Werkblad week 4

d.r.m.langers@pl.hanze.nl

14 oktober 2021

1 Werken met Weka

Dit werkblad bevat in tegenstelling tot werkbladen 1 t/m 3 geen nieuwe theorie, maar geeft je de gelegenheid om de theorie toe te passen en te oefenen met Weka.

Gebruik waar mogelijk de standaardinstellingen (bv. parameters van algoritmen, of waarden van random seeds), tenzij het voor de oplossing noodzakelijk is om deze te wijzigen. Bijbehorende databestanden zijn te downloaden vanaf BlackBoard.

1.1 Medische diagnostiek

De bestanden AUTISM.ARFF, BREAST-CANCER.ARFF, DIABETES.ARFF, HEART.ARFF, HEPATITIS.ARFF, SCHIZO.ARFF en SICK.ARFF bevat gegevens omtrent de diagnostiek van patiënten met diverse aandoeningen. In alle gevallen bevat het laatste attribuut dichotome klasse-labels. Pas de onderstaande tien algoritmen op deze datasets toe (met standaardinstellingen, tenzij anders vermeld) en vergelijk hun prestaties op grond van tienvoudige 10-fold cross-validation.

- Zero-R
- One-R
- J48-Tree
- Naive Bayes
- k-Nearest Neighbor
- Classificatie via Clustering, gebaseerd op (Simple) k-Means
- Random Forest
- Logistische regressie
- AdaBoost

- Vote, gebaseerd op majority voting, met alle negen bovenstaande algoritmen als base-classifiers.

(Hint: installeer het Classification via Clustering algoritme indien nodig vanuit Tools > Package manager als onderdeel van het classificationViaClustering pakket.) Beantwoord daarna de onderstaande vragen.

Opgave 1. Waarom kan het Id3-Tree algoritme niet in bovenstaand lijstje worden meegenomen? (Hint: installeer indien nodig vanuit Tools > Package manager het pakket simpleEducationalLearningSchemes.)

Opgave 2. Welke combinatie van dataset en algoritme behaalt de hoogste nauwkeurigheid en welke andere combinatie de laagste? Hoe groot zijn deze maximale en minimale nauwkeurigheden?

Opgave 3. Welke combinatie(s) van dataset en algoritme geven een significant hogere error rate dan Zero-R behaalt op diezelfde dataset?

Opgave 4. Welk algoritme heeft, gemiddeld over alle zeven datasets, de hoogste oppervlakte onder de Receiver Operating Characteristic (ROC) curve? Hoe groot is deze gemiddelde Area Under the Curve?

1.2 Granulocyten

Het bestand GRANULOCYTES.ARFF bevat een dataset met gemeten celdiameters (d, in μm) en celvolumes (V, in μm^3) van dertig humane witte bloedcellen. Deze zijn onder te verdelen in twee groepen: neutrofiele en eosinofiele granulocyten.

Gebruik eerst de visualisatie-opties van Weka om een indruk te krijgen van de spreiding van de datapunten. Pas vervolgens logistische regressie toe om deze dataset te fitten aan de functie $p = g(a_0 + a_1 \cdot d + a_2 \cdot V)$, waarbij g de logistische functie $g(x) = \frac{1}{1+e^{-x}}$ aanduidt.

Opgave 5. Wat zijn de waarden van de resulterende coëfficiënten a_i?

Opgave 6. Hoe groot is de nauwkeurigheid (d.w.z. accuracy) van het voorgaande logistische regressie model op grond van resubstitutie van de trainingsdata? En hoe groot is deze op grond van Leave-One-Out kruisvalidatie?

Opgave 7. Hoe groot is de oppervlakte onder de Receiver Operating Characteristic curve van het voorgaande logistische regressie model op grond van resubstitutie van de trainingsdata? En hoe groot is deze op grond van leave-one-out kruisvalidatie?

Opgave 8. Is deze dataset lineair separabel, dat wil zeggen zijn de klassen perfect te onderscheiden middels een rechte lijn? Motiveer je antwoord.

Stel vervolgens de cost van een eosinofiele granulocyt die mis-geclassificeerd wordt als neutrofiele granulocyt vier keer zo hoog als die van een neutrofiele granulocyt die mis-geclassificeerd wordt als eosinofiele granulocyt.

Opgave 9. Leidt cost-sensitive classification voor het logistische model met leave-one-out kruisvalidatie tot een verandering van de confusion matrix t.o.v. het eerdere resultaat bij gelijke costs? En hoe zit dit voor cost-sensitive learning? Leg in je eigen woorden uit wat het verschil is tussen de begrippen cost-sensitive classification en cost-sensitive learning.

