CHAPITRE 2

ETAT DE L’ART

Introduction :

Dans notre thème, nous nous intéressons à la classification hiérarchique pour la détection d'intrusions dont l'objectif est l'application de cet algorithme pour détecter les éventuelles intrusions dans les réseaux locaux.

Selon la documentation que nous avons pu consulter, Nous avons constaté une pénurie de l'utilisation de cet algorithme dans ce domaine.

Dans ce chapitre nous allons présenter une collection des travaux réalisés dans le domaine de data mining pour la détection d'intrusion.

La section 1 présente les IDS basé sur la technique de Machine à vecteurs de support, section 2 et 3 présentent l'utilisation des algorithmes du data mining pour les IDS, section 4 et 5 présentent respectivement la construction des modèles pour la détection d'intrusion et la détection d'intrusion dans les réseaux sans fils, et la dernière section présente D'autres domaines d'application et méthodes pour la détection d'intrusion.

1. Les IDS basé sur la technique de Machine à vecteurs de support (SVM) :

La première étude :

En 2002, S, Mukkamala, G, Janoski, et A, Sung [22] ont décrit des approches à la détection d'intrusion et la réduction des données de vérification à l'aide de machines à vecteurs de supports et réseaux de neurones, en utilisant l'ensemble de données de référence de la KDD (Knowledge Discovery et Data Mining) pour l'évaluation de la détection. L'approche est de former les machines à vecteurs de supports et réseaux de neurones pour apprendre les modèles de comportement normal et d'attaque. En outre, ils utilisent uniquement le (13 de 41) caractéristiques les plus importantes de données. Les deux SVM et les réseaux neuronaux offrent des performances très précises (99% et plus), avec SVMs régulièrement (soit avec 41 ou 13 caractéristiques) surpassent les réseaux neuronaux, en termes de temps de formation et la précision de détection.

La deuxième étude :

Dans [19], L Khan, M Awad et B Thuraisingham [2007] ont proposé une nouvelle approche nommée CTSVM (Clustering Trees based on SVM) pour la détection d'anomalie, c'est une solution évolutive pour détecter les anomalies réseau en utilisant les machines à vecteurs de support pour la classification et l'algorithme DGSOT (Dynamically Growing Self-Organizing Tree) pour le regroupement hiérarchique à travers une combinaison entre ces deux algorithmes.

Les principales contributions de ce travail sont les suivants : d'abord, pour réduire le temps de formation des SVM, ils ont proposé une nouvelle technique de sélection de vecteurs de support à l'aide de clustering. Deuxièmement, ils ont montré analytiquement le degré auquel leur approche est asymptotiquement plus rapide que pure SVM, et valide cette demande avec des résultats expérimentaux. Enfin, ils ont comparé leurs approches avec la sélection aléatoire, et Rocchio Bundling sur un ensemble de données de référence (1998 DARPA), et de démontrer des résultats impressionnants en termes de temps de formation, taux de faux positifs, taux de faux négatifs, et la précision de détection.

La troisième étude :

Certains chercheurs [Peter Scherer et al, 2011] [23] ont examiné les possibilités d'application des algorithmes de clustering et machines à vecteurs de supports (SVM) pour une utilisation dans l'IDS. Ils ont utilisé K-means, Farthest first, l'algorithme COBWEB et la technique SVM de type 1, connu aussi comme C-SVM. Pour chaque algorithme utilisé est affiché le taux de réussite à chaque classe d'attaque. Le meilleur taux de réussite moyen a algorithme SVM, plus de 99% (le meilleur de tous est l'algorithme SVM qui utilise le noyau RBF, il a un taux de réussite de 99,722%). Le taux de réussite moyen des autres algorithmes se situait entre 91,228% et 98,998%.

La quatrième étude :

Cette étude [21] [Bisen, M et Dubey, A, 2015] a proposé, un système de détection d'intrusion basée sur SVM (Support Vector Machine) qui combine l'algorithme génétique, l'algorithme de classification hiérarchique et la technique de SVM. L'algorithme génétique et l'algorithme BIRCH de classification hiérarchique sont utilisé pour le prétraitement de l'ensemble de données, pour le SVM pourrait les classer avec plus de précision. Le système proposé réduit le temps de formation et de réaliser également une meilleure classification des différents types d'attaques, L'ensemble de données KDD CUP 1999 a été utilisé pour évaluer ce système. Il a montré une meilleure performance dans la détection de différentes attaques ainsi que d'augmenter la précision des attaques probe et U2R.

