REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE MOULOUD MAMMERI DE TIZI-OUZOU
FACULTE DE GENIE ELECTRIQUE ET INFORMATIQUE
DEPARTEMENT INFORMATIQUE

Mémoire de Fin d'étude
En vue de l'obtention du diplôme de Master en Système informatique
Option : Système Informatique

Thème

Application des Algorithmes Génétiques dans la Détection d'intrusion Réseau

Proposé et Dirigé par :

- **M^r CHAIEB.Y**

Réalisé par :

- **M^elle DEBIANE Thiziri**
- **M^elle DIDAOUI Zina**

PROMOTION : 2014/2015

Remerciements

Nous remercions le bon dieu de nous avoir mis sur la route du savoir

Nous remercions infiniment notre promoteur : Monsieur CHAIEB Yazid, pour nous avoir encadrées durant cette année, Nous tenons à lui exprimer notre profonde gratitude pour sa disponibilité tout au long de la réalisation de ce mémoire, ainsi pour l'orientation, la confiance, l'aide et le temps qu'il a bien voulu nous consacrer.

Nous remercions également les membres du jury pour nous avoir honorées en jugeant notre travail.

Nous tenons à saluer la peine et l'effort fournis par l'ensemble de nos professeurs afin d'assurer notre formation tout au long de notre cursus universitaire et leur disons de ce fait, merci.

Nous exprimons notre infinie gratitude à nos chers parents en reconnaissance de leurs sacrifices, dévouement, soutien et encouragements.

Enfin, nous remercions tous ceux qui ont contribué de prés ou de loin à réaliser ce modeste travail.

Ces quelques mots ne traduisent guère tout ce que nous avons pu recevoir de la part de chacun d'entre eux, mais nous souhaitons néanmoins qu'ils y trouvent l'expression de notre infinie reconnaissance.

Dédicaces

Je dédie ce modeste travail à :

Mes très chers parents qui m'ont toujours soutenue et encouragée dans tout ce que j'entreprends et qui ont fait de moi ce que je suis aujourd'hui.

Mes adorables sœurs aux quelles je tiens énormément : Thileli, Thanina, Tassadit et Thinhinane.

Ma chère binôme : Zina, sans qui ce travail ne serait pas accompli.

Ma meilleure amie : Nassima, qui a toujours été présente pour moi.

Mon seul et Grand Amour : Brahim, qui m'apporte le bonheur, la joie et l'amour chaque jour que Dieu fait.

Toute personne ayant contribuée de prés ou de loin au bon déroulement de ce projet.

Thiziri

Dédicaces

Je dédie ce modeste travail à :

Mes très chers parents qui ne cessent de m'encourager, me soutenir et veiller à ce que je ne manque de rien pour réussir. Je leur dois tout.

Mes très chers frères que j'aime beaucoup : Ilyes, Adel.

Mes sœurs qui comptent énormément pour moi et que j'adore plus que tout : Loubna, Assia.

Mes anges : Dihia, Nadir, Sonia.

Mon seul et Grand Amour : Athmane, qui m'apporte le bonheur, la joie et l'amour chaque jour que Dieu fait.

Mes beaux-frères : Samir et Mustapha ainsi que leurs familles.

Mes oncles et tantes.

Ma chère binôme Thiziri, sans qui ce travail ne serait pas accomplie.

Toutes les personnes auxquelles je tiens et qui me sont chères, elles se reconnaîtront.

Zina

Table des matières

Table des matières

Table des matières

Introduction Générale

Introduction

Les systèmes informatiques et les réseaux sont devenus des outils indispensables pour la société actuelle. Ils sont aujourd'hui déployés dans tous les secteurs professionnels. Initialement, isolés les uns des autres, ces systèmes informatiques sont devenus interconnectés et le nombre de points d'accès ne cesse de croître. Ce développement phénoménal est accompagné également par la croissance du nombre d'utilisateurs, qui ne sont pas forcément pleins de bonnes intentions vis-à-vis de ces systèmes informatiques. Ils peuvent exploiter les vulnérabilités des réseaux et les systèmes pour essayer d'accéder à des informations sensibles dans le but de les lire, les modifier ou les détruire, portant atteinte au bon fonctionnement du système.

Problématique

L'importance de sécurité des systèmes informatiques motive les angles divers de la recherche dont le but principal est de fournir de nouvelles solutions prometteuses qui ne pourraient être assurées par des méthodes classiques. Les systèmes de détection d'intrusions sont l'une de ces solutions qui permettent la détection des utilisations non autorisées, les mauvaises utilisations et les abus dans un système informatique par les utilisateurs externes ainsi que ses utilisateurs internes. Le défi dans le domaine de la sécurité informatique et plus précisément dans les systèmes de détection d'intrusions est de pouvoir déterminer la différence entre un fonctionnement normal et un fonctionnement avec intrus.

Objectif du travail

L'objectif de ce travail s'articule autour de ce domaine dont il consiste à sécuriser un réseau à l'aide d'un système de détection d'intrusion comportementale à base d'algorithme génétique.

- Le premier chapitre est un chapitre descriptif pour la sécurité des réseaux, sur lequel on va définir les menaces, les logiciels malveillants et une politique de sécurité, les principaux mécanismes de sécurité ainsi que l'application de la cryptographie.
- Le second chapitre est consacré à présenter le concept du Data Mining qui a conduit a la détection d'intrusion, une architecture globale d'un IDS, la définition et le mode de fonctionnement de ce dernier. Ainsi la classification des IDS et enfin la méthode de détection d'une intrusion.

- Le troisième chapitre illustre concrètement le processus d'application des algorithmes génétiques pour la sécurité informatique
- Le quatrième chapitre présente l'implémentation de l'algorithme ainsi que les résultats obtenus a la fin.

Nous terminerons ce mémoire en présentant nos conclusions ainsi que les perspectives de ce travail.

Chapitre I

La Sécurité Informatique

1. Introduction

La sécurité informatique est de nos jours devenue un problème majeur dans la gestion des réseaux d'entreprise ainsi que pour les particuliers toujours plus nombreux à se connecter à Internet. La transmission d'informations sensibles et le désir d'assurer la confidentialité de celles-ci est devenue un point primordial dans la mise en place de réseaux informatiques.

2. Définition [1]

La sécurité informatique désigne un ensemble de techniques et de bonnes pratiques pour protéger les ordinateurs et les données qui y sont stockées ainsi que d'assurer la confidentialité des transmissions d'informations. Si elles sont élaborées par des spécialistes, les plus simples doivent être connues et mises en œuvre par tous les utilisateurs.

3. Donnée - information – connaissance [2]

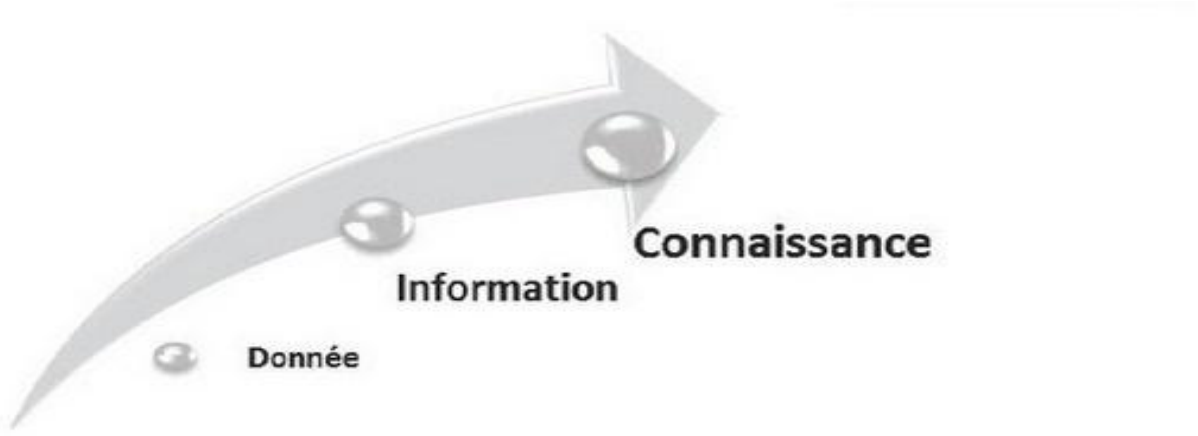


Figure 1 : Donnée-Information-Connaissance

La donnée est une valeur dans un champ. Elle peut être sous une forme cognitive, informatique ou dans des documents sous forme de texte .... (Temps = " Pluie " ou Pluie = " Oui " ...). Une donnée peut exprimer une mesure, un coût, une désignation, un état, etc..

L'information est une donnée avec une valeur particulière et une ou plusieurs significations, et parfois différentes selon les personnes et/ou le contexte (il pleut, il fait mauvais...).

La connaissance permet de traiter, comprendre des données ou des informations. Elle donne un sens à la donnée, qui devient alors une information, raisonne et agit ou fait agir en fonction. (Tiens ! Il pleut, donc il fait mauvais, je vais prendre mon parapluie).

4. Représentation de connaissances

4.1 Définition :

La représentation des connaissances désigne un ensemble d'outils et de technologies destinés d'une part à représenter et d'autre part à organiser le savoir humain pour l'utiliser et le partager. **[3]**

4.2 Qu'est-ce qu'une représentation

La représentation d'une entité sur laquelle on souhaite opérer est nécessairement une approximation de cette entité. Si la représentation devait posséder toutes les propriétés de l'entité représentée, il s'agirait de l'entité elle-même ! Par exemple une carte représente le territoire dans le cadre d'un processus de recherche d'un itinéraire. Une bonne carte n'est pas celle qui représente tout, mais celle qui met de côté les éléments dont n'a pas besoin pour accomplir cette opération. **[4]**

En résumé, une représentation est une structure de symboles pour décrire un modèle (une approximation) monde dans le contexte d'une tâche particulière.

4.3 Fondements logiques de la représentation des connaissances [4]

Les modèles de représentation des connaissances que nous allons étudiés reposent essentiellement sur des théories issues de la logique (en tant que discipline). En effet, pour manipuler des connaissances explicites, un système doit utiliser un langage formel de représentation, le plus efficacement possible. Toute expression de ce langage s'établit à l'aide de symboles dont les associations sont régies par des règles qui forment la syntaxe de la représentation. Si à toute expression syntaxiquement correcte de la représentation on fait correspondre une situation de l'univers de référence, on adjoint une sémantique à ce formalisme de représentation. Cette sémantique s'exprime souvent en terme booléens : la situation est vraie (dans l'univers considéré) ou elle n'est pas vraie.

Modèles de représentation des connaissances généralement issus de la logique mathématique

Manipulation de connaissances explicites :

❖ **Langage formel de représentation (symboles)**

Système formel

- Syntaxe :

– **Langage** : alphabet (ensemble de symboles) et procédé de formation d'expressions (grammaire)

– **Système de déduction** : règles pour construire de nouvelles formules à partir d'axiomes (vérités arbitraires)

- Sémantique

– **Règles d'évaluation**: pour associer une valeur (vraie ou faux) à toute formule du langage.

– Généralement compositionnelle : l'interprétation de la formule x=(y+3)/z dépend de celle de x, de =, de y, etc.

– Remarque : le langage naturel n'est pas compositionnel (le sens de l'expression « tout à l'heure » ne provient pas de la composition du sens de « tout », « à » et « l'heure »)

- Correction : tout ce qui est déductible à partir des axiomes est vrai (théorèmes).
- Complétude : tout ce qui est vrai peut être déduit à partir des axiomes
- Décidabilité : étant donnée une formule, il existe un procédé de calcul (algorithme) qui permet de dire en temps fini si c'est un théorème ou non

a) Logique propositionnelle :

- Système formel décidable dont la sémantique est compositionnelle
- Inconvénients

– Temps de décision exponentiels

– Expressivité limitée

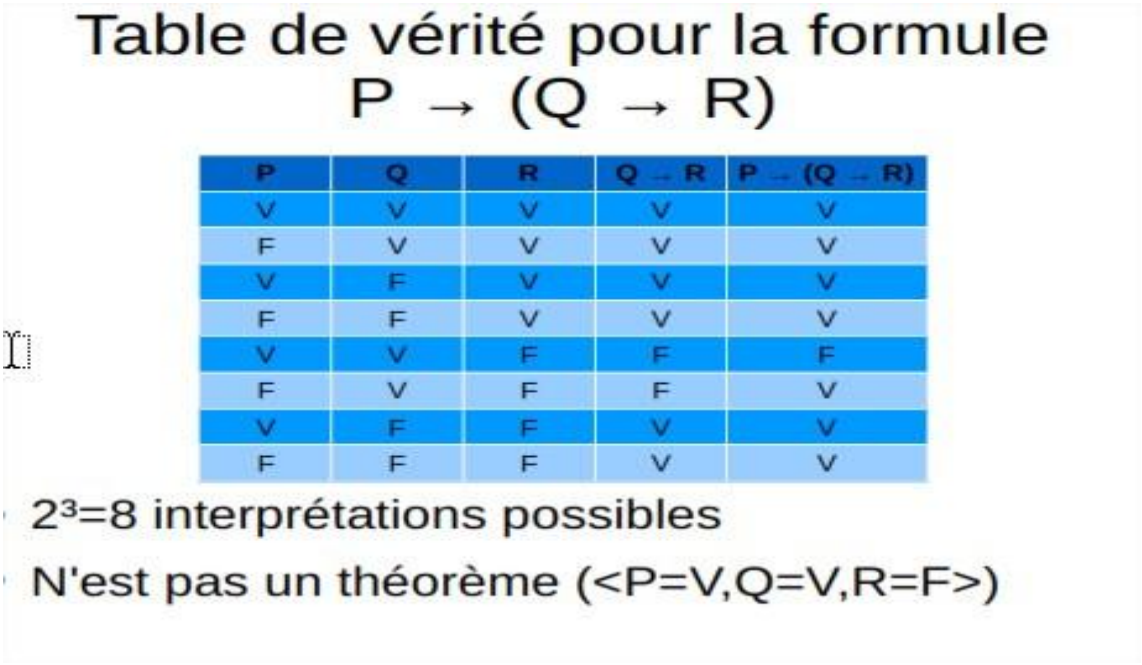


Table I.1 : Logique Propositionnelle

b) Logique du premier ordre :

- Relations entre objets
- Prédicats : lion(x) = "x est un lion"
- Quantificateurs : ∀ x (lion(x) → roi (x)) = "tous les lions sont des rois",

∃y (félin(y) ∧ roi(y)) = "certains félins sont des rois"

❖ **Compromis efficacité/expressivité :**

- Restreindre la logique pour avoir des algorithmes déductifs complets et rapides.
- Abandonner l'exigence de complétude.

4.4 Représentation des connaissances à base de catégories et d'objets : les ontologies

4.4.1 Systèmes experts

Il existe différentes manières d'envisager la représentation des connaissances. Une première approche consiste à utiliser des représentations à base de faits et de règles. Cette approche, appelée ingénierie des connaissances, a été employée dans la mise au point des premiers systèmes experts, notamment Mycin qui utilisait des outils inspirés des logiques multi évaluées et qui est parvenu à de bons résultats, prouvant ainsi l'intérêt des logiques pour représenter des connaissances.

4.4.2 Ontologies

Une ontologie définit des concepts abstraits (principes, idées, objets, temps, espace, etc.) et des relations. Elle inclut généralement une organisation hiérarchique des concepts pertinents et des relations qui existent entre ces concepts, ainsi que des règles et des axiomes qui les contraignent. Les ontologies sont généralement destinées au partage et à l'échange des connaissances qu'elles représentent et sont souvent conçues en vue d'être réutilisées dans le contexte d'une autre tâche.

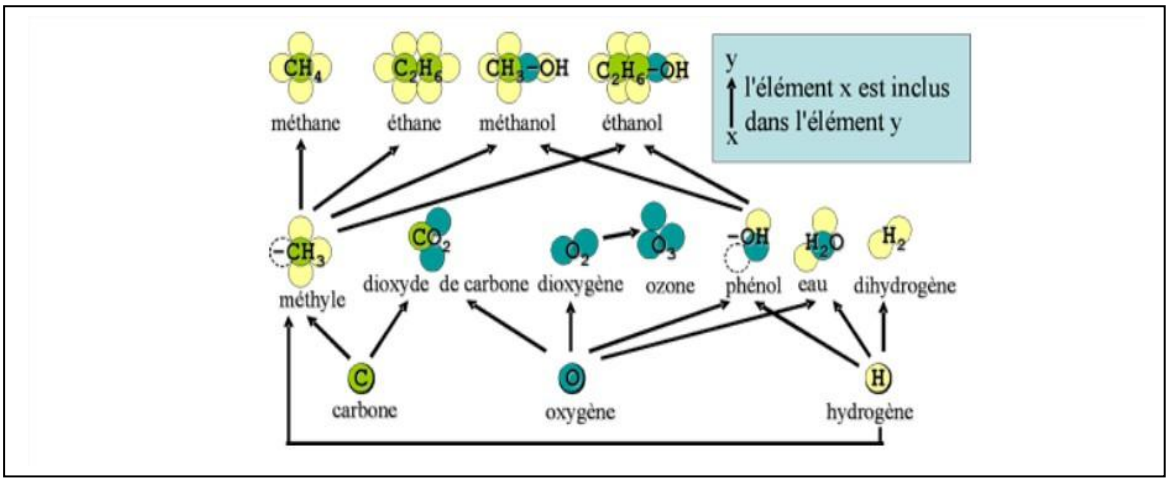


Figure 2 : Exemple d'ontologie en chimie qui montre une ontologie de domaine au formalisme simple

4.5 Représentation graphique des connaissances : les réseaux sémantiques

Les réseaux sémantiques, moins formels que les représentations logiques et plus directement développés par l'intelligence artificielle tirent leur origine de certains résultats de psychologie cognitive sur la mémoire associative et se veulent un modèle de représentation du contenu sémantique des concepts des langues qui soit psychologiquement valide.

Les réseaux sémantiques sont fondés sur la notion simple de graphe, formé de nœuds — représentant les concepts — reliés par des arcs. Cette unité de base qu'est le concept n'acquiert tout son sens que par les relations qui le lient aux autres concepts. Les arcs du graphe représentent alors des relations (généralement binaires) entre ces concepts. Les nœuds, comme les arcs sont étiquetés.

Il n'existe pas un, mais différents modèles de réseau sémantique, dont le point commun est de définir un formalisme de représentation graphique déclaratif pour représenter des connaissances et/ou pour raisonner sur ces connaissances

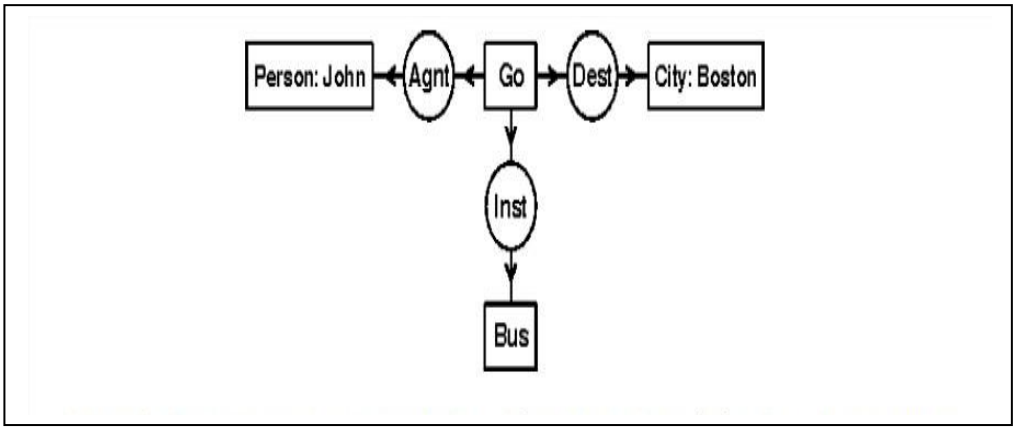


Figure 3 Graphe Conceptuel

5. Sécurisation de données

5.1 Définition [5]

De nombreux systèmes d'information sont distribués sur le net, et font pour remplir leur tâche appel à des composants distribués : bases de données, services, etc. En quelques décennies, le mode de gestion de tels systèmes est passé d'une vision système d'information centralisé, avec contrôle d'accès, le plus souvent localisé sur un seul site, à une vision système centralisé accessible depuis l'extérieur, mettant l'accent sur la sureté des protocoles d'échange. Les paradigmes de S.I en vogue sont actuellement d'une part le cloud, dans lequel les informations sont conservées dans des fermes de serveurs, migrent et se dupliquent de manière transparente à l'utilisateur, et d'autre part un paradigme de type « documents et services », c'est-à-dire des architecture de systèmes totalement distribuées, et ouvertes à l'extérieur.

