Détection d'anomalies distribuée adaptative via machine learning

MAZEL Johan

Directeurs de thèse : Y. Labit, P. Owezarski

28 avril 2009

LAAS

Besoins/Motivations

- Croissance d'Internet et évolution de son traffic
  $\Rightarrow$ Mutation de la demande de QoS : best effort vers différents niveaux de QoS garantis.
- Impact des anomalies sur les QoS.
- Anomalies :
  - Définition : phénomène qui perturbe le fonctionnement d'un réseau.
  - Différents types d'anomalies :
    - Anomalies légitimes, ex : foules subites (flashcrowds).
    - Anomalies illégitimes, ex : attaques de Déni de Service Distribué (DdSD).

$\Rightarrow$ Besoin de détecter les anomalies de façon fiable.

Problématique

Forte variabilité du traffic rend la détection délicate :

$\Rightarrow$ Nombreux faux positifs.

$\Rightarrow$ Possibilité de faux négatifs.

État de l'art de la détection d'anomalies

- Statistiques simples (comptages par ex.).
- Statistiques avancées basées sur la variabilité et les propriétés d'échelle du traffic.
  - Entropie.
  - Densité spectrale (Hussain 2003)
  - Modèles de Markov (Ye 2000)
  - Augmentation de la correlation du trafic (Jin Yuan 2004)
  - Décomposition en ondelettes de la distribution d'énergie (Li 2003)
  - Analyse multi-échelle par la décomposition en ondelettes (Barford Jung 2002)

Objectifs

Approche pour la détection d'anomalies

- Détection d'anomalies par machine learning
  - Introduction d'un système de détection à deux étapes : détection puis classification.
  - Adaptation à l'évolution du traffic et détection de nouvelles anomalies via machine learning.
- Collaboration des noeuds pour améliorer la fiabilité de la détection via corrélation des résultats locaux.

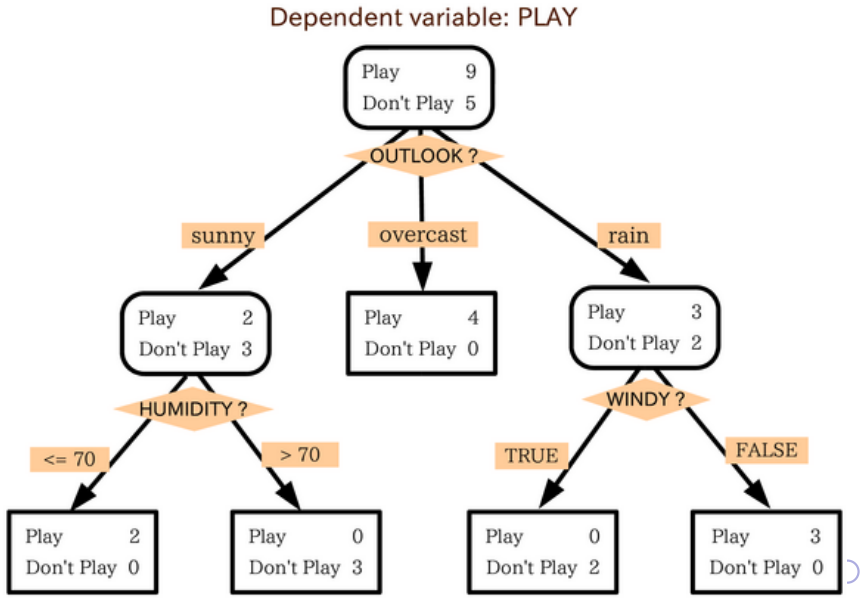
Outil annexe : système distribué de métrologie passive et active.

Qu'est-ce que le machine learning ?

- "Supervised/Semi-supervised learning"
  - Principe : établissement automatique de règles à partir de données préalablement marquées.
  - Algorithmes : decision tree, SVM (Support Vector Machines),...
  - Ex. : ADAM : Détection d'intrusion par semi-supervised learning.