2. Data mining et les systèmes de détection d'intrusion :

La première étude :

Le système de détection d'intrusion (IDS) est utilisé pour protéger la confidentialité des données, l'intégrité et la disponibilité du système contre les différents types d'attaques. Data mining est un artifice efficace appliqué à la détection d'intrusion pour déterminer un nouveau

plan à partir des données de réseau massives aussi bien qu'il permet de réduire la pression des compilations manuelles des modèles de comportement normal et anormal. Dans cet article [1] et [2], de nombreuses techniques data mining ont été proposées pour améliorer le taux de détection du système de détection d'intrusion tels que, les méthodes de partitionnement, les méthodes hiérarchiques, modèle basé sur les méthodes de classification et de leurs divers types.

La deuxième étude :

Dans [25] SELIYA Naeem et al [2007] ont décrit les méthodes d'apprentissage non supervisés pour la détection d'intrusion réseau. Ils ont étudié plusieurs algorithmes de classification non supervisée basée sur un centroïde pour la détection d'intrusion, et proposait une heuristique d'auto-étiquetage simple mais efficace pour détecter les clusters normaux et d'attaques de données d'audit de trafic réseau. Les algorithmes de clustering étudiés incluent, k-means, Mixture-Of-Spherical Gaussians, Self-Organizing Map, et Neural-gaz. L'étude empirique constituée de différents algorithmes de clustring est effectué avec les données de trafic de réseau obtenues à partir du projet de détection d'intrusion du DARPA 1998, Et l'analyse comparative et l'évaluation des algorithmes de clustering donnent des taux de détection d'intrusions raisonnables. En outre, la méthode heuristique d'auto-étiquetage a démontré l'utilité et la faisabilité des méthodes de clustring par la détection de nouveaux types d'attaques et l'aide des techniques de classification dans la réalisation de meilleures performances de détection d'intrusion.

La troisième étude :

Certains chercheurs [NIEVES, Jose F. et JIAO, Yu Cathy, 2009] [27] ont examiné les différentes méthodes de détection d'intrusion réseau à travers une étude comparative avec plus d'emphase sur les méthodes d'apprentissage non supervisé pour la détection d'anomalies. L'algorithme K-mean a été choisi pour évaluer la performance de la méthode d'apprentissage non supervisé à l'aide de la Coupe Kdd 1999 le jeu de données réseau. Ils ont utilisé des méthodes de clustering pour la détection d'anomalies dans la détection d'intrusion réseau, et ont atteint un taux élevé de détection des attaques (y compris les attaques inédites) tout en gardant un taux de fausse alarme faible sans avoir besoin de passer par la procédure d'étiquetage.

3. Optimisation de l'algorithme de k-mean :

Afin d'améliorer l'algorithme de clustring « k-mean » RANJAN, Ravi et SAHOO, G [2014] [30] ont proposé une nouvelle façon de détection d'intrusion utilisant l'algorithme de

clustering k-medoids avec certaines modifications de celui-ci. L'algorithme spécifié un nouveau mode de sélection de médoïde initiale et prouvé d'être mieux que K-means pour la détection d'anomalies. L'approche proposée est d'avoir de nombreux avantages sur l'algorithme existant qui remédie essentiellement les inconvénients de la dépendance sur les centroïdes initiales, la dépendance du nombre de clusters non pertinents. L'algorithme a été capable de trier ces problèmes et a été en mesure de fournir des taux de détection élevés et taux de faux négatif faible.

4. La construction des modèles pour la détection d'intrusion :

La première étude :

Certains chercheurs [LEE, Wenke et STOLFO, Salvatore J, 2000] [3] ont proposé un framework, MADAM ID, pour construire les modèles automatisés de détection d'intrusion. L'idée principale est d'appliquer d'abord les programmes du data mining pour vérifier les données pour calculer les motifs fréquents, extraire des caractéristiques, puis utiliser des algorithmes de classification pour calculer les modèles de détection. Pour faciliter l'adaptabilité et l'extensibilité, ils ont proposé aussi l'utilisation de méta-apprentissage comme un moyen de construire un modèle combiné qui intègre des preuves à partir de modèles de base multiples. Ils ont étendu les règles d'association de base et des épisodes d'algorithmes fréquents pour tenir compte des exigences particulières dans l'analyse des données d'audit. leurs expériences ont montré que les motifs fréquents extraits à partir des données d'audit peuvent être utilisés comme modèles de détection d'anomalies utilisateur fiables, et de lignes directrices pour la sélection des fonctions statistiques temporelles pour construire des modèles de classification efficaces. Ils ont discuté également brièvement leur expérience dans la conversion des modèles de détection produits par les programmes du data mining hors ligne à des modules en temps réel d'IDSs existant.

