TEKNILLINEN KORKEAKOULU
Tietotekniikan osasto
Teknillisen fysiikan laitos

Kimmo Kiviluoto

Tilinpäätösten analysointi itseorganisoivalla kartalla

Diplomi-insinöörin tutkintoa varten tarkastettavaksi jätetty diplomityö.
Työn valvoja: Professori Erkki Oja
Työn ohjaaja: Diplomi-insinööri Pentti Bergius
Otaniemessä 9.5.1996.

TEKNILLINEN KORKEAKOULU **DIPLOMITYÖN TIIVISTELMÄ**

Tekijä:	Kimmo Kiviluoto	
Työn nimi:	Tilinpäätöstietojen analysointi itseorganisoivalla kartalla	
Title in English:	Analyzing Financial Statements with the Self-Organizing Map	
Päivämäärä: 9.5.1996		**Sivumäärä:** 60
Osasto:	Tietotekniikan osasto	**Professuuri:**
Laitos:	Teknillisen fysiikan laitos	Tik-61 Informaatiotieteet
Työn valvoja:	Professori Erkki Oja	
Työn ohjaaja:	Diplomi-insinööri Pentti Bergius	

Tilinpäätöstietojen analysointi on yksi yritystutkimuksen kulmakivistä. Pelkistetyimmässä muodossaan analyysissä pyritään luokittelemaan yritykset konkurssiyrityksiin ja terveisiin yrityksiin niiden antamien tilinpäätösten perusteella. Tässä diplomityössä etsittiin menetelmiä, joilla luokittelu voitaisiin luotettavasti toteuttaa.

Tutkimus toteutettiin yhteistyössä Kera Oy:n kanssa, ja siinä käytetty aineisto koostui tietyn Kera Oy:n asiakasyritysten osajoukon tilipäätöksistä. Tehtyjen rajausten jälkeen aineistoon jäi 4 898 tilinpäätöstä yhteensä 1 137 yritykseltä; näistä 304 oli ajautunut konkurssiin.

Työ jakautuu kahteen osaan. Ensimmäisessä osassa tarkastellaan konkurssi-ilmiötä; päämääränä on löytää konkurssiyrityksiä ja terveitä yrityksiä erottavia tekijöitä. Työkaluna käytettiin erästä neuroverkkomallia, itseorganisoivaa karttaa, jonka avulla nimellisesti korkeadimensioinen ilmiö pystytään muokkaamaan visuaaliseen tarkasteluun soveltuvaan muotoon.

Työn toisessa osassa vertaillaan erilaisten luokittelumenetelmien soveltuvuutta tilinpäätösaineiston käsittelyyn. Luokittelussa käytettiin kahta eri kriteeriä: tavallista virheiden kokonaismäärän minimointiin pyrkivää luokittelukriteeriä sekä ns. Neyman-Pearsonin kriteeriä, jossa yhden luokan luokitteluvirhe asetetaan tietyn suuruiseksi ja muiden luokkien luokitteluvirheet pyritään minimoimaan tämän rajoituksen puitteissa. Käytännössä arvokkaita tuloksia saatiin ainoastaan Neyman-Pearsonin kriteeriä käyttäen.

Luokitteluun käytettiin seuraavia menetelmiä: kolme itseorganisoivaan karttaan perustuvaa luokitinta, oppiva vektorikvantisointi, lähimmän naapurin luokitin, lineaarinen diskriminanttianalyysi sekä kvadraattinen diskriminanttianalyysi; menetelmä oppivan vektorikvantisoinnin modifioimiseksi Neyman-Pearsonin kriteerille kehitettiin tämän työn yhteydessä. Parhaat tulokset saatiin itseorganisoivaan karttaan perustuvalla luokittimella.

Avainsanat: itseorganisoiva kartta, oppiva vektorikvantisointi, tilinpäätösanalyysi, konkurssien ennustaminen

HELSINKI UNIVERSITY OF TECHNOLOGY **ABSTRACT OF THE MASTER'S THESIS**

Author:	Kimmo Kiviluoto	
Title of the thesis:	Analyzing Financial Statements with the Self-Organizing Map	
Title in Finnish:	Tilinpäätöstietojen analysointi itseorganisoivalla kartalla	
Date: 9.5.1996		**Number of pages:** 60
Faculty:	Information technology	**Professorship:**
Department:	Technical physics	Tik-61 Information sciences
Supervisor:	Professor Erkki Oja	
Instructor:	M.Sc.(tech.) Pentti Bergius	

Analysis of financial statements is one of the cornerstones of the analysis of an enterprise. The crudest level of the analysis is classifying the companies to healthy and bankrupt on the basis of their financial statements. In this diploma thesis, methods to reliably carry out this classification were investigated.

The study was conducted in co-operation with Kera Ltd, a service company that specializes in financing and development of small and medium-sized enterprises in Finland. The material consists of a certain segment of Kera Ltd's customer companies. The total number of financial statements used is 4 898; these have been given by 1 137 companies, of which 304 have gone bankrupt.

The study consists of two parts. In the first part, the phenomenon of going bankrupt is examined. The aim has been to find out differences between the companies that go bankrupt and those that do not. The method that was used was a neural network paradigm called Self-Organizing Map, which can be applied to transform nominally high-dimensional data better-suited for visual analysis.

In the second part of the study, several classification methods were compared on the financial statement data. Two different classification criteria were used: the usual criterion that aims to minimize the total number of misclassifications, and the so-called Neyman-Pearson criterion that fixes the relative error of one class and aims to minimize the classification error in other classis within this constraint. Results of practical value were obtained only using the Neyman-Pearson criterion.

In the classification, the following classification methods were used: three classifiers based on the self-organizing map, learning vector quantization, nearest-neighbor classifier, linear discriminant analysis, and quadratic discriminant analysis; the method to modify learning vector quantization for the Neyman-Pearson criterion was an original contribution of this study. The best-performing classifier was one of the classifiers based on the self-organizing map.

Keywords: self-organizing map, learning vector quantization, financial analysis, bankruptcy prediction

Kiitokset

Lämpimin kiitokseni kaikille työtovereille mukavasta työilmapiiristä ja monenmoisesta avustuksesta sekä tämän tutkimuksen tekemiseen liittyvissä että monissa muissakin asioissa. Aivan erityisesti haluaisin kiittää Jorma Laaksosta QDA- ja kNN-luokitustulosten laskemisesta.

Kiitokset myös työn ohjaajalle Pentti Bergiukselle sekä valvojalle Erkki Ojalle hyvin arvokkaista ohjeista ja kommenteista työn eri vaiheissa.

Inkeri Hyväristä, ravintola Kvarkin pitkäaikaista emäntää haluaisin puolestani kiittää niistä oivallisista olosuhteista, jotka hän järjesti pyöreiden pöytien ääressä istuskeluun – yksi niistä asioista, jotka ovat tehneet opiskelusta paljon muutakin kuin vain opiskelua.

Lopuksi kiitän vaimoani Minnaa kannustuksesta sekä kärsivällisyydestä.

Otaniemessä 9.5.1996

Kimmo Kiviluoto

Sisältö

Luku 1

Johdanto

Tässä diplomityössä tutkitaan erään neuroverkkopohjaisen data-analyysimenetelmän, itseorganisoituvan kartan, sekä joidenkin muiden menetelmien käyttämistä tilinpäätöstietojen analysoinnissa. Päämäärät ovat kahtaalla: työssä pyritään paitsi selvittämään yrityksen konkurssiin johtavan kehityksen näkymistä tilinpäätöstiedoissa, myös löytämään hyviä keinoja käyttää itseorganisoivaa karttaa tilastollisen analyysin työkaluna.

