Universidade de São Paulo
SCC - 230 Inteligência Artificial
Profa. Solange Oliveira Rezende

Base de dados “Wine Quality”

8598861 - Bernardo Simões Lage G. Duarte
8122585 - Eder Rosati Ribeiro
8936993 - Gabriel Luiz Ferraz Souto
8936648 - Giovani Ortolani Barbosa
8531887 - Giovanni Robira
8937271 - Rafael Bueno da Silva

São Carlos - SP
Novembro de 2017

Base de Dados

A base de dados[1] escolhida apresenta é dividida em duas bases menores, uma delas é referente à análise sensorial de vinho branco e a outra de vinho tinto. Existem 4898 instâncias na base com 12 atributos no total (11 de entrada e 1 de saída). Os dados podem ser utilizados tanto para problemas de classificação quanto de regressão.

O objetivo da aplicação de aprendizado de máquina nesta base é prever a qualidade do vinho através dos 11 atributos psicoquímicos de entrada. Também podem ser extraídos conhecimentos se existe uma relação forte entre um determinado atributo de entrada e a respectiva classificação

Os 11 atributos de entrada presentes na base foram obtidos através de testes psicoquímicos, já o atributo de saída foi classificado de acordo com uma análise sensorial de especialistas.

Atributos

Entrada

1. Acidez fixada
2. Acidez volátil
3. Ácido cítrico
4. Açúcar residual
5. Cloreto
6. Dióxido de enxofre livre
7. Dióxido de enxofre total
8. Densidade
9. pH
10. Sulfatos
11. Álcool

Todos os atributos são valores numéricos.

Saída

1. Qualidade

É um valor entre 0 (ruim) e 10 (bom) que indica a qualidade do vinho.

As classes não estão ordenadas e estão desbalanceadas, havendo um maior número de vinhos intermediários do que vinhos ruins ou bons.

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Proposta

A nossa intenção é trabalhar com a base da dados citada acima e classificar os vinhos pela qualidade. Fizemos alguns testes no Weka, e pretendemos trabalhar com o algoritmo kNN (IBK). A escolha o kNN foi feita a partir do estudo do problema, analisamos os possíveis algoritmos, onde percebemos que as melhores possibilidades seriam kNN, C4.5 e Multi-Layer Perceptron (MLP). Mais uma vez contamos com a ferramenta Weka para nos orientar, e analisamos o kNN e o MLP, porém não conseguimos executar o C4.5 (algoritmo J48 no Weka). Percebemos que o kNN, com parâmetros padrões tem coeficiente de relação de 0.5623 e 60.1022% de erro absoluto relativo, enquanto o MLP apresenta 0.5204 e 80.381% , respectivamente. Como o kNN teve melhores resultados, queremos encontrar relações mais forte entre os atributos de entrada a qualidade, a fim de encontrarmos melhores resultados.

A princípio não foram feitas modificações na base, apenas formalizamos a entrada para ser compatível com o Weka. Não tendo, por hora, que realizar uma etapa de pré-processamento. Vale ressaltar também que não faltam dados dentro da base, e todos os dados são válidos.

Estudaremos a possibilidade de utilizar outros algoritmos também, a fim de poder comparar os resultados. Tendo em base que o nosso problema é de classificação, estudamos os métodos citados nos slides e nas aulas (kNN, Naive Bayes, C4.5, SVM e Multi-Layer Perceptron) e com poucos testes percebemos que o kNN é o que apresenta melhores resultados com testes simples. Estudando os parâmetros e explorando o Weka, queremos descobrir qual seu melhor resultado frente a esses algoritmos.