Les résultats du programme d'évaluation de la DARPA 1998 ont montré que leurs modèles de détection effectués sont les meilleurs systèmes construits en utilisant les manuels approches d'ingénierie des connaissances.

La deuxième étude :

Dans [20], Mrutyunjaya Panda et Manas Ranjan Patra [2009] ont pour but de présenter une méthodologie pour reconnaître les attaques au cours des activités normales dans un système. Ils ont proposé une nouvelle classification via un algorithme de classification nommé sequential information bottleneck (sIB), pour construire un modèle efficace de détection d'intrusion de réseau. Ils ont comparé leur méthode proposée avec d'autres

algorithmes de clustering comme X-Means, Farthest First, Filtered clusters, DBSCAN, K-Means, et EM (Expectation Maximization) en utilisant le sous-ensemble des données de référence la Coupe KDD 1999, et ont trouvé des résultats montrent que leur méthode est efficace en termes de précision de détection, le faible taux de faux positifs par rapport aux autres méthodes existantes.

La troisième étude :

Dans [24] CHEN, Zhenguo et ZHU, Dongmei [2011] ont choisi l'algorithme de classification basé sur la structure hiérarchique, pour former le profil de comportement normal sur les enregistrements d'audit et d'ajuster le profil en temps opportun que le comportement du programme a changé, ils ont obtenu des informations des appels système et d'utiliser la méthode de pondération de fréquence tf-idf pour le prétraitement. Lors du test de leur processus de formation en DARPA ils ont obtenu le taux moyen de 6,33% de faux positifs et le taux de détection était de 92,03%. Donc un taux de détection plus élevé et un taux de faux positifs d'ensemble de données DARPA est faible. Ce qui montre que l'expérience est faisable et efficace.

5. La détection d'intrusion dans les réseaux sans fils :

La première étude :

La détection d'intrusion dans les MANET est difficile parce que ces réseaux changent leurs topologies dynamiquement, manque de points de concentration où le trafic agrégé peut-être analysé, utiliser des protocoles d'infrastructure qui sont sensibles à la manipulation, et compter sur les communications sans fil, intermittents bruyants. Cette étude [31] [STERNE, Daniel, et al, 2005] présente une architecture distribuée de détection d'intrusion coopérative qui répond à ces défis tout en facilitant la détection précise des attaques MANET spécifiques et conventionnelles. L'architecture est organisée comme une hiérarchie dynamique dans lequel des données de détection sont acquise à feuilles et sont progressivement agrégées, réduits et analysés comme ils coulent vers le haut en direction de la racine. La hiérarchie est automatiquement configurée au besoin en utilisant des techniques de regroupement dans lequel les têtes de cluster sont sélectionnées en fonction de la topologie et d'autres critères. L'utilité potentielle de l'architecture a été illustré par l'intermédiaire d'une série de scénarios d'utilisation, y compris la détection délibérée du chute de paquets, les attaques contre les protocoles MANET de routage, et les attaques sur le réseau conventionnel et protocoles des couches supérieures, y compris les attaques avec des cibles ou des sources distribuées.

La deuxième étude :

En 2010 Certains chercheurs [MAMUN, Mohammad Saiful Islam et KABIR, AFM Sultanul] ont proposé [32] une nouvelle architecture de l'IDS pour réseaux de capteurs ad hoc basée sur la conception de superposition hiérarchique (hierarchical overlay design (HOD)). Ils ont proposé un mécanisme de réponse aussi selon l'architecture proposée. Leur conception de l'IDS améliore d'autres designs dans la façon dont il distribue la tâche totale de détection d'intrusion. Leur modèle découple le travail total de détection d'intrusion dans une hiérarchie à quatre niveaux qui se traduit par une structure hautement économie d'énergie, et qui améliore l'évolutivité du réseau à grande zone géographique et utiliser les deux techniques de détections d'intrusion (Signature et anomalie).ils ont divisé chaque zone de nœuds de capteurs dans la région hexagonale (comme des cellules GSM). Les nœuds de capteurs dans chacune de la zone hexagonale sont surveillés par un nœud de cluster, Chaque nœud de cluster est ensuite contrôlé par un nœud régional, à son tour, les nœuds régionaux seront contrôlés et surveillés par la station de base. En raison du modèle hiérarchique, le système de détection fonctionne de manière très structurée et capable de détecter toute intrusion efficacement. Ainsi est capable d'identifier les attaques connues ainsi que les attaques inconnues.