1.1 Taustaa

Konkursiin johtavan kehityksen oireiden havaitseminen mahdollisimman aikaisin on erittäin arvokasta niin yrityksen itsensä kuin rahoittajienkin kannalta. Konkurssi on yleensä huonoin ja kallein ratkaisu kaikille osapuolille. Yrityksen toiminnan loppuminen merkitsee omistajille ja rahoittajille rahallisia menetyksiä, yrityksen henkilöstö menettää työpaikkansa, alihankkijaverkostossa katoaa asiakas tai toimittaja, ja samalla häviää suuri osa yrityksen prosesseihin sitoutuneesta henkisestä pääomasta. Vaikka konkursseja voidaan pitää myös siinä mielessä myönteisinä tapahtumina, että ne karsivat huonoja yrityksiä ja antavat siten hyvin toimiville paremmat toimintaedellytykset, näin syntyvä hyöty ei kompensoi konkursseista aiheutuvia haittoja.

Konkurssiuhan torjuminen onnistuu sitä paremmin, mitä varhaisemmassa vaiheessa kielteiseen suuntaan johtava kehitys havaitaan. Myös uhan torjumiseen tarvittavien toimenpiteiden rajuus vähenee, jos niihin ryhdytään riittävän aikaisin. Tärkeä tavoite olisi siksi kehittää mahdollisimman varhaisen varoituksen antava järjestelmä *(early warning system)*, joka voisi perustua niin yrityksen reaali- kuin rahaprosessin tunnuslukujen seurantaan, miksei molempiinkin. Karkeimmillaan tällainen järjestelmä perustuisi pelkän rahaprosessin vuosittaiseen seuraamiseen, jolloin syötetietoina käytettäisiin tilinpäätöstunnuslukuja – kuten tässä työssä on tehty.

Edellä kuvatut tekijät muodostavat yhden lähtökohdan tälle työlle. Toinen lähtökohta on neuroverkkotutkimus. Neuroverkot ovat osoittautuneet hyödyllisiksi lukuisilla sovellusalueilla, muun muassa hahmontunnistuksessa, monimutkaisten prosessien mallintamisessa ja ohjauksessa sekä tilastojen analysoinnissa. Niitä on myös sovellettu konkurssien ennustamiseen tilinpäätöstietojen pohjalta – ongelma on jopa muodostunut eräänlaiseksi testipenkiksi, jolla neurolaskentaa ja perinteisiä tilastomatemaattisia malleja vertaillaan toisiinsa.

Aiemmissa tutkimuksissa, joissa neurolaskennan menetelmiä sovelletaan tilinpäätösanalyysiin, on kuitenkin eräitä piirteitä, joiden vuoksi aihe on edelleenkin tutkimisen arvoinen. Ensinnäkin, valtaosa aiemmista tutkimuksista pohjautuu varsin suppeisiin aineistoihin ja eri malleilla saatujen tulosten erot pieniä, joten tulosten yleistettävyydestä on vaikea sanoa mitään varmaa. Lisäksi tässä tutkimuksessa käytettyjen neuroverkkomenetelmien soveltuvuutta tilinpäätösaineistoon on tutkittu vain vähän – valtaosa aiemmista neuroverkkoja käyttävistä tutkimuksista perustuu ns. monikerros-perceptron-verkkoon, tässä tutkimuksessa taas keskeisinä työkaluina ovat itseorganisoiva kartta ja oppiva vektorikvantisointi.

1.2 Tutkimusongelma

Keskeinen kysymys tarkastellussa sovelluksessa on seuraava: miten yrityksen tulevaisuudennäkymät heijastuvat tilinpäätöstietoihin? Erityisesti, onko tilinpäätöstietojen avulla mahdollista havaita sellaiset muutokset yrityksen toimintatavoissa, jotka ennakoivat konkurssiriskin kasvamista?

Pelkistettynä ja ahtaasti rajattuna sama kysymys voidaan muotoilla näin: millä todennäköisyydellä yritys, jonka tähänastiset tilinpäätökset tunnetaan, menee konkurssiin tietyn ajanjakson sisällä? Tällöin ongelmaa voidaan tarkastella joko luokitteluna terveisiin ja konkurssiyrityksiin tai konkurssin todennäköisyyden estimointina.

Tarkasteltavan sovelluskohteen lisäksi myös tässä työssä käytettävissä menetelmissä riittää paljon tutkimista. Toimivilla työkaluilla, jotka soveltuvat datan laadulliseen analyysiin tai luokitteluun ja ennusteiden laatimiseen, olisi paljon käyttöä tilinpäätösanalyysin lisäksi monella muullakin sovellusalueella.

1.3 Tutkimuksen tavoitteet

Tutkimuksen ensisijainen tavoite on kehittää menetelmä, joka aiempia keinoja luotettavammin ja varhaisemmassa vaiheessa havaitsee yrityksen ajautumisen konkurssiuralle. Tavoitteeseen pääsemiseksi tutkitaan, miten eri tunnusluvut suhtautuvat toisiinsa, ja miten konkurssiriski riippuu tunnusluvuista. Lisäksi etsitään keinoja luokitella tilinpäätökset terveisiin ja konkurssiyritysten antamiin.

Työkaluna tunnuslukujen ja konkurssiriskin välisten suhteiden tarkastelussa käytetään itseorganisoivaa karttaa. Sen käyttöön liittyen vertaillaan erilaisten datan esikäsittelytapojen vaikutusta kartan muotoutumiseen, kun päämääränä on laatia kartta, jolla riippuvuuksia voisi parhaiten tarkastella. Tavoitteena on löytää eroja terveiden ja konkurssiyritysten käyttäytymisessä, joita voisi jatkossa hyödyntää.

Luokittelussa vertaillaan erilaisten luokittelumenetelmien toimivuutta tilinpäätösaineiston analysoinnissa. Näin pyritään löytämään tehtävään parhaiten sopivat työkalut selvittäen samalla, miten neurolaskentamenetelmiä voisi tämän tapaiseen ongelmaan tehokkaimmin soveltaa.

1.4 Tutkimuksen rajaukset

Tutkimuksen merkittävin rajaus on, että yrityksen tulevaisuudennäkymien arviointi pelkistetään karkeimmalle mahdolliselle tasolle: meneekö yritys lähitulevaisuudessa konkurssiin vai ei? Huomiota ei siis kiinnitetä lainkaan esimerkiksi siihen, kuinka tuloksekkaasti yritys toimii.

Tutkimuksessa ei myöskään laajalti kokeilla erilaisia tunnuslukuyhdistelmiä, vaan tarkastelu perustetaan yhteen ainoaan neljän eri tunnusluvun muodostamaan yhdistelmään. Perustana tunnuslukuyhdistelmän valinnassa on ollut aiempi kirjallisuus ja tutkimuksen yhteistyökumppanin, Kera Oy:n asiantuntijoiden mielipiteet sekä Kera Oy:ssä tehdyistä selvityksistä saadut kokemukset. Tunnusluvut on pyritty valitsemaan siten, että ne olisivat mahdollisimman riippumattomia yrityksen toimialasta.

Käytetty aineisto rajoittuu tiettyyn osajoukkoon Kera Oy:n asiakasyrityksistä, joka muodostaa varsin valikoidun otoksen kaikista Suomessa toimivista yrityksistä. Aineiston rajauksia on esitelty laajemmin luvussa 5.

Tutkimuksessa käytetään vain pientä osaa kaikista tilinpäätösaineiston luokitteluun ehdotetuista menetelmistä. Vertailumenetelmät itseorganisoivalle kartalle ja oppivalle vektorikvantisoinnille on valittu lähinnä sillä perusteella, miten yleisiä ja tunnettuja ne ovat, ja miten hyviä tuloksia niillä on luokittelussa yleensä saatu.

Luku 2

Katsaus toteutettuihin konkurssitutkimuksiin

2.1 Konkurssi

Juridisesti konkurssilla tarkoitetaan prosessia, jossa yrityksen koko ulosmittauskelpoinen omaisuus jaetaan oikeuden päätöksellä mahdollsimman oikeudenmukaisesti yrityksen velkojien kesken. Yritys voidaan hakea konkurssiin joko yrityksen omasta tai sen velkojien aloitteesta, kun yritys on osoittautunut kykenemättömäksi vastaamaan tekemistään maksusitoumuksista. Konkurssia muistuttava tilanne on myös sellainen, jossa yritys lopettaa toimintansa ja sen omaisuus jaetaan velkojien kesken ilman, että oikeus olisi antanut konkurssituomiota. Syynä tällaiseen menettelyyn voi olla joko velkojien ja yrityksen tekemä sovitteluratkaisu, tai konkurssin raukeaminen oikeudessa sen vuoksi, että konkurssipesässä ei ole riittävästi varoja kattamaan edes konkurssimenettelyn aiheuttamia kustannuksia.