La sécurité des systèmes d'information (SSI) est l'ensemble des moyens techniques, organisationnels, juridiques et humains nécessaires et mis en place pour conserver, rétablir, et garantir la sécurité du système d'information. Assurer la sécurité du système d'information est une activité du management du système d'information. Aujourd'hui, la sécurité est un enjeu majeur pour les entreprises ainsi que pour l'ensemble des acteurs qui l'entourent. Elle n'est plus confinée uniquement au rôle de l'informaticien. Sa finalité sur le long terme est de maintenir la confiance des utilisateurs et des clients. La finalité sur le moyen terme est la cohérence de l'ensemble du système d'information. Sur le court terme, l'objectif est que

chacun ait accès aux informations dont il a besoin. La norme traitant des SMSI est l'ISO 27001 qui insiste sur Confidentiality – Integrity – Availability, c'est-à-dire en français Disponibilité – Intégrité - Confidentialité.

Les systèmes informatiques sont au cœur des systèmes d´information. Ils sont devenus la cible de ceux qui convoitent l'information. Assurer la sécurité de l'information implique d'assurer la sécurité des systèmes informatiques.

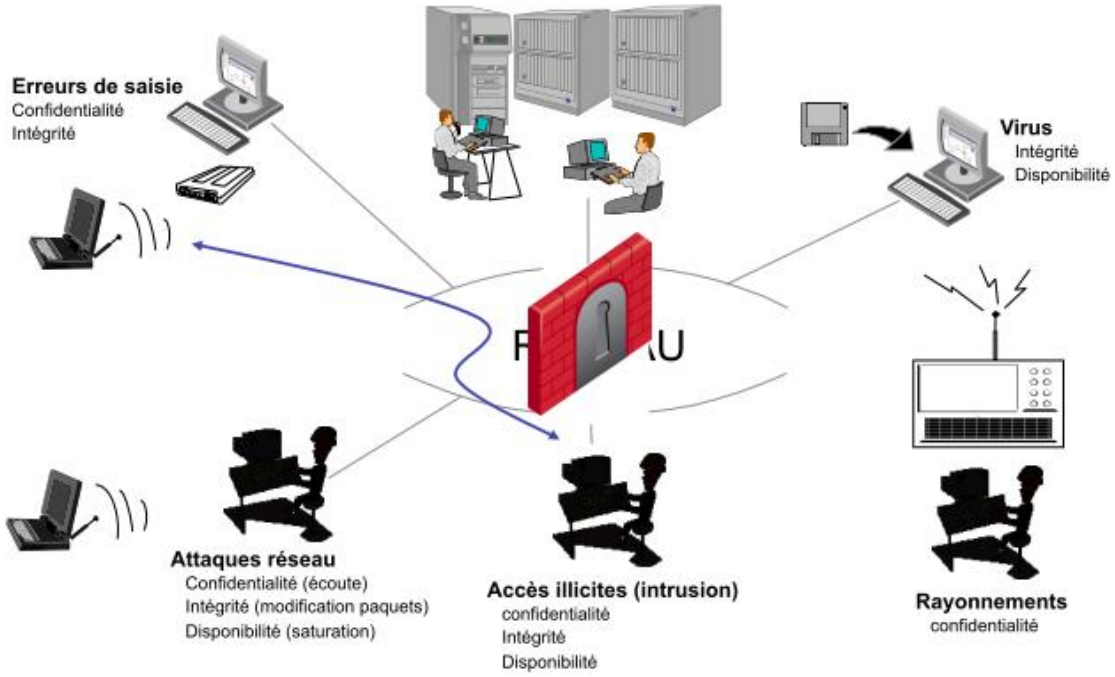


Figure 4 : Les attaques sur les systèmes informatiques

5.2 Les principales attaques [6]

Les attaques se divisent, selon leurs types sur quatre catégories :

5.2.1 Les attaques par programmes malveillants : Un logiciel malveillant (malware en anglais) est un logiciel développé dans le but de nuire un système informatique. Voici les principaux types de programmes malveillants :

- **Le virus :** Un virus est un code (programme) malveillant qui s'attache à un autre programme dans le but d'exécuter une fonction non souhaitée sur un ordinateur. Du fait qu'un virus se greffe à un programme légitime, son activation se produit grâce à l'utilisateur qui lance le programme infecté. Un virus peut rester inactif pendant un certain temps et ne s'activer qu'après une période de temps. Une fois activé, un virus

peut accéder au disque à la recherche de nouveaux exécutables à infecter. Un virus peut être inoffensif comme il peut exécuter des opérations dangereuse (effacement de fichiers, arrêts de services, …etc.) Initialement, les virus se répondaient par les disquettes et les modems. Actuellement ils se répondent par les disques amovibles, le réseau et le courrier électronique. Ce dernier est actuellement le moyen de propagation le plus répondu

- **Le ver (worm en anglais):** Un ver est un code autonome particulièrement dangereux qui se réplique (répondre) de façon indépendante en exploitant les vulnérabilités du réseau. Contrairement aux virus, les vers n'ont pas besoin de programme hôte pour se répliquer, ni de la participation de l'utilisateur pour s'activer. Ils peuvent se répondre rapidement sur le réseau. Ils sont responsables des attaques les plus destructrices sur Internet (comme SQL slammer en 2001).
- **Le wabbit :** Un wabbit est un type de logiciel malveillant qui s'auto réplique. Contrairement aux virus, il n'infecte pas les programmes ni les documents. Contrairement aux vers, il ne se propage pas par les réseaux. En plus de s'auto répliquer rapidement, les wabbits peuvent avoir d'autres effets malveillants. Un exemple de wabbit est la bombe fork, du nom de la commande Unix exploitée : fork.
- **Le cheval de Troie (trojan en anglais) :** C'est une application qui exécute autre chose que ce que son utilisateur pense exécuter. D'apparence écrite pour une fonction légitime, elle effectue des traitements cachés à l'insu de l'utilisateur. Les jeux sont les véhicules de cheveux de Troie les plus utilisés. Quand un jeu infecté est lancé, ce dernier fonctionne correctement, mais le cheval de Troie attaché est installé dans le système et continu a fonctionner même si le jeux est arrêté. Les chevaux de Troie peuvent causer des dommages immédiats, fournir un accès distant à la machine ou effectuer de l'espionnage (envoie de tous les caractères tapés par l'utilisateur pour retrouver des mots de passes)
- **La porte dérobée (backdoor en anglais) :** Dans un logiciel une porte dérobée (de l'anglais backdoor, littéralement porte de derrière) est une fonctionnalité inconnue de l'utilisateur légitime, qui donne un accès secret au logiciel. En sécurité informatique, la porte dérobée peut être considérée comme un type de cheval de Troie. Une porte dérobée peut être introduite soit par le développeur du logiciel, soit par un tiers, typiquement un pirate informatique. La personne connaissant la porte dérobée peut l'utiliser pour surveiller les activités du logiciel, voire en prendre le contrôle (par

exemple, par contournement de l'authentification). Enfin, selon l'étendue des droits que le système d'exploitation donne au logiciel contenant la porte dérobée, le contrôle peut s'étendre à l'ensemble des opérations de l'ordinateur.

- **Le logiciel espion (spyware en anglais) :** Un logiciel espion (espiogiciel, mouchard ou en anglais spyware) est un logiciel malveillant qui s'installe dans un ordinateur dans le but de collecter et transférer des informations sur l'environnement dans lequel il s'est installé, très souvent sans que l'utilisateur n'en ait connaissance. L'essor de ce type de logiciel est associé à celui d'Internet, qui lui sert de moyen de transmission de données.
- **Le keylogger (enregistreur de touches) :** Un enregistreur de frappe ou keylogger peut être assimilé à un matériel ou à un logiciel espion qui a la particularité d'enregistrer les touches frappées sur le clavier sous certaines conditions et de les transmettre via les réseaux. Par exemple, certains enregistreur de frappe analysent les sites visités et enregistrent les codes secrets et mots de passe lors de la saisie.
- **L'exploit :** Dans le domaine de la sécurité informatique, un exploit est un programme permettant à un individu d'exploiter une faille de sécurité informatique dans un système d'exploitation ou un logiciel que ce soit à distance (remote exploit) ou sur la machine sur laquelle cet exploit est exécuté (local exploit). Prononcer comme en anglais « explo-ï-te » et non « exploi », le mot provenant de exploitation (de faille informatique) et non pas du fait de réaliser un quelconque exploit extraordinaire.
- **Le rootkit :** ensemble de logiciels permettant généralement d'obtenir les droits d'administrateur sur une machine, d'installer une porte dérobée, de truquer les informations susceptibles de révéler la compromission des données compromises (altérées), et d'effacer les traces laissées par l'opération dans les journaux systèmes.

5.2.2 Les attaques par messagerie : En dehors des nombreux programmes malveillants qui se propagent par la messagerie électronique, il existe des attaques spécifiques à celles-ci :

- **Le pourriel (spam en anglais) :** un courrier électronique non sollicité, la plupart du temps de la publicité. Ils encombrent le réseau, et font perdre du temps à leurs destinataires.

- **L'Hameçonnage (phishing en anglais):** un courrier électronique dont l'expéditeur se fait généralement passer pour un organisme financier et demandant au destinataire de fournir des informations confidentielles.

- **Le Canular informatique (hoax en anglais) :** un courrier électronique incitant généralement le destinataire à retransmettre le message à un de ses contacts sous divers prétextes. Ils encombrent le réseau, et font perdre du temps à leurs destinataires.

Dans certains cas, ils incitent l'utilisateur à effectuer des manipulations dangereuses sur son poste (suppression d'un fichier prétendument lié à un virus par exemple).

5.2.3 Les attaques sur le réseau :

- **Intrusion :** l'intrusion dans un système informatique a généralement pour but la réalisation d'une menace et est donc une attaque. Les conséquences peuvent être catastrophiques : vol, fraude, incident diplomatique, chantage…

  Le principal moyen pour prévenir les intrusions est le pare-feu ('firewall'). Il est efficace contre les fréquentes attaques de pirates amateurs, mais d'une efficacité toute relative contre des pirates expérimentés et bien informés. Une politique de gestion efficace des accès, des mots de passe et l'étude des fichiers « log »(traces) est complémentaire.

- **Ecoute du réseau (sniffing):** il existe des logiciels qui, permettent d'intercepter certaines informations qui transitent sur un réseau local, en retranscrivant les trames dans un format plus lisible (Network packet sniffing). C'est l'une des raisons qui font que la topologie en étoile autour d'un hub n'est pas la plus sécurisé, puisque les trames qui sont émises en « broadcast » sur le réseau local peuvent être interceptées.

  De plus, l'utilisateur n'a aucun moyen de savoir qu'un pirate à mis son réseau en écoute. L'utilisation de switchs (commutateurs) réduit les possibilités d'écoute mais en inondant le commutateur, celui-ci peut se mettre en mode « HUB » par sécurité.

- **Le déni de service (DOS):** l'attaquant n'obtient pas un accès au système informatique sur le réseau mais il parvient à mettre en panne certains composants stratégiques (le serveur de messagerie, le site web, etc). Le but d'une telle attaque n'est pas de dérober des informations sur une machine distante, mais de paralyser un service ou un réseau complet. Les utilisateurs ne peuvent plus alors accéder aux

ressources. Les deux exemples principaux, sont le « Ping flood » ou l'envoie massif de courriers électroniques pour saturer une boite aux lettres (mailbombing). La meilleure parade est firewall ou la répartition des serveurs sur un réseau sécurisé.

- **Le déni de service distribué (DDOS) :** Une attaque de dénie de service distribuée (DDOS) est une attaque de dénie de service produite par plusieurs sources coordonnées entre elles. Plusieurs techniques (exemples) : Ping of death, L'attaque Smurf, TCP SYN Flood

- **IP Spoofing :** Usurpation d'adresse IP, on fait croire que la requête provient d'une machine autorisée. Une bonne configuration du routeur d'entrée permet d'éviter qu'une machine extérieure puisse se faire pour une machine interne.

- **DNS Spoofing :** pousse un serveur DNS à accepter l'intrus. Pour l'éviter, il est intéressant de séparer le DNS du LAN de celui de l'espace publique.

- **Man-in-the-middle :** L'intrus est positionné au milieu d'une communication entre deux parties dans le but de l'espionner ou de la modifier. Un intrus peut fournir des services de point d'accès wifi ou de passerelle réseau pour centraliser les communications à son niveau. La victime peut se connecter et utiliser normalement le réseau sans savoir que le flux passe par une tiers personne.

5.2.4 Les attaques sur les mots de passe : peuvent consister à faire de nombreux essais jusqu'à trouver le bon mot de passe.

- **L'attaque par dictionnaire :** le mot testé est pris dans une liste prédéfinie contenant le mot de passe les plus courants et aussi des variantes de ceux-ci (à l'envers, avec un chiffre à la fin, etc). ces listes sont généralement dans toutes les langues les plus utilisées, contiennent des mots existants, ou des diminutifs (comme par exemple 'powa ' pour 'power', ou 'G0d' pour 'god').

- **L'attaque par force brute:** toutes les possibilités sont faites dans l'ordre jusqu'à trouver la bonne solution (par exemple de 'aaaaaa' jusqu'à 'ZZZZZZ' pour un mot de passe composé de six caractères alphabétiques).

  Exemple : L0phtCrack, est une application de force brute pour retrouver les mots de passes Windows server. Décryptage d'une clef wep sous windows

5.3 Les critères de la sécurité [7]

Pour une bonne sécurisation des données, on devrait déterminer les mesures de protection pour chaque scénario bien /menace/incidence. Un service un mécanisme ou une procédure peut constituer une mesure de protection si elle empêche ou réduit la probabilité que ne soient exploités les points vulnérables d'un réseau.

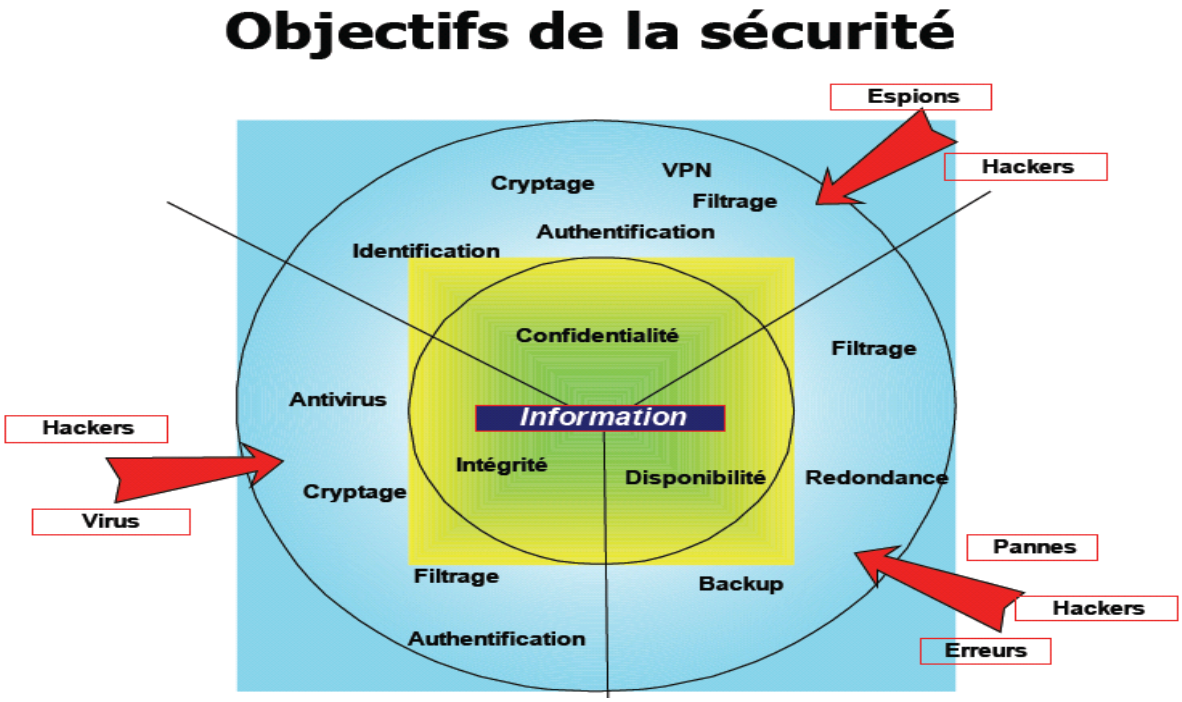


Figure 5 : Objectif de la sécurité

Les mesures de protection sont habituellement effective dans aux moins dans l'un des domaines suivants :

5.3.1 L'authentification

S'assurer que l'origine du message soit correctement identifiée

- ✓ Assurer le receveur que le message émane de la source qui prétend avoir envoyé ce message.
- ✓ Assurer l'authenticité des entités participantes: chacune des entités est celle qui prétende l'être
- ✓ Empêcher la perturbation de la connexion par une tierce partie qui se fait passer pour une entité légitime (émission ou réception non autorisée).

Mécanismes utilisés: Cryptage, signature numérique (mécanisme permettant d'authentifier l'auteur d'un document électronique et de garantir son intégrité en informatique c'est une clé de chiffrement et de déchiffrement d'un document électronique), Notarisation (permet la vérification et l'archivage des preuves d'échanges et d'archivage électroniques par un tiers de confiance agréé. Cette technique améliore la sécurité des échanges et de l'archivage électronique dans la mesure où elle assure le suivi et l'archivage des transactions émises et reçues (l'intégrité, l'origine, la date et la destination des données).

5.3.2 Contrôle d'accès

Empêcher l'utilisation non autorisée d'une ressource (serveur, application, etc.)

- ✓ Définir qui a le droit d'accéder aux ressources
- ✓ Déterminer sous quelles conditions ceci peut avoir lieu ?
- ✓ Défini ce qu'une entité est autorisée de faire lors de l'accès à une ressource
- ✓ Signaler les failles de sécurité relatives à l'OS

Mécanismes utilisés: Authentification, signature numérique, pare-feu, mécanismes propres aux OS.

5.3.3 Confidentialité

- ✓ Protection de l'information transmise contre les attaques passives, et protection des flux de données contre l'analyse.
- ✓ Préservation du secret des données transmises. Seulement les entités communicantes sont capables d'observer les données.

Il existe plusieurs niveaux de confidentialité :

- ✓ Protection de toutes les données échangées tout au long d'une connexion.
- ✓ Protection des données contenues au niveau d'un seul bloc de donnée.
- ✓ Protection de quelques champs des données échangées (pour une connexion ou un seul bloc de donnée)
- ✓ Protection de l'information (source, destination, etc.) qui peut être déduite à partir de l'observation des flux de données échangés.

