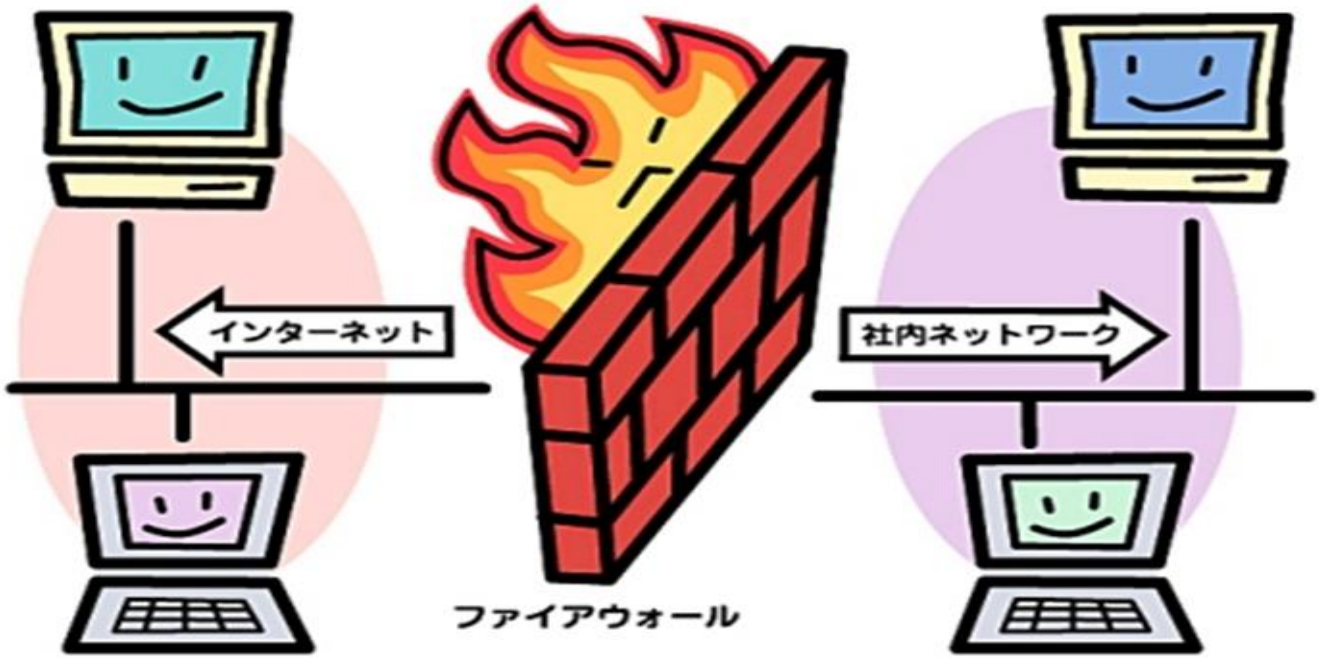
ランダムフォレストを用いたフィルタリングルール

木下研究室　201503821　野末大貴

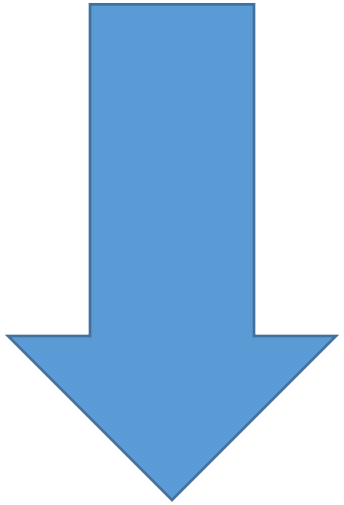
研究背景

不正アクセスによるアカウントの乗っ取り
不正アクセス対策としてファイアウォールがあげられる



