Läheisyysfunktioista sekamuotoisen datan luokittelussa

Janne Lumijärvi

Tampereen yliopisto
Tietojenkäsittelytieteiden laitos
Tietojenkäsittelyoppi
Pro gradu –tutkielma
Elokuu 2003

Tampereen yliopisto
Tietojenkäsittelytieteiden laitos
Janne Lumijärvi: Läheisyysfunktioista sekamuotoisen datan luokittelussa
Pro gradu -tutkielma, 75 sivua, 8 liitesivua
Elokuu 2003

Tutkielmassa tarkastellaan kirjallisuudessa esitettyjä läheisyysfunktioita ja niiden soveltamista luokittelussa. Erityisesti tarkoituksena oli pohtia läheisyysfunktioiden soveltuvuutta heterogeenisen datan käsittelyyn. Heterogeenisellä datalla tarkoitetaan dataa, joka on kuvattu mitta-asteikoiltaan nominaalisilla sekä ordinaalisilla ja kvantitatiivisilla attribuuteilla. Niin sanotuissa homogeenisissä läheisyysfunktioissa on usein ongelmana nominaalisten attribuuttien sopimaton käsittely. Heterogeenisissä funktioissa on pyritty siihen, että kukin attribuutti käsitellään erikseen sille sopivalla tavalla. Tutkielmassa tarkastellaan funktioiden ominaisuuksia ja esitetään euklidisen etäisyys- ja kosinietäisyysfunktion sekä kuuden heterogeenisen läheisyysfunktion välillä suoritetun keskinäisen vertailun tulokset. Tehtävänä oli luokitella esimerkkejä lähimmän naapurin menetelmällä, joka on yksinkertainen havaintoperustaiseen oppimiseen perustuva menetelmä. Vertailussa käytettiin 36 aineistoa. Tulosten perusteella perinteisten heterogeenisten funktioiden ja homogeenisen euklidisen etäisyysfunktion välillä ei ole merkittäviä eroja sekamuotoisen datan luokittelussa. Keskimäärin parhaita tuloksia saatiin HVDM-funktiolla (Heterogeneous Value Difference Metric), jossa nominaalisen datan käsittely poikkeaa selvästi perinteisistä heterogeenisistä funktioista. HVDM-funktioon esitetään tarkennus luokittelutuloksen parantamiseksi. Tutkielmassa tarkastellaan myös hieman ennen ja jälkeen varsinaisen luokittelun tehtäviä toimenpiteitä, kuten datan esikäsittelyä sekä tulosten arviointia.

Avainsanat ja -sanonnat: läheisyysfunktiot, etäisyysfunktiot, samankaltaisuusfunktiot, luokittelu, havaintoperustainen oppiminen, mitta-asteikot.

Sisällys

Kiitokset

Kiitokset ohjaajilleni ja erityinen kiitos FT Erkki Pesoselle (Kuopion yliopisto) tutkimusmateriaalin luovuttamisesta.

1. Johdanto

Olioiden *luokitteleminen* sekä ryhmittely eli *klusterointi* ovat tavallisia ongelmia hyvin monilla aloilla. Klusteroinnissa on tarkoitus löytää datasta keskenään homogeenisten olioiden ryhmiä eli klustereita ja sillä on sovelluksia muun muassa astronomiassa, psykiatriassa ja arkeologiassa [Everitt et al., 2001]. Klusterointi on *ohjaamatonta luokittelua*, jossa luokkarakennetta ei tiedetä etukäteen. *Ohjatussa luokittelussa* kyse on uusien tapausten sijoittamisesta jo valmiiksi tunnettuihin luokkiin. Esimerkiksi lääketieteellisissä asiantuntijajärjestelmissä [Auramo et al., 1993; Laurikkala, 2001a] pyritään potilaan tietojen perusteella tekemään diagnoosi aiempien diagnoosien perusteella. Luokittelu ja ryhmittely ovat esimerkkejä *tiedonlouhinnasta*. Tiedonlouhinnassa tarkoituksena on etsiä datasta säännöllisyyksiä sekä pyrkiä datan avulla hankkimaan tietämystä ja parantamaan päätöksentekoa tulevaisuudessa [Hand et al., 2001; Mitchell, 1999; Vesanto and Hollmén, 2002]. Hieman samasta asiasta on kysymys *koneoppimisessa* [Mitchell, 1997], jossa ideana on, että algoritmit parantavat suoritustaan ja kehittyvät automaattisesti kokemuksen avulla.

Erilaiset luokittelutehtävät perustuvat usein olioiden väliseen *erilaisuuteen* tai *samankaltaisuuteen* tai yleisemmin *läheisyyteen*. Tarkastellaan esimerkkinä niin sanottua *havaintoperustaista oppimista* (instance-based learning) [Mitchell, 1997], joka tunnetaan myös nimellä *muistipohjainen päättely* (memory-based reasoning) [Stanfill and Waltz, 1986]. Siinä on tarkoituksena luokitella olioita suoraan muistissa olevien olioiden perusteella. Oliota kutsutaan tavallisesti *esimerkiksi*. Uutta esimerkkiä luokiteltaessa muistista etsitään samankaltaisia esimerkkejä kuin luokiteltava esimerkki. Tämän jälkeen luokiteltava esimerkki sijoitetaan siihen luokkaan, johon sen kanssa samankaltaiset esimerkit kuuluvat. Luokittelu perustuu siis uuden esimerkin ja muistissa olevien esimerkkien välisiin läheisyyksiin, jotka määritellään *etäisyysfunktion* avulla. Siis sitä samankaltaisempia esimerkit ovat, mitä pienempi etäisyys niiden välillä on. Yhtä lailla voitaisiin käyttää *samankaltaisuusfunktiota*.

Sen määrittäminen, kuinka etäällä (tai lähellä) kaksi reaalimaailman oliota ovat toisistaan, ei tietenkään ole yksiselitteistä. Kahden kokonaisluvun välisen etäisyyden määrittelemiseen on olemassa hyvin vakiintunut tapa, mutta sen laskeminen, kuinka paljon esimerkiksi kaksi potilasdiagnoosia eroavat toisistaan, on monestakin syystä vaikeaa. Kun määritellään objektiivisesti reaalimaailman olioiden välisiä läheisyyksiä, tulee aluksi tarkastella tapaa, jolla oliot on kuvattu. Tyypillisesti olio aluksi ositetaan ominaisuuksiin eli *attribuutteihin* ja eri mittaustekniikoiden avulla mitataan kullekin ominaisuudelle jonkinlainen symbolinen esitys, joka on yleensä binääri-, kokonais- tai reaaliluku. Näin olio voidaan esittää vektorin avulla, jonka alkioina ovat mittaustulokset olion ominaisuuksista. Tällainen esitystapa on välttämätön etäisyys- ja samankaltaisuusfunktioiden kannalta. Tavallisesti funktio määrittää syötteenä saatujen kahden oliovek-

torin välisen läheisyyden reaalilukuna. Luonnollisesti tavoitteena on, että lasketut arvot kertoisivat mahdollisimman hyvin olioiden todellisista suhteista. Reaalimaailman data voi olla luonteeltaan hyvinkin vaihtelevaa esimerkiksi attribuuttien mitta-asteikoiltaan ja arvoalueiden laajuudeltaan, mikä asettaa haasteita etäisyyden laskemiselle.

Tässä tutkimuksessa on tarkoituksena perehtyä funktioihin, jotka määräävät kahden olion välisen etäisyyden ja samankaltaisuuden. Ideana on ennen kaikkea pohtia funktioiden soveltuvuutta sekamuotoisen eli heterogeenisen datan läheisyyksien määrittelemisessä. Heterogeenisellä datalla tarkoitetaan tässä yhteydessä dataa, joka on kuvattu monitasoisilla nominaalisilla sekä ordinaalisilla ja kvantitatiivisilla attribuuteilla. Tutkielmassa tarkastellaan funktioiden ominaisuuksia ja suoritetaan tilastollinen vertailu funktioiden välillä. Tehtävänä on luokitella joukko erityyppisiä aineistoja. Luokittelussa käytetään läheisyyteen perustuvaa menetelmää, jonka osana käytetään eri läheisyysfunktioita. Tämän jälkeen vertaillaan eri funktioilla saatuja luokittelutuloksia.

Tutkielman rakenne on seuraava. Luvussa 2 tarkastellaan joitain reaalimaailman datan tyypillisiä ominaisuuksia ja ratkaisuja datan käsittelyssä ilmeneviin ongelmiin. Erityisesti tarkastellaan attribuuttien mitta-asteikoita, joilla on olennainen merkitys läheisyysfunktioiden kannalta. Luvussa 2 esitetään lisäksi datan merkintä- ja esitystavat, joita käytetään tässä työssä. Luvussa 3 käsitellään luokittelua ja havaintoperustaista oppimista. Luvussa 4 tarkastellaan läheisyyden, etäisyyden ja samankaltaisuuden käsitteitä. Luvussa 5 perehdytään joihinkin homogeenisiin ja luvussa 6 joihinkin heterogeenisiin etäisyys- ja samankaltaisuusfunktioihin. Luvussa 7 tarkastellaan tunnuslukuja, joiden avulla luokittelutuloksia voidaan arvioida. Luvusta 8 alkaa tutkielman empiirinen osa, jossa luokitellaan erityyppisiä aineistoja *k-lähimmän naapurin menetelmällä,* joka on yksinkertainen havaintoperustaiseen oppimiseen perustuva menetelmä. Lähimmän naapurin menetelmän osana käytetään vuorotellen kahdeksaa läheisyysfunktiota, joista kuusi on heterogeenisiä. Vertailun vuoksi mukana on lisäksi kaksi homogeenistä funktiota. Testeissä käytetään 36 valmiiksi luokiteltua aineistoa, joista suurin osa on reaalimaailmaa kuvaavia aineistoja ja loput keinotekoisia koneoppimismenetelmien testaamisessa usein käytettyjä aineistoja. Kukin aineisto luokitellaan uudelleen käyttäen ristiinvalidointia. Funktioiden luokittelutuloksia vertaillaan tilastollisten testien avulla. Tuloksia analysoidaan tarkemmin luvussa 9 ja luvussa 10 tehdään yhteenveto.

2. Tiedonlouhintaa edeltävistä toimista

Ennen kuin reaalimaailman dataa voidaan käyttää luokittelutehtävissä tulee aineisto käytännössä aina esiprosessoida ja tarvittaessa tehdä sille joitain muunnoksia. Muunnoksista tyypillinen on *imputointi* eli puuttuvien tietojen korvaaminen, joka joudutaan tekemään, mikäli käytetyt algoritmit vaativat, että aineisto on täydellistä. Aineistosta saattaa olla myös järkevää karsia pois poikkeavat havainnot eli *kohina*, jotta luokittelualgoritmin toiminta ei häiriintyisi niiden takia. Lisäksi esimerkiksi attribuuttien arvojen normalisointi parantaa yleensä etäisyys- ja samankaltaisuusfunktioiden toimintaa ja näin ollen myös luokittelutulosta.

Aivan ensimmäisenä täytyy luonnollisesti kerätä datakokoelma. Tämän jälkeen

- valitaan käsiteltävä datajoukko,
- suoritetaan tälle joukolle esiprosessointi, ja
- tehdään tarvittavat muunnokset.

Nämä vaiheet ovat oleellisen tärkeitä, jotta varsinainen luokittelu saavuttaisi sille asetetut tavoitteet.

Kun ollaan tekemisissä reaalimaailmaa kuvaavan datan kanssa, ongelmia luonnollisesti esiintyy runsaasti [Hand et al., 2001; Kubat et al., 1998; Laurikkala et al., 2001]. *Tiedonlouhinnassa*, kuten luokittelussa, on usein kysymys niin sanotun sekundäärisen datan uudelleen käytöstä. Tällainen data on alun perin kerätty toiseen tarkoitukseen, ja näin ollen se ei ole valmista käsiteltäväksi. Datan analyysi ja mahdollinen esikäsittely on tehtävä huolellisesti.

Tarkennetaan tässä vaiheessa hieman ”datan” määrittelyä. Myöhemmin käsiteltävät läheisyysfunktiot nimittäin soveltuvat ainoastaan attributatiiviselle datalle. Olio tulee siis voida esittää vektorin avulla, jossa vektorin alkiot ovat mittaustuloksia olion valituille ominaisuuksille. Tällaisesta oliota kuvaavasta vektorista on kirjallisuudessa useita eri nimityksiä, joista tavallisimpia ovat *havainto* tai *esimerkki*, jota termiä käytettäköön vastaisuudessa. Aineisto eli joukko esimerkkejä voidaan näin ollen esittää matriisin avulla, jossa rivit kuvaavat olioita ja sarakkeet olioiden ominaisuuksia. Tällaista matriisia kutsutaan havainto- tai esimerkkimatriisiksi. Usein tiedonlouhinnan lähtökohtana ovat relaatiotietokannat tai taulukot, joten matriisiesitystapa on varsin luonnollinen tämän muotoiselle datalle. Tavallisesti havaintomatriisin elementit ovat reaali- tai kokonaislukuja, mutta ne voivat toki olla myös esimerkiksi lukujoukkoja tai merkkijonoja. Tässä työssä kuitenkin oletetaan, että kaikki elementit ovat reaali- tai kokonaislukuja. Esiteltävistä etäisyysfunktioistakaan ei löydy yhtä lukuunottamatta mekanismeja käsittelemään muun tyyppistä dataa. Niin sanottuun karteesiseen avaruusmalliin perustuva CSM-funktio [Ichino and Yaguchi, 1994] (katso kohta 6.4) pystyy käsittelemään myös lukujoukkoja ja -intervalleja.

Tässä luvussa määritellään datan esitys- ja merkintätavat, joita käytetään jatkossa. Datan ominaisuuksista tarkastellaan ennen kaikkea attribuuttien mitta-asteikkoja, joihin palataan lukuisia kertoja myöhemmin etäisyys- ja samankaltaisuusfunktioiden yhteydessä. Lisäksi tarkastellaan joitain datan muunnosmenetelmiä, joita tarvitaan myöhemmin käytännön dataa käsiteltäessä (katso luku 8).

2.1. Havaintomatriisi

Yleensä tiedonlouhinta-algoritmien lähtökohta on, että tutkittava aineisto esitetään *havaintomatriisina*. Havaintomatriisi on tietyn sopimuksen mukainen havaintoaineiston esitys [Puntanen, 1998], joka esitetään tavallisesti kuvan 2.1 mukaisesti tai lyhyemmin $E = (e_{ij})$.

$$E = \begin{array}{c} \\ r_1 \\ r_2 \\ \dots \\ r_n \end{array} \begin{array}{c} \begin{array}{cccc} v_1 & v_2 & \dots & v_m \end{array} \\ \begin{pmatrix} e_{11} & e_{12} & \dots & e_{1m} \\ e_{21} & e_{22} & \dots & e_{2m} \\ \dots & \dots & \dots & \dots \\ e_{n1} & e_{n2} & \dots & e_{nm} \end{pmatrix} \end{array}$$

Kuva 2.1. Havaintomatriisi.

Kuvassa 2.1 havaintomatriisin E sarakkeet (pystyvektorit) v_1, $v_{2, \dots,} v_m$ kuvaavat olioiden ominaisuuksia. Pystyvektoreita kutsutaan *muuttujiksi* tai *attribuuteiksi*. Matriisin rivit r_1, r_2, …, r_n (vaakavektorit) kuvaavat puolestaan olioita. Vaakavektoreita kutsutaan yleensä *havainnoiksi* tai *esimerkeiksi*. Voidaan ajatella, että matriisi E virittää *havainto- tai olioavaruuden*. Tässä työssä etäisyys- ja samankaltaisuusfunktioiden yhteydessä esimerkistä käytetään yleensä merkintää e_i, x tai y, attribuutista symbolia a ja esimerkkien x ja y attribuutin a arvoihin viitataan merkinnöillä x_a ja y_a. Matriiseja merkitään jatkossa isolla kursivoidulla kirjaimella E. Olioavaruuden *dimensiolla* eli *ulottuvuudella* tarkoitetaan matriisin sarakemäärää eli attribuuttien lukumäärää. Havaintomatriisin riveillä ja sarakkeilla on eri merkitykset ja tästä syystä matriisia sanotaankin *kaksitilaiseksi* matriisiksi. *Yksitilainen* matriisi on sellainen matriisi, jonka riveillä ja sarakkeilla on sama merkitys. Tällainen on esimerkiksi *etäisyysmatriisi*, johon tutustutaan tarkemmin kohdassa 3.2.

2.2. Mitta-asteikoista

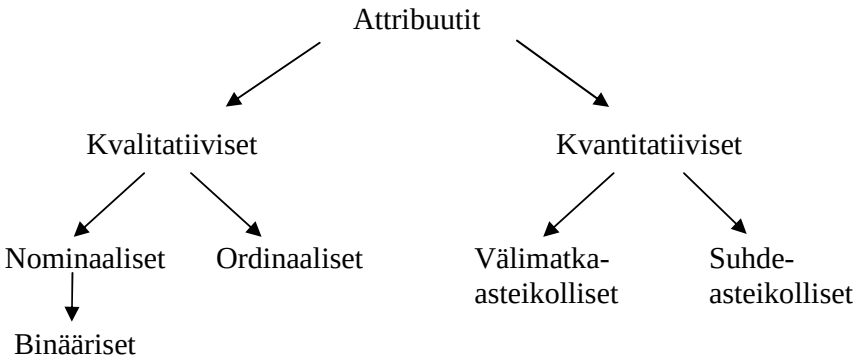
Attribuutin *mitta-asteikko* [Sharma, 1996] ilmaisee, kuinka paljon attribuutilla on informaatiota arvojensa välisistä suhteista. Reaalimaailmaa kuvaava data on moniulotteista ja harvoin mitta-asteikon suhteen homogeenistä. Huomioimalla mitta-asteikkojen ominaispiirteet etäisyys- ja samankaltaisuusfunktiossa vältetään tilanne, jossa käytettäi-

siin mitta-asteikolle sopimattomia operaatioita. Myöhemmin tarkastellaan etäisyysfunktioita, jotka hyödyntävät attribuuttien mitta-asteikkoinformaatiota.

Attribuutit jaetaan tavallisesti mitta-asteikkonsa perusteella *kvalitatiivisiin* sekä *kvantitatiivisiin* attribuutteihin. Tämän yleisemmän tason jaottelun jälkeen attribuutit jaotellaan yleensä vielä

1. *nominaalisiin,*
2. *ordinaalisiin,*
3. *välimatka-asteikollisiin* ja
4. *suhdeasteikollisiin* attribuutteihin.

Jaottelu on esitetty kuvassa 2.2. Kvalitatiivisiin attribuutteihin luetaan nominaaliset sekä ordinaaliset attribuutit. Kvantitatiivisiin luetaan välimatka-asteikolliset ja suhdeasteikolliset attribuutit. Tavallisesti heterogeenisissä läheisyysfunktioissa ordinaalisia attribuutteja ei käsitellä millään erityisellä tavalla, vaan ne käsitellään joko nominaalisten tai kvantitatiivisten attribuuttien tavoin. Yleensä läheisyysfunktiot eivät myöskään erottele välimatka-asteikollisia ja suhdeasteikollisia attribuutteja, koska ne ovat luonteeltaan hyvin lähellä toisiaan.



Kuva 2.2. Attribuuttien perinteinen jaottelu tilastotieteissä.

Mitta-asteikot lueteltiin edellä heikoimmasta vahvimpaan. Vahvempi sisältää aina myös heikomman asteikon ominaisuudet. Esimerkiksi suhdeasteikolliset attribuutit sisältävät kaikkien muiden mitta-asteikoiden ominaisuudet.

Nominaaliset attribuutit ovat attribuutteja, joiden arvoille ei voida määrittää järjestystä. *Väri* on esimerkki tällaisesta attribuutista. Eri värejä ei voi selvästikään laittaa yksiselitteiseen, yleisesti hyväksyttyyn järjestykseen. Etäisyyslaskentaa varten attribuutin arvot tulee kuitenkin *koodata* jollain tavalla. Nominaalisen attribuutin arvot voidaan

koodata periaatteessa millä tavalla tahansa, koska järjestystä niiden välillä ei ole. Helpoin ja yleisin tapa on esittää ne kokonaislukuina. Väriesimerkin tapauksessa voitaisiin määrätä esimerkiksi *punainen* = 0, *valkoinen* = 1, *musta* = 2 jne. Toisaalta menettämättä mitään tietoa voitaisiin määrätä *punainen* = 2, *valkoinen* = 1, *musta* = 0. Juuri nominaaliset attribuutit ovat ongelmallisimpia laskettaessa olioiden välisiä etäisyyksiä ja samankaltaisuuksia, koska ne kertovat ainoastaan, eroavatko kaksi oliota toisistaan. Ne eivät kerro sitä, millä tavalla tai kuinka paljon oliot eroavat. Näin ollen esimerkiksi nominaalisten arvojen erotusten vertailu ei anna mielekästä informaatiota olioiden läheisyydestä.

Nominaalisen attribuutin erikoistapaus on *binäärinen* attribuutti, joka voi saada kaksi eri arvoa. Binäärisiä attribuutteja on kahden tyyppisiä. Binäärinen attribuutti voi kertoa jonkin ominaisuuden laadun, jollaisesta attribuutista esimerkki on sukupuoli. Binäärinen attribuutti voi myös kuvata ilmeneekö oliossa jokin ominaisuus vai ei. Tällainen on esimerkiksi attribuutti, joka kertoo onko eläimellä siivet. Binäärisille attribuuteille on runsaasti samankaltaisuusfunktioita [Gower, 1985]. Joitakin niistä tarkastellaan kohdassa 5.2.

Ordinaaliset eli *järjestysasteikolliset* attribuutit ovat attribuutteja, joiden arvot voidaan asettaa mielekkääseen järjestykseen, mutta joiden kahden peräkkäisen arvon välimatka ei välttämättä ole aina sama. Esimerkkinä ordinaalisesta attribuutista voi ajatella vaikkapa taudin *vakavuusastetta*. Voidaan sanoa, että joku tila on vakavampi kuin joku toinen. Sen sijaan sitä, kuinka suuri ero taudin vakavuudessa on, ei voida objektiivisesti tämän attribuutin perusteella määrittää. Koodaustapa on ordinaalisilla attribuuteilla vapaa, kunhan arvojen järjestys säilyy. Kahden ordinaalisen arvon erotus kertoo enemmän kuin kahden nominaalisen, mutta sekään ei kerro muusta kuin arvojen järjestyksestä.

Välimatka- eli *intervalliasteikolliset* attribuutit poikkeavat ordinaalisista siinä, että niiden arvojen välit ovat yhtä suuret. Esimerkki välimatka-asteikollisesta attribuutista on lämpötila muilla kuin *Kelvin*-asteikoilla mitattuna. Attribuutti sisältää sekä nominaalisen, että ordinaalisen ominaisuuden. Lisäksi tiedetään, että esimerkiksi 100 *Celsiusta* tarkoittaa 50 Celsius-astetta lämpimämpää kuin 50 astetta. Toisaalta mitään tietoa ei menetetä, vaikka käytettäisiin Celsiuksen asemesta mitta-yksikkönä *Fahrenheit*-asteikkoa (122 ja 212 astetta). Näin ollen esimerkiksi 100 Celsius-astetta ei tarkoita kaksi kertaa niin lämmintä kuin 50 Celsius-astetta.

Suhdeasteikolliset attribuutit kertovat olioiden suhteesta kaikkein eniten. Paitsi, että ne sisältävät muiden asteikkojen ominaisuudet, suhdeasteikollisilla attribuuteilla on lisäksi *absoluuttinen nollapiste*, jossa ominaisuus ikään kuin häviää. Tämä tarkoittaa, että ne ilmaisevat myös olion suhteellisen eron toiseen olioon. Esimerkki suhdeasteikollisesta attribuutista on *paino* tai lämpötila mitattuna *Kelvin*-asteikolla. Esimerkiksi kahden kilogramman painoinen kivi painaa kaksi kertaa enemmän kuin yhden kilogramman painava.

Usein joudutaan pohtimaan myös ovatko kvantitatiiviset attribuutit luonteeltaan *diskreettejä* tai *jatkuvia*. Yleensä diskreetiksi käsitetään sellainen kvantitatiivinen attribuutti, jonka kahden arvon välissä voi olla vain äärellinen määrä mahdollisia arvoja. Jatkuvaksi käsitetään sellainen attribuutti, jonka kahden arvon välissä voi periaatteessa olla ääretön määrä erilaisia arvoja. Käytännössä tietenkään äärettömän tarkkoja mittauslaitteita ei ole. Vaikka attribuutti olisi luonteeltaan jatkuva, etäisyysfunktion kannalta yleensä vasta esitystapa ratkaisee, käsitelläänkö attribuutti diskreettinä vai jatkuvana. Jatkossa tarkoitetaan diskreeteillä sellaisia attribuutteja, joiden arvot on koodattu kokonaisluvuiksi tiedonkeruuvaiheessa ja jatkuvilla sellaisia, joiden arvot on tallennettu reaalilukuina. Näin ollen esimerkki jatkuvasta attribuutista voisi olla ihmisen paino kilogrammamitalla mitattuna grammojen tarkkuudella. Diskreetistä attribuutista esimerkki on ihmisen ikä vuosissa.

Etäisyys- ja samankaltaisuusfunktiot eivät yleensä hyödynnä kaikkien mitta-asteikkojen erityisominaisuuksia. Niin sanotut homogeeniset funktiot eivät hyödynnä tietoja eri mitta-asteikoista lainkaan, vaan käsittelevät kaikki attribuutit samalla tavalla. Useimmat heterogeeniset funktiot tyytyvät jaottelemaan attribuutit *nominaalisiin* ja *kvantitatiivisiin*. Tässä jaottelussa kvantitatiivisiin kuuluvat siis välimatka-asteikollisten ja suhdeasteikollisten lisäksi myös ordinaaliset attribuutit.

2.3. Datan muuntaminen ja redusointi

Tiedonlouhinnassa datalle joudutaan usein tekemään erilaisia *muunnoksia*, jotta valittu louhintamenetelmä sopisi paremmin ongelmanratkaisuun. Reaalimaailman dataa käsiteltäessä ongelmia aiheuttavat esimerkiksi erilaiset mittayksiköt ja attribuuttien erisuuret vaihteluvälit. Lisäksi arvoa jokaisen tutkittavan olion jokaiselle attribuutille ei välttämättä ole saatavilla tai pystytä mittaamaan. Etäisyysfunktion kannalta ongelmana on myös attribuutin painoarvo, eli kuinka paljon attribuutti vaikuttaa etäisyyslaskennassa.

Normalisointi eli *standardointi* on menetelmä, jolla havaintomatriisin attribuutit normalisoidaan eli niiden painoarvot yhdenmukaistetaan. Tällöin muunnetaan attribuuttien arvoja niin, että kaikkien attribuuttien arvot saadaan jakautumaan jollekin samanpituiselle välille. Oletetaan, että havaintomatriisin arvot ovat reaalilukuja. Yksinkertainen tapa normalisoida attribuutin a arvo x_a on käyttää kaavaa

$$z_a = \frac{|x_a - \min(a)|}{\max(a) - \min(a)}, \qquad (2.1)$$

missä siis arvon ja attribuutin minimin erotus jaetaan erotuksella $\max(a) - \min(a)$, joka on attribuutin *vaihteluväli*. Normalisoidun attribuutin z_a arvoalueeksi tulee näin ollen [0,1]. Ongelma tässä tavassa on se, että arvot, jotka poikkeavat merkittävästi muista (*outlier),* pääsevät dominoimaan normalisointia. Usein attribuutin arvot jaetaan keski-

hajonnalla. Wilson ja Martinez [1997] käyttävät neljää keskihajontaa. Koska normaalijakautuneessa datassa 95 prosenttia havaintoarvoista sijoittuu keskiarvon molemmille puolille kahden keskihajonnan sisään, suurin osa (97,5%) muunnetuista arvoista jää välille [min(a)/4σ, 1] ja loput (2,5%) ovat ykköstä suurempia.

Tilastotieteessä standardointi suoritetaan tyypillisesti siten, että attribuutin arvoista vähennetään attribuutin keskiarvo ja erotus jaetaan attribuutin keskihajonnalla:

$$z_a = \frac{|x_a - \bar{a}|}{\sigma} \tag{2.2}$$

Näin muunnetun attribuutin keskiarvo on nolla ja varianssi yksi [Sharma, 1996].

Huomattava on, että normalisointi sisältää oletuksen, että kaikki attribuutit ovat yhtä tärkeitä, eli että niillä on yhtä suuri vaikutus esimerkiksi esimerkkien välisiä etäisyyksiä ja samankaltaisuuksia laskettaessa. Näin ei aina ole, joten normalisointi voi olla yhtä väärä toimenpide, kuin normalisoimatta jättäminenkin. Perusongelmia dataa analysoitaessa onkin attribuuttien painoarvojen määrittely, jonka erikoistapaus normalisointikin on [Hand et al., 2001]. Kyseisen ongelman tarkempi pohdinta jää kuitenkin tämän tutkielman ulkopuolelle ja jatkossa oletetaan, että mitään painoarvoja ei ole attribuuteille määritelty poikkeuksena VDM-metriikka (katso kohta 5.4), joka rakentaa itse attribuuteille painoarvot luokkatietojen pohjalta.

