Instituto Tecnológico de Aeronáutica

Divisão de Engenharia de Computação

CTC-17: Inteligência Artificial

Projeto III - Aprendizado de Máquina

Aluno: Luiz Angel R. Rafael

Profº. responsável:
Paulo André Castro

28 de Outubro de 2016

1 Resultados Obtidos

1.1 Descrição dos classificadores

Os dois classificadores implementados foram: árvore de decisão e *a priori*. Na codificação, ambos os classificadores recebem dados dos filmes e das classificações, porém, apenas o classificador baseado em árvore de decisão levou em conta as informações dos usuários.

O algoritmo para construir a árvore de decisão utilizado foi o ID3[1], que foi o mesmo algoritmo ensinado em aula, e os atributos utilizados foram: gênero, idade e ocupação das pessoas e gênero dos filmes.

Como o classificador *a priori* leva em conta apenas as informações dos filmes (média truncada de todas as classificações para cada filme) e a árvore de decisão leva em conta, também, as informações dos usuários, é esperado que a taxa de acertos do classificador baseado em árvore de decisão seja maior.[2]

1.2 Dados e Resultados da comparação

Saída 1: *Resultado para o classificador baseado em árvore de decisão*

```
---------------------- DECISION TREE CLASSIFIER ----------------------
Twelve Monkeys (1995) : 4
Seven (Se7en) (1995) : 4
Miserables, Les (1995) : 4
Fight Club (1999) : 4
Godfather, The (1972) : 4
2001: A Space Odyssey (1968) : 4
Metropolis (1926) : 4
E.T. the Extra-Terrestrial (1982) : 4
Dancer in the Dark (2000) : 4
Psycho (1960) : 4
Gladiator (2000) : 4
Mission: Impossible 2 (2000) : 4
Exorcist, The (1973) : 4
Addams Family, The (1991) : 4
Joe's Apartment (1996) : 4
Willy Wonka and the Chocolate Factory (1971) : 4
Blue Lagoon, The (1980) : 4
Anaconda (1997) : 4
Gattaca (1997) : 4
Devil's Advocate, The (1997) : 4
Anastasia (1997) : 4
Titanic (1997) : 4

Execution time: 470.220999956 seg
```

Saída 2: *Resultado para o classificador a priori*

```
---------------------- A PRIORI CLASSIFIER ----------------------
Twelve Monkeys (1995) : 4
Seven (Se7en) (1995) : 4
Miserables, Les (1995) : 4
Fight Club (1999) : 4
Godfather, The (1972) : 5
2001: A Space Odyssey (1968) : 4
Metropolis (1926) : 4
E.T. the Extra-Terrestrial (1982) : 4
Dancer in the Dark (2000) : 4
Psycho (1960) : 4
Gladiator (2000) : 4
Mission: Impossible 2 (2000) : 3
Exorcist, The (1973) : 4
Addams Family, The (1991) : 3
Joe's Apartment (1996) : 2
Willy Wonka and the Chocolate Factory (1971) : 4
Blue Lagoon, The (1980) : 2
Anaconda (1997) : 2
Gattaca (1997) : 4
Devil's Advocate, The (1997) : 3
Anastasia (1997) : 4
Titanic (1997) : 4

Execution time: 5.60499978065
```

Saída 3: *Classificações dadas pelo aluno*

```
Twelve Monkeys (1995) : 4
Seven (Se7en) (1995) : 5
Miserables, Les (1995) : 5
Fight Club (1999) : 5
Godfather, The (1972) : 5
2001: A Space Odyssey (1968) : 4
Metropolis (1926) : 4
E.T. the Extra-Terrestrial (1982) : 3
Dancer in the Dark (2000) : 5
Psycho (1960) : 4
Gladiator (2000) : 2
Mission: Impossible 2 (2000) : 3
Exorcist, The (1973) : 4
Addams Family, The (1991) : 3
Joe's Apartment (1996) : 1
Willy Wonka and the Chocolate Factory (1971) : 4
Blue Lagoon, The (1980) : 1
Anaconda (1997) : 1
Gattaca (1997) : 4
Devil's Advocate, The (1997) : 4
Anastasia (1997) : 4
Titanic (1997) : 3
```

Tabela 1: *Matriz de confusão esperada (p_e = 10/22 = 0.4545)*

Real \ Previsto	1	2	3	4	5
1	1	2	0	0	0
2	0	1	0	0	0
3	0	1	2	1	0
4	0	0	1	5	3
5	0	0	0	4	1