Mécanisme utilisé: Cryptage

5.3.4 Intégrité

L'intégrité permet de certifier que les données, les traitements ou les services n'ont pas été modifiés, altérés ou détruits tant de façon intentionnelle qu'accidentelle L'altération est principalement occasionnée par le média de transmission mais peut provenir du système d'informations Il faut également veiller à garantir la protection des données d'une écoutes actives sur le réseau

Détecter si les données ont été modifiées depuis la source vers la destination

- ✓ Service orienté connexion: Protection contre la duplication, la destruction, l'insertion, la modification, le rejeu, le reclassement, etc.
- ✓ Service non orienté connexion: Protection contre la modification uniquement.

Mécanismes utilisés: cryptage, signature numérique, contrôle d'accès, contrôle d'intégrité

5.3.5 La disponibilité

- ✓ la propriété qu'un système ou une ressource du système soit accessible et utilisable suite à la demande d'une entité autorisée.
- ✓ protège un système pour assurer sa disponibilité particulièrement contre des attaques de dénie de service
- ✓ Dépend d'autres services comme le contrôle d'accès, l'authentification, …etc.

Mécanismes utilisés : Filtrage (pare-feu), antivirus, contrôle d'accès

5.4 Les techniques de sécurisation des données [8]

Sécuriser des données, consiste à les rendre incompréhensibles aux intrus qui veulent s'emparer des informations que vous jugez secrètes ou bien confidentielles. Pour cela différentes techniques de sécurisation ont été mises en œuvre pour atteindre ce but

5.4.1 L'authentification

L'authentification est la procédure qui consiste, pour un système informatique, à vérifier l'identité d'une entité (personne, ordinateur….), afin d'autoriser l'accès de cette entité à des ressources (systèmes, réseaux, applications…).

Plusieurs solutions simples sont mise en œuvre pour cela, comme l'utilisation d'un identifiant (login) et d'un mot de passe (password). L'authentification peut s'effectuer par un numéro d'identification personnel, comme le numéro inscrit dans une carte a puce, ou code PIN (Personal Identification Number).

Des techniques beaucoup plus sophistiqués, comme les empreintes digitales ou rétiniennes (Biométrie), les authentifieurs (clé USB,… etc) se développent de façon industrielle au début des années 2000. Cependant, leur utilisation est assez complexe et ne peut être mise en place que dans un contexte particulier, comme un centre de recherche de l'armé par exemple.

5.4.2 Firewalls et Anti virus

Un pare-feu est un logiciel (ou pièce électronique) chargé d'empêcher l'accès non autorisé à un système informatique. Les gens responsables de ces tentatives, appelés « Pirates », exploitent des failles de sécurité d'un système pour y accéder. Il existe deux types de pare-feu :

- **Un pare-feu matériel** est souvent un boîtier extérieur à l'ordinateur, qui sert de diviseur de connexion Internet et de mini-réseau à domicile.
- **Un pare-feu logiciel** Protection plus avancée, cloisonnant le système des intrusions/sorties non-autorisées sur lequel il est installé et les réseaux (internet...).

D'une certaine manière, les firewalls matériels sont plus efficaces car ils ne s'occupent que de protéger, tandis que les firewalls logiciels sont contrôlés par un ordinateur occupé par autre chose.

Un antivirus est un logiciel permettant de détecter les virus informatiques qui peuvent arriver sur nos ordinateurs.

5.4.3 Protection contre les attaques virales

- Utilisation d'au moins un antivirus, remis à jour très régulièrement et exécuté régulièrement.
- Utilisation d'un pare-feu pour filtrer ce qui vient d'extranet.
- Surveiller les activités système et réseau pour détecter tout comportement anormal
- En cas d'atteinte, isoler les équipements infectés et tenter des procédures de reprise
- Mettre à jour les systèmes dés qu'une faille de sécurité est découverte

5.4.4 Protection contre les attaques de reconnaissance

- Utiliser des procédures de cryptage (VPN par exemple)
- Utiliser des techniques d'isolement physique (commutateur au lieu du concentrateur
- Segmenter le réseau pour le décomposé en plusieurs domaines de diffusion
- Utiliser des logiciels et du matériel de détection d'attaques de reconnaissance (Antisniffer)
- Utiliser un pare-feu pour filtrer les paquets (exemple bloquer tous les paquets ICMP)

5.4.5 Protection contre les attaques d'accès

- Utiliser des mots de passes forts

    - ✓ 8 caractères au minimum.
    - ✓ Inclure des majuscules et des minuscules.
    - ✓ Inclure des chiffres.
    - ✓ Inclure des caractères spéciaux (?, !, @, &…).

- Désactiver un compte pendant une période après un nombre de tentatives de connexions échouées (temps d'attente monte graduellement).
- Imposer un seuil de protection minimal aux équipements de confiance (sécurisation des ports).
- Utiliser le cryptage

5.4.6 Protection contre les attaques de Dénis de Service

- Utilisation des équipements spécialisés permettant de :

- ✓ Examiner la taille des paquets.
- ✓ Compter le nombre de TCP SYN arrivant.
- ✓ Le contrôle de flux, …etc

- ➢ Mise à jour des systèmes.
- ➢ Utilisation des techniques de gestion de la QoS (Détection de la baisse de la QoS).

6. Application de cryptage pour la sécurité informatique [9]

6.1 Cryptage et décryptage

Les données lisibles et compréhensibles sans intervention spécifique sont considérées comme du texte en clair. La méthode permettant de dissimuler du texte en clair en masquant son contenu est appelée le cryptage. Le cryptage consiste à transformer un texte normal en charabia inintelligible appelé texte chiffré. Cette opération permet de s'assurer que seules les personnes auxquelles les informations sont destinées pourront y accéder. Le processus inverse de transformation du texte chiffré vers le texte d'origine est appelé le décryptage.

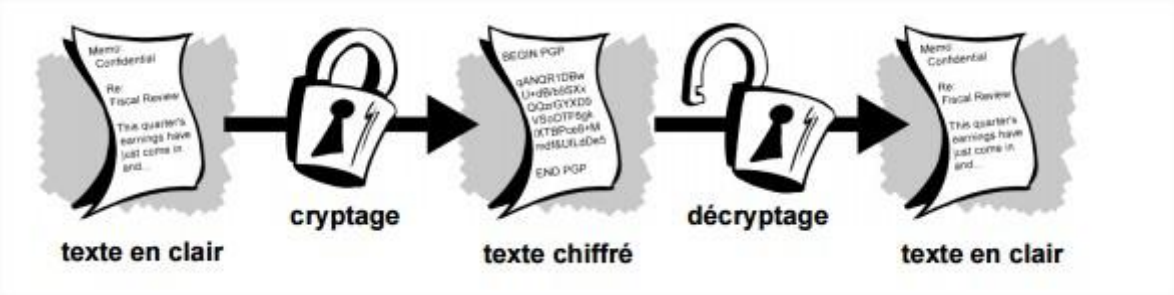


Figure 6 : Cryptage et décryptage

6.2 Définition de la cryptographie

La cryptographie est la science qui utilise les mathématiques pour le cryptage et le décryptage de données. Elle vous permet ainsi de stocker des informations confidentielles ou de les transmettre sur des réseaux non sécurisés (tels que l'Internet), afin qu'aucune personne autre que le destinataire ne puisse les lire. Alors que la cryptographie consiste à sécuriser les données, la cryptanalyse est l'étude des informations cryptées, afin d'en découvrir le secret. La cryptanalyse classique implique une combinaison intéressante de raisonnement analytique, d'application d'outils mathématiques, de recherche de modèle, de patience, de détermination et de chance. Ces cryptanalyses sont également appelés des pirates. La cryptologie englobe la cryptographie et la cryptanalyse.

6.3 Mécanismes de la cryptographie

Un algorithme de cryptographie ou un chiffrement est une fonction mathématique utilisée lors du processus de cryptage et de décryptage. Cet algorithme est associé à une clé (un mot, un nombre ou une phrase), afin de crypter le texte en clair. Avec des clés différentes, le résultat du cryptage variera également. La sécurité des données cryptées repose entièrement sur deux éléments : l'invulnérabilité de l'algorithme de cryptographie et la confidentialité de la clé. Un système de cryptographie est constitué d'un algorithme de cryptographie, ainsi que de toutes les clés et tous les protocoles nécessaires à son fonctionnement. PGP est un système de cryptographie.

6.4 Définition d'un Crypto-système

Un crypto-système est l'ensemble des deux méthodes de chiffrement et de déchiffrement utilisable en sécurité. Définit par les propriétés suivantes :

- Réalisation simple et rapide du chiffrement et du déchiffrement (pour atteindre des débits élevés).
- Éviter un encombrement important des clés.
- Une méthode de cryptographie (fonctions E et D) doit être stable. On ne peut la changer que très rarement.
- Elle est le plus souvent publiée (largement connue).
- Un crypto-système dépend de paramètres (clés) qui doivent pouvoir être modifiés aisément et fréquemment.
- On estime que la sécurité ne doit pas dépendre du secret des algorithmes E et D mais uniquement du secret des clés k et k' (exception pour le domaine militaire).

7. Conclusion

Le sécurité informatique sera effective dans la mesure ou l'on sait mettre en place des mesures de protection homogènes et complémentaires des ressources informatiques et de télécommunication, mais aussi de l'environnement qui les héberge. Aux aspects purement techniques de la sécurité, il faut associer la mise en œuvre efficace des procédures d'exploitation et de gestion. Par ailleurs, le gérant de l'organisme doit être formé aux mesures de sécurité et doit s'engager à les respecter.

Chapitre II
Fouille de données et Détection d'intrusion

1. Introduction

Selon R.L. Grossman dans le "Data Mining : Challenges and Opportunities for Data Mining During the Next Decade", il définit le data mining comme "concerné à découvrir des configurations, des associations, des changements, des anomalies, et statistiquement des structures et des événements significatifs dans les données". Simplement il offre la capacité de prendre des données et de tirer d'elle les configurations ou les déviations qui ne peuvent être vues facilement à l'œil nu. Un autre terme parfois utilisé est knowledge discovery.

Ils existent beaucoup de types différents d'algorithmes d'extraction de données pour inclure l'analyse de lien, le clustering, l'association, l'abduction de règle, l'analyse de déviation, et l'analyse de séquence.

Afin que nous puissions déterminer comment le data mining peut aider la détection anticipée d'intrusion il est important de comprendre comment les IDS actuels travaillent pour identifier une intrusion, et c'est ce que nous allons découvrir dans ce chapitre.

2. Data Mining

2.1 Introduction

Le Data Mining que l'on peut traduire par "fouille de données" apparaît au milieu des années 1990 aux États-Unis comme une nouvelle discipline à l'interface de la statistique et des technologies de l'information : bases de données, intelligence artificielle, apprentissage automatique. C'est un domaine pluridisciplinaire permettant, à partir d'une très importante quantité de données brutes, d'en extraire de façon automatique ou semi-automatique des informations cachées, pertinentes et inconnues auparavant en vue d'une utilisation industrielle ou opérationnelle de ce savoir.

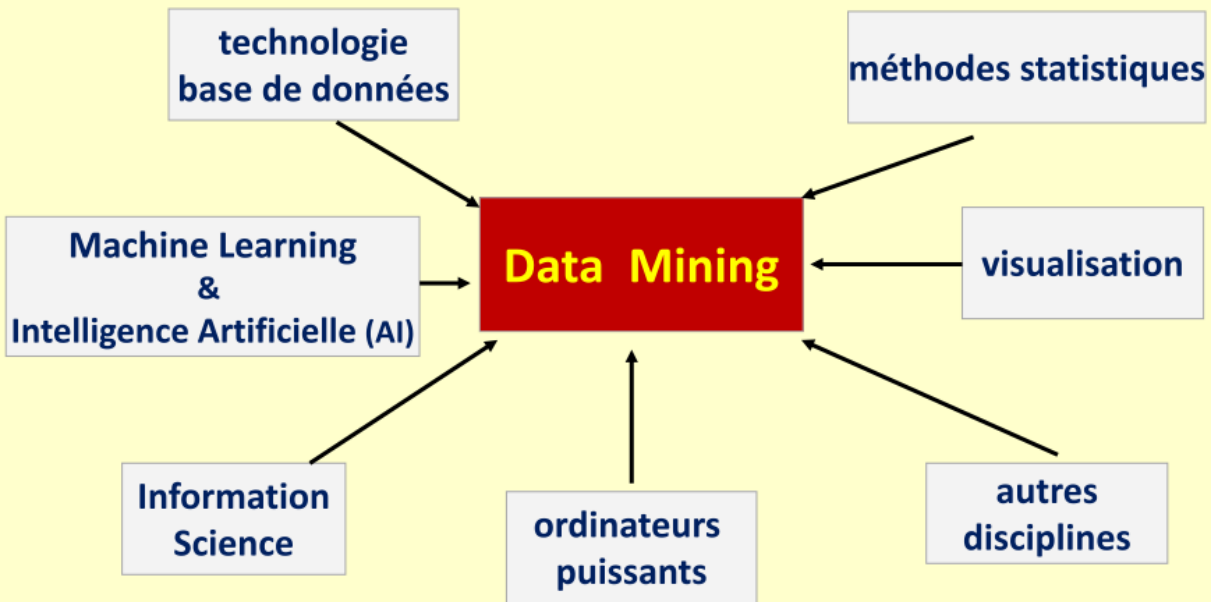


Figure 1 : fouille de données.

2.2 Définition [10]

La Fouille de données (Data Mining) est un Ensemble de techniques d'exploration de données permettant d'extraire d'une base de données des connaissances sous la forme de modèles de description afin de :

- Décrire le comportement actuel des données et/ou
- Prédire le comportement futur des données

Extraction de connaissance à partir de données (Knowledge Discovery in Databases – KDD) :

- Cycle de découverte d'information regroupant la conception de grandes bases de données ou entrepôts de données (Data Warehouse)
- Tous les traitements à effectuer pour extraire de l'information des données
- L'un de ces traitements est la Fouille de données (Data Mining)

Le Data Warehouse esr une base de données d'aide à la décision qui est entretenue de manière séparée de la base de données opérationnelle de l'organisation. Data warehousing: Le processus de construction et d'utilisation du data warehouse .

2.3 Découverte de connaissances dans les bases de donnés (knowledge Discovery in Databases = KDD)

2.3.1 Définition [11]

Le KDD est un processus inductif, itératif et interactif d'identification d'une structure de données valide, nouvelle, potentiellement utile et finalement compréhensible.

- Itératif: nécessite plusieurs passes
- Interactif: l'utilisateur est dans la boucle du processus
- Valide: valable dans le futur
- Nouvelle: non prévisible
- Utile: permet à l'utilisateur de prendre des décisions
- Compréhensible: présentation simple

Ce processus utilise le raisonnement inductif.

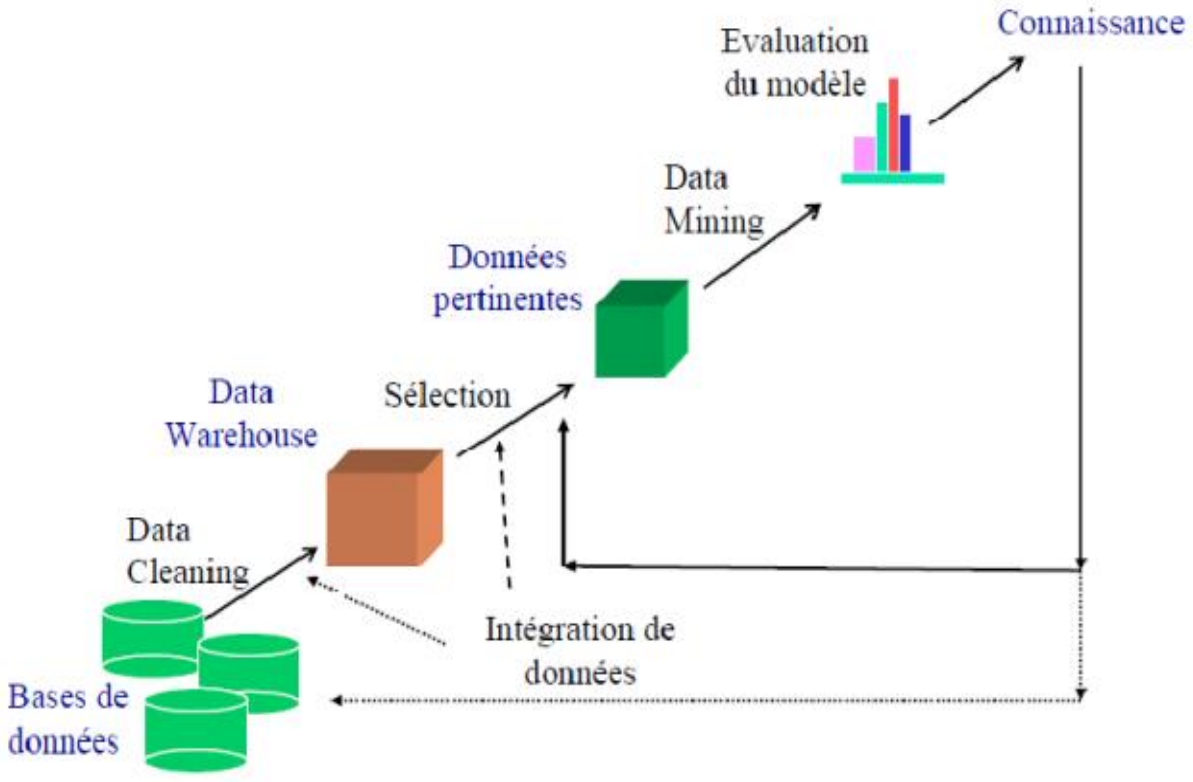


Figure 2: Le processus de découverte de connaissance (KDD)

2.3.2 Démarche méthodologique du KDD [12]

- Comprendre l'application
  - Connaissances à priori, objectifs, etc.
- Sélectionner un échantillon de données
  - Choisir une méthode d'échantillonnage
- Nettoyage et transformation des données
  - Supprimer le «bruit» : données superflues, marginales, données manquantes, etc.
  - Effectuer une sélection d'attributs, réduire la dimension du problème, etc.
- Appliquer les techniques de fouille de données
  - Choisir le bon algorithme
- Visualiser, évaluer et interpréter les modèles découverts
  - Analyser la connaissance (intérêt)
  - Vérifier sa validité (sur le reste de la base de données)
  - Réitérer le processus si nécessaire
- Gérer la connaissance découverte
- La mettre à la disposition des décideurs
- etc.

2.4 Architecture d'un système type de Data Mining

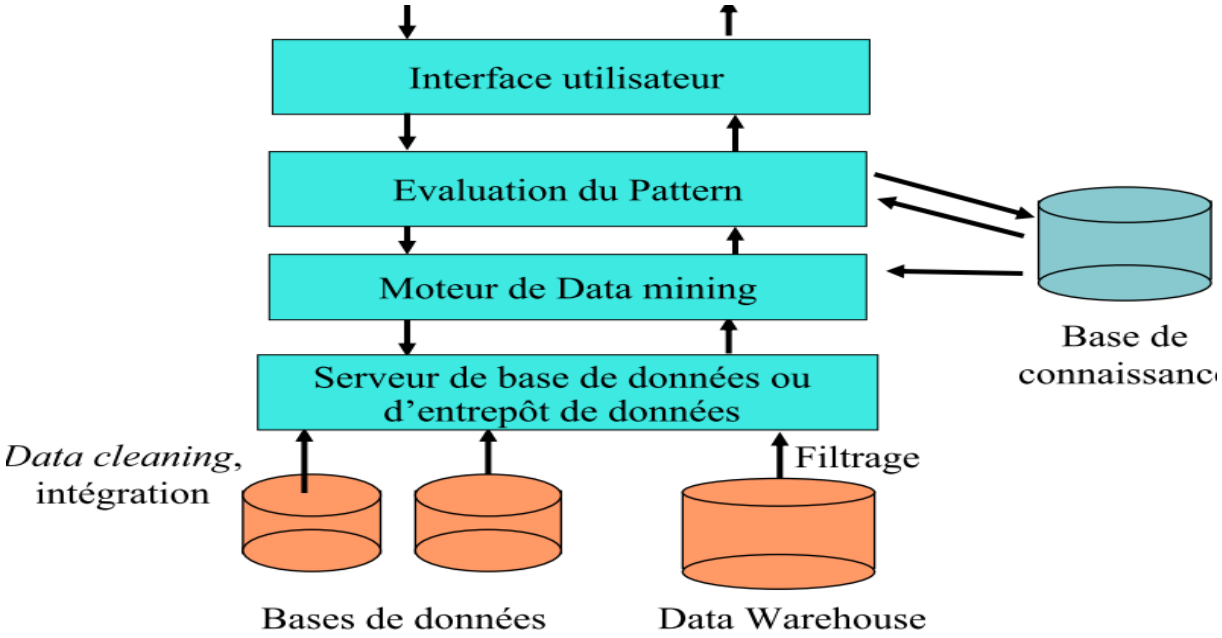


Figure II.3 : Architecture d'un système type de Dta Mining.

