Recuperação inteligente da informação
Relatório atividade E-1.5

André Pacheco

Doutorado em Ciência da Computação
Programa de pós-graduação em informática
Universidade Federal do Espírito Santo

Conteúdo

1 Introdução

A atividade E-1.5 tem como objetivo a realização da classificação da famosa base de dados da IRIS [1] utilizando o algoritmo k-vizinhos mais próximos (KNN) (Fukunaga and Narendra, 1975) internamente implementado no aLine. Portanto, neste relatório são descritos os procedimentos para realizar tal tarefa.

2 Configuração dos dados para o aLine

Para utilizar o KNN do aLine, primeiramente é necessário converter as *features* da base de dados da IRIS para o formato *Matrix Market*. Sendo assim, é criado o arquivo `iris_data.mtx` para armazenar esses valores. Além disso é necessário gerar os índices de cada amostra para treinamento e teste. Neste trabalho foi separado 30% para testes e 70% para treinamento. Esses valores são gerados aleatoriamente e armazenados nos arquivos `iris_test.txt` e `iris_train.txt`, respectivamente. Por fim, no arquivo `iris_class.txt` são armazenados todos os *labels* da base relacionados a cada amostra disponível na *matrix market*.

Gerados todos os arquivos mencionados anteriormente, o aLine é executado da seguinte forma:

[1] https://archive.ics.uci.edu/ml/datasets/iris

```
$ aLine --classifier --algorithm knn --features iris_data.mtx
--train iris_train.txt --test iris_test.txt --labels iris_class.txt
 -k $valK$ -o iris_class_predict.txt
```

onde a variável `$valK$` é o valor do parâmetro k do KNN e o arquivo `iris_class_predict.txt` são as amostras classificadas pelo KNN.

3 Resultados de classificação

Os resultados da classificação da base de dados da IRIS pelo aLine/KNN são descritos na Tabela 1. Como pode ser observado, o valor de k foi variado, e para cada valor, o algoritmo foi executado 30 vezes (com cada partição escolhida de forma aleatória em cada execução). O desempenho é mensurado pela acurácia de classificação [2].

Tabela 1. Acurácia de classificação da IRIS de acordo com o KNN

k	**Média (%)**	**Devio (%)**
3	95.55	2.75
5	96.07	2.60
7	95.85	2.34
9	96.66	2.54
11	96.96	2.25
13	97.33	2.17
17	95.48	3.27
23	94.44	4.01

Observado os resultados apresentados na Tabela 1, pode-se notar que valores de k até 13, os resultados são bem próximos. Todavia, quando k assume valores como 17 e 23 o kNN começa a deteriorar o seu resultado. Isso ocorre por conta do *trade-off* entre variância e *bias*, ou seja, com k suficientemente grande, o kNN 'decora' a base de dados, diminuindo a variância, porém 'decora' os padrões de treinamento, favorecendo o *bias* (Friedman et al., 2001).

Além da classificação, também foi calculado a densidade espacial total e média [3] de acordo com Salton et al. (1975). O valores das densidades são descritos na Tabela 2.

[2] As métricas de *recall* e *precision* retornadas pelo aLine não estão sendo calculadas de forma correta. Por isso, foi implementado uma função para ler o arquivo de saída e informar a acurácia da classificação. Esse problema já foi reportado para o monitor da disciplina.

[3] Outro ponto a destacar é que neste modo, o aLine não retorna os valores de similaridade das amostras. Neste caso, foi necessário uma função para calcular as similaridades par a par de cada a amostra para calcular a densidade espacial. Sendo assim, é os valores das similaridade são calculadas duas vezes (uma no KNN interno e outra no script da atividade), o que poderia ser evitado se o aLine retornasse esses valores neste modo de operação.

Tabela 2. Densidade espacial da base de dados IRIS

Densidade espacial	**Valor**
Total	21351.87
Média	0.94

Como pode ser observado, de acordo com a densidade espacial media, de maneira geral, as amostras da IRIS possui uma similaridade alta.

4 Conclusão

Nesta atividade a base de dados da IRIS foi classificada utilizando o aLine e sua implementação interna do KNN. Como pode ser observado pelos resultados, os valores de k foram variados e a classificação foi realizada de maneia satisfatória. Embora o arquivo com o resultado da classificação retornado pelo o aLine é coerente com a base, as métricas apresentadas pelo framework não estão corretas. Além disso, foi apontada uma sugestão, ainda em relação ao aLine, para facilitar o cálculo da densidade espacial neste modo de operação.

Bibliografia

Friedman, J., Hastie, T., Tibshirani, R., 2001. The elements of statistical learning. volume 1. Springer series in statistics New York.

Fukunaga, K., Narendra, P.M., 1975. A branch and bound algorithm for computing k-nearest neighbors. IEEE Transactions on Computers 100, 750–753.

Salton, G., Wong, A., Yang, C.S., 1975. A vector space model for automatic indexing. Communications of the ACM 18, 613–620.