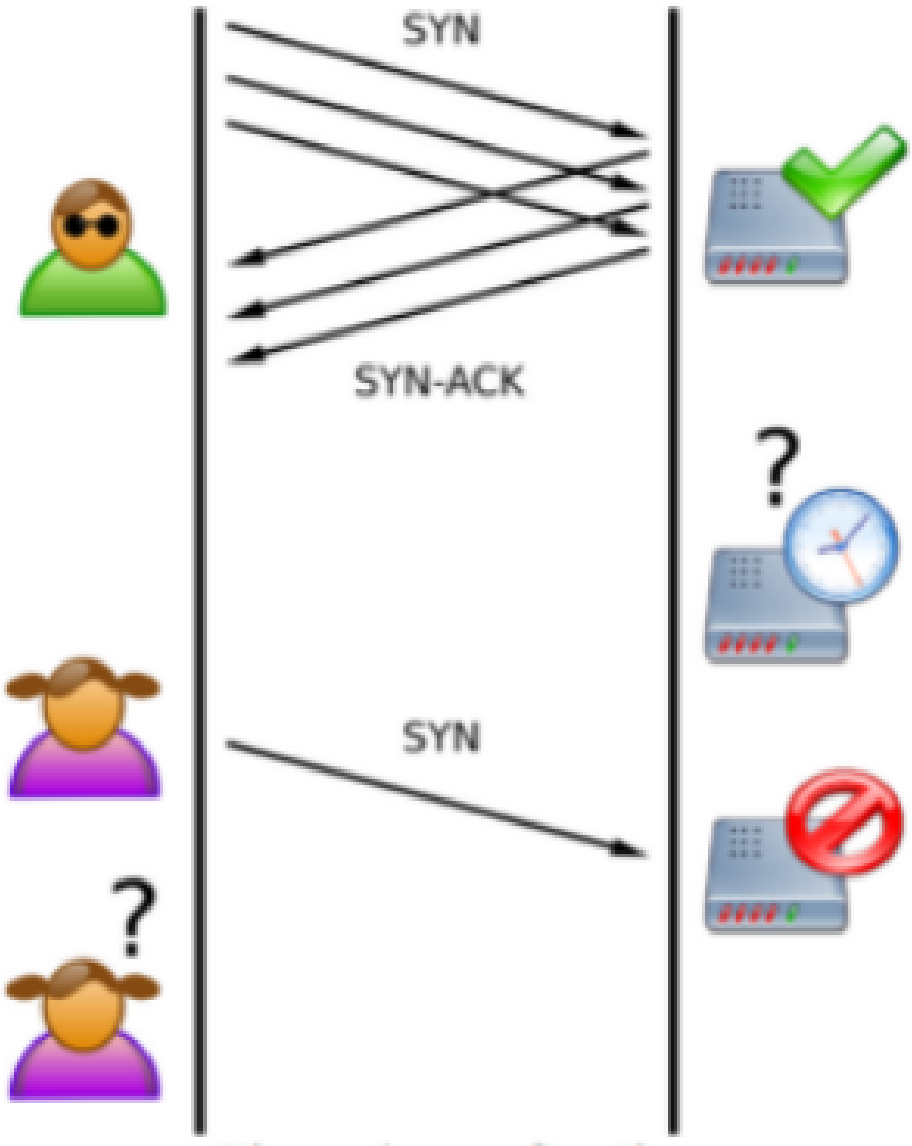


Figure 2:SYN Flooding[25]

Etant donné que ces connexions semi-ouvertes consomment des ressources mémoires au bout d'un certain temps, la machine est saturée et ne peut plus accepter de connexion. Ce type de déni de service n'affecte que la machine cible.

- Le DDoS :

"Distributed Denial of Service" ou déni de service distribué est un type d'attaque très évolué visant à faire planter ou à rendre muette une machine en la submergeant de trafic inutile. Plusieurs machines à la fois sont à l'origine de cette attaque (c'est une attaque distribuée) qui vise à anéantir des serveurs, des sous réseaux, etc. D'autre part, elle reste très difficile à contrer ou à éviter. C'est pour cela que cette attaque représente une menace que beaucoup craignent.

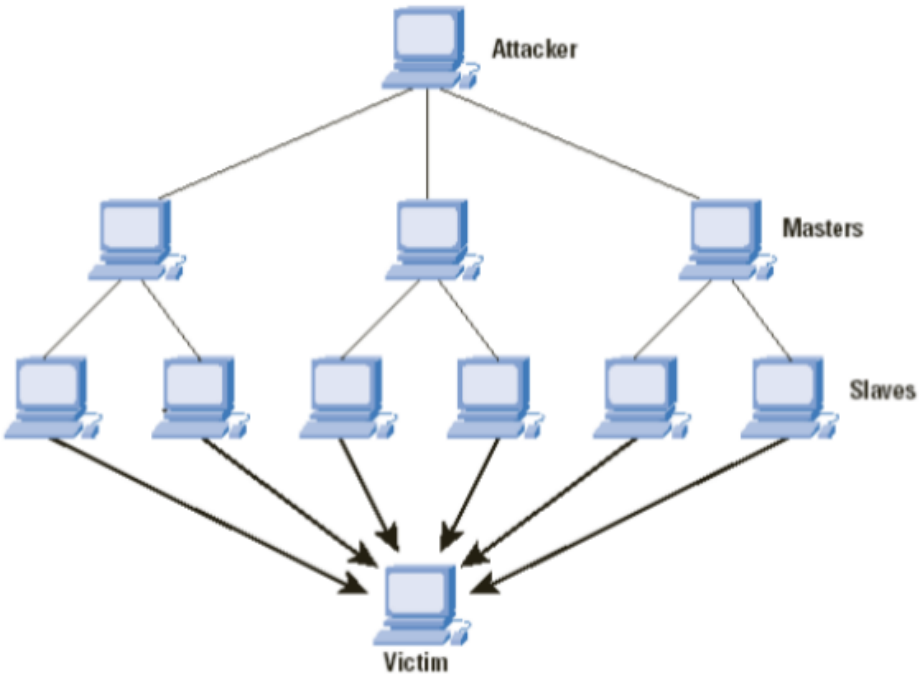


figure03:DDoS[25]

2.4.3 Les attaques en chiffres:

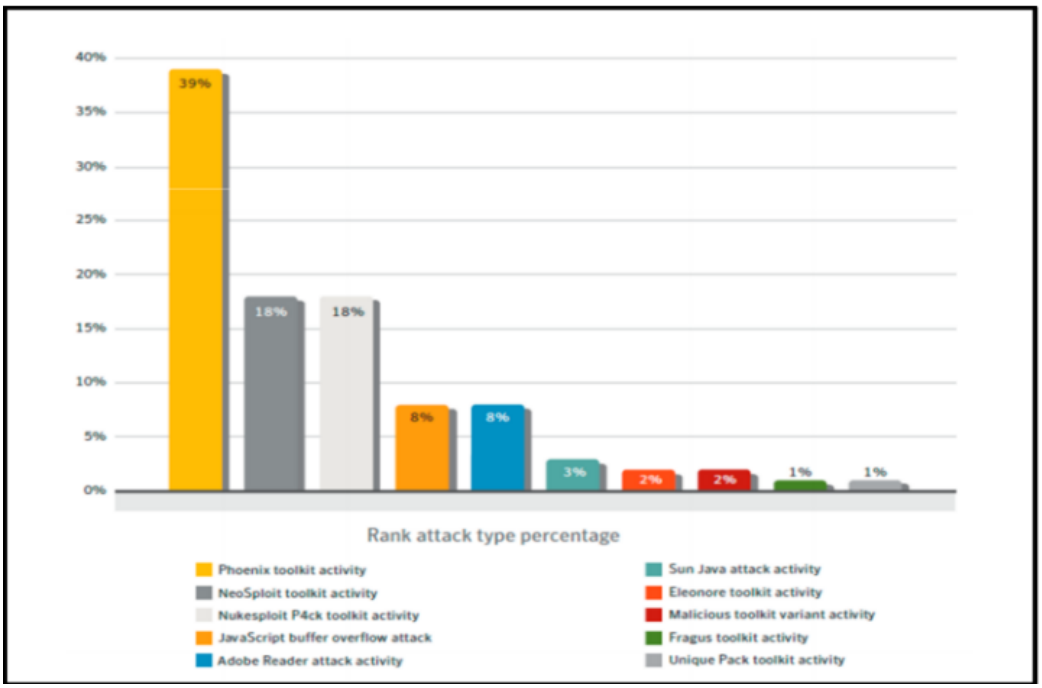
Un rapport construit à partir de données relevées sur le réseau Global Intelligence Network, ayant utilisé quelque 240000 capteurs dans plus de 200 pays nous a permis de tirer les conclusions suivantes :

- Le vol d'informations s'est proliféré très rapidement, du fait de l'utilisation à grande échelle des réseaux sociaux

- Ce dernier a couté 2.2 millions d'euros en France, en 2010, et 7.2 millions de dollars aux états unis la même année.

- L'utilisation grandissante des techniques de camouflages de malware à savoir les Rootkits , afin de mettre en œuvre des menaces de plus en plus furtives et difficilement détectables.

- L'explosion de l'utilisation des kits d'attaques, que ce soit par des experts ou bien des débutants, a permis d'exploiter des vulnérabilités Java afin d'attaquer plusieurs plateformes développées par feu Sun Microsystems.[26]



figuere04: Usage en hausse des kits d'attaques, notamment pour exploiter les failles Java[26]

2.5 Les Catégories de détection d'intrusion:

2.5.1 Les Systèmes de détection d'intrusion réseau:

Les systèmes de détection d'intrusion réseau (NIDS) sont placés à un ou plusieurs points stratégiques du réseau pour surveiller le trafic vers et depuis tous les appareils du réseau. Il effectue une analyse du trafic passant sur l'ensemble du sous-réseau et fait correspondre le trafic qui est transmis sur les sous-réseaux à la bibliothèque d'attaques connues. Une fois qu'une attaque est identifiée ou qu'un comportement anormal est détecté, l'alerte peut être envoyée à l'administrateur. Un exemple de NIDS serait de l'installer sur le sous-réseau où se trouvent les pare-feu afin de voir si quelqu'un essaie de pénétrer dans le pare-feu. Idéalement, il faudrait analyser tout le trafic entrant et sortant, mais cela pourrait créer un goulot d'étranglement qui nuirait à la vitesse globale du réseau.[27]

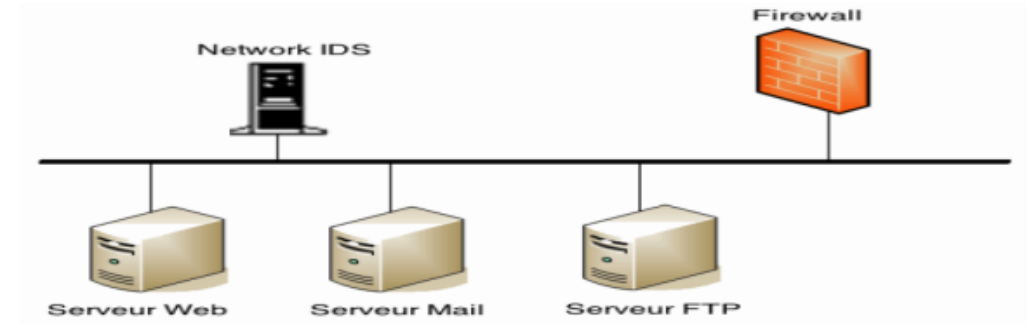


figure05: Système de détection d'intrusion réseau.[27]

On peut placer les capteurs dans deux endroits différents :

- A l'intérieur du pare-feu :

Si les capteurs se trouvent à l'intérieur du pare-feu, il sera plus facile de dire si le pare-feu a été mal configuré et nous pouvons ainsi savoir si une attaque est venue par ce pare-feu.

- A l'extérieur du pare-feu :

Les capteurs placés à l'extérieur du pare-feu servent à la détection et l'analyse d'attaques. Il offre l'avantage d'écrire dans les logs, ainsi l'administrateur voit ce qu'il doit modifier dans la configuration du pare-feu.

2.6 Les Systèmes de détection d'intrusion sur l'hôte :

Les systèmes de détection d'intrusion sur l'hôte (HIDS) s'exécutent sur des hôtes ou des périphériques individuels sur le réseau. Un HIDS surveille uniquement les paquets entrants et sortants de l'appareil et alerte l'utilisateur ou l'administrateur si une activité suspecte est détectée. Il prend un instantané des fichiers système existants et le fait correspondre à l'instantané précédent. Si les fichiers système critiques ont été modifiés ou supprimés, une alerte est envoyée à l'administrateur pour enquête. Un exemple d'utilisation de HIDS peut être observé sur des machines critiques, qui ne sont pas censées modifier leurs configurations.

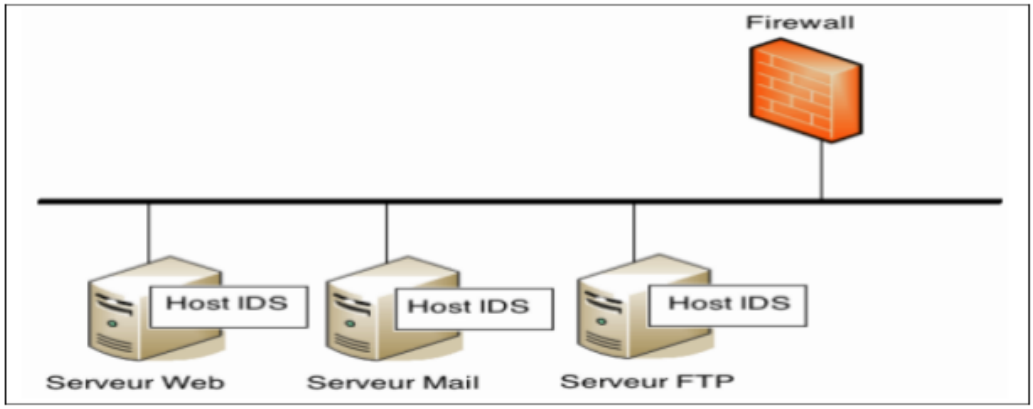


Figure 6 : Le Système de détection d'intrusion hôte.[27]

2.7 La Classification d'un IDS:

Les systèmes de prévention des intrusions peuvent être classés en quatre types différents :[28]

2.7.1 Système de prévention des intrusions basé sur le réseau (NIPS) :

surveille l'ensemble du réseau pour détecter tout trafic suspect en analysant l'activité du protocole.

2.7.2 Système de prévention des intrusions sans fil (WIPS) :

surveillez un réseau sans fil pour détecter tout trafic suspect en analysant les protocoles de réseau sans fil.

2.7.3 Analyse du comportement du réseau (NBA) :

examine le trafic réseau pour identifier les menaces qui génèrent des flux de trafic inhabituels, telles que les attaques par déni de service distribué (DDoS), certaines formes de logiciels malveillants et les violations de politique.

2.7.4 Système de prévention des intrusions basé sur l'hôte (HIPS) :

progiciel installé qui surveille un hôte unique pour détecter toute activité suspecte en analysant les événements se produisant au sein de cet hôte.

2.8 Les Méthodes de détection:

La majorité des systèmes de prévention des intrusions utilisent l'une des trois méthodes de détection : analyse basée sur les signatures, basée sur les anomalies statistiques et analyse de protocole avec état. [29]

2.8.1 La Détection basée sur les signatures :

l'IDS basé sur les signatures surveille les paquets dans le réseau et les compare avec des modèles d'attaque préconfigurés et prédéterminés connus sous le nom de signatures.

2.8.2 La Détection statistique basée sur les anomalies :

un IDS basé sur les anomalies surveillera le trafic réseau et le comparera à une base de référence établie. La ligne de base identifiera ce qui est "normal" pour ce réseau quel type de bande passante est généralement utilisé et quels protocoles sont utilisés. Il peut cependant déclencher une alarme de faux positif pour une utilisation légitime de la bande passante si les lignes de base ne sont pas configurées intelligemment. Les modèles d'ensemble qui utilisent le coefficient de corrélation de Matthews pour identifier le trafic réseau non autorisé ont obtenu une précision de 99,73

2.8.3 La Détection d'analyse de protocole avec état :

cette méthode identifie les déviations des états de protocole en comparant les événements observés avec des "profils prédéterminés de définitions généralement acceptées d'activité bénigne".

2.9 Les Principales approches en détection d'intrusions:

2.9.1 L'approche par scénario ou par signature:

Cette technique s'appuie sur la connaissance des techniques utilisées par les attaquants pour déduire des scénarios typiques. Elle ne tient pas compte des actions passées de l'utilisateur et utilise des signatures d'attaques existantes (ensemble de caractéristiques permettant d'identifier une activité intrusive : une chaîne alphanumérique, une taille de paquet inhabituelle, une trame formatée de manière suspecte, ...) .

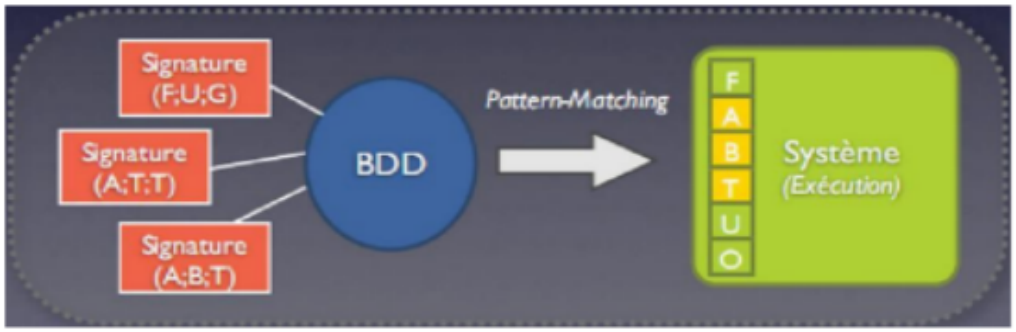


Figure 07 : Approche par scénario ou par signature

Cette technique se base sur:

- La recherche de motifs (pattern matching) :

C'est la méthode la plus connue et la plus facile à comprendre. Elle se base sur la recherche de motifs (chaînes de caractères ou suite d'octets) au sein du flux de données.