2.5 Techniques utilisées dans le Data Mining :[13]

Les techniques les plus communément utilisées dans le data mining sont :

2.5.1 L'Analyse du panier de la ménagère

L'analyse du panier de la ménagère est un moyen de trouver les groupes d'articles qui vont ensembles lors d'une transaction. C'est une technique de découverte de connaissances non dirigée (de type analyse de clusters) qui génère des règles et supporte l'analyse des séries temporelles (si les transactions ne sont pas anonymes). Les règles générées sont simples, faciles à comprendre et assorties d'une probabilité, ce qui en fait un outil agréable et directement exploitable par l'utilisateur métier.

2.5.2 Les arbres de décision

Les arbres de décision sont utilisés dans le cadre de la découverte de connaissances dirigée. Ce sont des outils très puissants principalement utilisés pour la classification, la description ou l'estimation. Le principe de fonctionnement est le suivant : pour expliquer une variable, le système recherche le critère le plus déterminant et découpe la population en sous populations possédant la même entité de ce critère. Chaque sous population est ensuite analysée comme la population initiale. Le modèle rendu est facile à comprendre et les règles trouvées sont très explicites. Ce système est donc très apprécié.

2.5.3 Le raisonnement basé sur la mémoire

Le raisonnement basé sur la mémoire (RBM) est une technique de prédiction et de classification utilisée dans le cadre de la découverte de connaissances dirigée. Elle peut être également utilisée pour l'estimation. Pour chaque nouvelle instance présentée, le système recherche le(s) voisin(s) le(s) plus proche(s) et procède ainsi à l'affectation ou estimation. L'avantage du RBM est qu'il est facile à mettre en œuvre, très stable (les nouvelles données n'entraînent pas de refaire fonctionner un système de calcul) et supporte tout type de données.

2.5.4 La détection automatique des clusters

La détection automatique de clusters est une technique de découverte de connaissances non dirigée (ou apprentissage sans supervision). Elle consiste à regrouper les enregistrements en fonction de leurs similitudes. Chaque groupe représente un cluster. C'est une excellente technique pour démarrer un projet d'analyse ou de data mining. Les groupes de similitudes permettront de mieux comprendre les données et d'imaginer comment les utiliser au mieux.

2.5.5 L'analyse des liens

L'analyse des liens est une technique de description qui s'inspire et repose sur la théorie des graphes. Elle consiste à relier des entités entre elles (clients, entreprises, . . . etc) par des liens. A chaque lien est affecté un poids, défini par l'analyse, qui quantifie la force de cette relation. Cette technique peut être utilisée pour la prédiction ou la classification mais généralement une simple observation du graphe permet de mener à bien l'analyse.

2.5.6 Les algorithmes génétiques

Les algorithmes génétiques sont utilisés dans la découverte de connaissances dirigée. Ils permettent de résoudre des problèmes divers, notamment d'optimisation, d'affectation ou de prédiction. Leur fonctionnement s'apparente à celui du génome humain. Le principe de fonctionnement est le suivant : les données sont converties en chaînes binaires (comme les chaînes d'ADN. Celles-ci se combinent par sélection, croisement ou mutation et donnent ainsi une nouvelle chaîne qui est évaluée. En fonction du résultat, les chaînes les plus faibles cèdent leur place aux plus fortes.

Cette technique est particulièrement intéressante pour résoudre des problèmes d'affectation ou des problèmes sur lesquels on peut poser une fonction d'évaluation car elle peut trouver des solutions optimisées parfois inexistantes dans les données d'origine.

2.5.7 Les réseaux de neurones

Les réseaux de neurones représentent la technique de Data Mining la plus utilisée. Pour certains utilisateurs, elle en est même synonyme. C'est une transposition simplifiée des neurones du cerveau humain. Dans leur variante la plus courante, les réseaux de neurones apprennent sur une population d'origine puis sont capables d'exprimer des résultats sur des données inconnues. Ils sont utilisés dans la prédiction et la classification dans le cadre de découverte de connaissances dirigée. Certaines variantes permettent l'exploration des séries temporelles et des analyses non dirigées (réseaux de Kohonen). Le champ d'application est très vaste et l'offre logicielle importante.

Cependant, on leur reproche souvent d'être une "boite noire" : il est difficile de savoir comment les résultats sont produits, ce qui rend les explications délicates, même si les résultats sont bons.

2.5.8 Les agents intelligents ou Knowbot

Les agents intelligents ou Knowbot sont des entités logicielles autonomes dont les plus récentes versions s'intègrent tout à fait dans le processus de Data Mining. Certains iront jusqu'à les considérer comme des outils de Data Mining. Certains d'entre eux, les plus élaborés, sont capables de suivre et mémoriser les mouvements, visites et achats sur Internet et permettent d'élaborer des profils d'utilisateurs pour leur faire des offres commerciales "un à un (one to one) ". L'utilisateur peut, quant à lui, lancer des appels d'offres et mises en concurrence automatiquement gérés par ces agents.

2.5.9 Le traitement analytique en ligne (TAEL)

Pour terminer ce tour d'horizon, nous évoquerons ici le TAEL (traitement analytique en ligne) car bien que ne faisant pas partie du data mining, il s'agit d'outils d'analyse de données souvent utiles en préalable au data mining. Le TAEL est une manière de présenter aux utilisateurs les données relationnelles afin de faciliter la compréhension des données et des formes importantes qu'elles recèlent. Ces outils s'appuient sur OLAP (On Line Analytical Process), MOLAP (Multidimensional OLAP), et ROLAP (Relational OLAP).

2.6 Objectif du Data Mining

Le Data Mining (DM) ou fouille de données est l'ensemble des techniques qui permettent de transformer les données en connaissances. Son but est de remplir l'une des tâches suivantes :

- La Classification
- Le Clustering (Segmentation)
- La Recherche d'associations
- La Recherche de séquences
- La Détection de déviation

2.6.1 Clustering

Partitionnement logique de la base de données en clusters

- Clusters : groupes d'instances ayant des caractéristiques communes
- Apprentissage non supervisé (classes inconnues)
- Pb : interprétation des clusters identifiés
- Applications: Economie (segmentation de marchés), médecine (localisation de tumeurs dans le cerveau), etc.

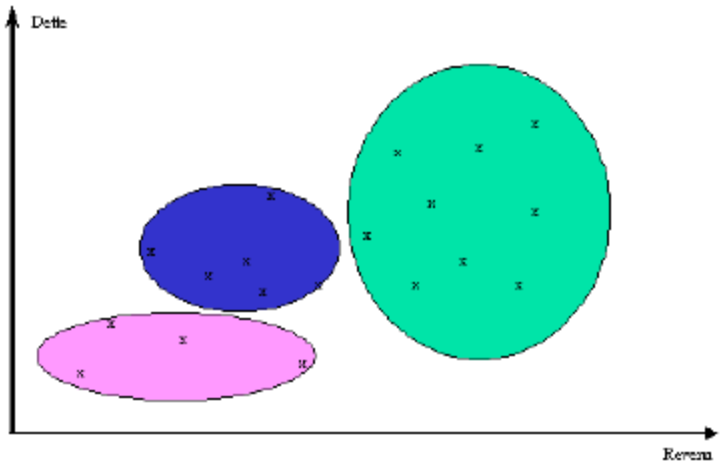


Figure 4: Clustering

2.6.2 Classification

Cette tâche permet de prédire si une instance de donnée est membre d'un groupe ou d'une classe prédéfinie.

- Classes : Groupes d'instances avec des profils particuliers
- Apprentissage supervisé: les classes sont connues à l'avance
- Applications: marketing direct (profils des consommateurs), grande distribution (classement des clients), médecine (malades/non malades), etc.

❖ Différence entre Classification et Clustring : le clustring permet de regrouper des éléments selon leur ressemblance, contrairement à la classification on n'a pas besoin de connaissances sur les classes dans lesquelles on regroupe les éléments.

2.6.3 Règles d'association

Les règles d'association ont été crées pour répondre à la problématique du panier de la ménagère. Étant donnée une base de transactions (les paniers), chacune composée d'items (les produit achetés), la découverte de règles d'association consiste à chercher des ensembles d'items fréquemment liés dans une même transaction. La recherche de règles d'associations s'avère donc très utile pour la découverte de relations.

- Exemple
  - BD commerciale : panier de la ménagère
  - Articles figurant dans le même ticket de caisse
  - Ex: achat de riz + jus ==> achat de poisson

2.6.4 Recherche de séquences

- Liaisons entre événements sur une période de temps
- Extension des règles d'association
  - Prise en compte du temps (série temporelle)
- Applications: marketing direct (anticipation des commandes), bioinformatique (séquences d'ADN), bourse (prédiction des valeurs des actions)
- Exemple
  - BD commerciale (ventes par correspondance)
  - Commandes de clients

2.6.5 Détection de déviation

Instances ayant des caractéristiques les plus différentes des autres

- Basée sur la notion de distance entre instances
- Expression du problème
- Temporelle : évolution des instances ?
- Spatiale : caractéristique d'un cluster d'instances ?

3. Détection d'intrusion

3.1 Introduction

Avec le développement des réseaux de communication, l'Internet est devenu l'infrastructure critique pour une société moderne. La croissance explosive d'utilisateurs d'Internet a motivé l'expansion rapide de commerce électronique et d'autres services en ligne. Malheureusement derrière la convenance et l'efficacité de ces services, les risques et les chances d'intrusions malveillantes sont aussi augmentés. La sécurité des systèmes informatiques est devenue un défi majeur dont l'objectif est d'assurer la disponibilité des services, la confidentialité et l'intégrité des données et des échanges.

3.2 Définition

- **Une attaque** : découverte systématique d'informations du réseau par des scans de port et balayage du réseau, tentative réelle d'intrusion dans un réseau [14]
- **Intrusion**: nous appellerons intrusion toute utilisation d'un système informatique à des fins autres que celles prévues, généralement dues à l'acquisition de privilèges de façon illégitime. L'intrus est généralement vu comme une personne étrangère au système informatique qui a réussi à en prendre le contrôle, mais les statistiques montrent que les

utilisations abusives (du détournement de ressources à l'espionnage industriel) proviennent le plus fréquemment de personnes internes ayant déjà un accès au système.[15]

- **Détection d'attaques** : afin de détecter les attaques que peut subir un système, il est nécessaire de disposer d'un logiciel spécialisé dont le rôle sera de surveiller les données qui transitent sur ce système et qui serait capable de réagir si des données semblent suspectes.

Les logiciels qui sont les plus à même d'effectuer cette tâche sont les systèmes de détection d'intrusion : les IDS.

- **Un système de détection d'intrusion (ou IDS : Intrusion Detection System) est** Un système de détection d'intrusions à un niveau très macroscopique peut être décrit comme un détecteur. Ce détecteur est un moteur d'analyse qui reçoit des données de trois sortes de ressources. L'analyse de ces données génère une décision d'évaluation de la probabilité que ces actions peuvent être considérées comme des symptômes d'intrusions. Ces données sont :
  - Des informations de configuration relatives à l'état actuel du système.
  - Des informations à long terme relatives à la technique utilisée pour détecter les intrusions par exemple une base de connaissances d'attaques.
  - Des informations venant du système à protéger qui sont les informations d'audit décrivant les événements qui apparaissent dans le système.

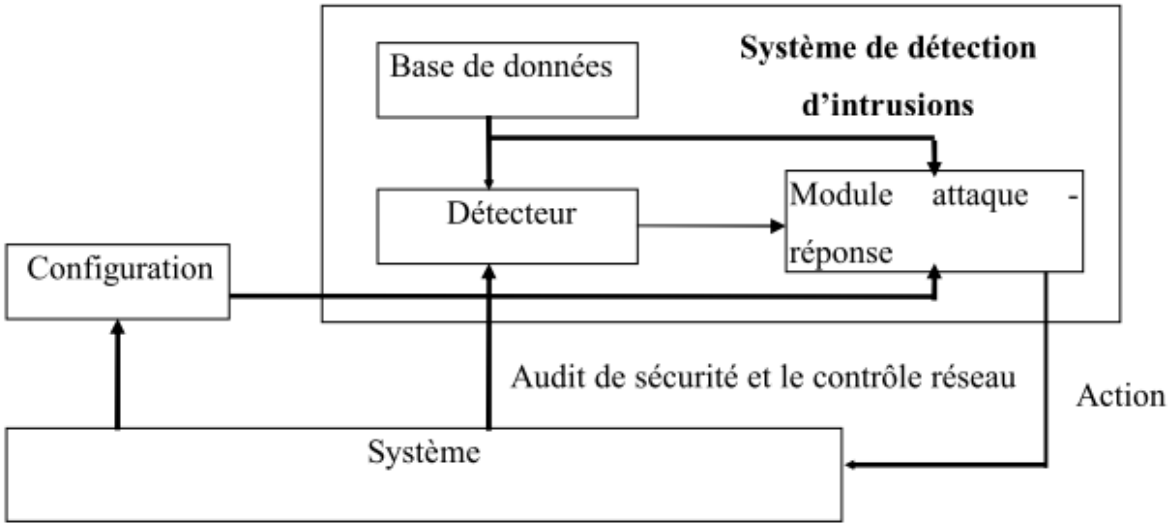


Figure 5: Description d'un système de détection d'intrusions.

3.3 Types des IDS

3.3.1 IDS réseaux

Le rôle essentiel d'un IDS réseau (NIDS) est l'analyse et l'interprétation des paquets circulant sur ce réseau. **[16]**

L'implantation d'un NIDS sur un réseau se fait de la façon suivante: des capteurs sont placés aux endroits stratégiques du réseau et génèrent des alertes s'ils détectent une attaque. Ces alertes sont envoyées à une console sécurisée, qui les analyse et les traitent éventuellement. Cette console est généralement située sur un réseau isolé, qui relie uniquement les capteurs et la console.

Les capteurs: Les capteurs placés sur le réseau sont placés en mode furtif (ou stealth mode), de façon à être invisibles aux autres machines. Pour cela, leur carte réseau est configurée en mode "promiscuous", c'est à dire le mode dans lequel la carte réseau lit l'ensemble du trafic, de plus aucune adresse IP n'est configurée. Un capteur possède en général deux cartes réseaux, une placée en mode furtif sur le réseau, l'autre permettant de le connecter à la console de sécurité. Du fait de leur invisibilité sur le réseau, il est beaucoup plus difficile de les attaquer et de savoir qu'un IDS est utilisé sur ce réseau.

Placer les capteurs : Il est possible de placer les capteurs à différents endroits, en fonction de ce que l'on souhaite observer. Les capteurs peuvent être placés avant ou après le pare-feu, ou encore dans une zone sensible que l'on veut protéger spécialement.

Si les capteurs se trouvent après un pare-feu, il leur est plus facile de dire si le parefeu a été mal configuré ou de savoir si une attaque est venue par ce pare-feu. Les capteurs placés derrière un pare-feu ont pour mission de détecter les intrusions qui n'ont pas été arrêtées par ce dernier. Il s'agit d'une utilisation courante d'un NIDS. Il est également possible de placer un capteur à l'extérieur du pare-feu (avant le firewall). L'intérêt de cette position est que le capteur peut ainsi recevoir et analyser l'ensemble du trafic d'Internet. Si vous placez le capteur ici, il n'est pas certain que toutes les attaques soient filtrées et détectées. Pourtant, cet emplacement est le préféré de nombreux experts parce qu'il offre l'avantage d'écrire dans les logs et d'analyser les attaques (vers le pare-feu...), ainsi l'administrateur voit ce qu'il doit modifier dans la configuration du pare-feu.

Les capteurs placés à l'extérieur du pare-feu servent à détecter toutes les attaques en direction du réseau, leur tâche ici est donc plus de contrôler le fonctionnement et la configuration du firewall que d'assurer une protection contre toutes les intrusions détectées (certaines étant traitées par le firewall).

Il est également possible de placer un capteur et un autre après le firewall. En fait, cette variante réunit les deux cas mentionnés ci-dessus. Mais elle est très dangereuse si on configure mal les capteurs et/ou le pare-feu, en effet on ne peut simplement ajouter les avantages des deux cas précédents à cette variante. Les capteurs IDS sont parfois situés à l'entrée de zones du réseau particulièrement sensibles (parcs de serveurs, données confidentielles…), de façon à surveiller tout trafic en direction de cette zone.

Les avantages des NIDS sont les suivants : les capteurs peuvent être bien sécurisés puisqu'ils se contentent d'observer le trafic et permettent donc une surveillance discrète du réseau, les attaques de type scans sont facilement détectées, et il est possible de filtrer le trafic.

Les NIDS sont très utilisés et remplissent un rôle indispensable, mais ils présentent néanmoins de nombreuses faiblesses. En effet, la probabilité de faux négatifs (attaques non détectées comme telles) est élevée et il est difficile de contrôler le réseau entier. De plus, ils doivent principalement fonctionner de manière cryptée d'où une complication de l'analyse des paquets. Pour finir, à l'opposé des IDS basés sur l'hôte, ils ne voient pas les impacts d'une attaque. Voici quelques exemples de NIDS : NetRanger, Dragon, NFR, Snort, ISSRealSecure.

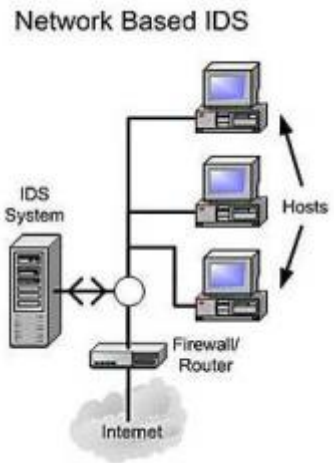


Figure 6: Architecture d'un NIDS

3.3.2 IDS Host

Les systèmes de détection d'intrusion basés sur l'hôte ou HIDS (Host IDS) analysent exclusivement l'information concernant cet hôte. Comme ils n'ont pas à contrôler le trafic du réseau mais "seulement" les activités d'un hôte ils se montrent habituellement plus précis sur les types d'attaques subies. De plus, l'impact sur la machine concernée est sensible immédiatement, par exemple dans le cas d'une attaque réussie par un utilisateur. Ces IDS utilisent deux types de sources pour fournir une information sur l'activité de la machine : les logs et les traces d'audit du système d'exploitation. Chacun a ses avantages : les traces d'audit sont plus précises et détaillées et fournissent une meilleure information alors que les logs qui ne fournissent que l'information essentielle sont plus petits. Ces derniers peuvent être mieux contrôlés et analysés en raison de leur taille, mais certaines attaques peuvent passer inaperçues, alors qu'elles sont détectables par une analyse des traces d'audit. **[16]**

Ce type d'IDS possède un certain nombre d'avantages : il est possible de constater immédiatement l'impact d'une attaque et donc de mieux réagir. Grâce à la quantité des informations étudiées, il est possible d'observer les activités se déroulant sur l'hôte avec précision et d'optimiser le système en fonction des activités observées. De plus, les HIDS sont extrêmement complémentaires des NIDS. En effet, ils permettent de détecter plus facilement les attaques de type "Cheval de Troie", alors que ce type d'attaque est difficilement détectable par un NIDS. Les HIDS permettent également de détecter des attaques impossibles à détecter avec un NIDS, car elles font partie de trafic crypté. Néanmoins, ce type d'IDS possède également ses faiblesses, qui proviennent de ses qualités : du fait de la grande quantité de données générées, ce type d'IDS est très sensible aux attaques de type DoS, qui peuvent faire exploser la taille des fichiers de logs. Un autre inconvénient tient justement à la taille des fichiers de rapport d'alertes à examiner, qui est très contraignante pour le responsable sécurité. La taille des fichiers peut en effet atteindre plusieurs Mégaoctets. Du fait de cette quantité de données à traiter, ils sont assez gourmands en CPU et peuvent parfois altérer les performances de la machine hôte. Enfin, ils ont moins de facilité à détecter les attaques de type hôte que les IDS réseaux. Les HIDS sont en général placés sur des machines sensibles, susceptibles de subir des attaques et possédantes des données sensibles pour l'entreprise. Les serveurs, web et applicatifs, peuvent notamment être protégés par un HIDS.

