

hyväksymispäivä arvosana

arvostelija

Puoliohjattu oppiminen

Jarmo Ahosola

Helsinki 04.11.2005

HELSINGIN YLIOPISTO

Tietojenkäsittelytieteen laitos

Tekijä		
Jarmo Ahosola		
Työn nimi		
Puoliohjattu oppiminen		
Oppiaine		
Tietojenkäsittelytiede		
Työn laji	Aika	Sivumäärä
Seminaari esitelmä	04.11.2005	12
Tiivistelmä		
<p>Puoliohjatulla oppimisella tarkoitetaan menetelmiä, jotka hyödyntävät luokittelijan oppimisessa luokitellun aineiston lisäksi luokittelematonta aineistoa. Tämä on mahdollista, jos aineiston jakaumien käyttäytyminen tunnetaan riittävän hyvin. Tässä artikkelissa tarkastellaan erilaisia tapoja hyödyntää luokittelematonta aineistoa oppimisen tukena ja esitetään kriteerejä, joiden perusteella voidaan arvioida soveltuuko puoliohjattu oppiminen tietyn ongelman ratkaisemiseen ja jos soveltuu, niin mitkä puoliohjatun oppimisen menetelmistä tarjoavat parhaat edellytykset ongelman ratkaisemiseen.</p>		
Avainsanat		
Puoliohjattu oppiminen, luokittelu, klusterointi		
Säilytyspaikka		
Muita tietoja		

Sisältö

1 Johdanto.....	1
2 Klusterointi geneettisellä algoritmilla.....	1
3 Sanojen merkityksen oppiminen.....	3
4 Odotusarvon maksimointi.....	4
5 Transduktiiviset tukivektorikoneet.....	4
6 Rinnakkaisoppiminen.....	4
7 Puoliohjatun oppimisen soveltaminen.....	5
8 Yhteenveto.....	7

1 Johdanto

Luokittelu on aikaa vaativaa työtä, minkä takia luokiteltu aineisto on vaikeammin saatavissa ja kallimpaa kuin luokittelematon raaka-aineisto. Pelkästään luokitellun aineiston käyttäminen luokittelijan opettamiseen tarkoittaa yleensä vain aineiston murto-osan hyväksikäyttämistä. Puoliohjatulla oppimisella tarkoitetaan menetelmiä, jotka hyödyntävät luokittelematonta aineistoa luokitellun aineiston tukena luokittelijan opettamisessa.

Koska luokittelematonta aineistoa ei ole luokiteltu, se ei suoraan auta luokittelun parantamisessa. Sen sijaan luokittelemattoman aineiston perusteella voidaan saada parempi käsitys aineiston jakaumista, mikä helpottaa eri mallien sovittamista aineistoa vasten. Jos esimerkiksi oletetaan, että aineiston jakauma on yhdistelmä jakauma luokkien jakaumista ja kunkin luokan jakauma on normaalijakautunut, voidaan luokittelemattoman aineiston perusteella oppia yhdistelmä jakauman sisältämien jakaumien parametrit ja luokitellun aineiston perusteella voidaan laskea kullekin löydetylle jakaumalle todennäköisyydet minkä luokan jakaumasta on kyse. Tällöin päästään parempin luokittelutuloksiin kuin jos vastaavat jakaumat yritettäisiin oppia pelkästään luokitellun aineiston perusteella. Tämä pitää paikkansa kuitenkin vain, jos taustalla tehdyt oletukset jakaumien luonteesta pitävät paikkansa. Jos jakaumat eivät vastaa oletusta, niin silloin luokittelemattoman aineiston määrän kasvattaminen lisää lopullisen opitun luokittelijan luokitteluvirhettä.[CCC03]