Eräs tyypillinen ongelma reaalimaailman dataa käsiteltäessä on puuttuvat tiedot. Tässä yhteydessä puuttuvalla datalla ei tarkoiteta sitä, että esimerkiltä puuttuu jokin ominaisuus, vaan tilannetta jolloin jollekin attribuutille ei ole mittaustietoa, ei edes sitä puuttuuko ominaisuus vai ei. Esimerkiksi lääketieteellisessä tiedonlouhinnassa puuttuvan datan ongelmaa on tutkittu paljon [Doyle et al., 1995]. Monet tiedonlouhintatekniikat vaativat, että havaintomatriisi on täydellinen. Tästä syystä datan puuttuvat arvot on usein imputoitava. *Imputoinnilla* tarkoitetaan puuttuvien arvojen paikkausta uusilla estimoiduilla arvoilla [Little and Rubin, 1986]. Oletetaan havaintomatriisin alkioiden olevan reaalilukuja. Yksinkertainen tapa estimoida attribuutin *a* puuttuvat arvot on käyttää esimerkiksi *a*:n *keskiarvoa*, *mediaania* tai *moodia* perusjoukossa. Jos datassa on olemassa luokkatiedot voidaan käyttää myös luokkakohtaisia tunnuslukuja. Hieman edellisiä monimutkaisempi menetelmä on *maksimiuskottavuusestimointi* (expectation maximization) [Hand et al., 2001]. Mikäli puuttuvia arvoja jollekin attribuutille on paljon, saattaa olla järkevämpää jättää kyseinen attribuutti tarkastelun ulkopuolelle. Samalla tavoin voidaan jättää tarkastelun ulkopuolelle esimerkit, joilla puuttuvia tietoja on paljon.

3. Luokittelusta ja havaintoperustaisesta oppimisesta

Hyvin monentyyppiset tiedonlouhintatehtävät ovat palautettavissa *luokitteluongelmiksi*, joissa annetut esimerkit on sijoitettava äärelliseen määrään toisensa poissulkevia luokkia [Quinlan, 1986]. Luokittelu voi olla joko *ohjattua* tai *ohjaamatonta*. Ohjatussa luokittelussa, esimerkiksi *k-lähimmän naapurin menetelmässä* (k-nearest neighbor method, KNN), luokat tiedetään etukäteen ja uudet tapaukset tulee leimata opetusjoukon ja luokiteltavan esimerkin etäisyyksien tai muun informaation perusteella johonkin jo tiedettyyn luokkaan. Ohjaamattomassa luokittelussa, esimerkiksi *klusteroinnissa* [Everitt et al., 2001; Jain and Dubes, 1988; Jain et al., 1999], luokkia ei tiedetä etukäteen ja tehtävänä on ryhmitellä tapaukset esimerkiksi etäisyyksien perusteella.

Luokitteluongelmat ratkaistaan tyypillisesti *koneoppimismenetelmien* avulla [Mitchell, 1997]. Koneoppimismenetelmissä ideana on tavallisesti etsiä data- eli *opetusjoukosta* hahmoja ja luokkarakenteita, joiden perusteella rakennetaan erilaisia luokitteluheuristiikoita, kuten esimerkiksi päättelypuita tai neuroverkkoja, joita voidaan käyttää uusien tapauksien eli *luokittelu- tai testijoukon* luokittelemiseen [Long, 2001]. Koneoppimismenetelmiä on sovellettu monella alalla. Suosittuja sovelluksia ovat esimerkiksi erilaiset lääketieteelliset päätöksenteon tukijärjestelmät eli asiantuntija-järjestelmät (expert system) [Auramo et al., 1993; Brasil et al., 2001; Laurikkala, 2001a].

Yksi tapa on luokitella esimerkit suoraan muistissa olevien esimerkkien eli opetusjoukon perusteella rakentamatta mitään heuristiikkaa tai mallia. Tällaista tapaa kutsutaan *havaintoperustaiseksi oppimiseksi* (instance-based learning) [Mitchell, 1997; Wilson and Martinez, 1997] tai *muistipohjaiseksi päättelyksi* (memory-based reasoning) [Stanfill and Waltz, 1986]. Luokittelu perustuu opetusjoukon esimerkkien ja luokiteltavien esimerkkien läheisyyksiin. Tässä menetelmässä varsinainen oppiminen tarkoittaa lähinnä esimerkkien tallettamista ja joidenkin tunnuslukujen laskemista, ja suurin työ tehdään vasta siinä vaiheessa, kun uusia esimerkkejä luokitellaan. Tämän takia havaintoperustaista oppimista kutsutaan myös *laiskaksi oppimiseksi.*

Tässä luvussa tarkastellaan muistipohjaisen päättelyn mahdollistavaa lähimmän naapurin menetelmää, jota käytetään myöhemmin (katso luvut 8 ja 9) arvioitaessa läheisyysfunktioiden keskinäistä paremmuutta. Lisäksi tarkastellaan läheisyysfunktioiden roolia tässä menetelmässä.

3.1. Lähimmän naapurin menetelmästä

KNN-menetelmässä (katso algoritmi 3.1) lähtökohta on, että opetusjoukko T on esitetty kuvan 2.1 mukaisena havaintomatriisina [Mitchell, 1997]. Lisäksi oletetaan esimerkki e_q, joka on tarkoitus luokitella opetusjoukosta saatavan tietämyksen perusteella. Käytännössä esimerkki e_q luokitellaan sen ja opetusjoukon esimerkkien välisten läheisyyksien perusteella. Yleensä läheisyyksiä ei tunneta etukäteen, joten luokitteluvaiheessa

lasketaan esimerkin e_q ja opetusjoukon esimerkkien väliset läheisyydet käyttäen jotain etäisyys- tai samankaltaisuusfunktiota. Jos on käytetty etäisyysfunktiota, niin tämän jälkeen haetaan opetusjoukosta k esimerkkiä joiden etäisyys esimerkkiin e_q on mahdollisimman pieni. Esimerkki e_q leimataan luokalla, jonka frekvenssi on k-lähimmän naapurin joukossa suurin.

Opetusalgoritmi

- Lisää jokainen opetusesimerkki joukkoon T

Luokittelualgoritmi

- Luokitellaan esimerkki e_q,
  - Oletetaan, että $e_1 \ldots e_k$ ovat esimerkin e_q k lähintä naapuria
  - Palautetaan ennustettu luokkaleima

$$\hat{f}(e_q) \leftarrow \arg\max_{c \in C} \sum_{i=1}^{k} \delta(c, f(e_i)),$$

missä $\delta(a,b) = 1$, jos $a = b$ ja $\delta(a,b) = 0$ muutoin ja C on luokkaleimojen joukko.

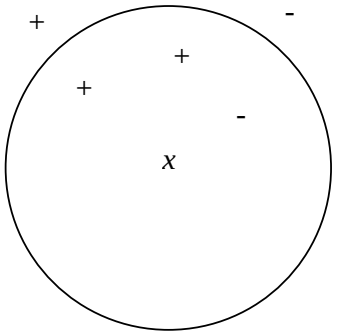
Algoritmi 3.1. Lähimmän naapurin luokittelu –algoritmi.

KNN-menetelmässä etsitään siis opetusjoukosta T esimerkin e_q k ($1 \leq k \leq |T|$) lähintä naapuria ja lähimpien naapurien joukosta valitaan yleisin luokka. Usein kuitenkin useamman luokan frekvenssi on sama. Tarkennetaan tästä syystä luokittelualgoritmia hieman. Olkoon S lähimpien naapurien saatu joukko. Lasketaan joukosta S luokkafrekvenssit. Olkoon S' yleisimpien luokkien (eli luokkien, joiden frekvenssi on suurin joukossa S) esimerkeistä muodostettava joukko. Jos S' koostuu yhden luokan esimerkeistä, valitaan tämä luokka. Jos joukossa S' on usean luokan edustajia, voidaan käyttää yhden lähimmän naapurin menetelmää ainoastaan joukossa S'. Jos tämäkään ei määrää yksikäsitteistä luokkaa, voidaan käyttää joko satunnaisvalintaa tai tarvittaessa jotain determinististä valitsemistapaa (esimerkiksi valitaan käsittelyjärjestyksessä ensimmäisen esimerkin luokka). Erikoistapaus on tilanne, jolloin k lähimmän naapurin etäisyys on sama kuin $k+1$ lähimmän naapurin etäisyys. Tilanne voidaan jättää huomioimatta mikäli se on harvinainen. Voidaan kuitenkin toimia myös siten, että valitaan joukkoon S $k + n$ lähimmät naapurit ($n \geq 1$), joiden etäisyys esimerkkiin e_q on sama kuin k lähimmän naapurin ja esimerkin e_q välinen etäisyys. Tämän jälkeen toimitaan kuten edellä.

Aika- ja muistivaatimuksia tarkasteltaessa KNN-menetelmän hyvänä puolena on opetusalgoritmin yksinkertaisuus ja nopeus. Opetusalgoritmissa opetusjoukon esimerkit tallennetaan muistiin ja mahdollisesti lasketaan joitain tunnuslukuja aineistosta. Suurten aineistojen kohdalla suuri talletustilavaatimus saattaa tosin muodostua ongelmaksi. Minimissään KNN vaatii tilaa $O(mn)$, missä m on attribuuttien ja n esimerkkien luku-

määrä. Luokittelualgoritmin aikavaatimus $O(mnd)$ riippuu opetusaineiston koosta ja käytetyn läheisyysfunktion aikavaatimuksesta d. Näin ollen tavallista monimutkaisempaa funktiota kuten esimerkiksi VDM-funktiota (katso kohta 5.4) käytettäessä luokittelualgoritmin suoritus saattaa kestää epäkäytännöllisen kauan, jos opetusjoukko on suuri. KNN-menetelmän aika- ja talletustilavaatimuksiin palataan kohdassa 8.2.

Kuvassa 3.1 on esitetty tilanne, jossa tehtävänä on luokitella olio *x* kolmen lähimmän naapurin menetelmällä joko positiiviseksi tai negatiiviseksi [Mitchell, 1997]. Olioavaruus on kaksiulotteinen ja kvantitatiivinen, joten käytetään tason geometriasta tuttua euklidista etäisyyttä (katso kohta 5.1). Havaitaan, että kolmen lähimmän naapurin joukossa on kaksi positiivista ja yksi negatiivinen. Tästä syystä *x* luokitellaan positiiviseksi.



Kuva 3.1. Kolmen lähimmän naapurin luokittelutilanne. Kuvassa *x* merkitsee luokiteltavaa oliota ja + ja – positiivisen ja negatiivisen luokan olioita.

3.2. Läheisyysfunktiot luokittelussa

Läheisyysfunktiot ovat luokittelussa keskeinen väline. Niiden tavoitteena on virittää luokittelun perustaksi todellisuutta vastaava tai muuten tarkoitukseen sopiva etäisyys- tai samankaltaisuusavaruus tutkittujen olioiden välille.

Tarkastellaan tilannetta, jossa lähtökohtana on relaatiotietokannan taulu (katso taulukko 3.1), jonka sisältämät esimerkit on tarkoitus ryhmitellä kahteen joukkoon siten, että joukkoihin sijoitetaan esimerkit, jotka ovat mahdollisimman lähellä toisiaan.

Taulukko 3.1. Esimerkkiaineisto.

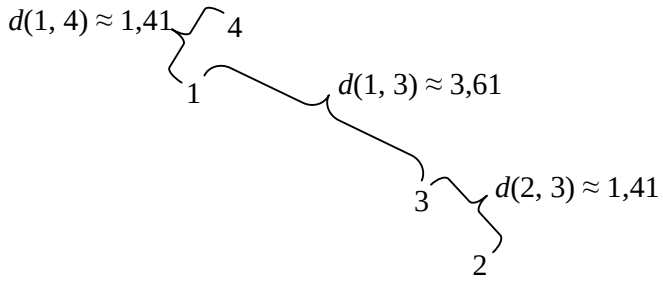
id	luokka	a1	a2
1	1	0	3
2	2	4	0
3	2	3	1
4	1	1	4

Lasketaan taulukon 3.1 esimerkkien väliset etäisyydet käyttäen normaalia euklidista etäisyyttä, jota tarkastellaan myöhemmin kohdassa 5.1. Näin saatu etäisyysmatriisi on esitettynä taulukossa 3.2. Huomattava on, että matriisin oikea yläkolmio voidaan jättää tyhjäksi, koska sama tieto saadaan jo vasemmasta alakolmiosta. Tämä johtuu etäisyysfunktion $d(x, y)$ ominaisuudesta $d(x, y) = d(y, x)$, missä $\forall x, y \in E$ (katso luku 4).

Taulukko 3.2. Etäisyysmatriisi taulukon 3.1 esimerkkiaineistolle.

	1	2	3	4
1	0			
2	5	0		
3	3,61	1,41	0	
4	1,41	5	3,61	0

Etäisyysmatriisia voidaan havainnollistaa esimerkiksi kuvan 3.2 tapaan. Kuvasta 3.2 nähdään, että oliot muodostavat kaksi selvästi erillistä ryhmää. Alkuperäisen aineiston luokkajako näkyy myös kuvassa 3.2, joten käytetty euklidinen etäisyysfunktio sopii siis varsin hyvin juuri tähän aineistoon. Aineisto olisi voinut luokittua huonommin, mikäli olisi valittu tehtävään huonosti sopiva funktio.



Kuva 3.2. Taulukon 3.2 etäisyysmatriisi havainnollistettuna tasolla.

Funktiot voidaan jaotella monilla tavoin ja yksi tapa on jakaa ne *homogeenisiin* ja *heterogeenisiin funktioihin*. Homogeenisiksi kutsutaan funktioita, jotka olettavat, että käsiteltävän aineiston attribuuttien mitta-asteikko on sama. Heterogeenisiksi puolestaan kutsutaan funktioita, jotka huomioivat erilaiset mitta-asteikot aineistossa ja hyödyntävät mitta-asteikkojen ominaisuuksia läheisyyslaskennassa. Homogeenisen funktion erikoistyyppi on *binäärinen funktio*, joka olettaa attribuuttien olevan binäärisiä, eli että jokaisella attribuutilla on vain kaksi mahdollista arvoa. Binäärifunktiot ovat useimmiten määritelty samankaltaisuusfunktiona. Seuraavassa luvussa käsitellään etäisyys- ja samankaltaisuusfunktioiden keskeisiä ominaisuuksia. Luvussa 5 tarkastellaan joitain homogeenisiä ja luvussa 6 joitain heterogeenisiä funktioita.

4. Läheisyys, etäisyys ja samankaltaisuus

Edellä on kuvattu tiedonlouhintaprosessin ensimmäisiä vaiheita, joiden jälkeen data on valmiina luokiteltavaksi. Ennen kuin käydään läpi tunnettuja etäisyys- ja samankaltaisuusfunktioita, tarkastellaan joitain etäisyys- ja samankaltaisuusfunktioiden yleisiä ominaisuuksia. Tässä luvussa määritellään etäisyyden, samankaltaisuuden ja metrisyyden käsitteet. Luvun määritelmät perustuvat Bobergin esitykseen [1999] ellei toisin mainita.

Vertailtaessa olioita käytetään yleensä termejä *läheisyys* (proximity), *samankaltaisuus* (similarity), *erilaisuus* (dissimilarity) ja *etäisyys* (distance). Läheisyys on yleisempi termi, joka voi tarkoittaa sekä samankaltaisuutta että erilaisuutta. Erilaisuus ja etäisyys tarkoittavat samaa asiaa. Usein kuitenkin vain etäisyys ja samankaltaisuus ovat määritelty täsmällisesti. Käsitteellä samankaltaisuus on päinvastainen merkitys etäisyyteen verrattuna. Mitä enemmän oliot muistuttavat toisiaan sitä suurempi niiden välinen samankaltaisuus ja sitä pienempi on niiden välinen etäisyys on.

Olkoot *x*, *y* ja *z* olioita avaruudessa *E* ja $\boldsymbol{R}$ tarkoittaa reaalilukujen joukkoa.

Määritelmä 4.1. Oletetaan, että d_0 on reaaliluku. Funktio $d : E \times E \to \boldsymbol{R}$ on etäisyysfunktio, jos

1. $d(x, y) \geq d_0 \qquad \forall x, y \in E$,
2. $d(x, x) = d_0 \qquad \forall x \in E$, ja
3. $d(x, y) = d(y, x) \quad \forall x, y \in E$.

Määritelmä 4.2. Oletetaan, että s_0 on reaaliluku. Funktio $s : E \times E \to \boldsymbol{R}$ on samankaltaisuusfunktio, jos

1. $s(x, y) \leq s_0 \qquad \forall x, y \in E$,
2. $s(x, x) = s_0 \qquad \forall x \in E$, ja
3. $s(x, y) = s(y, x) \quad \forall x, y \in E$.

Näin ollen etäisyysfunktiolla on alaraja d_0 (usein 0) ja samankaltaisuusfunktiolla yläraja s_0 (usein 1), jotka kuvaavat olioiden samantasoista vastaavuutta. Ehdot 2 ja 3 tunnetaan refleksisyys- ja symmetrisyysehtoina.

Määritelmä 4.3. Jos etäisyysfunktio $d : E \times E \to \boldsymbol{R}$ täyttää kolmioepäyhtälön ehdon

$$d(x, y) + d(y, z) \geq d(x, z) \quad \forall x, y, z \in E,$$

niin *d*:n sanotaan olevan *pseudometriikka E*:ssä, ja (*E*, *d*):n sanotaan olevan *pseudometrinen avaruus*.

Jos etäisyysfunktio ei toteuta kolmioepäyhtälöä, sanotaan etäisyysfunktion olevan ei-metrinen. Jos kolmioepäyhtälö ei päde, on *E*:ssä sellaiset oliot *x*, *y* ja *z*, että $d(x, y) > d(x, z) + d(z, y)$. Tämä tarkoittaa, että etäisyysarvo $d(x, y)$ ei ole lyhin etäisyys olioiden *x* ja *y* välillä, vaan on olemassa ”väliolio” *z*, joka ”lyhentää” etäisyyttä oli-

oiden *x* ja *y* välillä. Tämä tarkoittaisi, että lyhin polku pisteiden *x* ja *y* välillä olisi lyhyempi kuin viivan pituus pisteestä *x* pisteeseen *y*, mikä on mahdotonta tavallisessa euklidisessa avaruudessa.

Määritelmä 4.4. Jos etäisyysfunktio $d : E \times E \to \boldsymbol{R}$ on pseudometriikka ja lisäksi täyttää ehdon

$$d(x, y) = d_0 \Leftrightarrow x = y ,$$

niin *d*:tä sanotaan *metriikaksi E*:ssä.

Määritelmä 4.5. Jos etäisyysfunktio $d : E \times E \to \boldsymbol{R}$ on metriikka ja lisäksi täyttää *ultrametrisen epäyhtälön* ehdon

$$\max\{d(x, y), d(y, z)\} \geq d(x, z) \quad \forall x, y, z \in E ,$$

niin *d*:n sanotaan olevan *ultrametriikka E*:ssä, ja (*E*, *d*):n sanotaan olevan ultrametrinen avaruus. Huomioitava on, että ultrametrisessä epäyhtälössä etäisyysarvoista kahdella tulee olla sama arvo. Helposti on nähtävissä, että ultrametrinen epäyhtälö toteuttaa myös kolmioepäyhtälön ehdot, joten ultrametrinen avaruus on aina myös metrinen avaruus. Olioavaruus on harvoin ultrametrinen. Esimerkiksi taso $\boldsymbol{R}^2$ varustettuna tavallisella euklidisella etäisyydellä on metrinen, mutta ei ultrametrinen. Myöskään myöhemmin käsiteltävistä etäisyysfunktioista yksikään ei viritä ultrametristä avaruutta.

Aiemmin mainittiin, että voidaan käyttää samankaltaisuusfunktiota etäisyysfunktion asemesta. Kolmioepäyhtälö samankaltaisuusfunktiolle *s* on

$$|s(x, y) + s(y, z)| \cdot s(x, z) \geq s(x, y)s(y, z) \quad \forall x, y, z \in E . \tag{4.1}$$

Samankaltaisuusfunktio, joka toteuttaa kolmioepäyhtälön 4.1 ja ehdon

$$s(x, y) = s_0 \Leftrightarrow x = y \tag{4.2}$$

on metriikka. Ultrametrinen epäyhtälö samankaltaisuusfunktiolle *s* on

$$\min\{s(x, y), s(y, z)\} \leq s(x, z) \quad \forall x, y, z \in E . \tag{4.3}$$

Etäisyyden ja samankaltaisuuden välillä on monia hyödyllisiä suhteita, joiden avulla samankaltaisuudesta voidaan johtaa etäisyys ja päinvastoin [Späth, 1980].

Määritelmä 4.6. Funktio *s*(*x*, *y*) on metrinen samankaltaisuusfunktio, jos ja vain jos funktio $s_0 - s(x, y)$ on metrinen etäisyysfunktio.

Määritelmä 4.7. Jos $d(x, y)$ on metrinen etäisyysfunktio, jolla on äärellinen yläraja $D = \max\{d(x, y) | x, y \in E\}$, niin D – $d(x, y)$ on metrinen samankaltaisuusfunktio.

Määritelmä 4.8. Oletetaan, että $d(x, y) > 0$ kaikille esimerkeille x ja y. Tällöin $d(x, y)$ on metrinen etäisyysfunktio, jos ja vain jos $1/d(x, y)$ on metrinen samankaltaisuusfunktio.

Voidaan sanoa, että etäisyyden ja samankaltaisuuden käsitteet ovat yhtä eksakteja [Boberg, 1999]. Yhtenäisyyden vuoksi tässä tutkielmassa käytetään käsitettä etäisyys samankaltaisuuden asemesta. Samankaltaisuusfunktiot muunnetaan tarvittaessa etäisyysfunktioiksi käyttäen määritelmää 4.6.

5. Homogeenisiä etäisyysfunktioita

Tässä luvussa tutustutaan joihinkin tavallisiin läheisyysfunktioihin, joille on yhteistä homogeenisyys. Homogeenisyydellä tarkoitetaan tässä, että funktio osaa käsitellä sopivasti aineistoja, joissa data on kuvattu eri mitta-asteikoilla. Reaalimaailmaa kuvaavaa dataa käsiteltäessä homogeenisyys on selvä ongelma, koska usein tällainen data on kuvattu eri mitta-asteikoilla.

Kohdassa 5.1 esitetään yleisesti käytetty *Minkowski*-funktio ja sen erikoistapauksia, joista tavallisin on *euklidinen etäisyysfunktio*. Minkowski-funktiot toimivat varsin hyvin attribuuteille, jotka sisältävät välimatka- tai suhdeominaisuuden, mutta ordinaalisten ja varsinkin nominaalisten attribuuttien kanssa esiintyy ongelmia.

Kohdassa 5.2 tarkastellaan joitain tavallisia *binäärifunktioita*, jotka soveltuvat vain pelkillä binäärisillä attribuuteilla kuvatuille aineistoille. Binäärisellä attribuutillahan tarkoitetaan kaksitasoista nominaalista attribuuttia. Tällainen data on kätevää, koska yhtä attribuutin arvoa kohti tarvitaan vain yksi bitti talletustilaa. Kohdassa 5.3 tarkastellaan myös monitasoiselle nominaaliselle soveltuvaa yksinkertaista sovitusfunktiota.

Kohdassa 5.4 tutustutaan VDM-funktioon, jossa etäisyyslaskenta perustuu jokaiselle attribuutin arvo-luokka-parille laskettuun painoarvoon, joka on sitä suurempi, mitä enemmän kyseinen arvo korreloi luokan kanssa. VDM-funktio on käyttökelpoinen lähinnä vain nominaaliselle ja ordinaaliselle datalle.

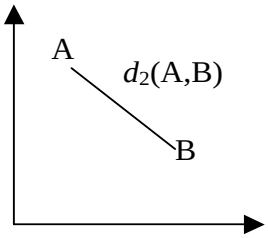
Kohdassa 5.5 tarkastellaan kosinietäisyysfunktiota, joka perustuu vektoreiden väliseen kulmaan. Funktio sopii tilanteisiin, joissa pyritään etsimään esimerkkejä, joiden *muodot* ovat toisiaan lähellä.

5.1. Minkowski-metriikat

Kohdassa 2.1 määriteltiin olioavaruus *E*, joka esitettiin $n \times m$-matriisin avulla $E = (e_{ij})$, missä $i = 1,2,...,n$ ja $j = 1,2,...,m$. Seuraavassa oletetaan, että *x* ja *y* ovat avaruuden *E* olioita ja attribuuttia *a* vastaavat arvot $x_a, y_a \in \boldsymbol{R}$.

Kaikkein tunnetuin etäisyysfunktio lienee *euklidiseen etäisyyteen* perustuva funktio. Ajatellaan esimerkit *x* ja *y* pisteiksi *m*-akselisessa koordinaatistossa, jossa akselit vastaavat attribuutteja. Euklidinen etäisyys on näiden kahden pisteen välille piirretyn janan pituus (katso kuva 5.1). Määritellään euklidinen etäisyysfunktio

$$d_2(x, y) = \sqrt{\sum_{a=1}^{m} (x_a - y_a)^2} \,. \qquad (5.1)$$



Kuva 5.1. Kaksiulotteiset esimerkit A ja B, ja niiden välinen euklidinen etäisyys havainnollistettuna kaksiulotteisella koordinaatistolla.

Euklidinen etäisyysfunktio on erikoistapaus *Minkowski-funktiosta*, joka määritellään

$$d_p(x, y) = \left(\sum_{a=1}^{m} |x_a - y_a|^p \right)^{\frac{1}{p}}, \; p \geq 1. \tag{5.2}$$

Yleisimmät Minkowski-etäisyyden erikoistapaukset ovat metriikat, jotka saadaan, kun $p = 1$, $p = 2$ ja $p = \infty$. Kun $p = 1$, saadaan

$$d_1(x, y) = \sum_{a=1}^{m} |x_a - y_a|, \tag{5.3}$$

joka tunnetaan nimillä *Manhattan-*, *City block-* ja *Taxicab*-etäisyys. Kun $p = \infty$, saadaan maksimietäisyys

$$d_\infty(x, y) = \max_{a=1,\ldots,m} \{ |x_a - y_a| \}. \tag{5.4}$$

Minkowski-funktion erikoistapauksille pätee, että jos $p \geq q$, niin $d_p(x, y) \leq d_q(x, y)$ kaikille olioille x ja y [Boberg, 1999]. Minkowski-funktioon perustuvat funktiot täyttävät määritelmän 4.4 mukaisesti metriikoiden vaatimukset [Gower and Legendre, 1986].

Euklidinen metriikka (ja muut Minkowski-metriikat) antaa mielekkäitä arvoja silloin, kun attribuutit ovat vähintään välimatka-asteikollisia. Esimerkiksi, jos aineistoon olisi kuvattu eri paikoista *kirsikoita, omenoita* ja *meloneja* attribuuttien *paino grammoina* ja *läpimitta sentteinä* avulla, ja tehtävänä olisi näiden perusteella luokittelua uusia tapauksia lajin mukaan, olisi luokittelu euklidisella etäisyydellä todennäköisesti menestyksellistä johtuen attribuuttien sopivista mitta-asteikoista (molemmat suhdeasteikollisia). Jos mukaan lisättäisiin esimerkiksi nominaalinen attribuutti *väri* koodattuna jollain kokonaisluvulla, vaikeutuisi luokittelu euklidisella etäisyydellä jonkin verran.

Värikoodien erotuksellahan ei selvästi ole mielekästä tulkintaa luokiteltaessa, ja jos väri olisi kuvattu erittäin tarkasti vaikkapa kokonaisluvuin 1-100, olisi todennäköistä, että euklidinen etäisyysfunktio harhautuisi. Pitää muistaa, että mikäli opetusjoukon esimerkillä *x* ja luokiteltavalla esimerkillä *y* kyseisen attribuutin *a* arvo olisi sama eli $x_a = y_a$, niin näissä tilanteissa euklidinen etäisyysfunktio ei erehtyisi. Hieman samankaltainen tilanne on myös ordinaalisten attribuuttien kohdalla, joskin nämä voivat joissain tilanteissa sopia nominaalisia paremmin. Mikäli esimerkiksi ordinaalisen attribuutin *taudin vakavuus* koodattujen arvojen väli on aina 1, ja jos kahden esimerkkiparin kohdalla kyseisen attribuutin arvojen erotukset ovat 1 ja 5, niin on todennäköistä, että ensimmäinen esimerkkipari on läheisempi kuin jälkimmäinen (tietenkin vain *taudin vakavuusaste* -attribuutin suhteen), minkä euklidinen etäisyysfunktiokin tunnistaa. Ongelmana on, että euklidinen etäisyysfunktio pyrkii ”hyödyntämään” myös eron suuruusluokkaa, jolla ei ordinaalisten attribuuttien tapauksessa ole merkitystä.

