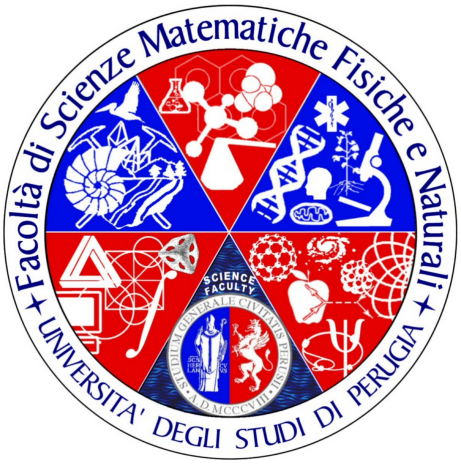
UNIVERSITÀ DEGLI STUDI DI PERUGIA

Facoltà di Scienze Matematiche, Fisiche e Naturali

Corso di laurea specialistica in Informatica



Corso di:
Sicurezza

Relazione di esame:
Bayesian Filters
(Applicazione dei filtri Bayesiani ad antispam ed IDS)

Studente:
Diego Russo - 231423
me@diegor.it
Flavio Vella - 233062
flavio.vella@dmi.unipg.it

Professore:
Prof. Stefano Bistarelli

Anno Accademico 2010/2011

Indice

Elenco delle figure

Sommario

AntiSpam ed **Intrusion Detection System** (IDS) svolgono un ruolo fondamentale nell'infrastruttura di rete. Queste due componenti, sebbene funzionalmente distanti, utilizzano strategie direttamente derivate dall'applicazione del Teorema di Bayes e delle tecniche di classificazione basate su di esso (i.e. classificatori bayesiani naive).

In questo elaborato verrà mostrato il Teorema di Bayes definendolo e spiegandone il significato; verranno mostrate le tecniche di classificazione basate su questo teorema ed infine verranno presentate due applicazioni che ne fanno uso, quali filtri antispam ed IDS.

1 Introduzione

L'utilizzo delle mail nella comunicazione tra soggetti e l'utilizzo di servizi telematici è oramai consuetudine. Con la diffusione della banda larga e dell'accesso ad internet un numero sempre più elevato di soggetti o entità fa uso di queste tecnologie, sia per motivi lavorativi che ludici. Contestualmente **l'utilizzo improprio di tali tecnologie è pratica altrettanto diffusa** (circa il 80-90% della mail mondiale è spam [2]).

Da un lato l'utilizzo della mail come strumento di pubblicità di massa non richiesta, dall'altro i tentativi di accesso non autorizzati a sistemi informativi per sottrarre informazioni, richiedono strumenti che non limitino la libertà o l'esigenze del soggetto autorizzato ed al tempo stesso tutelino il soggetto ed il sistema informativo da attività non autorizzate, illecite o criminose.

I software maggiormente utilizzati per contrastare i fenomeni sopracitati sono rispettivamente antispam ed Intrusion Detection System (IDS). L'efficacia di questi software dipende dall'**accuratezza delle policy** e dei parametri di configurazione applicati dall'amministratore di sistema. Generalmente la fruibilità di un servizio aumenta applicando policy e controlli meno restrittivi.

Tuttavia l'utilizzo di policy poco restrittive o l'applicazione di regole più stringenti potrebbero causare una **diminuzione del QoS (Quality of Service)**. Inoltre le tipologie di utenti che accedono al servizio sono varie, quindi gli strumenti a tutela del servizio stesso e del sistema oltre a garantire il corretto funzionamento devono prevedere **regole auto adattative** discriminando nella maniera più precisa possibile quali azioni sono lecite e quali non lo sono. Gli strumenti tradizionali applicano **regole statiche** impostate sullo studio a priori dell'uso del servizio e necessitano di una fase di *tuning* delle policy non sempre semplice ed efficace.