Play golf dataset

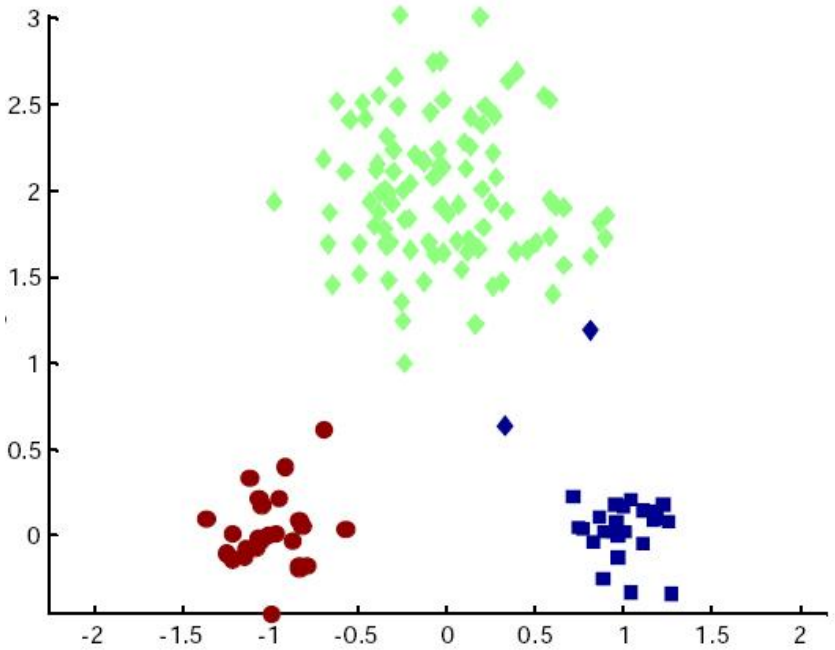
Independent variables				Dep. var
OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
sunny	85	85	FALSE	Don't Play
sunny	80	90	TRUE	Don't Play
overcast	83	78	FALSE	Play
rain	70	96	FALSE	Play
rain	68	80	FALSE	Play
rain	65	70	TRUE	Don't Play
overcast	64	65	TRUE	Play
sunny	72	95	FALSE	Don't Play
sunny	69	70	FALSE	Play
rain	75	80	FALSE	Play
sunny	75	70	TRUE	Play
overcast	72	90	TRUE	Play
overcast	81	75	FALSE	Play
rain	71	80	TRUE	Don't Play