Tabela 2: *Matriz de confusão do classificador baseado em árvore de decisão (p_o = 9/22 = 0.409)*

Real \ Previsto	1	2	3	4	5
1	0	0	0	3	0
2	0	0	0	1	0
3	0	0	0	4	0
4	0	0	0	9	0
5	0	0	0	5	0

Tabela 3: *Matriz de confusão do classificador a priori (p_o = 11/22 = 0.5)*

Real \ Previsto	1	2	3	4	5
1	0	3	0	0	0
2	0	0	0	1	0
3	0	0	2	2	0
4	0	0	1	8	0
5	0	0	0	4	1

Para o classificador a baseado em árvore de decisão, tem-se que:

Tabela 4: *Estatísticas do classificador baseado em árvore de decisão*

Taxa de acerto	$9/22 = 40.9\%$
Erro quadrático médio	$40/22 = 1.82$
Estatística kappa	$\frac{(p_o - p_e)}{(1 - p_e)}$ = -8.33%

Para o classificador *a priori*, tem-se que:

Tabela 5: *Estatísticas do classificador a priori*

Taxa de acerto	$11/22 = 50\%$
Erro quadrático médio	$14/22 = 0.64$
Estatística kappa	$\frac{(p_o - p_e)}{(1 - p_e)} = 8.33\%$

1.3 Discussão e sugestão de melhorias para o classificador

Pelos resultados acima obtidos, as avaliações sugeridas pelos dois classificadores foram bem diferentes, porém as taxas de acertos foram próxima entre eles. Como na base de dados havia, em sua maioria, classificações com *rating* 4 e 3, é possível que isso tenha influenciado o resultado das folhas da árvore a ter, em sua maioria, esses valores. Além disso, pode ser que a maior parte de filmes com um determinado gênero tivessem sido classificados como bons que como ruins, e como o atributo gênero do filme foi levado em consideração, isso afetou a previsão do algoritmo.

Uma situação possível na árvore de decisão seria: apesar de filmes como Blue Lagoon serem considerados ruins, eles levam a mesma classificação que filmes considerados bons, como The Godfather, por exemplo, pois ambos possuem o gênero Drama.

Para os classificador baseado em árvore de decisão, seria interessante implementar uma validação cruzada, sendo os exemplos divididos entre conjuntos de treinamento e de validação (melhoria no aprendizado dos algoritmos).

Já para o classificador *a priori*, mesmo que ele seja bastante simples, ele funciona bem para este problema de classificação de filmes, já que a as avaliações da maioria sobre algum filme impactam bastante numa avaliação futura. Por isso, ele se mostra bem eficiente, analisando a sua complexidade e o tempo de execução do algoritmo.

2 Conclusões

A formulação do problema é muito importante para decidir qual classificador utilizar e quais terão melhor custo-benefício: se compensa o custo de processamento do algoritmo dependendo da previsão que o classificador realize.

Para a questão de classificação de filmes, percebeu-se que, por causa do tempo de execução e da taxa de acertos da implementação com árvore de decisão, o classificador *a priori* seria mais vantajoso. Enquanto o classificador *a priori* demora cerca de segundos para executar, o baseado em árvore demora alguns minutos, mostrando uma previsão pior, com menor taxa de acerto e menor estatística kappa. Porém, para melhor análise de ambas as implementações, seriam necessários mais testes com outras pessoas e maior quantidade de filmes.

3 Descrição da Implementacao

O código do projeto foi feito na linguagem Python 2.7.11. Na pasta do projeto, encontra-se a seguinte composição:

```
....
data
    movies.dat
    ratings.dat
    users.dat
    README
source
    a_priori
        helper.py
        main.py
    decision_tree
        classes.py
        helper.py
        main.py
```

Os arquivos principais são aqueles cujos nomes começam com main. Os arquivos de classes (que especificam uma classe em Python) são todos exceto os principais e os arquivos helper.py. Estes são arquivos com funções auxiliares utilizadas por arquivos principais e/ou de classe.

Para obter resultado de algum algoritmo, basta executar python <NomeArquivoPrincial>:

```
> python main.py
```

Referências

[1] Decision Tree - Classification. [Online]. Available: http://www.saedsayad.com/decision_tree.htm

[2] CTC-17 - Inteligência Artificial. [Online]. Available: http://www.comp.ita.br/~pauloac/ctc17/index.html

4 Apêndice

O repositório *Git* que possui todo o código do projeto pode ser encontrado em

`https://github.com/Luizangel50/CTC-17_Lab3`