2.2 Konkurssitutkimuksen esihistoria ja klassikot

Jonkinlaista tilinpäätöstutkimusta on varmasti harjoitettu niin kauan kuin yritykset ovat tilinpäätöksiään julkistaneet, ja konkurssin todennäköisyyden arvioiminen on epäilemättä ollut alusta alkaen tutkimuksessa keskeinen kysymys. Tieteellisemmin asiaa ryhdyttiin tutkimaan 1930-luvun laman aikoihin; Prihti [1] sekä Laitinen [2] mainitsevat alan pioneereina Ramserin ja Fosterin (1931) sekä FitzPatrickin (1932), joita seurasi seuraavan kolmenkymmenen vuoden aikana joukko muita tutkimuksia. Yleinen lähestymistapa oli joidenkin tunnuslukujen keskiarvojen vertailu konkurssiin ajautuneiden ja terveiden yritysten välillä; joissakin tutkimuksissa tarkasteltiin myös tunnuslukujen muuttumista konkurssin lähestyessä.

Systemaattisemmin konkurssin ennustamista tilinpäätöstietojen perusteella alettiin tutkia 1960-luvun puolivälin jälkeen. Laitinen nostaa seuraavat kolme tutkimusta merkitykseltään yli muiden: Beaver (1966), Altman (1968) sekä Prihti (1975); lisäksi hän mainitsee Argentin tutkimukset (1976, 1983), jotka kuitenkin eroavat edellämainituista siinä, että ne perustuvat lähinnä muihin tekijöihin kuin tilinpäätöstunnuslukuihin.

2.2.1 Beaver: Ennustaminen yksittäisten tunnuslukujen perusteella

Beaver [3] valitsi tutkimukseensa 79 konkurssiyritystä ja 79 tervettä yritystä vastinpareittain siten, että jokaista konkurssiyritystä vastasi suurin piirtein samankokoinen ja samalla toimialalla samantapaisessa ympäristössä toimiva terve yritys. Beaver havaitsi, että näiden yritysryhmien tunnuslukujen keskiarvot erosivat toisistaan jo viisi vuotta ennen konkurssia; suurimmat erot olivat rahoitustuloksen ja vieraan pääoman sekä nettotuloksen ja koko pääoman suhteissa.

Tutkimuksessa selvitettiin myös, mitkä tunnusluvut olivat käyttökelpoisimpia konkurssin ennustamisessa. Parhaiksi osoittautuivat tunnusluvut rahoitustulos/vieras pääoma, nettotulos/koko pääoma sekä vieras pääoma/koko pääoma. Kokonaisluokitteluvirheiksi näille tunnusluvuille saatiin 13...19% vuotta ennen konkurssia ja 22...29% viittä vuotta ennen konkurssia, kun erotteluanalyysin parametri määritettiin puolella yrityspareista ja luokittelutulos laskettiin aineiston toiselle puolikkaalle.

Beaver esitti lisäksi, että konkurssiyritykset on vaikeampi tunnistaa kuin terveet yritykset. Tyypin I virhe eli konkurssiin ajautuneen yrityksen luokittelu terveeksi yritykseksi oli selvästi yleisempi kuin tyypin II virhe eli terveen yrityksen luokittelu konkurssiyritykseksi. Parhaalla yksittäisellä tunnusluvulla, rahoitustulos/vieras pääoma, nämä virheet olivat 22% ja 5% vuotta ennen konkurssia sekä 47% ja 3% neljää vuotta ennen konkurssia – tässä tyypin II virheen pieneneminen ennustejakson pidentyessä lienee otoksen pienuuden aiheuttamaa tilastollista harhaa. Tarkkuus pienenee siis niin nopeasti, että konkurssiyritystä ei yhden tunnusluvun perusteella pystytä tunnistamaan neljää vuotta ennen konkurssia. Rahoittajan kannalta tällaisen ennusteen kantama on riittämätön: usein lainan lyhennykset alkavat vasta kolmen vuoden kuluttua lainan myöntämisestä.

2.2.2 Altman: Yhdistelmätunnuslukujen käyttö

Systemaattisen menetelmän konkurssin ennustamiseen usean tunnusluvun perusteella esitti ensimmäisenä Altman [4]. Menetelmän matemaattisena perustana oli lineaarinen erotteluanalyysi (kts. kappale 4.2.2). Mallissa käytetyiksi tunnusluvuiksi valittiin kahdenkymmenen heuristisesti valitun kandidaatin joukosta seuraavat viisi:

x_1: nettokäyttöpääoma/koko pääoma

x_2: kertyneet voittovarat/koko pääoma

x_3: tulos ennen korkoja ja veroja/koko pääoma

x_4: oman pääoman markkina-arvo/vieraan pääoman kirjanpitoarvo

x_5: myynti/koko pääoma

Malli voidaan pelkistää ns. *Z-luvun* laskentakaavaan, jossa tunnusluvut kerrotaan tietyillä painokertoimilla ja lasketaan yhteen:

$$Z = 0,012x_1 + 0,014x_2 + 0,033x_3 + 0,006x_4 + 0,999x_5 \tag{2.1}$$

Kaavan perusteella koko mallia kutsutaan usein Z-malliksi. Jako konkurssiyrityksiin ja terveisiin yrityksiin tehdään vertaamalla Z-lukua johonkin kriittiseen arvoon Z_0, joka valitaan siten, että tyypin I ja II virheille saadaan haluttu suhde.

Altmanin 33 yritysparista muodostuvassa aineistossa kokonaisluokitteluvirheen minimoi kriittinen arvo $Z_0 = 2,675$. Testiaineistossa, jonka koko ja koostumus on tuntematon, tämä tuotti luokitteluvirheiksi vuotta ennen konkurssia 17% (kokonaisvirhe), 4% (tyyppi I) ja 21% (tyyppi II). Kuitenkin jo mallin rakentamisessa käytetyssä aineistossa kokonaisluokitteluvirhe kolmea vuotta ennen konkurssia oli 52%. Syynä näin suureen virheeseen saattaa olla se, että malli perustui varsin pieneen aineistoon, ja toisaalta malli rakennettiin ennenkaikkea minimoimaan virhe ennusteessa yhtä vuotta ennen konkurssia.

Altman on jatkotutkimuksissaan muokannut Z-malliaan useaan otteeseen. Tunnusluku x_5 jätettiin kokonaan pois liian toimialakohtaisena, ja tunnusluvussa x_4 oman pääoman markkina-arvo korvattiin oman pääoman kirjanpitoarvolla, jotta tunnuslukua voitaisiin soveltaa myös julkisesti noteeraamattomiin yrityksiin. Samalla mallin kertoimia muutettiin. Tehdyillä muutoksilla ei kuitenkaan ollut merkittävää vaikutusta mallin ennustamiskykyyn.

Z-mallin lisäksi Altman on kehittänyt myös ns. Zeta-mallin yhdessä Haldemanin ja Narayananin kanssa [5]. Malli perustuu seitsemään tunnuslukuun, joista yksikään ei ole mukana Z-mallissa; erona Z-malliin on lisäksi se, että Zeta-mallissa käytettävät tunnusluvut lasketaan oikaistuista tilinpäätöksistä. Mallin parametrit on estimoitu 53 yritysparin aineistosta, jossa konkurssiyritysten tiedot ovat konkurssia edeltäneestä tilinpäätöksestä. Ristiinvalidoidut virheet (ristiinvalidointimenetelmä on esitelty kappaleessa 6.2.1) yhtä vuotta ennen konkurssia ovat 9,0% (kokonaisvirhe), 7,5% (tyyppi I) ja 10,3% (tyyppi II). Pidemmän kantaman ristiinvalidoituja virheitä ei ole ilmoitettu – osasyynä tähän lienee se, että Zeta-malli on Zeta Services Inc. -yhtiön omaisuutta, minkä vuoksi myöskään sen kertoimia ei ole julkistettu.

