LAPPEENRANNAN-LAHDEN TEKNILLINEN YLIOPISTO LUT
School of Engineering Science
Laskennallisen tekniikan ja analytiikan koulutusohjelma
Kandidaatintyö

Emmi Huovila

Entropiaan ja samankaltaisuuteen perustuva muuttujan valintamenetelmä markkinointiin liittyvässä luokitteluongelmassa

Ohjaaja: TkT Pasi Luukka

TIIVISTELMÄ

Lappeenrannan teknillinen yliopisto

School of Engineering Science

Laskennallisen tekniikan koulutusohjelma

Emmi Huovila

Entropiaan ja samankaltaisuuteen perustuva muuttujan valintamenetelmä markkinointiin liittyvässä luokitteluongelmassa

Kandidaatintyö

2019

24 sivua, 3 kuvaa, 3 taulukkoa, 1 liite

Ohjaaja: TkT Pasi Luukka

Avainsanat: muuttujan valinta; luokittelu; entropia; samankaltaisuus

Kandidaatintyössä tavoitteena on tutustua entropiaan ja samankaltaisuuteen perustuvaan muuttujan valintamenetelmään sekä samankaltaisuuteen perustuvaan luokittelumenetelmään. Muuttujan valinnan tavoitteena on rajata datasta pois turhat tai vaikuttamattomat muuttujat. Luokittelussa luokitin opetetaan harjoitusdatalla ja testausdatalla testataan luokittimen toimivuus. Molempien menetelmien taustalla oleva teoria esitellään ja menetelmät käydään läpi esimerkein. Teoriaosuudessa esitellään myös luokitteluun liittyvä ristiinvalidointi sekä tulosten analysointiin käytettävät mittarit: herkkyys ja spesifisyys.

Menetelmiä käytetään luomaan ennustemalli markkinoinnissa kerätystä aineistosta. Tulokset esitellään kuvin ja taulukoin. Lopuksi analysoidaan muuttujien valinnan ja luokittelun onnistumista.

Sisältö

Symboli- ja lyhenneluettelo

A	joukko
C	luokka
f_A	jäsenyysfunktio
f_n	muuttuja datassa
H	sumean entropian arvo (fuzzy entropy)
i	luokkien määrä
j	näytteen numero datassa
J	luokan pituus
m	keskiarvo
n	muuttujien määrä
N	määrä
O	näyte datassa
p	vakiotermi similaarisuuden kaavassa
S	samankaltaisuusarvo
μ_A	fuzziness, kuuluvuus joukkoon A
w	paino (weight)
x	funktioon syötettävä muuttuja

Mittareihin liittyvät lyhenteet

FN	väärä negatiivinen, false negative
FP	väärä positiivinen, false positive
N	negatiiviset, negatives
P	positiiviset, positives
TN	oikea negatiivinen, true negative
TNR	spesifisyys, true negative rate
TP	oikea positiivinen, true positive
TPR	herkkyys, true positive rate

1 JOHDANTO

Työn tavoitteena on luoda ennustemalli markkinointidataan perustuen. Ennustemallissa poistetaan ylimääräiset muuttujat, joilla ei ole merkittävästi vaikutusta lopputulokseen. Ennustetta markkinoinnin onnistumisesta tarvitaan, jotta voidaan kohdentaa resursseja oikeaan paikkaan ja parantaa markkinointitulosta.

1.1 Tausta

Nykypäivänä kaikista asioista, kuten myös markkinoinnista, saadaan kerättyä paljon dataa. Jotta datan keräämisestä olisi hyötyä, kerätty data tulisi saada hyödynnettyä jotenkin. Markkinoinnissa datan hyödyntäminen tarkoittaa, että saadaan tuotetta tai palvelua markkinoitua mahdollisimman tehokkaasti ja taloudellisesti. On turhaa ja resursseja kuluttavaa markkinoida sellaiselle kohderyhmälle, joka ei ole kyseisestä tuotteesta tai palvelusta kiinnostunut.

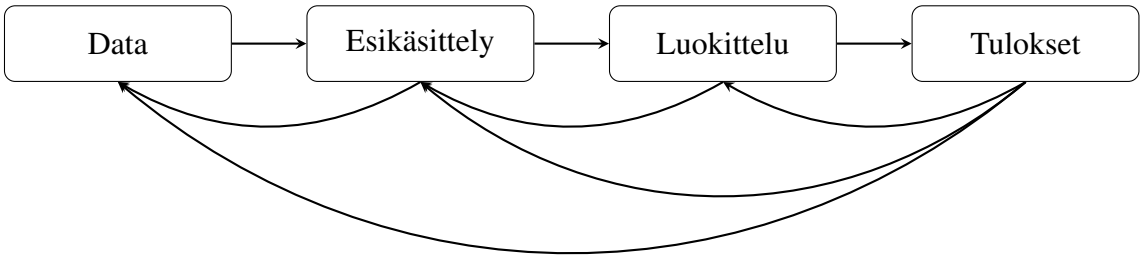
Markkinointi jakautuu kahteen osaan: massamarkkinointi ja suoramarkkinointi. Massamarkkinoinnissa mainontavälineitä ovat yleiset joukkotiedotusvälineet, kuten televisio, radio ja sanomalehdet. Suoramarkkinoinnissa markkinointi on kohdennettu tietyille ihmisille suuren väkijoukon sijaan. Markkinoinnissa käytetään suoria kanavia, kuten sähköposti- ja tekstiviestit sekä puhelut. Suoramarkkinoinnissa henkilöt, joille markkinoidaan, valitaan sen mukaan, että heidän ominaisuudet ja tarpeet vastaisivat sellaista, joka on potentiaalinen asiakas. Koska markkinointi on koko ajan kilpailullisempaa, massamarkkinoinnin tehokkuus on heikentynyt. Suoramarkkinoinnilla voidaan saada parempi prosenttiosuus tuotteen ostaneista suhteessa kaikkiin, joille markkinointi on kohdistunut kuin massamarkkinoinnilla. Nykyisin yrityksillä on paljon tietoa asiakkaista, jota voidaan käyttää ennusteen mallintamiseen, jolla voidaan markkinoida tuotetta tai palvelua oikeille ihmisille, jolloin markkinointi on tehokkaampaa, sillä markkinointiin ei mene ylimääräisiä resursseja, jos markkinoidaan ihmisille, joista ei todennäköisesti tule asiakkaita. [1]

1.2 Tutkimusongelma, tavoitteet ja rajaus

Tässä kandidaatintyössä tavoitteena on luoda ennustemalli markkinoinnissa kerätystä datasta sekä tutustua menetelmiin, joilla esikäsittely ja mallin luokittelu tehdään. Ongelma on rajattu tiettyyn dataan sekä tiettyihin menetelmiin. Tutkimusaineistona käytettiin tutkimuskäyttöön avointa Bank Marketing -aineistoa [2]. Muuttujan valintamenetelmäksi valittiin entropiaan

ja samankaltaisuuteen perustuva menetelmä. Luokittelussa käytettiin samankaltaisuusluokitinta.

