



WEKA-N EGINDAKO 2.PRAKTIKARI BURUZKO DOKUMENTAZIOA

| | |
|-------------------|--|
| IRAKASGAIA | <i>Datu Meatzaritza</i> |
| IRAKASLEA | <i>Basilio Sierra Araujo & Naiara Aginako Bengoa</i> |
| IKASTURTEA | <i>2022-2023</i> |
| EGILEA | <i>Oier Álvarez Parada</i> |
| DATA | <i>2022/11/17</i> |

AURKIBIDEA

| | |
|---|---|
| <i>AURKIBIDEA</i> | 2 |
| SARRERA | 3 |
| 1.ATALA: ERABILIKO DIREN DATU BASEAK AUKERATU | 3 |
| 2.ATALA: BAGGING ETA BOOSTING | 3 |
| 3.ATALA: VOTE | 4 |
| 4.ATALA: STACKING..... | 4 |
| 5.ATALA: ONDORIOAK..... | 4 |

SARRERA

Praktika honetan, aurreko praktikatik lortutako emaitzak hobetzen saiatuko gara. Horretarako, multisailkatzaile desberdinak erabiliko ditugu, bagging, boosting, vote eta stacking, esate baterako. Multisailkatzaile hauekin, 3 metodo erabiliko ditugu, hau da, hauen barne 3 sailkatzaile arrunta (aurreko praktikako batzuk) sartuko ditugu.

Dokumentazio honekin batera, excell taula bat bidali izan da, aurreko praktikako eta honetako emaitza guztiak dituen, dokumentazioan zehar aipatzen diren emaitzak ikusteko.

1.ATALA: ERABILIKO DIREN DATU BASEAK AUKERATU

Datuak aukeratzeko, gomendatzen da, aurreko praktikan emaitza hoberenak izan dutenak erabiltzea, konkretuki 2 convert hoberenak eta hauekin emaitza onenak izan duten 2 filtroak aukeratzea. Guztira lau datu base izango ditugu (4 .arff, aurreko praktikatik jadanik lortu genituenak). Nire kasuan, ez dut convert hoberena erabili, zeren eta, hoberena, *ordenatuak* karpetakoa baitzen, hau da, berez, convert bat aplikatu ez zaiona. Beraz, hurrengo bi hoberenak erabili ditut, Edge2 eta Enhance (honek *ordenatuak* bezalako emaitzak ematen ditu, gutxi gorabehera).

Edge2-ren kasuan, erabili diren bi filtroak PHOGFilter eta SimpleColorHistogram izan dira, hauek biak batasbesteko %85,5 eta %83,6 -ko asmatze tasekin.

Enhance-ren kasuan, EdgeHistogram eta PHOGFilter erabili dira, hauek, aldiz, %87,9 eta %82,8 -ko asmatze tasekin.

2.ATALA: BAGGING ETA BOOSTING

Multisailkatzaile hauek erabiltzeko, hauen barnean lehengo praktikan erabilitako beste sailkatzaile arrunt bat sartu beharko dizkiegu. Praktika osorako, hurrengo 3 sailkatzaile arruntak erabili ditut: 3-NN, NaiveBayes eta RandomForest. Izan ere, 3 sailkatzaile hauek izan dira emaitza onenak zituztenak. Beraz, lehen aipatutako .arff bakoitza hartuko dugu, eta bagging eta boosting aplikatuko diogu, sailkatzaile bakoitzeko behin. Guztira 24 emaitza lortuko ditugu -2 Convert * 2 Filtro * (3 Bagging iterazio + 3 Boosting iterazio)-.

Emaitza hauek lortzeko, lehenik eta behin Weka-n arff-a kargatu beharko dugu. Ondoren *Classifier* atalera jo beharko dugu eta han, sailkatzailea aukeratzeak, *meta* karpeta barruan dagoen Bagging edota Boosting hartuko dugu. Azkenik, honen konfigurazioan, erabiliko dugu sailkatzaile arrunta sartuko dugu.

3.ATALA: VOTE

Multisailkatzaile hau aurrekoen leku berean aukitu dezakegu eta haren konfigurazioa nahiko antzekoa da. Kasu honetan, ez dugu sailkatzaile bakoitzerako behin exekutatu beharrik, aldiz, hiru sailkatzaileak batera erabiliko dira, beraz, aurrekoen antzera, konfigurazioan 3 sailkatzaileak sartuko ditugu. Beraz, multisailkatzaile honerako, soilik 4 emaitza izango ditugu -2 Convert * 2 Filtro * 1 Iterazio bakarra-.

4.ATALA: STACKING

Honetarako *Vote* multisailkatzailearen pausu berak jarraitu beharko ditugu, baina honen kasuan, J48 eta RandomForest erabiltzea esaten digute. Jadanik RandomForest erabiltzen nuenez, soilik J48 gehitu behar izan da. Gainera *Stacking* multisailkatzailean hiru metasailkatzaile gehitu behar dira (aldi berean ez). Eskatzen digute metasailkatzaile hautariko bat SMO izatea, beraz hau eta beste bi hartu ditut, Logistics eta MultilayerPerceptron.

Guztira *Stacking*-en kasuan 12 emaitza izango ditugu -2 Convert * 2 Filtro * 3 Metasailkatzaile-.

5.ATALA: ONDORIOAK

Behin, multisailkatzaile guztien emaitzak ditugula, zenbait ondorio aterako ditugu. Hasteko, jasotako emaitzak, aurreko praktikako emaitzekin aldaratzen baditugu, hobekuntza asko egon ez dela ikus daiteke, kasu batean izan ezik. Esate baterako, *bagging*-ekin izan ditudan emaitzak, aurreko praktikakoak baino txarragoak izan dira (ez asko, <%1 txarragoak, gehienetan). *Boosting*-ekin gauza bera gertatu da, hobekuntzak %±1-koak inguru izan dira. *Vote*-ekin aldiz, hobekuntza signifikatiboagoak izan ditugu, ia %6-ra heldu dira. Beraz, hemendik ondoriozta dezakegu erabilitako datu baseetan, *Vote* multisailkatzaileak lan hobea egiten duela. *Stacking* multisailkatzailearekin, nahiko emaitza onak jaso dira, *Logistics* metasailkatzailearekin izan ezik, hau ez baita aukera ona izan gure datuentzat, izan ere, %13-ko emaitza okerragoa izatera heldu da. Aldiz, beste bi metasailkatzaileek lan ona egiten dute gure datuekin, hobekuntza (3%, 8%) tartean egon da.

Dena den, *Edge2*-ko *SimpleColor* filtroa destakatu beharko dugu, emaitza guztien artean. Izan ere, honek emaitza nahiko onak izan ditu multisailkatzaile guztiekin (*bagging* izan ezik), bestekin konparatuz. *Boosting*-ek >%3-ko hobekuntza izan du eta *Stacking*-en metasailkatzaile guztiak asmatze-tasa altuak izan dituzte, *Logistics* barne. Beraz, esan dezakegu, erabilitako filtro guztien artean, *SimpleColor* izan dela hoberena, izan ere, gure trafiko seinaleak, gehien bat koloreekin desberdintzen dira.