Graafeihin perustuvassa menetelmässä koko aineistosta rakennetaan puu, jonka kaarten pituuksien summa minimoidaan. Tämä edellyttää eri parametridimensioiden yhteismitallistamista eli algoritmia, joka voi määrittää naapuruston keskenäisiä etäisyyksiä. Teoria perustuu oletukseen, että muodostuvassa puussa samaan luokkaan kuuluvat tapaukset ovat lähellä toisiaan. Uusi tapaus, jota puussa ei ole voidaan lisätä puuhun lyhimmillä mahdollisella kaarella ja luokitella sen solmun perusteella mihin kaari yhdistetään.[ZLG03]

Hiukan vastaavaan ideaan perustuvat itseoppivat mallit. Itseoppivissa malleissa oppimiseen käytetään luokiteltua aineistoa. Kun malli on tältä pohjalta opittu, sillä luokitellaan koko luokittelematon aineisto ja pieni osa varminnan luokituksen saaneista luokittelemattomista tapauksista uskotaan oikein luokitelluiksi ja liitetään mukaan luokiteltuun aineistoon mallin antamalla luokalla. Näin syntynyttä uutta suurempaa luokiteltua aineistoa käytetään oppimaan uusi malli ja sama toistetaan kunnes koko luokittelematon aineisto on luokiteltu luotettavasti tai luokittelun virherajat heikkenevät alle raja-arvon. Tässä lähestymistavassa virheellisen luokittelun vaikutus kumuloituu iteraatio iteraatiolta aina seuraaviin mallien sukupolviin.[Xia05]

2 Klusterointi geneettisellä algoritmilla

Geneettisellä algoritmilla klusteroinnissa lähtökohtana muodostaa sekä luokitelluista, että luokittelemattomista aineistosta klustereita. Klusterin määrittää euklidisessa avaruudessa keskipiste. Kaikki avaruuden pisteet luokitellaan kuuluvan siihen klusteriin, jonka määrittävä piste on kyseistä avaruuden pistettä lähinnä. Klusterien muodostamisessa pyritään minimoimaan aineiston hajontaa

ja maksimoimaan klustereiden puhtautta luokitellun aineiston suhteen.[DBE99]

Hajonnan mittana voidaan käyttää neliövirheen minimointia (MSE), mutta Davies-Bouldin indeksi [JaD88] (DBI) antaa tähän tarkoitukseen parempia tuloksia, sillä se suosii tiiviitä ja hyvin erillään olevia klustereita.[DBE99]

Kun N pistettä jaetaan K klusteriin määritetään ensin kahteen klusteriin jakamisen mitta seuraavasti:

$$R_{j,k} = \frac{e_j + e_k}{D_{jk}},$$

missä $1 \leq j, k \leq K$; $j \neq k$; e_j ja e_k ovat klusterien C_j ja C_k keskihajontoja ja D_{jk} on euklidinen etäisyys klusterien C_j ja C_k välillä.

Jos m_j ja m_k ovat klustereiden C_j ja C_k keskipisteitä ja klusterissa C_j on N_j pistettä, niin silloin

$$e_j = \frac{1}{N_j} \sum_{x \in C_j} \|x - m_j\|^2 \quad \text{ja} \quad D_{jk} = \|m_j - m_k\|^2.$$

Määritetään termi R_k kullekin klusterille C_k

$$R_k = \max_{j \neq k} R_{j,k},$$

jolloin Davies-Bouldin indeksi saadaan määriteltyä muodossa

$$DB(K) = \frac{1}{K} \sum_{k=1}^K R_k$$

Klusterien puhtauden mittana voidaan käyttää päätöspuiden jakamiseen kehitettyä Gini-indeksiä [BFO84]. Päätöspuissa Gini-indeksillä arvioidaan väheneekö epäpuhtaus puun jakamisen perusteella ja tämän tiedon avulla puu pilkotaan siten, että syntyvien puiden puhtaus saadaan maksimoitua. Klusteroinnissa Gini-indeksi klusterille saadaan kaavasta

$$GiniP_j = 1.0 - \sum_{i=1}^k \left(\frac{P_{ji}}{N_j} \right)^2 \quad j \in 1, \dots, K,$$

missä P_{ji} on klusterin j luokkaan i kuuluvien pisteiden lukumäärä. N_j on klusterissa j olevien pisteiden kokonaismäärä. Tällöin klusteroinnin K klusteriin epäpuhtaus on

$$impurity = \frac{\sum_{j=1}^K T_{P_j} * GiniP_j}{N},$$

missä N on aineiston havaintopisteiden lukumäärä.