Gebruik het Cascade (Simple) k-Means algoritme in Weka om deze data te clusteren op grond van d en V. (Hint: installeer dit algoritme indien nodig vanuit Tools > Package manager als onderdeel van het cascadeKMeans pakket.)

Opgave 10. Hoeveel clusters worden door dit algoritme als optimaal gerapporteerd?

Gebruik het reguliere (Simple) k-Means algoritme in Weka om deze data te clusteren in twee clusters op grond van gestandaardiseerde(!) waarden voor d en V en maak hierbij gebruik van de taxicab-metriek. Evalueer de overeenkomst van de clusters met de klassen.

Opgave 11. Wat zijn de gestandaardiseerde coördinaten van de gevonden cluster centroïden? Hoeveel instances worden aan een verkeerd cluster toegekend?

1.3 Wie Is Het?

In het gezelschapsspel "Wie is het?" (zie https://nl.wikipedia.org/wiki/Wie_is_het%3F) dient de identiteit van een persoon op een kaartje te worden geraden aan de hand van een reeks ja/nee-vragen. Er zijn 24 kaartjes met verschillende personen. Gegeven de kenmerken in de tabel hieronder, eveneens beschikbaar in het bestand GUESSWHO.ARFF, is het doel om een beslisboom te genereren die leidt tot "optimaal spel" waarbij je in zo weinig mogelijk vragen tot de juiste identiteit komt.

		ALEX	ALFRED	ANITA	ANNE	BERNARD	BILL	CHARLES	CLAIRE	DAVID	ERIC	FRANS	GEORGE	HERMAN	JOE	MARIA	MAX	PAUL	PETER	PHILIP	RICHARD	ROBERT	SAM	SUSAN	TOM	TOTAL
HAIR STYLE	HAIR PARTITION	NO	YES	YES	NO	NO	NO	YES	NO	NO	NO	NO	NO	NO	NO	NO	NO	YES	NO	NO	YES	NO	YES	NO	6
	CURLY HAIR	NO	NO	NO	YES	NO	NO	NO	NO	NO	NO	YES	NO	YES	YES	NO	YES	NO	NO	YES	NO	NO	NO	NO	NO	6
	HAT	NO	NO	NO	NO	NO	YES	NO	NO	YES	NO	YES	NO	YES	NO	NO	YES	NO	NO	NO	NO	NO	NO	NO	NO	5
	BALD	NO	NO	NO	NO	NO	YES	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO	NO	NO	NO	NO	YES	NO	YES	NO	5
	HAIR STUFF	NO	NO	YES	NO	YES	NO	NO	YES	NO	NO	NO	YES	NO	NO	YES	NO	NO	NO	NO	NO	NO	NO	NO	NO	5
	LONG HAIR	NO	YES	YES	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO	NO	NO	NO	NO	YES	NO	4
HAIR COLOUR	GINGER HAIR	NO	YES	NO	NO	NO	YES	NO	YES	NO	NO	YES	NO	YES	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	5
	WHITE HAIR	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO	NO	YES	YES	NO	NO	NO	YES	YES	NO	5
	BROWN HAIR	NO	NO	NO	NO	YES	NO	NO	NO	NO	NO	NO	NO	NO	NO	YES	YES	NO	NO	NO	YES	YES	NO	NO	NO	5
	BLOND HAIR	NO	NO	YES	NO	NO	NO	YES	NO	YES	YES	NO	NO	NO	YES	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	5
	BLACK HAIR	YES	NO	NO	YES	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO	NO	YES	4
FACIAL ATTRIBUTES	BIG MOUTH	YES	NO	NO	NO	NO	NO	YES	NO	YES	YES	NO	YES	NO	NO	NO	YES	NO	YES	YES	NO	YES	NO	YES	NO	10
	BIG NOSE	NO	NO	NO	NO	YES	NO	NO	NO	NO	NO	NO	YES	NO	NO	YES	NO	YES	NO	YES	NO	NO	YES	NO	NO	5
	RED CHEEKS	NO	NO	YES	NO	NO	YES	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	YES	NO	NO	YES	NO	YES	NO	5
	BLUE EYES	NO	YES	YES	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	YES	NO	YES	NO	YES	NO	NO	YES	5
	SAD LOOKING	NO	YES	NO	NO	YES	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO	4
FACIAL HAIR	FACIAL HAIR	YES	YES	NO	NO	NO	YES	YES	NO	YES	NO	NO	NO	NO	NO	NO	YES	NO	YES	YES	YES	NO	NO	NO	NO	8
	MOUSTACHE	YES	YES	NO	NO	NO	NO	YES	NO	NO	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO	YES	NO	NO	NO	NO	5
	BEARD	NO	NO	NO	NO	NO	YES	NO	NO	YES	NO	NO	NO	NO	NO	NO	NO	NO	YES	YES	NO	NO	NO	NO	NO	4
OTHERS	GLASSES	NO	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO	NO	NO	YES	NO	NO	YES	NO	NO	NO	NO	YES	NO	YES	5
	EAR RINGS	NO	NO	NO	YES	NO	NO	NO	NO	NO	NO	NO	NO	NO	YES	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	2
	FEMALE	NO	NO	YES	YES	NO	NO	NO	YES	NO	NO	NO	NO	NO	YES	NO	NO	NO	NO	NO	NO	NO	YES	NO		5