Recentemente sono stati sviluppati strumenti per realizzare policy auto adattative che minimizzano le restrizioni di utilizzo ed accesso del servizio

o al sistema, garantendone contemporaneamente la sicurezza e il corretto funzionamento. Esistono diverse strategie per la realizzazione di software e strumenti dinamici ed una di queste è basata sull'applicazione diretta del **Teorema di Bayes** della probabilità condizionata [3]; le applicazioni più note riguardano i filtri e classificatori bayesiani.

Nell'ultimo decennio [9] tali strumenti sono stati introdotti in software antispam ed IDS. In linea di principio, per quanto riguarda i software antispam, l'applicazione valuta che il messaggio in ingresso sia spam secondo una probabilità calcolata basata sulla composizione del messaggio stesso. Analogamente gli IDS che utilizzano questo strumento, sollevano un allarme se la probabilità dell'azione in analisi è tale da considerarla illecita. Di seguito per tanto descriveremo il Teorema di Bayes che rappresenta la base teorica di questi strumenti ed infine presenteremo l'applicazione diretta a sistemi antispam ed IDS.

2 Principi

In questo capitolo verrà mostrata la teoria che sta alla base dei filtri Bayesiani. Si inizierà dapprima con il Teorema di Bayes per poi passare ad illustrare i classificatori bayesiani naive.

2.1 Teorema di Bayes

Il Teorema di Bayes, nella teoria probabilistica, mostra come calcolare la probabilità inversa di un certo evento. Il teorema verrà mostrato dopo un'analisi dei principi teorici sul quale si basa:

- **probabilità marginale**
- **probabilità condizionata**

La prima, detta anche **probabilità semplice**, non è altro che la probabilità di occorrenza di ogni evento indipendentemente dall'accadere di altri eventi. La probabilità marginale è indicata con **P** ed indica la probabilità di un evento. Per esempio **P(A)** indica la probabilità che l'evento A si verifichi. Questa probabilità è chiamata anche **probabilità a priori**, poiché calcola la probabilità di A prima di aver raccolto informazioni addizionali.

La **probabilità condizionata** invece è la probabilità di un qualche evento A, data l'occorrenza di qualche altro evento B. Questa probabilità si scrive come **P(A|B)**, si legge *la probabilità (condizionata) che accada A, data l'occorrenza di B* ed è definita dalla seguente formula:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{2.1}$$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \tag{2.2}$$

Questa probabilità è anche chiamata **probabilità a posteriori**, poiché calcolo la probabilità di A dopo che ho preso atto che l'evento B è accaduto.

Il Teorema di Bayes deriva a sua volta da altri due teoremi fondamentali della probabilità. Questi sono il **teorema della probabilità composta** che enuncia:

$$P(A \cap B) = P(B) \cdot P(A \mid B) = P(A) \cdot P(B \mid A) \tag{2.3}$$

ovvero la probabilità che l'evento A e l'evento B si verifichino contemporaneamente è uguale alla probabilità di uno dei due eventi moltiplicato con la probabilità dell'altro evento condizionato dall'occorrenza del primo. In caso di di indipendenza stocastica la 2.3 vale

$$P(A \cap B) = P(A) \cdot P(B) \tag{2.4}$$

L'altro teorema su cui si basa Bayes è il **teorema della probabilità assoluta** che enuncia:

*Se $A_i, ..., A_n$ formano una partizione dello spazio campionario di tutti gli eventi possibili Ω e **B** è un qualsiasi evento dipendente dagli eventi A_i, allora:*

$$P(B) = \sum_{i=1}^{n} P(A_i \cap B) = \sum_{i=1}^{n} P(A_i) \cdot P(B \mid A_i) \tag{2.5}$$

Date le premesse enunciate, il **Teorema di Bayes** viene spesso utilizzato per calcolare la **probabilità a posteriori** conseguentemente a delle osservazioni. La sua formulazione più nota è:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)} \tag{2.6}$$

dove:

- **P(A)**: è la **probabilità a priori** (o probabilità marginale) di A, ovvero senza tenere conto dell'evento B
- **P(A | B)**: è la **probabilità a posteriori** (o probabilità condizionata) di A dato B, poiché dipende dal verificarsi di B.
- **P(B | A)**: è la **probabilità condizionata** di B dato A
- **P(B)**: è la **probabilità a priori** (o probabilità marginale) di B ed è chiamata *costante di normalizzazione*

Intuitivamente, il teorema descrive il modo in cui le opinioni nell'osservare A siano arricchite dall'aver osservato l'evento B e descrive la relazione che intercorre *tra la probabilità condizionata dell'evento A dato B e la probabilità condizionata inversa dell'evento B dato A*.