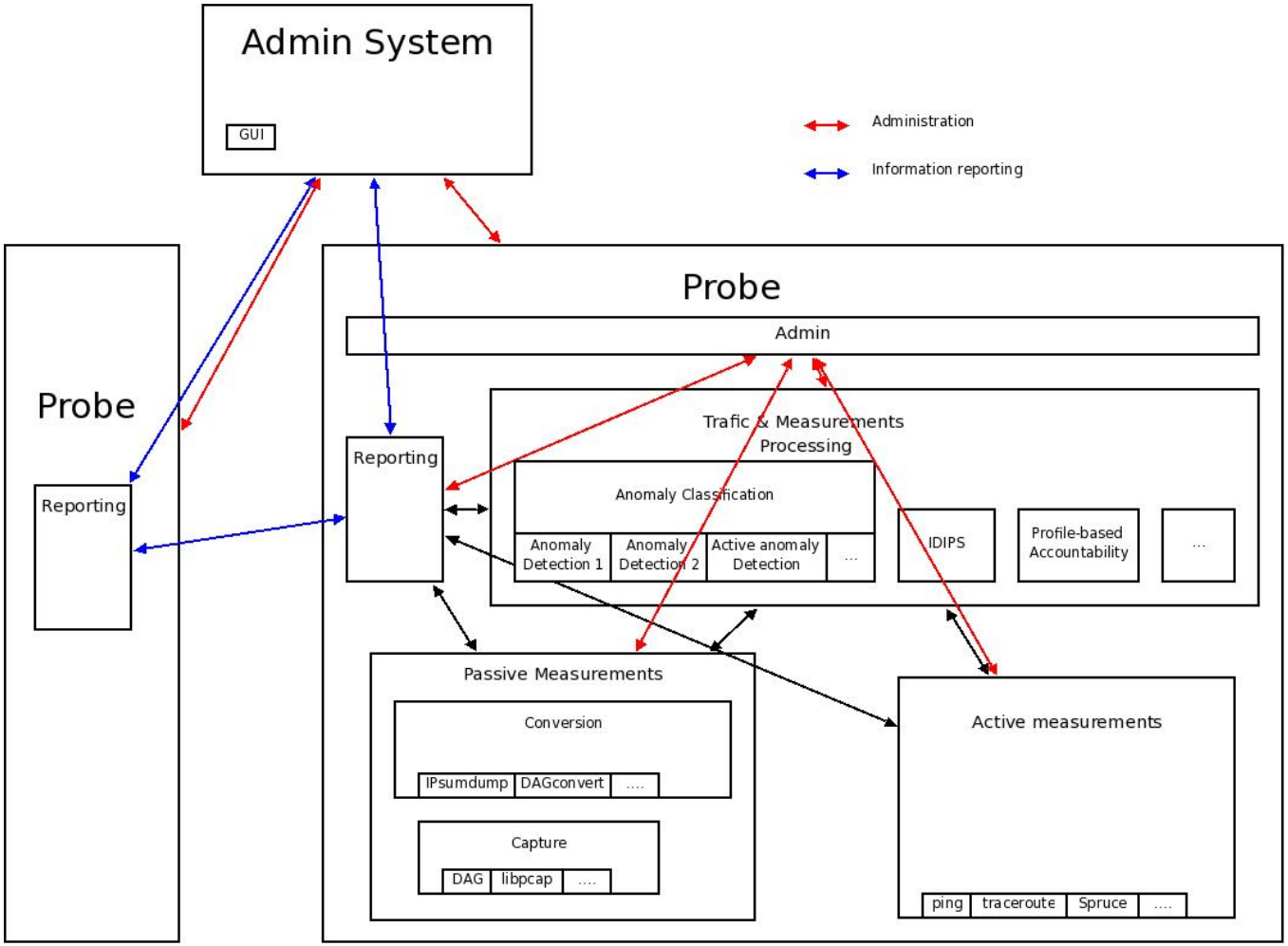
Qu'est-ce que le machine learning ?

- "Unsupervised learning"

- Principe : recherches de structures/similitudes au sein de données non marquées.
- Algorithmes : réduction dimensionnelle (PCA, MDS), estimation de densité (réseaux bayésiens, mixture network), clustering (k-means, hierarchical).
- Ex. : Lakhina/Crovella/Diot : Application d'unsupervised learning sur l'entropie de différents attributs (srcIP, dstIP, srcPort and dstPort).



Architecture proposée



Détection/classification d'anomalies

- Détection des anomalies configurée pour ne pas obtenir de faux négatifs.
- Classification des alarmes précédemment levées pour éliminer les faux positifs.

  $\Rightarrow$ Utilisation de 5 régles basées sur 30 attributs qui permettent de classifier les anomalies parmi 5 types.