Ongelmanratkaisu etenee Kuvan 1 mukaisesti. Ratkaisussa lähdetään liikkeelle datasta. Tarvittaessa dataa muokataan paremmin käsiteltävään muotoon, esimerkiksi tekstimuotoiset muuttujat numeroiksi. Seuraava vaihe on esikäsittely. Esikäsittely voi olla muuttujan valinta (feature selection), jossa datasta poistetaan muuttujat, jotka ovat merkityksettömiä tai tarpeettomia, tai muuttujan poistaminen (feature extraction), jossa muuttujia yhdistellään uusiksi muuttujiksi, jolloin muuttujien määrä vähenee [3]. Tässä kandidaatintyössä keskitytään muuttujan valintaan. Muuttujan valinnan jälkeen tehdään luokittelu. Luokittelusta saadaan tulokset, joiden perusteella voidaan analysoida esikäsittelyn onnistumista. Jokaisesta vaiheesta voidaan aina tarvittaessa palata takaisin mihin tahansa vaiheeseen sen mukaan, mitä on tarpeen tehdä. Vaiheita toistetaan kunnes ollaan tyytyväisiä lopputuloksiin.



Kuva 1. Ongelman ratkaisun vaiheet

2 KIRJALLISUUSKATSAUS

Dataa kerättäessä pyritään huomioimaan kaikki mahdolliset muuttujat, jotka voivat vaikuttaa kyseiseen tilanteeseen. Itse datan keräämisessä kerätään tiedot kaikista mahdollisista muuttujista, joita pidetään olennaisina ja jotka voivat vaikuttaa malliin. Data voi olla eri muotoista. Tutuin on niin sanottu tavallinen data (conventional data), joka on mahdollista ilmoittaa matriisimuodossa tai esimerkiksi puurakenteena, ja muuttujat ovat yhtenäisiä. Kokonaisuudessaan aineistot jakautuvat staattiseen dataan (static data) ja jatkuvasti päivittyvään dataan (streaming data). Staattinen data pysyy muuttumattomana koko käsittelyn ajan. Staattinen data jakautuu vielä tavalliseen dataan ja heterogeeniseen dataan (heterogeneus data). Staattisessa datassa voi olla yksinkertaiset muuttujat (flat features) tai rakenteiset muuttujat (structured features), esimerkiksi puurakenne. Heterogeenisessa datassa muuttujat voivat poiketa toisistaan, esimerkiksi samassa datassa voi olla kuva- sekä tekstimuotoisia muuttujia. Jatkuvasti päivittyvässä datassa aineisto muuttuu koko ajan, eikä voida etukäteen tietää, kuinka suuri data on tai onko se ääretön. Jatkuvasti päivittyvässä datassa muuttujan valintamenetelmäksi ehdotetaan yhdellä ajolla suoritettavaa menetelmää ja valintamenetelmän tavoitteena on määrittää, pitäisikö dataan lisätä uusi muuttuja tai poistaa vanhentunut muuttuja. [4]

Tavallisessa datassa kerätystä datasta mallinnetaan ennustemalli $y = f(x_1, \ldots, x_n)$, jossa riippumattomilla (selittävillä) muuttujilla $x_1, \ldots, x_n$ ennustetaan riippuvaa (selitettävää) muuttujaa y. Diskreetissä datassa ennuste tehdään luokittelumallin avulla ja jatkuvassa datassa regressiomallilla. Muuttujan valintaa käytetään siihen, että saadaan eroteltua ne muuttujat, jotka oikeasti vaikuttavat merkitsevästi malliin ja voidaan jättää huomiotta ne muuttujat, jotka eivät olennaisesti vaikuta malliin tai ovat korreloituneet muiden muuttujien kanssa.

Aineiston tyypin perusteella valitaan soveltuva muuttujan valintamenetelmä. Muuttujan valinnan tavoitteena on pienentää datan dimensiota. Muuttujan valintamenetelmät voidaan jakaa kolmeen ryhmään: suodatus- (filter), käärintä- (wrapped) ja sulautusmenetelmät (embedded). Tyypillinen suodatusmenetelmä koostuu kahdesta vaiheesta. Ensin lasketaan muuttujan tärkeys, joko yksilöllisesti tai joukoissa. Vähiten tärkeimmät muuttujat poistetaan. Suodatusmenetelmät ovat laskennallisesti tehokkaampia kuin käärintämenetelmät, mutta koska suodatusmenetelmät eivät ole yhteydessä oppimismenetelmään, valitut muuttujat eivät välttämättä ole optimaaliset valitulle oppimismenetelmälle. Käärintämenetelmä on myös kaksivaiheinen, mutta se hyödyntää oppimisalgoritmia toisin kuin suodatusmenetelmä. Ensimmäisessä vaiheessa etsitään alaryhmä muuttujien joukosta ja toisessa vaiheessa arvioidaan oppimismenetelmän avulla muuttujien ennustavuutta. Vaiheita toistetaan lopetusehtoon asti. Käärintämenetelmät eivät sovellu korkeadimensioisille aineistoille, sillä hakutilavuus n muuttujalle on 2^n. Sulautusmenetelmä yhdistää suodatus- ja käärintämenetelmän. Sulautus-

menetelmä käyttää oppimisalgoritmia muuttujienvalintaan, mutta se on kuitenkin laskennallisesti paljon tehokkaampi kuin käärintämenetelmä. Yleisimmin käytettyjen sulautusmenetelmien tavoitteena on samanaikaisesti sovittaa malli minimoimalla virheet ja pakottaa muuttujien kertoimet pieniksi. [4]–[6]