Il teorema può essere generalizzato prendendo in considerazione più eventi, $P(A_i \mid B)$. Infatti considerando lo spazio degli eventi $A_1, A_2, A_3, ..., A_n$ (chiamato partizione dello spazio degli eventi) la 2.6 si modifica come segue:

$$P(A_i \mid B) = \frac{P(B \mid A_i) \cdot P(A_i)}{P(B)} = \frac{P(B \mid A_i) \cdot P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j) \cdot P(A_j)} \tag{2.7}$$

Tutte le componenti hanno lo stesso significato della 2.6. Come detto precedentemente il problema deriva dalla definizione di **probabilità condizionata**. Infatti partendo dalle formule 2.1 e 2.2 si ha che:

$$P(A \mid B) \cdot P(B) = P(A \cap B) = P(B \mid A) \cdot P(A) \tag{2.8}$$

Dividendo ogni termine per P(B) si Teorema di Bayes è verificato.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A) \cdot P(A)}{P(B)} \tag{2.9}$$

2.2 Classificazione Bayesiana Naive

Un classificatore, in cluster analysis o in analisi statistica, risolve il problema di individuare delle sotto-popolazioni di individui con determinate caratteristiche o features in un insieme più grande, con l'uso eventuale di un sotto-insieme di individui noti discriminati a priori (insieme di training).

Una tipologia di classificatori è basata sulla applicazione del Teorema di Bayes: questi tipi di classificatori sono noti come classificatori bayesiani. Il principio che regola questa tipologia di classificatori si basa sul fatto che alcuni individui appartengono ad un classe di interesse con una data probabilità sulla base di certe osservazioni. Tale probabilità è calcolata assumendo che le caratteristiche osservate possano essere tra loro dipendenti o indipendenti; in questo secondo caso il classificatore bayesiano è detto naive o idiota in quanto assume ingenuamente che la presenza o l'assenza di una particolare caratteristica in una data classe di interesse non è correlata alla presenza o assenza di altre caratteristiche semplificando notevolmente il calcolo.

Descriviamo di seguito il modello probabilistico che sta alla base dei classificatori bayesiani naive. Data una classe di interesse C ed un insieme di caratteristiche $F_1, ..., F_n$ di un individuo si vuole conoscere con quale probabilità

questo appartiene a C. Ovvero si vuole conoscere:

$$p(C|F_1, \cdots, F_n). \tag{2.10}$$

Si vuole per tanto conoscere

$$probabilita' \, a \, posteriori \; = \frac{probabilita' \, a \, priori \; * \; verosimiglianza}{osservazioni} \tag{2.11}$$

Più formalemente per Bayes

$$2.10 = \frac{p(C) * p(F_1, \cdots, F_n|C)}{p(F_1, \cdots, F_n)} \tag{2.12}$$

Tralasciando il denominatore che rappresenta una costante di normalizzazione è necessario valutare la probabilità $p(F_1, ..., F_n|C)$. Applicando al numeratore (che per la probabilità composta 2.3 equivale a $p(C, F_1, \cdots, F_n)$) più volte la definizione di probabilità condizionata 2.1 2.2 si ottiene

$$p(C, F_1, \cdots, F_n) = p(C)p(F_1|C)p(F_2|C, F_1)p(F_3|C, F_1, F_2) \cdots p(F_n|C, F_1, \cdots, F_{n-1}) \tag{2.13}$$