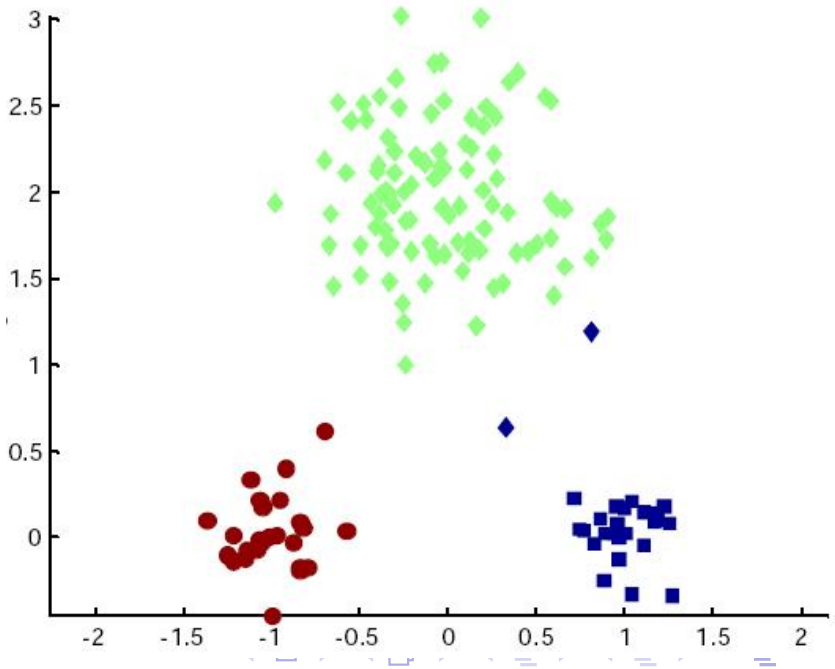
Application du machine learning à la détection d'anomalies

- Supervised/Semi-supervised learning
  - Établit automatiquement des règles à partir de données marquées
    ⇒ Permet de définir les seuils du système de classification.
    ⇒ Problème : Requiert des données marquées
  - Suppose que l'on connaît l'anomalie
    ⇒ Empêche la découverte de nouvelles anomalies.

Application du machine learning à la détection d'anomalies

- Unsupervised learning : solution adoptée
  - Utilisation du clustering pour :
    - Utiliser des attributs explicites (contrairement à PCA).
  - Deux représentations possibles du traffic dans le clustering :

- Un ou plusiers cluster(s) pour le trafic ≪ normal ≫ et chaque outlier/point isolé représente une anomalie.
- Un ou plusiers cluster(s) pour le trafic ≪ normal ≫ et autant d'autres clusters que d'anomalies.



Application du machine learning à la détection d'anomalies

- Unsupervised learning/Clustering
  - Méthodes d'identification comportement normal/anormal :
    - Intervention manuelle systématique.
    - Supervised learning pour identifier le traffic normal et intervention manuelle pour les anomalies.
  - Découverte de nouvelles anomalies.
    - Importance du choix des attributs.
      $\Rightarrow$ Découverte de nouveaux attributs.

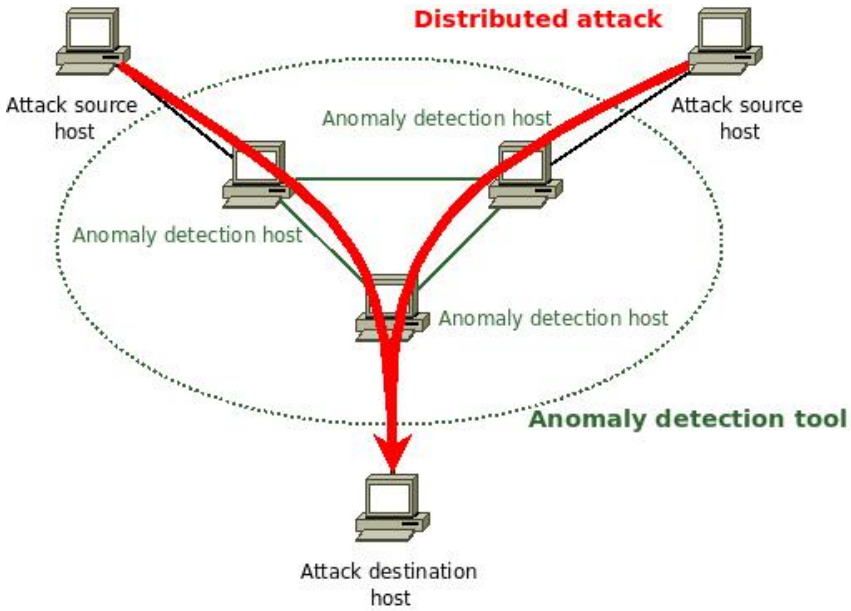
Détection d'anomalies distribuée

- Différents types de système de détection d'anomalies distribués:
  - Système distribué avec un nœud central : DIDS (Distributed Intrusion Detection System).
  - Système totalement distribué : EMERALD (Détection d'intrusion).
- Protocoles de communication/reporting existants : IPFIX

