ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-HCM
KHOA CÔNG NGHỆ THÔNG TIN

KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG
Bài tập Thực hành 3

CLASSIFICATION

Nhóm thực hiện

1. Hồng Thanh Hoài 1612855
2. Huỳnh Minh Huấn 1612858

Giáo viên lý thuyết
PGS.TS Lê Hoài Bắc

Giáo viên hướng dẫn
Nguyễn Ngọc Thảo

Tháng 05 năm 2019

Lời cảm ơn

Trong quá trình thực hiện bài tập này, nhóm chúng em đã nhận được rất nhiều sự giúp đỡ cũng như hỗ trợ từ các thầy cô Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM và các bạn bè trong trường. Nhóm chúng em xin bày tỏ lòng cảm ơn chân thành đến mọi người vì đã hướng dẫn, chỉ bảo rất tận tình.

Đặc biệt, nhóm chúng em xin bày tỏ lòng biết ơn sâu sắc đến các thầy cô khoa Công nghệ thông tin, cụ thể hơn là thầy Lê Hoài Bắc đã giảng dạy rất kĩ lưỡng để chúng em nắm rõ kiến thức và cô Nguyễn Ngọc Thảo đã hướng dẫn chúng em thực hiện bài tập này rất nhiệt tình.

Một lần nữa, chúng em xin bày tỏ lòng biết ơn sâu sắc đến với các thầy cô và bạn bè.

Tháng 05 năm 2019,
Đại học Khoa học Tự nhiên, ĐHQG-HCM.

Mục lục

1 Giới thiệu nhóm và phân công công việc

1.1 Giới thiệu nhóm

Nhóm gồm 2 thành viên.

STT	Họ và tên	MSSV	Email	SĐT
1	Hồng Thanh Hoài	1612855	hthoai1006@gmail.com	0965596807
2	Huỳnh Minh Huấn	1612858	minhhuanhuynh289@gmail.com	0824540646

1.2 Phân công công việc

STT	Họ và tên	Công việc
1	Hồng Thanh Hoài	Câu 1, 2, 3, 4.
2	Huỳnh Minh Huấn	Câu 5, 6, 7, 8.

2 Nội dung

2.1 Câu 1

Bảng 1: Bảng thống kê độ chính xác phân lớp của mỗi giải thuật trên tập dữ liệu `mushroom.arff`.

Giải thuật	**Accuracy**	**Detailed Accuracy By Class**							
		Class edible				**Class poisonous**			
		TP Rate	**FP Rate**	**Precision**	**Recall**	**TP Rate**	**FP Rate**	**Precision**	**Recall**
LR	99.7537	0.998	0.003	0.998	0.998	0.997	0.002	0.997	0.997
J48	99.8768	1.000	0.003	0.998	1.000	0.997	0.000	1.000	0.997
IBk (KNN=1)	99.8768	0.998	0.000	1.000	0.998	1.000	0.002	0.997	1.000
IBk (KNN=4)	99.1379	0.988	0.005	0.995	0.988	0.995	0.012	0.988	0.995

2.2 Câu 2

Bảng 2: Hai giải thuật A và B.

(a) Giải thuật A.

Accuracy: 99.8768%		
Class	TP Rate	FP Rate
edible	1.000	0.003
poisonous	0.997	0.000

(b) Giải thuật B.

Accuracy: 99.8768%		
Class	TP Rate	FP Rate
edible	0.998	0.000
poisonous	1.000	0.002

Có thể thấy giải thuật A tương ứng với J48, còn giải thuật B tương ứng với IBk (kNN=1) ở câu 1. Ta có nhận xét sau:

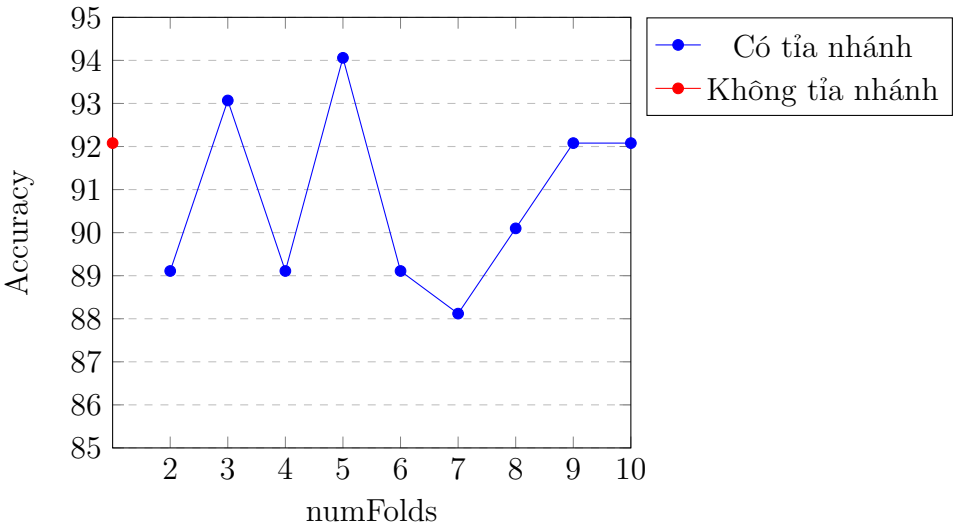
- Giải thuật A:
  - Lớp *edible* có $Precision = 0.998$, nghĩa là tỉ lệ nấm phân vào lớp *edible* là đúng trên tổng số nấm được phân vào lớp *edible* bằng 0.998.
  - Lớp *poisonous* có $Precision = 1$, nghĩa là tỉ lệ nấm phân vào lớp *poisonous* là đúng trên tổng số nấm được phân vào lớp *poisonous* bằng 1.
- Giải thuật B:
  - Lớp *edible* có $Precision = 1$, nghĩa là tỉ lệ nấm phân vào lớp *edible* là đúng trên tổng số nấm được phân vào lớp *edible* bằng 1.

– Lớp *poisonous* có $Precision = 0.997$, nghĩa là tỉ lệ nấm phân vào lớp *poisonous* là đúng trên tổng số nấm được phân vào lớp *poisonous* bằng 0.997.

Đối với bài toán nhận diện nấm độc, giải thuật B tốt hơn vì có *Precision* của lớp *edible* cao hơn. Ở đây, ta thà chấp nhận có một vài loại nấm ăn được bị phân vào lớp có độc (giải thuật B), còn hơn là có độc mà được phân vào lớp ăn được (giải thuật A) sẽ rất nguy hiểm.

2.3 Câu 3

Biểu đồ thể hiện độ chính xác phân lớp.



2.4 Câu 4

Tham số *numFolds* giúp ta xác định lượng dữ liệu sử dụng cho việc tỉa nhánh giảm lỗi. Số lượng *numFolds* ta điền vào sẽ được dùng cho việc tỉa nhánh, phần còn lại dùng để xây dựng cây. Tỉa nhánh giúp giảm bớt số lượng lá cây trung gian để cây quyết định đơn giản và dễ hiểu hơn.

Ảnh hưởng của *numFolds* đến độ chính xác của cây quyết định là không tuyến tính. Ở đây độ chính xác của cây quyết định thay đổi lên xuống khi ta tăng dần *numFolds*. Độ chính xác cao nhất có được khi $numFolds = 5$ – ở khoảng giữa.

2.5 Câu 5

Việc tỉa nhánh giảm lỗi giúp tăng độ chính xác của cây quyết định khi chọn được *numFolds* phù hợp. Dựa vào đồ thị, ta có thể thấy khi *numFolds* quá nhỏ – nghĩa là lượng dữ liệu dùng cho việc tỉa nhánh ít – thì kết quả không tốt. Tương tự, khi *numFolds* quá lớn – nghĩa là lượng dữ liệu dùng cho việc xây dựng cây ít – thì cây xây

dựng được không tốt, nên hiệu quả mà việc tỉa nhánh mang lại không đáng kể. Do đó, cần chọn một giá trị trung dung để cây xây dựng được tốt, và hiệu quả mà việc tỉa nhánh giảm lỗi mang lại có thể cải thiện tốt độ chính xác của cây quyết định.

Việc tỉa nhánh giảm lỗi không chỉ giúp đơn giản hóa cây quyết định, mà còn giúp tăng độ chính xác của cây xây dựng được do giảm nguy cơ dữ liệu "quá khớp" với *training set*. Nghĩa là cây xây dựng quá phức tạp nên có thể đúng 100% với *training set* nhưng khi dùng cho *test set* thì kết quả xấu. Đây chính là hiện tượng *overfitting* trong máy học. Vậy, việc tỉa nhánh giúp giải quyết hiện tượng *overfitting* khi dùng thuật toán J48 để học.