L'IDS comporte une base de signatures où chaque signature contient les protocoles et ports utilisés par une attaque spécifique ainsi que le motif qui permettra de reconnaître les paquets suspects.

De manière analogue, cette technique est également utilisée dans les anti-virus. En effet un anti-virus ne peut reconnaître un virus que si ce dernier est reconnu dans sa base de signatures virale, d'où la mise à jour régulière des anti-virus.

- Recherche de motifs dynamiques:

Le principe de cette méthode est le même que précédemment mais les signatures des attaques Évoluent dynamiquement. L'IDS est de ce fait doté de fonctionnalités d'adaptation et d'apprentissage.

- Analyse de protocoles:

Cette méthode se base sur une vérification de la conformitédes flux,ainsi que sur l'observation des champs et paramètres suspects dans les paquets. L'analyse protocolaire est souvent implémentée par un ensemble de préprocesseurs (programmes ou plug-in), où chaque préprocesseur est chargé d'analyser un protocole particulier (FTP, HTTP, ICMP, ...). Du fait de la présence de tous ces préprocesseurs, les performances dans un tel système s'en voient fortement dégradées (occupation du processeur). L'intérêt fort de l'analyse protocolaire est qu'elle permet de détecter des attaques inconnues, contrairement au pattern matching qui doit connaître l'attaque pour pouvoir la détecter.

- Analyse heuristique et détection d'anomalies:

Le but de cette méthode est, par une analyse intelligente, de détecter une activité suspecte ou toute autre anomalie (une action qui viole la politique de sécurité définie dans l' IDS) . Par exemple : une analyse heuristique permet de générer une alarme quand le nombre de pings vers un réseau ou hôte est très élevé ou incessant (Ping de la mort).

- L'approche comportementale (AnomalyDetection):

Cette technique consiste à détecter une intrusion en fonction du comportement passé de l'utilisateur. Il faut préalablement dresser un profil utilisateur à partir de ses habitudes et déclencher une alerte lorsque des événements hors profil se produisent. Cette technique peut être appliquée non seulement à des utilisateurs mais aussi à des applications et services.

Plusieurs métriques (paramètres) sont possibles la charge CPU, le volume de données échangées, le temps de connexion sur des ressources, la répartition statistique des protocoles et applications utilisés, les heures de connexion, ...

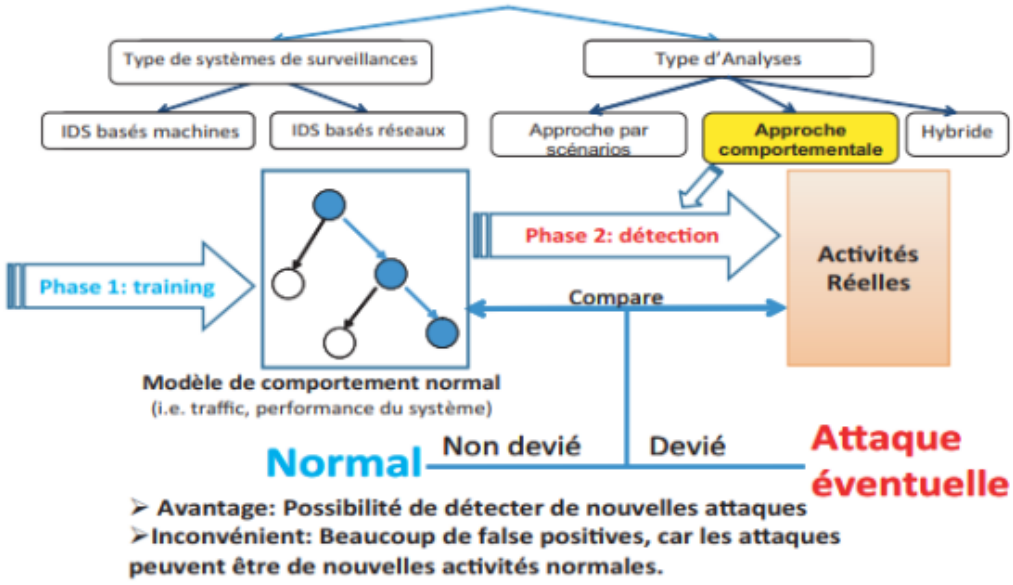


Figure 08 : L'Approche comportementale

2.9.2 L'approche de détection d'anomalies:

Dans le système d'utilisation abusive, le signature des attaques connues sont stockées dans la base de données. Toutes les données similaires à ces données sont classées comme des attaques. Détection d'une anomalie fait référence à la connaissance statistique de l'activité normale. Le L'approche de détection d'anomalies peut être classée en détection d'anomalies semi-supervisée et non supervisée.

- Les approches de détection d'anomalies semi-supervisées:

elles nécessitent un ensemble de données d'entraînement normal à partir desquelles ils ont trouvé le profil de comportement normal. Si les données d'entraînement contiennent des attaques cachées à l'intérieur, l'approche peut ne pas détecter de futurs cas de ces attaques.

- Une anomalie non supervisée:

Les approches de détection établissent le profil du comportement normal avec des données d'entraînement non étiquetées composées à la fois des données normales et échantillons anormaux. Les intrusions correspondent à des écarts par rapport à l'activité normale du système. Le système de

détection d'anomalies a taux élevé de fausses alarmes positives/négatives par rapport à une mauvaise utilisation systèmes de détection.

2.10 L'architecture d'un IDS :

Cette section décrit les trois composants qui constituent classiquement un système de détection d'intrusions. La Figure illustre les interactions entre ces trois composants.[30]

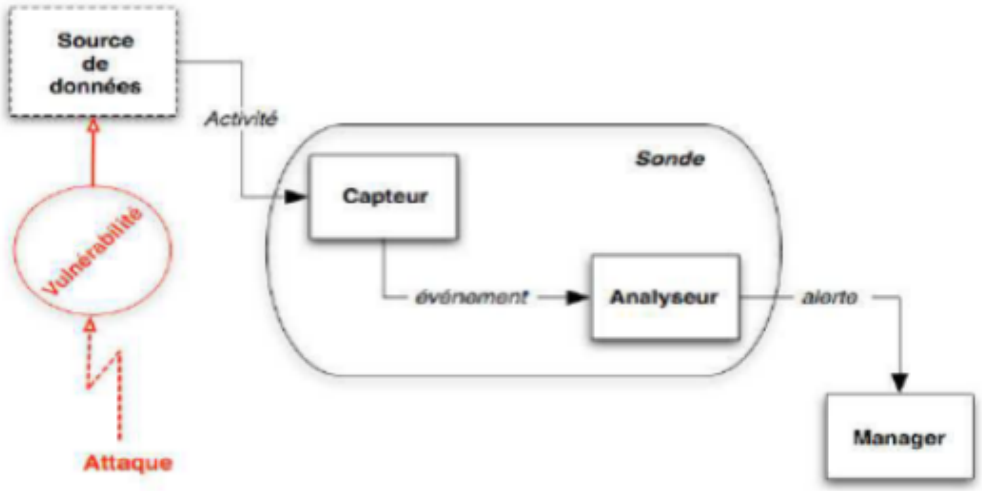


Figure 9 : L'architecture d'un IDS.[30]

2.10.1 Le Capteur :

Le capteur observe l'activité du système par le biais d'une source de données et fournit à l'analyseur une séquence d'événements qui renseignent de l'évolution de l'état du système. Le capteur peut se contenter de transmettre directement ces données brutes, mais en général un prétraitement est effectué. On distingue classiquement trois types de capteurs en fonction des sources de données utilisées pour observer l'activité du système : les capteurs système, les capteurs réseau et les capteurs applicatifs.

2.10.2 L'analyseur :

L'objectif de l'analyseur est de déterminer si le flux d'événements fourni par le capteur contient des éléments caractéristiques d'une activité malveillante.

2.10.3 Me manager :

Le manager collecte les alertes produites par le capteur, les met en forme et les présente à l'opérateur. Éventuellement, le manager est chargé de la réaction à adopter qui peut être :

- Le confinement de l'attaque, qui a pour but de limiter les effets de l'attaque.
- L'éradication de l'attaque, qui tente d'arrêter l'attaque.
- Le Recouvrement, qui est l'étape de restauration du système dans un état sain.
- Le Diagnostic, qui est la phase d'identification du problème

2.11 Mise en place d'un IDS :

Le positionnement de l'IDS : Il existe plusieurs endroits stratégiques où il convient de placer un IDS. Le schéma suivant illustre un réseau local ainsi que les trois positions que peut y prendre un IDS :[31]

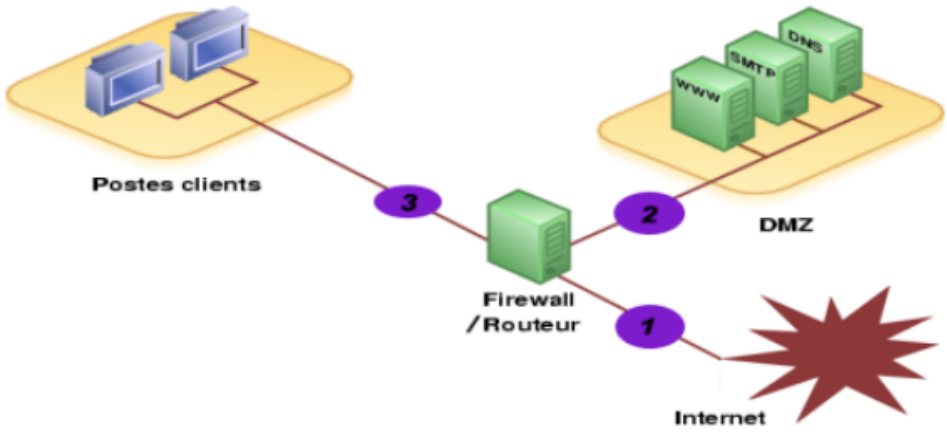


Figure 10: La position des IDS.[31]

-La Position (1):

Sur cette position, l'IDS va pouvoir détecter l'ensemble des attaques frontales, provenant de l'extérieur, en amont du firewall. Ainsi, beaucoup d'alertes seront remontées ce qui rendra les logs difficilement consultables.

- Position (2):

Si l'IDS est placé sur la DMZ, il détectera les attaques qui n'ont pas été filtrées par le firewall et qui relèvent d'un certain niveau de compétence. Les logs seront ici plus clairs à consulter puisque les attaques ne seront pas recensées.

-Position (3):

L'IDS peut ici rendre compte des attaques internes, provenant du réseau local de l'entreprise. Il peut être judicieux d'en placer un à cet endroit étant donné le fait que 80 pourcent des attaques proviennent de l'intérieur. De plus, si des trojans ont contaminé le parc informatique (navigation peu méfiante sur internet) ils pourront êtres ici facilement identifiés pour être ensuite éradiqués.

2.12 Critères de Choix D'un IDS:

Les systèmes de détection d'intrusion sont devenus indispensables lors de la mise en place d'une infrastructure de sécurité opérationnelle. Ils s'intègrent donc toujours dans un contexte et dans une architecture imposants des contraintes très diverses. Certains critères imposant le choix d'un IDS peuvent être dégagés:

2.12.1 La Fiabilité :

Les alertes générées doivent être justifiées et aucune intrusion ne doit pouvoir lui échapper.

2.12.2 La Réactivité :

Un IDS doit être capable de détecter les nouveaux types d'attaques le plus rapidement possible; pour cela il doit rester constamment à jour. Des capacités de mise à jour automatique sont indispensables.

2.12.3 La Facilité de mise en œuvre et adaptabilité :

Un IDS doit être facile à mettre en œuvre , surtout s'adapter au contexte dans lequel il doit opérer. Il est inutile d'avoir un IDS émettant des alertes en moins de 10 secondes si les ressources nécessaires à une réaction ne sont pas disponibles pour agir dans les mêmes contraintes de temps.

2.12.4 La Performance :

la mise en place d'un IDS ne doit en aucun cas affecter les performances des systèmes surveillés. De plus, il faut toujours avoir la certitude que l'IDS a la capacité de traiter toute l'information à sa disposition (par exemple un IDS réseau doit être capable de traiter l'ensemble du flux pouvant se présenter à un instant donné sans jamais supprimer de paquets) car dans le cas contraire il devient trivial de masquer les attaques en augmentant la quantité d'information.

2.13 Conclusion :

Dans ce chapitre, nous avons abordé en détails les systèmes de détection d'intrusions IDS ainsi que les attaques et leurs types. Nous avons parlé des principales Méthodes de détection ainsi que des approches de détection d'intrusions. Enfin, nous avons conclu que le problème de l'IDS est un problème de classification, et plus précisément c'est un problème de sélection d'attributs car c'est dans ce sens que nous pourrons améliorer les résultats (augmenter la précision de la classification et réduction du temps de calcul par exemple).

Chapitre 3

Les méta-heuristiques

3.1 Introduction :

Durant ces vingt dernières années de nouveaux types d'algorithmes, appelés métaheuristiques ont vu le jour et ne cessent de se développer. Ces algorithmes, ont été introduit , jusqu'à une certaine période, on les appelait les heuristiques modernes . Ils ont été développés pour résoudre des problèmes d'optimisation combinatoire de plus en plus complexes, et sur lesquelles les méthodes exactes commencent à montrer leurs limites. En effet, ces dernières effectuent une recherche exhaustive de l'espace de solution afin de trouver la solution optimale, ce qui les rend inutilisables pour des problèmes de la vie réelle.

Les métaheuristiques sont basées sur des heuristiques afin d'explorer l'espace de solutions de manière efficace et non exhaustive, surtout si ce dernier est très grand. Ils sont une évolution logique dans le temps des heuristiques . Le terme métaheuristique est une composition de deux mots d'origines grecques heuristique qui a pour sens trouver et méta qui veut dire à un niveau supérieur . Parmi ces algorithmes on peut citer : la recherche Tabou, les algorithmes génétiques, les intelligences en essaims (Les colonies de fourmis, Les essaims d'abeilles), etc.

Afin de résoudre un problème d'optimisation combinatoire, on utilise généralement les heuristiques spécialisées qui sont développées afin de résoudre un problème particulier d'optimisation combinatoire. Leur principal inconvénient est qu'elles ne peuvent être appliquées qu'à un problème donnée et que le résultat obtenu ne pourra pas être appliqué sur une autre classe de problèmes différents.

3.2 Définition des méta-heuristiques:

Une métaheuristique est un algorithme d'optimisation visant à résoudre des problèmes d'optimisation difficile (souvent issus des domaines de la recherche opérationnelle, de l'ingénierie ou de l'intelligence artificielle) pour lesquels on ne connaît pas de méthode classique plus efficace.

Les métaheuristiques sont généralement des algorithmes stochastiques itératifs, qui progressent vers un optimum global (c'est-à-dire l'extremum global d'une fonction), par échantillonnage d'une fonction objectif. Elles se comportent comme des algorithmes de recherche, tentant d'apprendre les caractéristiques d'un problème afin d'en trouver une approximation de la meilleure solution (d'une manière proche des algorithmes d'approximation).

Il existe un grand nombre de métaheuristiques différentes, allant de la simple recherche locale à des algorithmes complexes de recherche globale. Ces méthodes utilisent cependant un haut niveau d'abstraction, leur permettant d'être adaptées à une large gamme de problèmes différents [32].

3.3 Propriétés:

Ce sont des propriétés qui caractérisent la plupart des méta-heuristiques [33]:

- L'objectif est d'explorer efficacement l'espace de recherche afin de trouver des solutions quasi optimales.
- Les techniques qui constituent les algorithmes méta-heuristiques vont des simples procédures de recherche locale aux processus d'apprentissage complexes.
- Les algorithmes méta-heuristiques sont approximatifs et généralement non déterministes.
- Les méta-heuristiques ne sont pas spécifiques à un problème.
- Les méta-heuristiques sont des stratégies qui guident le processus de recherche.

3.4 Classification des méta-heuristiques :

Il existe une grande variété de méta-heuristiques [34] et un certain nombre de propriétés par rapport auxquelles on les classe [33].

3.4.1 Recherche locale vs recherche globale:

Une approche consiste à caractériser le type de stratégie de recherche.[33]Un type de stratégie de recherche est une amélioration des algorithmes de recherche

locale simples. Un algorithme de recherche locale bien connu est la méthode d'escalade qui est utilisée pour trouver des optimums locaux. Cependant, l'escalade ne garantit pas de trouver des solutions optimales globales. De nombreuses idées de méta-heuristiques ont été proposées pour améliorer l'heuristique de recherche locale afin de trouver de meilleures solutions. Ces méta-heuristiques incluent le recuit simulé, la recherche tabou, la recherche locale itérée, la recherche de voisinage variable et GRASP.[33] Ces méta-heuristiques peuvent être classées en méta-heuristiques de recherche locale ou globale.

Les autres méta-heuristiques de recherche globale qui ne sont pas basées sur la recherche locale sont généralement des méta-heuristiques basées sur la population, comme l'optimisation par colonies de fourmis, le calcul évolutif, l'optimisation par essaims de particules (PSO), l'algorithme génétique et algorithme d'optimisation du pilote [35].

3.4.2 Solution unique ou basée sur la population:

Une autre dimension de classification est la solution unique par rapport aux recherches basées sur la population [33]. Les approches de solution unique se concentrent sur la modification et l'amélioration d'une solution candidate unique ; les méta-heuristiques à solution unique incluent le recuit simulé, la recherche locale itérée, la recherche de voisinage variable et la recherche locale guidée. Les approches basées sur la population maintiennent et améliorent plusieurs solutions candidates, en utilisant souvent les caractéristiques de la population pour guider la recherche ; les méta-heuristiques basées sur la population comprennent le calcul évolutif, les algorithmes génétiques et l'optimisation des essaims de particules. Une autre catégorie de méta-heuristiques est l'intelligence en essaim qui est un comportement collectif d'agents décentralisés et auto-organisés dans une population ou un essaim. L'optimisation des colonies de fourmis, l'optimisation des essaims de particules, l'optimisation cognitive sociale sont des exemples de cette catégorie [36] [37].