Tavalliselle datalle soveltuvat muuttujan valintamenetelmät voidaan ryhmitellä samankaltaisuuteen (similarity), informaatioteoriaan (information theoretical), harvaan oppimiseen (sparse learning), tilastotieteeseen (statistical) ja muihin menetelmiin perustuviin menetelmiin. Ohjatussa oppimisessa datan samankaltaisuus johdetaan luokkatiedosta ja ohjaamattomassa oppimisessa hyödynnetään etäisyysmittareita. Samankaltaisuusmetodit ovat erinomaisia ja yksinkertaisia sekä ohjatun että ohjaamattoman oppimisen ongelmiin. Samankaltaisuusmenetelmien huonona puolena on, että ne eivät kykene reagoimaan muuttujien päällekkäi-syyksiin. Informaatioteoreettisissa menetelmissä pyritään maksimoimaan muuttujien merkitys sekä minimoimaan muuttujien päällekkäisyys. Menetelmät soveltuvat vain diskreeteille datoille. Informaatioteoriaan perustuvan menetelmän etuna verrattuna samankaltaisuus menetelmään on, että se kykenee myös huomioimaan muuttujien päällekkäisyyden, mutta haittana on, että monet informaatioteoriamenetelmät soveltuvat vain ohjattuun oppimiseen. Harvassa oppimisessa minimoidaan sovitusvirhettä ehtojen avulla ja pakotetaan muuttujien kertoimet pieniksi, jolloin ne voidaan poistaa. Harvan oppimisen suosio on lisääntynyt, sillä se sulauttaa muuttujan valinnan tyypilliseen oppimisalgoritmiin, esimerkiksi lineaarinen regressio. Kuitenkin harvassa oppimisessa on vielä useampia ongelmia, jotta se toimisi hyvin kaikissa tapauksissa. Tilastotieteeseen perustuvat menetelmät ovat usein suodatusmenetelmiä, sillä ne mittaavat muuttujien merkitystä tilastotieteen mittarien avulla. Tilastotieteen menetelmät ovat yksinkertaisia ja niiden laskennalliset kustannukset ovat matalia, mutta ne eivät huomioi muuttujien päällekkäisyyksiä ja ne toimivat vain diskreeteille aineistoille. [4]

Tässä kandidaatintyössä käytettyä dataa on käytetty myös useissa muissa julkaisuissa. 11 muuttujalla ja 1000 näytteellä tehdyssä tutkimuksessa [7] käytettiin 4 eri luokittelumenetelmää. Käytetyt menetelmät olivat päätöspuu (decision tree), naiivi Bayes (naive Bayes), tukivektorikone (support vector machine) ja perceptron-neuroverkko (perceptron neural network). Näillä menetelmillä herkkyydet vaihtelivat välillä 94-97%, spesifisyys 95-98% ja tarkkuus 79-90%.

2.1 Entropiaan ja samankaltaisuuteen perustuva muuttujan valintamenetelmä

Seuraavassa on esitelty taustateoriaa entropiaan ja samankaltaisuuteen perustuvan muuttujan valintamenetelmään. Teorian jälkeen menetelmä esitellään vaiheittain ja käydään samanaikaisesti läpi esimerkin avulla.

2.1.1 Sumea joukko ja entropia

Sumean joukon (fuzzy set) on määritellyt L. A. Zadeh vuonna 1965. Klassisessa matematiikassa luokittelu on määritelty siten, että x joko kuuluu tai ei kuulu joukkoon A. Sumeassa joukossa joukkoon kuulumisen raja ei ole näin tarkka. Arvon x kuulumista joukkoon A voidaan kuvata jäsenyysfuntiolla (membership function) $f_A(x)$. Klassisessa matematiikassa jäsenyysfunktio saa joko arvon 1 ($x \in A$) tai arvon 0 ($x \notin A$). Sumeassa joukossa myös jäsenyysfunktion arvot, eli joukkoon kuulumisen aste, välillä [0,1] ovat mahdollisia. [8]

Esimerkki 1. Joukkoon A kuuluvat paljon lukua 1 suuremmat luvut. Tällöin sumean joukon teoria antaa jäsenyysfunktioille arvot: [8]

$f_A(0) = 0$

$f_A(1) = 0$

$f_A(5) = 0.01$

$f_A(10) = 0.2$

$f_A(100) = 0.95$

$f_A(500) = 1.$

Entropia kertoo informaation määrän datassa. Entropia toimii mittarina mittaamaan informaatiota, valintaa ja epävarmuutta. Shannon on määritellyt entropian kaavan

$$H = -K \sum_{i=1}^{N} q_k \log q_k, \tag{1}$$

jossa q_k on tapahtuman k todennäköisyys ja K vakio. Kahden vaihtoehdon tapauksessa entropia voidaan määritellä kaavalla

$$H = -(q \log q + (1-q) \log(1-q)), \tag{2}$$

jossa q on tapahtuman 1 todennäköisyys ja $1-q$ on tapahtuman 2 todennäköisyys. [9]

2.1.2 Menetelmän kuvaus vaiheittain ja esimerkki

Sumean entropian (fuzzy entropy) muuttujan valintamenetelmä voidaan jakaa viiteen vaiheeseen. Datassa ensimmäisissä sarakkeissa on eri muuttujia $f_1, f_2, \ldots, f_n$ ja viimeisessä sarakkeessa on näytteen luokkatieto (label) C.