Minkowski-metriikat eivät ole invariantteja attribuuttien arvojen skaalaamiselle. Tämä tarkoittaa, että siirryttäessä yhdestä mittayksiköstä toiseen olioiden välisten etäisyyksien suhteet muuttuvat. Esimerkiksi, jos attribuutin *a* mittayksikkö muutetaan metristä senttimetriin, kasvavat kaikkien olioavaruuden *E* olioiden väliset etäisyydet, ja lisäksi kokonaisetäisyyttä laskettaessa attribuutin *a* painoarvo suhteessa muihin attribuutteihin kasvaa. Kohdassa 2.3 tarkasteltiin datan muunnosmenetelmiä, joita voidaan käyttää etäisyyslaskennan yhteydessä. Muuttujan painoarvot voidaan tasata normalisoinnin avulla, jolloin eliminoidaan mitta-asteikon vaihdon vaikutus etäisyyslaskentaan. Mikäli aineisto ei ole valmiiksi normalisoitua, voidaan normalisointi lisätä funktioon itseensä. Määritellään normalisoitu euklidinen etäisyysfunktio

$$d_2(x,y) = \sqrt{\sum_{a=1}^{m}\left(\frac{x_a - y_a}{std_a}\right)^2}\,, \tag{5.5}$$

missä normalisoimistapa std_a voi olla esimerkiksi $vaihteluväli_a = \max(a) - \min(a)$ tai 4σ (katso kohta 2.3).

5.2. Binäärisiä samankaltaisuusfunktioita

Binäärinen data on nopeasti käsiteltävää ja vähän tilaa vievää, joten usein sekamuotoinen data redusoidaan binääriseksi. Kohdassa 2.2 havaittiin, että binäärisiä attribuutteja voi olla kahdentyyppisiä, eli jonkun ominaisuuden laatua ilmaisevat attribuutit, ja sellaiset attribuutit, jotka vain ilmaisevat ilmeneekö jokin ominaisuus oliossa vai ei. Tätä tietoa voidaan käyttää hyväksi samankaltaisuusfunktiossa. Binääriselle datalle esitetyt läheisyysfunktiot ovat tavallisesti samankaltaisuusfunktioita [Gower, 1985]. Etäisyysfunktioiksi ne on muutettavissa määritelmän 4.6 avulla. Binääriset funktiot perustuvat taulukossa 5.1 esitettyyn nelikenttään.

Taulukko 5.1. Nelikenttä binäärisen data samankaltaisuuslaskentaan.

	$y_j = 1$	$y_j = 0$	yhteensä
$x_j = 1$	a	b	a+b
$x_j = 0$	c	d	c+d
yhteensä	a+c	b+d	a+b+c+d

Taulukkoa 5.1 tulkitaan siten, että esimerkiksi a tarkoittaa niiden attribuuttien lukumäärä, joille esimerkeillä *x* ja *y* on kummallekin arvo 1. Kaikkein tyypillisin samankaltaisuusfunktio binääridatalle on niin sanottu *sovituskerroin* (Matching Coefficient, MC), joka määritellään

$$mc(x, y) = \frac{\mathrm{a} + \mathrm{d}}{\mathrm{a} + \mathrm{b} + \mathrm{c} + \mathrm{d}}. \tag{5.6}$$

MC-funktio olettaa, että a- ja d-tapauksilla on samankaltaisuuden laskemisen kannalta yhtäläinen merkitys (a-tapaus tarkoittaa tilannetta, jossa $x_a = y_a = 1$, ja d-tapaus tilannetta, jossa $x_a = y_a = 0$). Näiden merkitys ei kuitenkaan aina ole sama, koska monesti se, että molemmilta esimerkeiltä puuttuu sama ominaisuus (d), ei kerro välttämättä mitään esimerkkien samankaltaisuudesta. Yleisesti tunnettu *Jaccardin kerroin* (Jaccard's Coefficient, JC) jättää tällaiset tapaukset huomiotta. Määritellään JC

$$jc(x, y) = \frac{\mathrm{a}}{\mathrm{a} + \mathrm{b} + \mathrm{c}}. \tag{5.7}$$

Datalle voi olla ominaista, että kahden esimerkin välillä jonkun ominaisuuden vastaavuus (*a* ja *d*) on samankaltaisuuden laskemisen kannalta merkitsevämpää kuin eivastaavuus (*b* ja *c*). Jollekin datalle taas tilanne voi olla päinvastoin. Tästä syystä on esitetty joukko samankaltaisuusfunktioita, jotka painottavat eri tavalla osumia ja eiosumia. *Rogers ja Tanimoto* [1960] korostavat ei-osumien (kaava 5.8) ja *Gower ja Legendre* [1986] puolestaan osumien merkitystä (kaava 5.9).

$$rt(x, y) = \frac{a + d}{a + 2(b + c) + d}, \tag{5.8}$$

$$gl(x, y) = \frac{a + d}{a + \frac{1}{2}(b + c) + d}. \tag{5.9}$$

Edellisten funktioiden ideoiden pohjalta voidaan määritellä datan ominaisuuksien perusteella erilaisia funktioita, jotka vastaavat kulloisenkin luokitteluongelman tarpeisiin. Esimerkiksi *Sokal ja Sneath* yhdistelevät Jaccardin ja Rogersin ja Tanimoton kertoimen [Gower, 1985] ominaisuuksia:

$$ss(x,y) = \frac{a}{a + 2(b+c)}. \tag{5.10}$$

Edellä tarkasteltiin vain muutamaa samankaltaisuusfunktioita binääriselle datalle. Eri tarkoituksiin sopivia binäärifunktioita on kymmeniä [Gower and Legendre, 1986]. Mitään yksiselitteistä ohjetta jonkun funktion sopivuudesta esimerkiksi johonkin luokitteluongelmaan ei voida antaa. Funktion valinta tulisikin tehdä yleensä aineiston ominaisuuksien perusteella. Tämä pätee binäärisen datan lisäksi muunkin tyyppiseen dataan. Edellä mainittujen samankaltaisuusfunktioiden ilmeinen ongelma on, etteivät ne pysty käsittelemään muuta kuin binääristä dataa. Jos reaalimaailman data on monitasoisempaa, joudutaan tekemään informaatiota hävittäviä muunnoksia, mikäli kyseisiä funktioita halutaan käyttää.

5.3. Yksinkertainen sovitusfunktio monitasoiselle nominaaliselle datalle

Edellä kuvattua MC-funktiota hieman yleistäen voidaan määritellä *sovitus* (overlap) -funktio [Wilson and Martinez, 1997] nominaaliselle attribuutille, joilla on useampi kuin yksi taso seuraavasti:

$$s\_overlap1_a(x,y) = \begin{cases} 1, \text{ jos } x_a = y_a \\ 0, \text{ muutoin} \end{cases}. \tag{5.11}$$

ja vastaavasti esimerkkiparille *x* ja *y*

$$s\_overlap1(x,y) = \frac{\sum_{a=1}^{m} s\_overlap1_a(x,y)}{m}, \tag{5.12}$$

missä *m* on attribuuttien lukumäärä.

Jos ei haluta laskea negatiivisia täsmäyksiä mukaan, voidaan Jaccardin kerrointa hieman yleistäen määritellä seuraavanlainen funktio [Gower, 1971] attribuutille *a*:

$$s\_overlap2_a(x,y) = \begin{cases} 1, \text{ jos } x_a = y_a \text{ ja } x_a \neq 0 \\ 0, \text{ muutoin} \end{cases} \tag{5.13}$$

ja esimerkkiparille *x* ja *y*

$$s\_overlap2(x,y) = \frac{\sum_{a=1}^{m} s\_overlap2_a(x,y)}{\sum_{a=1}^{m} \delta_a(x,y)}, \tag{5.14}$$

missä

$$\delta_a(x,y) = \begin{cases} 0, \text{ jos } x_a = y_a = 0 \\ 1, \text{ muutoin} \end{cases}. \tag{5.15}$$

Tietenkin vastaavanlaiset funktiot voidaan muodostaa myös mittaamaan etäisyyttä samankaltaisuuden asemesta. Esimerkiksi funktiota *s_overlap*1_a(*x*,*y*) vastaava etäisyysfunktio määritellään

$$d\_overlap_a(x,y) = \begin{cases} 0, \text{ jos } x_a = y_a \\ 1, \text{ muutoin} \end{cases}. \tag{5.16}$$

Joissain funktioissa (katso kohta 6.4) käytetään esimerkiksi nominaalisten attribuuttien määrittelyjoukon kokoa (eli mahdollisten arvojen lukumäärää) $domain_a$ nominaalisten attribuuttien ”standardoimiseksi” ($d\_overlap_a$ / $domain_a$). Toisin sanoen attribuutin painoarvo on sitä pienempi, mitä suurempi $domain_a$ on, eli mitä todennäköisempi on tilanne $x_a \neq y_a$.

Edellä mainitut sovitusfunktiot ottavat huomioon nominaalisen attribuutin intervallittomuuden, joten voisi olettaa, että ne onnistuisivat nominaalisten aineistojen luokittelussa euklidista etäisyysfunktiota paremmin. Ne ovat kuitenkin varsin yksinkertaisia, eivätkä käytännön testit osoita, että niillä saataisiin merkittävästi parempia tuloksella kuin euklidisella etäisyysfunktiolla (katso luvut 8 ja 9).

5.4. VDM-metriikka monitasoiselle nominaaliselle datalle

Edelliset funktiot ovat olettaneet, että jokainen attribuutti ja attribuutin arvo on etäisyyttä määritettäessä tasavertaisessa asemassa. *VDM*-funktiossa (*Value Difference Metric*) perusideana on, että attribuutin arvon merkitys luokittelussa riippuu siitä, kuinka voimakkaasti arvo rajoittaa sitä, mihin luokkaan esimerkki voi kuulua [Stanfill and Waltz, 1986]. Toisin sanoen ideana on määrittää jokaiselle attribuutin arvo-luokka-parille (x_a, c) painoarvo sen mukaan, kuinka todennäköistä on, että arvo x_a esiintyy luokassa c. Etäisyys on pienin mahdollinen (eli 0) silloin, kun jokaisessa luokassa on samassa suhteessa arvoja x_a ja y_a. Poikkeavaa edellisiin funktioihin nähden on se, että

VDM käyttää luokkainformaatiota hyväkseen etäisyyslaskennassa. Tästä seuraa, että VDM ei ole käyttökelpoinen muissa kuin sellaisissa tilanteissa, joissa opetusjoukon luokkatiedot tunnetaan etukäteen.

Käytetään alkuperäisestä VDM-metriikasta [Stanfill and Waltz, 1986] hieman yksinkertaistettua versiota [Wilson and Martinez, 1997], joka määritellään seuraavasti:

$$VDM_a(x, y) = \sum_{c=1}^{C} \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|^q = \sum_{c=1}^{C} \left| P_{a,x,c} - P_{a,y,c} \right|^q , \qquad (5.17)$$

missä

- $N_{a,x}$ on niiden esimerkkien lukumäärä opetusjoukossa *T*, joilla on sama attribuutin *a* arvo kuin esimerkillä *x*,
- $N_{a,x,c}$ on niiden esimerkkien lukumäärä opetusjoukossa *T*, joilla on sama attribuutin *a* arvo kuin esimerkillä *x* ja, jotka kuuluvat luokkaan *c*,
- *C* on luokkien lukumäärä,
- *q* on vakio, yleensä 1 tai 2, ja
- $P_{a,x,c}$ n ehdollinen todennäköisyys, että luokka on *c*, olettaen, että attribuutin *a* arvo on sama kuin esimerkillä *x*. Kuten nähtiin kaavasta 5.17, todennäköisyys $P_{a,x,c}$ määritellään

$$P_{a,x,c} = \frac{N_{a,x,c}}{N_{a,x}} , \qquad (5.18)$$

missä $N_{a,x}$ on summa $N_{a,x,c}$ yli kaikkien luokkien, eli

$$N_{a,x} = \sum_{c=1}^{C} N_{a,x,c} \qquad (5.19)$$

ja summa $P_{a,x,c}$ yli kaikkien luokkien on 1 määrätyille luokalle *a* ja esimerkille *x*. VDM-funktio täyttää metriikan vaatimukset [Juhola and Laurikkala, 2001], joten esimerkiksi $VDM_a(x, y) = 0$, kun $x_a = y_a$.

Edellä kuvattu VDM-algoritmi on yksinkertaistettu versio alkuperäisestä algoritmista [Stanfill and Waltz, 1986], jossa huomioidaan myös attribuuttien suhteellinen paino. HVDM-algoritmi, joka esitellään myöhemmin ei sisällä myöskään attribuuttipainoja. Sovelluskohtaisesti attribuuttipainot voidaan kuitenkin useimmissa tapauksissa lisätä toteutettavaan algoritmiin [Wilson and Martinez, 1997].

Ongelmaksi VDM-metriikassa muodostuvat tilanteet, joissa syöte-esimerkissä on opetusjoukossa esiintymätön arvo. Tällöin $N_{a,x,c}$ tulee olemaan 0 kaikille luokille *c* ja tästä seuraa, että myös $N_{a,x}$ on 0, eikä arvoa $P_{a,x,c}$ ole määritelty. Attribuuteille on mah-

dotonta määrittää todennäköisyyttä tässä tilanteessa, joten todennäköisyydeksi tulee valita jokin kiinteä arvo. Perusteltuja vaihtoehtoja täksi arvoksi ovat joko 0 tai 1/*C* [Wilson and Martinez, 1997].

Jos VDM-metriikkaa käytetään laskemaan etäisyyttä kvantitatiivisille attribuuteille, on todennäköistä, että päädytään edellä kuvattuun tilanteeseen. Kun kvantitatiivisen attribuutin arvoalue on laaja, niin on todennäköistä, että kaikki sen arvot opetusjoukossa ovat uniikkeja. Tällöin arvon x_a etäisyys mistä tahansa muusta arvosta on 1 ja etäisyys itsestään on 0. Näistä syistä VDM-metriikka on sopimaton käsittelemään kvantitatiivisia ja varsinkaan jatkuvia attribuutteja.

Eräs ratkaisu tähän ongelmaan on *kvantisoida* eli *diskretisoida* kvantitatiiviset attribuutit [Skubacz and Hollmén, 2000]. Diskretisoinnissa kvantitatiiviselle attribuutille määritellään diskreetit arvoalueet, ja näin kvantitatiivisesta attribuutista saadaan muunnettua ordinaalinen tai nominaalinen. Jos esimerkiksi tiedetään, että diskreetti kvantitatiivinen attribuutti *a* voi saada arvoja alueelta [0,999], voidaan muodostaa neljä ryhmää A([0,249]), B([250,499]), C([500,749]) ja D([750,999]). Ryhmän nimen jälkeen suluissa on se arvoalue, johon kyseiseen ryhmään kuuluva arvo sijoittuu (eli esimerkiksi kuuluakseen ryhmään A tulee arvon x_a täyttää ehto $0 \leq x_a \leq 249$. Tämän pohjalta voidaan määritellä uusi attribuutti *a*', jonka arvot saadaan koodaamalla ne edellä mainitulla tavalla attribuutin *a* arvoista. Diskretisoinnissa menetetään kuitenkin arvokasta tietoa, joka on kvantitatiiviselle attribuutille ominaista. Jotta jokaisen attribuuttityypin ominaisuudet tulisivat hyödynnetyksi, Wilson ja Martinez [1997] ovat kehittäneet joukon metriikoita heterogeeniselle datalle, joiden perustana on VDM, ja joissa kvantitatiivisten attribuuttien käsittelyyn on kiinnitetty aiempaa enemmän huomiota. Näistä HVDM-funktiota käsitellään kohdassa 6.5.

5.5. Kosinietäisyysfunktio

Kosinietäisyysfunktiota on käytetty etäisyyden laskemiseen erityisesti tiedonhaussa [Salton, 1989]. Kosinietäisyysfunktio poikkeaa varsin paljon muista tässä työssä mainituista funktioista, koska siinä etäisyys määritellään vektoreiden välisen kulman kosinin [Puntanen, 1998] avulla. Kun attribuutteja on *m* kappaletta, määritellään vektoreiden (esimerkkien) *x* ja *y* välisen kulman kosini

$$\cos(x, y) = \frac{\sum_{a=1}^{m} x_a y_a}{\sqrt{\sum_{a=1}^{m} x_a^{\,2} \cdot \sum_{a=1}^{m} y_a^{\,2}}}. \qquad (5.20)$$

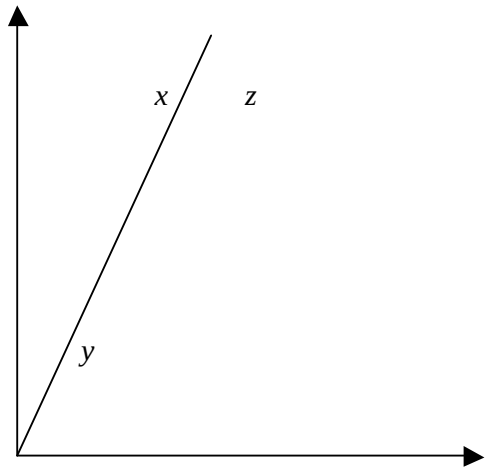
Määritellään kosinietäisyysfunktio seuraavasti [Everitt et al., 2001]:

$$d\_cos(x, y) = 1 - |\cos(x, y)|. \qquad (5.21)$$

Kulman kosinilla on muun muassa seuraavat ominaisuudet:

- $\cos(x, y) = 1$, jos on olemassa sellainen luku $\lambda > 0$, että $x = \lambda y$ ja
- $\cos(x, y) = -1$, jos on olemassa sellainen luku $\mu < 0$, että $x = \mu y$.

Kosinietäisyysfunktio katsoo siis monikerrat samankaltaisiksi ja tästä syystä funktio virittää hyvin erilaisen etäisyysmatriisin muihin tässä esitettyihin funktioihin verrattuna. Tarkastellaan tilannetta kuvan 5.2 avulla, jossa kyseessä on kaksiulotteinen olioavaruus ja siinä esimerkit *x, y* ja *z*. Vaikuttaisi, että esimerkit *x* ja *z* ovat lähimpänä toisiaan. Kuitenkin esimerkkivektoreiden *x* ja *y* välinen kulma on lähellä nollaa ja pienempi kuin esimerkkien *x* ja *z* välinen, ja siis kosinietäisyysfunktion perusteella $d(x, y) < d(x, z)$. Kosinietäisyysfunktio virittääkin usein reaalimaailmaa vastaamattoman etäisyysmatriisin, joten sen käyttäminen ei ole aina perusteltua. Edellä mainittujen ominaisuuksiensa vuoksi kosinietäisyysfunktio on käyttökelpoinen tehtävissä, joissa on tavoitteena löytää *muodoltaan* samankaltaisia esimerkkejä. Kaksi esimerkkiä ovat samanmuotoiset silloin, kun attribuuttien suhteet esimerkkien sisällä ovat samankaltaiset. Esimerkiksi, jos *x*' = (2,4) ja *y*' = (8,16) niin $d\_cos(x, y) = 0$.



Kuva 5.2. Kaksiulotteinen olioavaruus esitettynä euklidisella tasolla. Euklidisessa avaruudessa esimerkit *x* ja *z* ovat selvästi lähinnä toisiaan, mutta kosinietäisyysfunktion mukaan esimerkit *x* ja *y* ovat toisilleen lähimmät.

6. Heterogeenisiä etäisyysfunktioita

Kohdassa 2.2 tarkasteltiin attribuuttien jaottelua ja eri attribuuttityyppien välisiä eroja. Läheisyysfunktioita voidaan tarkastella sen mukaan, miten ne huomioivat eri mitta-asteikot aineistossa. Homogeeniset funktiot, joita käsiteltiin edellä, olettavat, että mitta-asteikko on sama jokaisella attribuutilla, kun taas heterogeeniset läheisyysfunktiot huomioivat eri mitta-asteikot. Muutkin tekijät kuin mitta-asteikko aiheuttavat ongelmia etäisyyslaskennassa. Attribuuttien toisistaan eroavat mittayksiköt saadaan yhdenmukaisiksi normalisoinnin avulla. Lisäksi on pohdittava jokaisen attribuutin painoarvoa eli sitä, kertooko jokin attribuutti olioiden välisestä erosta enemmän kuin jokin toinen attribuutti. Etäisyysfunktioihin voidaan toteutettaessa lisätä esimerkiksi kiinteät painoarvot tätä varten. Tässä työssä oletetaan kuitenkin, että jokaisella attribuutilla on sama painoarvo 1.

Seuraavassa esitellään joitain heterogeeniselle datalle soveltuvia etäisyysfunktioita. Kohdissa 6.1, 6.2, 6.3 ja 6.4 tarkasteltavat funktiot perustuvat nominaalisten attribuuttien osalta yksinkertaiseen sovitusfunktioon (katso kohta 5.3). Kohdassa 6.5 käsiteltävä HVDM puolestaan perustuu VDM-funktioon (katso kohta 5.4). Koska usean funktion nimelle ei ole vakiintunutta suomennosta, funktioista käytetään englanninkielisiä lyhenteitä.

6.1. Gowerin funktio

Gowerin esittämä samankaltaisuusfunktio [Gower, 1971] voidaan muuntaa etäisyysfunktioksi määritelmän 4.6 perusteella. Gowerin funktio on todistetusti metrinen [Gower, 1971], mikäli datassa ei ole puuttuvia arvoja tai dikotomisia attribuutteja. Nominaalisten attribuuttien etäisyys määritellään

$$s\_overlap_a(x,y) = \begin{cases} 1, & \text{jos } x_a = y_a \text{ ja } x_a \neq 0 \\ 0, & \text{muutoin} \end{cases}. \tag{6.1}$$

Tässä yhteydessä dikotomisella attribuutilla tarkoitetaan attribuuttia, jonka kohdalla negatiivista osumaa ei katsota samankaltaisuuden kannalta merkitseväksi (katso kohta 5.2). Jos molemmilta esimerkeiltä puuttuu jokin ominaisuus, Gowerin funktio ei tulkitse esimerkkejä samankaltaisiksi, vaan näissä tilanteissa kyseinen attribuutti jätetään huomioimatta etäisyyttä laskettaessa. Gowerin metriikka määrittelee kvantitatiivisen attribuutin *a* arvojen x_a ja y_a välisen samankaltaisuuden seuraavasti:

$$s\_diff(x_a, y_a) = \frac{1 - |x_a - y_a|}{vaihteluväli_a}. \tag{6.2}$$

Funktio samankaltaisuuden laskemiseen määritellään

$$gower_a(x,y) = \begin{cases} 0, & \text{jos } x_a \text{ tai } y_a \text{ tuntematon} \\ s\_overlap_a(x,y), & \text{jos } a \text{ nominaalinen} \\ s\_diff_a(x,y), & \text{jos } a \text{ kvantitatiivinen.} \end{cases} \quad (6.3)$$

Funktio kahden esimerkin kokonaisetäisyyden laskemiseksi, kun attribuuttien lukumäärä on *m*, määritellään kaavassa 6.4. Samankaltaisuus muutetaan etäisyydeksi määritelmän 4.6 perusteella ja näin saadaan seuraava funktio:

$$d\_gower(x,y) = 1 - \frac{\sum_{a=1}^{m} gower_a(x,y)}{\sum_{a=1}^{m} \delta_a(x,y)}. \quad (6.4)$$

Funktio $\delta_a(x,y)$ määritellään

$$\delta_a(x,y) = \begin{cases} 0, & \text{jos } x_a \text{ tai } y_a \text{ tuntematon tai, jos } x_a = y_a = 0 \\ 1, & \text{muutoin.} \end{cases} \quad (6.5)$$

Kokonaisetäisyydestä *d_gower*(*x*, *y*) voidaan laskea vielä neliöjuuri. Tämän operaation avulla laskettu etäisyysmatriisi on *euklidinen* [Everitt et al., 2001; Gower and Legendre, 1986] olettaen, että aineistossa ei ole puuttuvia tietoja.

Gowerin funktio poistaa tavallisten Minkowski-funktioiden ongelman, jossa nominaalisten attribuuttien arvoista lasketaan erotus. Toisaalta sen tapa käsitellä nominaalisia attribuutteja on vielä varsin yksinkertainen. Aineistoissa, jossa ei ole nominaalisia attribuutteja, Gowerin funktio toimii Manhattan-metriikan (katso kohta 5.1) tavoin.

6.2. HEOM-funktio

HEOM-funktio (Heterogeneous Euclidean-Overlap function) [Aha et al., 1991; Wilson and Martinez, 1997] muistuttaa paljon Gowerin funktiota. Funktio määrittelee etäisyyden attribuutin *a* arvojen x_a ja y_a välillä seuraavasti:

$$heom_a(x,y) = \begin{cases} 1, & \text{jos } x_a \text{ tai } y_a \text{ tuntematon,} \\ d\_overlap_a(x,y), & \text{jos } a \text{ on nominaalinen,} \\ rn\_diff_a(x,y) & \text{muutoin.} \end{cases} \quad (6.6)$$

Tuntemattomat attribuutin arvot käsitellään palauttaen etäisyysarvo 1, joka on useimmiten myös maksimietäisyys, jonka funktio voi palauttaa. Funktiot $d\_overlap_a$ ja rn_diff_a määritellään seuraavasti:

$$d\_overlap_a(x,y) = \begin{cases} 0, & \text{jos } x_a = y_a \\ 1, & \text{muutoin} \end{cases}, \tag{6.7}$$

$$rn\_diff_a(x,y) = \frac{|x_a - y_a|}{vaihteluväli_a}, \tag{6.8}$$

jossa $vaihteluväli_a = \max_a - \min_a$, jossa $\min_a$ ja $\max_a$ ovat minimi- ja maksimiarvot attribuutille *a* opetusjoukossa. Joskus puhutaan myös attribuutin *a määrittelyjoukon* (domain) koosta samassa merkityksessä. Määritellään $domain_a$ tässä työssä kuitenkin attribuutin *a* kaikkien mahdollisten arvojen lukumääräksi. Jatkuvalle attribuutille *a* määritellään tällöin $domain_a = \infty$.

Käyttäen yllä olevaa määritelmää etäisyysarvolle d_a saadaan arvo väliltä [0,1], mikäli kvantitatiivisen attribuutin arvo sijaitsee datasta lasketun arvoalueen sisällä. Tässä työssä käytetään luokittelussa ristiinvalidointia (katso luku 8), jolloin edellä kuvattu tilanne on mahdollinen ja näin ollen on myös mahdollista, että $d_a > 1$. Wilsonin ja Martinezin [1997] mukaan tämä ei kuitenkaan ole iso ongelma, sillä ensinnäkin kyseiset tilanteet ovat harvinaisia ja toiseksi tällaisissa tilanteissa voi olla myös hyväksyttävää, että etäisyysarvo on poikkeuksellisen suuri. Kokonaisetäisyys *heom*(*x*, *y*) esimerkeille *x* ja *y* on

$$heom(x,y) = \sqrt{\sum_{a=1}^{m} heom_a(x,y)^2}\ . \tag{6.9}$$

HEOM eroaa etäisyysmitaksi muutetusta Gowerin funktiosta vain vähän. Toisin kuin Gowerin funktiossa, ei HEOM-funktiossa oteta huomioon nominaalisten attribuuttien negatiivisten osumien painoarvoa. HEOM-funktiossa kokonaisetäisyys saadaan laskemalla neliöjuuri attribuuttien arvojen erotusten neliöiden summasta, joten tässä suhteessa se muistuttaa euklidista etäisyysfunktiota (katso kohta 5.1). Aiemmin todettiin Gowerin funktion muistuttavan Manhattan-funktiota.

6.3. Estabrook-Rogers –funktio

Estabrook ja Rogers [1966] määrittelevät samankaltaisuusfunktion, joka muistuttaa joiltain osin Gowerin funktiota. Käytetään tästä eteenpäin funktiosta nimeä ER-funktio ja muunnetaan samankaltaisuusfunktio etäisyysfunktioksi määritelmän 4.6 perusteella. Esimerkkien *x* ja *y* nominaaliselle attribuutille *a* voidaan käyttää kaavaa $s\_overlap_a$

(katso kohta 6.1) ja kvantitatiivisten attribuuttien kohdalla käytetään kaavaa $n\_diff_a(x, y)$.

$$n_a(x,y) = \begin{cases} 0, & \text{jos } x_a \text{ tai } y_a \text{ tuntematon} \\ s\_overlap_a(x,y), & \text{jos } a \text{ nominaalinen} \\ n\_diff_a(x,y), & \text{muutoin} \end{cases} \quad (6.10)$$

$$k_a(x,y) = |x_a - y_a|, \quad (6.11)$$

$$n\_diff_a(x,y) = \begin{cases} \dfrac{2u_a + 1 - k_a(x,y)}{2u_a + 2 + k_a(x,y)u_a}, & \text{jos } k_a(x,y) \le u_a \\ 0, & \text{muutoin.} \end{cases} \quad (6.12)$$

Muuttujalla u_a määritellään suurin mahdollinen kvantitatiivisten attribuuttien arvojen erotuksen itseisarvo. Tätä suuremmat erotukset tulkitaan siten, että attribuuttien välillä ei ole minkäänasteista samankaltaisuutta eli $n\_diff_a(x, y) = 1$. ER-funktio sopii siis tilanteisiin, jolloin voidaan ajatella jokin etäisyysraja, jolloin ei ole mielekästä enää puhua attribuuttien samankaltaisuudesta. Edellisillä funktioilla etäisyys $d_a(x, y) < 1$ aina, jos oletetaan, että attribuuttien arvojen erotus ei ole sama kuin attribuutin arvoalueen pituus eli $k_a < vaihteluväli_a$. Arvoksi u_a ei ole suositeltavaa valita suurempaa arvoa kuin $vaihteluväli_a - 2$ [Estabrook ja Rogers, 1966]. Mikäli tällainen arvo u_a valittaisiin, ei funktio antaisi koskaan etäisyysarvoa 1. Tämä heikentäisi funktion kykyä erotella esimerkkejä toisistaan [Estabrook and Rogers, 1966].