UC-Irvine Machine Learning Repository[MuA92] aineistoilla tehtyjen testien perusteella Gini-indeksi soveltuu tähän tarkoitukseen paremmin kuin muut epäpuhtauden mittarit kuten esimerkiksi virheellisesti luokiteltujen tapausten lukumäärä. Gini-indeksi suosii puhtaita klustereita luokitteluvirheen kustannuksella.[DBE99]

Geneettisten algoritmien soveltamisessa edellä esitettyjä tavoitefunktioita (object function) voidaan suoraan käyttää jalostetun yksilön elinkykyisyyden mittarina (fitness function). Demiriz et al käyttivät sellaisenaan GALib:in[Wal96] geneettisiä algoritmeja soveltaen elitististä strategiaa, joka sallii parhaiden yksilöiden säilyä sellaisenaan seuraavaan sukupolveen. Perimäksi Demiriz et al valitsivat Kd reaalilukua, missä K on klusterien lukumäärä ja d dimensioiden lukumäärä. Valitut luvut kuvastivat suoraan valittavien klustereiden keskipisteiden koordinaatteja. Muodostuneet klusterit saivat luokan sen mukaan mitä luokiteltuja tapauksia klusterissa oli eniten.[DBE99]

Induktiivisissa tapauksissa Davies-Bouldin indeksin ja Gini-indeksin käyttö yhdessä ei tuottanut merkittävästi erilaisia tuloksia kuin pelkän Gini-indeksin käyttö. Sen sijaan transduktiivisissa tapauksissa Davies-Bouldin indeksin ja Gini-indeksin käyttö yhdessä tuotti selkeästi parhaat tulokset verratuista menetelmistä.[DBE99]

Demirez et al havainto, jonka mukaan algoritmi, joka suosii muita parametreja luokitteluvirheen kustannuksella on parempi kuin algoritmi, joka minimoi luokitteluvirhettä selittyy sillä, että heidän taustalla ollut malli jakaumien käyttäytymisestä oli testattujen aineistojen kohdalla oikea. Näissä tapauksissa luokitteluvirheen optimointi paremmaksi olisi aiheuttanut ylioppimista luokitellun aineiston perusteella. Sen sijaan valitut parametrit hyödynsivät luokittelemattoman aineiston massan tuomaa lisäinformaatiota ja siten estivät ylioppimista pelkän luokitellun aineiston sisältämän informaation suuntaan.

3 Sanojen merkityksen oppiminen

David Yarowskyn esittämä algoritmi[Yar95] monimerkityksellisten sanojen merkityksen asiayhteydestä tunnistamisen oppimiseksi on esimerkki itseoppivasta lähestymistavasta puoliohjatussa oppimisessa. Algoritmin taustaoletuksena on, että samassa asiayhteydessä käytetään vain yhtä merkitystä monimerkityksellisistä sanoista.[GCY92] Tämä oletus mahdollistaa puoliohjatun oppimisen soveltamisen ongelman ratkaisemiseen, sillä silloin asiayhteyksiä voidaan käyttää tarkasteltavina kokonaisuuksina, joista haetaan merkitystä selittäviä piirteitä.