Esimerkki 2. Muuttujan valintamenetelmä

f_1	f_2	f_3	C
3.5	5	1.1	1
3.7	0.1	1.2	1
3.9	3	0.8	1
1.3	0.7	2.8	2
2	4	3	2
1.6	6	3.3	2

Ensimmäisenä data normalisoidaan eli skaalataan välille [0,1]. Normalisointi voidaan tehdä esimerkiksi min-max-normalisoinnilla, jossa kaavalla

$$d' = \frac{(d - min(a)) \cdot (max_{new}(a) - min_{new}(a))}{max(a) - min(a)} + min_{new}(a) \tag{3}$$

lasketaan joukosta a arvosta d uusi arvo d' välille $[min_{new}(a), max_{new}(a)]$, jossa $min_{new}(a) = 0$ ja $max_{new}(a) = 1$. Tästä saadaan kaavalle yksinkertaisempi muoto

$$d' = \frac{d - min(a)}{max(a) - min(a)}.[10] \tag{4}$$

Esimerkki 2. jatkuu

f_1	f_2	f_3	C
0.8462	0.8305	0.1200	1
0.9231	0	0.1600	1
1.0000	0.4915	0	1
0	0.1017	0.8000	2
0.2692	0.6610	0.8800	2
0.1154	1.0000	1.0000	2

Seuraavaksi jokaiselle luokalle C_i lasketaan keskiarvovektori m_i:

$$m_i = [m_{1i}, m_{2i}, \ldots, m_{ni}], \tag{5}$$

$$m_{ni} = \frac{1}{J} \sum_{j=1}^{J} f_{nj}. \tag{6}$$

Esimerkki 2. jatkuu

Esimerkissä luokkia on kaksi, joten lasketaan kaksi keskiarvovektoria m_1 ja m_2.

$$m_{11} = \frac{1}{3} \sum_{j=1}^{3} f_{1j} = \frac{1}{3} \cdot (f_{11} + f_{12} + f_{13}) = \frac{1}{3} \cdot (0.8462 + 0.9231 + 1) = 0.9231$$

$$m_{21} = \frac{1}{3} \sum_{j=1}^{3} f_{2j} = \frac{1}{3} \cdot (f_{21} + f_{22} + f_{23}) = \frac{1}{3} \cdot (0.8305 + 0 + 0.4915) = 0.4407$$

$$m_{31} = \frac{1}{3} \sum_{j=1}^{3} f_{3j} = \frac{1}{3} \cdot (f_{31} + f_{32} + f_{33}) = \frac{1}{3} \cdot (0.12 + 0.16 + 0) = 0.0933$$

$$m_1 = [m_{11}, m_{21}, m_{31}] = [0.9231, 0.4407, 0.0933]$$

$$m_2 = [m_{12}, m_{22}, m_{32}] = [0.1282, 0.5876, 0.8933]$$

Jokaiselle näytteelle jokaisessa luokassa lasketaan samankaltaisuusvektorit S_j keskiarvovektorin avulla

$$S_j = s(O_j^i, m_i) \tag{7}$$

$$S_j = \sqrt[p]{1 - \left| (O_j^i)^p - (m_i)^p \right|}. \tag{8}$$

Esimerkki 2. jatkuu

$$S_j = 1 - |O_j^i - m_i|, \quad \text{kun } p = 1$$

$$S_1 = 1 - \big|[0.8462, 0.8305, 0.1200] - [0.9231, 0.4407, 0.0933]\big|$$

$$= [0.9231, 0.6102, 0.9733]$$

f_1	f_2	f_3
0.9231	0.6102	0.9733
1.0000	0.5593	0.9333
0.9231	0.9492	0.9067
0.8718	0.5141	0.9067
0.8590	0.9266	0.9867
0.9872	0.5876	0.8933

Neljäs vaihe on laskea jokaiselle muuttujalle sumean entropian arvot $h_1, h_2, \ldots, h_n$ samankaltaisuusarvojen avulla. De Luca ja Termini ovat kehittäneet sumeaa entropiaa mittaavan kaavan

$$H_A = -\sum_{j=1}^{J} \mu_A(x_j)\ln(\mu_A(x_j)) + (1-\mu_A(x_j))\ln(1-\mu_A(x_j)) \text{ [11].} \tag{9}$$

De Lucan ja Terminin kaava (7) ei toimi tilanteessa, jolloin similaarisuusarvo μ_A on 0 tai 1, sillä $\ln(0) = -\infty$. Tällöin kyseisessä muuttujassa ei ole epävarmuutta eikä De Lucan ja Terminin kaava ole soveltuva käytettäväksi. μ_A:n arvo 0.5 kuvaa suurinta epävarmuutta.

Uudemman sumean entropian mittarin ovat kehittäneet Parkash, Sharma ja Mahajan:

$$H(A;w) = \sum_{j=1}^{J} w_j \left(\sin\frac{\pi\mu_A(x_j)}{2} + \sin\frac{\pi(1-\mu_A(x_j))}{2} - 1 \right) \tag{10}$$

ja

$$H(A;w) = \sum_{j=1}^{J} w_j \left(\cos\frac{\pi\mu_A(x_j)}{2} + \cos\frac{\pi(1-\mu_A(x_j))}{2} - 1 \right) \text{ [12].} \tag{11}$$

Esimerkki 2. jatkuu

Sumean entropian arvot laskettuna De Lucan ja Terminin Kaavalla 9

$$H_k = -\sum_{j=1}^{J} s_{jk}\ln(s_{jk}) + (1-s_{jk})\ln(1-s_{jk})$$

$$\begin{aligned}
H_1 = -(&\, s_{11}\ln(s_{11}) + (1-s_{11})\ln(1-s_{11}) \\
&+ s_{21}\ln(s_{12}) + (1-s_{12})\ln(1-s_{12}) \\
&+ s_{31}\ln(s_{13}) + (1-s_{13})\ln(1-s_{13}) \\
&+ s_{41}\ln(s_{14}) + (1-s_{14})\ln(1-s_{14})) \\
&+ s_{41}\ln(s_{15}) + (1-s_{15})\ln(1-s_{15})) \\
&+ s_{41}\ln(s_{16}) + (1-s_{16})\ln(1-s_{16})) \\
H_1 = -(&\, 0.9231\cdot\ln(0.9231) + (1-0.9231)\ln(1-0.9231) \\
&+ 1.0000\cdot\ln(1.0000) + (1-1.0000)\ln(1-1.0000) \\
&+ 0.9231\cdot\ln(0.9231) + (1-0.9231)\ln(1-0.9231) \\
&+ 0.8718\cdot\ln(0.8718) + (1-0.8718)\ln(1-0.8718)) \\
&+ 0.8590\cdot\ln(0.8590) + (1-0.8590)\ln(1-0.8590)) \\
&+ 0.9872\cdot\ln(0.9872) + (1-0.9872)\ln(1-0.9872)) \\
H_1 &= 1.4008 \\
H_2 &= 3.1887 \\
H_3 &= 1.3986
\end{aligned}$$