Pour finir, voici quelques HIDS connus: Tripwire, WATCH, DragonSquire, Tiger, Security Manager…

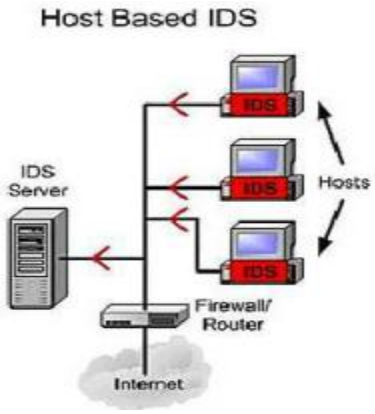


Figure 7: Architecture d'un HIDS

3.3.3 IDS Hybride

Les IDS hybrides rassemblent les caractéristiques des NIDS et HIDS. Ils permettent, en un seul outil, de surveiller le réseau et les terminaux. Les sondes sont placées en des points stratégiques, et agissent comme NIDS et/ou HIDS suivant leurs emplacements.[17] Toutes ces sondes remontent alors les alertes à une machine qui va centraliser le tout, et agréger/lier les informations d'origines multiples. Ainsi, on comprend que les IDS hybrides sont basés sur une architecture distribuée, où chaque composant unifie son format d'envoi. Cela permet de communiquer et d'extraire des alertes plus pertinentes. Les avantages des IDS hybrides sont multiples :

- Moins de faux positif.
- Meilleure corrélation (la corrélation permet de générer de nouvelles alertes à partir de celles existantes).
- Possibilité de réaction sur les analyseurs.

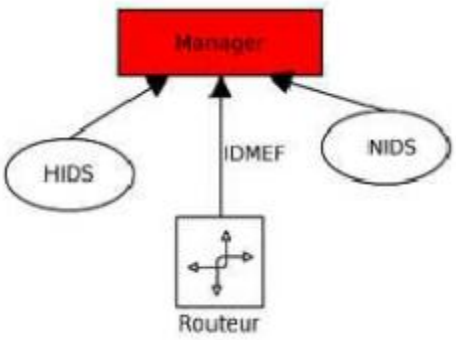


Figure 8: Architecture d'un IDS Hybride

3.3.4 Système de prévention d'intrusion IPS

Un système de prévention d'intrusion est un dispositif capable de détecter des attaques, connues et inconnues, et de les empêcher d'être réussies.[17] L'IPS n'est pas un observateur : il fait partie intégrante du réseau. Il est placé en ligne et examine tous les paquets entrants ou sortants.

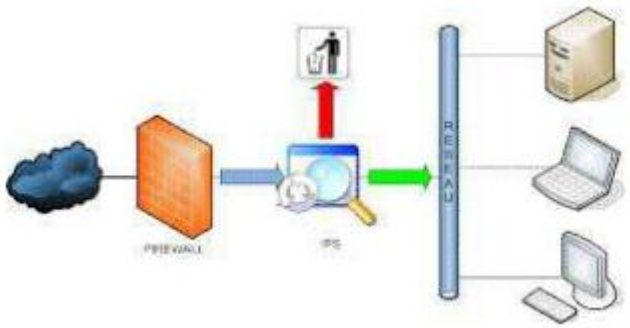


Figure 9 : Architecture d'un IPS

3.4 Mode de fonctionnement d'un IDS

Il faut distinguer deux aspects dans le fonctionnement d'un IDS : le mode de détection utilisé et la réponse apportée par l'IDS lors de la détection d'une intrusion. Il existe deux modes de détection, la détection d'anomalies et la reconnaissance de signatures. De même, deux types de réponses existent, la réponse passive et la réponse active.

3.4.1 Mode de détection

a) **La détection d'anomalie**: Elle consiste à détecter des anomalies par rapport à un profil "de trafic habituel". La mise en œuvre comprend toujours une phase d'apprentissage au cours de laquelle les IDS vont "découvrir" le fonctionnement "normal" des éléments surveillés. Ils sont ainsi en mesure de signaler les divergences par rapport au fonctionnement de référence.

b) **La reconnaissance de signature:** Cette approche consiste à rechercher dans l'activité de l'élément surveillé les empreintes (ou signatures) d'attaques connues. Ce type d'IDS est purement réactif ; il ne peut détecter que les attaques dont il possède la signature. De ce fait, il nécessite des mises à jour fréquentes.

   De plus, l'efficacité de ce système de détection dépend fortement de la précision de sa base de signature. Une signature permet de définir les caractéristiques d'une attaque, au niveau des paquets ou au niveau protocole.

3.4.2 Réponse passive et active

Il existe deux types de réponses, la réponse passive et la réponse active.

a) **La réponse passive :** La réponse passive d'un IDS consiste à enregistrer les intrusions détectées dans un fichier de log qui sera analysé par le responsable sécurité. Certains IDS permettent de logger l'ensemble d'une connexion identifiée comme malveillante. Ceci permet de remédier aux failles de sécurité pour empêcher les attaques enregistrées de se reproduire, mais elle n'empêche pas directement une attaque de se produire.
b) **La réponse active :** La réponse active au contraire a pour but de stopper une attaque au moment de sa détection.

3.5 Classification des IDS

Les différents systèmes de détection d'intrusions disponibles peuvent être classés selon plusieurs critères, qui sont :

- La méthode de détection.
- Le comportement du système après la détection.
- La source des données.
- La fréquence d'utilisation.

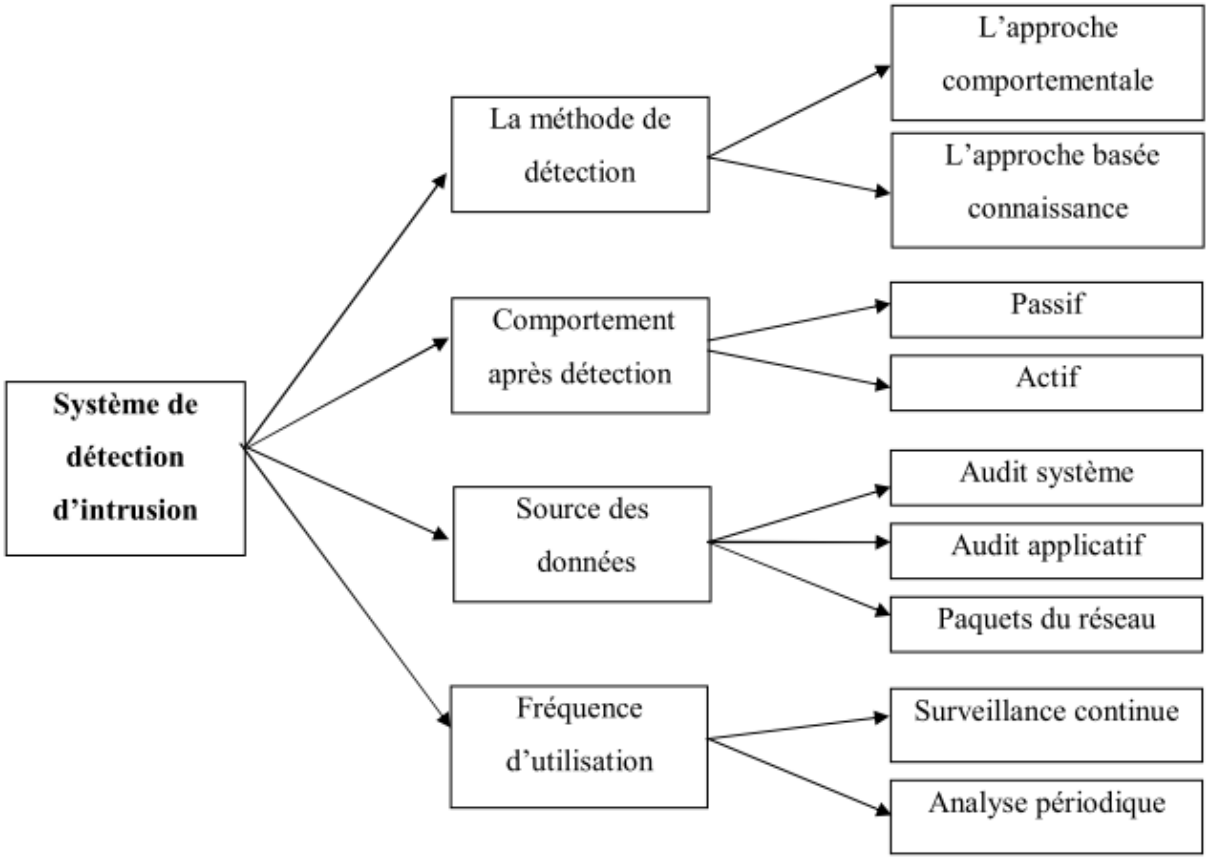


Figure 10: Classification d'un système de détection d'intrusion

3.6 Les types de menaces réseaux

Cette section donne un aperçu des quatre grandes catégories d'attaques réseau. Chaque attaque sur un réseau peut être placée dans l'un de ces groupes

a) **Déni de service (DoS)** : (en anglais Denial of Service) est une Attaque d'un serveur informatique destinée à l'empêcher de remplir sa fonction.[18] La méthode la plus classique consiste à faire crouler le serveur sous une masse de requêtes généralement mal formées à dessein pour entraîner une réponse anormale et paralysante. L'attaque utilise très souvent une multitude de "PC zombies" travaillant de concert, infectés par des backdoors ou chevaux de Troie, et mobilisables à distance par un pirate.

   Il est aussi possible de bloquer à distance des routeurs en tirant parti de failles de leur software, exemple : apache, smurf, neptune, ping of death, back, mail bomb, UDP storm etc

b) **Remote to User Attacks (R2L**) : Une attaque à distance est une attaque dans laquelle l'utilisateur envoie des paquets sur le réseau Internet vers une machine à laquelle il n'a pas accès, ceci afin d'exposer les vulnérabilités des machine et exploiter les failles et les privilèges qu'un utilisateur local pourrait avoir sur l'ordinateur, exemple : xlock, guest, xnsnoop, phf, sendmail dictionary etc.

c) **User to Root Attacks (U2R):** ces attaques consistent en l'exploitation des failles, le hacker commence par utiliser le système comme un compte utilisateur normal et essaye d'exploiter les vulnérabilités du système afin de gagner en privilèges, exemple : perl, xterm.

d) **Probing:** Probing, ou sondage, est une attaque ou le Hacker scanne ou sonde une machine ou un réseau de machine , afin de déterminer les failles et les vulnérabilités qui peuvent être exploitées afin d'endommager le système , cette technique est souvent utilisée dans le data mining, exemple : saint, portsweep, mscan, nmap etc.

4. Application de Data Mining dans la détection d'intrusion

L'exploration de données est une activité d'extraction d'information dont le but est de découvrir des faits cachés contenus dans des bases de données. En utilisant une combinaison d'apprentissage automatique, d'analyse statistique, des techniques de modélisation et de la technologie de base de données, le data mining trouve des motifs et des relations subtiles dans les données et en déduit des règles qui permettent la prévision de résultats futurs.

Le but de la détection d'intrusion est de détecter les violations de la sécurité des systèmes d'information. La détection d'intrusion est une approche passive de la sécurité car il surveille les systèmes d'information et soulève des alarmes lorsque des violations de sécurité sont détectées.

Les méthodes de fouille de données les plus utilisées pour la détection d'intrusions sont la classification et le clustering. Avant de présenter de ces méthodes nous allons détailler une étape commune à ces méthodes : la sélection d'attributs

4.1 Sélection d'attributs

Un choix pertinent des attributs impacte sur la qualité des résultats obtenue. Les méthodes de fouille de données sont plus efficaces s'il existe des connaissances sur les

attributs de domaine, sur la priorité de ces attributs, sur les attributs moins importants et les relations éventuelles entre les attributs.

4.2 IDS et classification

Les IDS basés sur les techniques de classification ont pour but de classifier les trafics d'un réseau en deux classes : "général" et "intrusion". La classification nécessite un apprentissage. La précision de cet apprentissage assure la diminution du taux de faux positifs (i.e. les cas normaux classés comme intrusions) et du taux de faux négatifs (i.e. les intrusions classées comme normales).

4.3 IDS et Clustering

Le clustering est une technique de fouille de données permettant de regrouper les éléments selon leur ressemblance. Contrairement à la classification, il n'y a pas besoin de connaissances a priori sur les classes dans lesquelles on regroupe les éléments. On n'a pas besoin des données labélisées (par les classes prédéfinis) pour l'apprentissage.

Exemple d'application

Un système de détection d'intrusions basé sur la détection d'anomalies contrôle les activités du système afin de les classer comme normales ou anomalies. Le but est l'exploitation des techniques de data mining pour extraire des anomalies à partir des grandes quantités de données du trafic réseau. Parmi les travaux existants, on peut citer ADAM « Audit Data Analysis and Mining »qui est un système de détection d'intrusions qui exploite des techniques de data mining pour construire des profils du trafic réseau normaux.

ADAM utilise les règles d'association pour construire des profils du trafic de réseau normaux qui seront employées par la suite pour détecter les comportements incorrects de trafic de réseau. Pour détecter des anomalies, ADAM extrait les règles d'association à partir des données du trafic réseau et qui seront comparées aux profils du réseau. Si n'importe quelle règle d'association produite à partir des données de trafic de réseau rassemblées n'est pas incluse dans les profils, alors cette règle est considérée comme une indication d'un comportement incorrect.

5. Conclusion

Dans ce chapitre, nous avons présenté le système de détection d'intrusions et nous avons également étudié d'une manière détaillée les différents types d'IDS selon différents critères de classification avec la présentation générale des différentes techniques utilisées pour la détection d'intrusions. Afin d'obtenir un système de détection d'intrusions compétent et efficace, il est souhaitable d'utiliser les deux techniques de détection comportementale et basée connaissances en parallèle pour surmonter les problèmes liés à chacune de ces deux techniques de détection. Cependant, les systèmes de détection d'intrusions commercialisés emploient seulement la technique de détection basée connaissance, ce qui motive les différents efforts de recherche dans le domaine de la détection d'anomalies.

Pour cette raison, différentes approches sont utilisées pour implémenter la technique de la détection d'anomalies. Parmi ces approches diverses, nous nous somme intéressées à l'approche du data mining. Cette approche qui présente beaucoup d'aspects intéressants pour le développement d'un système de détection d'intrusions efficace.

Chapitre III
Application des Algorithmes Génétiques pour la Sécurité Informatique

I. Introduction

Les algorithmes génétiques sont des algorithmes d'optimisation s'appuyant sur des techniques dérivées de la génétique et des mécanismes d'évolution de la nature

1. Historique :

Au siècle dernier, Charles Darwin observa les phénomènes naturels et fit les constatations suivantes [19]:

- L'évolution n'agit pas directement sur les êtres vivants ; elle opère en réalité sur les chromosomes contenus dans leur ADN.
- L'évolution a deux composantes : la sélection et la reproduction.
- La sélection garantit une reproduction plus fréquente des chromosomes les plus forts.
- La reproduction est la phase durant laquelle s'effectue l'évolution.

Dans les années 60s, John H. Holland expliqua comment ajouter de l'intelligence dans un programme informatique avec les croisements (échangeant le matériel génétique) et la mutation (source de la diversité génétique).

Il formalisa ensuite les principes fondamentaux des algorithmes génétiques :

- La capacité de représentations élémentaires, comme les chaînes de bits, à coder des structures complexes.
- Le pouvoir de transformations élémentaires à améliorer de telles structures.

Et récemment, David E. Goldberg ajouta à la théorie des algorithmes génétiques les idées suivantes :

- Un individu est lié à un environnement par son code d'ADN.
- Une solution est liée à un problème par son indice de qualité.

2. Terminologie :

- Les chromosomes : sont les éléments à partir desquels sont élaborées les solutions (mutation et croisement génétiques).
- La population : (génération) est l'ensemble des chromosomes

- La reproduction : est l'étape de combinaison des chromosomes

3. Principe :

Pour utiliser les techniques génétiques il faut faire deux opérations :

1- Une fonction de codage de données en entrée sous forme d'une séquence de bits.
2- Trouver une fonction U(x) pour pouvoir calculer l'adaptation d'une séquence de bits x.

Après avoir trouvé ces deux fonctions on peut appliquer l'AG :

1- Générer aléatoirement quelques séquences de bits pour composer la soupe initiale.
2- Mesurer l'adaptation de chacune des séquences présentes.
3- Reproduction des séquences en fonction de son adaptation.
4- Faire l'opération de croisement aléatoirement de quelque paire de séquences, et ce sera :

- Une séquence est composée de la 1^{ere} partie de la 1^{ere} séquence et de la 2^{nd} partie de la 2^{nd} séquence.
- Et une séquence est composée de la 2^{nd} partie de la 1^{ere} séquence et de la 1^{ere} partie de la 2^{nd} séquence.

5- Faire l'opération de mutation d'un bit choisi aléatoirement dans une ou plusieurs séquences.
6- Retour à l'étape 2 (mesurer l'adaptation à nouveau).

4. les opérateurs

4.1 Codage binaire et réel des variables :

Pour résoudre un problème il faut d'abord coder les paramètres, un gène Correspond à une variable d'optimisation Xi, et un ensemble de gène correspond à un chromosome, un individu a un ou plusieurs chromosomes et une population c'est un ensemble d'individus.

Dans l'informatique nous utilisons un codage binaire (0 et 1), par exemple un gène est un entier long (32 bits).

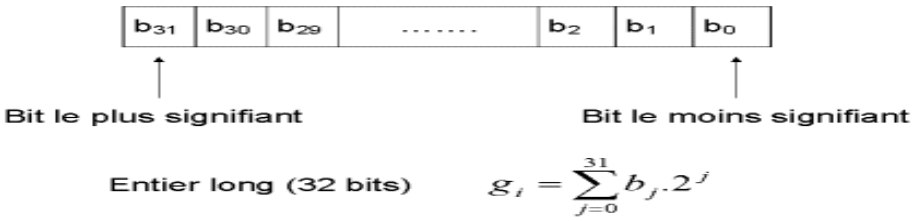
Les cinq niveaux d'organisation schéma du codage de variable d'optimisation Xi de notre AG

Un des avantages du codage binaire est que l'on peut facilement coder les objets : réels, des entiers, des chaînes de caractères …etc.