Con l'assunzione naive della indipendenza condizionale [5] si assume che ogni caratteristica F_i sia condizionalmente indipendente da ogni altra caratteristica F_j con $i \neq j$. Per cui

$$p(F_i|C, F_j) = p(F_i|C) \tag{2.14}$$

con $i \neq j$; per la probabilità composta 2.4 possiamo quindi

$$p(C, F_1, \cdots, F_n) = p(C)p(F_1|C)p(F_2|C) \cdots p(F_n|C) = p(C) \prod_{i=1}^{n} p(F_i|C). \tag{2.15}$$

Questo modello abbinato ad un decisore costituisce un classificatore bayesiano naive. L'apprendimento bayesiano naive richiede il calcolo della probabilità $P(C|F_n)$ per tutte le F_i che nel caso di un problema reale risulta intrattabile. Per semplificare si considerano solo le F_i che massimizzano la probabilità $P(C|F_n)$. Con questo tipologia di decisore, chiamato Maximum A Posteriori (MAP), il classificatore bayesiano naive assume la seguente:

$$arg\, max_C \, p(C) \prod_{i=1}^{n} p(F_i|C). \tag{2.16}$$

3 Applicazioni

In questo capitolo verranno mostrate due importanti applicazioni basate sul Teorema di Bayes: **antispam** e **Intrusion Detection System**, detti **IDS**.

3.1 Antispam

Lo **spam** è l'uso dei sistemi di messaggistica elettronica per inviare indiscriminatamente messaggi indesiderati. Oggi lo spam coinvolge molti canali di trasmissione elettronica: instant messaging, usenet, search engine, blogs, wiki, social network e molti altri.

Quello che però ha avuto il maggior successo è l'**lo spam via email**, definito anche come **junk mail**. É un sottoinsieme dello spam che coinvolge una serie di messaggi tendenzialmente uguale inviati a migliaia di destinatari in un lasso di tempo relativamente breve utilizzando come mezzo di trasporto la posta elettronica. Lo spam è iniziato a diventare un problema quando internet fu aperto all'uso pubblico, intorno alla metà degli anni 90. Con il passare del tempo è cresciuto esponenzialmente arrivando ad occupare oggi circa l'80-90% di tutte le mail inviate nella rete mondiale [2].

La legislazione sullo spam è trattata in maniera diversa a livello nazionale. Nonostante in alcuni stati ci siano restrizioni legislative, lo spam al giorno d'oggi ha ancora la migliore ed ha dei costi elevatissimi. Si pensi che nel 2009 lo spam ha avuto un costo di circa 130miliardi di dollari, 42 solo negli Stati Uniti [1].

Ci sono varie tecniche di spam, quali: l'**image spam** dove consiste in una immagine GIF o JPEG per offuscare del testo; c'è il **blank spam**, dove il corpo del messaggio è inesistente: lo scopo di questo attacco è collezionare

una serie di indirizzi email validi per poi mandare altra tipologia di spam o venderli ad altri enti.

Lo spam inoltre può essere utilizzato anche per attaccare alcuni enti o servizi: a tal proposito sono stati creati **virus** e **worms** che una volta infettato un computer, inviavano email a tutti gli indirizzi conosciuti di quella macchina. Lo scopo era quello di infettare altri computer o semplicemente anche di raccogliere una lista di mail valide.

Un'altra tecnica per creare liste di email valide consiste nella scansione del web: un indirizzo mail corrisponde ad una **regex** e tramite un software apposito che scansiona pagine web si possono trovare migliaia di indirizzi mail che *matchano* la regex creata. Per contrastare questo fenomeno molti utenti hanno pubblicato la propria mail sotto forma di immagine oppure scrivendola in maniera differente dal solito, per *rompere* il match con la regex. Per esempio la mail *me@diegor.it* può essere scritta come segue:

- *me chiocciola diegor punto it*
- *me at diegor dot it*
- *me[@]diegor[.]it*
- *me@NOSPAM.diegor.it*
- ...