Lopuksi valitun kynnysehdon (treshold) avulla valitaan, mitkä muuttujista poistetaan. Ne muuttujat, joiden sumean entropian arvo on suurempi kuin kynnysehto, poistetaan datasta. [13]

Esimerkki 2. jatkuu

Jos kynnysehtona käytetään arvoa 2, poistetaan datasta muuttuja f_2, sillä $h_2 = 3.1887 > 2$

2.2 Luokittelu samankaltaisuuden perusteella

Luokittelun tarkoitus on selvittää, mihin luokkaan tietty näyte kuuluu. Valmista dataa käytetään opettamiseen ja testaamiseen. Oppimista voi olla ohjattu oppiminen (supervised learning), jossa datassa on jokaisella näytteellä tieto sen luokasta, jota voidaan hyödyntää oppimisessa. Ohjaamattomassa oppimisessa (unsupervised learning) datassa ei ole luokkatietoa, jolloin dataa luokitellaan ryhmiin ja samankaltaisen datapisteiden voidaan olettaa olevan yksi luokka. Tässä luokkaa ei voida tarkistaa. [14]

Ensimmäisenä data jaetaan kahteen osaan, harjoitus- ja testausdata. Seuraavaksi otetaan käsittelyyn harjoitusdata. Harjoitusdatasta lasketaan keskiarvovektorit $m_i = [m_{1i}, m_{2i}, \ldots, m_{ni}$ jokaiselle luokalle C_i. Keskiarvovektorit ovat samalla ideaalivektorit datalle.

Esimerkki 3. Luokittelu

Käytetään samaa dataa kuin Esimerkissä 2, joka jaetaan harjoitus- ja testausdataan:

f_1	f_2	f_3	C	
0.8462	0.8305	0.1200	1	harjoitusdata
0.9231	0	0.1600	1	
0	0.1017	0.8000	2	
0.2692	0.6610	0.8800	2	
1.0000	0.4915	0	1	testausdata
0.1154	1.0000	1.0000	2	

Keskiarvovektorit, jotka ovat myös ideaalivektorit:

$$m_1 = [0.8846, 0.4153, 0.1400]$$

$$m_2 = [0.1346, 0.3814, 0.8400]$$

Luokittelu tehdään testausdatalla. Jokaiselle näytteelle O_j testausdatassa lasketaan samankaltaisuusarvot näytteen ja ideaalivektorien m_i välillä kaavalla

$$S_{ji}\langle O_j, m_i\rangle = \frac{1}{n}\sum_{i=1}^{n} \sqrt[p]{1 - \left|O_j^p - m_i^p\right|}. \tag{12}$$

Samankaltaisuusarvo lasketaan jokaiselle luokalle. Näyte kuuluu siihen luokkaan, jonka samankaltaisuusarvo on suurin. [13]

Esimerkki 3. jatkuu

Lasketaan samankaltaisuusarvot kaavalla (10) ($p = 1$):

Näyte O_1:

$$\begin{aligned}
S_{11}\langle O_1, m_1\rangle &= S_{11}\langle[1.0000, 0.4915, 0], [0.8846, 0.4153, 0.1400]\rangle \\
&= \frac{1}{3}((1-|1.0000-0.8846|)+(1-|0.4915-0.4153|)+(1-|0.4153-0.1400|) \\
&= 0.8894 \\
S_{12}\langle O_1, m_2\rangle &= S_{12}\langle[1.0000, 0.4915, 0], [0.1346, 0.3814, 0.8400]\rangle \\
&= \frac{1}{3}((1-|1.0000-0.1346|)+(1-|0.4915-0.3814|)+(1-|0.4153-0.8400|) \\
&= 0.3948
\end{aligned}$$

$S_{11} > S_{12}$, joten testausdatan näyte 1, kuuluu luokkaan 1.

Näyte O_2:

$$S_{21}\langle O_2, m_1\rangle = 0.2620$$
$$S_{22}\langle O_2, m_2\rangle = 0.7340$$

$S_{22} > S_{21}$, joten testausdatan näyte 2, kuuluu luokkaan 2.

2.3 Ristiinvalidointi ja tulosten analysointimittarit

Validoinnin tavoitteena on estää ennusteen ylisovittaminen (overfitting). Ylisovittunut ennuste reagoi liikaa harjoitusdatan satunnaisuuksiin eikä siten pysty ennustamaan tulevaisuudessa oikein eri aineistolla. Ristiinvalidoinnissa (cross validation) luokittelussa käytettävä harjoitusdata jaetaan vielä kahteen osaan, harjoitus- ja validointidata. Harjoitusdataa käytetään luokitteluun. Validointidata ei ole sama kuin testausdata. Harjoitusdatalla luodaan malli, validointidatalla arvioidaan mallia ja testausdatalla selvitetään ennusteen toimimista kokonaisuudessaan. Monte Carlo -ristiinvalidoinnissa harjoitusdata jaetaan satunnaisesti harjoitus- ja validointidataan. Datan jakaminen toistetaan N kertaa. [15], [16]

2.3.1 Herkkyys ja spesifisyys

		OIKEA LUOKKA: P	OIKEA LUOKKA: N
ENNUSTETTU LUOKKA	P	oikea positiivinen	väärä positiivinen
	N	väärä negatiivinen	oikea negatiivinen