Zeta-malli perustuu Z-mallin tavoin lineaariseen erottelufunktioon. Mallin kehittäjät ovat kokeilleet myös neliöllistä erottelufunktiota mutta todenneet lineaarisen funktion antavan parempia tuloksia käytetyllä 53 yritysparin aineistolla.

2.2.3 Prihti: Konkurssiteorian kehittäminen

Siinä missä Beaver ja Altman tarkastelivat konkurssin ennustamista puhtaan empiirisesti, Prihti [1] kehitti väitöskirjassaan teoreettisen kehikon konkurssi-ilmiölle, josta saatuja kriittisiä tunnuslukujen arvoja hän sitten testasi empiirisesti.

Prihti esitti työssään kolme konkurssihypoteesia:

1. Toimivilla yrityksillä kukin yksittäinen investointi ja siten myös investoinneista muodostuva aikasarja tuottaa tulorahoitusta vähintään investoinneista syntyvien maksuvaatimusten verran; tästä voidaan poiketa vain yksittäisissä tapauksissa lyhyellä tähtäyksellä. Sen sijaan konkurssiyrityksillä tapahtuu jonakin ajankohtana investointien nettotuloissa pysyvä muutos epäsuotuisaan suuntaan eli investointien tulorahoituksella ei pystytä kattamaan niistä syntyneitä maksuvaateita; poikkeuksia tähänkin sääntöön saattaa olla yksittäisissä tapauksissa ja lyhyellä tähtäyksellä.

2. Konkurssi tapahtuu sillä hetkellä, kun yrityksen lisäluoton tarve ylittää sen yhteenlasketun eri sidosryhmiltä saatavissa olevan enimmäislisäluoton.

3. Yritys voi saada yksipuolista lisäluottoa – toisin sanoen, jättää jotkin velvoitteensa hoitamatta – joltakin sidosryhmältä sitä helpommin, mitä alempana sidosryhmä on konkurssisäädösten määräämässä etuoikeusjärjestyksessä. Enimmäislisäluottomäärien täytyttyä kuitenkin mikä hyvänsä sidosryhmä on valmis hakemaan yrityksen konkurssiin sillä hetkellä, jolloin se ottaa vielä lisää yksipuolista lisäluottoa.

Ensimmäisen konkurssihypoteesin, investointien tuoton riittävyyden, Prihti yhdisti tunnuslukuun tulos verojen jälkeen/koko pääoma. Toiseen konkurssihypoteesiin, konkurssin ajankohdan määräytymiseen, liitettiin kaksi tunnuslukua: quick[1]/koko pääoma sekä vieras pääoma/koko pääoma. Kolmannelle konkurssihypoteesille, yksipuolisen lisäluoton käyttöjärjestykselle, ei tilinpäätöstietojen perusteella voitu laskea vastaavaa tunnuslukua.

Tunnusluvuista muodostettiin painotettu summa kertoimilla 0,049; 0,021; -0,021; kertoimet saatiin estimointiaineistolla, jossa oli 31 konkurssiyritystä ja 59 toimivaa yritystä. Summaa verrattiin teoreettisin perustein saatuun kriittiseen arvoon -4,55 jonka alle jääneet yritykset luokiteltiin konkurssiyrityksiksi. Vertailuryhmässä, jossa oli 18 konkurssi- ja 28 tervettä yritystä, saatiin tällöin ennusteen virheiksi vuotta ennen konkurssia 21,7% (kokonaisvirhe), 0,0% (tyyppi I) ja 35,7% (tyyppi II) sekä neljää vuotta ennen konkurssia 42,1% (kokonaisvirhe), 60,0% (tyyppi I) ja 35,7% (tyyppi II).

Prihtin mukaan keskeiset konkurssiriskiin vaikuttavat tekijät ovat (1) tulorahoituksen riittävyys eli kannattavuus, (2) maksuvalmius sekä (3) vakavaraisuus; käytettyyn erottelufunktioon valitut tunnusluvut mittaavat juuri näitä. Monessa myöhemmässä tutkimuksessa konkurssiriskiä arvioidaan samojen kolmen tekijän avulla, vaikkakin niihin yhdistetyt tunnusluvut saattavat olla toiset.

Mielenkiintoinen seikka Prihtin tutkimuksessa on, että toisin kuin useissa aiemmissa tutkimuksissa – kuten Beaverin ja Altmanin – siinä ei käytetty vastinparimenettelyä, jossa jokaiselle aineistoon kuuluvalle konkurssiyritykselle etsitään sitä mahdollisimman paljon muistuttava toimiva yritys. Vastinparimenettelyä oli kritisoitu erityisesti sen vuoksi, että kullekin konkurssiyritykselle mahdollisimman samankaltaisen terveen yrityksen etsiminen vääristää terveiden yritysten otosta: esimerkiksi konkurssiyritysten ja terveiden yritysten suuruusluokat saattavat keskimäärin erota toisistaan, mutta vastinparimenettelyssä tämä ero ei tule esille. Lisäksi terveitä yrityksiä on huomattavasti enemmän kuin konkurssiyrityksiä, joten jos erottelufunktion kriittinen arvo valitaan vastinparimenettelyyn perustuvan otoksen perusteella niin, että kokonaisluokitteluvirhe minimoituu, saadaan harhainen estimaatti. Tosin myös Prihtin tutkimuksessa konkurssiyritykset olivat suhteellisesti selvästi yliedustettuina.

[1]Prihti määrittelee quickin seuraavasti: quick = rahoitusomaisuus - lyhytaikainen vieras pääoma. Kyseessä ei siis ole nykyisin yleisessä käytössä oleva quick ratio, joka määritellään seuraavasti: quick ratio = rahoitusomaisuus/(lyhytaikaiset velat - saadut ennakkomaksut)

2.3 Tuoreempia tutkimuksia

Edellä esiteltyjen tutkimusten jälkeen konkurssin ennustamista tilinpäätöstietojen perusteella on tutkittu runsaasti niin Suomessa kuin muuallakin. Tuoreemmista tutkimuksista tässä luodaan katsaus vain kahteen kotimaiseen työhön, joiden yhteydessä tarjoutuu mahdollisuus käsitellä useaa käsillä olevan tutkimuksenkin kannalta olennaista ongelmaa.

2.3.1 Laitinen: pienten ja keskisuurten yritysten konkurssit

Laitinen [2] on tutkinut perusteellisesti pienten ja keskisuurten yritysten konkursseja, jotka poikkeavat suuryritysten kaatumisesta monin tavoin. Myös käsillä oleva tutkimus keskittyy pk-yrityksiin; siksi on paikallaan tutustua tarkemmin Laitisen tekemiin havaintoihin.

Joitakin Laitisen mainitsemia pk-yritysten ominaispiirteitä on lueteltu seuraavassa; silloin kun aiheellista, lopussa on mainittu, miten Laitisen tutkimuksessa kyseinen seikka otettiin huomioon:

- Pk-yritysten joukossa on paljon nuoria ja nopeasti kasvavia yrityksiä, joiden toiminta poikkeaa selvästi vakiintuneemmista yrityksistä; näiden yritysten konkurssiriski on myös poikkeuksellisen suuri. – Aineistoon otettiin mukaan vain yli kymmenvuotiaita yrityksiä.

- Mukana on paljon erittäin pieniä yrityksiä, jotka eivät usein työllistä edes yhtä henkilöä täysipäiväisesti; myös nämä yritykset ovat selvästi tavanomaista alttiimpia konkursseille. – Pienimmät yritykset rajattiin aineiston ulkopuolelle.

- Pk-yritykset toimittavat usein tilinpäätöksensä puuttellisin erittelyin ja liitetiedoin, lykkäävät tilinpäätösten antamista tilikautta pidentämällä tai muutoin, tai jättävät tilinpäätöksen kokonaan antamatta. Tälläistä ilmenee terveilläkin pk-yrityksillä, mutta erityisen usein näin menettelevät vaikeuksiin ajautuneet yritykset. – Tutkimuksessa tarkasteltiin vuosien asemesta tilikausia ennen konkurssia, missä viimeinen tilikausi määriteltiin viimeisen ennen konkurssia annetun tilinpäätöksen mukaisesti.