Nonostante queste accortezze, ci sono siti (come forum, social network, etc) che pubblicano le mail dei propri utenti in maniera chiara permettendo così agli spammer di *catturarli*. Una volta che un indirizzo entra in una di queste liste, l'unica cosa che può aiutare è un software da installare sul computer o fornito dal provider di posta elettronica: software **antispam**.

Un antispam è un software che controlla ogni messaggio in entrata in un sistema e decide se il messaggio è **spam** o **ham** (ovvero tutti quei messaggi considerati *buoni*). Esistono diverse tecniche antispam, ognuna delle quali non assicura una soluzione definitiva al problema dello spam ed ognuna di essa ha un *trade-off* tra l'incorretto rifiuto di email legittime ed il non rifiuto di tutto lo spam. Inoltre anche i costi associati in termini di tempo e sforzo non possono essere trascurati.

Esistono quattro categorie principali di tecniche antispam:

- **richiedono l'intervento di una persona**: disabilitare l'HTML nelle mail, modificare la rappresentazione del proprio indirizzo su pagine pubbliche, evitare di rispondere allo spam, segnalare lo spam, etc...

- **possono essere automatizzate dall'amministratore del server di posta**: meccanismi di autenticazione e reputazione, controlli basati su checksum, liste nere basate su DNS, greylisting, controllo HELO/EHLO, controllo inverso sui DNS, filtri basati su regole, SMTP proxy, filtri statistici, etc...

- **possono essere automatizzati da chi invia la posta**: controlli sui nuovi utenti e clienti, limitare il backscatter, bloccare o intercettare la porta SMTP, controllo del FROM, accordi TOS (Terms of service) restrittivi, etc...

- **sono impiegate da ricercatori e funzionari delle forze dell'ordine**: legislazioni e pene più severe, analisi dei siti che fanno spam

Una delle strategie di tecnica antispam a livello server è rappresentata dall'uso di filtri statistici, alcuni dei quali basati sul Teorema di Bayes, denominati **filtri bayesiani**

I filtri bayesiani una volta configurati non richiedono manutenzione straordinaria da parte degli amministratori: gli utenti marcano i messaggi come **spam** o **ham** ed i filtri apprendono da questi giudizi (feedback). Tale tipologia di filtro si adatta velocemente al cambiamento del contenuto dello spam senza nessun particolare intervento da parte degli amministratori. Questi filtri inoltre possono usare gli *header* del messaggio così che possano fare ulteriori controlli sul trasporto del messaggio. I filtri più semplici utilizzano una solo parola per calcolare se il messaggio è uno spam o un ham; altri, più evoluti, considerano gruppi di due o più parole, aumentando la precisione diminuendo casi di **falsi positivi** e/o **falsi negativi**.

Il Teorema di Bayes può essere applicato per decidere se un messaggio è *spam* o *ham*. I filtri per fare ciò usano il **classificatore naive di Bayes**.

La prima applicazione ad usare i filtri bayesiani fu rilasciata nel 1996: questa non faceva altro che ordinare la mail in una directory. Invece la prima pubblicazione in merito uscì nel 1998: 'A Bayesian approach to filtering junk e-mail' [9]. Non tardarono ad arrivare le prime applicazioni commerciali ad implementare questa tipologia di filtri.

Molte parole hanno una specifica probabilità di occorrenza nelle mail di spam o nelle mail di ham (basti pensare alla parola '*viagra*'). I filtri non conoscono queste probabilità a priori ed hanno bisogno di apprendere in modo da poter funzionare correttamente. Per fare ciò, l'utente indica manualmente quale mail è spam o no. Per tutte le parole nelle mail di apprendimento, il filtro regola le probabilità che ogni parola appare in uno spam o in un ham.

Dopo una fase di training, le probabilità delle parole sono usate per calcolare la probabilità che una mail con un particolare insieme di parole all'inter-

no appartenga ad una delle categorie individuate (spam, ham, altro). Ogni parola contenuta nella mail contribuisce alla probabilità che quella mail sia spam (magari escludendo articoli, congiunzioni, etc..). Questo contributo è chiamato **probabilità a posteriori** ed è calcolata con il Teorema di Bayes 2.6. Successivamente, la probabilità che la mail sia uno spam è calcolata su tutte le parole della mail e se il totale supera una certa soglia (per esempio 95%) il filtro marca la mail come spam.