Kuva 2. Kahden luokan luokittelu

Kahden luokan luokittelussa luokat ovat positiivinen ja negatiivinen. Ennustetut luokat ovat myös positiivinen ja negatiivinen. Tällöin ennusteella on neljä mahdollista tulosta, jotka on esitetty matriisimuodossa Kuvassa 2. Jos esimerkkinäyte on positiivinen ja ennuste on positiivinen, tulos on oikea positiivinen (true positive, TP) tai jos ennuste on negatiivinen, tulos on väärä negatiivinen (false negative, FN). Jos esimerkkinäyte on negatiivinen ja ennuste on negatiivinen, tulos on oikea negatiivinen (true negative, TN) tai jos ennuste on positiivinen, tulos on väärä positiivinen (false positive, FP). Näiden muuttujien avulla voidaan laskea mallin herkkyys (sensitivity) ja spesifisyys (specificity). Herkkyys (true positive rate, TPR) on oikeiden positiivisten osuus ennusteessa

$$\text{herkkyys} = \frac{\text{true positives (TP)}}{\text{positives (P)}} = \frac{\text{true positives (TP)}}{\text{true positives (TP) + false negatives (FN)}}. \tag{13}$$

Spesifisyys (true negative rate, TNR) on oikeiden negatiivisten suhde kaikkiin negatiivisiin

$$\text{spesifisyys} = \frac{\text{true negatives (TN)}}{\text{negatives (N)}} = \frac{\text{true negatives (TN)}}{\text{true negatives (TN) + false positives (FP)}}. \tag{14}$$

Tarkkuudessa (accuracy) huomioidaan kaikki oikeat tulokset

$$\text{tarkkuus} = \frac{\text{true positives (TP) + true negatives (TN)}}{\text{positives (P) + negatives (N)}}. \text{ [17]} \tag{15}$$

3 AINEISTO JA OHJELMISTOT

Käytetty aineisto on tutkimuskäyttöön avoin (vuosina 2008-2010 Portugalissa kerätty) Bank Marketing -data. Portugalilainen pankki keräsi vuosina 2008-2010 tietoa 17 eri markkinointikampanjasta, jotka tapahtuivat pääasiassa puhelimitse. Pankki kontaktoi noin 80 000 asiakasta sekä keräsi 59 ominaisuutta jokaisesta asiakkaasta. Datalle tehtiin esikäsittelyä, jossa datasta poistettiin asiakkaat, joiden markkinoinnin lopputulos (output) ei ollut binäärinen, onnistunut tai epäonnistunut. Muuttujat, joilla ei ollut merkitystä, poistettiin, jolloin muuttujien määrä väheni 59:stä 16:teen, joista osa on binäärimuotoisia (kyllä/ei) ja osassa on useampi vaihtoehto ja lisäksi tulosmuuttuja. Muuttujat on esitelty Liitteessä 1. Viimeiseksi poistettiin asiakkaat, joissa oli puuttuvia tietoja ja lopulliseksi asiakkaiden määräksi jäi 45 211. [2]

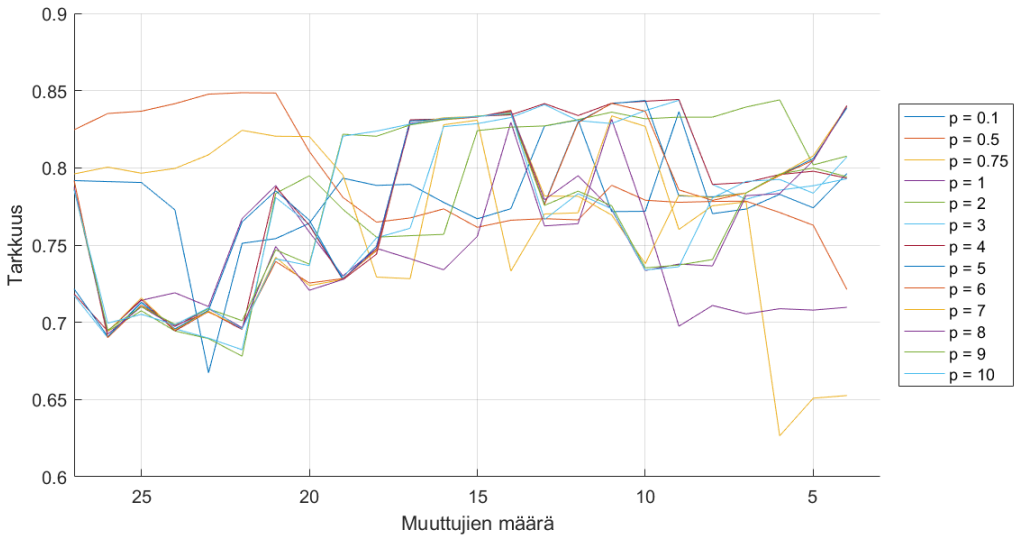
Esikäsittelyssä data muutettiin matriisimuotoon ja kaikille muuttujille annettiin numeroarvot. Muuttuja 16 ”edellisen kampanjan tulos” poistettiin matriisista kokonaan eikä sitä huomioitu laskennassa, koska suuri osa arvoista oli tuntemattomia. Kokonaisuudessaan matriisissa oli 29 muuttujaa, sillä jokainen työ muutettiin omaksi muuttujakseen. Muuttujan poiston jälkeen kaikki asiakkaat, joiden tiedoissa oli tuntemattomia muuttujia poistettiin. Laskennassa käytettävään dataan jäi 30 907 asiakasta.

Aineistossa on kaksi luokkaa: markkinointi on onnistunut tai markkinointi ei ole onnistunut. Onnistuneiden markkinointien prosenttiosuus on 11.7 %. Datan esikäsittelyn jälkeen sama osuus on 14.6 %.

Laskenta tehtiin MATLAB R2018b -ohjelmistolla. Laskennassa käytettiin valmiita funktioita muuttujan valinnassa sekä luokittelussa [13].

4 TULOKSET

Data järjestettiin luokkien mukaan ja molemmista luokista otettiin 70 % harjoitusdataan ja 30 % testausdataan. Harjoitusdatalla tehtiin muuttujien valinta. Luokittelussa harjoitusdata jaettiin satunnaisesti puoliksi harjoitus- ja validointidataksi 20 kertaa. Tulokset on esitetty De Lucan ja Terminin sumean entropian mittarilla (Kaava 8) laskettuna, mutta tulokset olivat samoja myös Parkash et al:n kaavoilla 10 ja 11. Kuvassa 3 on validoinnin keskitarkkuus muuttujan valinnan aikana. Muuttujan valinnassa käytettiin similaarisuusarvon laskennassa arvoa $p = 0.1 - 10$. Kuvasta 3 nähdään, että tarkkuudet vaihtelevat eri parametrin p arvoilla lasketuina. Parhaat tulokset saadaan, kun $p = 3$ tai $p = 4$. Rajakohtana usealla eri p-parametrilla on 11 muuttujaa, jonka jälkeen tulokset heikkenevät. Vaikka tarkkuus näyttää hyvältä myös vähemmillä muuttujilla, tulokset eivät ole kelpaavia, sillä suuri tarkkuus johtuu hyvästä spesifisyydestä. Datassa on toista luokkaa vain noin 15 %. Tämän luokan ennustettavuuden heikkentyessä, toinen luokka ennustuu paremmin ja se saa tarkkuuden vaikuttamaan hyvältä.



