Uczenie klasyfikatorów rozpoznawania włamań do sieci z wykorzystaniem WEKA

dr inż. Joanna Kołodziejczyk

8.01.2017

1 Cel laboratoriów

Porównanie skuteczności różnych klasyfikatorów w uczeniu się rozpoznawania włamań wariant binarny (było włamanie lub nie) oraz wariant wielowyjściowy (różne typy włamań).

2 Założenia eksperymentu

2.1 Zbiory trenujące

1. KDDTrain+20percent.arff (liczba rekordów: 25192, 41 atrybuty + klasa, dwie klasy: normal, anomaly)
2. KDDTrain+20percent-23class-nodiff.txt (liczba rekordów: 25192, 41 atrybuty + klasa, 23 klasy)
3. KDDTrain+20percent-23class-nodiff.txt zmienić na KDDTrain+20percent- 5class-nodiff.txt zgodnie z instrukcją z zajęć projektowych nr 2(liczba rekordów: 25192, 41 atrybuty + klasa, 5 klas: normal, dos, ur2, r2l, probe)

2.2 Algorytmy klasyfikacji do przetestowania

W ramach eksperymentu należy przetestować następujące klasyfikatory: perceptron wielowarstwowy (sieć neuronowa), naiwny Bayes, drzewo J.48 oraz jako odnośnik wykorzystać klasyfikator ZeroR. W niektórych klasyfikatorach należy zmieniać parametry, które mogą mieć wpływ na wyniki. I tak w sieci neuronowej wielowarstwowej należy przetestować różną liczbę neuronów w warstwie ukrytej oraz różne wielkości współczynnika uczenia. W algorytmie J 48 można manipulować współczynnikiem wpływającem na liczbę przycięć w drzewie (mniejsze wartości skutkują większym przycinaniem) oraz minimalną liczbą rekordów w liściu.

- ZeroR
- MultilayerPerceptron

1. TrainingTime = 50, hiddenLayers = a ($a = (attribs + classes)/2$) i learningRate = 0.3
2. TrainingTime = 50, hiddenLayers = o ($o = classes$) i learningRate = 0.3
3. TrainingTime = 50, hiddenLayers = t ($t = attribs + classes$) i learningRate = 0.3
4. TrainingTime = 50, hiddenLayers = i, t (dwie warstwy ukryte) i learningRate = 0.3
5. TrainingTime = 50, hiddenLayers = a ($a = (attribs + classes)/2$) i learningRate = 0.8
6. TrainingTime = 50, hiddenLayers = a ($a = (attribs + classes)/2$) i learningRate = 0.9, decay=true

- NaiveBayes
- J48

1. confidenceFactor 0.25 i minNumObj = 2
2. confidenceFactor 0.1 i minNumObj = 2
3. confidenceFactor 0.5 i minNumObj = 2
4. confidenceFactor 0.25 i minNumObj = 10

2.3 Porównanie wyników — parametry porównania

Aby porównać otrzymane z klasyfikacji wyniki należy zestawić nastepujące parametry (miary jakości klasyfikacji):

- procent poprawnie klasyfikowanych rekordów
- powierzchnię pod krzywą ROC
- czas tworzenia modelu (User CPU time training)
- czas testowania (User CPU time testing)

Podawane wartości mają przedstawiać średnią z liczby uruchomień i odchylenie standardowe.

3 Zadania do wykonania

Do wykonania eksperymentu należy wykorzystać Experimenter w narzędziu Weka. W zakładce Setup należy otworzyć nowy eksperyment i podać następujące jego parametry:

1. Ustalić sposób testu na walidację krzyżową 5 krotną
2. Typ zadania: classification

3. Liczba powtórzeń eksperymentu równa 5.

4. Dodać odpowiednie pliki z danymi uczącymi.

5. Wybrać listę testowanych klasyfikatorów zgodnie z opisem podanym w punkcie 2.2

W sprawozdaniu należy zamieścić:

1. Krótki opis eksperymentu z warunkami brzegowymi.

2. Wyniki zestawić w postaci tabel.

3. Wykonać wykresy odpowiednie do uzyskanych wyników.

4. Zapisać wnioski, które wynikają z eksperymentu. Należy pamiętać, że nie wszystkie różnice są istotne statystycznie i należy o tym też napisać. Bardzo proszę o rozsądne i przemyślane wnioski.

3.1 Sprawozdanie

Sprawozdanie w formacie i o nazwie *imie_ nazwisko.pdf* należy przesłać na adres jkolodziejczyk[at]ajp.edu.pl. Tytuł maila: Sprawozdanie 2 z ISPAS.