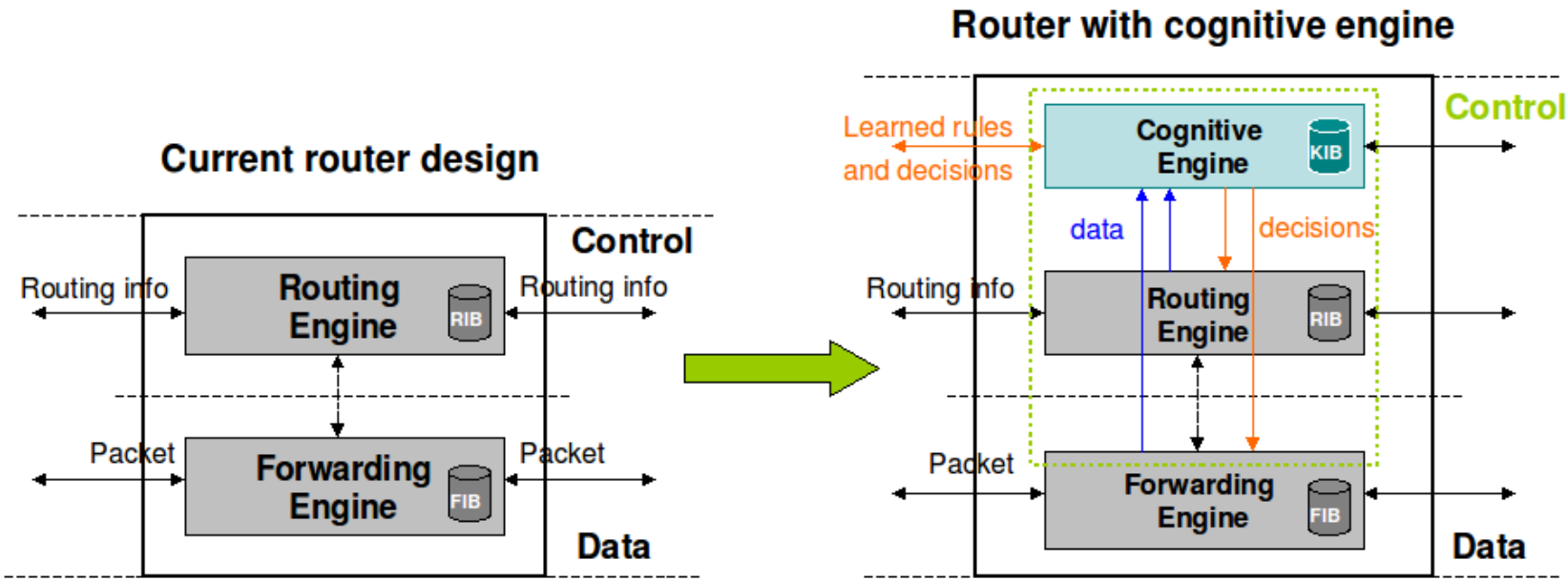
Détection distribuée d'anomalies

- Collaboration des hôtes.

  ⇒ Élaboration d'un protocole collaboratif de détection d'anomalie.



Intégration dans le projet ECODE



Intégration dans le projet ECODE

Technical Objective	**Ref.**	**Use Case**
TO1: Adaptive traffic sampling and management, path performance monitoring, and intrusion and attack/anomaly detection	a1	Adaptive traffic sampling and management
	a2	Path performance monitoring
	a3	Cooperative intrusion and attack/anomaly detection
TO2: Path availability, network recovery and resiliency, and profile-based accountability	b1	Path availability
	b2	Network recovery and resiliency
	b3	Profile-based accountability
TO3: Routing system scalability and quality	c1	Routing system scalability and routing system quality (convergence, stability/ robustness, and stretch)