Il training iniziale può essere spesso rifinito quando il software identifica un falso positivo o un falso negativo: questo permette al software di adattarsi automaticamente alla naturale evoluzione dello spam. Molti antispam combinano l'utilizzo di questi filtri con delle euristiche avendo così una maggiore accuratezza.

I filtri bayesiani possono essere applicati nei seguenti modi:

- per calcolare la probabilità che il messaggio è spam, sapendo che una parola data appare nel messaggio
- per calcolare la probabilità che il messaggio è spam, prendendo in considerazione tutte le sue parole (o un sottoinsieme significativo)

Formalmente, supponiamo di avere un messaggio che contiene la parola *'viagra'*. La maggior parte delle persone che ricevono questa mail sanno che è un messaggio di spam, poiché è una proposta di vendita di pillole miracolose che non sono altro che un agglomerato di zuccheri. Comunque il software antispam non fa un analisi semantica del messaggio, dunque non riesce a capirne il significato e tutto quello che può fare è calcolare la sua probabilità. É possibile applicare direttamente il Teorema di Bayes 2.6.

$$P(S \mid W) = \frac{P(W \mid S) \cdot P(S)}{P(W \mid S) \cdot P(S) + P(W \mid H) \cdot P(H)} \tag{3.1}$$

dove:

- $\mathbf{P(S \mid W)}$ è la probabilità che un messaggio è considerato spam sapendo che la parola *'viagra'* sia contenuta in esso
- $\mathbf{P(S)}$ è la probabilità a priori che un dato messaggio sia spam
- $\mathbf{P(W \mid S)}$ è la probabilità che la parola *'viagra'* compaia nel messaggio di spam
- $\mathbf{P(H)}$ è la probabilità a priori che un dato messaggio **non** sia spam (*'H'* sta per *ham*)

- $\mathbf{P(W \mid H)}$ è la probabilità che la parola *'viagra'* compaia in un messaggio *ham* (non spam)

Assumiamo che la probabilità di ogni messaggio spam sia l'80% [2] avremo le seguenti probabilità $P(S) = 0.8; P(H) = 0.2$. Comunque la maggior parte dei software antispam non partono da questa supposizione e prendono queste due probabilità in egual misura $P(S) = 0.5; P(H) = 0.5$

I filtri che usano questa ipotesi si chiamano *'imparziali'* e la 3.1 risulterà:

$$P(S \mid W) = \frac{P(W \mid S)}{P(W \mid S) + P(W \mid H)} \tag{3.2}$$

La 3.2 è chiamata *spamicity o spaminess* della parola *'viagra'*. Il valore $P(W \mid S)$ usato nella formula è approssimato alla frequenza dei messaggi contenenti la parola *'viagra'* nei messaggi identificati come *spam* durante la fase di apprendimento. In maniera analoga $P(W \mid H)$ è approssimato alla frequenza dei messaggi contenenti la parola *'viagra'* nei messaggi identificati come *ham* nella fase di apprendimento. Per far si che queste approssimazioni abbiano senso il numero di messaggi di apprendimento deve essere sufficientemente grande e suddiviso in egual misura tra spam ed ham.

Ovviamente è errato prendere in considerazione una sola parola per capire se un messaggio è spam o meno. Dunque i filtri bayesiani più comuni combinano la *'spaminess'* di un insieme significativo di parole contenute nel messaggio in modo tale da calcolare la probabilità globale che il messaggio sia spam o meno. In questo caso il filtro bayesiano è realizzato con un classificatore bayesiano naive dove le parole presenti nel messaggio sono considerate eventi indipendenti. Questa assunzione non è corretta nei linguaggi naturali, infatti la probabilità di trovare un aggettivo è influenzata dalla probabilità di avere un nome. Con questa assuzione il filtro bayesiano segue il modello probabilistico 2.16

La probabilità così calcolata è comparata con una **soglia**: se è più *bassa* il messaggio è un *ham*, altrimenti uno *spam*.