Pour résoudre il nous faut un espace de recherche fini : $0 \leq g_i \leq g_{max}$ i =1 a n

avec :g_{max} = 2**32 –1=4294967295 valeurs discrètes.

Chaque gène est codé par 32 bits.



Les formules de codage et de décodage sont :

$g_i = (x_i - x_{imin} / x_{imax} - x_{min}) g_{max}$.

$xi = x_{imin} + (x_{imax} - x_{imin}) g_i / g_{max}$

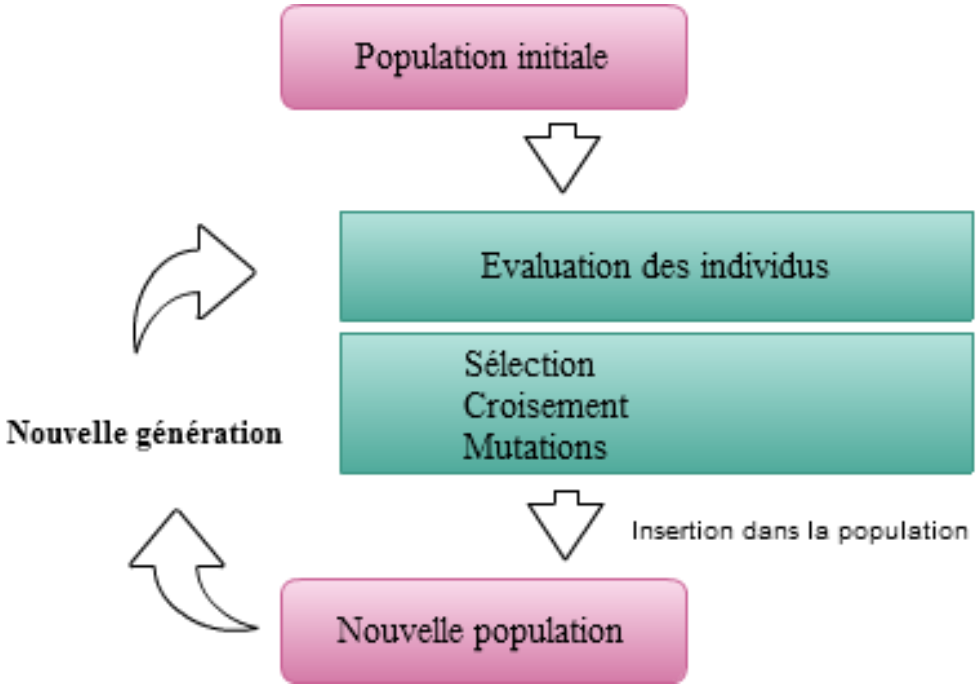
Contrairement au codage binaire, dans le codage réel il n'y a pas d'opération de conversion vers le réel et le binaire, et alors l'AG soit plus dépendant du problème, chacune des composantes correspond à une inconnue du problème. Et pour l'insertion des chromosomes soit :

Par l'insertion des chromosomes directement dans la population de la génération suivante.

- Par "l'élitisme " : c'est l'insertion du meilleur chromosome dans la population de la génération suivante et de compléter l'individu par des chromosomes d'uns manière traditionnelle.
- Par "la population sans double " : c'est l'insertion du chromosome dans la population de la génération suivante à condition qu'il soit différent de tous les chromosomes de la nouvelle génération.

4.2 Sélection

Représentation schématique du fonctionnement de l'AG



Il y' a plusieurs méthodes de sélection, citons quelques-unes :

a. Roulette de casino :

C'est la sélection naturelle la plus employée pour l'AG binaire. Chaque chromosome occupe un secteur de roulette dont l'angle est proportionnel à son indice de qualité. Un chromosome est considéré comme bon aura un indice de qualité élevé, un large secteur de roulette et alors il aura plus de chance d'être sélectionné.

b. N/2 –élitisme :

Les individus sont triés selon leur fonction d'adaptation, seul la moitié supérieure de la population correspondant aux meilleurs composants est sélectionnée, nous avons constates que la pression de sélection est trop forte, il est important de maintenir une diversité de gènes pour les utilise dans la population suivante et avoir des populations nouvelles quant on les combine.

c. par tournoi :

Choisir aléatoirement deux individus et on compare leur fonction d'adaptation (combattre) et on accepte la plus adapte pour accéder à la génération intermédiaire, et on répète cette opération jusqu'à remplir la génération intermédiaire (N/2 composants). Les individus qui gagnent à chaque fois on peut les copier plusieurs fois ce qui favorisera la pérennité de leurs gènes.

4.3 Croisement :

Le phénomène de croisement est une propriété naturelle de l'ADN, et c'est analogiquement qu'on fait les opérations de croisement dans les AG.

a. croisement binaire :

- **croisement en un point :**

On choisit au hasard un point de croisement, pour chaque couple (fig. 1). Notons que le croisement s'effectue directement au niveau binaire, et non pas au niveau des gènes. Un chromosome peut donc être coupé au milieu d'un gène.

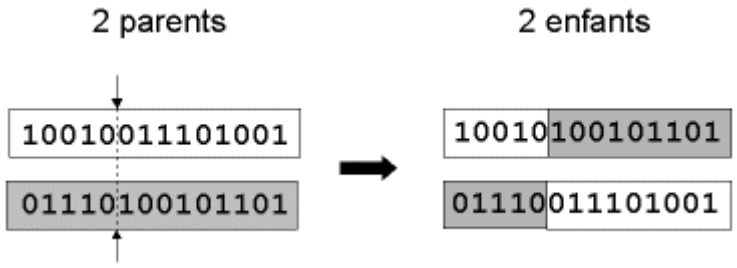


Figure 1: représentation schématique du croisement en 1 point. Les chromosomes sont bien sûr généralement beaucoup plus longs.

a. croisement en deux points:

On choisit au hasard deux points de croisement (Fig. 2). Par la suite, nous avons utilisé cet opérateur car il est généralement considéré comme plus efficace que le précédent [Beasley, 1993b]. Néanmoins nous n'avons pas constaté de différence notable dans la convergence de l'algorithme.

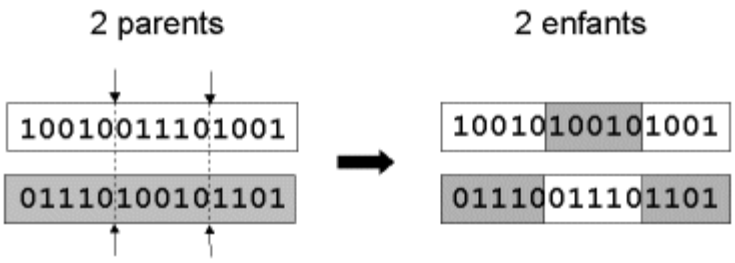


Figure 2: représentation schématique du croisement en 2 points.

Notons que d'autres formes de croisement existent, du croisement en k points jusqu'au cas limite du croisement uniforme.

b. Croisement réel :

Le croisement réel ne se différencie du croisement binaire que par la nature des éléments qu'il altère : ce ne sont plus des bits qui sont échangés à droite du point de croisement, mais des variables réelles.

c. Croisement arithmétique :

Le croisement arithmétique est propre à la représentation réelle. Il s'applique à une paire de chromosomes et se résume à une moyenne pondérée des variables des deux parents.

Soient [ai, bi, ci] et [aj, bj, cj] deux parents, et p un poids appartenant à l'intervalle [0, 1], alors les enfants sont [pai + (1-p)aj, pbi + (1-p)bj, pci + (1 - p)cj] ...

Si nous considérons que p est un pourcentage, et que i et j sont nos deux parents, alors l'enfant i est constitué à p% du parent i et à (100-p)% du parent j, et réciproquement pour l'enfant j.

4.4 Mutation :

Nous définissons une *mutation* comme étant l'inversion d'un bit dans un chromosome (Fig. 3). Cela revient à modifier aléatoirement la valeur d'un paramètre du dispositif. Les mutations jouent le rôle de bruit et empêchent l'évolution de se figer. Elles permettent d'assurer une recherche aussi bien globale que locale, selon le poids et le nombre des bits mutés. De plus, elles garantissent mathématiquement que l'optimum global peut être atteint.

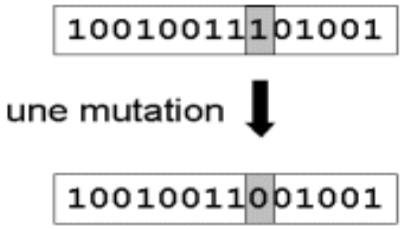


Figure 3: représentation schématique d'une mutation dans un chromosome.

D'autre part, une population trop petite peut s'homogénéiser à cause des erreurs stochastiques : les gènes favorisés par le hasard peuvent se répandre au détriment des autres. Cet autre mécanisme de l'évolution, qui existe même en l'absence de sélection, est connu sous le nom de dérive génétique. Du point de vue du dispositif, cela signifie que l'on risque alors d'aboutir à des dispositifs qui ne seront pas forcément optimaux. Les mutations permettent de contrebalancer cet effet en introduisant constamment de nouveaux gènes dans la population [Beasley, 1993b].

Comment réaliser notre opérateur mutation ? De nombreuses méthodes existent. Souvent la probabilité de mutation p_m par bit et par génération est fixée entre 0,001 et 0,01. On peut prendre également $p_m=1/l$ où l est le nombre de bits composant un chromosome. Il est possible d'associer une probabilité différente de chaque gène. Et ces probabilités peuvent être fixes ou évoluer dans le temps.

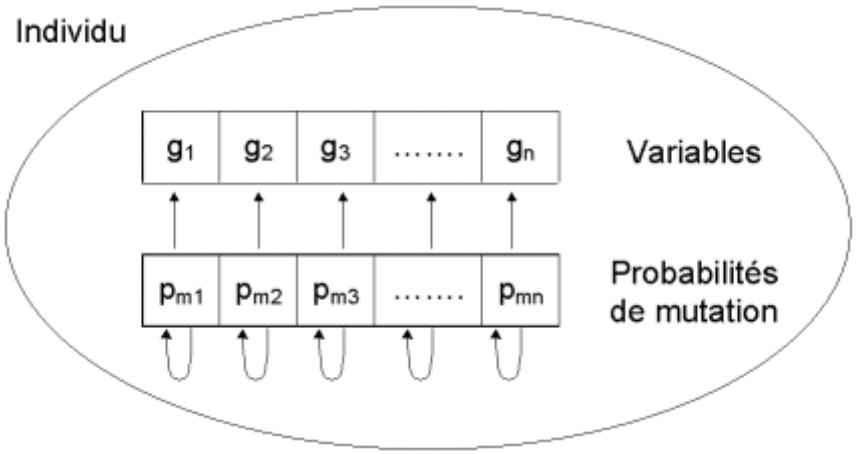


Figure 4: principe de l'auto-adaptation. A chaque variable est associée sa propre probabilité de mutation, qui est elle-même soumise au processus d'évolution. L'individu possède donc un second chromosome codant ces probabilités.

Après divers essais, ils ont abouti à la méthode d'*auto-adaptation* des probabilités de mutation [Bäck, 1992]. Si dans un environnement stable il est préférable d'avoir un taux de mutation faible, la survie d'une espèce dans un environnement subissant une évolution rapide nécessite un taux de mutation élevé permettant une adaptation rapide. Les taux de mutation d'une espèce dépendent donc de leur environnement [Wills, 1991].

Pour prendre en compte cette formulation biologique et l'adapter à notre cas, ils ont introduit dans chaque individu (dispositif) un second chromosome (ensemble de paramètres) dont les gènes (paramètres) représentent les probabilités de mutation de chaque gène du premier chromosome (Fig. 4). Ce second chromosome est géré de façon identique au premier, c'est-à-dire qu'il est lui-même soumis aux opérateurs génétiques (croisement et mutation). Cela revient à fixer les probabilités assurant la modification des valeurs des paramètres du composant en fonction des valeurs d'un ensemble d'autres paramètres (les probabilités de mutation).

Lors de la genèse, les probabilités de mutation sont posées égales à 0,1 (valeur qui a paru la meilleure après plusieurs essais). Au cours du déroulement de l'algorithme, les gènes et les individus ayant des probabilités de mutation trop élevées ont tendance à disparaître. De même, les gènes ayant des probabilités de mutation trop faibles ne peuvent pas évoluer favorablement et tendent à être supplantés. Les probabilités de mutation dépendent donc du gène considéré et de la taille de la population. De plus, elles évoluent au cours du temps. Il y a donc auto-adaptation des probabilités de mutation.

a. Mutation binaire :

La mutation binaire s'applique à un seul chromosome. Un bit du chromosome est tiré au hasard. Sa valeur est alors inversée.

Il existe une variante où plusieurs bits peuvent muter au sein d'un même chromosome. Un test sous le taux de mutation est effectué non plus pour le chromosome mais pour chacun de ses bits : en cas de succès, un nouveau bit tiré au hasard remplace l'ancien.

b. mutation réelle :

La mutation réelle ne se différencie de la mutation binaire que par la nature de l'élément qu'elle altère : ce n'est plus un bit qui est inversé, mais une variable réelle qui est de nouveau tirée au hasard sur son intervalle de définition.

c. mutation non uniforme :

La mutation non uniforme possède la particularité de retirer les éléments qu'elle altère dans un intervalle de définition variable et de plus en plus petit. Plus nous avançons dans les générations, moins la mutation n'écarte les éléments de la zone de convergence. Cette mutation adaptative offre un bon équilibre entre l'exploration du domaine de recherche et un affinement des individus.
Le coefficient d'atténuation de l'intervalle est un paramètre de cet opérateur.

II. Application des Algorithmes génétiques pour la détection d'Intrusion :

1. IDS à base d'algorithmes génétiques :

L'application des algorithmes génétiques pour la détection d'intrusion semble être un domaine prometteur. Nous allons parler de la motivation et de la mise en œuvre des détails dans cette section :

Les algorithmes génétiques peuvent être utilisés pour faire évoluer des règles simples pour le trafic réseau (Sinclair, Pierce, et Matzner 1999). Ces règles sont utilisées pour différencier les connexions réseau normales de connexions anormales. Ces connexions anormales se réfèrent à des événements avec une probabilité d'intrusions. Les règles stockées dans la base de règles sont généralement sous la forme suivante (Sinclair, Pierce, et Matzner 1999):

SI {état} ALORS {acte}

Pour les problèmes que nous avons présentées ci-dessus, la condition se réfère généralement à un match entre la connexion réseau actuelle et les règles en IDS, tels que les adresses IP source et destination et les numéros de port (utilisés dans les protocoles de réseau TCP / IP), la durée de la connexion, protocole utilisé, etc., indiquant la probabilité d'une intrusion. Le champ d'acte se réfère généralement à une action définie par les politiques de sécurité

Le champ d'acte se réfère généralement à une action définie par les politiques de sécurité au sein d'une organisation, tels que des rapports d'alerte à l'administrateur du système, l'arrêt de la connexion, un message de connexion dans le dossier de vérification du système, ou la totalité de ce qui précède. Par exemple, une règle peut être définie comme:

SI { la connexion a les informations suivantes:

L'adresse IP source 124.12.5.18; l'adresse IP destination: 130.18.206.55;

Le numéro de port de destination: 21; temps de connexion: 10,1 secondes}

Alors {arrêter la connexion}

Cette règle peut être expliqué comme suit: si il existe une demande de connexion de réseau avec l'adresse IP source 124.12.5.18, adresse IP destination 130.18.206.55, numéro de port de destination 21, et le temps de connexion 10,1 secondes, Alors arrêter l'établissement de cette connexion. Ceci est parce que l'adresse IP 124.12.5.18 est reconnu par l'IDS comme l'une des adresses IP sur la liste noire; par conséquent, toute demande de service initiée à partir d'elle est rejetée.

L'objectif final de l'application des Algorithmes Génétiques est de générer des règles qui détectent les connexions anormales. Ces règles sont testées sur des liens historiques et sont utilisés pour filtrer les nouvelles connexions pour trouver le trafic réseau suspect.

Dans cette mise en œuvre, le trafic de réseau utilisé pour les AG est un ensemble de données pré-classés qui différencie les connexions réseau normales des connexions anomales. Cet ensemble de données sont collectées à l'aide de renifleurs de réseau (un programme utilisé pour le trafic réseau d'enregistrement sans faire quelque chose de nocif) tels que tcpdump (http://www.tcpdump.com) ou Snort (http://www.snort.com). L'ensemble des données est classé manuellement sur la base de connaissances d'experts. Il est utilisé pour l'évaluation de la condition lors de l'exécution de l'AG. En commençant AG avec seulement un petit ensemble de règles générées de façon aléatoire, nous pouvons générer un ensemble de données plus vaste qui contient des règles pour les IDS. Ces règles sont des solutions «assez bonnes» pour les AG et peuvent être utilisées pour filtrer de nouveaux trafics réseaux.

2. Représentation de données :

Afin d'exploiter pleinement le niveau suspect, nous devons examiner tous les domaines liés à une connexion réseau spécifique. Pour plus de simplicité, nous ne considérons que certains attributs évidents pour chaque connexion. La définition de règles (pour les protocoles TCP / IP) est indiqué dans le tableau 1.

La règle correspondante de l'attribut "Exemple de valeur" dans le tableau 1 peut être traduit par:

SI {la connexion a les informations suivantes:

L'adresse IP source 209.11 ?? ??;.. l'adresse IP de destination:130.18.176+?.?? ;

Le numéro de port source: 42335; Le numéro de port de destination: 80;

Le temps de connexion: 482 secondes;

La connexion est interrompue par l'expéditeur; le protocole utilisé est TCP;

L'auteur a envoyé 7320 octets de données; et le répondeur envoyé 38 891 octets de données}

ALORS {arrêter la connexion}

Attribut	Plage de valeur	Valeur d'exemple	Descriptions
Adresse IP source	0.0.0.0~255.255.255.255	d1.0b.**.** (209.11.??.??)	Un sous-réseau avec l'adresse IP 209.11.0.0 à 209.11.255.255
Adresse IP destination	0.0.0.0~255.255.255.255	82.12.b*.** (130.18.176+?.??)	Un sous-réseau avec l'adresse IP 130.18.176.0 à 130.18.255.255
Numéro du Port source	0~65535	42335	Numéro de port source de la connexion
Numéro du port de destination	0~65535	00080	Numéro de port de destination indique que cela est un service http
La durée	0~99999999	00000482	La durée de la connexion est de 482 secondes
L'état	1~20	11	La connexion est interrompue par l'initiateur, à usage interne
Le protocole	1~9	2	Le protocole pour cette connexion est TCP
Nombre d'octets envoyés par l'expéditeur	0~9999999999	0000007320	L'expéditeur envoie 7320 octets de données
Nombre d'octets envoyés par le répondeur	0~9999999999	0000038891	Les répondeurs de 38 891 octets de données

Tableau III.1: Définition de la règle pour la connexion et la gamme des valeurs de chaque champ

On peut convertir l'exemple ci-dessus sous la forme de chromosome, tel que décrit dans la figure.