Kuva 3. Muuttujien määrän vaikutus luokittelutarkkuuteen

Tulosten jatkokäsittelyyn otetaan parametrin p arvolla 3 lasketut tulokset 9 muuttujalla. Taulukossa 1 on lueteltu 9 eniten vaikuttavaa muuttujaa satunnaisessa järjestyksessä ja Taulukossa 2 on lueteltu järjestyksessä poistetut muuttujat. Taulukossa 3 on taulukoituna testausdatalla tehdyn testauksen tulokset. Keskitarkkuus paranee hieman sekä keskiherkkyys paranee huomattavasti, mutta testausdatalla lasketut tulokset ovat silti heikkoja ennustettavuuden kannalta.

Muuttujan nimi
maksamaton velka
keskimääräinen saldo vuodessa
yhteydenoton kesto
yhteydenottojen määrä kampanjan aikana
yhteydenottojen määrä ennen tätä kampanjaa
työn tyyppi:
työtön
kodinhoitaja
yrittäjä
opiskelija

Taulukko 1. 9 eniten vaikuttavaa muuttujaa, kun $p = 3$

	Muuttujan nimi		Muuttujan nimi
27.	asuntolaina	18.	ikä
26.	naimisissa	17.	virkamies
25.	naimaton	16.	laina
24.	koulutus	15.	eronnut
23.	viimeinen yhteydenottopäivä	14.	yhteydenottotapa
22.	viimeinen yhteydenottokuukausi	13.	asiakaspalvelija
21.	johtaja	12.	eläkeläinen
20.	asentaja	11.	kuluneet päivät edellisestä yhteydenotosta
19.	haalarityöntekijä	10.	itsenäinen ammatinharjoittaja

Taulukko 2. Aineistosta poistetut muuttujat muuttujan valinnan aikana, kun $p = 3$

5 JOHTOPÄÄTÖKSET

Tuloksissa päädyttiin siihen, että paras ennuste saadaan, kun parametri $p = 3$ ja huomioidaan 9 muuttujaa, jotka on lueteltu Taulukossa 1. Vaikuttavat muuttujat ovat järkeviä, sillä esimerkiksi yhteydenottokuukaudella tai -päivällä ei luultavasti ole kovin paljoa merkitystä. Toisaalta taas yhteydenoton kesto vaikuttaa ennusteeseen, sillä oletettavasti, jos yhteydenotto on kestänyt pidempään, asiakas on ollut kiinnostunut, ja jos taas kesto on ollut lyhyt, asiakas ei ole ollut kiinnostunut.

	Alkutilanne	9 muuttujalla
Keskitarkkuus 95 % luottamusvälillä	$87.5 \pm 0.1\%$	$89.5 \pm 0.1\%$
Keskiherkkyys 95 % luottamusvälillä	$19.6 \pm 0.1\%$	$53.8 \pm 0.1\%$
Keskispesifisyys 95 % luottamusvälillä	$99.1 \pm 0.1\%$	$95.6 \pm 0.1\%$
Varianssi	$7.5 \cdot 10^{-6}$	$10.4 \cdot 10^{-6}$
Hajonta	$2.7 \cdot 10^{-3}$	$3.2 \cdot 10^{-3}$
p	0.35	1.6
m	0.6	1.85

Taulukko 3. Testausdatalla saadut luokittelutulokset

Markkinoinnissa tärkeämpää on saada markkinoitua oikeille asiakkaille kuin saavuttaa täydellinen tarkkuus. Sairauksien ennustamisessa spesifisyys on tärkeämpää kuin herkkyys, sillä herkkyys ei huomioi vääriä positiivisia. Markkinoinnissa halutaan saada markkinointi kohdennettua niille, jotka ovat potentiaalisia asiakkaita. Tällöin ei ole haitallista, vaikka joukossa olisi vääriä positiivisia, kunhan mahdollisimman moni oikea positiivinen kuuluu joukkoon. Herkkyys mittaa oikeiden positiivisten osuutta kaikista positiivisista. Alkutilanteessa herkkyys on huono, noin 20 %. Spesifisyys puolestaan on erittäin korkea. 9 muuttujalla herkkyys on 54 %, joten se on parantunut muuttujien valinnan jälkeen. Spesifisyys on heikentynyt 3.5 %-yksikköä. Esimerkissä 4 on tehty vertailua suora- ja massamarkkinoinnin välillä saaduilla tuloksilla. Esimerkin perusteella tulokset eivät siis ole kovin hyviä suoramarkkinoinnin kannalta.

Esimerkki 4. Kustannukset ja tuotto suora- ja massamarkkinoinnilla

Puhelun hinta 15 €/h

Keskimääräinen talletus 1000 €

Voittoprosentti 7 %

Puhelun kesto keskimäärin 12 min

Puhelut yhteensä 100 kpl

Onnistuneet markkinoinnit 10 kpl

Epäonnistuneet markkinoinnit 90 kpl

Herkkyys 53 %

Spesifisyys 96 %

Suoramarkkinoinnissa soitetaan positiivisiksi ennustetuille.

	Puhelujen määrä	Kulut	Tulot	Tuotto
FN	4	12	280	268
FP	5	15	0	-15
TN	86	258	0	-258
TP	5	15	350	335
Suoramarkkinoinnin tuotto				320
Massamarkkinoinnin tuotto				400

Tässä tilanteessa massamarkkinointi on tuottavampi vaihtoehto. Jos herkkyys olisi 65 % tai parempi, olisi suoramarkkinointi taloudellisempi.