3.2 Intrusion Detection System

Gli Intrusion Detection System sono dispositivi largamente diffusi per rilevare attacchi o accessi non autorizzati a sistemi informativi. Il loro scopo è analizzare l'attività del sistema informativo tramite la lettura dei log o monitoraggio dei pacchetti di rete discriminando azioni o accessi leciti da quelli non autorizzati o abusivi e rilevando tramite allarmi eventuali anomalie. Un modello generale di IDS è mostrato in 3.1. Le risorse del sistema

(applicazioni, computer e rete) sono protette e sia l'accesso che l'uso sono regolamentate da opportune policy che associano a ciascun entità utilizzatore (utente o host) le azioni ammissibili. L'IDS processa, al fine di rilevare una eventuale intrusione o un azione non autorizzata, dati di audit che descrivono le azioni delle entità e lo stato del sistema informativo durante il suo ciclo di funzionamento. Sulla base delle strategie di analisi dei dati audit gli IDS

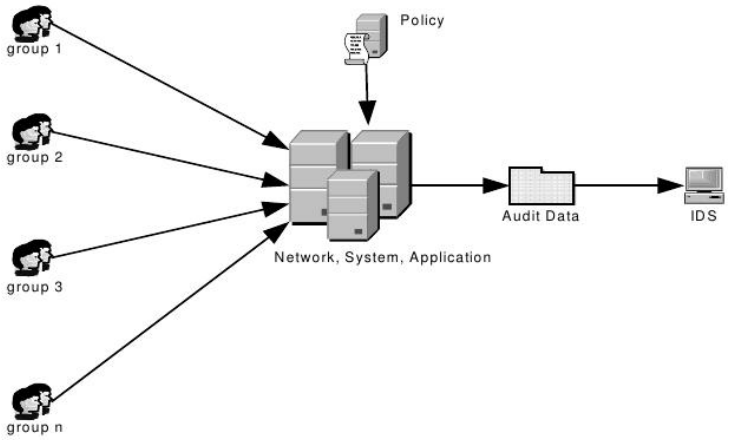


Figura 3.1: Modello generale di IDS.

generalmente sono classificati in due tipologie [8]:

- Misuse o knowledge-based IDS: mirano ad individuare uno stato del sistema o sequenze di azioni che sono precedentemente già stati identificati come intrusione. Queste tipologie di IDS devono conoscere a priori lo stato del sistema o la sequenza di azioni non ammissibili.

- Anomaly o behavior-based IDS: presumono che un intrusione può essere rilevata osservando una differenza sull'uso del sistema informativo rispetto al suo funzionamento abituale. L'IDS confronta quindi il modello di funzionamento abituale con l'attività corrente e genera un avviso ogni volta si verifichi un certo grado di divergenza rispetto al modello abituale.

Rispetto ai misuse IDS, gli anomaly IDS non richiedono regole statiche fornendo un comportamento dinamico tramite l'uso di metodi statistici ed euristici. La bontà dei metodi di classificazione dei dati di audit per questo tipo di IDS risulta strategica.

Tuttavia gli anomaly IDS presentano dei svantaggi:

- **falso negativo**: se non associati a regole statiche generali un utilizzo corretto ma estremamente raro del sistema informativo potrebbe generare un falso allarme
- **feature**: la scelta delle feature da monitorare risulta strategica in quanto condiziona la reattività del sistema agli allarmi.

Dato un insieme ordinato di azioni $S = e_1, e_2, \cdots, e_n$, il meccanismo di classificazione valuta ogni evento $e_i \in S$ se è un'azione ammissibile o meno. Le feature di ogni evento e_i sono confrontate con le corrispondenti presenti nei k modelli usati per il confronto. Il risultato di questa comparazione è un valore di output o_i che rappresenta la deviazione del valore della feature dell'evento dal valore *'normale'* atteso.