Kun attribuutteja on m kappaletta, kahden esimerkin x ja y kokonaisetäisyys määritellään seuraavasti:

$$er(x,y) = 1 - \frac{\sum_{a=1}^{m} n_a(x,y)}{m}. \quad (6.13)$$

ER-funktiolla on seuraavia ominaisuuksia.

Määritelmä 6.1.

1. Kun $k_a = 0$, $n_a(x,y) = 1$.
2. $n_a(x,y)$ vähenee, kun k_a vähenee u_a:n pysyessä vakiona.
3. $n_a(x,y)$ kasvaa, kun k_a kasvaa u_a:n pysyessä vakiona.
4. $n_a(x,y) = 0$, kun $d_a > k_a$.

ER-funktio muistuttaa edellä käsiteltyjä Gower- ja HEOM-funktiota. Nominaaliset attribuutit käsitellään kaikissa kolmessa funktiossa samalla tavoin käyttäen yksinkertaista sovitusfunktiota. Suurin ero on siinä, että Gower- ja HEOM-funktio toimivat kvantitatiivisten attribuuttien osalta kuten Minkowski-metriikat (normalisoinnilla varustettuna). Kuten kaavasta 6.12 nähdään, ER-funktiossa kvantitatiiviset attribuutit käsitellään hieman eri tavalla. Tarkastellaan miten tämä vaikuttaa kokonaisetäisyyteen. Lisätään funktioihin d_1 (Manhattan) ja d_2 (euklidinen) normalisointi ja jakaminen attribuuttien lukumäärällä, jolloin näidenkin funktioiden etäisyydet saadaan skaalattua välille [0,1]. Oletetaan, että

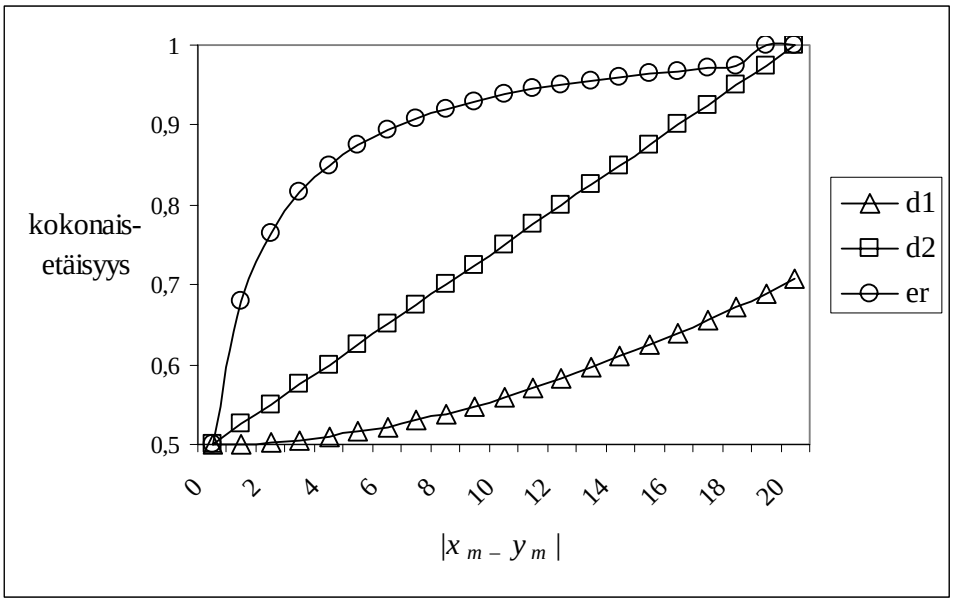
$$\sum_{a=1}^{m-1} d_a(x,y) = \sum_{a=1}^{m-1} d_a(x,y)^2 = \sum_{a=1}^{m-1} n\_diff_a(x,y) = 1, \text{ ja että} \tag{6.14}$$

$$d_1(x,y) = \frac{\sum_{a=1}^{m-1} d_a(x,y) + d_m(x,y)}{m}, \tag{6.15}$$

$$d_2(x,y) = \sqrt{\frac{\sum_{a=1}^{m-1} d_a(x,y)^2 + d_m(x,y)^2}{m}}, \text{ sekä} \tag{6.16}$$

$$er(x,y) = 1 - \frac{\sum_{a=1}^{m-1} n\_diff_a(x,y) + n\_diff_m(x,y)}{m}, \tag{6.17}$$

missä m on attribuuttien lukumäärä ja kaikki attribuutti ovat kvantitatiivisia. Kuvasta 6.1 nähdään miten kokonaisetäisyydet d_1, d_2 ja er muuttuvat, kun $|x_m - y_m|$ kasvaa. Euklidinen etäisyys kasvaa selvästi hitaammin kuin Manhattan-etäisyys. ER-funktiolle puolestaan on ominaista, että pieniä etäisyyksiä tulee vähän ja suuria paljon.



Kuva 6.1. Euklidisen etäisyys- (d1), Manhattan- (d2), ja ER-funktion (er) etäisyysjakauman tarkastelua.

ER-funktion alkuperäiset sovellukset liittyvät *biologiseen luokitteluun* eli *taksonomiaan* [Estabrook and Rogers, 1966]. Funktiosta on kehitetty myös *ekologiseen luokitteluun* paremmin sopiva versio, jossa puuttuvien tietojen käsittely on hieman toisenlainen [Legendre and Chodorowski, 1977].

6.4. Karteesiseen avaruusmalliin perustuva CSM-funktio

Edellä on kuvattu metriikoita ja funktioita, jotka sopivat datoille, joissa attribuuttien arvot ovat ainoastaan datapisteitä. Attribuuttien arvot voivat olla myös esimerkiksi intervalleja tai joukkoja. Tällaista dataa varten Ichino ja Yaguchi [1994] ovat kehittäneet metriikan, jota kuitenkin voidaan soveltaa myös pelkästään datapisteitä sisältäviin aineistoihin, jollaisia tässä työssä käsiteltävät aineistot kaikki ovat. Funktio perustuu matemaattiseen malliin, jota nimitetään *karteesiseksi avaruusmalliksi* (Cartesian Space Model, CSM). Jatkossa Ichinon ja Yaguchin metriikkaa kutsutaan CSM-funktioksi.

Kohdassa 2.1 tarkasteltiin tapaa esittää havaintomateriaali matriisina, jonka alkioina on mittaustulokset kullekin attribuutille. Edellä kuvatut metriikat ja funktiot ovat kaikki olettaneet, että nämä mittaustulokset eli attribuuttien arvot ovat datapisteitä eli kokonais- tai reaalilukuja. Attribuuttien arvot voivat yhtä hyvin olla suljettuja intervalleja tai äärellisiä joukkoja. Esimerkiksi jos tarkkaa mittaustulosta ei saada, voidaan arvioida tehtyjen mittausten perusteella suljettu väli johon oikea tulos todennäköisesti sijoittuu. Intervallityyppiset attribuutin arvot merkitään $x_a = [i, j]$, mikä tarkoittaa että

$i \le x_a \le j$. Nominaaliset attribuutit voidaan vastaavasti esittää pisteiden sijasta joukkoina. Joukkotyyppiset attribuutin arvot merkitään $x_a = \{a_1, a_2, ..., a_n\}$.

CSM-malli määritetään kahden operaattorin, karteesisen liitoksen (Cartesian Join, CJ) ja karteesisen yhdisteen (Cartesian Meet, CM) avulla. Määritellään attribuutin arvo x_a *tapahtumaksi* (event). CJ- ja CM-operaatioiden tulos on myös tapahtuma.

CJ lasketaan eri attribuuttityypeille eri tavoilla. Jos attribuutti *a* on kvantitatiivinen, niin *a*:ta vastaavien tapahtumien x_a ja y_a CJ määritellään suljettuna intervallina

$$x_a \oplus y_a = [\min(x_{aL}, y_{aL}), \max(x_{aU}, y_{aU})], \tag{6.18}$$

missä x_{aL} ja x_{aU} ovat suljetun intervallin x_a ala- ja yläraja sekä min(a_1, a_2, …, a_n) ja max(a_1, a_2, …, a_n) operaattoreita, jotka palauttavat minimin ja maksimin arvoista a_1, a_2,…,a_n. Jos attribuutti *a* on nominaalinen määritellään CJ joukkojen x_a ja y_a unionina

$$x_a \oplus y_a = x_a \cup y_a. \tag{6.19}$$

Vastaavasti myös CM määritellään eri attribuuttityypeille eri tavoilla. Jos *a* on kvantitatiivinen, CM määritellään

$$x_a \otimes y_a = \begin{cases} \text{Ø}, & \text{jos } x_a \text{ ja } y_a \text{ erillisiä} \\ [\max(x_{aL}, y_{aL}), \min(x_{aU}, y_{aU})], & \text{muutoin.} \end{cases} \tag{6.20}$$

Jos *a* on nominaalinen, CM määritellään

$$x_a \otimes y_a = x_a \cap y_a. \tag{6.21}$$

Kahdelle attribuutin arvolle x_a ja y_a määritellään

$$\phi_a(x, y) = |x_a \oplus y_a| - |x_a \otimes y_a| + \gamma(2|x_a \otimes y_a| - |x_a| - |y_a|), \tag{6.22}$$

missä $0 \le \gamma \le 0{,}5$, ja missä $|x_a|$ tarkoittaa intervallin x_a pituutta, jos attribuutti *a* on jatkuva kvantitatiivinen, tai joukon x_a mahdollisten arvojen lukumäärää, jos attribuutti *a* on diskreetti kvantitatiivinen, ordinaalinen tai nominaalinen. Tyhjälle joukolle $|\varnothing| = 0$.

Parametrilla γ kontrolloidaan sitä, korostuuko sisä- vai ulkopuolinen etäisyys. Jos $\gamma = 0$, niin saadaan $\phi(x_a, y_a) = |x_a \oplus y_a| - |x_a \otimes y_a|$. Tällöin $\phi(x_a, y_a)$ kertoo vain ulkopuolisesta etäisyydestä, jos tapahtumat x_a ja y_a ovat erillisiä. Esimerkiksi, kun $x_1 = [0,3]$, $y_1 = [12,15]$, $x_2 = [0,5]$ ja $y_2 = [10,15]$, niin $\phi(x_1, y_1) = \phi(x_2, y_2)$.

Jos $\gamma = 0.5$, niin saadaan $\phi(x_a, y_a) = |x_a \oplus y_a| - (|x_a - y_a|)/2$. Tällöin huomioidaan myös sisäpuolinen etäisyys. Näin ollen yllä mainituille erillisille tapahtumille x_1, y_1, x_2 ja y_2 pätee $\phi(x_1, y_1) > \phi(x_2, y_2)$.

Esimerkkien *x* ja *y* välinen kokonaisetäisyys, kun attribuutteja on *m* kappaletta, määritellään

$$csm(x, y) = \left[\sum_{a=1}^{m} \psi_a(x, y)^p \right]^p \quad . \tag{6.23}$$

Luvulla *p* määrätään etäisyysjakauman muoto. Funktiota $\phi(x_a, y_a)$ ei käytetä sellaisenaan, vaan siitä johdetaan normalisoitu versio $\psi(x_a, y_a)$, joka määritellään

$$\psi_a(x, y) = \phi_a(x, y)/U_a \,, \tag{6.24}$$

Jatkuvien kvantitatiivisten attribuuttien kohdalla U_a tarkoittaa vaihteluväliä. Diskreettien kvantitatiivisten ja kvalitatiivisten attribuuttien kohdalla U_a tarkoittaa mahdollisten arvojen lukumäärää.

Kun CSM-metriikkaa sovelletaan vain datapisteitä sisältäville aineistoille, toimii se hyvin samankaltaisesti kuin aiemmin käsitellyt heterogeeniset funktiot. Pelkästään datapisteitä sisältävässä aineistossa x_a ja y_a ovat aina erillisiä, jos $x_a \neq y_a$. Tällöin ei ole mielekästä puhua sisä- ja ulkopuolisista etäisyyksistä, koska ne ovat käytännössä aina samat. Seuraavassa $|x_a - y_a|$ tarkoittaa erotuksen $x_a - y_a$ itseisarvoa (toisin kuin edellä kaavassa 6.22).

Datapisteaineistoissa nominaalisille attribuuteille

$$\psi_a(x, y) = \begin{cases} 0, & \text{kun } x_a = y_a \\ (2 - 2\gamma)/domain_a \,, & \text{muutoin.} \end{cases} \tag{6.25}$$

Parametrista γ riippuen siis $1/domain_a \leq \psi_a(x, y) \leq 2/domain_a$, kun $x_a \neq y_a$. Kokonaisetäisyyden kannalta nominaalisen attribuutin merkitys on sitä pienempi, mitä suurempi U_a on. Tätä ominaisuutta aiemmin käsitellyillä heterogeenisillä funktioilla ei ole. Tällainen menettely voi olla joissain tilanteissa hyödyllinen. Jos nominaalisella attribuutilla on paljon mahdollisia arvoja ja hajonta on suurta, on todennäköistä, että usein $\psi_a(x, y) = 1$, mikä voi korostaa liikaa attribuutin merkitystä luokittelussa. Normalisoimalla tulos määrittelyjoukon koolla saadaan pienempiä kokonaisetäisyyksiä ja nominaalisten attribuuttien merkitys kokonaisetäisyyttä laskettaessa vähenee.

Diskreeteille kvantitatiivisille ja ordinaalisille attribuuteille

$$\psi_a(x,y) = \begin{cases} 0, & \text{kun } x_a = y_a \\ \left(\left|x_a - y_a + 1\right| + \gamma(-2)\right) / domain_a, & \text{muutoin.} \end{cases} \quad (6.26)$$

Ordinaalisten ja diskreettien kvantitatiivisten attribuuttien käsittely eroaa vain hieman aiemmista funktioista. Parametrilla γ voidaan määritellä pienin ja suurin mahdollinen etäisyys silloin kun $x_a \neq y_a$. Etäisyys on pienin mahdollinen silloin, kun $|x_a - y_a| = 1$ ja suurin silloin, kun $|x_a - y_a| = vaihteluväli_a$. Jos $\gamma = 0{,}5$, niin pienin mahdollinen etäisyys on 1 / $domain_a$ ja suurin mahdollinen etäisyys $|x_a - y_a|$ / $domain_a$. Tällöin suurin mahdollinen etäisyys ei voi olla koskaan 1. Jos $\gamma = 0$, niin pienin mahdollinen etäisyys on 2 / $domain_a$ ja suurin mahdollinen etäisyys $|x_a - y_a + 1|$ / $domain_a$.

Jatkuville kvantitatiivisille attribuuteille

$$\psi_a(x,y) = \begin{cases} 0, & \text{kun } x_a = y_a \\ \left(\left|x_a - y_a\right|\right) / vaihteluväli_a, & \text{muutoin,} \end{cases} \quad (6.27)$$

joten jatkuvat kvantitatiiviset attribuutit käsitellään kaikissa tilanteissa samalla tavalla kuin HEOM-funktiossa.

Datapisteaineistoissa CSM-metriikka on siis hyvin paljon HEOM- ja Gowerin funktion kaltainen. Parametrilla γ voidaan kontrolloida funktion antamia tuloksia hieman, mutta käytännössä datapisteaineistoissa γ on merkityksetön, koska tällaisissa aineistoissa ei ole mielekästä puhua sisä- ja ulkopuolisista etäisyyksistä. Käytännössä suurin ero on siinä, että CSM-funktio ”normalisoi” nominaaliset attribuutit niiden mahdollisten arvojen lukumäärällä. Lukujen 8 ja 9 luokittelutehtävissä käytettävässä CSM-metriikassa $\gamma = 0{,}5$.

6.5. VDM-funktioon perustuva heterogeeninen HVDM-funktio

Edellä havaittiin, että euklidinen etäisyys on sopimaton nominaalisten attribuuttien käsittelyyn ja toisaalta VDM ei erityisen hyvin sovellu kvantitatiivisten attribuuttien käsittelyyn. Ongelman ratkaisemiseksi Wilson ja Martinez [1997] ovat kehittäneet HVDM-metriikan (Heterogeneous Value Difference Metric), joka yhdistää euklidisen metriikan ja VDM-metriikan ideat heterogeeniseksi etäisyysfunktioksi. HVDM-funktio on huomattu käyttökelpoiseksi esimerkiksi lääketieteellisissä sovelluksissa [Montani et al., 2000]. HVDM määritellään

$$hvdm(x,y) = \sqrt{\sum_{a=1}^{m} hvdm_a(x,y)^2}, \quad (6.28)$$

missä m on attribuuttien lukumäärä. Funktio $hvdm_a(x,y)$ palauttaa esimerkkien x ja y attribuutin a arvojen välisen etäisyyden, ja määritellään

$$hvdm_a(x,y) = \begin{cases} 1, & \text{jos } x_a \text{ tai } y_a \text{ on tuntematon,} \\ norm\_vdm_a, & \text{jos } a \text{ on nominaalinen,} \\ norm\_diff_a, & \text{jos } a \text{ on kvantitatiivinen.} \end{cases} \quad (6.29)$$

Funktio $hvdm_a(x,y)$ käyttää seuraavaa funktiota silloin, kun kyseessä on kvantitatiivinen attribuutti:

$$norm\_diff_a(x,y) = \frac{|x_a - y_a|}{4\sigma_a}, \quad (6.30)$$

missä σ_a on attribuutin a arvojen keskihajonta.

Attribuuttien normalisointihan on tarpeellista silloin, kun halutaan, että jokaisella attribuutilla on sama vaikutus etäisyyttä laskettaessa. Esimerkiksi HEOM-metriikassa kvantitatiiviset attribuutit normalisoitiin attribuutin vaihteluvälillä. Ongelmia esiintyy silloin kun syöte-esimerkissä on arvo, joka ei esiinny opetusjoukossa. Tällöin etäisyysarvo voi karata suuremmaksi kuin 1. Toinen ongelma vaihteluvälillä normalisoitaessa on mahdollinen *poikkeava arvo* (outlier) opetusjoukossa. Esimerkiksi, jos attribuutti saa arvoja suurimmaksi osaksi alueelta [1, 100], mutta joukkoon sisältyy yksi poikkeava arvo 1000, niin HEOM-etäisyydeksi tulee lähes aina arvo, joka on alle 0,1. Tästä syystä robustimpi normalisoimistapa on käyttää keskihajontaa. Toisaalta, koska nominaalisella attribuutilla etäisyys voi olla maksimissaan 1, tulisi myös kvantitatiivisen attribuutin maksimietäisyyden olla suunnilleen vastaava, etteivät kvantitatiiviset attribuutit dominoisi etäisyyslaskennassa. HVDM-metriikassa kvantitatiiviset attribuutit normalisoidaan neljällä keskihajonnalla, jolloin normaalisti jakautuneessa datassa 95 % prosenttia arvoista sijoittuu keskiarvon molemmille puolille kahden keskihajonnan sisälle [Wilson and Martinez, 1997].

Funktioksi $norm\_vdm_a$ on esitetty kolme vaihtoehtoa, jotka määritellään seuraavasti [Wilson and Martinez, 1997]:

$$norm\_vdm1_a(x,y) = \sum_{c=1}^{C} \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|, \quad (6.31)$$

$$norm\_vdm2_a(x,y) = \sqrt{\sum_{c=1}^{C} \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|^2} \text{ , ja} \quad (6.32)$$

$$norm\_vdm3_a(x,y) = \sqrt{C \cdot \sum_{c=1}^{C} \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|^2} \,. \qquad (6.33)$$

Funktioiden *norm_vdm*1 ja *norm_vdm*2 ero on vastaavanlainen kuin on Manhattan- ja euklidisen metriikan (katso kohta 5.1) välillä. Funktion *norm_vdm*2 tuottamien etäisyyksien jakauma on muodoltaan sellainen, että hyvin suuria tai vastaavasti hyvin pieniä etäisyyksiä tulee vähemmän suhteessa keskisuuriin etäisyyksiin, mitä ominaisuutta funktiolla *norm_vdm*1 ei ole. Funktio *norm_vdm*3 lisää funktioon *norm_vdm*2 vielä luokkien määrällä kertomisen. Funktio *norm_vdm*2 on hypoteettisesti robustimpi kuin *norm_vdm*1 ja *norm_vdm*3 [Wilson and Martinez, 1997] ja myös tässä työssä käytetään jatkossa funktiota *norm_vdm*2.

HVDM näyttäisi sisältävän hyvin luontevan käsittelyn nominaalisille ja kvantitatiivisille attribuuteille. Alun perin ordinaaliset attribuutit on käsitelty kvantitatiivisina. Tämä ei kuitenkaan ole ongelmatonta, koska ordinaaliset attribuutit eivät sisällä välimatkatietoa. Käytetty euklidisen etäisyyden keskihajonnalla standardoitu versio kuitenkin nimenomaan vaatisi vähintään sen, että muuttujan peräkkäisten arvojen erotus on vakio kaikilla muuttujan arvoilla. Näin ollen ordinaaliset attribuutit on koodattava tavalla, että etäisyysarvoista saataisiin mielekkäitä. Joskus ei ole kuitenkaan järkevää käyttää euklidista etäisyyttä ordinaalisiin attribuutteihin, vaan käsitellä ne nominaalisten muuttujien tavoin. Tästä johtuen testeissä (katso luku 8) on kokeiltu HVDM-funktiosta kahta eri versiota, joista toisessa ordinaaliset attribuutit käsitellään kvantitatiivisten attribuuttien (HVDM) ja toisessa nominaalisten attribuuttien tavoin (MHVDM). Luvussa 8 esitellään saatuja tuloksia, jotka osoittavat, että ordinaalisten käsittely nominaalisten tavoin antaa harvoin huonompia tuloksia kuin alkuperäinen käsittelytapa. Joidenkin aineistojen kohdalla erot ovat selviä MHVDM-funktion hyväksi.

7. Luokittelutulosten arviointi

Keskeinen osa tiedonlouhintaprosessia on tulosten evaluointi. Se on tärkeää erityisesti lääketieteellisissä asiantuntijajärjestelmissä, joissa järjestelmän luotettavuus ja hyväksyttävyys ovat keskeisessä asemassa [Smith et al., 2003]. Sen lisäksi, että sovellusalueen asiantuntija tarkastelee saatuja tuloksia ja arvioi niiden laatua, voidaan käyttää erilaisia objektiivisia evaluointimenetelmiä [Lavrač, 1999]. Tämä tarkoittaa yleensä erilaisten tunnuslukujen laskemista. Tässä luvussa tarkastellaan joitain tavallisia menetelmiä ja tunnuslukuja, joiden avulla voidaan arvioida luokittelutuloksia.

Lähimmän naapurin luokittelijan kykyä voidaan arvioida esimerkiksi luokittelemalla opetusaineisto uudelleen ja vertailemalla luokittelijan ennustamia luokkaleimoja opetusaineiston todellisiin luokkiin. Käytettäessä ainoastaan opetusjoukkoa on kuitenkin vaarana, että *ylioppimisen* seurauksena saadaan harhaisia tuloksia. Usein opetusjoukko on vain otos perusjoukosta, koska perusjoukkoa ei ole saatavilla tai se on niin valtavan kokoinen, että laskentatehon rajallisuuden takia ollaan pakotettuja käyttämän vain pientä osaa siitä. Kun luokittelutietämys sisältää liikaa opetusjoukolle tyypillisiä, mutta perusjoukossa harvinaisia piirteitä, ovat pelkästään opetusjoukosta hankitut tulokset liian hyviä. Tästä syystä arviointiin käytetään erillistä testijoukkoa eli opetusjoukossa käyttämätöntä dataa. *Ristiinvalidointi* [Schaffer, 1993] on hyödyllinen menetelmä tähän tarkoitukseen. Siinä datajoukko on jaettu erillisiin osiin ja vuorotellen yksi näistä toimii testijoukkona ja loput opetusjoukkona. Termi *n-kertainen ristiinvalidointi* tarkoittaa, että data jaetaan *n* erilliseen osaan ristiinvalidointia varten. Joukoista yksi toimii vuorollaan testijoukkona ja muiden $n - 1$ joukon unioni opetusjoukkona. Yleensä kaikkien *n* validointiaskeleen tulokset kootaan yhteen.

Perustermi luokittelijaa arvioidessa on *luokitteluharha* (misclassification) eli tilanne, kun luokittelija luokittelee esimerkin virheellisesti. *Virhetaajuus* (error rate) tarkoittaa luokitteluvirheiden lukumäärää suhteessa kaikkiin luokittelutilanteisiin. Olkoon t_i testijoukon T esimerkin e_i todellinen luokka ja p_i luokittelijan ennuste. Määritellään virhetaajuus:

$$virhetaajuus = \frac{\sum_{i=1}^{n} \begin{cases} 1, \text{ jos } t_i \neq p_i \\ 0, \text{ jos } t_i = p_i \end{cases}}{n} \cdot 100\%, \qquad (7.1)$$

missä testijoukon koko $n = |T|$.

Vastaavasti *tarkkuus* (accuracy) = 100% – *virhetaajuus* eli oikein menneiden luokittelujen lukumäärä suhteessa kaikkiin luokittelutilanteisiin määritellään

$$tarkkuus = \frac{\sum_{i=1}^{n} \begin{cases} 0, \text{ jos } t_i \neq p_i \\ 1, \text{ jos } t_i = p_i \end{cases}}{n} \cdot 100\% \qquad (7.2)$$

Virhetaajuus ja tarkkuus ovat useissa tilanteissa kuitenkin riittämättömiä tunnuslukuja arvioitaessa luokittelijaa. Arvioinnissa joudutaan lisäksi ottamaan huomioon esimerkiksi luokkajakauma, eri virhetyypit ja luokitteluvirheen kustannus [Hollmén et al., 2000]. Jos aineistossa on useampia luokkia, voidaan virhetaajuus laskea jokaiselle luokalle erikseen.

Jos halutaan tutkia eri virhetyyppejä, voidaan luokittelun tulokset esittää *sekaannusmatriisin* (confusion matrix) avulla (katso taulukot 7.1 ja 7.2). Sekaannusmatriisin perusteella voidaan laskea erilaisia tunnuslukuja. Tunnuslukujen perustana on binäärinen luokitteluongelma eli tilanne, jolloin mahdollisia luokkia on kaksi. Usein käytetään lääketieteellisestä diagnosoinnista tuttuja käsitteitä positiivinen ja negatiivinen. Mahdollisia tilanteita on näin ollen neljä. Luokittelutapaus voi olla *oikea positiivinen* (True Positive, TP), *väärä positiivinen* (False Positive, FP), *oikea negatiivinen* (True Negative, TN) tai *väärä negatiivinen* (False Negative, FN). Yhteenveto binäärisestä luokittelusta voidaan esittää taulukon 7.1 kaltaisena sekaannusmatriisina.

Taulukko 7.1. Binäärinen sekaannusmatriisi.

	Todellinen luokka		
Ennustettu luokka	Positiivinen ($C+$)	Negatiivinen ($C-$)	
Positiivinen ($R+$)	Oikeat positiiviset (TP)	Väärät positiiviset (FP)	Positiiviset ennusteet
Negatiivinen ($R-$)	Väärät negatiiviset (FN)	Oikeat negatiiviset (TN)	Negatiiviset ennusteet
	Positiiviset esimerkit ($n+$)	Negatiiviset esimerkit ($n-$)	Kaikki esimerkit

Taulukon 7.1 lukuja käytetään harvoin sellaisenaan, vaan näiden pohjalta lasketaan suhdelukuja, joita on helpompi vertailla. Helpointa on laskea kaikkien oikeiden ennusteiden suhde kaikkiin luokittelutapauksiin. Tällöin saadaan jo äsken määritelty tarkkuus, joka nelikentän avulla määritellään seuraavasti:

$$\begin{aligned} tarkkuus &= \frac{TP + TN}{TP + FP + TN + FN} \cdot 100\% \\ &= \frac{TP + TN}{n} \cdot 100\%. \end{aligned} \qquad (7.3)$$

Luokittelutarkkuus ei ole kuitenkaan aina riittävä mitta, vaan usein tulee ottaa huomioon luokkajakauma ja eri virhetyypit. Jos aineistossa on hyvin epätasainen luokkajakauma luokittelutarkkuus kertoo ainoastaan suuremman luokan tarkkuudesta. Lisäksi erityisesti lääketieteellisessä diagnosoinnissa, jossa luokitellaan potilas jonkun taudin suhteen positiiviseksi tai negatiiviseksi, on merkityksellistä tutkia eri virhetyyppien esiintymistiheyksiä. On ilmeistä, että näissä tilanteissa on pyrittävä minimoimaan sekä väärien positiivisten, että väärien negatiivisten luokitusten lukumäärä. Eräs käytetty tunnusluku on *oikeiden positiivisten osuus* (True Positive Rate, TPR), joka tarkoittaa oikein menneiden positiivisten luokittelutilanteiden osuutta kaikista positiivisista esimerkeistä. Määritellään TPR seuraavasti:

$$TPR = \frac{TP}{TP + FN} \cdot 100\% \,. \tag{7.4}$$

Vastaavasti *oikeiden negatiivisten osuus* (True Negative Rate, TNR) määritellään

$$TNR = \frac{TN}{TN + FP} \cdot 100\% \,. \tag{7.5}$$

TPR ja TNR ovat eräänlaisia yleistermejä. Edellinen tarkoittaa siis tarkkuutta positiivisten esimerkkien joukossa ja jälkimmäinen tarkkuutta negatiivisten joukossa. Eri tieteenaloissa termeille on eri nimityksiä. Lääketieteessä käytetään termejä *sensitiivisyys* = *TPR* (sensitivity) ja *spesifisyys* = *TNR* (specificity). Tiedonhaussa TPR:ää ja sensitiivisyyttä vastaava termi on *muistettavuus* (recall). Tiedonhaussa käytetään myös termiä *tarkkuus* (precision), joka määritellään kaavasta 7.3 poiketen

$$tarkkuus_{TH} = \frac{TP}{TP + FP} \cdot 100\% \,, \tag{7.6}$$

joka tarkoittaa sitä, kuinka suuri osa positiivisista ennusteista on oikeasti positiivisia.