Yarowsky valitsi piirteiksi monimerkityksellistä sanaa edeltävän sanan, monimerkityksellistä sanaa seuraavan sanan, sekä monimerkityksellisestä sanasta mitattuna alle k sanan etäisyydellä olevan sanan. Luokiteltuna aineistona käytettiin sanakirjan määritelmiä monimerkityksellisten sanojen eri merkityksille. Luokittelemattomana aineistona käytettiin hyvin suurta aineistoa eri lähteistä kerättyä englanninkielistä tekstiä. Aineisto jaettiin kunkin n -monimerkityksellisen sanan suhteen n luokkaan ja luokittelemattomaan aineistoon. Kunkin luokan sisältämät piirteet arvoitettiin sen mukaan kuinka vahvasti piirteet pystyivät tukemaan hypoteesia, jonka mukaan näyte kuului kyseiseen luokkaan. Näiden piirteiden arvoitusten perusteella kaikki näytteet luokiteltiin ja ne luokittelemattomat

näytteet, joiden vahvin luokittelua tukeva piirre oli yli raja-arvoa O voimakkaampi kuin muuta luokittelua tukeva vahvin piirre liitettiin osaksi luokkaa. Vastaavasti ne luokitellut näytteet, joiden vahvin nykyistä luokkaa tukeva piirre oli alle raja-arvoa U voimakkaampi kuin muuta luokittelua tukeva vahvin piirre siirrettiin osaksi luokittelematonta aineistoa. Tämän jälkeen piirteet arvotettiin uudestaan ja samat vaiheet toistettiin, kunnes koko aineiston luokittelu konvergoitui stabiiliin tilaan. Vahvinta piirrettä käytettiin mittarina piirteiden yhteisen vahvuuden sijasta, koska sen ansiosta piirteiden keskenäisiä korrelatioita ei tarvinnut huomioida. Raja-arvon U käyttäminen salli oppimisen alkuvaiheessa tapahtuneiden virheiden korjaantumisen myöhemmissä iteraatioissa, sekä itsekorjaantumisen alunperin väärin luokitelluissa opetusnäytteissä.[Yar95]

Monimerkityksellisten sanojen merkityksen oppiminen asiayhteyden kautta on hyvä esimerkki siitä miten luokittelematonta aineistoa voidaan käyttää hyväksi näytteiden välisten naapurustoverkkojen rakentamisessa, jolloin luokittelematon aineisto kuvaa topologian, jonka avulla luokitellun aineiston luokittelua voidaan laajentaa kattamaan tapauksia, joita pelkästään luokitelluista tapauksista opittavissa olevan tiedon perusteella ei voi lainkaan luokitella.

4 Odotusarvon maksimointi

Odotusarvon maksimointi[DLR77] (EM) on vanhin puoliöhjatun oppimisen menetelmistä. Menetelmän oletuksena on, että kukin opittavista luokista muodostaa tunnistettavan jakauman parametriavaruudessa. Käyttämällä koko opetusaineistoa menetelmä pyrkii tunnistamaan minkälaisia päällekkäisiä jakaumia parametriavaruudessa on ja käyttämällä luokiteltua aineistoa menetelmä valitsee kullekin jakaumalle luokan odotusarvon maksimoinnin perusteella.[Xia05]

5 Transduktiiviset tukivektorikoneet

Ohjatussa oppimisessa etsitään parametriavaruudesta taso, joka parhaiten leikkaa aineiston oikeisiin luokkiin. Tällaisia yhtä hyviä tasoja on ääretön määrä. Menetelmässä luokittelematon aineisto lisätään parametriavaruuteen, jonka jälkeen ohjatun oppimisen kannalta yhtä hyvistä parhaista tasoista valitaan se, joka on mahdollisimman kaukana lähimmästä parametriavaruuden havaintopisteestä. Menetelmän ajatus on, että eri luokkien välillä on suurempi ero kuin samaan luokkaan kuuluvien havaintoyksilöiden välillä. Tällä perusteella havaintoavaruuden leikkaava taso, joka on mahdollisimman etäällä lähimmästä parametriavaruuden havaintopisteestä kulkee todennäköisemmin luokkien välistä kuin läheltä havaintopisteitä kulkevat tasot. Valittavan tason löytämiseksi tarvittavien tukivektorikoneiden parametrien selvittäminen on NP-kova ongelma, mutta useita lupaavia aproksimointimenetelmiä tämän ongelman ratkaisemiseksi on kehitetty. [Xia05]