Un evento e_i è quindi considerato nella maniera seguente:

$$EC(o_1, o_2, \cdots, o_k, I) = \begin{cases} \text{e is normal:} & \sum_{i=1}^{k} o_i \leq I \\ \text{e is anomalous:} & \sum_{i=1}^{k} o_i \geq I \end{cases} \tag{3.3}$$

dove:

- EC è la funzione di *Event Classification*
- I è una soglia di valutazione

Il sistema appena presentato è condizionato dalla scelta dei modelli e dalla scelta del parametro di soglia.

Recentemente sono stati teorizzati IDS che utilizzano i classificatori bayesiani per discriminare gli eventi [6] [9] [8] [7] [4] [10]. In tali IDS, in linea di principio, viene calcolata con quale probabilità un evento con determinati valori di features che lo caratterizzano possa essere considerato *normal* o *anomalous*. Più nel dettaglio, in accordo con **??** ogni evento e ricade in tre categorie $C = anomalous, normal$ con una certa probabilità. In maniera analoga ai classificatori bayesiani implementati nei sistemi antispam si vuole calcolare $p(C = normal|F_1, \cdots, F_n)$ ovvero con quale probabilità un evento e è classificato ad esempio come *normal* sulla base delle osservazioni delle features $F_1, \cdots, F_n$ per cui si vuole calcolare:

$$p(C = normal|F_1, \cdots, F_n) = \frac{p(C = normal) * p(F_1, \cdots, F_n|C)}{p(F_1, \cdots, F_n)} = p(C = normal) \prod_{i=1}^{n} p(F_i|C = normal) \tag{3.4}$$

Dove $P(C)$ rappresenta la probabilità a priori che un dato evento e sia considerato ammisibile dal sistema. Tale valore è costrutito a partire dalla fase di training e dai rilevamenti storici dell'IDS. La $p(F_1, \cdots, F_n|C = normal)$ rappresenta invece la funzione di verosimiglianza.

In conclusione, anche in questa applicazione viene utilizzato un decisore di tipo MAP che massimizza la probabilità che un evento e appartenga ad una determinata categoria. Tale strategia cerca di evitare condizioni in cui l'IDS non riesca a classificare e (categorie equiprobabili). Alcuni sistemi IDS in caso di incertezza non tentano di classificare e di conseguenza reaggire al singolo un singolo evento ma valutano una sequenza di eventi. Un allarme è sollevato nel caso in cui la probabilità di questa sequenza tenda alla categoria *anomalous*.

Bibliografia

[1] Cost of spam 2009. http://ferris.com/2009/01/28/cost-of-spam-is-flattening-our-2009-predictions/.

[2] The maawg email metrics program. http://www.maawg.org/email_metrics_report.

[3] T. Bayes and R. Price. An Essay towards solving a Problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFRS. *Philosophical Transactions*, 53:370, 1763.

[4] D.J. Burroughs, L.F. Wilson, and G.V. Cybenko. Analysis of distributed intrusion detection systems using Bayesian methods. In *pcc*, pages 329–334. IEEE, 2002.

[5] A.P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):1–31, 1979.

[6] C. Kruegel, D. Mutz, W. Robertson, and F. Valeur. Bayesian event classification for intrusion detection. 2003.

[7] Y. Liu, C. Comaniciu, and H. Man. A Bayesian game approach for intrusion detection in wireless ad hoc networks. In *Proceeding from the 2006 workshop on Game theory for communications and networks*, pages 4–es. ACM, 2006.

[8] R.S. Puttini, Z. Marrakchi, and L. Mé. A Bayesian Classification Model for Real-Time Intrusion Detection. In *AIP Conference Proceedings*, pages 150–162. Citeseer, 2003.

[9] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A Bayesian approach to filtering junk e-mail. In *Learning for Text Categorization: Papers from the 1998 workshop*, volume 62, 1998.

[10] A.A. Sebyala, T. Olukemi, and L. Sacks. Active platform security through intrusion detection using naive bayesian network for anomaly detection. In *London Communications Symposium*. Citeseer, 2002.