Mineração de Dados e Aprendizado de Máquina

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

Base de Dados (Wine Quality)

- Base de dados referente à análise sensorial de vinhos branco
- Possível aplicar Classificação e Regressão.
- As classes estão desbalanceadas e fora de ordem.

Instâncias: 4898

Atributos: 12 (sendo um deles, o atributo de classe)

Wine Quality (Atributos)

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
12. Qualidade

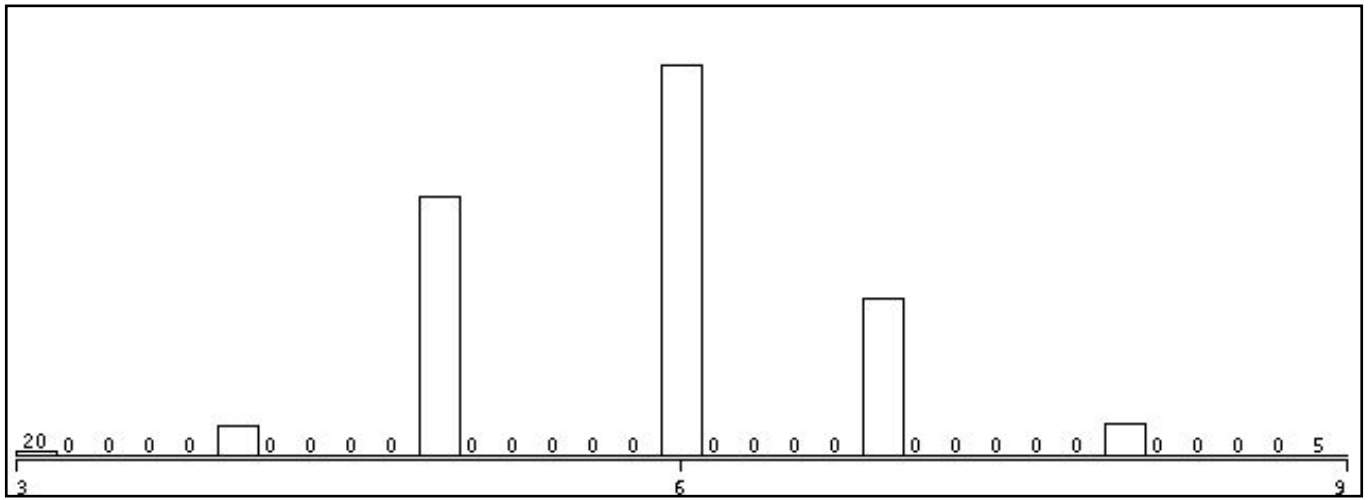
- Atributos de Entrada
  - Valores numéricos
  - Definidos por testes psicoquímicos
- Saída (Classificação)
  - Qualidade
  - Valores de 0 (ruim) a 10 (bom)
  - Obtidos através de análise sensorial de especialistas

Proposta

- Fazer classificação dos atributos da base de dados (vinhos)
- Extrair relações entre os 11 atributos de entrada e a qualidade do vinho
- Utilização de diferentes técnicas
  - k-Nearest-Neighbors (kNN) com 3 valores diferentes de k
  - Multi-Layer Perceptron (MLP)
  - Support Vector Machines (SVM)
  - Árvore de Decisão
  - Naive Bayes
  - Regra de Associação

Pré-processamento

- Atributos de entrada foram normalizados respeitando a mesma escala
- A qualidade é medida de 0 a 10, porém alguns dos valores não aparecem na base de dados
    - Classes 5, 6 e 7 são altamente dominantes

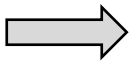


Pré-processamento

- A qualidade é medida de 0 a 10, porém alguns dos valores não aparecem na base de dados
  - Discretização das classes, agrupando valores
  - Manipulações tentando evitando overfitting e underfitting na base
  - Menos valores de classes possíveis (2 classes possíveis)
    1. intervalo [0-6] = ruim
    2. intervalo [7-10] = bom