La troisième étude :

La conception proposée dans cette étude [33] [JADIDOLESLAMY, Hossein, *et* al, 2011] présente une architecture de détection d'intrusion complète (IDA), et un modèle distribué pour la détection d'intrusion dans WSN (Wireless Sensor Networks). La principale contribution de cette architecture est sa structure hiérarchique, c'est-à-dire qu'il est conçu et applicable, dans un, deux ou trois niveaux y compris les SIDSs (sur les nœuds de capteurs), CIDSs (sur le cluster de têtes) et le WSNIDS (sur le serveur central), conformément au domaine d'application et son niveau de sécurité requis. Cette architecture a des propriétés principales telles que la robustesse, l'évolutivité, en réponse, l'extensibilité et la correspondance incrémentale le long avec les changements de l'environnement et de ses nouvelles conditions. En outre, l'IDA se concentre sur l'intégration des outils accessibles dans la zone de sécurité des réseaux informatiques (comme IDSs).

6. D'autres domaines d'application et méthodes pour la détection d'intrusion:

La première étude :

Cette étude [TÖLLE, Jens, NIGGEMANN, Oliver, 2000] [28] présente une description d'un système de soutien à la détection des intrusions et des anomalies du réseau en analysant

et en visualisant les flux de trafic dans les réseaux informatiques. Ce système se compose d'une unité de collecte de paquets, une unité de regroupement et de construction de graphe, le module de visualisation et le module de génération d'événement. Il a indiqué que le regroupement graphique offre des modèles qui permettent de visualiser et détecter automatiquement les anomalies dans le trafic réseau. Cette approche de la détection d'intrusion aide à la détection de faux positifs.

La deuxième étude :

Dans [26] BRAHMI, I., BEN YAHIA, S., et SLIMAI, Y [2008] ont proposé une nouvelle approche de détection d'intrusion qui se base sur la stratégie d'extraction de règles associatives génériques de classification. Ils ont ensuite implémenté cette stratégie dans un nouveau système de détection d'intrusion appelé IDS-GARC. Ce système utilise une approche comportementale, puisqu'il est capable de détecter de nouvelles attaques et de caractériser ainsi les nouveaux comportements. Les expérimentations réalisées sur des bases de données d'audit ont montré que le système IDS-GARC est plus efficace et robuste, puisqu'il génère peu de fausses alarmes et détecte plus de vraies intrusions.

La troisième étude :

Cette étude [4] [HILAS, Constantinos S. et MASTOROCOSTAS, 2008] appartient au domaine de télécommunication servi à résoudre les problèmes de la détection des fraudes de télécommunications en appliquant les techniques d'apprentissage supervisé et non supervisé. Cinq modèles d'utilisateurs différents sont comparés au moyen de deux techniques d'apprentissage, à savoir le perceptron multicouche (un classifieur linéaire de type réseau neuronal) et la classification ascendante hiérarchique. Les profils sont utilisés en tant que méthode de caractérisation de l'utilisateur, afin de discriminer l'usage légitime d'un usage frauduleux dans un environnement de télécommunications. Toutes les données proviennent de comptes d'utilisateurs réels floués dans un réseau de télécommunications. Le classificateur de réseau de neurones a très bien performé sur le problème, donner des mesures de taux de TP (true positive) mieux que 80% de FP (false positive) inférieure à 2%, Ainsi, le regroupement hiérarchique a été appliqué sur les mêmes ensembles de données en utilisant deux mesures de distance différentes, à savoir la distance euclidienne et la corrélation. La conclusion dominante est que le pourcentage le plus élevé de la classification correcte se produit lorsque la corrélation est utilisée en tant que mesure de la distance. Toutes les deux approches ont donné de meilleurs résultats avec Profile1. Cela nous a fourni assez de confiance pour utiliser Profile1 comme le comportement d'utilisateur approprié.

7. Les critiques :

Nous allons proposer quelques remarques sur les travaux ci-dessus, Elles sont résumées dans le tableau 1.