3.4.3 Hybridation et algorithmes mémétiques:

Une méta-heuristique hybride est une méta-heuristique qui combine une méta-heuristique avec d'autres approches d'optimisation, telles que des algorithmes de programmation mathématique, de programmation par contraintes et d'apprentissage automatique. Les deux composants d'une méta-heuristique hybride peuvent s'exécuter simultanément et échanger des informations pour guider la recherche. D'autre part, les algorithmes mémétiques

représentent la synergie d'une approche évolutive ou de toute approche basée sur la population avec des procédures d'apprentissage individuel ou d'amélioration locale séparées pour la recherche de problèmes. Un exemple d'algorithme mémétique est l'utilisation d'un algorithme de recherche locale au lieu d'un opérateur de mutation de base dans les algorithmes évolutionnaires. [38]

3.4.4 Méta-heuristiques parallèles :

Une méta-heuristique parallèle est une méta-heuristique qui utilise les techniques de programmation parallèle pour exécuter plusieurs recherches méta-heuristiques en parallèle ; ceux-ci peuvent aller de simples schémas distribués à des exécutions de recherche simultanées qui interagissent pour améliorer la solution globale.

3.4.5 Méta-heuristiques inspirées de la nature et basées sur des métaphores :

Un domaine de recherche très actif est la conception de méta-heuristiques inspirées de la nature. De nombreuses méta-heuristiques récentes, en particulier les algorithmes basés sur le calcul évolutionnaire, s'inspirent des systèmes naturels. La nature agit comme une source de concepts, de mécanismes et de principes pour la conception de systèmes informatiques artificiels pour faire face à des problèmes informatiques complexes. Ces méta-heuristiques comprennent le recuit simulé, les algorithmes évolutionnaires, l'optimisation des colonies de fourmis et l'optimisation des essaims de particules. Un grand nombre de méta-heuristiques plus récentes inspirées par les métaphores ont commencé à susciter des critiques dans la communauté des chercheurs pour avoir caché leur manque de nouveauté derrière une métaphore élaborée.[39]

3.4.6 Cadres d'optimisation des méta-heuristiques (MOF) :

Un MOF peut être défini comme " un ensemble d'outils logiciels qui fournissent une mise en œuvre correcte et réutilisable d'un ensemble de méta-heuristiques, et les mécanismes de base pour accélérer la mise en œuvre de ses heuristiques subordonnées partenaires (y compris éventuellement des encodages de solution et des opérateurs spécifiques à la technique) , qui sont nécessaires pour résoudre un cas de problème particulier en utilisant les techniques fournies".[40]

Il existe de nombreux outils d'optimisation candidats qui peuvent être considérés comme un MOF de fonctionnalités variables : Comet, EvA2, evolvica, Evolutionary :: Algorithm, GAPlayground, jaga, JCLEC, JGAP, jMetal, n-genes, Open Beagle, Opt4j, ParadisEO/EO , Pisa, Watchmaker, FOM, Hypercube, HotFrame, Templar, EasyLocal, iOpt, OptQuest, JDEAL, Optimization Algorithm Toolkit, HeuristicLab, MAFRA, Localizer, GALIB, DREAM, Discropt, MALLBA, MAGMA, Metaheuristics.jl, UOF et Opta-Planificateur .[40]

3.5 Définition du Data mining :

L'exploration de données est un processus d'extraction et de découverte de modèles dans de grands ensembles de données impliquant des méthodes à l'intersection de l'apprentissage automatique, des statistiques et des systèmes de bases de données. L'exploration de données est un sous-domaine interdisciplinaire de l'informatique et des statistiques dont l'objectif global est d'extraire des informations (avec des méthodes intelligentes) d'un ensemble de données et de transformer les informations en une structure compréhensible pour une utilisation ultérieure 41][42][43][44]. L'exploration de données est l'étape d'analyse du processus de "découverte des connaissances dans les bases de données", ou KDD.[45] Outre l'étape d'analyse brute, cela implique également des aspects de base de données et de gestion des données, le prétraitement des données, des considérations de modèle et d'inférence, des mesures d'intérêt, des considérations de complexité, le post-traitement des structures découvertes, la visualisation et la mise à jour en ligne.

Figure 11: Les fonction du data maning.[45]

3.6 Les différentes méthodes du Data Mining [46]:

La fouille de données permet de faire:

3.6.1 L'association:

recherche de patterns au seins desquels un évènement est lié à un autre.

3.6.2 L'analyse de séquences:

recherche de patterns au seins desquels un évènement mène à un autre évènement futur.

3.6.3 La classification:

classer de nouveau items en fonction de leurs caractéristiques.

3.6.4 Le clustering:

trouver des groupes de faits précédemment inconnus.

3.7 Les processus du data maning:[47]

3.7.1 Définition du problème:

Quel est le but de l'analyse, que recherche-t-on ? Quels sont les objectifs ? Comment traduire le problème en une question pouvant servir de sujet d'enquête pour cet outil d'analyse bien spécifique ? À ce sujet, se souvenir que l'on travaille à partir des données existantes. La question doit être ciblée selon les données disponibles

3.7.2 Collecte des données:

Une phase absolument essentielle. On n'analyse que des données utilisables, c'est-à-dire "propres" et consolidées. On n'hésitera pas à extraire de l'analyse les données de qualité douteuse. Bien souvent, les données méritent d'être retravaillées. S'assurer au final que la quantité de données soit suffisante pour éviter de fausser les résultats. Cette phase de collecte nécessite le plus grand soin. Voir en seconde partie de l'article un cas concret de projet Datamining où la qualité de la collecte laisse un peu à désirer...

3.7.3 Construire le modèle d'analyse:

Ne pas hésiter à valider vos choix d'analyse sur plusieurs jeux d'essais en variant les échantillons. Une première évaluation peut nous conduire à reprendre les points 1 ou 2.

3.7.4 Étude des résultats:

Il est temps d'exploiter les résultats. Pour affiner l'analyse, on n'hésitera pas à reprendre les points 1, 2 ou 3 si les résultats s'avéraient insatisfaisants. C'est-à-dire qu'ils ne seraient pas en phase avec les objectifs fixés au temps 1.

3.7.5 Formalisation et diffusion:

Les résultats sont formalisés pour être diffuser. Ils ne seront utiles qu'une fois devenus une connaissance partagée. C'est bien là l'aboutissement de la démarche. C'est aussi là que réside la difficulté d'interprétation et de généralisation.

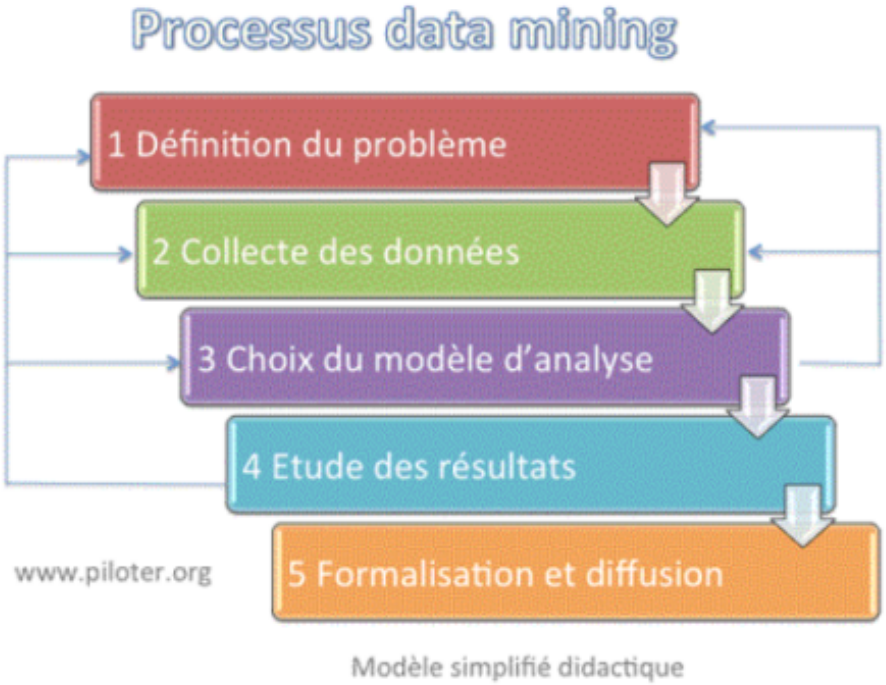


Figure 12: Les processus du data maning .[46]

3.8 Types du Data Mining:

Le Data Mining peut être effectuée sur les types de données suivants [48] :

3.8.1 Base de données relationnelle :

Une base de données relationnelle est une collection de plusieurs ensembles de données formellement organisés par des tables, des enregistrements et des colonnes à partir desquelles les données sont accessibles de différentes manières sans avoir à reconnaître les tables de la base de données. Les tableaux transmettent et partagent des informations, ce qui facilite la recherche, la création de rapports et l'organisation des données.

3.8.2 Entrepôts de données :

Un entrepôt de données est la technologie qui collecte les données de diverses sources au sein de l'organisation pour fournir des informations commerciales significatives. L'énorme quantité de données provient de plusieurs endroits tels que le marketing et les finances. Les données extraites sont utilisées à des fins analytiques et aident à la prise de décision pour une organisation commerciale. L'entrepôt de données est conçu pour l'analyse des données plutôt que pour le traitement des transactions.

3.8.3 Référentiels de données :

Le référentiel de données fait généralement référence à une destination pour le stockage des données. Cependant, de nombreux professionnels de l'informatique utilisent le terme plus clairement pour désigner un type spécifique de configuration au sein d'une structure informatique. Par exemple, un groupe de bases de données, où une organisation a conservé divers types d'informations

3.8.4 Base de données relationnelle objet :

Une combinaison d'un modèle de base de données orienté objet et d'un modèle de base de données relationnelle est appelée un modèle relationnel objet. Il prend en charge les classes, les objets, l'héritage, etc. L'un des principaux objectifs du modèle de données relationnel objet est de combler l'écart entre la base de données relationnelle et les pratiques de modèle orienté objet fréquemment utilisées dans de nombreux langages de programmation, par exemple, C++, Java, C, etc.

3.8.5 Base de données transactionnelle :

Une base de données transactionnelle fait référence à un système de gestion de base de données (SGBD) qui a le potentiel d'annuler une transaction de base de données si elle n'est pas effectuée correctement. Même s'il s'agissait d'une capacité unique il y a très longtemps, aujourd'hui, la plupart des systèmes de bases de données relationnelles prennent en charge les activités de bases de données transactionnelles.

3.9 Avantage et inconvénient du Data Mining [48]:

3.9.1 Avantages du Data Mining :

- La technique de Data Mining permet aux organisations d'obtenir des données basées sur la connaissance.

- Data Mining permet aux organisations d'apporter des modifications lucratives au fonctionnement et à la production.

- Par rapport à d'autres applications de données statistiques, Data Mining est rentableLe Data Mining aide au processus décisionnel d'une organisation.

- Il facilite la découverte automatisée des modèles cachés ainsi que la prédiction des tendances et des comportements.

- Elle peut être induite dans le nouveau système ainsi que dans les plateformes existantes.

- Il s'agit d'un processus rapide qui permet aux nouveaux utilisateurs d'analyser facilement d'énormes quantités de données en peu de temps.

3.9.2 Inconvénients du Data Mining :

- Il est probable que les organisations vendent des données utiles sur les clients à d'autres organisations contre de l'argent. Selon le rapport, American Express a vendu les achats par carte de crédit de ses clients à d'autres organisations.

- De nombreux logiciels d'analyse d'exploration de données sont difficiles à utiliser et nécessitent une formation préalable pour fonctionner.

- Différents instruments d'exploration de données fonctionnent de manière distincte en raison des différents algorithmes utilisés dans leur conception. Par conséquent, la sélection des bons outils d'exploration de données est une tâche très difficile.

- Les techniques d'exploration de données ne sont pas précises, de sorte qu'elles peuvent entraîner des conséquences graves dans certaines conditions.

3.10 Définition de la classification:

La classification est une fonction Data Mining qui affecte des éléments d'une collection à des catégories ou classes cibles. L'objectif de la classification est de prédire avec précision la classe cible pour chaque cas dans les données. Par exemple, un modèle de classification pourrait être utilisé pour identifier les

demandeurs de prêt comme présentant des risques de crédit faibles, moyens ou élevés

Bien que les outils d'analyse de données mettent davantage l'accent sur le libre-service, il est toujours utile de savoir quelle opération Data Mining convient à vos besoins avant de commencer une opération Data Mining. [49]

3.11 Les types de la classification :

Les techniques du Data Mining se présentent sous deux formes principales : supervisée (également appelée prédictive ou dirigée) et non supervisée (également appelée descriptive ou non dirigée). Les deux catégories englobent des fonctions capables de trouver différents modèles cachés dans de grands ensembles de données.[50]

3.11.1 Classification supervisé:

Les techniques du Data Mining supervisée sont appropriées lorsque vous avez une valeur cible spécifique que vous souhaitez prédire sur vos données. Les cibles peuvent avoir deux ou plusieurs résultats possibles, ou même être une valeur numérique continue (plus sur cela plus tard).

Pour utiliser ces méthodes, vous disposez idéalement d'un sous-ensemble de points de données pour lesquels cette valeur cible est déjà connue. Vous utilisez ces données pour créer un modèle de ce à quoi ressemble un point de données typique lorsqu'il a l'une des différentes valeurs cibles. Vous appliquez ensuite ce modèle aux données pour lesquelles cette valeur cible est actuellement inconnue. L'algorithme identifie les "nouveaux" points de données qui correspondent au modèle de chaque valeur cible. Maintenant, clarifions cela avec quelques démonstrations spécifiques [50] :

- Classification :

En tant que méthode du Data Mining supervisée, la classification commence par la méthode décrite ci-dessus. Imaginez que vous êtes une société de cartes de crédit et que vous souhaitez savoir quels clients sont susceptibles de ne pas payer leurs paiements au cours des prochaines années.

Vous utilisez les données sur les clients qui ont et n'ont pas fait défaut pendant de longues périodes en tant que données de construction (ou données de formation) pour générer un modèle de classification. Vous exécutez ensuite ce modèle sur les clients qui vous intéressent. Les algorithmes rechercheront les clients dont les attributs correspondent aux modèles d'attributs des défaillants/non

défaillants précédents, et les classeront en fonction du groupe auquel ils correspondent le plus. Vous pouvez ensuite utiliser ces regroupements comme indicateurs des clients les plus susceptibles de faire défaut.

De même, un modèle de classification peut avoir plus de deux valeurs possibles dans l'attribut cible. Les valeurs peuvent être n'importe quoi, des couleurs de chemise qu'ils sont les plus susceptibles d'acheter, des méthodes promotionnelles auxquelles ils répondront (courrier, e-mail, téléphone) ou s'ils utiliseront ou non un coupon.

- Régression :

La régression est similaire à la classification, sauf que les valeurs de l'attribut ciblé sont numériques plutôt que catégorielles. L'ordre ou l'ampleur de la valeur est significatif d'une certaine manière.

Pour reprendre l'exemple de la carte de crédit, si vous vouliez savoir quel seuil d'endettement les nouveaux clients sont susceptibles d'accumuler sur leur carte de crédit, vous utiliseriez un modèle de régression.

Fournissez simplement les données des clients actuels et passés avec leur niveau d'endettement précédent maximum comme valeur cible, et un modèle de régression sera construit sur ces données de formation. Une fois exécuté sur les nouveaux clients, le modèle de régression fera correspondre les valeurs d'attribut avec les niveaux d'endettement maximaux prévus et attribuera les prédictions à chaque client en conséquence.

Cela pourrait être utilisé pour prédire l'âge des clients avec des données démographiques et d'achat, ou pour prédire la fréquence des réclamations d'assurance.

- Détection D'une Anomalie :

La détection d'anomalies identifie les points de données atypiques d'une distribution donnée. En d'autres termes, il trouve les valeurs aberrantes. Bien que des techniques d'analyse de données plus simples que l'exploration de données à grande échelle puissent identifier les valeurs aberrantes, les techniques de détection d'anomalies d'exploration de données identifient des modèles d'attributs beaucoup plus subtils et les points de données qui ne se conforment pas à ces modèles. La plupart des exemples d'utilisations de détection d'anomalies impliquent la détection de fraude, comme pour les compagnies d'assurance ou de cartes de crédit.

- Classification non supervisé (segmentation):

Data Mining non supervisée ne se concentre pas sur des attributs prédéterminés et ne prédit pas non plus une valeur cible. Au contraire, l'exploration de données non supervisée trouve une structure et une relation cachées entre les données.

- Regroupement :

La technique du Data Mining la plus ouverte, les algorithmes de clustering, trouve et regroupe les points de données présentant des similitudes naturelles.Ceci est utilisé lorsqu'il n'y a pas de groupements naturels évidents, auquel cas les données peuvent être difficiles à explorer. Le regroupement des données peut révéler des groupes et des catégories dont vous n'aviez pas connaissance auparavant. Ces nouveaux groupes peuvent être adaptés à d'autres opérations d'exploration de données à partir desquelles vous découvrirez peut-être de nouvelles corrélations.

- Association:

Fréquemment utilisés pour l'analyse du panier de consommation, les modèles d'association identifient des concurrences communes parmi une liste d'événements possibles. L'analyse du panier de consommation examine tous les articles disponibles sur un support particulier, tels que les produits sur les étagères des magasins ou dans un catalogue, et trouve les produits qui sont couramment vendus ensemble. Cette opération produit des règles d'association. Une telle règle pourrait être une déclaration déclarant que 80 pourcent des personnes qui achètent du charbon de bois, de la viande de hamburger et des petits pains achètent également du fromage en tranches" ou, dans un exemple moins "panier de marché", 90 pourcent des citoyens de Detroit qui s'enracinent pour le Les Tigers, les Lions et les Pistons favorisent également les Red Wings par rapport aux autres équipes de hockey.

De telles règles peuvent être utilisées pour personnaliser l'expérience client afin de promouvoir certains événements ou actions. Cela peut être accompli en organisant les rayons des magasins avec des articles associés à proximité ou en suivant les mouvements des clients sur un site Web en temps réel pour leur présenter des liens de produits pertinents.