Usein joudutaan luokittelemaan aineistoja, joissa luokkia on enemmän kuin kaksi. Äsken tarkasteltuja tunnuslukuja voidaan silti käyttää myös useampiluokkaisissa aineistoissa. Taulukossa 7.2 esitetään sekaannusmatriisi aineistosta, jossa on C luokkaa. Taulukkoa tulkitaan siten, että r_{ii} tarkoittaa oikein luokiteltujen luokan c_i esimerkkien lukumäärää. Luokkaan c_j kuuluvien esimerkkien lukumäärä n_i aineistossa määritellään

$$n_j = \sum_{i=1}^{C} r_{ij} \,. \tag{7.7}$$

Taulukko 7.2. Sekaannusmatriisi aineistoille, joissa *n* luokkaa.

	Todellinen luokka			
Ennustettu luokka	c_1	c_2	…	c_n
c_1	r_{11}	r_{12}	…	r_{1n}
c_2	r_{21}	r_{22}	…	r_{2n}
…	…	…	…	…
c_n	r_{n1}	r_{n2}	…	r_{nn}

Binääriselle luokitteluongelmalle ajateltuja tunnuslukuja voidaan käyttää useamman luokan aineistoissa, koska useamman luokan luokitteluongelma voidaan ajatella sarjana binäärisiä ongelmia. Binääristen tunnuslukujen siirtäminen useamman luokan aineistoihin ei ole kuitenkaan täysin ongelmatonta, koska käsitteiden negatiivinen ja positiivinen käyttäminen on epäselvää tilanteessa, jossa aineisto ei ole jakautunut negatiiviseen ja positiiviseen luokkaan, vaan luokkia on useampia. Ongelmaa voidaan lähestyä tarkastelemalla yksittäistä luokittelutilannetta. Voidaan ajatella, että tällaisessa tilanteessa käsiteltävän esimerkin luokka on ”positiivinen”. Sen sijaan ”negatiivisen” luokan määritteleminen ei ole yksiselitteistä. Vaihtoehtoja ovat, että

1. muut luokat ovat kaikki omia joukkojaan, tai että
2. muut luokat muodostavat ”negatiivisen” luokan yhtenä joukkona.

Jos ennuste menee oikein, tulkitaan tilanne käsiteltävän esimerkin luokan kannalta oikeaksi positiiviseksi. Kaikkien muiden luokkien kannalta tilanne on oikea negatiivinen, mikä pätee negatiivisen luokan molemmissa määritelmissä. Jos ennuste on väärin, tulkitaan tilanne käsiteltävän esimerkin luokan kannalta vääräksi negatiiviseksi. Sen luokan kannalta, johon luokittelija esimerkin ennustaa kuuluvan, tilanne on väärä positiivinen. Tämän jälkeen on pohdittava, onko tilanne muiden kuin todellisen ja ennustetun luokan kannalta oikea negatiivinen. Alkuperäisessä binäärisessä tilanteessa oikea negatiivinen tarkoittaa tilannetta, jossa esimerkki luokitellaan oikein negatiiviseen luokkaan kuuluvaksi. Moniluokkaisessa aineistossa negatiivisen luokan ensimmäisen määrittelyn mukaan tässä tilanteessa ei tapahdu oikeaa negatiivista ennustetta. Sen sijaan jälkimmäisen määritelmän mukaan tilanne on kaikkien muiden paitsi todellisen ja ennustetun virheellisen luokan osalta oikea negatiivinen. Vaikka ennuste on virheellinen, kuuluu ennustettu luokka silti n – 2 luokan kannalta negatiiviseen luokkaan ja näiden osalta tilanne on näin ollen oikea negatiivinen. Käytännössä on järkevämpi käyttää ensimmäistä tapaa. Jos käytettäisiin jälkimmäistä tapaa, niin *TN*-frekvenssi kasvaisi erittäin suureksi, vaikka luokitteluvirheitä esiintyisi paljonkin, koska luokitteluvirheen sattuessakin *TN*-frekvenssiin summataan $n-2$.

Määritelmä 7.1. Moniluokkaisessa luokittelutilanteessa TN luokalle c_k on muiden luokkien TP-arvojen summa.

Taulukon 7.2 perusteella annetaan vielä määritelmät tunnusluvuille moniluokkaisessa ympäristössä. Aineistossa, jossa on *C* luokkaa voidaan määritellä *TP* luokan c_i esimerkkien joukossa

$$TP_i = r_{ii}, \tag{7.8}$$

missä TP_i tarkoittaa kaikkia tilanteita, joissa ennustettu luokka on sama kuin todellinen luokka c_i. Koko aineistoa vastaava *TP* määritellään

$$\begin{aligned} TP &= \sum_{i=1}^{C} TP_i \\ &= n - TN \end{aligned}, \tag{7.9}$$

joka tarkoittaa siis kaikkia luokittelutilanteita, joissa ennustettu luokka on sama kuin todellinen luokka. Luokittelutarkkuus moniluokkaiselle aineistoille voidaan määritellä

$$tarkkuus = \frac{\sum_{i=1}^{C} TP_i}{n} \cdot 100\%. \tag{7.10}$$

Vastaavasti *FP* luokan c_i esimerkkien joukossa määritellään

$$FP_i = \sum_{\substack{j=1 \\ i \neq j}}^{C} r_{ij}, \tag{7.11}$$

ja koko aineistossa

$$FP = \sum_{i=1}^{C} FP_i. \tag{7.12}$$

FP_i tarkoittaa siis tilanteita, joissa esimerkki on luokiteltu kuuluvaksi luokkaan c_i, mutta esimerkin todellinen luokka on joku muu kuin kyseinen luokka c_i.

Moniluokkaisessa ongelmassa *FN* määritellään luokan c_i esimerkkien joukossa

$$FN_j = \sum_{\substack{j=1 \\ i \neq j}}^{C} r_{ij} \tag{7.13}$$

ja kaikkien esimerkkien joukossa vastaavasti

$$FN = \sum_{i=1}^{C} FN_i \,. \tag{7.14}$$

TN luokan c_i sisällä on

$$TN_i = \sum_{\substack{j=1 \\ i \neq j}}^{C} r_{jj} \,, \tag{7.15}$$

ja kaikkien esimerkkien joukossa vastaavasti

$$\begin{aligned} TN &= \sum_{i=1}^{C} TN_i \,. \\ &= n - TP \end{aligned} \tag{7.16}$$

8. Etäisyysfunktioiden vertailu

Tutkielman tarkoituksena oli, paitsi tarkastella kirjallisuudessa esitettyjen etäisyysfunktioiden ominaisuuksia, niin myös vertailla funktioita käytännön luokitteluongelmien avulla. Suurimpana kiinnostuksen kohteena oli, miten funktiot luokittelevat heterogeenisiä aineistoja ja onko funktioiden välillä merkittäviä eroja. Tässä luvussa esitetään testien tulokset. Luku rakentuu siten, että kohdassa 8.1 kuvataan testiasetelma. Kohdassa 8.2 kuvataan lyhyesti testejä varten kirjoitettu ohjelma, ja tarkastellaan hieman funktioiden aika- ja muistivaatimuksia. Kohdassa 8.3 tarkastellaan testeissä käytettyjä aineistoja. Kohdassa 8.4 esitetään testien tulokset. Tuloksia pohditaan tarkemmin luvussa 9.

8.1. Testiasetelma

Funktioiden vertailun pohjaksi luokiteltiin joukko aineistoja etäisyyksien perusteella käyttäen kutakin funktiota vuorollaan etäisyysavaruuden virittäjänä. Luokittelijana käytettiin kohdassa 3.1 kuvattua *k*:n lähimmän naapurin menetelmää. Luokittelu ja ristiinvalidointi suoritettiin aluksi yhden, kolmen ja viiden lähimmän naapurin perusteella. Funktioiden välinen vertailu tehtiin kuitenkin vain kolmen lähimmän naapurin luokittelun perusteella, koska sillä saadut luokittelutulokset vaikuttivat keskimäärin parhailta. Tasapelitilanteissa valittiin käsittelyjärjestyksessä ensimmäisen esimerkin luokka. Lähimmän naapurin menetelmässä käytettiin luvuissa 5 ja 6 kuvattuja

- euklidista etäisyys-,
- kosinietäisyys-,
- Gowerin,
- HEOM-,
- ER-,
- CSM-,
- HVDM- ja
- MHVDM-funktiota.

Tunnusluvut laskettiin 10-kertaisen ristiinvalidoinnin perusteella. Jokainen aineisto jaettiin kymmeneen lähes yhtä suureen osajoukkoon C_1, C_2,…, C_{10}. Harva joukko oli jaollinen kymmenellä, joten käytännössä jakojäännös sijoitettiin mahdollisimman tasaisesti osajoukkoihin alkaen joukosta C_1. Näin ollen osajoukkojen koot poikkesivat toisistaan maksimissaan yhdellä, mikä on suhteellisen pieni ero, koska osajoukot olivat vähintään useamman kymmenen esimerkin kokoisia. Osajoukoista yksi toimi vuorollaan testijoukkona ja muiden yhdeksän joukon unioni opetusjoukkona. Jotkut aineistot sekoitettiin käyttäen yksinkertaista satunnaisotantaa, jotta jokaisessa osajoukossa kunkin luokan suhteellinen frekvenssi olisi suunnilleen sama kuin koko aineistossa. Joissa-

kin aineistoissa oli puuttuvia tietoja. Useimmiten puuttuvat tiedot imputoitiin luokan tai koko aineiston keskiluvuilla. Jos joltain esimerkiltä puuttui selvästi suurin osa tiedoista (yli 50%), jätettiin kyseinen esimerkki pois aineistosta. Samoin, jos jonkun attribuutin arvo puuttui selvästi suurimmalta osalta esimerkeistä, jätettiin kyseinen attribuutti testien ulkopuolelle.

Euklidinen etäisyys- ja kosinietäisyysfunktio otettiin mukaan testeihin, jotta nähtäisiin, miten funktiot, jotka eivät ota huomioon attribuuttien mitta-asteikkoja, selviytyvät heterogeenisen aineiston luokittelusta, ja eroavatko tarkkuudet merkittävästi heterogeenisillä funktioilla saaduista.

Euklidiseen etäisyys- ja kosinietäisyysfunktioon lisättiin standardointi, joka suoritettiin HVDM-funktion tapaan (katso kohta 6.5) neljällä keskihajonnalla. Vaikka kaikki käytetyt aineistot olivat joko valmiiksi täydellisiä tai täydennettyjä, lisättiin kaikkien funktioiden toteutukseen puuttuvien tietojen käsittely jatkotutkimuksia varten. Mikäli vähintään toinen attribuutin arvo puuttui, tulkittiin etäisyydeksi 1. Funktioissa käytetyt keskihajonta, vaihteluväli ja VDM-pohjaisissa funktioissa hyödynnettävät VDM-todennäköisyydet laskettiin jokaisessa ristiinvalidointiaskelmassa kulloisestakin opetusjoukosta.

Varsinaiset tilastolliset vertailut tehtiin käyttäen sekä kokonaistarkkuuksia että luokkien TPR-lukujen mediaania. Lisäksi testit tehtiin erikseen kaikille aineistoille ja heterogeenisille aineistoille. Heterogeeniseksi määriteltiin aineisto, joka oli kuvattu vähintään yhdellä monitasoisella nominaalisella attribuuteilla ja vähintään yhdellä ordinaalisella tai kvantitatiivisella attribuutilla. Näin ollen testejä suoritettiin neljä kappaletta (katso taulukko 8.1). Vertailussa käytettiin Friedmanin ja Wilcoxonin testejä. Wilcoxonin testien merkitsevyystasolle eli niin sanotulle *p*-arvolle tehtiin Bonferroni- ja Kounias-korjaukset testin epäparametrisen luonteen takia. Testausta käsitellään tarkemmin alakohdassa 8.4.2.

Taulukko 8.1. Suoritetut vertailut.

	Aineistot	Tunnusluku
Testi 1	Kaikki	*tarkkuus*
Testi 2	Heterogeeniset	*tarkkuus*
Testi 3	Kaikki	TPR_{MED}
Testi 4	Heterogeeniset	TPR_{MED}

8.2. Testiohjelmasta

Testejä varten kirjoitettiin Java-kielellä ohjelma, jonka suoritti lähimmän naapurin luokittelun ja ristiinvalidoinnin, sekä laski ja koosti luokittelutarkkuudet. Myös MatLab-ohjelmistoa ja sen skriptikielen käyttöä testiohjelman toteuttamiseksi harkittiin, mutta

Javaan päädyttiin erityisesti siksi, että luokittelualgoritmin suoritusajan huomattiin olevan merkittävästi pienempi Javalla toteutettuna. Friedmanin ja Wilcoxonin testeissä käytettiin SPSS-ohjelmistoa.

Suurin osa testeissä käytetyistä aineistoista (katso taulukko 8.3) oli suhteellisen pieniä korkeintaan muutaman tuhannen esimerkin aineistoja (ja tästä syystä epätyypillisiä tiedonlouhinta-aineistoja). Muutamissa aineistoissa esimerkkejä oli kuitenkin yli 10000 kappaletta, joten myös algoritmien tehokasta toteuttamista sekä laskenta-ajan että muistinkäytön kannalta jouduttiin pohtimaan. Euklidisen etäisyys-, HEOM ja HVDM-funktion talletustila- ja aikavaatimuksia ovat pohtineet Wilson ja Martinez [1997]. Heidän laskelmansa euklidisesta etäisyys- ja HEOM-funktiosta pätevät myös ER-, CSM-, ja Gowerin funktioon, koska niissäkin esimerkkiparin välisen etäisyyden laskemiseen kuluu vakioaika.

Tarkastellaan aluksi muiden paitsi VDM-pohjaisten funktioiden vaatimuksia. Tallennustilaa kaikki funktiot vievät suunnilleen O(*mn*), missä *m* on attribuuttien lukumäärä ja *n* esimerkkien lukumäärä. Myös opetusalgoritmin aikavaatimus on suunnilleen O(*mn*). Joissain funktioissa opetusvaiheessa lasketaan vaihteluväli, minkä aikavaatimus on O(*mn*), ja joissakin keskihajonta, minkä aikavaatimus on O(2*mn*) ≈ O(*mn*). Luokittelutilanteessa käydään läpi jokainen esimerkkipari opetus- ja testiaineistosta. Tämä vie aikaa suunnilleen O(*mn*).

HVDM-funktiossa talletustilaa tarvitaan aineiston vaatiman tilan O(*mn*) lisäksi O(*mvC*) VDM-todennäköisyyksiä $P_{a,x,c}$ varten. Tässä *v* tarkoittaa keskimääräistä määrittelyjoukon kokoa nominaalisilla attribuuteilla ja *C* luokkien lukumäärää. Opetusalgoritmi lukee aineiston ja laskee arvot $P_{a,x,c}$, joten se vaatii aikaa myös O(*mn* + *mvC*). Luokittelualgoritmin aikavaatimus on O(*mnC*), koska funktiossa $vdm_a(x, y)$ suoritetaan suunnilleen luokkien lukumäärän verran operaatioita. VDM-todennäköisyydet ovat tallennettuna hajautustaulussa, joten niiden haku pystytään tekemään vakioajassa. MHVDM-funktiossa talletustilaa vaaditaan O(*mn* + *muC*), missä *u* on keskimääräinen nominaalisten ja ordinaalisten attribuuttien määrittelyjoukkojen summa. Opetusalgoritmi vie aikaa suunnilleen O(*mn* + *muC*) ja luokittelualgoritmi HVDM-funktion tapaan suunnilleen O(*mnC*).

Taulukossa 8.2 on yhteenvetona algoritmien talletustila- ja aikavaatimukset. Luonnollisesti *r*-kertaisessa ristiinvalidoinnissa algoritmit joudutaan suorittamaan *r* kertaa. Käytännössä suurin osa testiaineistoista ristiinvalidoitiin muutamassa minuutissa. Erityisesti VDM-pohjaissa funktioissa suoritusaika kuitenkin kasvaa jyrkästi, kun opetusjoukon koko suurenee. Pahimmillaan ristiinvalidointi kestikin HVDM- ja MHVDM-funktioilla reilut 38 tuntia shakki-aineistolla, jossa oli noin 28000 esimerkkiä ja kuusi monitasoista nominaalista attribuuttia (katso taulukko 8.3).

Taulukko 8.2. Funktioiden talletustila- ja aikavaatimukset.

Funktio	Talletustila	Opetusalgoritmi	Luokittelualgoritmi
Kosini	O(mn)	O(mn)	O(mn)
CSM	O(mn)	O(mn)	O(mn)
ER	O(mn)	O(mn)	O(mn)
Euklidinen	O(mn)	O(mn)	O(mn)
Gower	O(mn)	O(mn)	O(mn)
HEOM	O(mn)	O(mn)	O(mn)
HVDM	O($mn + mvC$)	O($mn + mvC$)	O(mnC)
MHVDM	O($mn + muC$)	O($mn + muC$)	O(mnC)

8.3. Aineistoista

Aineistoja oli 36 kappaletta. Sopivia aineistoja ei ollut helppo löytää, ja erityisen vaikeaa oli löytää sopivia heterogeenisiä aineistoja. Aineistot kerättiin eri lähteistä, useimmat kuitenkin tunnetusta UCI-tietokannasta [Blake and Merz, 1998]. Lisäksi joukossa oli paljon lääketieteellisiä aineistoja. Kaikkien aineistojen yhteenveto on taulukossa 8.3 ja taulukossa 8.4 on aineistojen luokkafrekvenssit. Liitteessä 1 on jokaisesta aineistoista lyhyt yhteenveto, johon kuuluu lähdetietojen lisäksi aineiston ja sen luokitteluongelman kuvaus sekä kuvaus aineistolle mahdollisesti tehdyistä muunnoksista.

Kuten aiemmin on tullut jo esille, oli joissain aineistoissa alun perin puuttuvia tietoja. Vaikka läheisyysfunktioilla on mahdollista käsitellä myös puuttuvia tietoja, jätettiin kyseisen ominaisuuden tarkastelu tämän työn ulkopuolelle. Näin ollen puuttuvia tietoja sisältävät aineistot täydennettiin (katso kohta 2.3). Puuttuvien tietojen kohdalla oli kolme eri menettelytapaa. Ensin aineistoista poistettiin muutamia esimerkkejä, joilta puuttui merkittävä osa (suunnilleen yli 50 %) tiedoista. Toiseksi poistettiin sellaiset attribuutit, joilta puuttui paljon arvoja, tai joiden ei katsottu olevan merkittäviä luokittelun kannalta. Päätös attribuuttien poistamisesta tehtiin osin omasta toimesta ja joissain aineistoissa asiantuntijoiden tietämystä hyödyntäen. Kolmanneksi, mikäli aineistoissa oli puuttuvia tietoja suhteellisen pieni määrä, täydennettiin data imputoimalla. Imputointi tehtiin keskiluvuilla ja pääosin luokkien sisällä (muutamassa tapauksessa koko aineistossa). Aineistoja ei normalisoitu etukäteen, koska normalisointi sisältyi läheisyysfunktioihin.

Taulukosta 8.4 nähdään aineistojen luokkafrekvenssit. Luokittelutarkkuuksia (katso taulukot 8.5 ja 8.6) katsottaessa tulee huomioida, että mikäli aineistoissa on joku tai joitain selvästi muita yleisimpiä luokkia, kuvaa luokittelutarkkuus lähinnä luokittelutarkkuutta näiden yleisten luokkien sisällä. Tällöin se ei anna lainkaan kuvaa tarkkuudesta pienten luokkien sisällä. Tästä johtuen tarkasteltiin myös kunkin aineiston kohdalla luokkien sisäisiä TPR-arvoja. TPR-arvoista otettiin mediaani, joka antaa yleisku-

van funktion kyvystä luokitella sekä suurempia että pienempiä luokkia (katso taulukot 8.7 ja 8.8). Useimmat aineistoista ovat kaksiluokkaisia. Tällaisessa aineistossa TPR-mediaani on käytännössä kahden luokan TPR-arvojen keskiarvo. Näin ollen esimerkiksi alokkaat-aineiston kohdalla TPR-mediaani on selvästi heikompi kuin luokittelutarkkuus, koska toinen luokista on hyvin pieni ja tästä syystä vaikeasti luokiteltava. Lisäksi joissain useampiluokkaisissa aineistoissa pieniä luokkia on suhteellisen paljon (esimerkiksi hedelmäpeli- ja mahatautiaineisto). Näiden aineistojen kohdalla TPR-mediaani oli hyvin pieni (hedelmäpeliaineistossa joidenkin funktioiden kohdalla 0%), koska aineistojen marginaalisten luokkien TPR-arvot olivat lähimmän naapurin menetelmällä luokiteltaessa pieniä. Pienten luokkien luokittelemisen ongelmia on käsitellyt esimerkiksi [Laurikkala, 2001b].

Taulukko 8.3. Yhteenveto aineistoista.

Aineisto	Esimerkit	Luokat	Attribuutit	Nominaaliset		Ordinaaliset	Kvantitatiiviset	
				Binääriset	Monitasoiset		Diskreetit	Jatkuvat
Alokkaat	9004	2	39	39	0	0	0	0
Assistentit	151	3	5	2	2	0	1	0
Australia	690	2	14	4	4	0	6	0
Cleveland	303	2	13	8	0	0	4	1
Eturauhassyöpä	380	2	7	2	1	1	1	2
Guatemala	3334	4	3	2	1	0	0	0
Hedelmäpeli	345	5	4	1	3	0	0	0
Huimaus	914	10	38	11	1	10	11	5
ICU	200	2	19	14	1	1	3	0
Ihotauti	366	6	34	1	0	32	1	0
Inkontinenssi	529	5	14	8	0	0	6	0
Kööpenhamina	1681	2	3	0	1	2	0	0
Lasit	214	6	9	0	0	0	9	0
Led	1000	10	7	7	0	0	0	0
Led+kohina	1000	10	24	24	0	0	0	0
Liput	194	8	28	12	6	0	9	1
Ljubljana	286	2	9	3	1	5	0	0
Luotonhakijat	1000	2	20	2	6	5	7	0
Mahatauti	1333	13	16	9	4	1	1	1
Munkit1	432	2	6	2	4	0	0	0
Munkit2	432	2	6	2	4	0	0	0
Munkit3	432	2	6	2	4	0	0	0
Palkat	534	2	10	3	3	0	3	1
Plasebo	141	2	14	8	5	0	1	0
Profylaksi	166	2	14	8	5	0	1	0
Promoottorit	106	2	57	0	57	0	0	0
Ristinolla	958	2	9	0	9	0	0	0
Rytmihäiriö	452	2	278	73	0	0	205	0
Shakki	28056	18	6	0	6	0	0	0
Sillat	105	6	11	2	4	4	1	0
SPECT	267	2	22	22	0	0	0	0
Titanic	2201	2	3	2	1	0	0	0
Tyreoosi	3772	3	21	15	0	0	6	0
USA	30162	2	13	1	6	1	5	0
WDBC	569	2	30	0	0	0	0	30
WPBC	198	2	32	0	0	0	0	32

Taulukko 8.4. Aineistojen luokkafrekvenssit järjestettynä laskevaan järjestykseen.

Aineisto	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Alokkaat	8565	439																
Assistentit	52	50	49															
Australia	383	307																
Cleveland	164	139																
Eturauhassyöpä	227	153																
Guatemala	1883	799	519	133														
Hedelmäpeli	306	22	11	4	2													
Huimaus	313	146	130	120	65	41	39	31	16	13								
ICU	160	40																
Ihotauti	112	72	61	52	49	20												
Inkontinenssi	323	140	33	18	15													
Kööpenhamina	968	713																
Lasit	76	70	29	17	13	9												
Led	109	107	107	105	103	99	98	95	89	88								
Led+kohina	116	116	114	107	103	97	89	89	86	83								
Liput	69	47	40	17	9	9	2	1										
Ljubljana	201	85																
Luotonhakijat	700	300																
Mahatauti	614	270	125	63	59	54	50	22	22	19	15	11	9					
Munkit1	216	216																
Munkit2	290	142																
Munkit3	228	204																
Palkat	289	245																
Plasebo	79	62																
Profylaksi	122	44																
Promoottorit	53	53																
Ristinolla	626	332																
Rytmihäiriö	245	207																
Shakki	4553	4194	3597	2854	2796	2166	1985	1712	1433	683	592	471	390	246	198	81	78	27
Sillat	44	16	13	11	11	10												
SPECT	212	55																
Titanic	1490	711																
Tyreoosi	3488	191	93															
USA	22654	7508																
WDBC	357	212																
WPBC	170	28																

8.4. Tuloksista

Tässä kohdassa esitetään tulokset siten, että aluksi esitetään kolmen lähimmän naapurin menetelmällä ja ristiinvalidoimalla saadut luokittelutarkkuudet ja luokkien TPR-arvojen mediaanit. Sitten esitetään Friedmanin ja Wilcoxonin testien tulokset, joita siis tehtiin yhteensä neljä kappaletta (katso taulukko 8.1).

8.4.1. Luokittelutarkkuudet

Taulukossa 8.5 on esitetty luokittelutarkkuudet kaikille aineistoille. Jokaiselle funktiolle on laskettu tarkkuuksien keskiarvo ja mediaani. Keskiarvon ja mediaanin perusteella kosinietäisyysfunktion tarkkuus on selvästi muita huonompi. Keskiarvojen perusteella HVDM- ja MHVDM-funktiot näyttävät erottuvan hieman muista funktiosta, mutta niiden mediaanit eivät eroa juurikaan muista funktioista. Muiden kuin kosinietäisyysfunktioiden tarkkuuksien mediaanit sijoittuvat kaikki 78,88 ja 79,65 prosenttiyksikön väliin. Yksittäisissä aineistoissa suurimpia eroja funktioiden välillä esiintyy assistentit-, led+kohina-, promoottorit-, ristinolla-, ja shakki-aineistossa sekä kolmessa munkit-aineistoissa. Assistentit-aineistoa (sisältää yhden kvantitatiivisen attribuutin) lukuun ottamatta kaikki edellä mainitut aineistot sisältävät pelkästään nominaalisia attribuutteja. Pääsääntöisesti näiden kahdeksan aineiston joukossa parhaita tuloksia antavat HVDM- ja MHVDM-funktiot. Munkit-aineistojen kohdalla luokittelutarkkuudet vaihtelevat poikkeuksellisella tavalla. Näissä aineistoissa joku tai jotkut funktiot antavat yleiseen tasoon nähden selvästi heikompia tuloksia.

Taulukossa 8.6 on esitetty luokittelutarkkuudet heterogeenisille aineistoille. Erot funktioiden välillä ovat varsin pieniä. HVDM-, MHVDM- ja euklidinen etäisyysfunktio ovat kolmen tarkimman joukossa sekä keskiarvoja että mediaaneja tarkasteltaessa.

Taulukko 8.5. Luokittelutarkkuudet kaikille aineistoille (%).