Lähteet

[1] C. X. Ling ja C. Li, “Data Mining for Direct Marketing: Problems and Solutions”, teoksessa *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, sarja KDD’98, New York, NY: AAAI Press, 1998, s. 73–79. url: http://dl.acm.org/citation.cfm?id=3000292.3000304.

[2] S. Moro, R. Laureano ja P. Cortez, “Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology”, teoksessa *Proceedings of the European Simulation and Modelling Conference - ESM’2011*, P. N. et al., toim., Guimaraes, Portugal: EUROSIS, lokakuu 2011, s. 117–121.

[3] Z. M. Hira ja D. F. Gillies, “A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data”, *Advances in Bioinformatics*, vol. 2015, s. 1–13, 2015. DOI: 10.1155/2015/198363. url: https://doi.org/10.1155/2015/198363.

[4] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang ja H. Liu, “Feature Selection”, *ACM Computing Surveys*, vol. 50, nro 6, s. 1–45, joulukuu 2017. DOI: 10.1145/3136625. url: https://doi.org/10.1145/3136625.

[5] I. Guyon, M. Nikravesh, S. Gunn ja L. A. Zadeh, toim., *Feature Extraction*. Springer Berlin Heidelberg, 2006. DOI: 10.1007/978-3-540-35488-8. url: https://doi.org/10.1007/978-3-540-35488-8.

[6] C. Iyakaremye, P. Luukka ja D. Koloseni, “Feature selection using Yu’s similarity measure and fuzzy entropy measures”, teoksessa *2012 IEEE International Conference on Fuzzy Systems*, IEEE, kesäkuu 2012. DOI: 10.1109/fuzz-ieee.2012.6250817. url: https://doi.org/10.1109/fuzz-ieee.2012.6250817.

[7] S. Kurnaz, “Hybrid Datamining Approaches to Predict Success of Bank Telemarketing Anas Nabeel Falih AL-Shawi”, 2019.

[8] L. Zadeh, “Fuzzy sets”, *Information and Control*, vol. 8, nro 3, s. 338–353, kesäkuu 1965. DOI: 10.1016/s0019-9958(65)90241-x. url: https://doi.org/10.1016/s0019-9958(65)90241-x.

[9] C. E. Shannon, “A Mathematical Theory of Communication”, *Bell System Technical Journal*, vol. 27, nro 3, s. 379–423, heinäkuu 1948. DOI: 10.1002/j.1538-7305.1948.tb01338.x. url: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

[10] K. Jain ja K. Bhandare, “Min max normalization based data perturbation method for privacy protection”, *International Journal of Computer & Communication Technology*, vol. 2, s. 45–50, tammikuu 2011.

[11] A. D. Luca ja S. Termini, “A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory”, *Information and Control*, vol. 20, nro 4, s. 301–312, toukokuu 1972. DOI: 10.1016/s0019-9958(72)90199-4. url: https://doi.org/10.1016/s0019-9958(72)90199-4.

[12] O. Parkash, P. Sharma ja R. Mahajan, “New measures of weighted fuzzy entropy and their applications for the study of maximum weighted fuzzy entropy principle”, *Information Sciences*, vol. 178, nro 11, s. 2389–2395, kesäkuu 2008. DOI: 10.1016/j.ins.2007.12.003. url: https://doi.org/10.1016/j.ins.2007.12.003.

[13] P. Luukka, “Feature selection using fuzzy entropy measures with similarity classifier”, *Expert Systems with Applications*, vol. 38, nro 4, s. 4600–4607, huhtikuu 2011. DOI: 10.1016/j.eswa.2010.09.133. url: https://doi.org/10.1016/j.eswa.2010.09.133.

[14] D. Poole ja A. Mackworth, *Artificial Intelligence: Foundations of Computational Agents*, 2. painos. Cambridge, UK: Cambridge University Press, 2017, ISBN: 978-0-521-51900-7. url: http://artint.info/2e/html/ArtInt2e.html.

[15] J. Shao, “Linear Model Selection by Cross-validation”, *Journal of the American Statistical Association*, vol. 88, nro 422, s. 486–494, kesäkuu 1993. DOI: 10.1080/01621459.1993.10476299. url: https://doi.org/10.1080/01621459.1993.10476299.

[16] Q.-S. Xu ja Y.-Z. Liang, “Monte Carlo cross validation”, *Chemometrics and Intelligent Laboratory Systems*, vol. 56, nro 1, s. 1–11, huhtikuu 2001. DOI: 10.1016/

s0169-7439(00)00122-2. url: https://doi.org/10.1016/s0169-7439(00)00122-2.

[17] T. Fawcett, “An introduction to ROC analysis”, *Pattern Recognition Letters*, vol. 27, nro 8, s. 861–874, kesäkuu 2006. DOI: 10.1016/j.patrec.2005.10.010. url: https://doi.org/10.1016/j.patrec.2005.10.010.

Taulukot

Kuvat

Liite 1: Dataan kuuluvat muuttujat ja niiden selitykset

	muuttujan nimi	selitys
1	ikä	numero
2	työ	virkamies, tuntematon, työtön, johtaja, kodinhoitaja, yrittäjä, opiskelija, haalarityöntekijä, itsenäinen ammatinharjoittaja, eläkeläinen, asentaja, asiakaspalvelija
3	siviilisääty	naimisissa, eronnut, naimaton
4	koulutus	1. asteen, 2.asteen, 3.asteen
5	maksamaton velka	kyllä, ei
6	keskimääräinen saldo vuodessa	numero (euroina)
7	asuntolaina	kyllä, ei
8	laina	kyllä, ei
9	yhteydenottotapa	tuntematon, puhelin, matkapuhelin
10	viimeinen yhteydenottopäivä	numero
11	viimeinen yhteydenottokuukausi	kuukaudet
12	yhteydenoton kesto	numero (sekunteina)
13	yhteydenottojen määrä kampanjan aikana	numero
14	kuluneet päivät edellisestä yhteydenotosta	numero (-1, jos ei aiempia yhteydenottoja)
15	yhteydenottojen määrä ennen tätä kampanjaa	numero
16	edellisen kampanjan tulos	tuntematon, muu, epäonnistunut, onnistunut
17	tulos	kyllä, ei