6 Rinnakkaisoppiminen

Rinnakkaisoppiminen (co-training) on menetelmä, jossa koulutetaan saman aineiston perusteella kaksi eri luokittelijaa, joista kummankin opetuksessa käytetään eri osaa aineiston kuvaavista

tiedoista. Menetelmän oletuksena on, että kumpikin kuvaava osa on yksinään riittävä aineistosta oppimiseksi ja kuvaavat osat ovat toisistaan riippumattomia annettuna luokka. Luokitellun aineiston perusteella opetetaan ensin kumpikin luokittelijoista. Tämän jälkeen kummankin luokittelijan vahvimmin luokittelemia tapauksia liitetään toisen luokittelijan opetusaineistoon ja luokittelijat opetetaan uudestaan.[BMi98]

Ero itseoppivaan lähestymistapaan on siinä, että itseoppivassa lähestymistavassa sama luokittelija laajentaa uskomustaan kullakin iteraatiolla vahvistuvien käsitysten perusteella, mutta rinnakkaisoppimisessa luokittelun perusteena oleva tieto on luokittelijoilla erilainen, minkä ansiosta ensimmäiselle luokittelijalle helppo luokittelutehtävä voi olla toiselle luokittelijalle vaikea ja siten luokittelutieto minkä ensimmäinen luokittelija luokittelemattomalle tapaukselle antaa voi olla toiselle luokittelijalle hyvin arvokas samalla kun luokittelun oikeellisuuden todennäköisyys pysyy korkeana. Itseoppivassa lähestymistavassa uuden luokittelun tapauksen informaatioarvo luokittelun kannalta on verrannollinen luokittelun laajentamisessa otettuun riskiin.

Rinnakkaisoppiminen soveltuu tilanteisiin, joissa luokittelun tueksi kuvaavaa tietoa saadaan erillisistä lähteistä. Esimerkiksi tunnistus videomateriaalista voi perustua erikseen pelkkään ääniraitaan tai pelkkään kuvamateriaalin analysointiin. Tällöin luokiteltuna opetusmateriaalina luokittelijoille voidaan käyttää tapauksia, jotka ovat selkeitä luokaltaan ja luokittelemattomassa aineistossa olevat tapaukset, joissa on kuvassa esteitä tai äänessä häiriöitä auttavat näiden piirteiden oppimista vaikkein luokittelevälle luokittelijalle helpommin luokittelevan luokittelijan avulla. [BMi98]

Sen jälkeen kun kaksi erillistä luokittelijaan on saatu koulutettua, muodostetaan näistä kolmas luokittelija, joka luokittelee tapaukset koulutettujen luokittelijoiden luokitteluvarmuuksien tulon perusteella. Tätä luokittelijaa käytetään lopullisena luokittelijana.[BMi98]

Blum ja Mitchell tutkivat rinnakkaisoppimisen soveltamista yliopistojen nettisivujen luokittelemiseksi kurssisivuihin ja muihin sivuihin käyttämällä erillisinä kuvaavina piirteinä nettisivujen sisältöä ja nettisivuihin viittaavien hyperlinkkien sisältöä. Näiden aineistojen riippumattomuuden perusteena oli se, että viittaukset olivat eri tekijän sivulla kuin sivun sisältö. Heidän tulosten perusteella rinnakkaisoppimisella saatiin luokittelu virhe lopullisella luokittelijalla alle puoleen virheestä, joka saatiin luokittelijalla, joka koulutettiin käyttämällä koko kuvaavien piirteiden joukkoa.[BMi98]

7 Puoliöhjatun oppimisen soveltaminen

Luokittelemattoman tiedon hyödyntäminen luokitellun tiedon lisäksi oppimisen tukena edellyttää jotakin seuraavista asioista:

- 1) Malli opittavan aineiston sisäisistä suhteista pitää paikkansa

- 2) Virheellisellä mallilla optimaalinen tulos saavutetaan arvoilla, jotka ovat luokitellusta aineistosta opittavien arvojen luokittelemattomasta aineistosta opittavien arvojen puolella
- 3) Aineisto sisältää hyvin paljon parametreja suhteessa luokitellun aineiston kokoon, minkä seurauksena hajonnan pienentämisen merkitys on suurempi kuin biasin huononemisen merkitys luokittelemattomasta aineistosta oppimisen seurauksena

Riippumatta mallin oikeellisuudesta luokitteluennusteiden hajonta pienenee kun luokitellun aineiston lisäksi oppimisen tukena käytetään luokittelematonta aineistoa. Tämän takia puoliöhjattu oppiminen soveltuu erityisen hyvin tekstin luokitteluun ja kuvan tunnistukseen, vaikka taustalla oleva malli ei olisikaan täysin oikea. Mallin ollessa väärä bias kasvaa kun luokittelemattoman aineiston osuus suhteessa luokiteltuun aineistoon kasvaa. Jos luokittelematonta aineistoa ja luokiteltua aineistoa lisätään opetusaineistoon säilyttäen niiden keskenäinen suhde samana, päästään lähemmäksi mallin mahdollistamaa optimia hajonnan pienemisen ansiosta virheellisestä mallista huolimatta. Jos taustalla oleva malli on oikea, ei luokittelemattoman aineiston määrän lisääminen suhteessa luokiteltuun aineistoon opetusaineistossa voi huonontaa opittavan luokittelijan laatua, sillä edellytys tulosten huononemiselle on virheellinen malli. Tyypillinen virhe mallissa on mallin parametrien keskenäisen riippumattomuuden oletus. Tällainen mallin virhe on riittävä, jotta luokittelemattoman aineiston määrän kasvattaminen suhteessa luokitellun aineiston määrään opetusaineistossa voi heikentää opittavan luokittelijan laatua.[CCC03]

Edellä kohta 2 on mainittu täydellisyyden takia, sillä se voi selittää miksi tulokset paranevat luokittelemattoman aineiston lisäämisen vaikutuksesta vaikka malli olisi väärä ja hajonnan pienemisen merkitys biasin huononemisen merkitykseen verrattuna pieni. Mallin oikeellisuutta ei siis voi päätellä sen perusteella, että rajallinen määrä luokittelemattomia tapauksia opetusaineistoon lisättynä parantaa luokittelijan laatua.

Ongelman luonteesta riippuen eri puoliöhjatun oppimisen menetelmien soveltuvuus ongelman ratkaisemiseen vaihtelee. Useimmat menetelmät tuottavat huonoja tuloksia silloin, jos luokkien välinen optimaalinen raja kulkee tiheän näytteiden keskittymän kautta. Tällainen tilanne syntyy esimerkiksi silloin, jos kahden tunnistettavan luokan tapaukset ovat normaalijakautuneita, mutta kyseiset jakaumat menevät selvästi päällekkäin. Tällaisen ongelman ratkaisemiseen odotusarvon maksimointi sekoitemallien perusteella (EM with generative mixture models) soveltuu hyvin. Ensisijaisesti mallin valinnassa tulee käyttää menetelmää, jonka oletukset vastaavat ongelman rakennetta. Jos luokkien tuottama data muodostaa selkeitä klustereita odotusarvon maksimointi sekoitemallien perusteella on lupaava vaihtoehto. Jos parametrit luontaisesti jakautuvat kahdeksi erilliseksi ryhmäksi, rinnakkaisopetus (co-training) voi soveltua ongelman ratkaisemiseen. Jos havainnot joilla on samoja piirteitä kuuluvat tyypillisesti samaan luokkaan, niin graafeihin perustuvat ratkaisut ovat perusteltuja. Jos ongelman ratkaisussa jo käytetään tukivektorikoneita, niin puoliöhjatun oppimisen laajennus transduktiivisiin tukivektorikoneisiin (TSVM) on luontevaa. Jos ennestään on olemassa luokitellun aineiston perusteella opetettava monimutkainen luokittelija, niin itseoppivan lähestymistavan rakentaminen sen päälle on vaihtoehto, jota on syytä harkita.[Xia05]