- Extraction De Caractéristiques :

L'extraction d'entités crée de nouvelles entités basées sur les attributs de vos données. Ces nouvelles fonctionnalités décrivent une combinaison de modèles de valeur d'attribut significatifs dans vos données. Si la violence, l'héroïsme et les voitures rapides étaient les caractéristiques d'un film, alors le long métrage peut être action , apparenté à un genre ou à un thème. Ce concept peut être utilisé pour extraire les thèmes d'un document en fonction des fréquences de certains mots clés.

La représentation des points de données par leurs caractéristiques peut aider à compresser les données (en échangeant des dizaines d'attributs pour une caractéristique), à faire des prédictions (les données avec cette caractéristique ont souvent ces attributs également) et à reconnaître des modèles. De plus, les fonctionnalités peuvent être utilisées comme de nouveaux attributs, ce qui peut améliorer l'efficacité et la précision des techniques d'apprentissage supervisé (classification, régression, détection d'anomalies, etc.).

Connaître vos objectifs et les techniques appropriées pour les atteindre peut aider vos opérations d'exploration de données à se dérouler de manière fluide et efficace. Différentes données sont appropriées pour différentes informations et comprendre ce que vous demandez à vos analystes de données accélère le processus pour tout le monde

3.12 Données de test classifées :

ce processus aide à prédire les étiquettes de classe des instances de test d'entrée. L'échantillon de test classifié la prédiction de la performance du système est mesurée. La mesure est prise en termes de précision, de temps et d'espace complexité.[65][66]

3.12.1 La matrice de confusion :

est rassemble en lignes les observations et en colonnes les prédictions. voila l'explication dans le tableau:

		Normal	Anormal(attaque)
		La classe détectée	
La classe réelle	Normal	Vrai négative TN(True negative)	Faux positive FP(False positive)
	Anormal(attaque)	Faux négative FN(False negative)	Vrai positive TP(True positive)

Tableau 1 : matrice de confusion

- Un vrai négatif (TN): est une activité normale correctement classée.
- Un faux positif (FP): est une activité normale mal classée .
- Un faux négatif (FN): est une attaque non détectée .
- Un vrai positif (TP): est une attaque correctement détectée.

3.12.2 L'exactitude ou le taux de réussite :

c'est à dir la détection est correcte,il représente le rapport entre les enregistrements bien classés et la totalité d'enregistrements de test.Il est calculé à partir de la formule suivante :

$$Exactitude = \frac{TP + TN}{TP + TN + FP + FN} * 100\%$$

3.12.3 Le rappel :

c'est le taux des intrusions correctement détectées par rapport au nombre total d'intrusions.Il est calculé à partir de la formule suivante :

$$Rappel = \frac{TP}{TP + FN} * 100\%$$

3.12.4 La précision :

est la proportion de prédictions de positifs qui sont en effet des positifs.

$$Prcision = \frac{TP}{TP + FP} * 100\%$$

3.12.5 L'entropie :

est le taux moyen auquel l'information est produite par une source stochastique de données, Ou, c'est une mesure de l'incertitude associée à une variable aléatoire.

$$\text{Entropie} = \textit{précession} * log_2\ \textit{Precession}$$

3.12.6 Taux MC (mal classés) :

est représenté le nombre des classes qui non pas classé correcte par notre modèle.

$$\text{taux MC (mal classés)} = \frac{(FP+FN)}{nombre\ totale\ de\ connexion} * 100$$

3.12.7 Taux BC (bien classés) :

est représenté le nombre des classes qui on classé correcte par notre modèle.

$$\text{taux BC (bien classée)} = \frac{(VP+VN)}{nombre\ totale\ de\ connexion} * 100$$

3.12.8 Fmesure (Moyenne harmonique) :

c'est une métrique qui combine le rappel et la précision en un nombre compris entre 0 et 1. Elle donne une évaluation de synthèse de la classification.

$$F - mesure = \frac{2 * Prcision * Rappel}{Prcision + Rappel}$$

3.12.9 Fitness :

Il existe de nombreux livres traitant de l'essaim de particules et de la définition des fonctions de fitness. Cependant, la fonction de fitness est une fonction qui mappe les valeurs de vos particules à une valeur réelle qui doit récompenser les particules proches de votre critère d'optimisation.

$$\text{fitness}=(\text{accuracy}+\text{fscore})/2$$

3.13 Définition de l'algorithme k-nearest-neighbor (KNN):

Un algorithme k-plus proche voisin(k-nearest-neighbor), souvent abrégé k-nn, est une approche de la classification des données qui estime la probabilité qu'un point de données soit membre d'un groupe ou de l'autre en fonction du groupe dans lequel se trouvent les points de données les plus proches. Le k-nearest-neighbor est un exemple d'algorithme "d'apprentissage paresseux", ce qui signifie qu'il ne construit pas de modèle à l'aide de l'ensemble d'apprentissage tant qu'une requête sur l'ensemble de données n'est pas effectuée.[51]

3.13.1 Sélection des paramètres:

Le meilleur choix de k dépend des données ; généralement, des valeurs plus élevées de k réduisent l'effet du bruit sur la classification,[52] mais rendent les frontières entre les classes moins distinctes. Un bon k peut être sélectionné par diverses techniques heuristiques (voir optimisation des hyperparamètres). Le cas particulier où la classe est prédite comme étant la classe de l'échantillon d'apprentissage le plus proche (c'est-à-dire lorsque k = 1) est appelé l'algorithme du plus proche voisin. La précision de l'algorithme k-NN peut être sévèrement dégradée par la présence de caractéristiques bruyantes ou non pertinentes, ou si les échelles des caractéristiques ne sont pas cohérentes avec leur importance. De nombreux efforts de recherche ont été consacrés à la sélection ou à la mise à l'échelle des caractéristiques afin d'améliorer la classification. Une approche particulièrement populaire[citation nécessaire] est l'utilisation d'algorithmes évolutifs pour optimiser la mise à l'échelle des fonctionnalités.[53] Une autre approche populaire consiste à mettre à l'échelle les caractéristiques par l'information mutuelle des données d'entraînement avec les classes d'entraînement. [citation nécessaire]

Dans les problèmes de classification binaire (deux classes), il est utile de choisir k comme un nombre impair car cela évite les votes à égalité. Une façon populaire de choisir le k empiriquement optimal dans ce contexte consiste à utiliser la méthode bootstrap.[54]

3.13.2 Le principe de l'algorithme des K plus proches voisins (KNN)[55]:

L'intuition derrière l'algorithme des K plus proches voisins est l'une des plus simples de tous les algorithmes de Machine Learning supervisé :

Étape 1 : Sélectionnez le nombre K de voisins.

Étape 2 : Calculez la distance

$$\sum_{i=1}^{n} |x_i - y_i| \qquad \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Euclidienne *Manhattan*

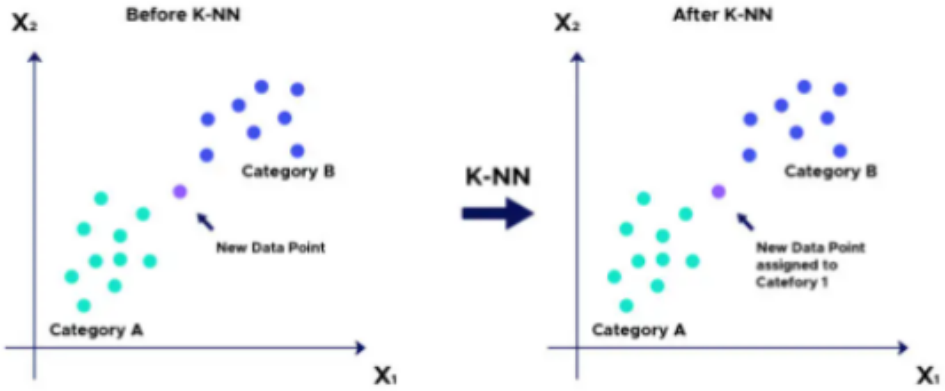
Du point non classifié aux autres points.

Étape 3 : Prenez les K voisins les plus proches selon la distance calculée.

Étape 4 : Parmi ces K voisins, comptez le nombre de points appartenant à chaque catégorie.

Étape 5 : Attribuez le nouveau point à la catégorie la plus présente parmis ces K voisins.

Étape 6 : Notre modèle est prêt :



3.13.3 Les Avantages [55]:

- L'algorithme est simple et facile à mettre en œuvre.

- Il n'est pas nécessaire de créer un modèle, de régler plusieurs paramètres ou de formuler des hypothèses supplémentaires.

- L'algorithme est polyvalent. Il peut être utilisé pour la classification ou la régression.

3.13.4 Les Inconvénients [55] :

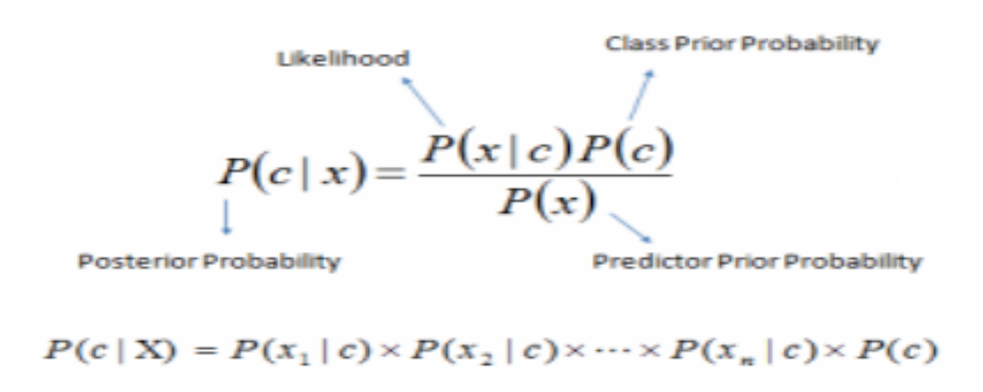
L'algorithme devient beaucoup plus lent à mesure que le nombre d'observation et de variables indépendantes augmente. Étant l'un des algorithmes les plus simples de Machine Learning, il est hautement implémenté pour développer des systèmes basés sur l'apprentissage, intuitifs et intelligents qui pourraient effectuer et prendre de petites décisions tout seuls. Cela rend les choses encore plus pratiques pour l'apprentissage et le développement et aide presque tous les types d'industries qui pourraient utiliser des systèmes, des solutions ou des services intelligents.

3.14 Définition du algorithme Naive Bayes:

Il s'agit d'une technique de classification basée sur le théorème de Bayes avec une hypothèse d'indépendance entre les prédicteurs. En termes simples, un classificateur Naive Bayes suppose que la présence d'une caractéristique particulière dans une classe n'est pas liée à la présence de toute autre caractéristique. Par exemple, un fruit peut être considéré comme une pomme s'il est rouge, rond et d'environ 3 pouces de diamètre. Même si ces caractéristiques dépendent les unes des autres ou de l'existence d'autres caractéristiques, toutes ces propriétés contribuent indépendamment à la probabilité que ce fruit soit une pomme et c'est pourquoi on l'appelle "Naif".

Le modèle Naive Bayes est facile à construire et particulièrement utile pour les très grands ensembles de données. En plus de sa simplicité, Naive Bayes est connu pour surpasser même les méthodes de classification les plus sophistiquées.Il s'agit d'une technique de classification basée sur le théorème de Bayes avec une hypothèse d'indépendance entre les prédicteurs. En termes simples, un classificateur Naive Bayes suppose que la présence d'une caractéristique particulière dans une classe n'est pas liée à la présence de toute autre caractéristique.

Le théorème de Bayes fournit un moyen de calculer la probabilité a posteriori P(c—x) à partir de P(c), P(x) et P(x—c). Regardez l'équation ci-dessous[61] :



- P(c—x) est la probabilité a posteriori de la classe (c, cible) compte tenu du prédicteur (x, attributs).

- P(c) est la probabilité a priori de la classe.

- P(x—c) est la vraisemblance qui est la probabilité du prédicteur pour une classe donnée.

- P(x) est la probabilité a priori du prédicteur.

3.14.1 Le principe de l'algorithme l'algorithme Naive Bayes [61] :

Comprenons le à l'aide d'un exemple. Ci-dessous, j'ai un ensemble de données d'entraînement sur la météo et la variable cible correspondante "Jouer" (suggérant des possibilités de jouer). Maintenant, nous devons classer si les joueurs joueront ou non en fonction des conditions météorologiques. Suivons les étapes ci-dessous pour l'exécuter.

Étape 1 : Convertir l'ensemble de données en un tableau de fréquence.

Étape 2 : Créez une table de probabilité en trouvant les probabilités comme la probabilité de couverture = 0,29 et la probabilité de jouer est de 0,64.

Étape 3 : Maintenant, utilisez l'équation bayésienne naïve pour calculer la probabilité a posteriori pour chaque classe. La classe avec la probabilité a posteriori la plus élevée est le résultat de la prédiction.

3.14.2 les avantages et les inconvénients de l'algorithme Naive Bayes [61]:

Avantages:

- Il est facile et rapide de prédire la classe de l'ensemble de données de test. Il fonctionne également bien dans la prédiction multi-classes - Lorsque l'hypothèse d'indépendance est valable, un classificateur Naive Bayes est plus performant que d'autres modèles comme la régression logistique et vous avez besoin de moins de données d'entraînement. - Il fonctionne bien dans le cas de variables d'entrée catégorielles par rapport aux variables numériques. Pour la variable numérique, une distribution normale est supposée (courbe en cloche, qui est une hypothèse forte).

Les inconvénients:

- Si la variable catégorielle a une catégorie (dans l'ensemble de données de test), qui n'a pas été observée dans l'ensemble de données d'apprentissage, le modèle attribuera une probabilité de 0 (zéro) et ne pourra pas faire de prédiction. Ceci est souvent connu sous le nom de fréquence zéro . Pour résoudre ce problème, nous pouvons utiliser la technique de lissage. L'estimation de Laplace est l'une des techniques de lissage les plus simples.

- Une autre limitation de Naive Bayes est l'hypothèse de prédicteurs indépendants. Dans la vraie vie, il est presque impossible d'obtenir un ensemble de prédicteurs complètement indépendants.

3.14.3 4 applications des algorithmes naïfs de Bayes :

- Prédiction en temps réel :

Naive Bayes est un classificateur avide d'apprentissage et il est certainement rapide. Ainsi, il pourrait être utilisé pour faire des prédictions en temps réel.

- Prédiction multi-classes :

cet algorithme est également bien connu pour la fonctionnalité de prédiction multi-classes. Ici, nous pouvons prédire la probabilité de plusieurs classes de variable cible.

- Classification de texte/filtrage de spam/analyse des sentiments :

les classificateurs Naive Bayes principalement utilisés dans la classification de texte (en raison de meilleurs résultats dans les problèmes multi-classes et de la règle d'indépendance) ont un taux de réussite plus élevé par rapport aux autres algorithmes. En conséquence, il est largement utilisé dans le filtrage anti-spam (identifier les spams) et l'analyse des sentiments (dans l'analyse des médias sociaux, pour identifier les sentiments positifs et négatifs des clients).

- Système de recommandation :

le classificateur Naive Bayes et le filtrage collaboratif créent ensemble un système de recommandation qui utilise des techniques d'apprentissage automatique et d'exploration de données pour filtrer les informations invisibles et prédire si un utilisateur aimerait ou non une ressource donnée.

3.15 Définition de l'algorithme d'optimisation par essaim de particules (PSO):

En informatique, l'optimisation par essaim de particules (PSO) est une méthode informatique qui optimise un problème en essayant de manière itérative d'améliorer une solution candidate par rapport à une mesure de qualité donnée. Il résout un problème en ayant une population de solutions candidates, ici appelées particules, et en déplaçant ces particules dans l'espace de recherche selon une formule mathématique simple sur la position et la vitesse de la particule. Le mouvement de chaque particule est influencé par sa meilleure position connue locale, mais est également guidé vers les meilleures positions connues dans l'espace de recherche, qui sont mises à jour au fur et à mesure que de meilleures positions sont trouvées par d'autres particules. Cela devrait déplacer l'essaim vers les meilleures solutions [56].

Dans PSO le comportement social est modélisé par une équation mathématique permettant de guider les particules durant leur processus de déplacement[57]. Le déplacement d'une particule est influencé par trois composantes : la composante d'inertie, la composante cognitive et la composante sociale. Chacune de ces composantes reflète une partie de l'équation voir figure13. [58]

1)La composante d'inertie : la particule tend à suivre sa direction courante de déplacement .

2) La composante cognitive : la particule tend à se diriger vers le meilleur site par lequel elle est déjà passée .

3) La composante sociale : la particule tend à se diriger vers le meilleur site atteint par ses voisines.

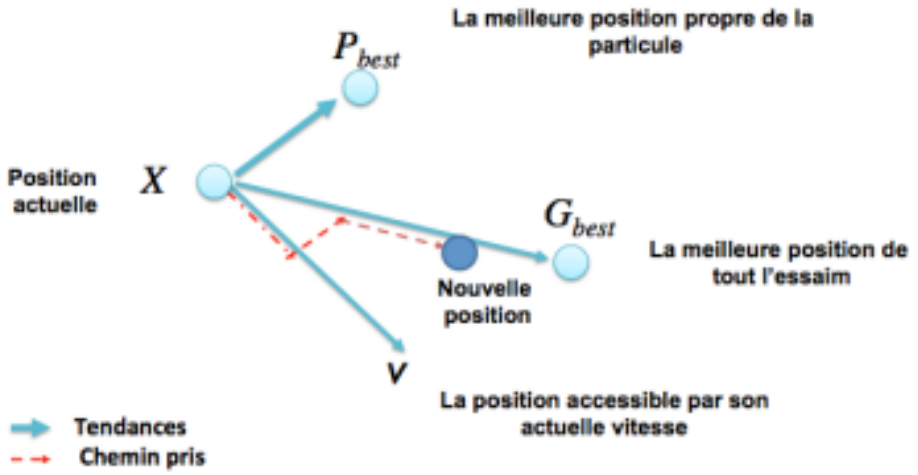


Figure13 : Déplacement d'une particule

3.15.1 Formalisation :

Une particule i de l'essaim dans un espace de dimension D est caractérisée, à l'instant t, par :

– X : sa position dans l'espace de recherche ;
– V : sa vitesse ;
– Pb : la position de la meilleure solution par laquelle elle est passée ;
– Pg : la position de la meilleure solution connue de tout l'essaim ;
– f (Pb) : la valeur de fitness de sa meilleure solution ;
– f (Pg) : la valeur de fitness de la meilleure solution connue de tout l'essaim. Le déplacement de la particule i entre les itérations t et t+1 se fait selon les deux équations 12. [57]

$$V(t+1)=v(t)+C1r1(Pb(t)-X(t))+C2r2(Pg(t)-X(t)) \quad 1$$

$$X(t+1)=X(t)+V(t+1) \quad 2$$

– C1 et C2: deux constantes qui représentent les coefficients d'accélération, elles peuvent être non constantes dans certains cas selon le problème d'optimisation posé .
– r1 et r2 : deux nombres aléatoires tirés de l'intervalle [0,1].