https://geekandsundry.com/wp-content/uploads/2016/05/GuessWhoCharacters4-680x412.png

Opgave 12. Als je dit probleem als een classificatieprobleem beschouwt, wat is dan de variabele in de dataset die de klasselabels bevat? Beschrijf het type van deze variabele.

Opgave 13. Wat is de entropie van de verdeling van mogelijke klasse-labels bij aanvang van het spel, nog vóór er vragen gesteld zijn? Motiveer op grond hiervan hoeveel ja/nee-vragen er gemiddeld minstens gesteld zullen moeten worden om de identiteit van de persoon op een kaartje te kunnen achterhalen.

Pas het Id3-Tree algoritme in Weka toe op deze dataset om een geschikte beslisboom te genereren. (Hint: installeer dit algoritme indien nodig vanuit Tools > Package manager als onderdeel van het simpleEducationalLearningSchemes pakket.)

Opgave 14. Hoe groot is de resubstitution error rate van de geconstrueerde beslisboom? Hoe groot is de error rate wanneer je kruisvalidatie uitvoert middels 10-fold cross-validation? Verklaar deze beide uitkomsten.

Opgave 15. Wat is volgens de Id3-Tree de beste eerste vraag om te stellen (d.w.z. het attribuut in de root node van de beslisboom)? Bereken de verwachtingswaarde van de entropie van mogelijke klasselabels ná beantwoording van deze eerste vraag.

Opgave 16. Stel dat de gegenereerde tree wordt gebruikt om een fictieve nieuwe instance te classificeren met de onderstaande eigenschappen. Missing values (aangeduid met "?") dien je daarbij te behandelen zoals het J48-Tree algoritme dat doet. Beredeneer stap voor stap welke klasse aan deze instance zal worden toegekend.

Name:	?		
Hair partition:	No	Big mouth:	No
Curly hair:	Yes	Big nose:	No
Hat:	No	Red cheeks:	No
Bald:	?	Blue eyes:	?
Hair stuff:	No	Sad looking:	No
Long hair:	No	Facial hair:	?
Ginger hair:	No	Moustache:	No
White hair:	No	Beard:	Yes
Brown hair:	?	Glasses:	?
Blond hair:	No	Ear rings:	No
Black hair:	?	Female:	No

Opgave 17. Bepaal door middel van exhaustive search aan welke subset van attributen Correlation-based Feature Selection de hoogste waardering geeft. Welke attributen kunnen volgens CFS het beste worden verwijderd, en hoeveel subsets zijn er geëvalueerd om tot die conclusie te komen? (Hint: installeer indien nodig vanuit Tools > Package manager het pakket attributeSelectionSearchMethods.)

In de volgende twee deelopgaven dien je het attribuut oogkleur (d.w.z. *blue eyes*) te voorspellen aan de hand van alle andere beschikbare gegevens in de dataset. Dat wil zeggen, stel oogkleur in als de klasse.

Opgave 18. Vergelijk de beslisboom voor het bepalen van oogkleur volgens het Id3-Tree algoritme met die van het J48-Tree algoritme zonder forward/backward pruning. Welke attributen worden gekozen in de root-nodes van deze twee trees? Leg uit welk verschil tussen deze algoritmen veroorzaakt dat de beide boomstructuren er in dit geval anders uitzien.

Opgave 19. Voor welke optimale waarde van de parameter k behaalt het k-Nearest Neighbor algoritme de hoogste nauwkeurigheid in het voorspellen van oogkleur (10-fold kruis-gevalideerd op alleen de trainingsdata)?