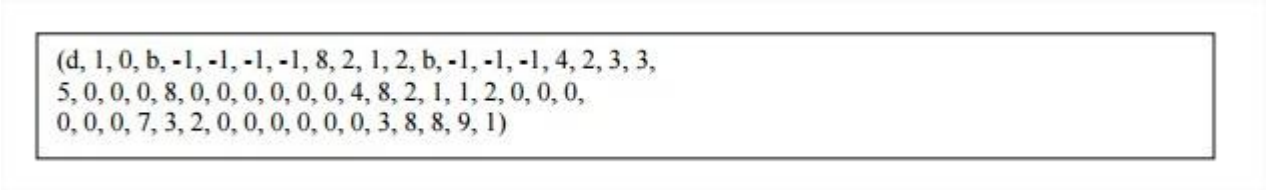


Figure 5 : Structure chromosome de l'exemple du Tableau 1

Au total, il ya cinquante-sept gènes de chaque chromosome. Pour plus de simplicité, nous utilisons des représentations hexadécimales pour les adresses IP. La règle peut être expliqué comme suit:

Si une connexion réseau avec l'adresse IP source 209.11 ?? ??. (209.11.0.0 ~ 209.11.255.255), L'adresse IP de destination 130.18.176. ?? (130.18.176.0 ~ 130.18.255.255), le numéro de port source 42335, numéro de port de destination 80, la durée de la connexion 482 secondes, l'état de terminaison 11 (la connexion terminée par l'auteur), utilise le protocole du type 2 (TCP), et l'expéditeur envoie 7320 octets de données, les intervenants envoie 38 891 octets de données, alors ceci est un comportement suspect et peut être identifié comme une intrusion potentielle. La validité réelle de cette règle sera examinée par la correspondance à l'ensemble de données historiques comprenant des raccordements marqués comme étant anormale ou normale. Si la règle est en mesure de trouver un comportement anormal, un bonus sera donné au chromosome actuel. Si la règle correspond à une connexion normale, une pénalité sera appliquée au chromosome. Clairement il n'ya pas de règle unique qui peut être utilisée pour séparer toutes les connexions anormales de connexions normales. La population a besoin d'évoluer pour trouver l'ensemble des règles optimales.

Dans l'exemple représenté dans le tableau 1, certaines cartes sauvages (le caractère "*" et le caractère '?') Sont utilisés et les gènes correspondants au sein du chromosome sont présentés comme -1. Ces cartes sauvages sont utilisées pour représenter une gamme appropriée de valeurs spécifiques (Crosbie et Spafford, 1995). Il est utile lorsqu'on représente un bloc de réseau (une plage d'adresses IP ou les numéros de port) dans une règle. Une fois l'information spatiale est incluse dans les règles, la capacité de l'IDS peut être grandement améliorée comme une intrusion qui peut être initié de nombreux endroits différents. La prise en compte

du temps de durée d'une connexion de réseau dans le chromosome assure l'incorporation de l'information temporelle pour les connexions réseau. La valeur maximale du temps de la durée est de 99999999 secondes, ce qui est plus d'un an. Ceci est utile pour identifier les intrusions parce que les intrusions complexes peuvent durer des heures, des jours, voire des mois.

L'algorithme génétique commence avec une population qui a des règles choisies au hasard. La population peut évoluer en utilisant les opérateurs de croisement et de mutations. En raison de l'efficacité de la fonction d'évaluation, les populations suivantes sont orientées vers des règles qui correspondent à des connexions intrusives. En fin de compte, quand l'algorithme s'arrête, des règles sont sélectionnées et ajoutés dans la base de règles IDS.

3. Les paramètres dans les Algorithmes génétiques :

Il ya beaucoup de paramètres à prendre en considération pour l'application des AG. Chacun de ces paramètres influence fortement l'efficacité de l'algorithme génétique. Nous allons discuter de la méthodologie et les paramètres connexes dans la section suivante.

3.1 La fonction d'évaluation :

La fonction d'évaluation est l'un des paramètres les plus importants de l'algorithme génétique. La mise en œuvre proposée diffère du système utilisé par (Crosbie et Spafford, 1995) dans ce cas la définition sur les calculs du résultat et de remise en forme est différente. Les étapes suivantes sont utilisées pour calculer la fonction d'évaluation. D'abord, le résultat global est calculé sur la base si un champ de la connexion correspond à l'ensemble de données pré-annoncé, puis multiplier le poids de ce domaine. La valeur assortie est réglé sur 1 ou 0.

$$Résulats = \sum_{i=1}^{57} Assorties * Poids_i$$

L'ordre des valeurs de poids dans la fonction est représenté sur la figure 4. Ces commandes sont classées selon différents champs de l'enregistrement de connexion tel qu'ils ont été rapporté par les renifleurs de réseau. Par conséquent, tous les gènes représentant le champ de l'adresse IP de la destination ont le même poids. Les valeurs réelles peuvent être finement réglées au moment de l'exécution.

L'idée de base derrière cette commande est l'importance des différents champs de paquets TCP / IP. Ce schéma est simple et intuitive. L'adresse IP de destination est la cible d'une intrusion alors que l'adresse IP source est à l'origine de l'intrusion. Ce sont les pièces les plus importantes de l'information nécessaires pour capturer une intrusion. Le numéro de port de destination indique aux applications que le système cible est en cours d'exécution (par exemple, le service FTP fonctionne habituellement sur le port 21). Certaines adresses IP sont les cibles les plus probables pour les intrusions, par exemple, les adresses IP pour les domaines militaires. Informations spécifiques à un domaine est moins important par rapport aux adresses IP source. D'autres paramètres tels que la durée, octets envoyés par l'expéditeur, octets envoyés par le récepteur, et l'État sont généralement moins importants que les champs ci-dessus, mais sont encore utiles. Les champs de numéro de protocole et le port source sont couramment dispensables et sont utilisés pour identifier certaines intrusions spécifiques.

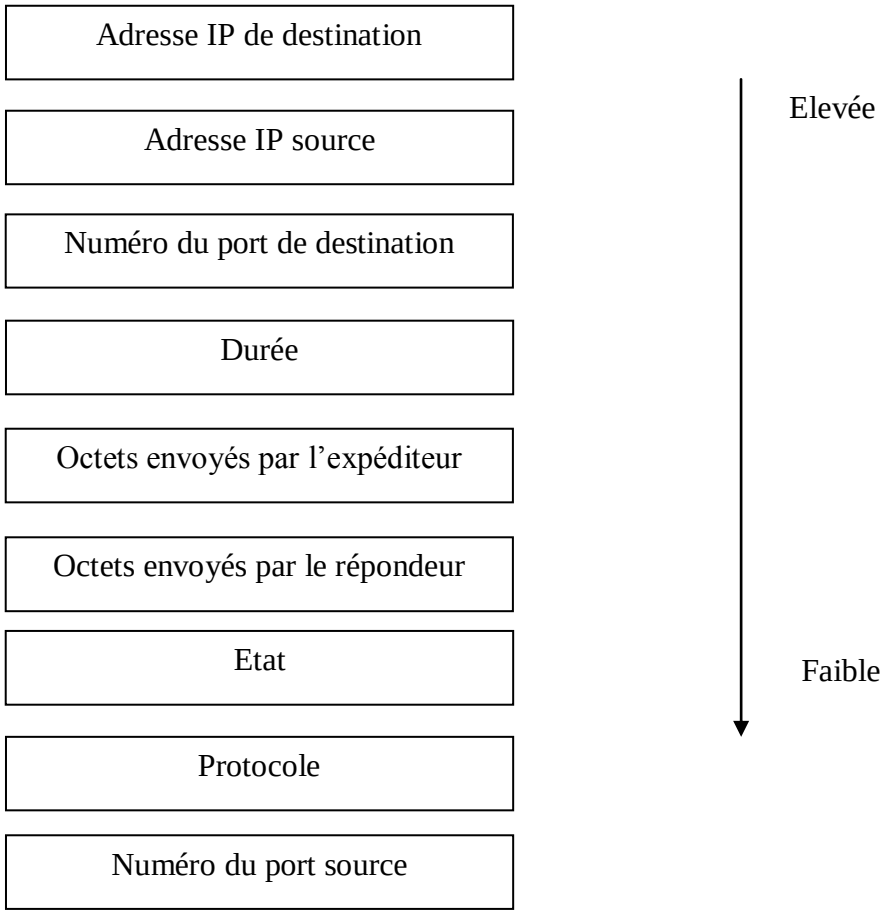


Figure 6 : Ordre de poids pour les champs dans la fonction d'évaluation

uspect est ensuite calculée en utilisant l'équation suivante. Le niveau suspect est un seuil qui indique la mesure dans laquelle les deux connexions réseaux sont considérées comme une

«correspondance». La valeur réelle de niveau suspect reflète les observations à partir de données historiques.

Δ = |Résultats – Niveau suspect |

Une fois le décalage produit, la valeur de pénalité est calculée en utilisant la différence absolue. Le classement dans l'équation indique si oui ou non une intrusion est facile à identifier.

Peine = $\left(\frac{\Delta * \text{Classement}}{100}\right)$.

L'aptitude d'un chromosome est calculée en utilisant la peine ci-dessus:

Remise en forme = 1- peine

De toute évidence, la gamme de la valeur de remise en forme est comprise entre 0 et 1. En définissant l'évaluation, nous avons incorporé à la fois l'aspect temporel et spatial des informations nécessaires à l'identification des intrusions réseau.

3.2 Autres Définition :

Une fausse positive : est une alerte pour une attaque qui n'a pas eu lieu.[16]

Une fausse négative : est une attaque qui n'a pas été détecté et n'a pas été filtré.[16]

3.3 Croisement et Mutation :

Les algorithmes génétiques classiques ont été utilisés pour identifier et faire converger les populations des hypothèses candidates à un seul optimum global. Pour ce problème, un ensemble de règles sont nécessaire comme base pour l'IDS. Comme mentionné précédemment, il est impossible d'identifié clairement si une connexion réseau est normale ou anormale simplement en utilisant une règle. Plusieurs règles sont nécessaires pour identifier les anomalies indépendantes, ce qui signifie que plusieurs bonnes règles sont plus efficaces qu'une seule meilleure règle (Sinclair, Pierce, et Matzner 1999). Une autre raison de trouver plusieurs règles est que parce qu'il ya tellement de possibilités de connexion réseau, un petit ensemble de règles sera loin d'être suffisant.

Pour l'utilisation de l'algorithme génétique, nous devons trouver la maximale locale (un ensemble de solutions "assez bonne") par opposition au maximum global (la meilleure solution) (Sinclair, Pierce, et Matzner 1999). Les techniques de nichage peuvent être utilisés pour trouver des maximales locales multiples (Miller et Shaw, 1996; voir aussi Sinclair, Pierce, et Matzner 1999). Il est basé sur l'analogie de la nature dans dans chaque environnement, il existe différents sous-espaces (niches) qui peuvent soutenir différents types de vie. D'une manière similaire, l'algorithme génétique peut maintenir la diversité de chaque population dans un domaine multimodal, qui se réfère à des domaines nécessitant l'identification de multiples optima. Deux méthodes de base, le surpeuplement et le partage peuvent être utilisés pour nichage. La méthode éviction utilise le membre le plus proche pour le remplacement de ralentir la population à converger vers un seul point dans les générations suivantes. La méthode de partage réduit l'aptitude des individus qui ont des membres très semblables et les forces des individus d'évoluer à d'autres maxima locaux qui peuvent être moins peuplée. Les métriques de similarité utilisées dans ces techniques peuvent être phénotype à génotype similaire tels que la distance de Hamming entre les représentations binaires ou phénotype similitude tels que la relation entre deux connexions réseau dans ce problème. Ce dernier est plus agité pour trouver des règles utilisées dans IDS. L'inconvénient de cette approche est qu'elle nécessite plus de connaissances spécifiques à un domaine (Miller et Shaw, 1996; voir aussi Sinclair, Pierce, et Matzner 1999).

L'opération de mutation devrait être significative au cours de l'évolution. Par exemple, chaque segment de l'adresse IP ne doit pas dépasser 255 (représentation décimale). Les mutations doivent être effectuées suivant les exigences spécifiées dans le tableau 1. Ces limitations peuvent être forcées en définissant des règles de mutation appropriées.

3.4 Autre paramètres :

Il ya aussi d'autres paramètres qui doivent être pris en considération, tels que le taux de mutation, taux de croisement, nombre de populations, et le nombre de générations. Ces paramètres doivent être ajustés en fonction de l'environnement de l'application du système et la politique de sécurité de l'organisation.

4. DataSet :

Pour mettre en œuvre l'algorithme et évaluer la performance du système, je propose les jeux de données standards employés dans KDD Coupe compétition 1999 »Computer Network Intrusion Détection".

Les ensembles de données KDD 99 de détection d'intrusion dépendent de la proposition DARPA 1998, qui est offerte aux concepteurs de systèmes de détection d'intrusion (IDS) avec une norme sur laquelle évaluer les différentes méthodologies ([21], [24]). Ainsi, une simulation est préparée à partir d'un réseau militaire fabriqué avec trois machines «cibles» exécutant différents services et systèmes d'exploitation. Ils ont également demandé trois machines supplémentaires pour usurper des adresses IP différentes pour générer du trafic réseau.

Une connexion est une série de paquets TCP commençant et se terminant à certaines périodes bien définies, entre lesquelles les inondations de données à partir d'une adresse IP source à une adresse IP cible sous un protocole bien défini. Il en résulte 41 caractéristiques pour chaque connexion.

Enfin, il reste un renifleur qui représente tout le trafic réseau à l'aide du format de vidage TCP. La période totale simulée est de sept semaines. Les connexions normales sont contournées exceptés dans un réseau militaire où les attaques sont classés dans l'un des quatre types: User Root; Remote to Local; Denial of Service; and Probe.

La KDD 99 détection d'intrusion de référence est composée de différents éléments :

kddcup.data;

kddcup.data_10_percent;

kddcup.newtestdata_10_percent_unlabeled;

kddcup.testdata.unlabeled;

kddcup.testdata.unlabeled_10_percent; corrigé.

Je propose d'utiliser "kddcup.data_10_percent" que la formation ensemble de données et "corrigée" que les tests ensemble de données. Dans ce cas, l'ensemble de la formation se compose de 494 021 dossiers dont 97 280 sont des enregistrements de connexion normales, tandis que l'ensemble de test contient 311 029 enregistrements parmi lesquels 60 593 sont des enregistrements de connexion normales. Le tableau 1 montre la répartition des types d'intrusion dans la formation et les jeux de données de test.

Dataset	normal	prob	Dos	u2r	r2l	Total
Train	97280	4107	391458	52	1124	494021
Test ("corrected")	60593	4166	229853	228	16189	311029

Tableau III.2: Distribution des Types d'intrusion dans les DataSets

Types de menaces prises en considérations dans KDD CUP 99 DATASET :

- **BackDoS:** Attaque par déni de service à l'encontre serveur web apache où un client demande une URL contenant de nombreuses Anti slash. Comme le serveur essaie de traiter ces demandes, il va ralentir et être incapable de traiter d'autres demandes.
- **LandDoS:** Un attaquant peut envoyer un paquet spécialement formatée qui peut causer le plantage d'un serveur distant, provoquant un déni de service. Certaines implémentations de TCP / IP sont vulnérables à des paquets qui sont fabriqués d'une manière particulière (un paquet SYN dans lequel l'adresse source et le port sont les mêmes que la destination c'est àdire, usurpée). Land est un outil d'attaque largement disponible qui exploite cette vulnérabilité.
- **NeptuneDoS:** Pour chaque connexion semi-ouverte faite à une machine, le serveur de tcpd ajoute un enregistrement à une structure de données décrivant toutes les connexions en cours. Cette structure de données est de taille finie, et il peut être fait un débordement intentionnellement en créant trop de connexions partiellement ouvertes. La structure de données de connexions semi-ouvertes sur le système de serveur de victime finira par combler; alors le système sera incapable d'accepter les nouvelles connexions entrantes jusqu'à ce que la table soit vidée. Normalement, il ya un délai d'attente associé à une connexion en attente, de sorte que les connexions semi-ouvertes finissent par expirer et le système de serveur de la victime seront guéris. Cependant, le système d'attaque peut simplement continuer à envoyer des paquets IP usurpées

demandant de nouvelles connexions plus rapides que le système de la victime peut expier les connexions en cours. Dans certains cas, le système peut épuiser la mémoire, crash, ou être rendu inopérant.

- **PodDoS:** Certains systèmes vont réagir de façon imprévisible lors de la réception des paquets IP surdimensionnés. Les réactions possibles comprennent le crash, la congélation et le redémarrage.
- **SmurfDoS:** Dans cette attaque, l'auteur envoie un ping IP (ou "echo mon message reviens à moi») demande au site de réception du paquet ping de préciser s'il sera diffusé à un nombre d'ordinateurs au sein du réseau local du site de réception. Le paquet indique également que la demande est d'un autre site, le site cible qui doit recevoir le déni de service. (Envoi d'un paquet avec l'adresse de retour de quelqu'un d'autre dans elle est appelée spoofing de l'adresse de retour). Le résultat sera un bon nombre de réponses ping inondations retour à l'innocent, l'hôte usurpée. Si l'inondation est assez grande, l'hôte usurpée ne sera plus en mesure de recevoir ou de distinguer le trafic réel.
- **TeardropDoS :** Ce type de déni de service exploite la façon dont le Protocol Internet (IP) nécessite un paquet qui est trop grand pour le prochain routeur et pour gérer la division en fragments. Le fragment de paquet identifie un décalage vers le début du premier paquet qui permet à l'ensemble du paquet à être remonté par le système de réception. Dans l'attaque de goutte d'eau, IP de l'attaquant met une valeur de décalage déroutant dans le second fragment ou plus. Si le système d'exploitation de réception n'a pas un plan de cette situation, il peut provoquer un crash du système.

5. Difficulté du choix des paramètres de l'algorithme génétique :

Le paramétrage est la principale difficulté des algorithmes génétiques, en effet, les différents paramètres doivent être optimisés pour chaque type de problème traité, en général les valeurs des paramètres sont réglés par étapes en fonction des résultats expérimentaux obtenus[20].

Il y a plusieurs types de paramètre à prendre en compte, il peut s'agir de la taille de la population initiale, du nombre de génération, ou bien du taux de mutation.

1) **Influence du nombre d'individus :**

Le nombre d'individus a une grande influence dans les algorithmes génétique, ainsi d'après différents tests effectués par cette étude, on constate que plus le nombre d'individus augment plus la qualité de la solution est meilleure, mais au bout d'un certain nombre d'individus l'évaluation reste constante. En revanche, le temps de calcul est plus important, en conclusion, il est important de fixer ce paramètre pour avoir le meilleur compromis entre la qualité de la solution et la rapidité de l'algorithme. En plus, un nombre d'individus trop petit pourrait conduire à un optimum local, de ce fait il est préférable d'avoir un nombre d'individus suffisamment grand pour avoir une grande diversité et éviter ce problème.

2) **Influence du nombre de générations :**

Les résultats obtenus par la même étude on montrés que le nombre de génération est directement proportionnel à la qualité de la solution, cependant, après un certain nombre de génération les solutions n'évoluent plus.

3) **Influence du taux de mutation :**

La mutation permet de préserver la diversité de la population, néanmoins, il est nécessaire de choisir un taux de mutation relativement faible pour éviter de tomber dans une recherche aléatoire, et ainsi conserver le principe de l'algorithme génétique qui est basé sur l'évolution des individus.

III. Conclusion

Les algorithmes génétiques grâce à leurs différentes opérations (sélection, croissement, mutation) qui s'inspirent principalement des mécanismes d'évolution de la nature, permettent de fournir rapidement des solutions proches de la solution optimale. Les résultats obtenus sont principalement influencés par la fonction objectif qui est au centre de tous les calculs, le choix du codage et les différentes implémentations des opérations génétiques permettent de créer plusieurs variantes de l'algorithme génétique, ces variantes combinées avec les bons paramètres vont déterminer le meilleur compromis entre la qualité de la solution et la rapidité de l'algorithme.

Chapitre IV
Conception et Réalisation

I. Introduction

Au cours de ce chapitre, on a évalué l'exécution de l'algorithme génétique sur un dataset en vue d'effectuer un apprentissage sur les différents champs d'une population de chromosomes représentant le trafic réseau dans le but de modéliser le trafic normal et d'y interpréter un ensemble de règles permettant de modéliser ce trafic normal. Les règles générées permettent de détecter tout comportement déviant de la normale.

