

6. Sailkatzaileen Konbinaketa

Irakasgaia: Konputazio Zientzien Metodo Matematikoak

Titulazioa: Informatikan Ingeniaria

Konputazio Zientzia eta Adimen Artifiziala saila

Universidad del País Vasco - Euskal Herriko Unibertsitatea

Sailkatzaileen konbinaketa

Zergatik eta nola?

- Zergatik?
 - Zenbait sailkatzailearen iragarpenak konbinatuz **emaitza hobekia lor daitezke** sailkatzaile horien artetik hoberena bakarrik erabiliz baino (azken erabakia hartu aurretik zenbait adituren iritzia konbinatzea) → **multisailkatzaileak**
 - Emaitzak hobetzeko garrantzitsua da **konbinatuko diren sailkatzaileak beren artean desberdinak** izatea; kasu berriak sailkatzerakoan errore-mota desberdinak egitea
- Nola?
 - **Hibridazioak**: sailkatzaile berriaren zatiak beste sailkatzaileez osatuta daude.
 - **Oinarrizko sailkatzaile bera** indusitu entrenamendurako azpimultzo desberdinetatik abiatuz: bagging, boosting
 - **Oinarrizko sailkatzaile desberdinak** indusitu

Hibridazioak

Sailkatzaile hibridoaren zatiak mota desberdineko sailkatzaileak dira

Adibideak

- **Lazy Bayesian Rules** (Zheng eta Webb, 1998): sailkatu beharreko kasuaren auzokide hurbilenekin Naïve Bayes eredua indusitzen da → **k-NN** + **Naïve Bayes** hibridatzen dira
- **Naïve Bayes Tree** (Kohavi, 1996): hostoetan naive Bayes sailkatzaileak dituzten zuhaitzak indusitzen dira. Sailkatu beharreko kasu berria hosto batera iristen denean, hostoari dagokion entrenamendurako azpidatubasearekin indusituko da Naïve Bayes sailkatzaile lokala → **Sailkapen zuhaitza** + **Naïve Bayes** hibridatzen dira

Oinarrizko sailkatzaile bera konbinatzea

- Entrenamendurako datubasetik **azpidatubase desberdinak** sortu, eta horietatik **sailkatzaile bana** induzitu
- Sailkatzaileen **iragarpenak konbinatu**, multisailkatzailearen azken iragarpena erabakitzeko. Adib. sarritasun handienez iragarritako klasea
- **Bagging** (Bootstrap AGGREGatING), (Breiman, 1994)
 - Entrenamendurako multzotik **bootstrap** aplikatuz lortuko dira entrenamendurako azpimultzoak
 - Sailkatzaileak **modu independentean indusituko dira**; ez dago elkarrekintzarik indukzio prozesuan
 - **Sailkatzaile ezegonkorretarako** (datuetan aldaketa txikiek aldaketa handiak sailkapenean) emaitza onak lortzen dira; zuhaitzak (ezegonkorrak), k -NN (egonkorrak)

Oinarrizko sailkatzaile bera. Bagging

Adibidea: Random forests (Breiman, 2001)

- Oinarrizko sailkatzaile moduan **sailkapen-zuhaitzak** erabiltzen dituen **Bagging** motako multisailkatzailea da
- Sailkapen-zuhaitz bakoitzaren indukzioa, **zoriz** aukeratutako aldagai iragarleen azpimultzo batean oinarrituko da → **random**
- Aldagaien azpimultzo horretan eta bootstrap lagineko kasuetan oinarrituz garatuko dira zuhaitzaren **adarrak**
- Sailkapen-zuhaitz **desberdinak**; basoa → **forest**
- Zuhaitzak bere osotasunean garatuko dira; **inausketarik ez**
- Asmatze-tasa altua eta datubase handiekin eraginkorra
- Zuhaitzak datuetara gehiegi egokitu? → **overfitting?**

Oinarrizko sailkatzaile bera. Boosting

Oinarrizko ideiak

- Multisailkatzailea osatuko duten sailkatzaileak **modu inkrementalean** garatzen dira
 - Sailkatzaile bat induzitu eta gero, bere asmatze-tasaren araberako **pisu bat egokituko zaio sailkatzaileari**
 - Gainera, **datubaseko kasuei pisuak egokituko zaizkie**; ebaluazio garaian gaizki sailkatuak izan diren kasuen pisuak handitu, eta ondo sailkatutakoenak txikitu
- Induzituko den hurrengo sailkatzailea aurrekoak **gaizki sailkatutako kasuekin trebeagoa** izatea da helburua
- Boosting algoritmo asko; datubaseko kasuei eta sailkatzaileei **pisuak egokitzeko moduan** desberdinak
- Inplementazio oso ezagun bat: **AdaBoost**. Sinplea, azkarra, “overfitting” ez

Sailkatzaileen iragarpenak nola konbinatu

Iragarpenen bozketa

Multisailkatzailea osatzen duten sailkatzaileek beren iragarpena egin dutenean, guztien arteko **konbinaketa** egin beharko da, multisailkatzailearen azken iragarpena emateko.

- **Gehiengoaren aldeko bozketa**: sarritasun handienekoa
- **Gehiengo sinplea**: gutxienez botoen erdia gehi bat
- **Gehiengoaren aldeko bozketa atalasea erabiliz**: atalaseak gutxieneko boto-kopurua ezarriko du
- **Gehiengoaren aldeko bozketa pisua erabiliz**: sailkatzaile bakoitzak emandako botoa sailkatzailearen ontasun-maila erabiliz ponderatua izango da azken bozketan

Bibliografia

- “Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid”, Ron Kohavi, KDDM-1996 → **Naïve Bayes Tree hibridoa**, artikulu motza, ulergarria, polita!
- “Bagging Predictors”, Leo Breiman (1996), Machine Learning Volume 24, Number 2, 123-140 → **Bagging**
- **Random Forests Classifier Description**: http://stat-www.berkeley.edu/users/breiman/RandomForests/cc_home.htm
- **“Bagging, Boosting, and C4.5”**, J. R. Quinlan (1996)
- http://en.wikipedia.org/wiki/Bootstrap_aggregating
- http://en.wikipedia.org/wiki/Random_forest
- <http://en.wikipedia.org/wiki/Boosting>