8 Yhteenveto

Puoliohjatulla oppimisella saadaan parannettua tuloksia suhteessa ohjattuun oppimiseen, jos taustalla oleva oletus mallista pitää riittävän hyvin paikkansa. Vaikka taustalla oleva malli olisi väärä, luokittelemattomien näytteiden lisääminen säilyttäen luokiteltujen ja luokittelemattomien tapauksien suhteen vakiona parantaa luokittelun tuloksia. Tämä johtuu siitä, että näytteiden suhteen pysyessä samana, virheellisestä mallista aiheutuva bias ei kasva, mutta näytteiden kokonaismäärän kasvaminen pienentää hajontaa, mikä parantaa luokittelutulosta.

Puoliohjattuja menetelmiä on useita erilaisia, joista oikean valitseminen riippuu tilanteesta. Jos luokat ovat erillään toisistaan, transduktiiviset tukivektorikoneet voivat löytää todennäköisen raon luokkien välistä. Jos jakaumat ovat vahvasti päällekkäisiä, sekoitemallit voivat soveltua ongelman ratkaisuun hyvin. Graafeihin perustuvat menetelmät soveltuvat tilanteisiin, joissa samaan luokkaan kuuluminen on todennäköistä naapurustossa lähellä olevilla tapauksilla. Itseoppivat menetelmät soveltuvat tilanteisiin, joissa on olemassaoleva ohjatun oppimisen ratkaisu ja rinnakkaisopetus aineistoille, joissa kuvaavat piirteet on luokiteltavissa kahdeksi erilliseksi piirrejoukoksi, jotka kumpikin itsessään riittäisivät aineistosta oppimiseen.

Puoliohjattu oppiminen ei ole ratkaisu kaikkeen, mutta erityisesti sovellusalueille, joissa on paljon parametreja suhteessa luokitellun aineiston kokoon menetelmät soveltuvat hyvin. Tällaisia alueita ovat esimerkiksi tekstin analysointi ja kuvan tunnistus.

Lähteet

- BFO84: L. Breiman, J. Friedman, R. Olshen, ja C. Stone, Classification and Regression Trees. Wadsworth International, California, 1984
- BMi98: Blum, A., & Mitchell, T., Combining labeled and unlabeled data with co-training: COLT: Proceedings of the Workshop on Computation Learning Theory, 1998
- CCC03: Cozman, F., Cohen, I., & Cirelo, M., Semi-supervised learning of mixture models. ICML-03, 20th International Conference on Machine Learningn., 2003
- DBE99: Demirez, A., Bennett, K., & Embrechts, M., Semi-supervised clustering using genetic algorithms. Proceedings of Artificial Neural Networks in Engineering, 1999
- DLR77: Dempster, A., Laird, N. & Rubin, D., Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B., 1977
- GCY92: Gale, W., K. Church, D. Yarowsky, A method for Disambiguating Word Senses in a Large Corpus. Computers and the Humanities, 26, pp 415-439, 1992
- JaD88: A.K. Jain ja R.C. Dubes, Algorithms for Clustering Data, 1988
- MuA92: P.M. Murphy ja D.W. Aha., UCI repository of machine learning databases., 1992
- Wal96: M. Wall, GAlib: A C++ Library of Genetic Algorithm Components. MIT, <http://lancet.mit.edu/ga/>, 1996
- Xia05: Xiaojin Zhu, Semi-Supervised Learning Literature Survey, 2005
- Yar95: Yarowsky, D., Unsupervised word sense disambiguation rivaling supervised methods. Proceedings of the 33rd Annual Meeting of the Association for Computer Linguistics (pp. 189-196), 1995
- ZLG03: Zhu, X., Lafferty, J., & Ghahramani, Z., Semi-supervised learning: From Gaussian fields to Gaussian processes (Technical Report CMU-CS-03-175)., 2003