Thèmes	Auteurs/ Années	Commentaires
Intrusion detection using neural networks and support vector machines	[MUKKAMALA, Srinivas et al, 2002]	- L'utilisation massive des SVM pour la détection d'intrusion.
A new intrusion detection system using support vector machines and hierarchical clustering.	[KHAN, Latifur et al, 2007]	- Le SVM est l'un des algorithmes de classification le plus de succès dans le domaine de l'exploration de données, mais son temps de formation limite son utilisation.
Using svm and clustering algorithms in ids systems	[SCHERER, Peter et al, 2011]	
An Intrusion Detection System Based On Support Vector Machine Using Hierarchical Clustering And Genetic Algorithm	[BISEN, Minakshi et al, 2015]	- La combinaison de technique SVM avec d'autres algorithmes de classification permet d'améliorer le temps de formation de SVM.
Clustering-based network intrusion detection	[ZHONG, Shi et al, 2007]	- Une attention considérable a été accordée à des approches d'exploration de données pour aborder la sécurité du réseau.
Data clustering for anomaly detection in network intrusion detection	[NIEVES, Jose F,2009]	- La détection d'intrusion non supervisée est plus appropriée pour la détection d'anomalie que les méthodes de détection d'intrusion basée sur la classification.
Classification, clustering and intrusion detection system	[JOSHI, Manish, 2012]	
A Study of Network Intrusion Detection by Applying Clustering Techniques	[Subaira.A.S et al, 2013]	- Les techniques de data mining aident les IDS à construire des comportements normaux et d'attaques.
A new clustering approach for anomaly intrusion detection	[RANJAN, Ravi et al, 2014]	- L'algorithme proposé dans cette étude élimine les inconvénients de l'algorithme K-means.
	[LEE, Wenke et al,	- ces approche réduit

A framework for constructing features and models for intrusion detection systems	2000]	considérablement la nécessité d'analyser manuellement et encoder des motifs d'intrusion. - La précision d'un modèle de classification dépend directement de l'ensemble des fonctionnalités fournies dans les données de formation. - Lorsque le profil de la normalité est formé, il est utilisé pour détecter les données inconnues.
A novel classification via clustering method for anomaly based network intrusion detection system	[PANDA, Mrutyunjaya et al, 2009]	
Hierarchic Clustering Algorithm Used for Anomaly Detecting	[CHEN, Zhenguo et al, 2011]	
A general cooperative intrusion detection architecture for MANETs	[STERNE, Daniel et al, 2005]	- La détection d'intrusion dans les MANET est difficile parce que ces réseaux changent leurs topologies dynamiquement. - Ces approches, ont porté presque exclusivement sur la détection des comportements malveillants par rapport aux protocoles de routage MANET. - Ajouter IDS, comme une nouvelle couche de défense de sécurité
Hierarchical design based intrusion detection system for wireless ad hoc sensor network	[MAMUN, Mohammad et al, 2010]	
A high-level architecture for intrusion detection on heterogeneous wireless sensor networks: hierarchical, scalable and dynamic reconfigurable	[JADIDOLESLAMY, Hossein, *et al, 2011*]	
Supporting intrusion detection by graph clustering and graph drawing	[TÖLLE, Jens et all, 2000]	- La Visualisation de la structure du trafic aidé à identifier le type et l'emplacement des anomalies du réseau.
Détection d'Intrusions Basée sur les Règles Associatives Génériques de Classification	[BRAHMI, I et al, 2008]	- La classification associative permet la concentration sur les règles de détection utiles.
An application of supervised and unsupervised learning approaches to telecommunications fraud detection	[HILAS et al, 2008]	- Utilisation de l'algorithme CAH, donne des meilleurs résultats

Tableau 2.1 : synthèse sur les études réalisées dans le domaine

Conclusion :

ce chapitre présente dans la section 1, les IDS basé sur la technique de Machine à vecteurs de support, section 2 et 3 détaillent l'utilisation des algorithmes du data mining pour les IDS, section 4 et 5 présentent respectivement la construction des modèles pour la détection d'intrusion et la détection d'intrusion dans les réseaux sans fils, et la dernière section présente D'autres domaines d'application et méthodes pour la détection d'intrusion.

D'après cette étude bibliographique, nous pouvons constater la grande utilisation de classification et algorithmes de clustring pour la détection d'intrusion. Et les résultats étonnants des études proposés prouvent le succès de ces études.

Dans le chapitre suivant, nous allons expliquer la conception de la méthode proposée.