Comparação de classificadores utilizando um dataset de classificação de cogumelos

Tiago Kenji Umemura

Universidade Tecnológica Federal do Paraná (UTFPR)
Campo Mourão – Paraná – Brasil

umemuratiago@gmail.com

Resumo. *Este artigo descreve os resultados obtidos com diferentes classificadores utilizando um dataset de classificação de cogumelos. Os classificadores utilizados foram Random Forest, SVM e nearest centroid classifier. O dataset utilizado divide as instâncias em duas classes, venenoso e comestível, e possui 22 características.*

1. Introdução

No experimento foi utilizado um dataset de classificação de cogumelos disponibilizado no Kaggle e que originalmente era mantido no UCI Machine Learning Repository, um website que mantêm bases de dados para serem utilizados pela comunidade de machine learning. O dataset de classificação de cogumelos possui 22 características e os cogumelos são classificados como comestível ou venenoso, assim é possível extrair quais características indicam maior probabilidade do cogumelo ser venenoso e também classificar os cogumelos em venenosos ou não com base nas características.

Kaggle também disponibiliza trabalhos que foram feitos utilizando esse dataset. Nesses trabalhos foram usados Random Forest, GBM e RPART para classificar as instâncias e mostrar quais características estão mais relacionadas a classe venenoso, porém nenhum deles utiliza todas as instâncias do dataset.

No experimento desenvolvido foi utilizado três classificadores para classificação das instâncias: Random Forest, Support Vector Machine (SVM) e nearest centroid classifier. Sendo que metade do dataset foi utilizada para treino e outra metade para teste.

2. Objetivo

O objetivo do experimento é comparar as taxas de acerto dos diferentes classificadores e também comparar as taxas com trabalhos já feitos do Kaggle. Além disso também foi analisado a relação de cada característica com a classe venenoso utilizando o algoritmo de Random Forest.

Os classificadores utilizados foram Random Forest, Support Vector Machine (SVM) e nearest centroid classifier, sendo que todos os classificadores do experimento fazem parte parte da biblioteca Scikit Learn para Python.

3. Fundamentação teórica

Random Forest é um algoritmo de aprendizagem de máquina baseado em árvores de decisão simples e métodos de aprendizagem em conjunto. O algoritmo consiste em um número arbitrário de árvores de decisões simples, que são utilizadas para determinar a saída final, ou decisão do que será avaliado.

Na classificação, o conjunto das árvores, também denominadas como floresta (forest), definem resultados que possibilitam a escolha da classe mais popular entre elas. Maior quantidade de floresta tende a melhorar a precisão da previsão.

Support Vector Machine (SVM) é um algoritmo de aprendizagem supervisionado e tem o objetivo de separar as instâncias de um dataset em duas classes. O algoritmo encontra uma linha de separação (hiperplano) que separa as instâncias em duas classes de acordo com suas características.

Nearest centroid classifier é um modelo de classificador que atribui cada classe a um centróide e a instância em análise pertence a classe do centróide mais próximo.

4. Método

O dataset disponibilizado está no formato CSV e os valores das características estão representados por letras. Primeiro foi necessário ler o arquivo utilizando a biblioteca Numpy após isso foi trocado os valores das características de letras para números, pois os algoritmos da biblioteca Scikit learn não permitem letras como parâmetros de entrada.

Após a leitura do dataset, foi executado os classificadores Random Forest, SVM e Nearest centroid no dataset. O dataset foi dividido duas partes: 4000 instâncias para treino e 4000 instâncias para teste.

O classificador Random Forest foi executado utilizando como paramêtro 10, 100 e 500 como número de árvores. Para cada valor do número de árvores o Random Forest foi executado 10 vezes e foi feito a média dessas 10 execuções, já que a taxa de acerto varia bastante. Na Random Forest também é possível verificar quais características são mais importantes para determinar se um cogumelo é venenoso.

Os classificadores SVM e Nearest centroid foram executados apenas uma vez pois suas taxas não variam como acontece com a Random Forest.

4.1. Experimentos

Random Forest			
Número de Árvores	**Mínimo**	**Máximo**	**Média**
10	0,5316	0,7376	0,5853
100	0,5858	0,8222	0,7746
500	0,6569	0,8182	0,7712

SVM	0,8247

Nearest Centroid	0.8162

5. Conclusão

Ao comparar os classificadores que foram executados utilizando 4000 instâncias para treino e teste, o SVM foi o que apresentou melhor desempenho. Nearest Centroid apresentou taxa um pouco abaixo do SVM.

A Random Forest apresentou o pior desempenho, com resultados variando bastante e mesmo com maior número de árvores a média de acertos ficou cerca de 5% abaixo dos outros classificadores, além ter maior tempo de execução conforme é aumentado a quantidade de árvores que devem ser geradas.

A Random Forest também permite analisar quais características estão mais relaciondas com a classe do cogumelo, venenoso ou não. Nesse dataset as características de maior importância foram: odor, gill size e stalk shape.

Referências

Kaggle (2017) “Mushroom Classification Safe to eat or deadly poison?”, https://www.kaggle.com/uciml/mushroom-classification, Junho.