Aineisto	Kosini	CSM	ER	Euklidinen	Gower	HEOM	HVDM	MHVDM
Alokkaat	95,10	95,10	95,10	95,10	95,10	95,10	95,10	95,10
Assistentit	49,67	35,76	40,40	49,01	40,40	40,40	51,66	51,66
Australia	82,46	83,62	84,64	82,32	85,22	86,09	84,06	84,06
Cleveland	82,51	81,52	78,88	83,17	81,85	83,50	81,52	81,52
Eturauhassyöpä	69,21	71,05	70,26	72,89	71,05	70,79	74,21	73,16
Guatemala	57,65	57,65	57,65	57,65	57,65	57,65	57,65	57,65
Hedelmäpeli	91,59	93,04	93,04	90,14	93,04	93,04	97,10	97,10
Huimaus	74,40	72,54	80,42	74,51	78,34	78,01	78,77	79,43
ICU	81,50	81,50	78,00	81,50	81,00	80,50	81,50	82,50
Ihotauti	94,81	96,99	97,54	95,63	96,72	96,99	95,63	98,09
Inkontinenssi	85,44	85,26	85,82	85,63	86,58	86,39	86,01	86,01
Kööpenhamina	58,66	58,36	59,85	58,36	58,36	58,36	58,36	58,36
Lasit	68,69	74,30	75,23	73,83	74,30	73,83	73,83	73,83
Led	71,10	71,60	71,60	70,90	71,60	71,60	71,10	71,10
Led+kohina	50,20	57,50	57,50	59,70	57,50	57,50	70,40	70,40
Liput	48,45	50,52	50,52	52,06	52,06	51,03	61,34	61,34
Ljubljana	66,08	64,69	66,43	65,73	70,98	67,13	67,13	72,38
Luotonhakijat	73,30	74,70	70,60	73,80	70,90	69,50	73,40	71,60
Mahatauti	59,79	59,49	61,97	60,47	56,79	60,02	61,59	61,37
Munkit1	87,50	88,43	99,77	91,20	99,77	99,77	79,86	79,86
Munkit2	96,30	79,63	56,48	81,94	56,48	56,48	88,66	88,66
Munkit3	76,85	84,03	99,07	95,60	99,07	99,07	100,00	100,00
Palkat	61,99	62,92	64,98	63,86	65,36	65,92	66,29	66,29
Plasebo	90,78	92,20	90,07	92,20	90,07	90,07	92,91	92,91
Profylaksi	89,76	91,57	90,36	90,36	90,36	90,36	89,16	89,16
Promoottorit	61,32	80,19	80,19	62,26	80,19	80,19	90,57	90,57
Ristinolla	74,11	98,96	98,96	83,92	98,96	98,96	87,79	87,79
Rytmihäiriö	64,82	65,04	62,17	65,04	64,82	62,61	65,49	65,49
Shakki	54,29	65,19	71,69	69,29	71,69	71,69	58,37	58,37
Sillat	60,95	61,90	63,81	60,95	59,05	58,10	61,90	64,76
SPECT	77,53	79,03	79,03	79,40	79,03	79,03	78,65	78,65
Titanic	78,92	78,87	78,87	78,87	78,87	78,87	78,87	78,87
Tyreoosi	95,23	97,19	92,47	94,54	94,11	93,37	95,15	95,15
USA	81,76	82,62	82,34	82,13	81,69	81,67	82,56	82,45
WDBC	94,73	96,84	94,38	96,66	96,84	96,84	96,66	96,66
WPBC	81,31	84,85	81,82	82,83	83,84	84,85	82,83	82,83
Keskiarvo	74,69	76,52	76,72	76,49	76,93	76,81	78,22	78,48
Mediaani	75,63	79,33	78,88	79,14	78,95	78,95	79,37	79,65

Taulukko 8.6. Luokittelutarkkuudet heterogeenisille aineistoille (%).

Aineisto	Kosini	CSM	ER	Euklidinen	Gower	HEOM	HVDM	MHVDM
Assistentit	49,67	35,76	40,40	49,01	40,40	40,40	51,66	51,66
Australia	82,46	83,62	84,64	82,32	85,22	86,09	84,06	84,06
Eturauhassyöpä	69,21	71,05	70,26	72,89	71,05	70,79	74,21	73,16
Huimaus	74,40	72,54	80,42	74,51	78,34	78,01	78,77	79,43
ICU	81,50	81,50	78,00	81,50	81,00	80,50	81,50	82,50
Kööpenhamina	58,66	58,36	59,85	58,36	58,36	58,36	58,36	58,36
Liput	48,45	50,52	50,52	52,06	52,06	51,03	61,34	61,34
Ljubljana	66,08	64,69	66,43	65,73	70,98	67,13	67,13	72,38
Luotonhakijat	73,30	74,70	70,60	73,80	70,90	69,50	73,40	71,60
Mahatauti	59,79	59,49	61,97	60,47	56,79	60,02	61,59	61,37
Palkat	61,99	62,92	64,98	63,86	65,36	65,92	66,29	66,29
Plasebo	90,78	92,20	90,07	92,20	90,07	90,07	92,91	92,91
Profylaksi	89,76	91,57	90,36	90,36	90,36	90,36	89,16	89,16
Sillat	60,95	61,90	63,81	60,95	59,05	58,10	61,90	64,76
USA	81,76	82,62	82,34	82,13	81,69	81,67	82,56	82,45
Keskiarvo	69,92	69,56	70,31	70,68	70,11	69,86	72,32	72,76
Mediaani	69,21	71,05	70,26	72,89	70,98	69,50	73,40	72,38

Tarkkuuksien lisäksi tutkittiin luokkien TPR-arvoja, koska ne antavat paremman kuvan funktioiden kyvystä luokitella muiden kuin suurimpien luokkien esimerkkejä. Kunkin aineiston eri luokkien TPR-arvoista laskettiin mediaanit, jotka on esitetty taulukoissa 8.7 (kaikki aineistot) ja 8.8 (heterogeeniset aineistot).

Tarkasteltaessa kaikkien aineistojen TPR-arvojen mediaaneja taulukossa 8.7 huomataan, että HVDM- ja MHVDM-funktioilla sekä keskiarvo että mediaani on hieman korkeampi kuin muilla funktioilla. Samat havainnot voidaan tehdä myös heterogeenisten aineistojen joukossa (taulukko 8.8). Yksittäisten aineistojen tarkkuuksissa on joitain huomattavia eroja erityisesti kaikkien aineistojen joukossa (taulukko 8.7). Esimerkiksi hedelmäpeli-aineistossa CSM-, ER- ja Gowerin funktiolla mediaani on 0%, kun se on muilla funktiolla 75-85%. Tämä tarkoittaa, että kyseiset kolme funktiota kykenivät luokittelemaan vain alle puolet aineiston luokista. Yksi selitys tähän on hedelmäpeli-aineiston luokkajakauma (katso taulukko 8.4). Hedelmäpeli-aineisto on esimerkki aineistosta, jossa on yksi selvästi muita suurempi luokka ja monta hyvin pientä luokkaa. Tyreoosi-aineisto (ER-funktion mediaani 0%) on vastaavanlainen.

Taulukko 8.7. Kaikkien aineistojen TPR-arvojen mediaanit (%).

Aineisto	Kosini	CSM	ER	Euklidinen	Gower	HEOM	HVDM	MHVDM
Alokkaat	49,99	49,99	49,99	49,99	49,99	49,99	49,99	49,99
Assistentit	48,00	30,61	40,82	48,00	40,82	40,82	51,92	51,92
Australia	82,36	83,31	84,22	82,10	84,91	85,79	83,80	83,80
Cleveland	82,25	81,34	78,46	83,14	81,53	83,22	81,39	81,39
Eturauhassyöpä	68,37	69,59	67,97	71,88	69,17	69,37	73,19	72,10
Guatemala	6,63	6,63	6,63	6,63	6,63	6,63	6,63	6,63
Hedelmäpeli	75,00	0,00	0,00	75,00	0,00	0,00	86,36	86,36
Huimaus	75,34	72,31	68,29	76,03	80,14	78,05	86,67	82,93
ICU	59,38	60,31	56,25	60,31	58,13	56,88	62,19	64,69
Ihotauti	98,15	98,86	99,55	98,86	99,55	99,55	98,86	100,00
Inkontinenssi	80,00	80,00	80,71	78,57	81,43	80,71	80,71	80,71
Kööpenhamina	54,05	53,89	56,01	53,89	53,89	53,89	53,89	53,89
Lasit	69,52	72,81	78,89	70,83	75,73	76,39	70,83	70,83
Led	70,65	71,67	71,67	71,70	71,67	71,67	71,67	71,67
Led+kohina	45,63	57,37	57,37	60,83	57,37	57,37	72,37	72,37
Liput	28,06	26,25	32,74	29,19	41,30	28,35	39,26	39,26
Ljubljana	56,52	54,85	55,41	56,95	61,02	56,93	57,27	61,68
Luotonhakijat	62,93	66,02	61,67	64,62	62,07	61,17	64,81	62,67
Mahatauti	9,09	10,53	12,00	13,64	11,11	8,00	13,64	11,11
Munkit1	87,50	88,43	99,77	91,20	99,77	99,77	79,86	79,86
Munkit2	96,16	76,56	43,33	77,57	43,33	43,33	86,34	86,34
Munkit3	77,32	83,40	99,12	95,73	99,12	99,12	100,00	100,00
Palkat	61,80	62,95	65,16	63,81	65,35	65,93	66,37	66,37
Plasebo	90,56	92,34	90,27	92,17	90,27	90,27	93,15	93,15
Profylaksi	83,59	88,45	85,45	85,45	85,45	85,45	86,08	86,08
Promoottorit	61,32	80,19	80,19	62,26	80,19	80,19	90,57	90,57
Ristinolla	70,29	98,49	98,49	78,08	98,49	98,49	86,62	86,62
Rytmihäiriö	62,61	62,96	60,08	62,85	62,79	60,34	63,33	63,33
Shakki	53,62	62,74	66,48	67,32	66,48	66,48	52,29	52,29
Sillat	47,73	52,88	45,38	47,73	38,64	38,64	48,18	56,92
SPECT	52,86	69,29	69,29	68,18	69,29	69,29	67,71	67,71
Titanic	68,73	68,22	68,22	68,22	68,22	68,22	68,22	68,22
Tyreoosi	73,12	75,27	0,00	74,19	65,59	58,06	83,87	83,87
USA	74,18	75,83	74,71	74,70	73,58	73,70	75,50	75,37
WDBC	93,60	95,95	93,32	95,71	96,14	95,95	95,71	95,71
WPBC	50,34	58,36	53,61	52,71	54,79	58,36	52,71	52,71
Keskiarvo	64,65	64,96	62,54	66,95	65,11	64,34	69,50	69,70
Mediaani	68,55	69,44	67,23	69,53	67,35	67,35	72,02	71,88

Taulukko 8.8. Heterogeenisten aineistojen TPR-arvojen mediaanit (%).

Aineisto	Kosini	CSM	ER	Euklidinen	Gower	HEOM	HVDM	MHVDM
Assistentit	48,00	30,61	40,82	48,00	40,82	40,82	51,92	51,92
Australia	82,36	83,31	84,22	82,10	84,91	85,79	83,80	83,80
Eturauhassyöpä	68,37	69,59	67,97	71,88	69,17	69,37	73,19	72,10
Huimaus	75,34	72,31	68,29	76,03	80,14	78,05	86,67	82,93
ICU	59,38	60,31	56,25	60,31	58,13	56,88	62,19	64,69
Kööpenhamina	54,05	53,89	56,01	53,89	53,89	53,89	53,89	53,89
Liput	28,06	26,25	32,74	29,19	41,30	28,35	39,26	39,26
Ljubljana	56,52	54,85	55,41	56,95	61,02	56,93	57,27	61,68
Luotonhakijat	62,93	66,02	61,67	64,62	62,07	61,17	64,81	62,67
Mahatauti	9,09	10,53	12,00	13,64	11,11	8,00	13,64	11,11
Palkat	61,80	62,95	65,16	63,81	65,35	65,93	66,37	66,37
Plasebo	90,56	92,34	90,27	92,17	90,27	90,27	93,15	93,15
Profylaksi	83,59	88,45	85,45	85,45	85,45	85,45	86,08	86,08
Sillat	47,73	52,88	45,38	47,73	38,64	38,64	48,18	56,92
USA	74,18	75,83	74,71	74,70	73,58	73,70	75,50	75,37
Keskiarvo	60,13	60,01	59,76	61,36	61,06	59,55	63,73	64,13
Mediaani	61,80	62,95	61,67	63,81	62,07	61,17	64,81	64,69

8.4.2. Luokittelutulosten tilastollinen testaus

Etäisyysfunktioiden luokittelutulosten tilastollinen testaus osoittautui monimutkaiseksi tehtäväksi, jossa oli huomioitava otosten riippuvuus, pieni otoskoko ja *p*-arvojen tulkinta. Tuloksia ei voinut selvästikään olettaa toisistaan riippumattomiksi, koska funktiot luokittelivat samat aineistot ja lisäksi testijoukot olivat kussakin aineistossa sisällöltään samat. Tällaisessa tilanteessa käytetään usein otosten riippuvuuden huomioivaa versiota varianssianalyysistä ja pareittaisia *t*-testejä varianssianalyysissä mahdollisesti havaitun merkitsevän *p*-arvon tuottaneiden pareittaisten erojen tunnistamiseen. Tässä tutkielmassa päädyttiin kuitenkin edellä mainittujen menetelmien parametrittomiin vastineisiin, koska heterogeenisten aineistojen määrä oli pieni (N=15) ja analyysissä haluttiin käyttää samoja testimenetelmiä sekä kaikille että heterogeenisille aineistoille. Friedmanin testi ja Wilcoxonin testi [Pett, 1997] tekevät löyhempiä jakaumaoletuksia ja ovat siten vastineitaan soveliaampia pienillä otoksilla, joissa parametristen testien oletukset eivät useinkaan toteudu.

Tulosten testaus suoritettiin tavanomaisen tapaan kahdessa vaiheessa. Aluksi tutkittiin Friedmanin testillä, onko muuttujien mediaaneissa tilastollisesti merkitseviä ($p < 0{,}05$) eroja. Testin nollahypoteesi H_0 oli siis muotoa H_0: $md_1 = md_2 = \ldots = md_8$ ja vaihtoehtoinen hypoteesi H_1 muotoa H_1: $md_i \neq md_j$, ainakin muuttujilla i ja j. Mikäli merkittävä ero havaittiin, testattiin kaikki 8!/(2!6!) = 28 muuttujaparia kaksisuuntaisella Wilcoxonin testillä, joka testaa, eroavatko muuttujien mediaanit. (H_0: $md_i = md_j$ ja H_1: $md_i \neq md_j$ muuttujille i ja j.) Testaus suoritettiin SPSS-ohjelmistolla. Friedmanin testin *p*-arvo arvioitiin Monte Carlo -menetelmällä ja Wilcoxonin testin merkitsevyys laskettiin tarkalla algoritmilla.

Wilcoxonin testeissä ei voitu käyttää alkuperäistä *p*-arvoa 0,05 niin sanotun monivertailu- eli monitestausongelman vuoksi. Monivertailuongelma (multiplicity effect) on tunnettu pitkään tilastotieteessä, mutta koneoppimisen ja tiedonlouhinnan alalla tähän ongelmaan on kiinnitetty laajemmin huomiota vasta viime aikoina [Salzberg, 1999]. Yleistäen monivertailuongelmalla tarkoitetaan mahdollisuutta saada toistetussa analyysissä yksi tai useampi merkitsevä tulos sattumalta [Hochberg and Tamane, 1987]. Esimerkiksi todennäköisyys, että 28 testissä on yksi tai useampia tyypin I virheitä (H_0 hylätään, kun H_0 on tosi) on $1 - (1 - 0.05)^{28} \approx 0{,}72$. Tulos perustuu tapahtumien todennäköisyyksien kertolaskusääntöön, jossa tapahtumat on oletettu toisistaan riippumattomiksi.

Merkitsevien erojen tunnistamiseen ja monivertailuongelman samanaikaiseen hallintaan varianssianalyysissä on esitetty useita menetelmiä kuten esimerkiksi Fisherin LSD-kriteeri [Hochberg and Tamane, 1987]. Valitettavasti nämä menetelmät eivät soveltuneet käytettäväksi Friedmanin testin epäparametrisen luonteen vuoksi. Tästä syystä merkitsevät erot paikannettiin parillisilla testeillä käyttäen Bonferroni- ja Kounias-korjattuja *p*-arvoja.

Bonferroni- ja Kounias-korjaukset approksimoivat todennäköisyyttä, jolla *k* testissä tapahtuu yksi tai useampi virhe. Olkoot A_1, A_2, … , A_k tapahtumia. Tapahtumalla tarkoitetaan satunnaiskokeen otosavaruuden osajoukkoa [Freund, 1971]. Tapahtumien todennäköisyyksien yhteenlaskusääntö voidaan ilmaista Boolen kaavan avulla [Hochberg and Tamane, 1987; Freund, 1971]

$$1-P\left(\bigcap_{i=1}^{k} A_i\right)=P\left(\bigcup_{i=1}^{k} A_i^C\right) \qquad (8.1)$$
$$=\sum_{i=1}^{k} P\left(A_i^C\right)+\sum_{i<j} P\left(A_i^C \cap A_j^C\right)+\sum_{i<j<k} P\left(A_i^C \cap A_j^C \cap A_k^C\right)-\ldots+(-1)^{n-1} P\left(\bigcap_{i=1}^{k} A_i^C\right),$$

missä A_i^C on tapahtuman A_i komplementti. Koska tiedetään, että yleisesti [Freund, 1971]

$$P\left(\bigcup_{i=1}^{k} A_i\right) \le \sum_{i=1}^{k} P\left(A_i\right), \qquad (8.2)$$

voidaan yhtälön 8.1 vasenta puolta approksimoida oikean puolen ensimmäisen termin avulla

$$1-P\left(\bigcap_{i=1}^{k} A_i\right) \leq \sum_{i=1}^{k} P\left(A_i^C\right), \tag{8.3}$$

josta termejä järjestelemällä saadaan Bonferronin epäyhtälö

$$P\left(\bigcap_{i=1}^{k} A_i\right) \geq 1-\sum_{i=1}^{k} P\left(A_i^C\right). \tag{8.4}$$

Kun määritellään A_i = { *i.* testi virheetön | *i.* testin H_0 tosi } ja $P(A_i) = 1- p$, voidaan korjattu *p*-arvo *p*' laskea ratkaisemalla yhtälö

$$1 - p \geq 1 - kp', \tag{8.5}$$

josta saadaan Bonferroni-korjaus (Bonferroni adjustment) [Pett, 1997; Salzberg, 1999; Hochberg and Tamane, 1987]

$$p' \geq p / k. \tag{8.6}$$

Yhtälöllä 8.6 korjattu *p*-arvo *p*' on 0,05 / 28 ≈ 0,002.

Bonferroni-approksimaatio vastaa tilannetta, jossa muuttujat oletetaan keskenään riippumattomiksi. Tästä syystä korjatut *p*-arvot ovat epärealistisen pieniä, kun muuttujilla on paljon keskinäistä riippuvuutta. Kouniaksen epäyhtälöön [Hochberg and Tamane, 1987]

$$P\left(\bigcap_{i=1}^{k} A_i\right) \geq 1-\sum_{i=1}^{k} P\left(A_i^C\right)+\max_j \sum_{i\neq j} P\left(A_i^C \cap A_j^C\right) \tag{8.7}$$

perustuva korjaus ottaa huomioon muuttujien pareittaiset todennäköisyydet. Erityisesti, jos $P(A_1{}^C) = P(A_2{}^C)$ ja $P(A_1{}^C \cap A_2{}^C) = P(A_i{}^C \cap A_j{}^C)$, missä $1 \leq i \neq j \leq k$, niin yhtälö 8.7 sievenee muotoon

$$P\left(\bigcap_{i=1}^{k} A_i\right) \geq 1-kP\left(A_1^C\right)+(k-1)P\left(A_1^C \cap A_2^C\right). \tag{8.8}$$

ja voidaan kirjoittaa edelleen muotoon

$$P\left(\bigcap_{i=1}^{k} A_i\right) \geq 1-kP\left(A_1^C\right)+(k-1)P\left(A_1^C \mid A_2^C\right)P\left(A_2^C\right) \tag{8.9}$$

todennäköisyyksien tulosäännön perusteella.

Tässä työssä käytetty Kounias-korjaus perustuu yhtälöön 8.9. Kun muuttujien pareittaisen riippuvuuden voimakkuutta arvioidaan Spearmanin korrelaation keskiarvolla $, \bar{s}$ saadaan Kounias-korjattu *p*-arvo *p*' seuraavasti

$$p' \geq p / (k - (k-1)\,\bar{s}\,). \tag{8.10}$$

Etäisyysfunktioita vertailtiin sekä luokittelutarkkuuksien, että TPR-mediaanien avulla. Tulokset esitetään että taulukkona että laatikko-jana-diagrammina. Diagrammissa (katso kuvat 8.1, 8.2, 8.3 ja 8.4) janat kuvaavat ei-poikkeavia havaintoja, jotka sijoittuvat kvartiilien ulkopuolelle. Janojen päätepisteiden ulkopuolella olevat pienet ympyrät kuvaavat poikkeavia havaintoja (outlier). Janojen päätepisteet ol_{ALA} ja $ol_{YLÄ}$ lasketaan seuraavasti:

$$ol_{ALA} = q_{25} - 1{,}5 \cdot iqr \text{ ja} \tag{8.11}$$

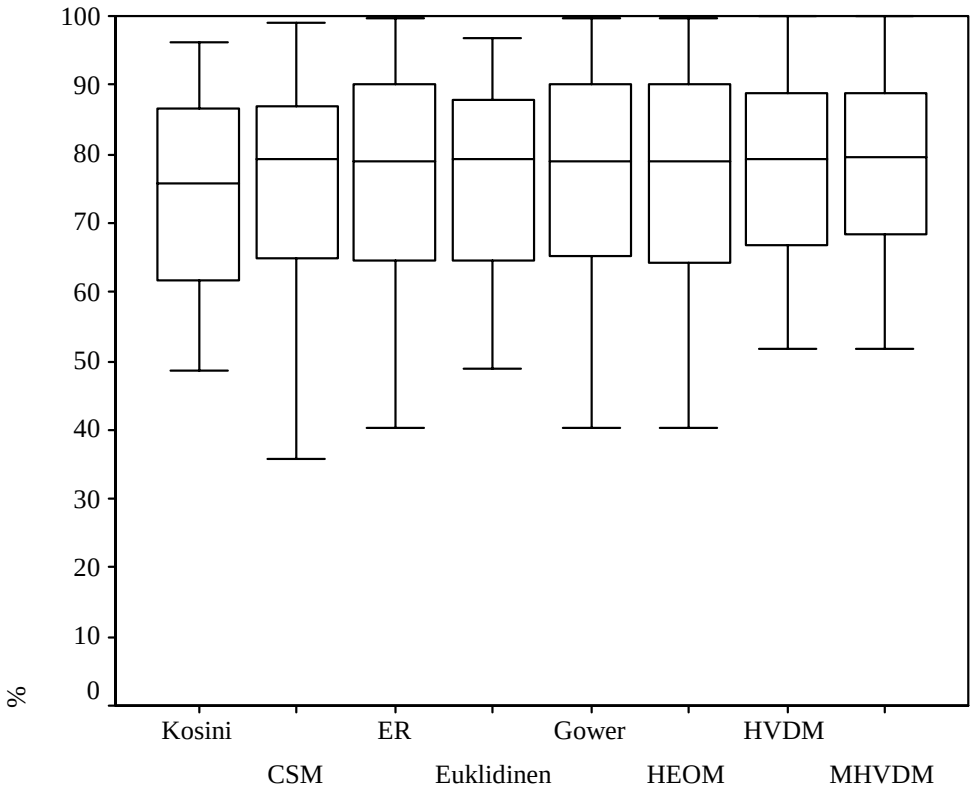
$$ol_{YLÄ} = q_{75} + 1{,}5 \cdot iqr, \tag{8.12}$$

missä *iqr* on kvartiiliväli ($q_{75} - q_{25}$) ja q_{75} on 75%- ja q_{25} 25%-kvartiili. Alempi jana piirretään seuraavasti. Haetaan lähin ol_{ALA}-arvoa oleva havainto ($> ol_{ALA}$) ja piirretään jana tähän havaintoon saakka. Esimerkiksi, jos $ol_{ALA} = 10$ ja ensimmäinen ol_{ALA}-arvoa pienempi havainto on 9,9 ja ensimmäinen ol_{ALA}-arvoa suurempi havainto on 10,1, niin jana ulottuu alakvartiilista havaintoon 10,1 ja havainto 9,9 katsotaan poikkeavaksi havainnoksi. Ylempi jana piirretään samalla periaatteella eli haetaan arvoa $ol_{YLÄ}$ lähin havainto ($< ol_{YLÄ}$) ja piirretään jana yläkvartiilista kyseiseen havaintoon.

Taulukossa 8.9 on Wilcoxonin pareittaisen testin tulokset kaikista aineistoista lasketuille luokittelutarkkuuksille ja kuvassa 8.1 on samaa asiaa havainnollistava laatikko-jana-diagrammi. Tässä testissä kosinietäisyysfunktion luokittelutarkkuus on merkittävästi heikompi kuin kaikkien muiden. Lisäksi MHVDM on parempi verrattuna euklidiseen etäisyysfunktioon.

Taulukko 8.9. Wilcoxonin testin tulokset kaikista aineistoista lasketuille tarkkuuksille. [a] Bonferroni-korjattu *p*-arvon merkitsevyystaso. [b] Kounias-korjattu *p*-arvon merkitsevyystaso. Tummennetun funktion luokittelutarkkuus oli parivertailussa Kounias-korjatun *p*-arvon perusteella merkittävästi pariaan korkeampi.

Testipari		Mediaanit		*p* -arvo	< 0,002[a]	< 0,01[b]
Kosini	**CSM**	75,63	79,33	0,000	X	X
Kosini	**ER**	75,63	78,88	0,000	X	X
Kosini	**Euklidinen**	75,63	79,14	0,000	X	X
Kosini	**Gower**	75,63	78,95	0,000	X	X
Kosini	**HEOM**	75,63	78,95	0,000	X	X
Kosini	**HVDM**	75,63	79,37	0,000	X	X
Kosini	**MHVDM**	75,63	79,65	0,000	X	X
CSM	ER	79,33	78,88	0,880		
CSM	Euklidinen	79,33	79,14	0,608		
CSM	Gower	79,33	78,95	0,559		
CSM	HEOM	79,33	78,95	0,500		
CSM	HVDM	79,33	79,37	0,143		
CSM	MHVDM	79,33	79,65	0,063		
ER	Euklidinen	78,88	79,14	0,846		
ER	Gower	78,88	78,95	0,452		
ER	HEOM	78,88	78,95	0,674		
ER	HVDM	78,88	79,37	0,104		
ER	MHVDM	78,88	79,65	0,038		
Euklidinen	Gower	79,14	78,95	0,542		
Euklidinen	HEOM	79,14	78,95	0,626		
Euklidinen	HVDM	79,14	79,37	0,013		
Euklidinen	**MHVDM**	79,14	79,65	0,009		X
Gower	HEOM	78,95	78,95	0,733		
Gower	HVDM	78,95	79,37	0,167		
Gower	MHVDM	78,95	79,65	0,032		
HEOM	HVDM	78,95	79,37	0,141		
HEOM	MHVDM	78,95	79,65	0,055		
HVDM	MHVDM	79,37	79,65	0,359		

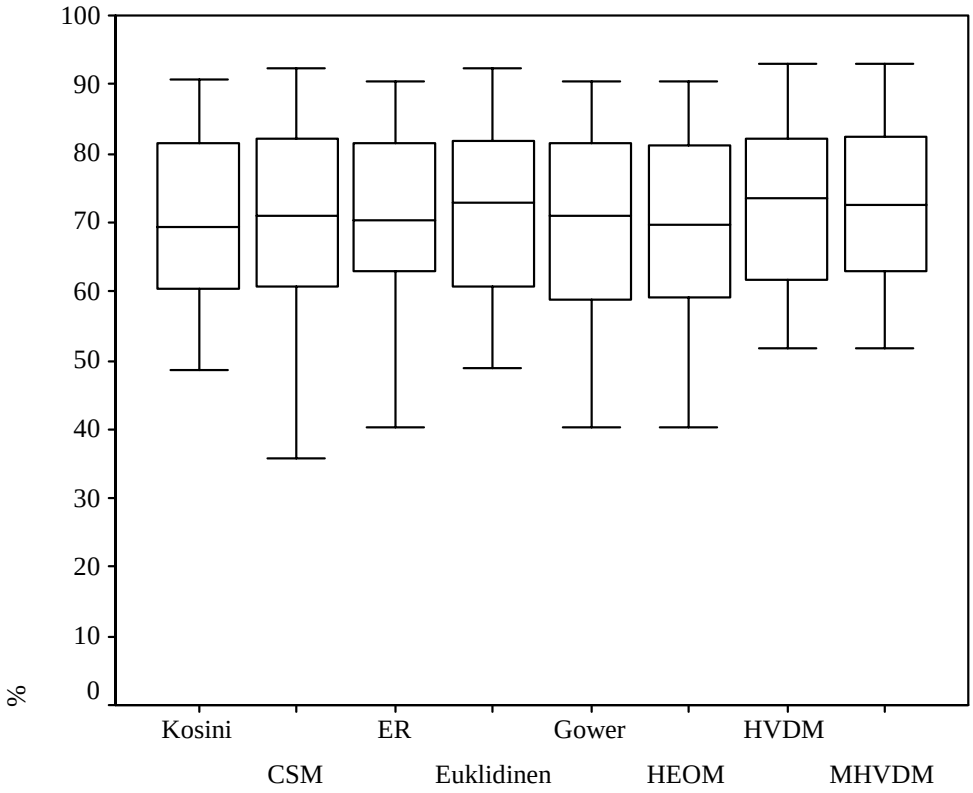


Kuva 8.1. Luokittelutarkkuuksien laatikko-jana-diagrammi kaikille aineistoille (%).