3.15.2 Les étapes de l'algorithme:

Les principales étapes de l'algorithme d'optimisation par essaim de particules (PSO) sont les suivantes :

Étape 1 : Attribuer les particules pour l'ensemble de l'espace de recherche en générant leurs positions, vitesses et topologie de communication.

Étape 2 : Créer les voisinages via la valeur initiale du rayon. Étape 3 : Diviser le traitement de l'algorithme PSO sur un ensemble de traitements : pour chaque traitement on attribue un thread.

Étape 4 : Attribuer les lots de particules aux threads.

Étape 5 : Lancer les traitements de tous les threads en parallèle pour une itération.

Étape 6 : Mettre à jour les voisinages selon les nouvelles positions des particules et la nouvelle valeur du rayon s'il y a lieu.

Étape 7 : Si le critère d'arrêt est satisfait, arrêter, sinon passer à l'étape 5.

Étape 8 : Le résultat est la meilleure solution obtenue parmi les threads.

3.16 Les notions du voisinage :

Le voisinage constitue la structure du réseau social. Le voisinage d'une particule représente avec qui chacune des particules va pouvoir communiquer. Il existe deux principaux types de voisinages :

3.16.1 Le voisinage géographique :

ce type de voisinage représente la proximité géographique, c'est la notion la plus naturelle du voisinage pour les essaims particulaires, les voisins sont considérés comme les particules les plus proches. Cependant, à chaque itération, les nouveaux voisins doivent être recalculés à partir d'une distance prédéfinie dans l'espace de recherche. C'est donc un voisinage dynamique qu'il convient de définir et d'actualiser à chaque itération. C'est ce type de voisinage qui a été retenu dans notre approche.

3.16.2 Le voisinage social :

ce type de voisinage représente la proximité sociale, les voisinages ne sont plus l'expression de la distance mais l'expression de l'échange d'informations, les voisins sont définis à l'initialisation et ne sont pas modifiés par la suite. Une fois le réseau des connexions sociales établi, il n'y a pas besoin de le réactualiser. C'est donc un voisinage statique. La modification de la formule

de vitesse (1) est réalisée en utilisant un nouveau terme dans l'équation. Il a été introduit par [59], son illustration paraît dans la figure 14 [60]:

$$V(t+1)=v(t)+C1r1(Pb(t)-X(t))+C2r2(Pg(t)-X(t))\ C3r3(Pn(t)-X(t))$$

où : Pn : la meilleure position du voisinage ;
C3 : le coefficient d'accélération, appelé aussi paramètre social ;
r3 : nombre aléatoire tiré de l'intervalle [0,1].

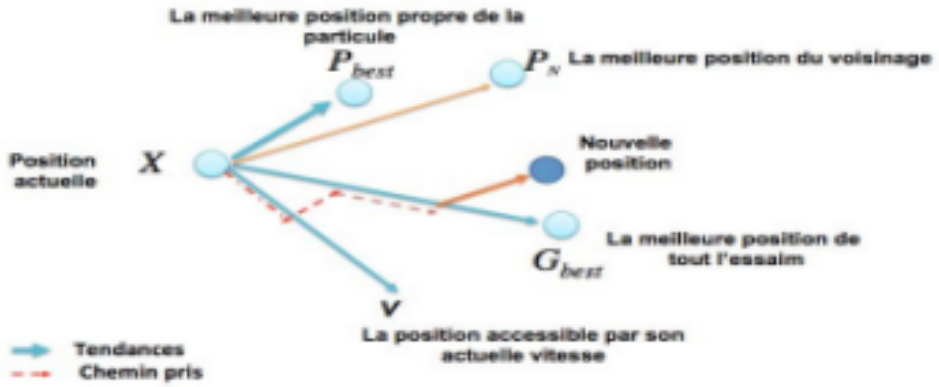


Figure14 : Déplacement d'une particule

3.17 Conclusion:

Les métaheuristiques constituent une classe de méthodes approchées adaptables à un grand nombre de problèmes d'optimisation combinatoire. Mais, si l'on a pu constater leur grande efficacité sur de nombreuses classes de problèmes, il existe en revanche très peu de résultats permettant de comprendre la raison de cette efficacité, et aucune méthode particulière ne peut garantir qu'une métaheuristique sera plus efficace qu'une autre sur n'importe quel problème. Concrètement, certaines métaheuristiques présentent l'avantage d'être simples à mettre en œuvre. D'autres sont plutôt bien adaptées à la résolution de certaines classes de problème, très contraints.

Chapitre 4

Implémentation et analyse des résultats

4.1 Introduction:

Dans ce chapitre nous allons détailler la méthodologie et le déroulement de notre programme, les moyens utilisés comme le data set nsl-kdd, la bibliothèque weka et Le logiciel Scene Builder. Nous avons programmé avec le langage de programmation java fx.

4.2 Le Benchmark utilisé:

4.2.1 NSL-KDD :

NSL-KDD est un ensemble de données proposé pour résoudre certains problèmes inhérents à l'ensemble de données KDD'99 qui sont mentionnés dans [62]. Bien que cette nouvelle version de l'ensemble de données KDD souffre encore de certains problèmes discutés par McHugh et ne soit peut-être pas un parfait représentant des réseaux réels existants, en raison du manque d'ensembles de données publics pour les IDS basés sur le réseau, nous pensons qu'il peut être appliqué comme un ensemble de données de référence efficace pour aider les chercheurs à comparer les différentes méthodes de détection d'intrusion.

De plus, le nombre d'enregistrements dans les ensembles d'apprentissage et de test NSL-KDD est raisonnable. Cet avantage rend abordable l'exécution des expériences sur l'ensemble complet sans qu'il soit nécessaire de sélectionner au hasard une petite partie. En conséquence, les résultats d'évaluation des différents travaux de recherche seront cohérents et comparables.

Les détails des attributs sont répertoriés dans le tableau suivant :

N°	Nom de l'attribut	Description
01	duration	Durée de la connexion(nb de secondes)
02	protocol_type	Type du protocol : TCP, UDP ou ICMP
03	service	Service du réseau sollicité sur l'hôte de destination(ftp, http,...etc)
04	flag	Statut de la connexion(REJ, RSTO,...etc)
05	src_ bytes	Nbr d'octets envoyés de la source vers la destination
06	dst_ bytes	Nbr d'octets envoyés de la destination vers la source
07	land	Vaut 1 si la connexion est de/vers le même hôte/port, 0 sinon
08	wrong_ fragment	Nbr de fragments erronés
09	urgent	Nbr de paquets urgents
10	hot	Nbr d'indicateurs hot
11	num_ failed_ logins	Nbr de logins échoués
12	logged_ in	Vaut 1 si login réussi, 0 sinon
13	num_ compromised	Nbr de cas de compromission (compromised condition)
14	root_ shell	Vaut 1 si un shell root est obtenu, 0 sinon
15	su_ attempted	Vaut 1 si une commande super utilisateur est tentée, 0 sinon
16	num_ root	Nbr d'accès en mode root
17	num_ file_ creations	Nbr d'opérations en création de fichiers
18	num_ shells	Nbr de shells lancés
19	num_ access_ files	Nbr d'opérations d'accès aux fichiers de contrôle
20	num_ outbound_ cmds	Nbr de commandes non autorisées dans les sessions ftp
21	is_ host_ login	Vaut 1 si le login fait partie de la list hot, 0 sinon

22	is_ guest_ login	Vaut 1 si le login fait partie de la list guest, 0 sinon
23	count	Nbr de connexions pour le même hôte
24	srv_ count	Nbr de connexions pour le même service
25	serror_ rate	% de connexions pour le même hôte ayant l'erreur SYN
26	srv_ serror_ rate	% de connexions pour le même service ayant l'erreur SYN
27	rerror_ rate	% de connexions pour le même hôte ayant l'erreur REJ
28	srv_ rerror_ rate	% de connexions pour le même service ayant l'erreur REJ
29	same_ srv _rate	% de connexions pour le même hôte utilisant le même service
30	diff_ srv_ rate	% de connexions pour le même hôte utilisant différents services
31	srv_ diff_ host_ rate	% de connexions pour le même service utilisant différents hôtes
32	dst_ host_ count	Nbr de connexions pour le même hôte
33	dst_ host_ srv_ count	Nbr de connexions pour le même hôte utilisant le même service
34	dst_ host_ same_ srv_ rate	% de connexions pour le même hôte utilisant le même service
35	dst_ host_ diff_ srv_ rate	% de connexions pour le même hôte utilisant différents services
36	dst_host_same_src_port_rate	% de connexions pour le même hôte ayant le port src
37	dst_ host_ srv_ diff_ host_ rate	% de connexions pour le même hôte en provenance de différents hotes
38	dst_ host_ serror_ rate	% de connexions pour le même hôte ayant l'erreur SYN
39	dst_host_srv_serror_rate	% de connexions pour le même hôte et service ayant l'erreur SYN
40	dst_ host_ rerror_ rate	% de connexions pour le même hôte ayant l'erreur REJ

Tableau 2: Liste des attributs de la base NSL KDD

4.2.2 La bibliothèque Weka:

Weka est une collection d'algorithmes d'apprentissage automatique pour les tâches d'exploration de données. Les algorithmes peuvent être appliqués directement à un ensemble de données ou appelés à partir de votre propre code Java. Weka contient des outils pour le prétraitement des données, la classification, la régression, le cluster, les règles d'association et la visualisation. L'entrée dans Weka doit être formatée selon le format de fichier relationnel d'attribut et avec le nom de fichier portant l'extension arff. Toutes les techniques de Weka reposent sur l'hypothèse que les données sont disponibles sous la forme d'un fichier plat ou d'une relation, où chaque point de données est décrit par un nombre fixe d'attributs (normalement, des attributs numériques ou nominaux, mais certains autres types d'attributs sont également pris en charge) . Weka fournit un accès aux bases de données SQL à l'aide de Java Database Connective et peut traiter le résultat renvoyé par une requête de base de données.

Weka donne accès à l'apprentissage en profondeur avec Deeplearning4j.[63] Il n'est pas capable d'explorer des données multi-relationnelles, mais il existe un logiciel séparé pour convertir une collection de tables de base de données liées en une seule table qui convient au traitement à l'aide de Weka.[64] Un autre domaine important qui n'est actuellement pas couvert par les algorithmes inclus dans la distribution Weka est la modélisation de séquences.

4.2.3 Définition de language java :

Java est un langage de programmation orienté objet, basé sur des classes et à usage général, conçu pour avoir moins de dépendances d'implémentation. C'est une plate-forme informatique pour le développement d'applications. Java est donc rapide, sécurisé et fiable. Il est largement utilisé pour développer des applications Java dans les ordinateurs portables, les centres de données, les consoles de jeux, les superordinateurs scientifiques, les téléphones portables, etc.

4.2.4 JavaFX :

JavaFX est une bibliothèque Java composée de classes et d'interfaces écrites en code Java natif. Les API sont conçues pour être une alternative conviviale aux langages Java Virtual Machine (Java VM), tels que JRuby et Scala.

4.2.5 Scene Builder:

Scene Builder est un outil interactif de conception d'interface graphique pour JavaFX. Créé par Oracle, il permet de construire rapidement des interfaces utilisateurs sans avoir besoin de les coder. Le logiciel est décliné en deux versions : l'une (8.x) destiné à JavaFX 8 et l'autre (9.0 et +) pour JavaFX 9 et plus.

4.3 Méthodologie :

Le système de détection d'intrusion basé sur l'apprentissage automatique propose est démontré dans la Fig.15. Les composants et leurs aspects fonctionnels sont expliqués dans cette section.

4.3.1 Insérer le Data set :

l'étape initiale du système consiste à produire un ensemble de données pour apprendre dans le système. Ici, le jeu de données KDD CUP 99 qui est utilisé. Il s'agit d'un énorme ensemble de données qui contient un total de 42 attributs dont un attribut représentant l'étiquette de classe comme exemple pour l'apprentissage.

il contienent une quantité importante d'informations sur les schémas d'attaque. Mais c'est une question qui concerne également la consommation des ressources. De plus, ces données peuvent contenir divers contenus bruyants, et attributs moins informatifs qu'il n'est pas nécessaire d'utiliser pour identifier un modèle d'attaque. Par conséquent, le processus suivant est utilisé pour optimiser la qualité des données et la réduction de dimension pour atteindre la précision et l'efficacité en termes de temps et consommation de mémoire.

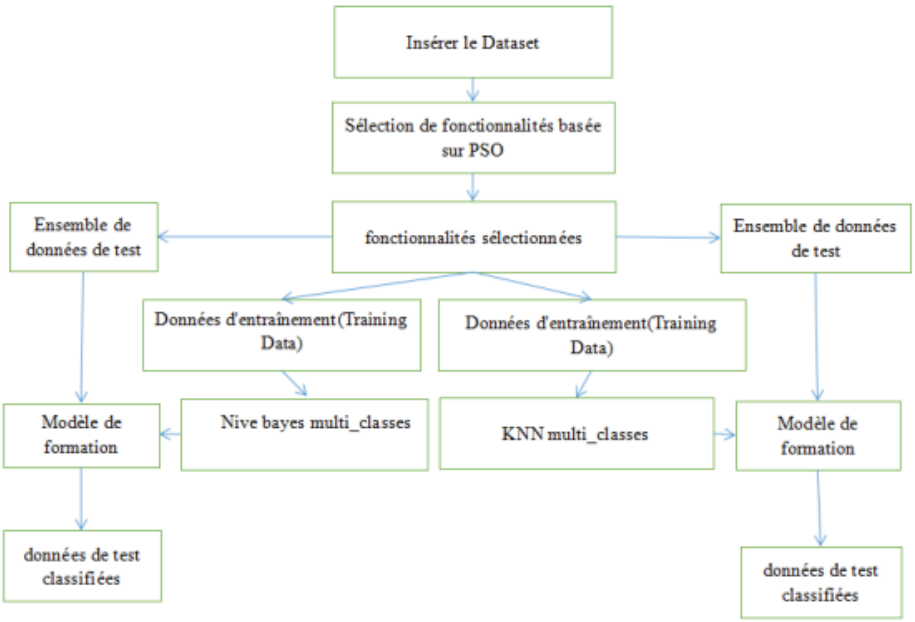


Figure15:Méthodologie du système de détection d'intrusion basé sur l'apprentissage automatique

4.3.2 Sélection sur fonctionnalités basée sur PSO :

Ce processus peut réduire un certain nombre de données , la taille de data set peut donc être modifiée. Ces données raffinées sont utilisé en outre avec l'algorithme PSO pour sélectionner les attributs essentiels.

4.3.3 fonctionnalités sélectionnées :

L'algorithme PSO est une technique d'optimisation. Elle est utilisée ici avec la base d'attributs pour comparer les valeurs cibles avec l'attribut individuel et classé tous les attributs en conséquence pour sélectionner les éléments essentiels parmi tous les attributs. Les 21 fonctionnalités les mieux classées sont sélectionnés ici pour l'expérimentation.

4.3.4 Données d'entraînement(Training Data) :

Les attributs sélectionnés vont être utilisées pour l'apprentissage et le test de l'algorithme d'apprentissage supervisé. Par conséquent, ces attributs sont subdivisées en deux sous-ensembles à savoir l'apprentissage et le test. Les 66% de la sélection fonctionnalités sont sélectionnées au hasard pour être utilisées pour l'apprentissage.

4.3.5 Ensemble de données de test(Test Data) :

Afin de tester le modèle proposé, les 34% restants des échantillons de données sont conservés séparément avec leurs étiquettes de classe utilisées pour le test. Pendant la validation, ces étiquettes de classe sont utilisées par l'algorithme de mesure des performances.

4.3.6 KNN multi-classes et le nive bayes multi-classes :

K plus proches voisins (KNN) et le nive bayes sont des classificateurs d'apprentissage supervisé. qui sert à classer les binaire, mais en utilisant diverses extensions, nous pouvons l'utiliser comme des classificateurs multiclasse. L'ensemble de données d'apprentissage est utilisé avec le KNN et nive bayes multiclasse. Le KNN fonctionne ici comme un concept un contre tous pour l'apprentissage. Le Knn multi-classe formé accepte en outre le jeu de données de test et reconnaît les modèles du jeu de données de test , Il en même pour algorithme nive bayes.

4.3.7 Modèle de formation :

Lors de la formation du classificateurs KNN et naive bayes, différents facteurs sont calculeés et qui peuvent justifier critères de classement courant. KNN et naive bayes sont les plus populaires en terme de techniques de classification des données. Il peuvent être utilisé pour classer un ensemble de données linéaires et non linéaires.

L'objectif de KNN et naive bayes est de produire un modèle qui prédit la valeur cible des données qui se produisent dans le test dans lequel seuls les attributs ont été donnés. Le but de la classification dans KNN et naive bayes est de séparer les deux classes par une fonction préparée à partir d'un ensemble de données et produisant un classificateur qui prédira en outre les classes pour les données invisibles .