Antes do *resampling*

Label	Count
ruim	3838
bom	1060



Depois do *resampling*

Label	Count
ruim	2448
bom	2448

kNN

- **k = 3**

  Acurácia: 92.46%

```
    a    b   <-- classified as
 2134  314 |    a = ruim
   55 2393 |    b = bom
```

```
                 TP Rate   FP Rate   Class
                 0,872     0,022     ruim
                 0,978     0,128     bom
Weighted Avg.    0,925     0,075
```

- **k = 5**

  Acurácia: 92.12%

```
    a    b   <-- classified as
 2115  333 |    a = ruim
   53 2395 |    b = bom
```

```
                 TP Rate   FP Rate   Class
                 0,864     0,022     ruim
                 0,978     0,136     bom
Weighted Avg.    0,921     0,079
```

- **k = 7**

  Acurácia: 92.03%

```
    a    b   <-- classified as
 2104  344 |    a = ruim
   46 2402 |    b = bom
```

```
                 TP Rate   FP Rate   Class
                 0,859     0,019     ruim
                 0,981     0,141     bom
Weighted Avg.    0,920     0,080
```

Naive Bayes

- Acurácia: 70.18%

```
   a    b   <-- classified as
1467  981 |    a = ruim
 479 1969 |    b = bom
```

	TP Rate	FP Rate	Class
	0,599	0,196	ruim
	0,804	0,401	bom
Weighted Avg.	0,702	0,298	

- Observações
  - Acredita-se que o algoritmo pode ser melhor
  - Parâmetros não estudados na matéria foram mantidos valores default do Weka

MLP

- Acurácia: 80.15%

```
   a    b   <-- classified as
1778  670 |    a = ruim
 302 2146 |    b = bom
```

	TP Rate	FP Rate	Class
	0,726	0,123	ruim
	0,877	0,274	bom
Weighted Avg.	0,801	0,199	

- Resultado satisfatório, porém não é o melhor especificamente para essa base de dados
- Assim como Naive Bayes, algum valores foram mantidos default

SVM

- Acurácia: 73.51%

```
   a    b   <-- classified as
1670  778 |    a = ruim
 519 1929 |    b = bom
```

	TP Rate	FP Rate	Class
	0,682	0,212	ruim
	0,788	0,318	bom
Weighted Avg.	0,735	0,265	

- Muito abaixo dos resultados já alcançados
  - Algoritmo não é bom para resolver esse problema

Árvore de Decisão (J48)

- Acurácia: 91.05%

```
   a    b   <-- classified as
2156  292 |    a = ruim
 146 2302 |    b = bom
```

	TP Rate	FP Rate	Class
	0,881	0,060	ruim
	0,940	0,119	bom
Weighted Avg.	0,911	0,089	

- Árvore resultante possui 285 nós-folha
- Desempenho bom
- Representação visual do problema

Regra de Associação (JRip)

- Acurácia: 86.85%

```
   a    b   <-- classified as
2043  405 |    a = ruim
 239 2209 |    b = bom
```

	TP Rate	FP Rate	Class
	0,835	0,098	ruim
	0,902	0,165	bom
Weighted Avg.	0,868	0,132	

- Foram geradas 60 Regras de associação
- Acurácia inesperada pelo grupo
- Pré-processamento visando o algoritmo JRip pode melhorar sua acurácia
- Exemplo de regra gerada:

```
(residual sugar >= 0.025806) and (alcohol >= 0.774194)
and (density <= 0.114273) => quality=bom (129.0/12.0)
```

Resultados

Método	Acurácia (%)
kNN - 3 vizinhos	92.46
kNN - 5 vizinhos	92.12
kNN - 7 vizinhos	92.03
Árvore de Decisão (*J48*)	91.05
Regra de Associação (*JRip*)	86.85
SVM	80.15
Naive Bayes	73.51
MLP	70.18

Conclusão

- Os algoritmos são melhores que classificadores aleatórios (Acurácia > 50%)
- kNN trás melhores resultados para o problema citado nesse trabalho
  - Com 3 vizinhos apresentou melhor resultado
  - Aprendizado incremental
  - Rápido para classificar
  - Menor quantidade de processamento
- Árvore de Decisão também pode ser adotado
  - O algoritmo tem acurácia próxima ao kNN
  - Gera uma visualização do resultado