- Konkurssit ovat selvästi yleisempiä pk-yrityksillä kuin suurilla yrityksillä, joten vertailuaineistona käytettävien terveiden yritysten joukossa saattaa olla runsaastikin sellaisia yrityksiä, jotka ovat periaatteessa konkurssiyrityksiä ja joiden konkurssi vain odottaa tapahtumistaan.

- Toisin kuin suuryritykset, pk-yritykset hakevat varsin usein itse itsensä konkurssiin. Tämä saattaa tapahtua jo ennen kuin yrityksen kunto on heikentynyt niin paljon, että se näkyisi tilinpäätöstiedoissa.

- Pk-yritysten joukossa on paljon tapauksia, joissa yritys haettiin konkurssiin, mutta joissa konkurssi raukesi esimerkiksi sovitteluratkaisun tai varojen puutteen vuoksi; näissä tapauksissa yritys yleensä lopetti toimintansa, mutta konkurssituomiota ei annettu. – Tutkimuksessa vain tuomitut konkurssit laskettiin konkursseiksi.

- Pk-yrityksen konkurssi aiheutuu suuryrityksiä useammin tekijöistä, jotka eivät suoraan liity yritykseen itseensä; esimerkkinä Laitinen mainitsee hyväkuntoisen yrityksen kaatumisen yhteen yllättävään, riittävän suureen luottotappioon.

Laitinen löysi pk-yritysten konkursseille kutakuinkin yhtä paljon syitä kuin aineistossa oli konkurssejakin. Konkurssien taustalla olevista tekijöistä useimmin esille tulivat huono yritysjohto, virheellinen investointi sekä idänkaupan supistuminen – tutkimusaineisto oli 1980-luvulta. Jos konkursseille pitäisi etsiä yhtä yhteistä nimittäjää, niin se olisi *yritysjohdon virheet, jotka johtavat heikkoon sopeutumiskykyyn.*

Laitinen havaitsi myös, että yritykset eivät käyttäydy samalla tavoin konkurssin lähestyessä. Faktorianalyysin avulla löytyi kolme erilaista konkurssiyritystyyppiä:

- *Krooninen konkurssiyritys*, jota vaivaa pitkäaikainen, lopulta konkurssiin johtava sairaus. Kaikki tunnusluvut heikkenevät tasaisesti konkurssia lähestyttäessä, joten tämä konkurssiyritystyyppi on helppo tunnistaa.

- *Tulorahoituskonkurssiyritys*, jolla vielä kolme tilikautta ennen konkurssia menee hyvin, mutta jolla sen jälkeen tulorahoitus äkillisesti romahtaa. Omavaraisuus ja maksuvalmius ovat tyypillisesti tyydyttävällä tasolla viimeisessäkin ennen konkurssia annetussa tilinpäätöksessä, joten näitä yrityksiä on vaikea tunnistaa Prihtin mallin mukaisella lineaarisella erottelufunktiolla, joka mittaa kannattavuutta, omavaraisuutta ja maksuvalmiutta.

- *Akuutti konkurssiyritys*, jolla kaikki tunnusluvut ovat kohtuullisella tasolla kahta tilikautta ennen konkurssia, mutta romahtavat sitten. Tämän tyypin yritykset on helppo tunnistaa viimeisen tilinpäätöksen perusteella, mutta aikaisempi tunnistaminen on erittäin vaikeaa.

Laitisen aineistossa eri tyyppisiä konkurssiyrityksiä oli kutakin suurin piirtein yhtä paljon. Vain ensimmäinen tyyppi oli helppo tunnistaa, joten Laitinen veti tästä sen johtopäätöksen, että konkurssin ennustaminen luotettavasti yli kaksi vuotta ennen konkurssia ei ole mahdollista pelkästään tilinpäätöstietojen perusteella.

Tutkimuksessa muodostettiin myös ennuste sille, meneekö yritys konkurssiin. Menetelmänä oli tässäkin tapauksessa lineaarinen erottelufunktio, mutta käytetyt tunnusluvut olivat toiset kuin Prihtin tai Altmanin malleissa. Ristiinvalidoiduiksi luokitteluvirheiksi saatiin yhtä tilikautta ennen konkurssia 13,8% (kokonaisvirhe), 15,0% (tyyppi I) ja 12,5% (tyyppi II) sekä neljää tilikautta ennen konkurssia 45,0% (kokonaisvirhe), 62,5% (tyyppi I) ja 27,5% (tyyppi II); aineisto koostui 40 konkurssiyrityksestä ja niille valituista mahdollisimman samanlaisista terveistä vastinpariyrityksistä.

2.3.2 Sundgren ja Wallin: logistinen regressio konkurssin ennustamisessa

Sundgrenin ja Wallinin tutkimus [6] poikkeaa monessa mielessä muista tässä esitetyistä konkurssitutkimuksista. Se perustui poikkeuksellisen laajaan aineistoon, noin kolmeen tuhanteen yritykseen, kun tyypillisessä konkurssitutkimuksessa aineisto kattaa vain joitakin kymmeniä yrityksiä. Siinä estimoitiin myös konkurssin todennäköisyyttä, sen sijaan että pelkästään luokiteltaisiin kaksiarvoisesti konkurssi- ja ei-konkurssiyrityksiin. Myös mallin perustana käytetty matemaattinen menetelmä, logistinen regressio (kts. esimerkiksi [7]), poikkeaa muista tässä esitellyistä.

Tutkimuksen otanta-asetelma oli teknisesti paremmin toteutettu kuin edellä esitellyissä tutkimuksissa. Käytetty aineisto koostui Tilastokeskuksen ositetulla satunnaisotannalla muodostaman yritysjoukon oikaisemattomista tilinpäätöksistä. Otoksen rakenne oli hyvin lähellä kaikkien suomalaisten yritysten muodostaman joukon rakennetta. Kuten perusjoukkoonkin, myös otokseen kuului joitakin yrityksiä, joiden tunnusluvut poikkesivat huomattavan paljon "normaaleista" yrityksistä. Nämä muodostivat ongelman, johon myös käsillä olevassa tutkimuksessa törmättiin; siksi on paikallaan tutustua Sundgrenin ja Wallinin kommentteihin tällaisten poikkeavien havaintojen *(outliers)*[2] käsittelystä.

Useimmat tilastolliset mallit – muun muassa edellä esitellyissä tutkimuksissa käytetty lineaarinen erotteluanalyysi tai tässä käytetty logistinen regressio – olettavat datan olevan multinormaalijakautunutta. Tilinpäätöstunnuslukujen kohdalla oletus ei kuitenkaan pidä paikkaansa; Sundgren ja Wallin viittaavat tässä Son (1987) sekä Ezzamelin ja Mar-Molineron (1991) tekemiin tutkimuksiin. Suuri osa normaalisuudesta poikkeamisesta voidaan laittaa poikkeavien havaintojen tiliin, vaikkakaan ne eivät kokonaan riitä ei-normaalisuutta selittämään. Seurauksena tästä on, että ilman poikkeavia havaintoja multinormaalijakautuneelle datalle suunnitellut mallit toimisivat yleensä riittävän hyvin.