Taulukossa 8.10 ja kuvassa 8.2 on vastaavat tulokset ainoastaan heterogeenisten aineistojen joukossa. Merkitseviä eroja näkyy tässä hieman enemmän kuin kaikilla aineistoilla. Kosinietäisyysfunktio on tässäkin heikompi muihin verrattuna. Lisäksi HVDM on euklidista etäisyys- ja HEOM-funktiota parempi. Euklidinen etäisyysfunktio on MHVDM-funktiota merkittävästi parempi. MHVDM-funktion tarkkuudessa on merkitsevä ero Gower- ja HEOM-funktioon verrattuna.

Taulukko 8.10. Wilcoxonin testin tulokset heterogeenisistä aineistoista lasketuille tarkkuuksille. [a] Bonferroni-korjattu *p*-arvon merkitsevyystaso. [b] Kounias-korjattu *p*-arvon merkitsevyystaso. Tummennetun funktion luokittelutarkkuus oli parivertailussa Kounias-korjatun *p*-arvon perusteella merkittävästi pariaan korkeampi.

Testipari		Mediaanit		*p* -arvo	< 0,002[a]	< 0,02[b]
Kosini	**CSM**	69,21	71,05	0,001	X	X
Kosini	**ER**	69,21	70,26	0,000	X	X
Kosini	**Euklidinen**	69,21	72,89	0,000	X	X
Kosini	**Gower**	69,21	70,98	0,000	X	X
Kosini	**HEOM**	69,21	69,50	0,000	X	X
Kosini	**HVDM**	69,21	73,40	0,000	X	X
Kosini	**MHVDM**	69,21	72,38	0,000	X	X
CSM	ER	71,05	70,26	0,426		
CSM	Euklidinen	71,05	72,89	0,203		
CSM	Gower	71,05	70,98	0,735		
CSM	HEOM	71,05	69,50	0,760		
CSM	HVDM	71,05	73,40	0,026		
CSM	MHVDM	71,05	72,38	0,024		
ER	Euklidinen	70,26	72,89	0,714		
ER	Gower	70,26	70,98	0,850		
ER	HEOM	70,26	69,50	0,518		
ER	HVDM	70,26	73,40	0,151		
ER	MHVDM	70,26	72,38	0,055		
Euklidinen	Gower	72,89	70,98	0,339		
Euklidinen	HEOM	72,89	69,50	0,375		
Euklidinen	**HVDM**	72,89	73,40	0,004		X
Euklidinen	MHVDM	72,89	72,38	0,013		X
Gower	HEOM	70,98	69,50	0,700		
Gower	HVDM	70,98	73,40	0,041		
Gower	**MHVDM**	70,98	72,38	0,004		X
HEOM	**HVDM**	69,50	73,40	0,017		X
HEOM	**MHVDM**	69,50	72,38	0,005		X
HVDM	MHVDM	73,40	72,38	0,640		

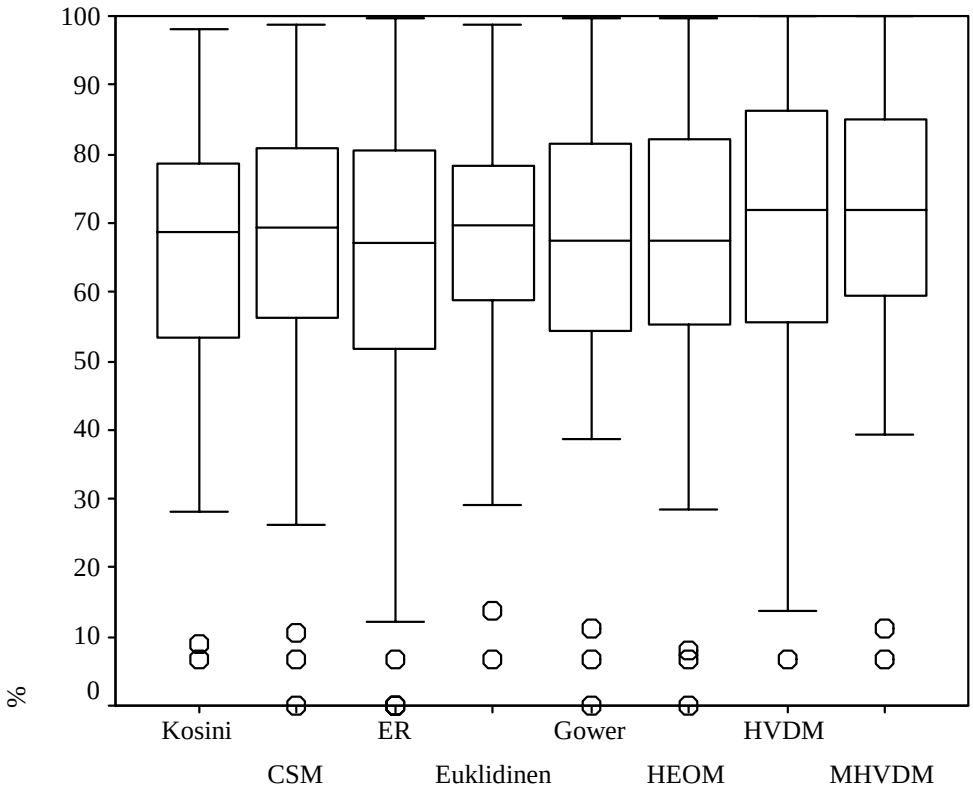


Kuva 8.2. Luokittelutarkkuuksien laatikko-jana-diagrammi heterogeenisille aineistoille.

Taulukossa 8.11 on Wilcoxonin pareittaisen *t*-testin tulokset kaikista aineistoista lasketuille TPR-mediaaneille ja kuvassa 8.3 on samaa asiaa havainnollistava laatikko-jana-diagrammi. Testin perusteella kosinietäisyysfunktion TPR-mediaani on merkittävästi pienempi kuin CSM-, euklidisen etäisyys-, HVDM- ja MHVDM-funktion vastaava arvo. Sen sijaan kosinietäisyysfunktion TPR-mediaani on merkittävästi korkeampi kuin ER-, Gowerin ja HEOM-funktion vastaava arvo. Lisäksi HVDM- ja MHVDM-funktion TPR-mediaani on euklidiseen etäisyysfunktioon verrattuna merkitsevästi korkeampi.

Taulukko 8.11. Wilcoxonin testin tulokset kaikista aineistoista lasketuille TPR-mediaaneille. [a] Bonferroni-korjattu *p*-arvon merkitsevyystaso. [b] Kounias-korjattu *p*-arvon merkitsevyystaso. Tummennetun funktion luokittelutarkkuus oli parivertailussa Kounias-korjatun *p*-arvon perusteella merkittävästi pariaan korkeampi.

Testipari		Mediaanit		*p* -arvo	< 0,002[a]	< 0,007[b]
Kosini	**CSM**	68,55	69,44	0,000	X	X
Kosini	ER	68,55	67,23	0,002		X
Kosini	**Euklidinen**	68,55	69,53	0,000	X	X
Kosini	Gower	68,55	67,35	0,000	X	X
Kosini	HEOM	68,55	67,35	0,000	X	X
Kosini	**HVDM**	68,55	72,02	0,000	X	X
Kosini	**MHVDM**	68,55	71,88	0,000	X	X
CSM	ER	69,44	67,23	0,546		
CSM	Euklidinen	69,44	69,53	0,583		
CSM	Gower	69,44	67,35	0,531		
CSM	HEOM	69,44	67,35	0,944		
CSM	HVDM	69,44	72,02	0,076		
CSM	MHVDM	69,44	71,88	0,035		
ER	Euklidinen	67,23	69,53	0,286		
ER	Gower	67,23	67,35	0,079		
ER	HEOM	67,23	67,35	0,442		
ER	HVDM	67,23	72,02	0,016		
ER	MHVDM	67,23	71,88	0,020		
Euklidinen	Gower	69,53	67,35	0,915		
Euklidinen	HEOM	69,53	67,35	0,581		
Euklidinen	**HVDM**	69,53	72,02	0,002		X
Euklidinen	**MHVDM**	69,53	71,88	0,005		X
Gower	HEOM	67,35	67,35	0,132		
Gower	HVDM	67,35	72,02	0,157		
Gower	MHVDM	67,35	71,88	0,124		
HEOM	HVDM	67,35	72,02	0,063		
HEOM	MHVDM	67,35	71,88	0,047		
HVDM	MHVDM	72,02	71,88	0,820		



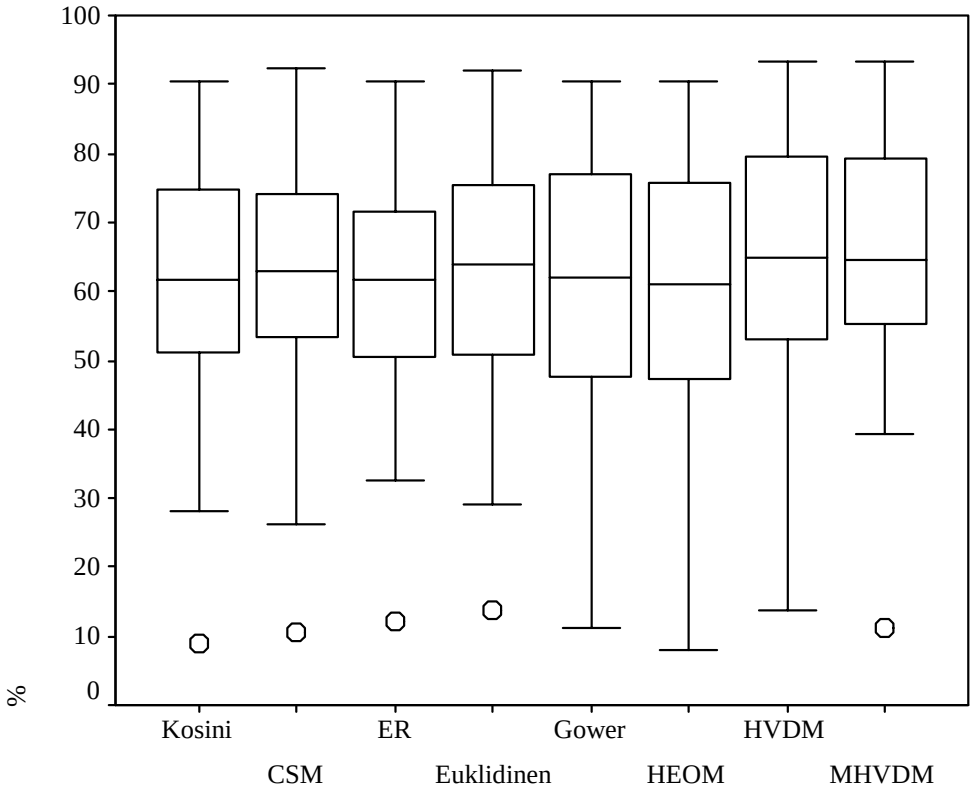
Kuva 8.3. TPR-mediaanien laatikko-jana-diagrammi kaikille aineistoille. Diagrammissa ympyrät tarkoittavat poikkeavia arvoja (outlier).

Kuvassa 8.3 huomataan, että ER- ja HVDM-funktion kohdalla alajana jatkuu selvästi pidempään kuin muilla funktioilla. Tämä ei kuitenkaan ei ole merkittävää tuntien janan piirtomekanismin (katso kaavat 8.11 ja 8.12).

Taulukossa 8.12 ja kuvassa 8.4 on vastaavat tulokset ainoastaan heterogeenisten aineistojen joukossa. Merkitseviä eroja on enemmän kuin aiemmissa testeissä. Kosinietäisyysfunktion TPR-mediaani on merkittävästi pienempi verrattuna CSM-, euklidiseen etäisyys-, Gowerin-, HVDM-, ja MHVDM-funktioon. Kosinietäisyysfunktion TPR-mediaani on merkittävästi suurempi verrattuna ER- ja HEOM-funktioon. Lisäksi HVDM-funktion TPR-mediaani on kaikkia muita paitsi CSM-, MHVDM- ja Gowerin funktiota merkittävästi korkeampi. MHVDM-funktion TPR-mediaani on kaikkia muita paitsi HVDM-funktiota merkittävästi suurempi. Myös kuvassa 8.4 muutaman laatikon alajanat ovat poikkeuksellisen pitkät, mikä on kuitenkin käytännössä merkityksetöntä.

Taulukko 8.12. Wilcoxonin testin tulokset heterogeenisistä aineistoista lasketuille TPR-mediaaneille. [a] Bonferroni-korjattu *p*-arvon merkitsevyystaso. [b] Kounias-korjattu *p*-arvon merkitsevyystaso. Tummennetun funktion luokittelutarkkuus oli parivertailussa Kounias-korjatun *p*-arvon perusteella merkittävästi pariaan korkeampi.

Testipari		Mediaanit		*p* -arvo	$< 0{,}002^a$	$< 0{,}02^b$
Kosini	**CSM**	61,80	62,95	0,005		X
Kosini	ER	61,80	61,67	0,000	X	X
Kosini	**Euklidinen**	61,80	63,81	0,000	X	X
Kosini	**Gower**	61,80	62,07	0,000	X	X
Kosini	HEOM	61,80	61,17	0,002		X
Kosini	**HVDM**	61,80	64,81	0,000	X	X
Kosini	**MHVDM**	61,80	64,69	0,000	X	X
CSM	ER	62,95	61,67	0,599		
CSM	Euklidinen	62,95	63,81	0,454		
CSM	Gower	62,95	62,07	0,714		
CSM	HEOM	62,95	61,17	0,714		
CSM	HVDM	62,95	64,81	0,057		
CSM	**MHVDM**	62,95	64,69	0,010		X
ER	Euklidinen	61,67	63,81	0,104		
ER	Gower	61,67	62,07	0,380		
ER	HEOM	61,67	61,17	0,791		
ER	**HVDM**	61,67	64,81	0,001	X	X
ER	**MHVDM**	61,67	64,69	0,004		X
Euklidinen	Gower	63,81	62,07	0,839		
Euklidinen	HEOM	63,81	61,17	0,094		
Euklidinen	**HVDM**	63,81	64,81	0,000	X	X
Euklidinen	**MHVDM**	63,81	64,69	0,010		X
Gower	HEOM	62,07	61,17	0,083		
Gower	HVDM	62,07	64,81	0,024		
Gower	**MHVDM**	62,07	64,69	0,017		X
HEOM	**HVDM**	61,17	64,81	0,001	X	X
HEOM	**MHVDM**	61,17	64,69	0,001	X	X
HVDM	MHVDM	64,81	64,69	0,945		



Kuva 8.4. TPR-mediaanien laatikko-jana-diagrammi heterogeenisille aineistoille. Diagrammissa ympyrät tarkoittavat poikkeavia arvoja (outlier).

9. Huomioita tuloksista

Testit tehtiin käyttäen kahta eri tunnuslukua ja erikseen kaikille ja ainoastaan heterogeenisille aineistoille. Suurimpia eroja funktioiden välillä oli testeissä, jotka tehtiin pelkästään heterogeenisille aineistoille. Heterogeeniset aineistot olivat myös odotetusti vaikeampia luokitella kuin muut aineistot. Homogeenisissä kvantitatiivisissa aineistoissa funktiot toimivat suunnilleen samalla tavoin eli kuten Manhattan- tai euklidinen etäisyysfunktio (poikkeuksena kosinietäisyys- ja ER-funktio). Näissäkin aineistoissa (lasit, WDBD ja WBPC) funktioiden välillä on pieniä eroja. Erot johtuvat funktioiden eri normalisointitavoista. Jatkotutkimuksissa saattaisi olla järkevämpää käyttää samoja normalisoimismenetelmiä kaikissa aineistoissa, jotta normalisoimistavan vaikutus funktion toimivuuteen saataisiin poistettua.

Kokonaisuutena luokittelusta parhaiten suoriutuvat funktiot olivat HVDM ja MHVDM. Yhtä poikkeusta lukuunottamatta kumpikaan näistä funktioista ei ollut pariaan merkittävästi huonompi Wilcoxonin testin perusteella. Poikkeus oli tilanne, jossa vertailtiin luokittelutarkkuuksia heterogeenisille aineistoille (katso taulukot 8.6 ja 8.10). Kyseisessä tilanteessa euklidinen etäisyysfunktion mediaani oli testin perusteella MHVDM-funktiota merkittävästi korkeampi. Toisaalta taulukosta 8.6 nähdään, että tässä tilanteessa MHVDM-funktion luokittelutarkkuuksien keskiarvo oli noin kaksi prosenttiyksikköä euklidisen etäisyysfunktion vastaavaa korkeampi. Lisäksi suurin osa yksittäisten aineistojen luokittelutarkkuuksista oli korkeampi MHVDM-funktiolla kuin euklidisella etäisyysfunktiolla. Näistä syistä tässä tilanteessa ei ole järkevää olettaa euklidisen etäisyysfunktion toimivan MHVDM-funktiota paremmin, vaikka Wilcoxonin testillä saadaankin merkitsevä ero euklidisen etäisyysfunktion hyväksi. Edellistä lukuun ottamatta HVDM- ja MHVDM-funktio olivat muita funktiota tarkempia erityisesti 15 heterogeenisen aineiston joukossa. Tässä joukossa funktioiden mediaanit olivat suurimmassa osassa parivertailuja pariaan korkeampia. Lisäksi funktioiden keskiarvot olivat selvästi parhaita. Funktiot saavuttivat myös selvästi parhaan tarkkuuden led+kohina- ja promoottorit-aineistossa, jotka on kuvattu pelkästään nominaalisilla attribuuteilla.

Tulokset eivät ole yllättäviä, kun tiedetään VDM-pohjaisten funktioiden tapa painottaa nominaalisia attribuutteja luokkatietojen perusteella. Tarkastellaan esimerkkiä, jossa $x_a = i$ ja $y_a = j$, missä $i \neq j$ ja a on nominaalinen. Oletetaan, että kaikki esimerkit, joilla on attribuutin arvona joko i tai j, kuuluvat luokkaan c. Tässä tilanteessa $vdm_a(x, y) = 0$, mutta useimmilla heterogeenisillä funktioilla etäisyys $d_a(x, y) > 0$ (monesti $d_a = 1$). Käytännössä edellä mainittu tilanne voisi esiintyä esimerkiksi, kun kaksi erilaista oiretta (o1 ja o2) liittyvät usein erääseen tautiin (t1). Kolmas oire (o3) liittyy tyypillisesti toiseen tautiin (t2). Olkoot x, y ja z potilaat, joilla x_a = o1, y_a = o3 ja z_a = o2. Tiedetään, että potilas x sairastaa tautia t1 ja potilas y tautia t2. Tehtävänä on ennustaa potilaan z tau-

ti *a*:n perusteella lähimmän naapurin menetelmällä. Tässä tilanteessa VDM määrittelisi potilaat *x* ja *z* keskenään samankaltaisiksi ja potilaat *y* ja *z* keskenään erilaisiksi. Johtopäätös olisi, että *y*:llä on tauti t1, mikä olisi todennäköisesti oikea johtopäätös, tuntien taudin tyypilliset oireet. Perinteisillä heterogeenisillä funktioilla kaikki kolme potilasta olisivat yhtä erilaisia, joten oikeaan diagnoosiin päädyttäisiin vain sattumalta. VDM-pohjaisten funktioiden etu tyypillisiin heterogeenisiin funktioihin verrattuna on siis selvä tällaisissa luokittelutilanteissa. Ongelmana on se, että VDM-funktioita ei voida käyttää esimerkiksi klusteroinnissa tai yleensäkään tilanteissa, joissa ei ole käytettävissä opetusjoukkoa, josta VDM-todennäköisyydet lasketaan.

VDM toimi erityisen huonosti nominaalisessa munkit1-aineistossa, jossa VDM-funktioiden luokittelutarkkuus ja TPR-mediaani olivat selvästi matalammat kuin tyypillisillä heterogeenisillä funktioilla. Munkit1 on kaksiluokkainen aineisto $c = \{c0, c1\}$, jossa on kuusi nominaalista attribuuttia a1, a2, a3, a4, a5 ja a6. Se, kuuluuko esimerkki *x* luokkaan c1, määräytyy sen perusteella onko esimerkille voimassa $x_a = x_b$, kun a = a1 ja b = a2, tai $x_a = 1$, kun a = a5 [Blake and Merz, 1998]. Jos kumpikaan ehto ei ole voimassa, esimerkki kuuluu luokkaan c0. Muilla attribuuteilla ei ole merkitystä luokan määräytymisen kannalta. VDM-todennäköisyydet koko aineistossa on esitetty taulukossa 9.1. VDM-funktion kannalta ongelmana on, että merkittäviä esimerkkejä erottavia attribuutteja on vain yksi (a5). Esimerkiksi $vdm_a(x, y) = 0$, kun a = a1 tai a = a2, aina riippumatta siitä, mitkä attribuuttien a1 ja a2 arvot ovat. Tietenkin ristiinvalidoinnissa kunkin validointiaskeleen opetusjoukon frekvenssit olivat hieman erilaisia, koska joukkojen valitsemiseen käytettiin yksinkertaista satunnaisotantaa. Tällöin satunnaisesti muutkin attribuutit kuin a5 saattoivat erotella esimerkkejä.

Taulukko 9.1. VDM-todennäköisyydet munkit1-aineistossa.

	a1			a2			a3		a4			a5				a6	
c	1	2	3	1	2	3	1	2	1	2	3	1	2	3	4	1	2
c0	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,00	0,67	0,67	0,67	0,50	0,50
c1	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50	1,00	0,33	0,33	0,33	0,50	0,50

Kosinietäisyysfunktio suoriutui luokittelusta kaikkein heikoimmin. Luokittelutarkkuuksia vertailtaessa (katso taulukot 8.9 ja 8.10) kosinietäisyysfunktio oli jokaisessa seitsemässä parivertailussa pariaan heikompi. TPR-mediaaneja vertailtaessa (katso taulukot 8.11 ja 8.12) tilanne ei ollut aivan näin selvä, mutta suurimmassa osassa näitäkin parivertailuja kosinietäisyysfunktio oli pariaan heikompi. Kosinietäisyysfunktion läheisen sukulaisen korrelaatiokertoimen käyttöä reaalimaailman datan etäisyyslaskennassa on kritisoitu [Fleiss and Zubin, 1969]. Kritiikki koskee lähinnä sitä, että korrelaatiokertoimeen perustuva funktio määrittää kaikki lineaarisesti riippuvat esimerkit samankaltaisiksi, eli jos $y = ax + b$, missä $a, b \in \boldsymbol{R}$, niin korrelaatiokertoimeen perustuva etäisyys $d\_corr(x, y) = 0$. Etäisyyden määrittäminen tällä tavalla ei ole reaalimaailman datan

kanssa yleensä sopivaa. Kritiikki ei täysin päde kosinietäisyysfunktioon, koska edellä mainituille esimerkeille x ja y etäisyys $d\_cos(x, y) \neq 0$, jos $b \neq 0$ ($d\_cos(x, y) = 0$ vain, jos $b = 0$). Toisaalta laskiessaan etäisyyden esimerkkivektorien välisen kulman perusteella, kosinietäisyysfunktio kykenee tunnistamaan esimerkit, joiden muoto (katso kohta 5.5) on sama, vaikka esimerkkien attribuuttien arvoilla olisi suuriakin eroavaisuuksia. Tulosten perusteella tästä ominaisuudesta ei käytettyjen aineistojen kohdalla ollut kuitenkaan hyötyä lukuun ottamatta munkit2-aineistoa, jossa kosinietäisyysfunktion tarkkuus on selvästi muita korkeampi.

Tarkasteltaessa taulukon 8.5 luokittelutarkkuuksia huomataan, että monen nominaalisia attribuutteja sisältävän aineiston kohdalla euklidisen etäisyysfunktion ja niin sanottujen tyypillisten heterogeenisten etäisyysfunktioiden (CSM, ER, Gower ja HEOM) väliset erot luokittelutarkkuudessa ovat pieniä tai eroja ei ole lainkaan. Tyypillisenä heterogeenisenä funktiona voidaan pitää funktiota, jossa nominaalisten attribuuttien arvojen x_a ja y_a etäisyys on 0, jos $x_a = y_a$, ja jokin vakio c (yleensä $c = 1$), jos $x_a \neq y_a$. Ordinaalisten ja kvantitatiivisten attribuuttien kohdalla kyseisissä funktiossa käytetään ER-funktiota lukuun ottamatta arvojen välistä erotusta $d = x_a - y_a$. Lisäksi funktioiden välillä on pieniä eroja muun muassa attribuuttien normalisoinnissa. Tarkasteltaessa lähemmin kyseessä olevien aineistojen nominaalisia attribuutteja huomataan, että suurin osa niistä on binäärisiä. Vain ordinaalisilla, kvantitatiivisilla ja binäärisillä attribuuteilla kuvatussa aineistossa euklidinen etäisyysfunktio ja tyypillinen heterogeeninen etäisyysfunktio toimivat hyvin samalla tavalla. Osoitetaan seuraavassa, että tällaisissa aineistoissa euklidisen ja tyypillisten heterogeenisen etäisyysfunktion tulokset ovat samat, mikäli attribuutit normalisoidaan samalla tavalla.

Olkoon E joukko esimerkkejä, jotka on kuvattu attribuuteilla A. Joukko A koostuu binäärisistä (B) sekä ordinaalisista ja kvantitatiivisista (Q) attribuuteista ($A = B \cup Q$, $m = |A|$). Koska etäisyyslaskennan tulos ei riipu attribuuttien järjestyksestä, A voidaan järjestää siten, että $k = |B|$ ensimmäistä attribuuttia ovat binäärisiä ja $l = m - k = |Q|$ viimeistä attribuuttia kuuluvat joukkoon Q. Saadaan järjestetty joukko $A' = \{b_1, b_2, ..., b_k, q_1, q_2, ..., q_l\}$, missä $b_i \in B \ (i = 1,2,...,k)$ ja $q_j \in Q \ (j = 1,2,...,l)$. Nyt esimerkkien $x, y \in A'$ etäisyys voidaan laskea euklidisella etäisyydellä (D_E) ja tyypillisellä heterogeenisellä (D_H) etäisyysfunktiolla seuraavasti:

$$D_E(x, y) = \sqrt{\sum_{a=1}^{k} d_E(x_a, y_a) + \sum_{a=k+1}^{m} d_E(x_a, y_a)} \text{ ja}$$

$$D_H(x, y) = \sqrt{\sum_{a=1}^{k} d_H(x_a, y_a) + \sum_{a=k+1}^{m} d_E(x_a, y_a)},$$

missä $d_E(x_a, y_a) = (x_a - y_a)^2$ ja $d_E(x_a, y_a) = \begin{cases} 0, & \text{jos } x_a = y_a \\ 1, & \text{muuten.} \end{cases}$

Väite: $D_E(x, y) = D_H(x, y) \quad \forall x, y \in A'$.

Todistus: Tarkastellaan funktioiden monotonisia muunnoksia $D'_E = D_E^2$ ja $D'_H = D_H^2$. Nyt riittää osoittaa, että $d_E(x_a, y_a) = d_H(x_a, y_a)$, kun a = 1, …, k, sillä edellä olevissa kaavoissa jälkimmäiset summat ovat identtiset. Taulukossa 9.2 on kaikki mahdolliset arvot funktioille d_E ja d_H binäärisessä tilanteessa. Nyt

$$d_E(x_a, y_a) = d_H(x_a, y_a) \Leftrightarrow D'_E(x, y) = D'_H(x, y) \Leftrightarrow D_E(x, y) = D_H(x, y) \square$$

Taulukko 9.2. Funktioiden d_E ja d_H arvot binäärisessä tilanteessa.

x_i	y_i	d_E	d_H
0	0	0	0
0	1	1	1
1	0	1	1
1	1	0	0

Teoriassa euklidinen etäisyysfunktio on tyypillisiä heterogeenisiä funktioita huonompi käsittelemään monitasoisia nominaalisia attribuutteja, koska se käyttää nominaalisille attribuuteille sopimattomia operaatioita. Euklidisessa etäisyysfunktiossa nominaalisen attribuutin a arvojen x_a ja y_a välinen etäisyys määräytyy erotuksen $x_a - y_a$ perusteella, vaikka arvoilla ei ole järjestystä. Tuloksista (katso taulukot 8.10 ja 8.12) nähdään kuitenkin, että myös sellaisissa heterogeenisissä aineistoissa, joissa on monitasoisia nominaalisia attribuutteja, perinteisten heterogeenisten ja euklidisen etäisyysfunktion välillä ei ole merkittävää eroa edellisten hyväksi. Euklidinen etäisyysfunktio jopa osoittautuu joitain perinteisiä heterogeenisiä funktioita tarkemmaksi myös monitasoisia nominaalisia attribuutteja sisältävissä aineistoissa. Tähän voi olla monia syitä. Yksi syy on kenties nominaalisten attribuuttien vähäinen merkitys luokittelu kannalta. Kuitenkin nominaaliset attribuutit monimutkaisemmin käsittelevien HVDM- ja MHVDM-funktioiden muita funktioita merkittävästi parempien luokittelutulosten perusteella nominaalisilla attribuuteilla on merkitystä ainakin joissain aineistoissa. Toiseksi nominaalisten attribuuttien määrittelyjoukot eivät olleet kovin laajoja, jolloin aineistot eivät välttämättä olleet tarpeeksi haastavia euklidiselle etäisyydelle. Lisäksi useimmissa aineistoissa nominaalisten attribuuttien koodaukseen käytettiin peräkkäisiä kokonaislukuja, mikä oli luultavasti suotuisampaa euklidisen etäisyysfunktion kannalta kuin vaikka täysin satunnaisten lukujen käyttö.