4.3.8 Les fonctionnalité de notre algorithme PSO en détaille dans cette Méthodologie:

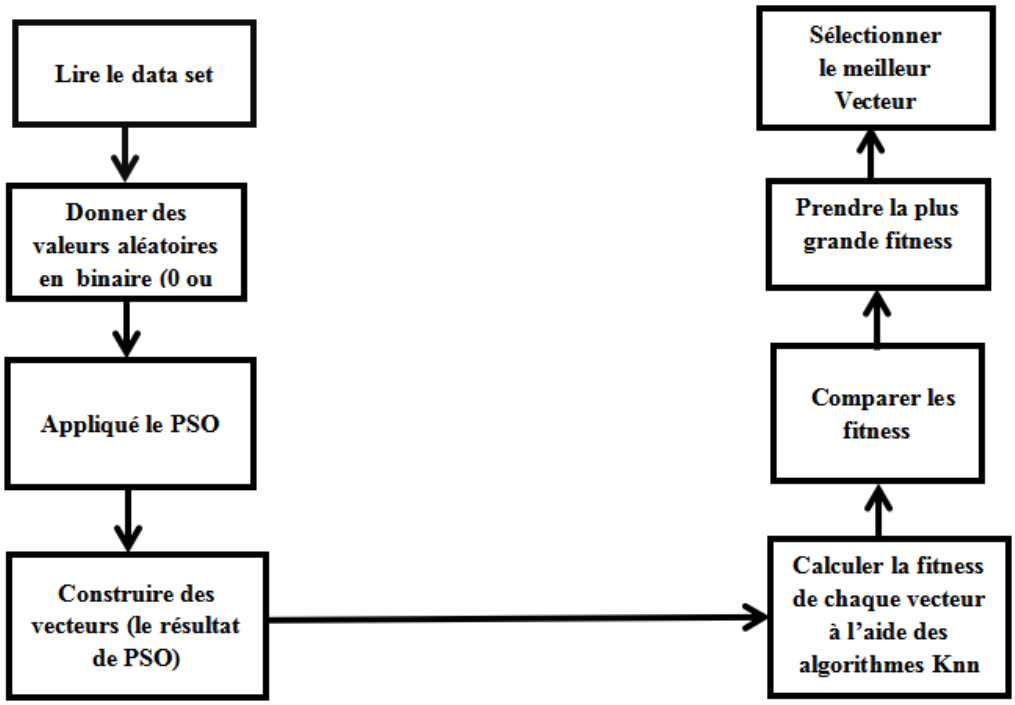


Figure16: Les fonctionnalité de notre algorithme PSO pour la détection d'intrusion

1- Lire le data:

L'algorithme PSO lit les données de data set .

2- Donner des valeurs aléatoires en binaire (0 ou 1) aux attributs:

L'algorithme PSO donne des valeurs aléatoire en binaire aux attributs de notre data set, à l'aide de la fonction random. donner le chiffre 1 à un attribut signifie qu'il a été choisi par l'algorithme PSO , le chiffre 0 signifie qu'il n'a pas été choisi par l'algorithme PSO.

Nous avons fait cette fonction afin de réduire la taille des donnée.

3- Appliqué le PSO:

dans cette étape on donne des nombre entiers à les paramètres Nbr essaim et Nb itération à chaque fois. Et pour les autres paramètres de notre algorithme PSO qui sont :

C1: est le facteur de contrôle d'accélération pour le Gbest (La meilleure position parmi tout les positions des particules)

C2 : facteur de contrôle d'accélération pour le Pbest (La meilleure position du particule)

C'est deux paramètres restent constants on donne des valeurs fixes :

C1 = 0 et C1 = 4

R1,R2 : Deux nombre aléatoires, générés indépendamment dans l'intervalle [0,1].

C'est quatre paramètres sont utilisés dans le calcul de la vitesse des particules.

4- Construire un vecteur (le résultat de PSO):

l'algorithme PSO sélectionne les attributs de manière aléatoire, c'est à dire qu'il prend un certain nombre d'attributs de chaque ligne de data set. Après cela, l'algorithme PSO crée un vecteur contenant 41 attributs.

5- Comparer les fitness:

aprés avoir calculé les fitness,l'algorithme pso compare les fitness.

6- Prendre la plus grande fitness:

Aprés la comparaison, l'algorithme Pso choisi le meilleur fitness, c'est à dire le fitness qui a la plus grande valeur.

7- Sélectionnée le meilleur vecteur:

l'algorithme Pso sélectionne le meilleur vecteur. Qui a les meilleures attributs..

4.3.9 Les fonctionnalité des algorithmes de classification le Knn et le nive bayes en détailledans cette Méthodologie:

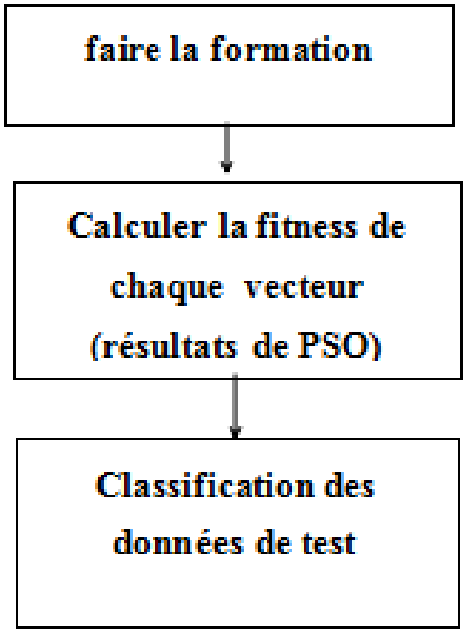


Figure17: Les fonctionnalité des algorithmes de classification(Knn et nive bayes)

1- faire la formation:

lors de la formation, le Knn et le nive bayes calculent les différents facteurs qui peuvent justifier les critères de classement.

2- Calculer la fitness de chaque vecteur (résultats de PSO):

L'algorithme PSO donne un ensemble de vecteur à l'algorithme KNN pour l'apprentissage et cet algorithme calcule la fitness pour chaque vecteur. La même chose pour l'algorithme nive bayes.

3- Faire le test à laid de milleure vecteur :

Cela signifie que les algorithmes Knn et nive bayes prennent le meilleur vecteur choisi pour effectuer le test. ils sont prédire les étiquettes de classe

des instances de test d'entrée.

4.3.10 Les objective de cette méthodologie:

1-Sélectionné les meilleurs caractéristique de data set par le pso.
2-Le Choix des meilleurs attributs est fait abase de la fonction fitness.
3-Assurer le meilleur résultat possible.
4-Gagner du temps.

4.4 L'environnement de programmation:

Nous avons choisi l'environnement de programmation NetBeans IDE 8.2 pour Windows, Le choix du langage java a été guidé par les avantages offerts par la programmation orientée objet d'une façon générale.

4.4.1 L'utislisation de la La bibliothèque Weka :

on utilisée La bibliothèque WEKA, pour la mise en œuvre des algorithmes de classification, qui sont:
1-L'algorithme Knn (k-nearest-neighbor)
2-L'algorithme nive bayes

4.4.2 Les caractéristi0ques techniques de la machine:

Les caractéristiques techniques de la machine sont représentées dans le tableau suivant :

Composent	valeurs
Processeur	Intel(R) CPUN3060GHz
Vitesse	1.60 GHz
Mémoire	4.00Go
Système d'exploitation	Windows 8.1 64 bits

Tableau 3 : Les caractéristiques techniques de la machine

4.5 Analyse des résultats:

Lors de l'exécution de notre programme on obtient la figure suivante:

Figure18: l'interface de notre programme

- la sélection de Data set:

Puis nous passons à la sélection du Data set pour choisir notre base de donnée et la connecter à notre application.

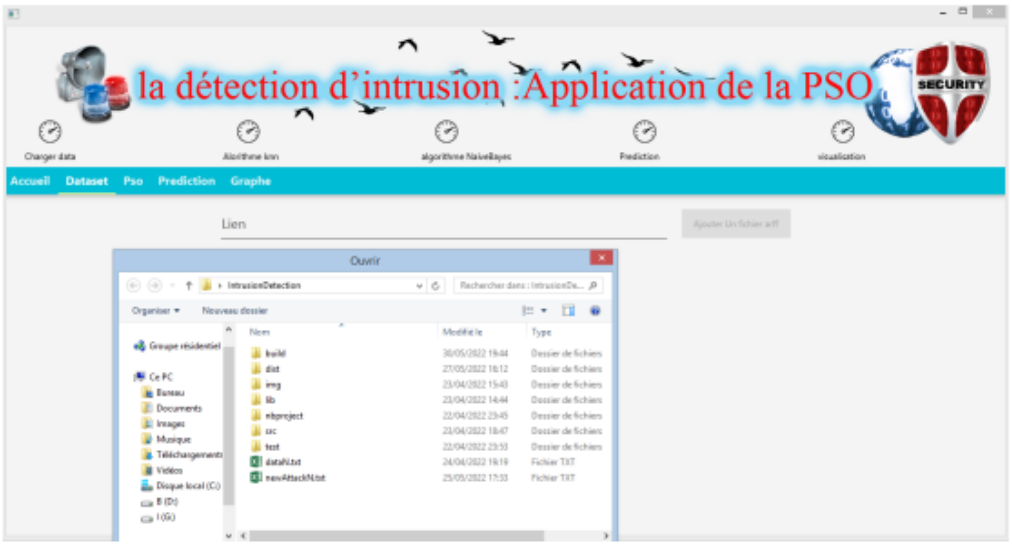


Figure19: le choix du Data set

4.5.1 Les paramètres du test:

Afin d'obtenir la meilleure exactitude (taux de réussite) possible, nous avons effectué plusieurs tests en changeant les paramètres suivants à chaque fois :

- nombre Essaim.
- nombre iteration.
- nombre K (le nombre des distances les plus proches de l' algorithme KNN).
- le choix des distances utilisées.
- la taille des donnée (200 lignes, 300lignes.....).

Les valeurs des autres paramètres communs dans les différentes expérimentations sont indiquées dans le tableau cidessous :

Paramètres de test	
Paramètres	**Valeur**
Nbre Essaim	40
Nbre étiration	10
Nbre k	10

Tableau 4 : Paramètres du test

On passe à la sélection PSO afin de saisir les paramètres dans notre application:

Figure 20 : les paramètres de test

4.5.2 Résultats obtenus :

- Les algorithmes classifiés (KNN et Nive Bayes) calculent la fitness pour chaque vecteur choisi par l'algorithme PSO, puis comparent les pourcentages de fitness et choisi le plus grand.

- L'algorithme PSO prend le vecteur qui à la plus grande fitness.

1- le tableau :

Le premier tableau représente les meilleures fitness qui sont calculées par les algorithmes de classification (KNN, naive bayes) :

La taille des données	La meilleure fitness pour l'algorithme KNN(en utilisant la distance euclidien) %	Le meilleur fitness pour l'algorithme KNN (en utilisant la distance manhattan) %	Le meilleur fitness pour naive bayes %
200	92.26	92.88	92.88
300	98.04	96.06	97.04
400	97.08	97.81	94.84
500	97.64	97.64	95.29
1000	97.94	97.90	92.06

Tableau 5: Le choix des meilleures fitness

On note que la plus grande fitness pour le vecteur est 98,04% à 300 lignes pour l'algorithme knn (la distance euclidien).

2- Le graphe de comparaison:

La Figure 22 montre la Comparaison de tempe d'éxecusion entre KNN et Le naive bayes pour obtient le meilleur finesse.

Dans ce cas les valeurs choisi pour les paramètre de l'algorithme Pso sont:

- Nbre iterations = 10
- Nbre Essaim= 40

Les valeurs choisi pour les paramètre de l'algorithme KNN sont:

- Nbr K=10
- La distance choisie est distance euclidienne
- La taille des données de data set choisi est : 1000 lignes

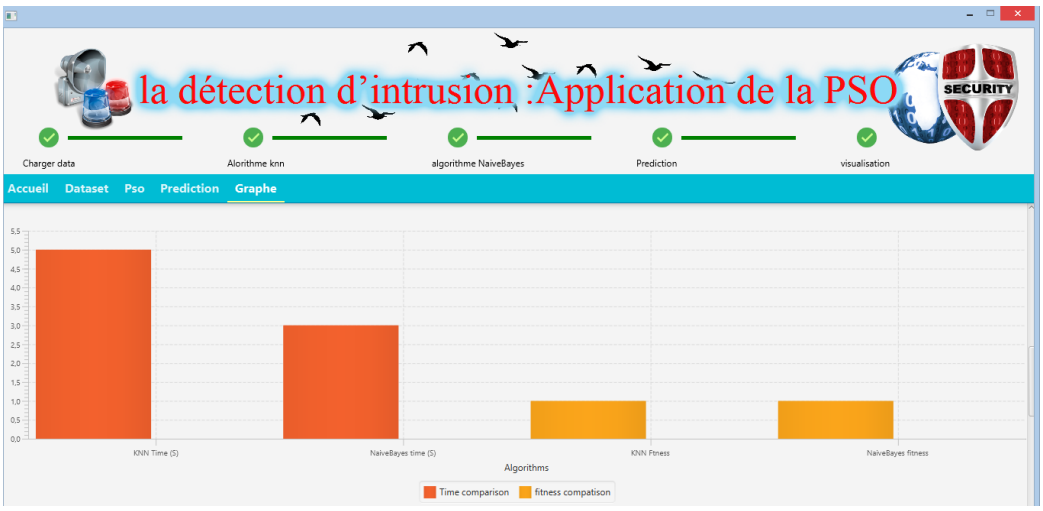


Figure 21:Comparaison de tempe d'éxecusion entre KNN et Le naive bayes pour obtient le meilleur finesse

- Les colonnes rouge représentent le tempe de comparaison. l'unité de temps est la seconde.

- Les colonnes orange représentent la meilleure valeur de fitness est choisie par chaque algorithme. Les valeurs de fitness dans la plage [0,1].

- Remarque:

1-L'algorithme de Knn a pris plus de temps pour comparer les fitness par rapport l'algorithme nive bayes.

2-On remarque que les valeurs de fitness sont proches, entre les deux algorithme Knn et le nive bayes.

4.5.3 Les résultats des notre algorithme PSO:

Dans ce cas, on donne des valeurs pour les paramètre de l'algorithme PSO:

- Nbre iterations = 10
- Nbre Essaim= 40

On plus les paramètre de l'algorithme KNN :

- Nbr K=10
- La distance choisie est distance euclidienne
- La taille des données de notre Data set est 500 lignes

1- Les résultats des notre algorithme PSO pour l'algorithme Knn:

Dans la figure 22 , nous pouvons voir les meilleures vecteur sélectionné par le pso, pour l'algorithme Knn :

Figure 22:Les résultats des notre algorithme PSO pour l'algorithme Knn

- Les couleurs rouges représentent les attributs qui ne sont pas choisis par le PSO pour l'algorithme Knn.

- Les couleurs vertes représentent les meilleures attributs choisis par le PSO pour l'algorithme Knn dans ce cas.

2- Les résultats des notre algorithme PSO pour l'algorithme nive bayes:

la figure 23 , représente le meilleure vecteur sélectionné par le pso, pour l'algorithme nive bayes :

Figure 23:Les résultats des notre algorithme PSO pour l'algorithme nive byes

- Les couleurs rouges représentent les attributs qui ne sont pas choisis par le PSO pour l'algorithme Naive Bayes.

- Les couleurs vertes représentent les meilleures attributs choisies par le PSO pour l'algorithme Naive Bayes dans ce cas.

4.5.4 Premier cas pour l'apprentissage : L'algorithme KNN

Dans ce cas, on donne des valeurs pour les paramètre de l'algorithme PSO:
- Nbre iterations = 10
- Nbre Essaim= 40

On plus les paramètre de l'algorithme KNN :
- Nbr K=10
- La distance choisie est distance euclidienne

1- le tableau :

le tableau suivant représente les Mesures de performance de l'algorithme KNN avec la distance euclidienne :

La taille des données	Exactitude %	Rappel %	Pricision %	Entropie %	F_score %	Fitness %
200	92.64	92.64	92.68	10.15	92.66	92.26
300	98.03	97.82	98.27	2.46	98.05	98.04
400	97.05	96.87	97.36	3.74	97.12	97.08
500	97.64	97.56	97.72	2.36	97.64	97.64
1000	97.94	97.95	97.95	0	100	97.94

Tableau 6 : L'évaluation des résultats obtenus par l'algorithme KNN avec la distance euclidien

- Remarque:

1-pour la taille des donnée 200 ligne, le pourcentage de l'exactitude est 92.64% et le Rappel est 92.64%.

2- pour la taille des donnée 1000 ligne, le pourcentage de l'exactitude est 97.94% et le Rappel est 97.95%.

Donc, l'augmentation du taille des données affecte positivement à la qualité des résultats.

2- Le graphe :

la Figure 24, représente les résultats de l'exactitude obtenue par l'algorithme Knn (La distance euclidienne) par rapport Les nombres d'itérations de l'algorithme PSO :

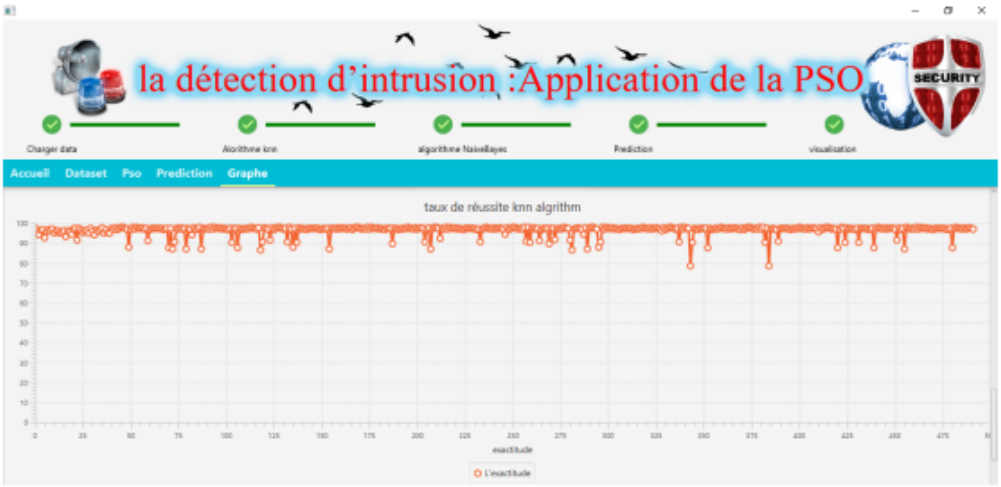


Figure 24: Taux de réussite ou bien l'exactitude de KNN (La distance euclidienne) par rapport Les nombres d'itérations de l'algorithme PSO

L'axe X représente Les nombres d'itérations de l'algorithme PSO et l'axe Y représente le pourcentage d'exactitude de l'algorithme Knn (La distance euclidienne) .