Sundgren ja Wallin viittaavat kirjallisuudessa esitettyihin menetelmiin poikkeavien havaintojen vaikutuksen eliminoimiseksi. Poikkeavat havainnot voidaan yksinkertaisesti poistaa datasta, mihin Sundgren ja Wallin päätyivät, tai niille voidaan antaa uusiksi arvoiksi lähimpänä sijaitsevien "normaalitapausten" arvot; datalle voidaan myös tehdä erilaisia muunnoksia, jotka saavat sen jakauman lähemmäs multinormaalistat. Menetelmiä on kritisoitu siitä, että niissä joko ongelman olemassaolo kokonaan kielletään, tai ne ovat liian ad hoc -luonteisia [8]. Vasta-argumenttina tähän on ehkä lupa todeta, että luonteeltaan toisenlaista dataa varten laadittujen tilastollisten menetelmien suora soveltaminen tilinpäätösaineistoon on sekin melko ad hoc -luonteinen ratkaisu. Käytössä oleva tilinpäätösmuoto ja siitä johdetut tunnusluvut perustuvat käytännössä erittäin toimivaksi osoittautuneeseen konventioon, mutta tämä ei suinkaan tarkoita sitä, etteikö tunnuslukuja voisi muokata toisenlaisiksi, jos se palvelee kulloisiakin päämääriä. Siten kaikki edellä esitetyt menetelmät ovat perusteltuja, jos niiden avulla pystytään luotettavammin luokittelemaan tilinpäätökset konkurssiyritysten antamiin ja terveiden yritysten antamiin. Luokitustulokset olisi kuitenkin aina evaluoitava perusjoukkoa hyvin edustavalla otoksella – jos tässä vaiheessa kaikkia tilinpäätöksiä ei voida käsitellä konstruoidulla luokittimella, ne on pystyttävä käsittelemään jollakin muulla tavalla, ja myös ne on otettava huomioon luokitustuloksissa.

Tutkimuksessa ennustettiin konkurssin todennäköisyyttä kahden-kolmen vuoden

[2]Sanan *outlier* suomenkielisiä vastineita ovat tässä käytetyn "poikkeavan havainnon" lisäksi "vieras havainto" ja "ulkopuolinen havainto".

kuluttua tilinpäätöksen antamisesta. Kaksi eri mallia muodostettiin: mallissa 1 osa tunnusluvuista muunnettiin lähemmäs normaalijakaumaa, mallissa 2 kaikki tunnusluvut otettiin mukaan sellaisenaan. Molemmat mallit perustuivat oikaisemattomiin tilinpäätöksiin, ilmeisesti käytännön pakosta, vaikka esimerkiksi Prihti oli omassa tutkimuksessaan todennut tilinpäätösten oikaisemisen parantavan ennustustarkkuutta.

Mallien avulla sekä estimoitiin konkurssin todennäköisyyttä että luokiteltiin yritykset konkurssiyrityksiin ja terveisiin yrityksiin. Käytännön tarpeiden kannalta konkurssin todennäköisyysestimaatti on luonnollisesti paljon arvokkaampi kuin pelkkä binäärinen luokittelutulos. Malli 1 tuottikin varsin hyvät konkurssitodennäköisyysestimaatit estimoinnissa käytetylle aineistolle. Sen sijaan testiaineistolla mallin tarkkuus oli heikohko: χ^2-testillä mitattuna nollahypoteesi, jonka mukaan estimoidut todennäköisyydet vastaisivat toteutuneita frekvenssejä, päädyttiin hylkäämään riskitasolla 0,1.

Binääriluokittelutulosten ilmoittamisessa on otettava huomioon, että konkurssiyrityksiä oli huomattavasti vähemmän kuin terveitä. Siksi kokonaisluokittelutarkkuus parani kutakuinkin samaa tahtia kuin tyypin II virhe pieneni ja tyypin I virhe kasvoi. Tulokset ilmoitettiin usealla eri virhetyyppien suhteellisilla kustannuksilla; tässä mainittakoon vain ne, jotka saatiin vertailuaineistolle kun eri tyyppien virheet olivat kutakuinkin yhtä suuret. Mallilla 1 virheet olivat 26,4% (kokonaisvirhe), 26,2% (tyyppi I) ja 31,5% (tyyppi II); mallilla 2 ne olivat 28,4% (kokonaisvirhe), 28,5% (tyyppi I) ja 36,0% (tyyppi II).

2.4 Konkurssin ennustaminen neurolaskennan menetelmin

Neurolaskennan kehittyminen 1980-luvulla toi sen piirissä syntyneet menetelmät pian työkaluiksi tilastolliseen analyysiin. 1990-luvun alusta lähtien neurolaskentamenetelmiä on sovellettu useassa eri tutkimuksessa (mm. [9, 10, 11, 12, 13, 14, 15, 16, 17, 18]) myös konkurssien ennustamiseen tilinpäätöstietojen perusteella. Valtaosa tutkimuksista perustuu ns. monikerrosperceptron-verkkoon *(multilayer perceptron, MLP)*, mutta myös itseorganisoituvaa karttaa on sovellettu ongelmaan ainakin kahdessa tapauksessa.

2.4.1 Konkurssiennusteet monikerrosperceptron-verkolla

Monikerrosperceptron-verkolla approksimoidaan tuntematonta funktiota $t(\mathbf{x})$, joka tässä tapauksessa kuvaa konkurssin todennäköisyyttä tilinpäätöstietovektorin $\mathbf{x}$ perusteella, verkon antamalla vasteella $y(\mathbf{x})$. Verkko opetetaan approksimoitavasta funktiosta tunnetuilla näytteillä $(\mathbf{x}^i, t^i)$ ns. backpropagation-opetussääntöä käyttäen. Opetuksessa pyritään löytämään verkolle sellaiset parametrit, että verkon vasteen neliöllinen virhesumma

$$\mathcal{E} = \frac{1}{2}\sum_i [y(\mathbf{x}^i) - t^i]^2 \tag{2.2}$$

minimoituu. Hyvä MLP-verkon ja backpropagation-opetussäännön esittely löytyy esimerkiksi viitteestä [19].

Yhteenveto kiinnostavimmista MLP-verkolla toteutetuista konkurssitutkimuksista on esitetty taulukossa 2.1. Useissa tutkimuksissa verrattiin MLPn ennustamiskykyä johonkin muuhun menetelmään; seuraavassa katsaus saatuihin tuloksiin

Tam ja Kiang [11] vertailivat keskenään MLPtä sekä ilman piiloneuroneita (MLP_0) että kymmenellä piiloneuronilla (MLP_{10}), lineaarista diskriminanttianalyysiä (LDA), logistista regressiota (logit), yhden ja kolmen lähimmän naapurin menetelmiä (1NN ja 3NN) sekä puuluokitinta (ID3) konkurssin ennustamisessa. Menetelmiä sovellettiin varsin standardinomaisesti; poikkeuksen teki MLPn opetusalgoritmi, jossa perus-backpropagation-algoritmia oli muunnettu siten, että virhefunktiossa otettiin huomioon tyypin I ja II luokitteluvirheitten erilaiset kustannukset. Ristiinvalidoidut tulokset ennustettaessa tilannetta seuraavana vuonna antoivat näille menetelmille seuraavan paremmuusjärjestyksen: MLP_{10}, MLP_0, LDA, logit, ID3, 3NN, 1NN. Kahden vuoden ennusteessa saatiin muuten samat tulokset, mutta LDA oli ohittanut MLP_0n. Yleensä varsin hyvin toimivien lähimmän naapurin menetelmien huonot tulokset selittänee tässä tapauksessa käytössä olleen aineiston suppeus – nämä menetelmät edellyttävät varsin laajaa opetusaineistoa, jotta niiden tulokset olisivat yleistettävissä myös opetusaineiston ulkopuolelle.

Udo [14] vertasi MLPtä lineaariseen regressioon, Fletcher ja Goss [12] logistiseen regressioon. Kummassakin tutkimuksessa todettiin MLPllä saatavan hivenen parempia tuloksia, vaikkakaan erot eivät olleet merkittäviä.