II. Outil de développement :

1. Eclipse

Eclipse **IDE** est un environnement de développement intégré libre (le terme*Eclipse* désigne également le projet correspondant, lancé par IBM) extensible, universel et polyvalent, permettant potentiellement de créer des projets de développement mettant en œuvre n'importe quel langage de programmation. Eclipse IDE est principalement écrit en Java (à l'aide de la bibliothèque graphique SWT, d'IBM), et ce langage, grâce à des bibliothèques spécifiques, est également utilisé pour écrire des extensions.

La spécificité d'Eclipse IDE vient du fait de son architecture totalement développée autour de la notion de plug-in (en conformité avec la norme OSGi) : toutes les fonctionnalités de cet atelier logiciel sont développées en tant que plug-in.

Plusieurs logiciels commerciaux sont basés sur ce logiciel libre, comme par exemple IBMLotus Notes 8, IBM Symphony ou Websphere Studio Application Developer.

2. Maven

Maven est un outil de construction de projets (build) open source développé par la fondation Apache, initialement pour les besoins du projet Jakarta Turbine. Il permet de faciliter et d'automatiser certaines tâches de la gestion d'un projet Java

Il permet de:

- d'automatiser certaines tâches : compilation, tests unitaires et déploiement des applications qui composent le projet
- de gérer des dépendances vis à vis des bibliothèques nécessaires au projet

- de générer des documentations concernant le projet

III. Langage d'implémentation utilisé :

Le langage **Java** est un langage de programmation informatique orienté objet créé par James Gosling et Patrick Naughton, employés de Sun Microsystems, avec le soutien de Bill Joy (cofondateur de Sun Microsystems en 1982), présenté officiellement le 23 mai 1995 au SunWorld.

La société Sun a été ensuite rachetée en 2009 par la société Oracle qui détient et maintient désormais Java.

La particularité et l'objectif central de Java est que les logiciels écrits dans ce langage doivent être très facilement portables sur plusieurs systèmes d'exploitation tels que UNIX, Windows, Mac OS ou GNU/Linux, avec peu ou pas de modifications. Pour cela, divers plateformes et frameworks associés visent à guider, sinon garantir, cette portabilité des applications développées en Java.

IV. Les interfaces de notre application :

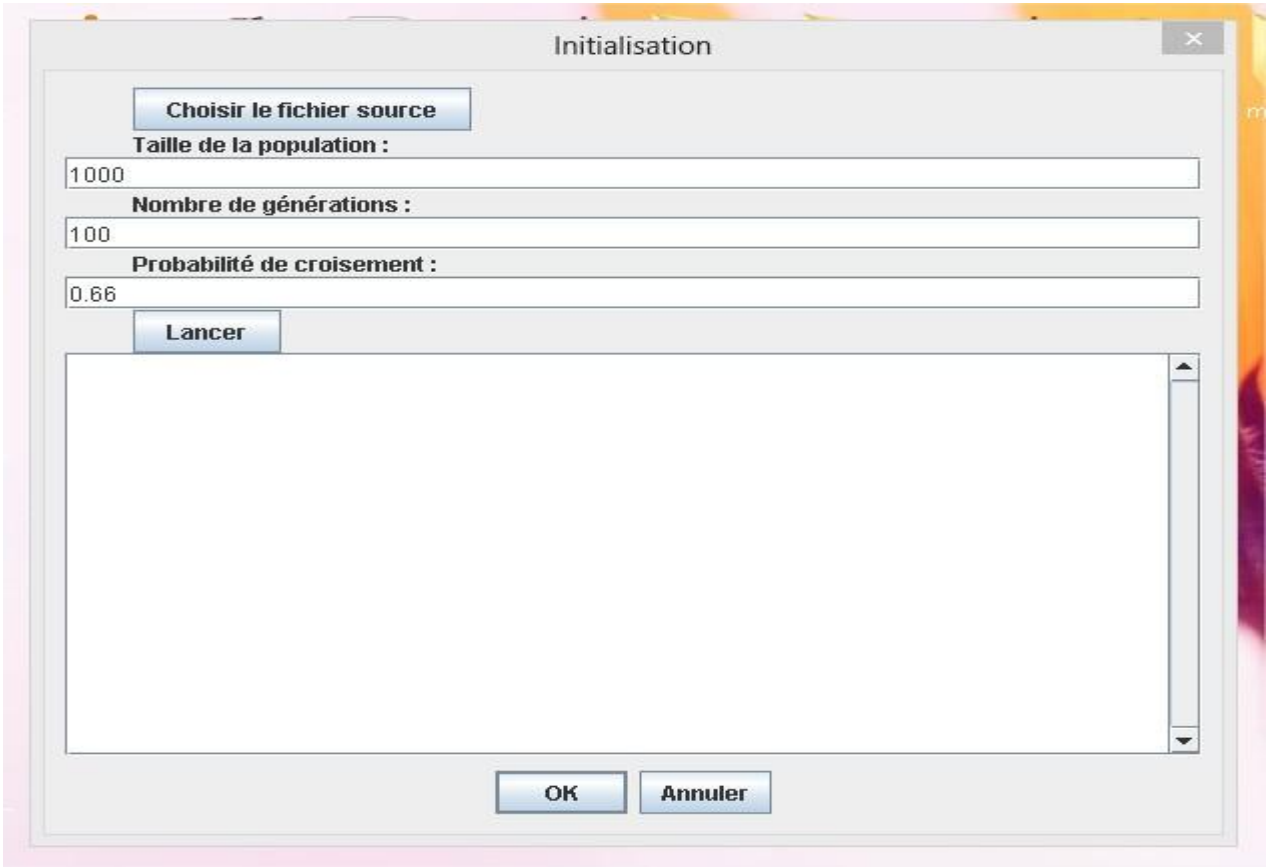


Figure 1 : Interface principale

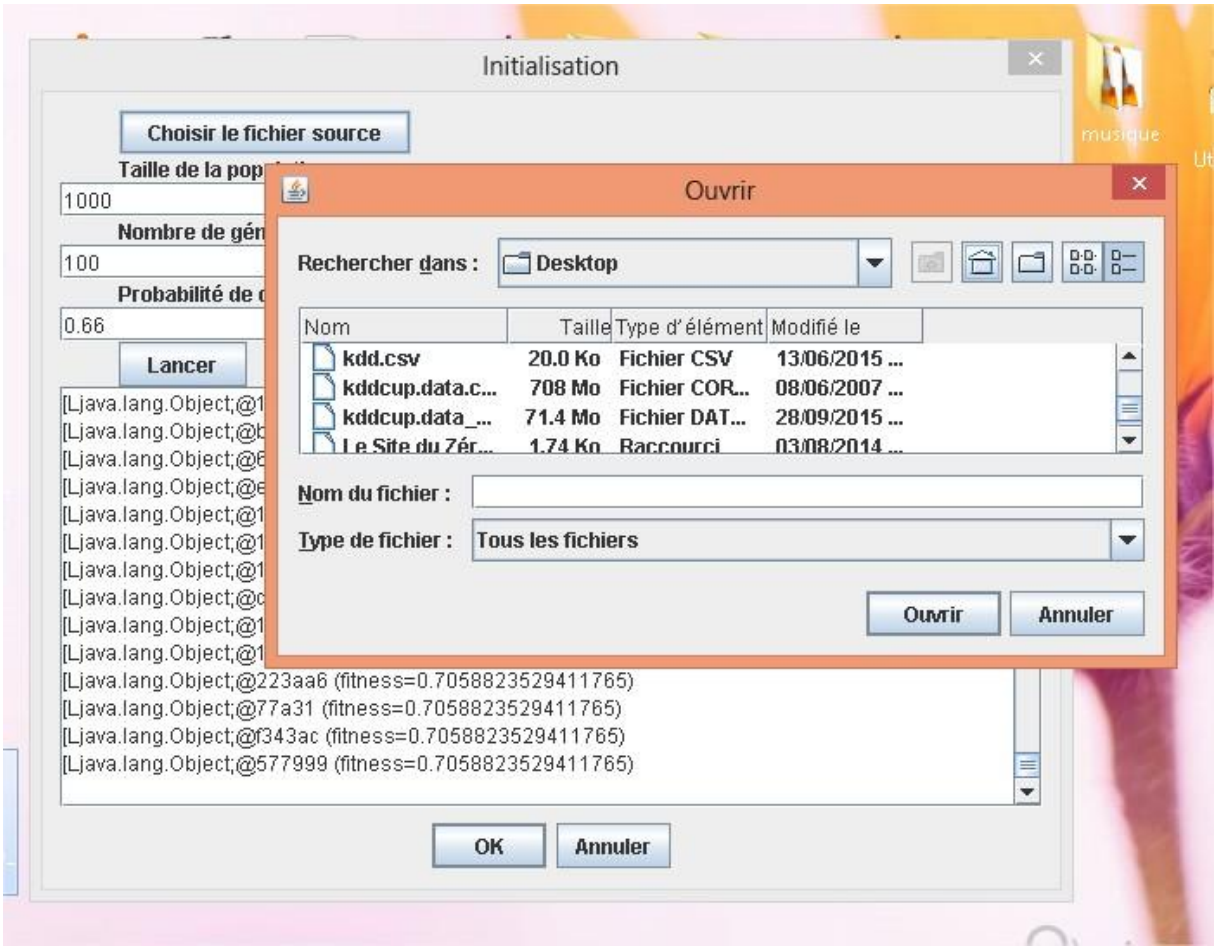


Figure 2 : choix du fichier source

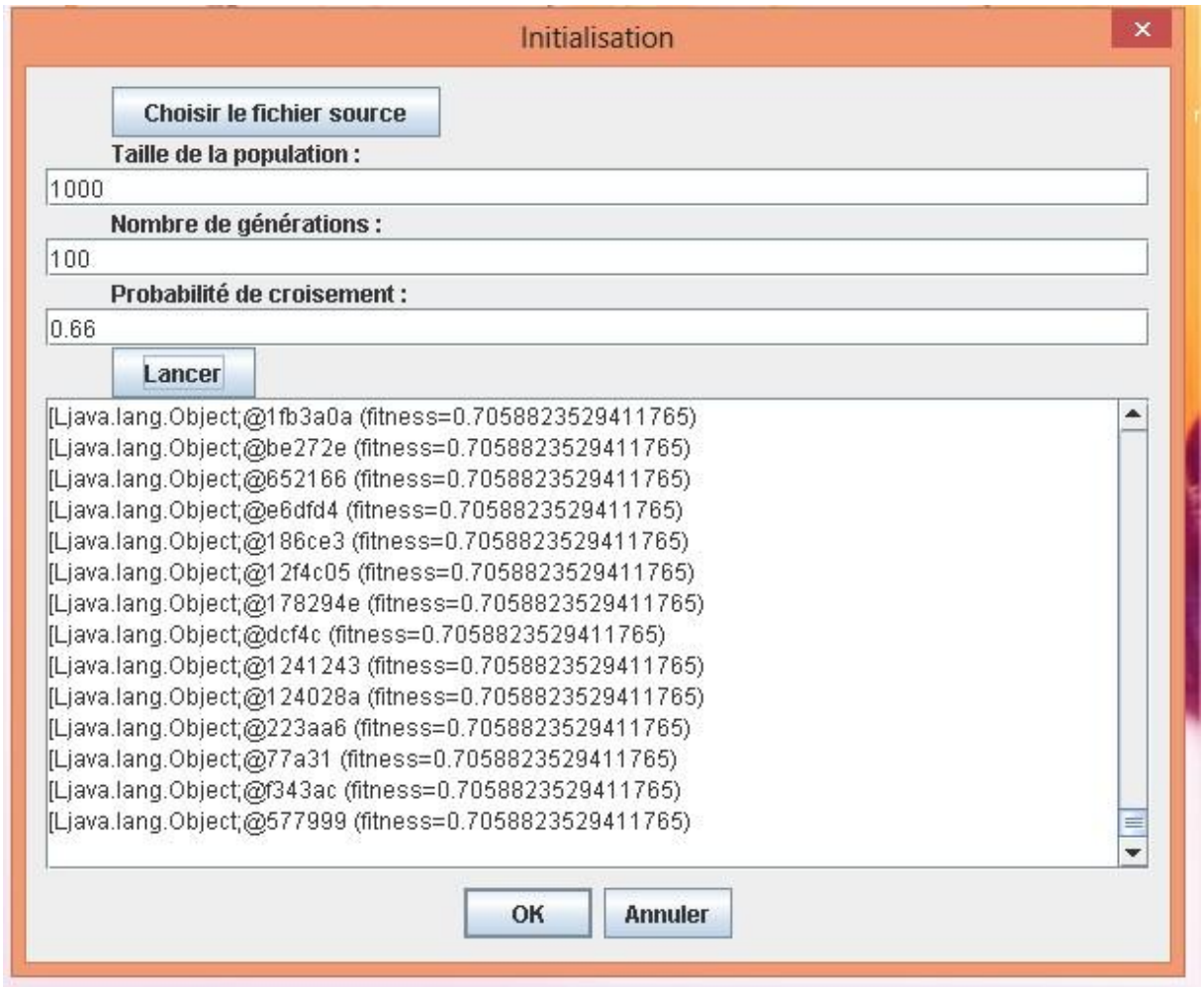


Figure 3 : les résultats

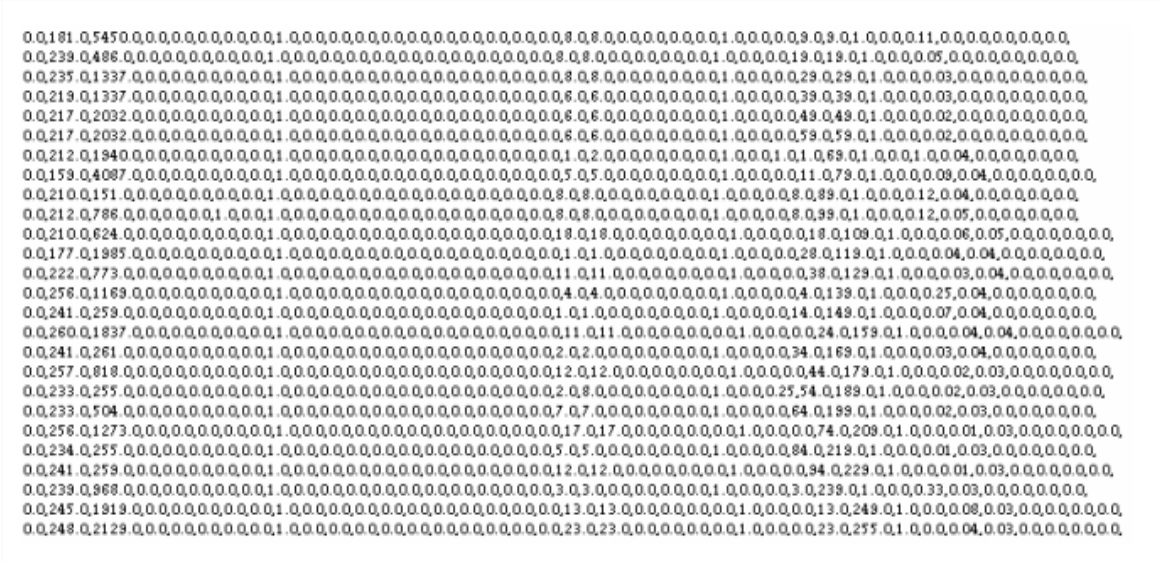


Figure 4 : la kdd 99 utilisée

Dataset	normal	probe	dos	u2r	r2l	Total
Train ("kddcup.data_10_percent")	97280	4107	391458	52	1124	494021
Test ("corrected")	60593	4166	229853	228	16189	311029

Figure 5 : La distribution des types d'attaque dans la DataSet

V. Conclusion :

Au cours de ce chapitre, on a évalué l'exécution de l'algorithme génétique sur un dataset en vue d'effectuer un apprentissage sur les différents champs d'une population de chromosomes représentant le trafic réseau dans le but de modéliser le trafic normal et d'y interpréter un ensemble de règles permettant de modéliser ce trafic normal. Les règles générées permettent de détecter tout comportement déviant de la normale.

Conclusion Générale

La défense en profondeur des réseaux passe par une bonne stratégie préventive pour penser ses réseaux et leurs interconnexions de façon sécurisée. Cette approche doit être complétée une fois le réseau en opération pour permettre de détecter des anomalies qui peuvent être révélatrices.

Ce travail nous a permis de nous familiariser avec les Systèmes de Détection d'Intrusion Réseau en appliquant un algorithme génétique sur un historique d'événements réseau et donc d'y effectuer une détection offline.

Comme perspective, on espère que l'ensemble de règles générées puissent être appliquées à un trafic online et que cet ensemble de règles puisse être enrichies au fur et à mesure de l'exécution de l'Algorithme Génétique

Liste des Figures

Chapitre I

Chapitre II

Chapitre III

Chapitre IV

Liste des tableaux

Liste des tableaux

Webliographie

Webliographie

[1]- http://ww2.ac-poitiers.fr/matice/spip.php?article156

[2]- http://aries.serge.free.fr/index.php?page=content/GC/SA8

[3]-https://fr.wikipedia.org/wiki/Repr%C3%A9sentation_des_connaissances

[4]-http://www.grappa.univ-lille3.fr/~champavere/Enseignement/0809/l3miashs/ia/rc-ws.pdf

[5]-http://people.rennes.inria.fr/Loic.Helouet/Sujets/SecuWebRules-2.pdf

[6]-http://www.ai-ps.info/share/Definition-Virus%20informatique-theme.pdf

[7]- http://dibai.free.fr/Securite/Securite/Docs/Mg1f.pdf

[8]- http://www.academiepro.com/uploads/cours/2015_09_16_cours_securite_v2.pdf

[9]-http://www.hoffmanncorporation.com/stoky/micro/dos/PGP-VN-ED-QK.pdf

[10]-http://www.lsis.org/espinasseb/Supports/DWDM-2013/8-IntroFouille-2009.pdf

[11]-http://www.memoireonline.com/10/13/7549/m_Une-contribution-du-datamining-la-segmentation-du-march-et-au-ciblage-des-offres--l-aide2.html

[12]- http://www.lifl.fr/~talbi/Cours-Data-Mining.pdf

[13]- http://depot-e.uqtr.ca/1423/1/030000495.pdf

[14]-http://liyun.free.fr/XP/IDS3.pdf

[15]-http://toubkal.imist.ma/bitstream/handle/123456789/9330/THESE-KARTIT.pdf?sequence=3

[16]- http://www.linuxfocus.org/Francais/May2003/article292.shtml

[17]- Lalaina KUHN, « VolP & Security :IPS» support de cours, Ecole d'Ingénieurs du Canton de Vaud.

[18]- http://www.futura-sciences.com/magazines/high-tech/infos/dico/d/internet-deni-service-2433/

[19]- http://produ.chez.com/badro/

[20]- http://dspace.univ-tlemcen.dz/bitstream/112/6004/1/Optimisation-de-la-QOS-dans-un-reseau-de-radio-cognitive-en-utilisant-les-algorithmes-genetiques.pdf

Résumé

Un système de détection d'intrusion (IDS) est un mécanisme écoutant le trafic réseau de manière furtive afin de repérer des activités anormales ou suspectes et permettant aussi d'avoir une action de prévention sur les risques d'intrusions. Les méthodes de détection d'intrusions reposent essentiellement sur deux approches : l'approche comportementale et l'approche par signatures. Chacune des deux présente des points forts, mais aussi des faiblesses qui sont les faux positifs et les faux négatifs. Notre objectif à était l'implémentation d'un algorithme génétique dans la détection d'intrusion réseau d'un trafic offline en utilisant l'approche comportementale optimisée.

En outre, il est important de noter que le risque nul d'être piraté n'existe pas et il faut s'avoir s'appuyer au mieux sur les outils (nouvellement) disponibles afin de tendre vers cet idéal.