研究目的

どの通信が危険かわからない

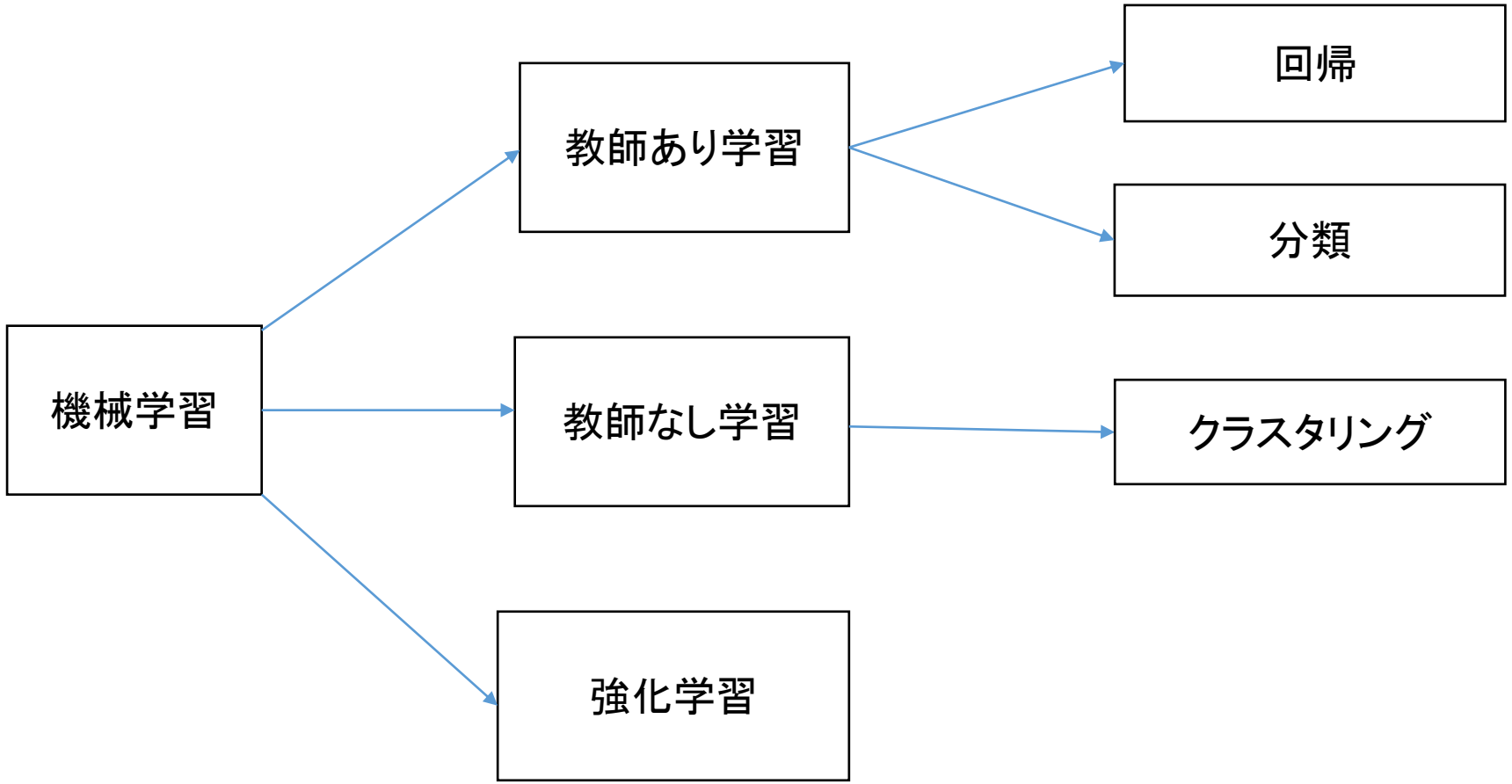


機械学習を用いたフィルタリングルールの最適化

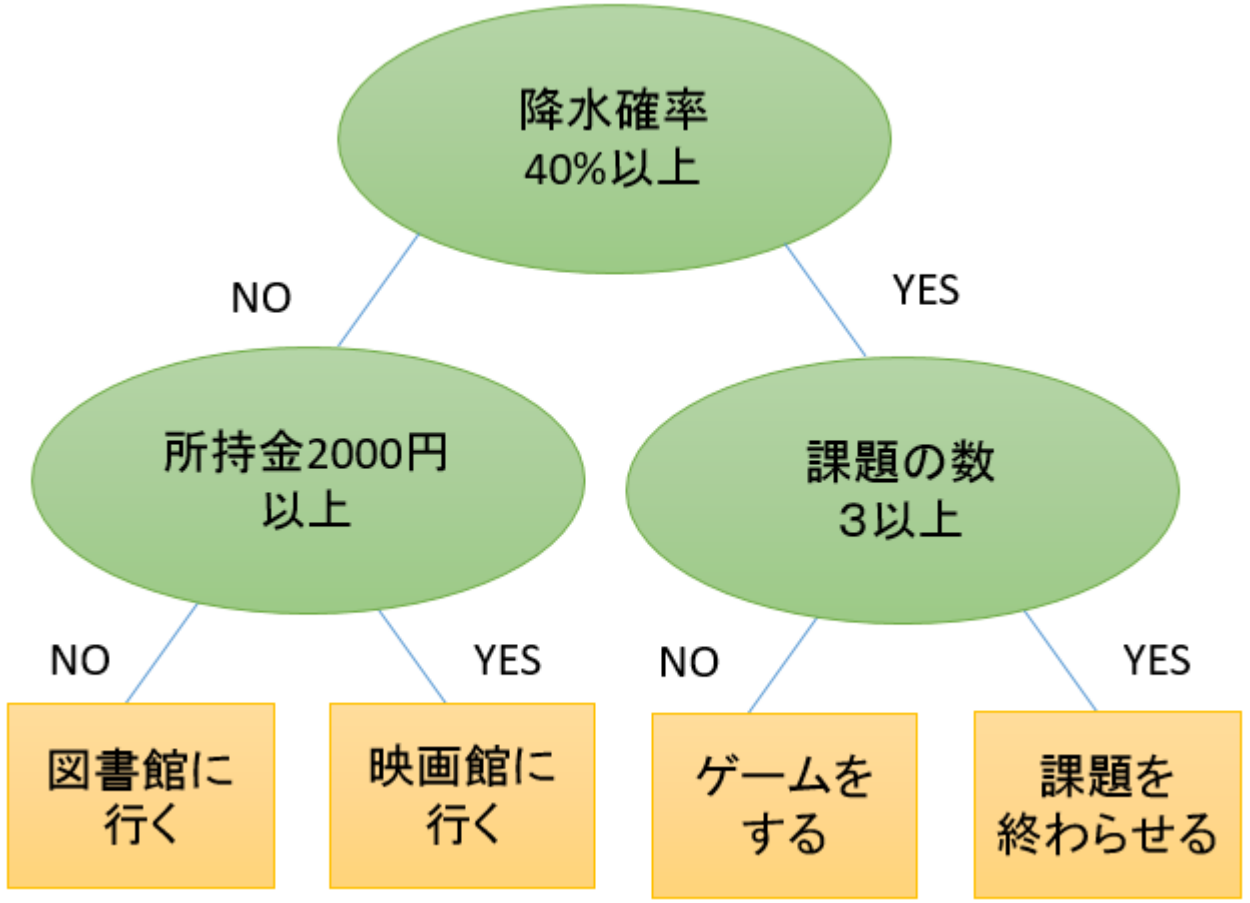
The Random Forest Based Detection of

Shadowsock's Traffic

機械学習（ランダムフォレスト）によってファイアーウォールのフィルタリングルールを作成する


機械学習
教師あり学習
教師なし学習
強化学習
回帰
分類
クラスタリング

ランダムフォレスト


降水確率
40%以上
NO
YES
所持金2000円
以上
課題の数
3以上
NO
YES
NO
YES
図書館に
行く
映画館に
行く
ゲームを
する
課題を
終わらせる