Taulukko 2.1: Yhteenveto joistakin MLPllä toteutetuista konkurssiennusteista (#TL = syötetietoina käytettyjen tunnuslukujen määrä; t = ennusteen kantama vuosina)

tekijä	#TL	opetusjoukko	testijoukko	t	virhe I	virhe II	kok.v.
Fletcher	3	18/18	ristiinval.	1			17,6
Tam	19	81/81	ristiinval.	1	8,6	12,3	10,5
				2	8,9	12,7	10,8
Udo	16	150/150	50/50	1	14,0	10,0	12,0
				2	16,0	18,0	17,0
Raghupathi	13	25/25	26/26	1			14,0
Tsukuda	11	44/44	12/12	1	16,7	13,0	14,9
Cadden	12	49/49	10/10	3	20,0	10,0	15,0

2.4.2 Konkurssiennusteet itseorganisoivalla kartalla

Itseorganisoivaa karttaa sovelsivat tilinpäätösanalyysiin tiettävästi ensimmäisinä Martín-del-Brío ja Serrano-Cinca [9]. Tutkimuksessa opetettiin kartta tilinpäätöstietojen perusteella, jonka jälkeen karttayksiköt leimattiin niille kuvautuvien yritysten enemmistön perusteella joko konkurssiyksiköiksi tai terveiksi yksiköiksi (samaa menetelmää on käytetty myös käsillä olevassa tutkimuksessa; yksityiskohtaisemmin se on esitelty kappaleessa 3.4.1). Tutkimalla karttayksiköitten arvoja komponenteittain (katso kappale 3.3) analysoitiin eri tunnuslukujen merkitystä konkurssiriskille ja tunnistettiin erilaisia yritystyyppejä. Lisäksi tutkimuksessa tarkasteltiin yritysten siirtymiä kartalla useampana perättäisenä tilivuotena.

Jatkotutkimuksessaan [20] Serrano-Cinca käytti kartalle muodostuneita konkurssi- ja terveiden yritysten alueita konkurssiyritysten luokitteluun. Menetelmää verrattiin LDAhan sekä useaan erilaiseen MLP-versioon siten, että mallit opetettiin 74 yrityksen aineistolla, jonka jälkeen ne evaluoitiin 55 yrityksen testiaineistolla; kummastakin yritysjoukosta noin puolet oli konkurssiyrityksiä. Kokonaisluokitteluvirheiksi saatiin LDAlle 25% ja muille malleille 18%.

Backin et al. tutkimus [10], muistutti monessa mielessä Serrano-Cincan työtä: SOMia käytettiin molemmissa samalla tavoin konkurssiyritysten luokitteluun ja tuloksia verrattiin MLP-verkkoon; sen sijaan LDA:n asemesta Back et al. ottivat kolmanneksi vertailumalliksi Boltzmannin kone -neuroverkkomallin (kts. esimerkiksi [19]). Vertailussa havaittiin MLPn toimivan luokittimena hiukan kahta muuta mallia paremmin.

2.5 Kommentteja aiemmista tutkimuksista

Edellä esiteltyjä tutkimuksia voi tarkastella monelta eri kannalta. Tutkimuksen tavoitteeksi voi asettaa yrityksen itsensä ja rahoittajien kannalta mahdollisimman luotettavan ennusteen tai varoitusjärjestelmän tuottamisen. Toisaalta pyrkimyksenä voi olla konkurssi-ilmiön valaiseminen. Kolmantena vaihtoehtona on yksinkertaisesti eri menetelmien vertailu ja testaus johonkin kiinnostavaan ongelmaan.

Ennuste tai varoitus on sitä arvokkaampi, mitä pidemmälle tulevaisuuteen sen tarkkuus säilyy käyttökelpoisena. Viimeisessä tilinpäätöksessä konkurssin oireet ovat yleensä selvimmin näkyvissä, mutta tässä vaiheessa ennusteen hyöty alkaa olla kyseenalainen. Siksi olisi ehkä perusteltua antaa pidemmän aikavälin ennusteille suurempi painoarvo kuin yleensä on tehty. Toinen seikka, mikä heikentää raportoitujen mallien käyttöarvoa, on niiden perustana oleva yritysaineisto, joka vain harvassa tapauksessa oli edustava otos kaikkien yritysten joukosta. Seurauksena saattaa olla liian optimistinen kuva mallien suorituskyvystä.

Konkurssi-ilmiön tarkastelun kannalta käytetyillä malleilla on suuria eroja. Jotkin mallit tuottavat pelkän ennusteen, jossa eri tekijöiden vaikutusta konkurssiriskiin on varsin hankalaa tulkita; toiset taas soveltuvat luonnostaan myös itse ilmiön syiden selvittelyyn. Jos näillä malleilla ei ole merkittävää eroa ennustustarkkuudessa, lienee perusteltua käyttää jälkimmäiseen ryhmään kuuluvaa mallia.

Eri tilastollisten menetelmien vertailu oli aiheena useassa esitellyssä tutkimuksessa. Saadut tulokset menettivät kuitenkin monessa tapauksessa merkityksensä käytetyn aineiston pienuuden vuoksi. Usein mallit evaluoitiin vain kerran muutaman kymmenen yrityksen aineistolla, jolloin jo yhden yrityksen luokittelu toisin muutti luokitteluvirhettä usean prosenttiyksikön verran. Ristiinvalidointi olisi monessa tapauksessa parantanut tulosten luotettavuutta.

Luku 3

Itseorganisoiva kartta (SOM)

Itseorganisoiva kartta *(Self-Organizing Map, SOM)* on yksi tunnetuimpia neurolaskentamalleja. Perusmuodossaan, jonka esitteli Kohonen vuonna 1982 [21], SOM kuuluu ns. ohjaamattoman oppimisen neurolaskentamenetelmiin, joista se lienee käytetyin.

Ohjaamattoman oppimisen menetelmissä neuroverkon avulla etsitään alkuperäiselle datalle jokin toinen esitysmuoto, joka säilyttää mahdollisimman uskollisesti alkuperäisen datan olennaiset piirteet mutta soveltuu jatkokäsittelyyn paremmin kuin se. SOM kuvaa alkuperäisen datajoukon karttahilalle pyrkien säilyttämään kuvauksessa naapurustosuhteet. Karttahilan dimensio on tyypillisesti matalampi kuin syöteavaruuden, jolloin SOMin voi ajatella muodostavan eräänlaisen epälineaarisen regression. Etenkin jos kartan dimensioksi valitaan kaksi, datajoukon visuaalinen tarkastelu onnistuu hyvin.

3.1 SOM-algoritmi

Merkitään tarkasteltavasta ilmiöstä saatuja havaintovektoreita $\mathbf{x}^i \subset \mathcal{X}, \quad i = 1, \ldots, N$ missä N on havainto- eli syötevektoreiden määrä, $\mathcal{X} \subset \mathbb{R}^d$ on syöteavaruus ja d sen dimensio. Muodostetaan lisäksi karttahila $\mathcal{L}$, joka koostuu yksiköistä $\mathbf{n}_j$. Karttahilan dimension ja hilarakenteen voi valita vapaasti, kunhan karttayksiköille on mahdollista määritellä metriikka; usein valitaan hilan dimensioksi kaksi, rakenteeksi suorakulmainen tai heksagonaalinen hila ja metriikaksi euklidinen metriikka.

Jokaisella karttayksiköllä $\mathbf{n}_j$ on painovektori $\mathbf{m}_j \in \mathbb{R}^d$, jonka dimensio on siis sama kuin syötevektoreiden. Painovektori liittää jokaiseen karttayksikköön vastinpisteen syöteavaruudessa, jolloin kullekin syötevektorille $\mathbf{x}^i$ voidaan hakea sitä parhaiten vastaava yksikkö $\mathbf{n}_c$, missä alaindeksi c saadaan kaavasta

$$c = \underset{j}{\operatorname{argmin}}(||\mathbf{x}^i - \mathbf{m}_j||) \tag{3.1}$$

Kaavassa (3.1) voidaan periaatteessa käyttää mitä hyvänsä metriikkaa; useimmiten tässäkin valitaan euklidinen metriikka.

SOMin opetuksessa painovektorit $\mathbf{m}_j$ alustetaan ensin satunnaisiin arvoihin syöteavaruudessa X. Syötevektorit käydään läpi vuorotellen useita kertoja siten, että kullekin syötevektorille $\mathbf{x}^i$ etsitään sitä parhaiten vastaava yksikkö $\mathbf{n}_c$, ja tämän jälkeen siirretään sekä parhaiten vastaavan yksikön $\mathbf{n}_c$ että sen naapuriyksiköiden painovektoreita jonkin verran syötevektoria $\mathbf{x}^i$ kohti:

$$\mathbf{m}_j := \mathbf{m}_j + \alpha h(c,j)(\mathbf{x}^i - \mathbf{m}_j) \tag{3.2}$$

Oppimisnopeusparametri $\alpha = \alpha(t)$ on tyypillisesti vähenevä ajan t funktio. Naapurustofunktio $h(c,j) = h(c,j;t)$ on vähenevä yksiköiden $\mathbf{n}_c$ ja $\mathbf{n}_j$ välisen etäisyyden $||\mathbf{n}_c - \mathbf{n}_j||$ funktio, joka lisäksi muuttuu ajan t mukana. Lisäksi vaaditaan, että $0 < \alpha < 1$ ja $0 \leq h \leq 1$.