3- Matrice de confusion:

Dans ces matrices nous avons le nombre de vrai classe. Prédire classe que l'algorithme KNN (La distance euclidienne) a détecté lors de l'apprentissage.

- Vrai classe: c'est à dire la classe d'origine pour chaque ligne.
- Prédire classe: la classe choisir par le classificateur KNN (La distance euclidienne)

Ces matrices montrent les résultats du meilleur vecteur choisi par l'algorithme Knn (La distance euclidienne), avec les différentes tailles des données de data set:

La taille des données	Vrai classe	Prédire classe	
		Attaque (anormale)	Normale
200	Attaque (anormale)	62	4
	Normale	6	64

Tableau 7.1: La matrice de confusion de L'algorithme KNN avec la distance euclidienne

La taille des données	Vrai classe	Prédire classe	
		Attaque (anormale)	Normale
300	Attaque (anormale)	88	0
	Normale	4	112

Tableau 7.2: La matrice de confusion de L'algorithme KNN avec la distance euclidienne

La taille des données	Vrai classe	Prédire classe	
		Attaque (anormale)	Normale
400	Attaque (anormale)	120	0
	Normale	8	144

Tableau 7.3: La matrice de confusion de L'algorithme KNN avec la distance euclidienne

La taille des données	Vrai classe	Prédire classe	
		Attaque (anormale)	Normale
500	Attaque (anormale)	154	2
	Normale	6	178

Tableau 7.4: La matrice de confusion de L'algorithme KNN avec la distance euclidienne

La taille des données	Vrai classe	Prédire classe	
		Attaque (anormale)	Normale
1000	Attaque (anormale)	332	4
	Normale	10	334

Tableau 7.5: La matrice de confusion de L'algorithme KNN avec la distance euclidienne

- Remarque :

1-pour la taille des donnée 1000 ligne, le nombre des attaques bien détectées est 332 attaques.

2-pour la taille des donnée 100 ligne, le nombre des attaques bien détectées est 34 attaques.

3-L'algorithme Knn (La distance euclidienne) a pu détecter plus de 298 attaques entre 1000 et 100 lignes.

Donc, Augmenter de la taille des données donne la possibilité de détecter plus d'attaque.

4.5.5 Deuxième cas pour l'apprentissage: l'algorithme KNN

Dans ce cas, on donne les même valeur que l'on donné à les paramètre de l'algorithme PSO dans le premier cas.

Et on change la distance pour l'algorithme KNN.

Les valeurs choisi pour les paramètre de l'algorithme KNN sont:

- Nbr K=10
- La distance choisie est distance manhattan

1- le tableau :

Le tableau suivant représente les Mesures de performance de l'algorithme KNN avec la distance manhattan :

La taille des données	Exactitude %	Rappe %	Pricision %	Entropie %	F_score %	Fitness %
200	92.64	92.64	93.58	0	93.11	92.88
300	96.07	95.84	96.27	5.26	96.06	96.06
400	97.79	97.65	98	2.85	97.82	97.81
500	97.64	97.56	97.72	3.23	97.64	97.64
1000	97.90	97.90	97.90	0	100	97.90

Tableau 8: L'évaluation des résultats obtenus par l'algorithme KNN avec la distance manhattan

Le résultat est presque équivalente entre L'algorithme KNN avec la distance euclidien et L'algorithme KNN avec la distance manhattan , juste dans L'entropie L'algorithme KNN avec la distance manhattan par exemple La taille des données est 200 lignes, Entropie=0% pour cent par contre L'algorithme KNN avec la distance euclidienne Entropie=10.15%.

2- Le graphe :

la Figure 25, représente les résultats de l'exactitude obtenue par l'algorithme Knn (la distance manhattan) par rapport Les nombres d'itérations de l'algorithme PSO:

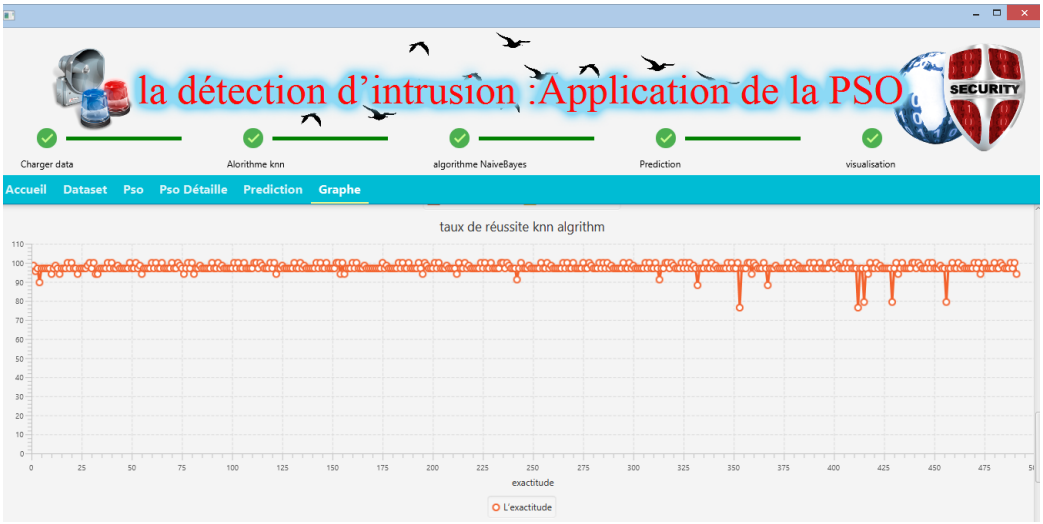


Figure 25: Taux de réussite ou bien l'exactitude de KNN (la distance manhattan) par rapport Les nombres d'itérations de l'algorithme PSO

L'axe X représente Les nombres d'itérations de l'algorithme PSO et l'axe Y représente le pourcentage d'exactitude de l'algorithme Knn (la distance manhattan) .

3- Matrice de confusion :

Dans ces matrices nous avons le nombre de vrai classe. Prédire classe que l'algorithme KNN (La distance manhattan) a détecté lors de l'apprentissage.

La taille des données	Vrai classe	Prédire classe	
		Attaque (anormale)	Normale
200	Attaque (anormale)	58	0
	Normale	10	68

Tableau 9.1: Matrice de confusion de l'algorithme KNN avec la distance manhattan

La taille des données	Vrai classe	Prédire classe	
		Attaque (anormale)	Normale
300	Attaque (anormale)	86	2
	Normale	6	110

Tableau 9.2: Matrice de confusion de l'algorithme KNN avec la distance manhattan

La taille des données	Vrai classe	Prédire classe	
		Attaque (anormale)	Normale
400	Attaque (anormale)	122	0
	Normale	6	142

Tableau 9.3: Matrice de confusion de l'algorithme KNN avec la distance manhattan

La taille des données	Vrai classe	Prédire classe	
		Attaque (anormale)	Normale
500	Attaque (anormale)	154	2
	Normale	6	178

Tableau 9.4: Matrice de confusion de l'algorithme KNN avec la distance manhattan

La taille des données	Vrai classe	Prédire classe	
		Attaque (anormale)	Normale
1000	Attaque(anormale)	330	2
	Normale	8	332

Tableau 9.5: Matrice de confusion de l'algorithme KNN avec la distance manhattan

la matrice de confusion de L'algorithme KNN (la distance euclidienne) a peu avancé par rapport à Matrice de confusion de L'algorithme KNN (la distance manhattan) dans la détection des attaques.

Les même remarques pour La matrice de confusion de L'algorithme KNN (la distance euclidienne) et la matrice de confusion de L'algorithme KNN (la distance manhattan). L'augmentation du nombre de ligne du data set donne la chance pour détecter les attaques.

4.5.6 Troisième cas pour l'apprentissage : Algorithme naive bayes

Dans ce cas, on donne des valeurs pour les paramètre de l'algorithme PSO:

- Nbre itérations = 10
- Nbre Essaim= 40

1- le tableau :

le tableau suivant représente les Mesures de performance de l'algorithme Nive bayes :

La taille des données	Exactitude %	Rappel %	Pricision %	Entropie %	F_score %	Fitness %
200	92.64	92.64	93.58	8.64	93.11	92.88
300	97.05	96.93	97.13	0	97.03	97.04
400	94.85	94.87	94.81	0	94.98	94.84
500	95.29	95.34	95.24	0	95.29	95.29
1000	92.05	92.05	92.1	10.93	100	92.06

Tableau 10: Évaluation des résultats obtenus L'algorithme naive bayes

D'aprés les résultats du tableau 10, on remarque que les résultets dans L'algorithme naive bayes sont de moins bonne qualité que L'algorithme KNN avec la distance euclidienne ou la distance manhattan.

par exemple pour L'algorithme naive bayes, La taille des données est 1000 lignes, Exactitude=92.05% et Rappel=92.05% par contre L'algorithme KNN avec la distance euclidienne Entropie=97.94% et Rappel=97.95% .

2- Le graphe :

La Figure 26, représente les résultats de l'exactitude obtenue par l'algorithme naive bayes par rapport Les nombres d'itérations de l'algorithme PSO :

Figure 26: Taux de réussite ou bien l'exactitude de naive bayes par rapport Les tentatives de l'algorithme PSO

L'axe X représente le nombre d'itération de l'algorithme PSO et l'axe Y représente le pourcentage d'exactitude de l'algorithme naive bayes .

3- Matrice de confusion :

Dans ces matrices nous avons le nombre de vrai classe. Prédire classe que l'algorithme nive bayes a détecté lors de l'apprentissage.

La taille des données	Vrai classe	Prédire classe	
		Attaque (anormale)	Normale
200	Attaque (anormale)	58	0
	Normale	10	68

Tableau 11.1: Matrice de confusion de L'algorithme naive bayes

La taille des données	Vrai classe	Prédire classe	
		Attaque (anormale)	Normale
300	Attaque (anormale)	88	2
	Normale	4	110

Tableau 11.2: Matrice de confusion de L'algorithme naive bayes

La taille des données	Vrai classe	Prédire classe	
		Attaque (anormale)	Normale
400	Attaque (anormale)	122	8
	Normale	6	136

Tableau 11.3: Matrice de confusion de L'algorithme naive bayes

La taille des données	Vrai classe	Prédire classe	
		Attaque (anormale)	Normale
500	Attaque (anormale)	154	10
	Normale	6	170

Tableau 11.4: Matrice de confusion de L'algorithme naive bayes

La taille des données	Vrai classe	Prédire classe	
		Attaque (anormale)	Normale
1000	Attaque (anormale)	306	32
	Normale	10	320

Tableau 11.5: Matrice de confusion de L'algorithme naive bayes

- Remarque :

1-pour L'algorithme KNN avec la distance euclidienne la taille des donnée 1000 ligne, le nombre des attaques bien détectées est 332 attaques.

2-Pour L'algorithme naive bayes la taille des donnée 1000 ligne, le nombre des attaques bien détectées est 306 attaques.

Le Knn a détecté plus de 25 attaque par rapport à l'algorithme nive bayes

4- Le graphe de comparaison:

La Figure 27, montre la comparaison entre matrice de confusion de L'algorithme KNN (la distance euclidienne) et matrice de confusion de L'algorithme Nive bayes.

Dans ce cas les valeurs choisi pour les paramètre de l'algorithme Pso sont:

- Nbre itérations = 10
- Nbre Essaim= 40

Les valeurs choisi pour les paramètre de l'algorithme KNN sont:

- Nbr K=10
- La distance choisie est distance euclidienne
- La taille des données de data set choisi est : 1000 lignes

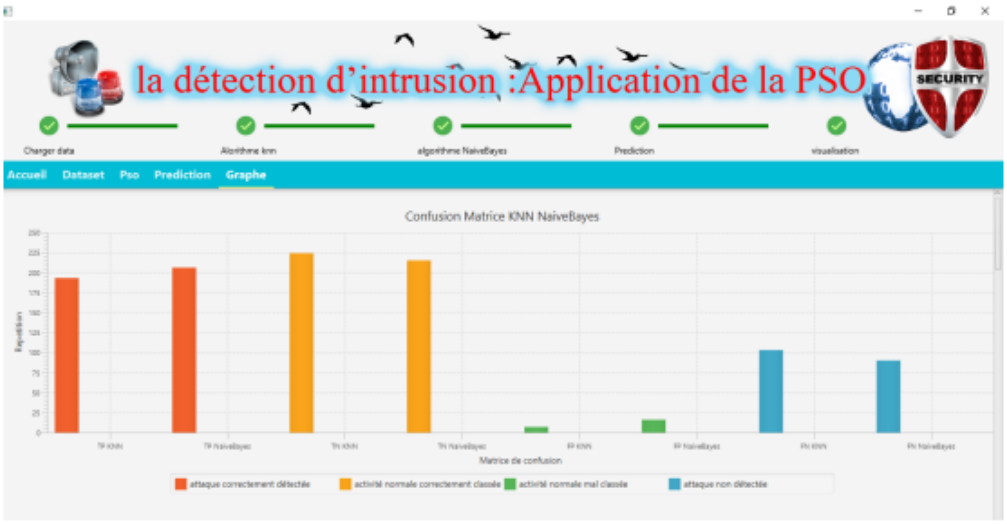


Figure 27:Comparaison de matrice de confusion entre KNN et Le naive bayes

Les colonnes en rouge représentent les attaques correctement détectée (True Positive (TP)).

Les colonnes en orange représentent les activité normale correctement classées(True negative(TN)).

Les colonnes en vert représentent les activité normale mal classées(false positive(FP)).

Les colonnes en bleu représentent les attaques non détectée (false negative(FN)) .

- la sélection du prédiction:

On passent à la sélection du prédiction pour donne un nouveau fichier qui contient des nouveaux attaques.Dans cette sélection, les algorithmes de classification(Knn et nive bayes) effectuent le test.

Dans la figure 28, les résultats des tests sont affichés:

Figure 28: la classification de KNN et nive bayes

- Le premier tableau représente les résultats des test de l'algorithme Knn.

- Le deuxième tableau représente les résultats des test de l'algorithme nive bayes.

1- Le graphe de comparaison :

La Figure 29 représente le comparaison entre les résultats des deux algorithmes dans la classification des classes.

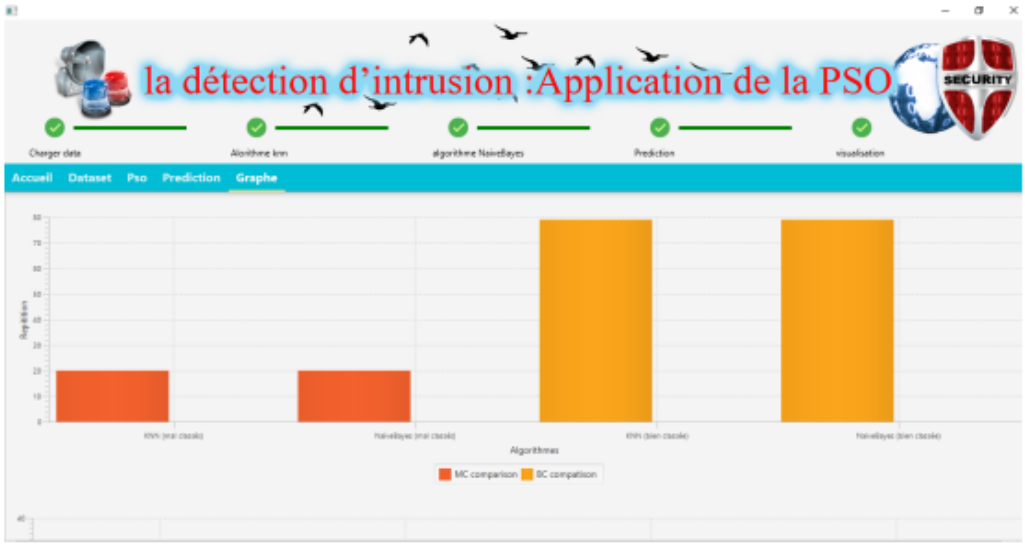


Figure 29: Comparaison de classification entre KNN et Le naive bayes

- Les colonnes en rouge représentent les pourcentages des classes qui ne sont pas bien classées (mal classés (MC)).

- Les colonnes en orange représentent les pourcentages des classes qui on correctement classées (bien classées (BC)).

Les résultats entre les deux algorithmes sont similaires.

4.6 Conclusion:

L'augmentation rapide des infrastructures et des applications numériques invite les graves problèmes de sécurité informatique. La solution existante pour ce contexte n'est pas très efficace et doit être amélioré en continu. Les techniques basées sur des règles ne sont pas non plus très utiles efficace en raison de chaque jour les schémas d'attaques peuvent être différent.

Ainsi, le travail proposé est motivé pour concevoir une technique d'exploration de données efficace et précise pour classer les modèles malveillants disponibles dans l'ensemble de données d'un IDS. Dans ce contexte le deux problèmes majeurs sont :

-la grande quantité de données à classer ayant un impact direct sur la performance des classifications

- l'efficacité des programmes mis en œuvre.

Par conséquent, dans ce modèle, deux phases de réduction de dimensionnalité sont appliquées ; la première lors du prétraitement des valeurs nulles qui ont été supprimées et la second est l'utilisation de l'algorithme PSO pour la sélection des meilleures caractéristiques parmi tout l'ensemble de données.

Le jeu de données contient un grand nombre d'instances de données et l'important quantité d'attributs, c'est-à-dire 41 attributs et 1 étiquette de classe. Ainsi, le PSO est utilisé pour réduire les fonctionnalités pour une classification efficace.