なぜ教師あり学習なのか

ファイアウォールでは、通信が攻撃かノーマルかを人間が決める決定事項であるため教師あり学習

なぜランダムフォレストなのか

木を作成する計算が並列化でき、高速な計算が可能

実装環境

- サイキットラーン
  ランダムフォレストのソースコードデータ
（武藤佳恭：アンサンブル機械学習,より）

- プロセッサ：intel
- OS：Ubuntu　17.10
- 使用言語：Python 2.7.15　　Anaconda

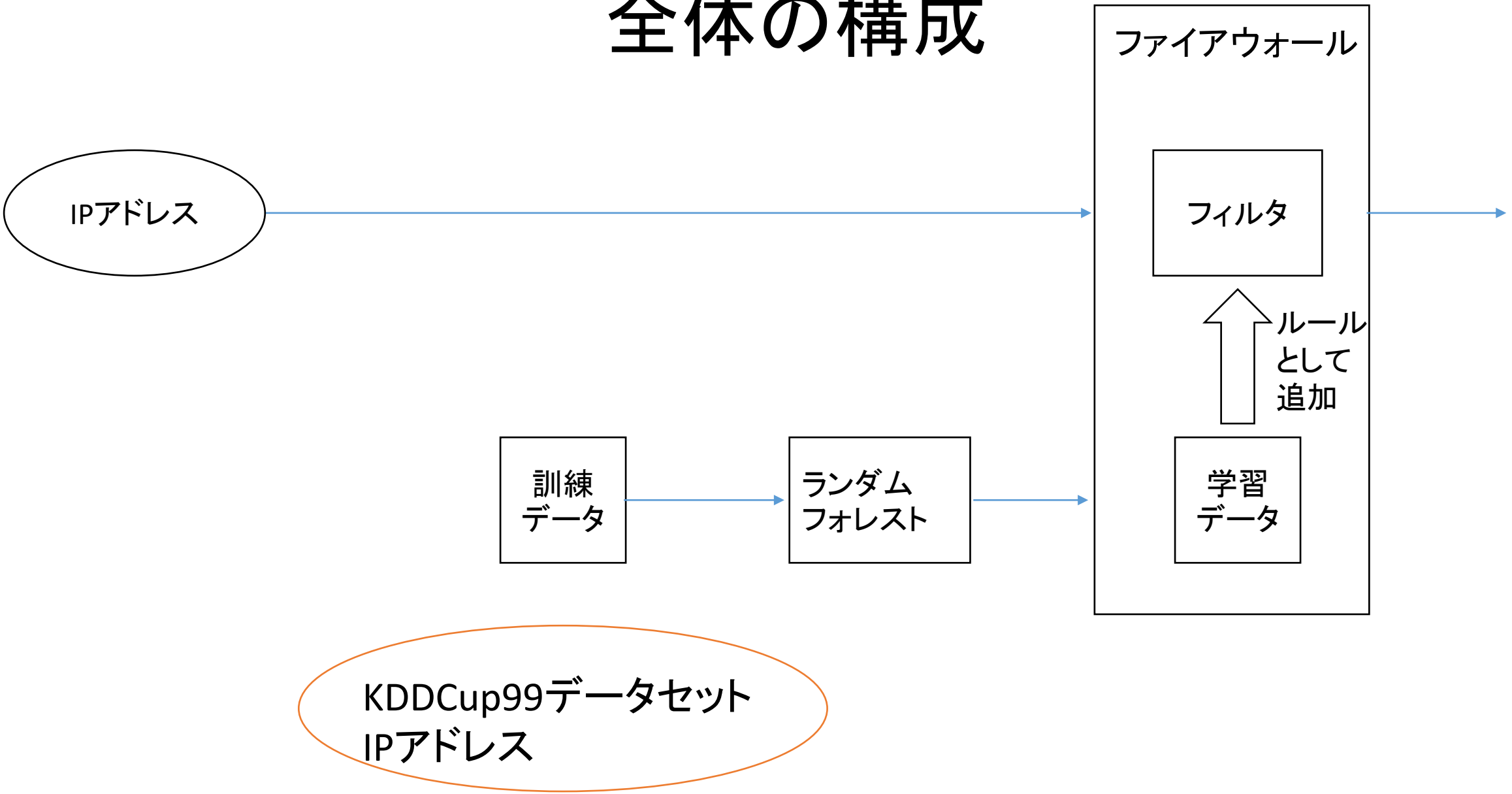
木の数と予測精度

木の数	予測精度
1	0.5161
50	0.9677
1000	0.9677

木の数が増えると精度が上がる

木の数	予測精度
1	0. 5161
2	0. 6129
5	0. 7742
10	0. 8710
20	0. 9032
30	0. 9354
50	0. 9677
1000	0. 9677

全体の構成


IPアドレス
ファイアウォール
フィルタ
ルール
として
追加
訓練
データ
ランダム
フォレスト
学習
データ
KDDCup99データセット
IPアドレス

訓練データの作成

- IPアドレス＋KDDCup99データセット

IPアドレス	hot	duration
192.168.14.29	1	0
192.168.91.175	3	0
192.168.72.178	2	0
192.168.141.90	0	2

出力

IPアドレス	ラベル	予測ラベル	判定
192.168.14.29	1	1	0
192.168.91.175	3	1	1
192.168.72.178	2	0	1
192.168.141.90	0	0	0

本来のラベルとランダムフォレストにより分類された予測ラベルを比較して、予測ラベルが正しく判定できたら0正しくない場合1と出力

結果（正規分布）

木の数	予測精度	学習時間[s]
1	0.9991	0.319
50	0.9997	10.571
1000	0.9997	216.855

最適な木の数は50本

結果（ランダム）

木の数	予測精度	学習時間[s]
1	0.9990	0.401
50	0.9997	10.761
1000	0.9997	206.285

今後の課題

IPアドレス	ラベル	予測ラベル	判定
192.168.14.29	1	1	O
192.168.91.175	0	3	1
192.168.72.178	2	O	1
192.168.141.90	0	O	O

自動で追加＋精度を1に近づける＋実際IPアドレスを使う