Monitasoisten nominaalisten attribuuttien käsittely euklidisella etäisyysfunktiolla ei ole edellisestä huolimatta sopivaa, koska etäisyys riippuu täysin siitä, miten arvot on

koodattu ja koodausta muuttamalla saadaan erilaisia tuloksia. Näin ollen saatujen tulosten perusteella ei voida sanoa, että euklidinen etäisyysfunktio soveltuisi yleisesti ottaen hyvin monitasoiselle nominaaliselle datalle. Tulokset kertovat vain sen, että perinteisten heterogeenisten funktioiden tapa käsitellä nominaalisia attribuutteja on liian yksinkertainen näin moniulotteisissa ja sekamuotoisissa aineistoissa. Perinteisten heterogeenisten funktioiden luokittelutarkkuuksien välillä ei ollut merkitseviä eroja, vaikka joidenkin yksittäisten aineistojen kohdalla selviäkin eroja oli.

10. Yhteenveto

Tutkielmassa tutustuttiin etäisyys- ja samankaltaisuusfunktioihin, ja vertailtiin niitä luokitellen erityyppisiä aineistoja lähimmän naapurin menetelmällä. Testeissä käytettiin 36 aineistoa, joista 15 oli niin sanottuja heterogeenisiä aineistoja, jotka sisälsivät sekä kvantitatiivisia että nominaalisia attribuutteja. Vertailussa tarkasteltiin luokittelutarkkuutta sekä oikein menneitä positiivisia ennusteita luokittain. Tulokset voidaan tiivistää seuraaviin neljään kohtaan.

1. HVDM- ja sen tarkennus MHVDM-funktio luokittelivat keskimäärin tarkimmin.
2. Perinteisten heterogeenisten funktioiden (CSM, ER, Gower ja HEOM) välillä ei ollut merkittäviä eroja.
3. Euklidisen etäisyysfunktion ja perinteisten heterogeenisten funktioiden välillä ei ollut merkittäviä eroja luokittelutarkkuudessa.
4. Kosinietäisyysfunktiolla saatiin keskimäärin kaikkein huonoimpia luokittelutarkkuuksia.

Vaikka keskimääräisiä tarkkuuksia tarkastelemalla HVDM ja MHVDM suoriutuivatkin kaikista parhaiten, paikoitellen yksittäisten aineistojen tarkkuuksissa muut funktiot ovat näitä kahta edellä. Tästä syystä etäisyysfunktion valinta tulisikin riippua aina käsiteltävästä aineistosta. Tiedetään myös, että kaikkien mahdollisten aineistojen joukossa kaikki luokittelijat toimivat keskimäärin yhtä hyvin [Wilson and Martinez, 1997]. Aineiston ominaisuuksia tutkimalla voidaan arvioida, mitä etäisyysfunktiota tulisi käyttää. Pohdittava on ainakin, käyttääkö funktiota, joka käsittelee nominaaliset attribuutit eri tavalla kuin kvantitatiiviset. Saatujen tulosten perusteella nominaalisia attribuutteja sisältävään aineistoon kannattaa nyt käsitellyistä funktioista valita yleensä HVDM, vaikka tähänkin on olemassa poikkeuksia. Jos aineistossa on ordinaalisia attribuutteja, saattaa olla järkevää käyttää MHVDM-funktiota, jossa ordinaaliset attribuutit käsitellään nominaalisten tavoin.

Tutkielmassa tarkasteltiin ja vertailtiin etäisyysfunktioita vain lähimmän naapurin menetelmän osana. Tulevaisuudessa funktioita tulisi vertailla myös muunlaisissa sovelluksissa, esimerkiksi klusteroinnissa ja neuroverkkojen kanssa. Tuleviin tutkimuksiin jää myös sen pohtiminen, mistä johtuvat tiettyjen funktioiden saavuttamat yleisestä linjasta poikkeavat tarkkuudet tietyissä aineistoissa. Lisäksi olisi syytä perehtyä tarkemmin niihin tilanteisiin, joissa muuten parhaiten suoriutuneilla HVDM- ja MHVDM-funktioilla oli ongelmia ja tutkia voisiko niitä muuttaa tai tarkentaa jollain tavalla, jotta näissäkin tilanteissa luokittelutarkkuus paranisi. Näissä tutkimuksissa voidaan hyödyntää testiohjelman tulostamia yksityiskohtaisia raportteja luokittelun kulusta eri aineistojen kohdalla. Muita tutkielman ulkopuolelle jääneitä ja tulevaisuudessa mahdollisesti selvitettäviä asioita ovat esimerkiksi normalisointimenetelmän sekä puuttuvien tietojen

ja kohinan vaikutus läheisyysfunktioiden toimintaan. Lisäksi voitaisiin pohtia lähimmän naapurin menetelmän hyvyyttä sekamuotoisen datan luokittelijana muihin koneoppimismenetelmiin verrattuna.

Viiteluettelo

[Aha et al. 1991] David W. Aha, Dennis Kibler and Marc K. Albert, Instance-based learning algorithms. *Machine Learning*, **6** (1991), 37-66.

[Auramo et al., 1993] Yrjö Auramo, Martti Juhola and Ilmari Pyykkö, An expert system for the computer-aided diagnosis of dizziness and vertigo. *Medical Informatics* **18** (1993), 293-305.

[Blake and Merz, 1998] Catherine L. Blake and Christopher John Merz, *UCI Repository of Machine Learning Databases*, University of California, Irvine, Department of Information and Computer Science, (1998). Available as http://www.ics.uci.edu/~mlearn/MLRepository.html.

[Boberg, 1999] Jorma Boberg, *Cluster Analysis – A Mathematical Approach with Applications to Protein Structures*. Ph.D. Thesis, University of Turku, 1999.

[Brasil et al., 2001] Lourdes Mattos Brasil, Fernando Mendez de Azevedo and Jorge Muniz Barreto, A hybrid expert system for diagnosis of epileptic crisis. *Artificial Intelligence in Medicine* **21** (2001), 227-233.

[Braun, 1995] W. John Braun, An Illustration of Bootstrapping Using Video Lottery Terminal Data. *Journal of Statistics Education* **3**, 2 (1995).

[Dawson, 1995] Robert J. MacG. Dawson, The "Unusual Episode" data revisited. *Journal of Statistics Education* **3**, 3 (Nov. 1995).

[Doyle et al., 1995] Howard R. Doyle, Bambang Parmanto, Paul W. Munro, Ignazio R. Marino, L. Aldrighetti, C. Doria, J. McMichael and J. J. Fung, Building clinical classifiers using incomplete observations – a neural network ensemble for hepatoma detection detection in patients with cirrhosis. *Methods of Information in Medicine* **34** (1995), 253-258.

[Estabrook and Rogers, 1966] George F. Estabrook and David J. Rogers, A general method of taxonomic description for a computed similarity measure. *BioScience* (Nov. 1966), 789-793.

[Everitt et al., 2001] Brian S. Everitt, Sabine Landau and Morven Leese, *Cluster Analysis*. Arnold, London, 2001.

[Fleiss and Zubin, 1969] Joseph L. Fleiss and Joseph Zubin, On the methods and theory of clustering. *Multivariate Behavioral Research* (Apr. 1969), 235-250.

[Freund, 1971] John E. Freund, *Mathematical Statistics*, Prentice-Hall, Englewood Cliffs, New Jersey, 1971.

[Gower, 1971] John C. Gower, A general coefficient of similarity and some of its properties. *Biometrics* **27** (Dec. 1971) 857-874.

[Gower, 1985] John C. Gower, Measures of similarity, dissimilarity, and distance. In: S. Kotz, N.L. Johson and C.B Read (eds.), *Encyclopedia of Statistical Sciences* **5** (1985), Wiley, New York, 397-405.

[Gower and Legendre, 1986] John C. Gower and P. Legendre, Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification* **3** (1986), 5-48.

[Hand et al., 2001] David Hand, Heikki Mannila and Padhraic Smyth, *Principles of Data Mining*. MIT Press, 2001.

[Hochberg and Tamane, 1987] Yosef Hochberg and Ajit C. Tamane, *Multiple Comparison Procedures*, Wiley, New York, 1987.

[Hollmén et el., 2000] Jaakko Hollmén, Michal Skubacz and Michiaki Taniguchi, Input dependent misclassification costs cost-sensitive classifiers. In: *Second International Conference on Data Mining 2000* (Jul. 2000), WIT Press, 495-503.

[Hosmer and Lemeshow, 2000] David W. Hosmer, Jr. and Stanley Lemeshow, *Applied Logistic Regression*. Wiley, 2000.

[Ichino and Yaguchi, 1994] Manabu Ichino and Hiroyuki Yaguchi, Generalized Minkowski metrics for mixed feature-type data analysis, *IEEE Transactions on Systems, Man, and Cypernetics* **24**, 4 (Apr. 1994), 698-708.

[Jain et al., 1999] Anil K. Jain, M.N. Murty and Patrick J. Flynn, Data clustering: A review. *ACM Computing Surveys* **31**, 3 (Sep. 1999), 264-323.

[Jain and Dubes, 1988] Anil K. Jain and Richard C. Dubes, *Algorithms for Clustering Data*. Prentice-Hall, New Jersey, 1988.

[Juhola and Laurikkala, 2001] Martti Juhola and Jorma Laurikkala, On metricity of heterogeneous Euclidean-overlap metric and heterogeneous value difference metric with missing values. Manuscript.

[Kentala, 1996] Erna Kentala, Characteristics of six otologic diseases involving vertigo. *The American Journal of Otology*, **17** (1996), 883-892.

[Kubat et al., 1998] Miroslav Kubat, Robert C. Holte and Stan Matwin, Machine learning for the detection of oil spills in satellite radar images. *Machine Learning* **30** (1998), 195-215.

[Laurikkala et al., 2001] Jorma Laurikkala, Martti Juhola, Seppo Lammi, Jorma Penttinen and Pauliina Aukee, Analysis of the imputed female urinary incontinence data for the evaluation of expert system parameters. *Computers in Biology and Medicine* **31** (2001), 239-257.

[Laurikkala, 2001a] Jorma Laurikkala, *Knowledge Discovery for Female Urinary Incontinence Expert System*. Ph.D. Thesis, Department of Computer and Information Sciences, University of Tampere, 2001.

[Laurikkala, 2001b] Jorma Laurikkala, Improving identification of difficult small classes by balancing class distribution. In: Silvana Quaglini, Pedro Barahona and Steen Andreassen (eds.), *Artificial Intelligence in Medicine: Eight European Conference on Artificial Intelligence in Medicine in Europe, Lecture Notes in Artificial Intelligence* **2101** (2001), Springer, Berlin, 63-66.

[Lavrač, 1999] Nada Lavrač, Selected techniques for data mining in medicine. *Artificial Intelligence in Medicine* **16** (1999), 3-23.

[Legendre and Chodorowski, 1977] Pierre Legendre and Andrzej Chodorowski, A generalization of Jaccard's association coefficient for Q analysis of multi-state ecological data matrices. *Ekologia Polska* **25**, 2 (1977), 297-308.

[Little and Rubin, 1986] Roderick J. A. Little and Donald B. Rubin, *Statistical Analysis with Missing Data*. Wiley, New York, 1987.

[Long, 2001] William J. Long, Medical informatics: Reasoning methods. *Artificial Intelligence in Medicine* **23** (2001), 71-87.

[Mitchell, 1997] Tom M. Mitchell, *Machine Learning*. McGraw-Hill, New York, 1997.

[Mitchell, 1999] Tom M. Mitchell, Machine learning and data mining. *Communications of the ACM* **42**, 11 (Nov. 1999), 30-36.

[Montani et al., 2000] Stefania Montani, Riccardo Bellazzi, Luigi Portinale and Mario Stefanelli, A multi-modal reasoning methodology for managing IDDM patients. *International Journal of Medical Informatics* **58**–**59** (2000), 243–256.

[Pett, 1997] Marjorie A. Pett: *Nonparametric Statistics for Health Care Research: Statistics for Small Samples and Unusual Distributions*, SAGE Publications, Thousand Oaks, 1997.

[Pesonen et al. 1994] Erkki Pesonen, J. Ikonen, Martti Juhola and Matti Eskelinen, Parameters for a knowledge base for accute appendicitis. *Methods of Information in Medicine* **33** (1994), 220-226.

[Puntanen, 1998] Simo Puntanen, *Matriiseja tilastotieteilijälle*. Matemaattisten tieteiden laitos, Tampereen yliopisto, 1998.

[Quinlan, 1986] J. Ross Quinlan, Induction of decision trees. *Machine Learning* **1** (1986), 81-106.

[Rasmussen et al. 1999] Finn Rasmussen, Malin Johansson and Hans Ole Hansen, Trends in overweight and obesity among 18-year-old males in Sweden between 1971 and 1995. *Acta Paediatrica* **88** (1999), 431-437.

[Rodríguez, 2003] Germán Rodríguez, Datasets used in the course Generalized linear models, University of Princetown, 2003. Available as: http://data.princeton.edu/wws509/datasets/default.htm.

[Rogers and Tanimoto, 1960] D. J. Rogers and T. T. Tanimoto, A computer program for classifying plants. *Science* **132** (1960), 1115-1118.

[Salton, 1989] Gerald Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, 1989.

[Salzberg, 1999] Steven L. Salzberg, On comparing classifiers: A critique of current research and methods, *Data Mining and Knowledge Discovery*, **1** (1999), 1-12.

[Schaffer, 1993] Cullen Schaffer, Selecting a Classification Method by Cross-Validation. *Machine Learning* **13**, 1 (Jan. 1993) 135-143.

[Sharma, 1996] Subhash Sharma, *Applied Multivariate Techniques*. Wiley, New York, 1996.

[Skubacz and Hollmén, 2000] Michal Skubacz and Jaakko Hollmén, Quantization of continuous input variables for binary classification. In: K. S. Leung, L.-W. Chan, and H. Meng (eds.), *Proceedings of the Second International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'2000), Lecture Notes in Computer Science* **1983** (2000), Springer, 42-47.

[Smith et al., 2003] A. E. Smith, Christopher D. Nugent, Sally I. McClean, Evaluation of inherent performance of intelligent medical decision support systems: utilising neural networks as an example. *Artificial Intelligence in Medicine* **27**, 1 (2003), 1-27.

[Späth, 1980] Helmuth Späth, *Cluster Analysis Algorithms for Data Reduction and Classification of Objects*. Ellis Horwood Publishers, 1980.

[Stanfill and Waltz, 1986] Craig Stanfill and David Waltz, Toward memory-based reasoning. *Communications of the ACM* **29**, 12 (Dec. 1986), 1213-1228.

[Vesanto and Hollmén, 2002] Juha Vesanto and Jaakko Hollmén, An automated report generation tool for the data understanding phase. In: Ajith Abraham, Lakhmi C. Jain and Janusz Kacprzyk (eds.), *Recent Advances in Intelligent Paradigms and Applications, Studies in Fuzziness and Soft Computing* **113** (2002), Springer.

[Viikki et al., 2002] Kati Viikki, Erna Kentala, Martti Juhola, Ilmari Pyykkö and Pekka Honkavaara. Generating decision trees from otoneurological data with a variable-grouping method, *Journal of Medical Systems*, **26** (2002), 415-425.

[Vlachos, 2003] Pantelis Vlachos, *StatLib*, Carnegie Mellon University, Pittsburgh, Department of Statistics, 2003. Available as: http://lib.stat.cmu.edu/index.php.

[Wilson and Martinez, 1997] D. Randall Wilson and Tony R. Martinez, Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research* **6**, 1 (1997), 1-34.

Liite

Tietoja aineistoista

Seuraavassa annetaan lyhyet kuvaukset testeissä käytetyistä aineistoista. Kuvauksissa mainitaan datan nimi, kuvaus ja luokitteluongelma. Lisäksi mainitaan datan lähdetiedot, mahdollinen internet-osoite sekä alkuperäiseen dataan mahdollisesti tehdyt muunnokset.

1.
Nimi: Alokkaat
Kuvaus: Taudinmääritysdataa armeijan kutsunnoissa käyneistä henkilöistä.
Luokitteluongelma: Onko henkilöllä kuulovauriota.
Lähde: [Rasmussen et al. 1999]

2.
Nimi: Assistentit
Kuvaus: Tietoja assistentin opetuksesta.
Luokitteluongelma: Assistentin opetuskyky.
Lähde: [Blake and Merz, 1998]
Url: ftp://ftp.ics.uci.edu/pub/machine-learning-databases/tae

3.
Nimi: Australia
Kuvaus: Australialaisten luotonhakijoiden henkilötietoja.
Luokitteluongelma: Onko henkilö luottokelpoinen.
Lähde: [Blake and Merz, 1998]
Url: ftp://ftp.ics.uci.edu/pub/machine-learning-databases/statlog/australian

4.
Nimi: Cleveland
Kuvaus: Diagnoosidataa sydänpotilaista
Luokitteluongelma: Onko henkilöllä sydänsairautta.
Lähde: [Blake and Merz, 1998]
Url: ftp://ftp.ics.uci.edu/pub/machine-learning-databases/heart-disease
Muunnokset: Puuttuvat arvot imputoitu luokkamoodeilla ja -mediaaneilla.

5.
Nimi: Eturauhassyöpä
Kuvaus: Tietoa eturauhassyöpäpotilaista.
Luokitteluongelma: Taudin eteneminen.
Lähde: [Hosmer and Lemeshow, 2000]
Url: http://www-unix.oit.umass.edu/~statdata/data/pros.dat
Muunnokset: Puuttuvat tiedot imputoitu luokkamoodeilla ja -mediaaneilla.

6.
Nimi: Guatemala
Kuvaus: Dataa Guatemalan maaseudulla synnyttaneistä äideistä.
Luokitteluongelma: Terveyspalvelujen tarjoaja.
Lähde: [Rodríguez, 2003]
Url: http://data.princeton.edu/wws509/datasets/default.htm#healthCare

7.
Nimi: Hedelmäpeli
Kuvaus: Hedelmäpelin pelitilanteita.
Luokitteluongelma: Voittokoodi.
Lähde: [Braun, 1995]
Url: http://www.amstat.org/publications/jse/datasets/vlt.txt

8.
Nimi: Huimaus
Kuvaus: Dataa erilaisista huimausoireista kärsivistä potilaista.
Luokitteluongelma: Kuulohäiriön tyyppi.
Lähde: [Kentala, 1996]
Muunnokset: Puuttuvat tiedot imputoitu luokkamediaaneilla ja -moodeilla.

9.
Nimi: ICU
Kuvaus: 200 tapauksen otos tutkimuksesta ICU-yksikössä olleista potilaista.
Luokitteluongelma: Henkilön vitaliteetti.
Lähde: [Vlachos, 2003]
Url: http://lib.stat.cmu.edu/DASL/Datafiles/ICU.html

10.
Nimi: Ihotauti
Kuvaus: Taudinmääritysdataa eryteemisiä ihotauteja sairastavista potilaista.
Luokitteluongelma: Ihotaudin tyyppi.
Lähde: [Blake and Merz, 1998]
Muunnokset: Muutama puuttuva tieto imputoitu koko aineiston mediaanilla.

11.
Nimi: Inkontinenssi
Kuvaus: Taudinmääritysdataa virtsan inkontinenssista kärsivistä naisista.
Luokitteluongelma: Inkontinenssin aiheuttaja.
Lähde: [Laurikkala et al., 2001]
Muunnokset: Puuttuvat tiedot imputoitu luokkamoodeilla- ja mediaaneilla.

12.
Nimi: Kööpenhamina
Kuvaus: Aineisto kuvaa asunto-oloja Kööpenhaminassa.
Luokitteluongelma: Naapurikontaktien lukumäärä.
Lähde: [Rodríguez, 2003]
Url: http://data.princeton.edu/wws509/datasets/default.htm#copen

13.
Nimi: Lasit
Kuvaus: Rikospaikkojen lasinytteistä koottua dataa.
Luokitteluongelma: Lasin tyyppi.
Lähde: [Blake and Merz, 1998]
Url: ftp://ftp.ics.uci.edu/pub/machine-learning-databases/glass

14.
Nimi: Led+kohina
Kuvaus: Kuvauksia 7-valoisesta led-näytöstä. Mukana 17 kohina-attribuuttia.
Luokitteluongelma: Ledien muodostama numero.
Lähde: [Blake and Merz, 1998]
Url: ftp://ftp.ics.uci.edu/pub/machine-learning-databases/led-display-creator

15.
Nimi: Led
Kuvaus: Datassa on kuvauksia 7-valoisesta led-näytöstä. Datan 7 attribuuttia sisältävät kukin 10 % kohinaa.
Luokitteluongelma: Ledien muodostama numero.
Lähde: [Blake and Merz, 1998]
Url: ftp://ftp.ics.uci.edu/pub/machine-learning-databases/led-display-creator

16.
Nimi: Liput
Kuvaus: Informaatiota eri maiden lipuista.
Luokitteluongelma: Alakulman kuvio.
Lähde: [Blake and Merz, 1998]
Url: ftp://ftp.ics.uci.edu/pub/machine-learning-databases/flags

17.
Nimi: Ljubljana
Kuvaus: Diagnoosidataa liittyen rintasyöpään.
Luokitteluongelma: Onko henkilöllä rintasyöpää.
Lähde: [Blake and Merz, 1998]
Muunnokset: Puuttuvat arvot imputoitu luokkamoodeilla.

18.
Nimi: Luotonhakijat
Kuvaus: Tietoa saksalaisista luotonhakijoista.
Luokitteluongelma: Henkilön luottokelpoisuus.
Lähde: [Blake and Merz, 1998]
Url: ftp://ftp.ics.uci.edu/pub/machine-learning-databases/statlog/german

19.
Nimi: Mahatauti
Kuvaus: Diagnoosidataa erilaisia mahatauteja sairastavista potilaista.
Luokitteluongelma: Mahataudin tyyppi.
Lähde: [Pesonen et al., 1994]
Muunnokset: Puuttuvat arvot imputoitu luokkamoodeilla ja mediaaneilla.

20.
Nimi: Munkit1
Kuvaus: Synteettinen aineisto. Jakautuu tiettyjen attribuuttien ja arvokombinaatioiden mukaan kahteen luokkaan. Loput atribuutit ovat vaikeuttamassa luokittelua.
Luokitteluongelma: Keinotekoinen.
Lähde: [Blake and Merz, 1998]
Url: ftp://ftp.ics.uci.edu/pub/machine-learning-databases/monks-problems

21.
Nimi: Munkit2
Kuvaus: Synteettinen aineisto. Jakautuu tiettyjen attribuuttien mukaan kahteen luokkaan. Loput atribuutit ovat vaikeuttamassa luokittelua.
Luokitteluongelma: Keinotekoinen.
Lähde: [Blake and Merz, 1998]
Url: ftp://ftp.ics.uci.edu/pub/machine-learning-databases/monks-problems

22.
Nimi: Munkit3
Kuvaus: Synteettinen aineisto. Jakautuu tiettyjen attribuuttien mukaan kahteen luokkaan. Loput atribuutit ovat vaikeuttamassa luokittelua. Luokka-attribuutilla 5% kohinaa.
Luokitteluongelma: Keinotekoinen.
Lähde: [Blake and Merz, 1998]
Url: ftp://ftp.ics.uci.edu/pub/machine-learning-databases/monks-problems

23.
Nimi: Palkat
Kuvaus: Tietoja palkansaajista.
Luokitteluongelma: Sukupuoli.
Lähde: [Vlachos, 2003]
Url: http://lib.stat.cmu.edu/datasets/CPS_85_Wages

24.
Nimi: Plasebo
Kuvaus: Dataa potilaista, jotka ovat läpikäyneet jonkin leikkauksen ja joilla on taipumusta pahoinvointiin. Pahoinvointia koitettu ehkäistä jollain lumelääkkeellä.
Luokitteluongelma: Onko tarvetta pelastuslääkkeeseen.
Lähde: [Viikki et al., 2002]
Muunnokset: Poistettu sellaiset attribuutit, joilla puuttuvia arvoja.

25.
Nimi: Profylaksi
Kuvaus: Dataa potilaista, jotka ovat läpikäyneet jonkin leikkauksen ja joilla on taipumusta pahoinvointiin. Pahoinvointia koitettu ehkäistä profylaksilla.
Luokitteluongelma: Onko tarvetta pelastuslääkkeeseen.
Lähde: [Viikki et al., 2002]
Muunnokset: Poistettu sellaiset attribuutit, joilla puuttuvia arvoja.

26.
Nimi: Promoottorit
Kuvaus: Data kuvaa dna-sekvenssejä.
Luokitteluongelma: Näkyykö sekvenssissä promoottoriaktiivisuutta.
Lähde: [Blake and Merz, 1998]

27.
Nimi: Ristinolla
Kuvaus: Ristinollan voitokkaista lopputilanteita
Luokitteluongelma: Pelin lopputulos.
Lähde: [Blake and Merz, 1998]
Url: ftp://ftp.ics.uci.edu/pub/machine-learning-databases/tic-tac-toe

28.
Nimi: Rytmihäiriö
Kuvaus: Henkilön ECG- ja joistain muista tiedoista koottua taudinmääritysdataa.
Luokitteluongelma: Sydämen rytmihäiriön tyyppi.
Lähde: [Blake and Merz, 1998]
Url: ftp://ftp.ics.uci.edu/pub/machine-learning-databases/arrhythmia/
Muunnokset: Poistettu attribuutti, jonka arvoista puuttui 83 prosenttia. Lisäksi muutama puuttuva arvo imputoitu koko aineiston mediaanilla.

29.
Nimi: Shakki
Kuvaus: Aineistoon kuvattu shakin pelitilanteita, jossa jäljellä valkoinen kuningas ja torni sekä musta kuningas.
Luokitteluongelma: Pelin lopputilanne.
Lähde: [Blake and Merz, 1998]
Url: ftp://ftp.ics.uci.edu/pub/machine-learning-databases/chess/king-rook-vs-king

30.
Nimi: Sillat
Kuvaus: Aineisto kuvaa Pittsburghilaisia siltoja.
Luokitteluongelma: Sillan tyyppi.
Lähde: [Blake and Merz, 1998]
Url: ftp://ftp.ics.uci.edu/pub/machine-learning-databases/bridges
Muunnokset: Poistettu yhden marginaalisen luokan esimerkit sekä esimerkit, joilla ei ollut luokkatietoja. Puuttuvat arvot imputoitu luokkamoodeilla ja -mediaaneilla.

31.
Nimi: SPECT
Kuvaus: Aineisto on koottu SPECT-kuvista, joilla tutkitaan sydäntä.
Luokitteluongelma: Onko sydänvikaa.
Lähde: [Blake and Merz, 1998]
Url: ftp://ftp.ics.uci.edu/pub/machine-learning-databases/spect

32.
Nimi: Titanic
Kuvaus: Tietoja Titanic-laivan tuhossa selvinneistä ja menehtyneistä.
Luokitteluongelma: Selviytyikö onnettomuudesta.
Lähde: [Dawson, 1995]
Url: http://www.amstat.org/publications/jse/v3n3/datasets.dawson.html

33.
Nimi: Tyreoosi
Kuvaus: Diagnoosidataa hypotyreoottisista potilaista.
Luokitteluongelma: Onko potilas hypotyreoottinen.
Lähde: [Blake and Merz, 1998]
Url: ftp://ftp.ics.uci.edu/pub/machine-learning-databases/thyroid-disease

34.
Nimi: USA
Kuvaus: Otos USA:n väestönlaskudatasta vuodelta 1994.
Luokitteluongelma: Henkilön vuositulot.
Lähde: [Blake and Merz, 1998]
Url: ftp://ftp.ics.uci.edu/pub/machine-learning-databases/adult
Muunnokset: Poistettu esimerkit, joilla puuttuvia arvoja.

35.
Nimi: WDBC
Kuvaus: Diagnoosidataa rintasyöpäpotilaista.
Luokitteluongelma: Onko syöpäkasvain hyvä- vai huonolaatuinen.
Lähde: [Blake and Merz, 1998]

36.
Nimi: WPBC
Kuvaus: Prognoosidataa rintasyöpäpotilaista.
Luokitteluongelma: Onko syöpä uusiutunut vai ei.
Lähde: [Blake and Merz, 1998]
Muunnokset: Muutama puuttuva kvantitatiivinen tieto korvattu attribuutin mediaanilla koko aineistossa.