Conclusion général:

Aujourd'hui, les attaques informatiques sont devenues une véritable menace pour les systèmes informatiques et les réseaux , ce qui nous a poussés à faire ce travail, et à travailler à développer un modèle de sécurité capable de faire face aux risques et menaces en détectant les tentatives malveillantes, qu'elles soient connues ou nouvelles. Pour atteindre cet objectif, nous avons utilisé l'algorthme pso, et plus précisément les algorithm de classification KNN et nive bayes .

Selon le principe de fonctionnement des systèmes de détection d'intrusion, on peut diviser ces systèmes en deux parties : premièrement, ceux qui détectent les actions malveillantes (on parle alors de l'approche par signature ou par scénario) et deuxièmement, qui détectent les anomalies (on parle alors de l'approche comportementale) , où la première approche compare le comportement du système avec l'utilisation d'attaques déjà connues, ce qui empêche De la découverte d'attaques nouvelles et développées par l'ennemi, quant à la deuxième approche, elle considère le mouvement comme normal, mais elle découvre le comportement qui ressort de ce mouvement, et cela rend le taux de les fausses alertes sont faibles et le taux de réussite beaucoup.

C'est la raison pour laquelle la sécurité est de la responsabilité de l'utilisateur et de la personne responsable du système .

Bibliography

[1] Schatz, Daniel; Bashroush, Rabih; Wall, Julie (2017). "Towards a More Representative Definition of Cyber Security".

[2] Journal of Digital Forensics, Security and Law. 12 (2). ISSN 1558-7215

[3] Beckers, K. (2015). Pattern and Security Requirements: Engineering-Based Establishment of Security Standards. Springer. p. 100. ISBN 9783319166643.

[4] Boritz, J. Efrim (2005). "IS Practitioners' Views on Core Concepts of Information Integrity". International Journal of Accounting Information Systems. Elsevier. 6 (4): 260–279. doi:10.1016/j.accinf.2005.07.001

[5] Hryshko, I. (2020). "Unauthorized Occupation of Land and Unauthorized Construction: Concepts and Types of Tactical Means of Investigation". International Humanitarian University Herald. Jurisprudence (43): 180–184. doi:10.32841/2307-1745.2020.43.40. ISSN 2307-1745.

[6] "Video from SPIE - the International Society for Optics and Photonics". dx.doi.org. doi:10.1117/12.2266326.5459349132001. Retrieved 202105-29

[7] "Communication Skills Used by Information Systems Graduates". Issues in Information Systems. 2005. doi:$10.48009/1_i is_2 005_3 11 - 317$. $ISSN$ $1529 - 7314$.

[8] Daniel Guinier, Sécurité et qualité des systèmes d'information : Approche systémique, Masson, 1992, 298 p. (ISBN 978-2-225-82686-3)

[9] McCarthy, C. (2006). "Digital Libraries: Security and Preservation Considerations". In Bidgoli, H. (ed.). Handbook of Information Security, Threats, Vulnerabilities, Prevention, Detection, and Management. Vol. 3. John Wiley Sons. pp. 49–76. ISBN 9780470051214.

[10] Guillon, F. Les politiques de sécurité - Enjeux et choix de société. L'Harmattan, Paris, déc. 2016

[11] ZDI Disclosure Policy Changes [archive], par Aaron Portnoy (Tipping Point), le 3 août 2010

[12] Cani, Andrea; Gaudesi, Marco; Sanchez, Ernesto; Squillero, Giovanni; Tonda, Alberto (24 March 2014). "Towards automated malware creation: code generation and code integration". Proceedings of the 29th Annual ACM Symposium on Applied Computing. SAC '14. New York, NY, USA: Association for Computing Machinery: 157–160. doi:10.1145/2554850.2555157. ISBN 978-1-4503-2469-4. S2CID 14324560,"computer virus – Encyclopædia Britannica". Britannica.com. Retrieved 28 April 2013

[13] "What are viruses, worms, and Trojan horses?". Indiana University. The Trustees of Indiana University. Retrieved 23 February 2015, Landwehr, C. E; A. R Bull;

[14] J.P McDermott; W. S Choi (1993). A taxonomy of computer program security flaws, with examples. DTIC Document. Retrieved 5 April 2012, "Trojan Horse Definition". Retrieved 5 April 2012.

[15] "Trojan horse". Webopedia. Retrieved 5 April 2012,"What is Trojan horse? – Definition from Whatis.com". Retrieved 5 April 2012, "Trojan Horse: [coined By MIT-hacker-turned-NSA-spook Dan Edwards] N." Archived from the original on 5 July 2017. Retrieved 5 April 2012

[16] Vincentas (11 July 2013). "Malware in SpyWareLoop.com". Spyware Loop. Retrieved 28 July 2013.

[17] McDowell, Mindi. "Understanding Hidden Threats: Rootkits and Botnets". US-CERT. Retrieved 6 February 2013.

[18] Henry, Alan. "The Difference Between Antivirus and Anti-Malware (and Which to Use)". Archived from the original on November 22, 2013.

[19] Boudriga, Noureddine (2010). Security of mobile communications. Boca Raton: CRC Press. pp. 32–33. ISBN 978-0849379420,Oppliger, Rolf (May 1997). "Internet Security: FIREWALLS and BEYOND". Communications of the ACM. 40 (5): 94. doi:10.1145/253769.253802. S2CID 15271915

[20] K. Salah, K. Sattar, Z. Baig, M. Sqalli, and P. Calyam, "Resiliency of opensource firewalls against remote discovery of lastmatching rules," in Proceedings of the 2nd International Conference on Security of Information and Networks, SIN '09, (New York, NY, USA), p. 186–192, Association for Computing Machinery, 2009.

[21] Peltier, Justin; Peltier, Thomas R. (2007). Complete Guide to CISM Certification. Hoboken: CRC Press. p. 210. ISBN 9781420013252.

[22] M. Afshar Alam; Tamanna Siddiqui; K. R. Seeja (2013). Recent Developments in Computing and Its Applications. I. K. International Pvt Ltd. p. 513. ISBN 978-93-80026-78-7,"Firewalls". MemeBridge. Retrieved 13 June 2014.John Pescatore (October 2, 2008). "This Week in Network Security History: The Firewall Toolkit". Archived from the original on April 29, 2016. Retrieved 2018-12-28,Marcus J. Ranum; Frederick Avolio. "FWTK history","What is Layer 7? How Layer 7 of the Internet Works". Cloudflare. Retrieved Aug 29, 2020.

[23] "What is an Intrusion Detection System (IDS)? — Check Point Software".,Martellini, Maurizio; Malizia, Andrea (2017-10-30). Cyber and Chemical, Biological, Radiological, Nuclear, Explosives Challenges: Threats and Counter Efforts. Springer. ISBN 9783319621081.

[24] dr/XPOSE2007/plebacco$_i ds$.

[25] Silvia Farraposo, Philippe Owezarski, Edmundo Monteiro Détection, classification et identification d'anomalies de trafic

[26] Memoire de fin d'etudes.Utilisation des métaheuristiques pour la résolution du problème de sélection d'attributs : Application à la détection d'intrusionssur

[27] Gurley., Bace, Rebecca (2001). Intrusion detection systems. [U.S. Dept. of Commerce, Technology Administration, National Institute of Standards and Technology]. OCLC 70689163.

[28] Scarfone, K. A.; Mell, P. M. (February 2007). "NIST – Guide to Intrusion Detection and Prevention Systems (IDPS)" (PDF). doi:10.6028/NIST.SP.800-94. Retrieved 2010-06-25.John R. Vacca (2010). Managing Information Security. Syngress. p. 137. ISBN 978-1-59749-533-2. Retrieved 29 June 2010.

[29] Michael E. Whitman; Herbert J. Mattord (2009). Principles of Information Security. Cengage Learning EMEA. ISBN 978-1-4239-0177-8. Retrieved 25

June 2010 , Engin Kirda; Somesh Jha; Davide Balzarotti (2009). Recent Advances in Intrusion Detection: 12th International Symposium, RAID 2009, Saint-Malo, France, September 23–25, 2009, Proceedings. Springer. p. 162. ISBN 978-3-642-04341-3. Retrieved 29 June 2010.

[30] mimoir online.Architecture Et Type De Système D'instruction - cloudfront.net sur: ¡https://d1n7iqsz6ob2ad.cloudfront.net¿

[31] mimoir online.Architecture Et Type De Système D'instruction - cloudfront.net sur: ¡https://d1n7iqsz6ob2ad.cloudfront.net¿

[32] V. Angel, La rugosité des paysages : une théorie pour la difficulté des problèmes d'optimisation combinatoire relativement aux métaheuristiques, thèse de doctorat de l'université de Paris-Sud, Orsay, 1998.

[33] Blum, C.; Roli, A. (2003). "Metaheuristics in combinatorial optimization: Overview and conceptual comparison". 35 (3). ACM Computing Surveys: 268–308.

[34] Bianchi, Leonora; Marco Dorigo; Luca Maria Gambardella; Walter J. Gutjahr (2009). "A survey on metaheuristics for stochastic combinatorial optimization" (PDF). Natural Computing. 8 (2): 239–287. doi:10.1007/s11047-008-9098-4. S2CID 9141490.

[35] D, Binu (2019). "RideNN: A New Rider Optimization Algorithm-Based Neural Network for Fault Diagnosis in Analog Circuits". IEEE Transactions on Instrumentation and Measurement. 68 (1): 2–26. doi:10.1109/TIM.2018.2836058. S2CID 54459927.

[36] Talbi, E-G. (2009). Metaheuristics: from design to implementation. Wiley. ISBN 978-0-470-27858-1

[37] M. Dorigo, Optimization, Learning and Natural Algorithms, PhD thesis, Politecnico di Milano, Italie, 1992

[38] Moscato, P. (1989). "On Evolution, Search, Optimization, Genetic Algorithms and Martial Arts: Towards Memetic Algorithms". Caltech Concurrent Computation Program (report 826)

[39] Sörensen, Kenneth (2015). "Metaheuristics—the metaphor exposed" (PDF). International Transactions in Operational Research. 22: 3–18. CiteSeerX 10.1.1.470.3422. doi:10.1111/itor.12001. Archived from the original (PDF) on 2013-11-02

[40] Moscato, P. (2012). "Metaheuristic optimization frameworks a survey and benchmarking". Soft Comput. 16 (3): 527–561. doi:10.1007/s00500-011-0754-8. hdl:11441/24597. S2CID 1497912

[41] "Data Mining Curriculum". ACM SIGKDD. 2006-04-30. Retrieved 2014-01-27.

[42] Clifton, Christopher (2010). "Encyclopædia Britannica: Definition of Data Mining". Retrieved 2010-12-09.

[43] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Archived from the original on 2009-11-10. Retrieved 2012-08-07.

[44] Han, Jaiwei; Kamber, Micheline; Pei, Jian (2011). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann. ISBN 978-0-12-381479-1

[45] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases" (PDF). Retrieved 17 December 2008.

[46] https://jafwin.com/2019/01/14/top-5-des-outils-les-plus-utilises-en-data-mining

[47] https://www.piloter.org/business-intelligence/datamining.htmmethode

[48] https://www.javatpoint.com/data-mining

[49] https://docs.oracle.com

[50] https://cloudtweaks.com/2014/09/supervised-unsupervised-data mining

[51] Stahl, Daniel (2011) "Miscellaneous Clustering Methods"

[52] Everitt, Brian S.; Landau, Sabine; Leese, Morven; and , in Cluster Analysis, 5th Edition, John Wiley Sons, Ltd., Chichester, UK

[53] Nigsch, Florian; Bender, Andreas; van Buuren, Bernd; Tissen, Jos; Nigsch, Eduard; Mitchell, John B. O. (2006). "Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization". Journal of Chemical Information and Modeling. 46 (6): 2412–2422. doi:10.1021/ci060149f. PMID 17125183.

[54] Hall, Peter; Park, Byeong U.; Samworth, Richard J. (2008). "Choice of neighbor order in nearest-neighbor classification". Annals of Statistics. 36 (5): 2135–2152. arXiv:0810.5276. Bibcode:2008arXiv0810.5276H. doi:10.1214/07-AOS537. S2CID 14059866.

[55] https://datascientest.com/knn

[56] Bonyadi, M. R.; Michalewicz, Z. (2017). "Particle swarm optimization for single objective continuous space problems: a review".

[57] KENNEDY J., EBERHART R., "Particle Swarm Optimization," Proceedings of the IEEE International Joint Conference on Neural Networks, IEEE Press, vol. 8, no. 3, pp. 1943–1948. 1995.

[58] COOREN Y., Perfectionnement d'un algorithme adaptatif d'Optimisation par Essaim Particulaire. Applications en génie médical et en électronique. Thèse de Doctorat, Université de Paris 12 Val de Marne, France. 2008.

[59] BOCHNEK B., FORY'S P., Structural optimization for post buckling behavior using particle swarms. Struct Multidisc Optim, p. 521-531. 2006.

[60] ELHAMI N., Contribution aux méthodes hybrides d'optimisation heuristiques : Distribution et application à l'interopérabilité des systèmes d'information. Thèse de Doctorat, Université Mohammed V Rabat, Maroc Université de Rouen, France, 2013.

[61] https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained

[62] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," Submitted to Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 2009.

[63] "Weka Package Metadata". SourceForge. 2017. Retrieved 2017-11-11.

[64] Reutemann, Peter; Pfahringer, Bernhard; Frank, Eibe (2004). "Proper: A Toolbox for Learning from Relational Data with Propositional and Multi-Instance Learners". 17th Australian Joint Conference on Artificial Intelligence (AI2004). Springer-Verlag. CiteSeerX 10.1.1.459.8443

[65] A. DJEFFAL, Cours Fouille de données avancée. Université Mohamed Khider Biskra, 20142015. www.abdelhamiddjeffal.net.

[66] M. LABONNE, A. OLIVEREAU, and D. ZEGHLACHE, "Automatisation du processus d'entraînement d'un ensemble d'algorithmes de machine learning optimisés pour la détection d'intrusion," 2018. cesarconference.org.

ملخص

تتزايد المتطلبات الحوسبة كل يوم ،هما يسمح بمضاعدد الهجومات ، لذلك من اجل تأمين الشبكات و الانظمة من المهاجمين والأنشطة الخبيثة العمل المقترح هو تقديم IDS محسن (نظام كشف التسلل).

IDS هي تقنية تعتمد على خوارزمية استخراج البيانات لتصنيف الأنماط الخبيثة. من أجل تنفيذ هذه التقنية يتم استخدام مجموعة بيانات NSL-KDD. تحتوي هذه المجموعة على 41 سمة وسمة فئة واحدة. يمكن لهذا البعد الضخم التأثير على أداء نظام IDS. لذلك يتم استخدام تقنية التحسين PSO(تحسين سرب الجسيمات). استخدام هذه الخوارزمية ، ترتيب جميع السمات وتحدد الميزات.

الميزات المحددة أصغر حجمًا مما يعني أنها تحتوي على 21 سمة وسمة فئة واحدة. في هذه الميزات المحددة ، يتم استخدام خوارزمية KNN لتصنيف البيانات. توضح النتائج التجريبية على أحجام مختلفة من مجموعة البيانات الأداء الفعال لنموذج البيانات المقترح. تتم مقارنة ذلك أيضًا بنموذج التصنيف Nive bayesذي الصلة.

المقارنة توضح تحليل الأداء النموذج المقترح دقيق ويستغرق وقتًا أقل لتصنيف الأنماط مقارنةً بنموذج KNN لكن استخدامات الذاكرة للنموذج المقترح أعلى فيما يتعلق بنموذج Nive bayes

Abstract

The new computational requirements are growing every day,and taken advantages of these services. But These Networks are not fully secured a significant amount of attacks can be deployed on these networks. Therefore to secure the network from the attackers and malicious activities the proposed work is motivated to deliver enhanced IDS (intrusion detection system). That IDS is a data mining algorithm based technique for classifying the malicious patterns. In order to implement this technique the NSL-KDD dataset is used. That dataset contains 41 attributes and 1 class attribute. This huge dimension can impact on the performance of IDS system. Therefore first the data processing technique is used to cleaning the data. After that the PSO (Particle swarm optimization) technique is used. Using this algorithm , rank all the attributes and select the features. The selected features are less in size means it contains 21 attributes and 1 class attribute. In this selected features the nive bayes algorithm is employed for classifying the data. The experimental results on different size of dataset demonstrate the effective performance of the proposed data model. That is also compared with the relevant k-NN classification model. The comparative performance analysis demonstrate the proposed model is accurate and less time consuming for classification of patterns as compared to the Nive bayes based model. But the memory usages of the proposed model are higher with respect to the k-NN model.

Résumé

Les besoins informatiques augmentent chaque jour, ils permettent de multiples le nombre des attaques, donc afin de sécuriser les réseaux et les systèmes contre les attaquants et les activités malveillantes, le travail proposé est de founir un IDS amélioré (Intrusion Detection System).

IDS est une technologie basée sur un algorithme d'exploration de données pour classer les modèles malveillants. Afin de mettre en œuvre cette technique, il est utilisé le Data set NSL-KDD. Cet ensemble contient 41 attributs et un attribut de catégorie. Cette énorme dimension peut affecter les performances d'un système IDS. La technologie PSO (Particle Swarm Optimization) est donc utilisée. À l'aide de cet algorithme, classez toutes les fonctionnalités et identifiez les fonctionnalités.

Les entités sélectionnées sont de plus petite taille, ce qui signifie qu'elles contiennent 21 attributs et un attribut de classe. Dans ces spécificités, l'algorithme KNN est utilisé pour classer les données. Les résultats expérimentaux sur différentes tailles de données démontrent la performance effective du modèle de données proposé. Ceci est également comparé au modèle de classification Nive bayes.

L'analyse des performances de comparaison montre que le modèle proposé est précis et prend moins de temps pour classer les modèles par rapport au modèle KNN, mais les utilisations de la mémoire du modèle proposé sont plus élevées par rapport au modèle Nive bayes.