2.6 Câu 6

Nhóm chọn kNN, k = 4 (weka.classifiers.lazy.IBk).

Bảng 3: Bảng đánh giá độ chính xác của của kNN bằng 10-fold cross validation.

Class	**TP Rate**	**FP Rate**	**Precision**	**Recall**
A	0.991	0.000	0.991	0.991
B	0.944	0.003	0.915	0.944
C	0.969	0.001	0.981	0.969
D	0.959	0.003	0.920	0.959
E	0.946	0.003	0.928	0.946
F	0.927	0.002	0.938	0.927
G	0.932	0.002	0.953	0.932
H	0.856	0.003	0.909	0.856
I	0.964	0.002	0.961	0.962
J	0.945	0.001	0.962	0.945
K	0.906	0.003	0.929	0.917
L	0.963	0.001	0.975	0.963
M	0.972	0.001	0.987	0.972
N	0.950	0.002	0.962	0.950
O	0.961	0.004	0.910	0.961
P	0.929	0.002	0.958	0.929
Q	0.955	0.002	0.955	0.955
R	0.936	0.004	0.900	0.936
S	0.975	0.001	0.980	0.975
T	0.970	0.001	0.966	0.970
U	0.988	0.001	0.980	0.988
V	0.969	0.002	0.949	0.969
W	0.979	0.001	0.985	0.979
X	0.963	0.002	0.960	0.963
Y	0.978	0.001	0.977	0.978
Z	0.978	0.001	0.986	0.978

2.7 Câu 7

- Về giải thuật nhóm đã chọn:
  - Giải thuật IBk (Instance-based learning algorithms) là nhóm giải thuật phân lớp dựa trên thể hiện, sử dụng bộ phân lớp kNN (K-nearest neighbours classifier). Một mẫu mới (dữ liệu test) sẽ được gán vào lớp có nhiều mẫu trong số k mẫu gần với nó nhất.
  - Giải thuật `kNN(D, d, k)`:
    * Tính khoảng cách (định nghĩa khoảng cách theo độ đo như Euclidian, Cosin,...) của d với tất cả các mẫu trong D.
    * Chọn k mẫu trong D "gần" d nhất, ký hiệu là P.
    * Gán d vào lớp có nhiều mẫu nhất trong số k mẫu láng giềng đó.
- Giải thuật IBk với bộ phân lớp kNN và k=4 vì đây là giải thuật đơn giản, dễ hiểu, thực thi tốt trong nhiều tình huống. Thời gian train model và test trong Weka thực hiện nhanh, cho độ chính xác cao (theo thực nghiệm của nhóm). Trong quá trình thử nghiệm, có hai giải thuật nổi trội khác là SMO và Random Forest. Tuy nhiên, SMO có thời gian thực thi quá lâu (gần nửa tiếng), còn Random Forest mặc dù cho Accuracy cao hơn kNN khi xây dựng mô hình nhưng lại thấp hơn kNN khi thử với một vài bộ test ngoài training set. Do đó nhóm quyết định chọn kNN là mô hình tối ưu nhất với bài toán.

2.8 Câu 8

- Bộ tham số được dùng để thực nghiệm:
  - $KNN = 4$
  - $batchSize = 100$
  - $distanceWeighting = Weight\ by\ 1/distance$
  - $meanSquared = true$
  - $nearestNeighbourSearchAlgorithm = KDTree$
- Khi chọn $distanceWeighting = Weight\ by\ 1/distance$ thì tỉ lệ phân lớp chính xác hơn. Chọn $nearestNeighbourSearchAlgorithm$ là $KDTree$ thì thời gian train model và đánh giá mô hình cũng như kiểm thử nhanh hơn.

3 Đánh giá

STT	Nội dung	Hoàn thành
1	Câu 1	100%
2	Câu 2	100%
3	Câu 3	100%
4	Câu 4	100%
5	Câu 5	100%
6	Câu 6	100%
7	Câu 7	100%
8	Câu 8	100%
Mức độ hoàn thành tổng thể của bài tập:		**100%**

Tài liệu

[1] Slide lý thuyết.

[2] Trang chủ của Weka: http://www.cs.waikato.ac.nz/ml/weka/

[3] J. Han and M. Kamber, *Data Mining, Concepts and Techniques, Second Edition*