Yleisesti käytettyjä naapurustofunktioita ovat kuplanaapurusto

$$h(c,j) = \begin{cases} 1 & \text{kun } ||\mathbf{n}_c - \mathbf{n}_j|| \leq \sigma \\ 0 & \text{muulloin} \end{cases} \tag{3.3}$$

sekä gaussinen naapurusto

$$h(c,j) = \exp\left(-\frac{||\mathbf{n}_c - \mathbf{n}_j||^2}{2\sigma^2}\right) \tag{3.4}$$

Kummassakin tapauksessa naapurustofunktion leveyden määrää parametri $\sigma = \sigma(t)$.

Edellä kuvattua perus-SOM-algoritmia voidaan muunnella lukuisin tavoin. Kartan topologia voidaan määritellä suorakulmaisen verkon sijasta heksagonaalisena hilana, kartan dimensio voi vaihdella, on mahdollista muodostaa toroidinmuotoisia karttoja tai jopa vaihdella hilarakennetta ja dimensiota kartan eri osissa. Oppimisnopeusparametri α voi pienetä lineaarisena tai käänteisenä ajan funktiona, samoin sen alku- ja loppuarvot on mahdollista valita varsin vapaasti, kunhan vain pieneneminen ei tapahdu liian nopeasti tai loppuarvo jää liian suureksi. Sama pätee naapurustofunktion leveysparametriin σ. Algoritmi on hämmästyttävän robusti eri parametrivalintojen suhteen.

Käytännössä hyvin toimivaksi on osoittautunut kartan opettaminen kahdessa vaiheessa [22]. Ensimmäisessä vaiheessa tavoitteena on saada kartta järjestymään. Siksi α voi olla aluksi suhteellisen suuri, jopa lähellä yhtä, ja naapurustofunktion on oltava riittävän leveä; hyvä valinta σ:n alkuarvoksi on noin puolet kartan halkaisijasta. Kumpikin parametri pienenee kohti nollaa, esimerkiksi lineaarisena ajan funktiona. Ensimmäisen vaiheen pituus voi olla 1 000 ... 10 000 opetusaskelta, kun yhdellä opetusaskeleella tarkoitetaan yhden syötevektorin esittämistä kartalle ja kartan opettamista sen perusteella.

Toisessa vaiheessa tavoitteena on saada järjestynyt kartta sovittautumaan dataan mahdollisimman hyvin. Parametrien alkuarvojen tulee olla selvästi pienemmät kuin ensimmäisessä vaiheessa, esimerkiksi $\alpha_0 = 0,02$ ja $\sigma = 3$, joista parametrit pienenevät opetuksen edistyessä nollaan. Koska oppiminen tapahtuu toisessa vaiheessa hitaammin kuin ensimmäisessä vaiheessa, opetusta on jatkettava selvästi pitempään: opetusaskelia on syytä olla vähintään 10 000, mutta niitä voi hyvin olla enemmänkin, esimerkiksi 100 000 askelta on yleinen valinta.

3.2 SOMin käyttöön liittyviä kysymyksiä

3.2.1 Datan esikäsittely

Syötevektoreiden komponentit voivat sisältää hyvin erityyppistä dataa, ja komponenttien varianssit saattavat vaihdella paljonkin. Tällöin se komponentti, jonka varianssi on suurin, dominoi euklidisia etäisyyksiä laskettaessa, jolloin sillä on myös suurin vaikutus kartan muotoutumiseen. Joskus näin halutaankin tapahtuvan; joskus taas kaikille vektorikomponenteille halutaan antaa sama painoarvo. Siksi sovelluskohtaisesti on mietittävä, miten vektorit esikäsitellään. Joitakin vaihtoehtoja ovat varianssin normalisointi, yksikköjananormalisointi sekä histogrammin tasoitus.

Varianssin normalisointi

Yleinen ja toimivaksi havaittu [23, 22] tapa esikäsitellä syötevektorit on normalisoida niiden varianssit yhteen; samalla yleensä vektorit muunnetaan nollakeskiarvoisiksi.

Tarkastellaan syötevektoreiden $\mathbf{x}^i$, $i = 1, \ldots, N$ komponenttia k. Sen keskiarvo $\overline{x}_k$ on

$$\overline{x}_k = \frac{1}{N} \sum_{i=1}^{N} x_k^i \tag{3.5}$$

ja keskihajonta S_k

$$S_k = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_k^i - \overline{x}_k)^2} \tag{3.6}$$

Tällöin varianssin normalisointi voidaan esittää muodossa

$$x_k^i := \frac{x_k^i - \overline{x}_k}{S_k} \tag{3.7}$$

Yksikköjananormalisointi

Varsin paljon varianssin normalisoinnin kaltainen esikäsittelytapa on siirtää ja skaalata vektorit yksikköjanalle, jolloin kunkin komponentin pienin arvo on 0 ja suurin 1. Merkitään syötevektoreiden k:nnen komponentin pienintä arvoa $x_k^{\min}$ ja suurinta $x_k^{\max}$. Tällöin yksikköjananormalisointi voidaan esittää muodossa

$$x_k^i := \frac{x_k^i - x_k^{\min}}{x_k^{\max} - x_k^{\min}} \tag{3.8}$$

Yksikköjananormalisointi eroaa varianssin normalisoinnista selvimmin siinä, miten muusta aineistosta poikkeavat arvot vaikuttavat tulokseen. Tarkastellaan esimerkin vuoksi sadan syötevektorin tapausta, jossa vektoreiden k:s komponentti on muutoin jakautunut tasan välille $[0; 10]$, mutta joukossa on yksi poikkeava vektori $\mathbf{x}^*$, jolle

$x_k^* = 1\ 000$. Tällöin yksikköjananormalisoitujen ei-poikkeavien vektoreiden k:nnet komponentit ovat hyvin lähellä toisiaan (välillä [0; 0,01]), kun varianssin normalisointi säilyttää niiden välisiä eroja enemmän (komponentit sijaitsevat välillä [-0,15; -0,05]). Sen sijaan poikkeava komponentti, jolle yksikköjananormalisointi antaa arvoksi 1, jää varianssin normalisoinnissa selvästi kauemmas, pisteeseen 9,90. Yksikköjananormalisointi siis hukkaa ei-poikkeavien vektoreiden eroja; varianssin normalisointi taas jättää yhdeltä komponentiltaan poikkeavat vektorit kauas muista, vaikka niiden muut komponentit eivät muusta aineistosta poikkeaisikaan.

Histogrammin tasoitus

Histogrammin tasoituksessa ei kiinnitetä lainkaan huomiota siihen, kuinka kaukana vektorit komponenteittain toisistaan ovat, vaan ratkaisevaa on se, missä järjestyksessä ne ovat. Tarkastellaan tilannetta, jossa on N syötevektoria, joiden k:nnen komponentin histogrammi tasoitetaan q osaan. Järjestetään vektorit siten, että $x_k^1 \leq x_k^2 \leq \cdots \leq x_k^N$. Tällöin ensimmäiset $\lceil N/q \rceil$ vektoria saavat k:nnen komponenttinsa uudeksi arvoksi 1; tässä $\lceil \cdot \rceil$ tarkoittaa pyöristystä ylöspäin lähimpään kokonaislukuun, joka on suurempi tai yhtäsuuri kuin argumentti. Seuraavat $\lceil N/q \rceil$ vektoria saavat k:nnen komponenttinsa uudeksi arvoksi 2 ja niin edelleen. Katkaisukohdat valitaan siten, että vektorit, joilla on yhtäsuuri k:s komponentti saavat aina saman arvon, ja seuraava $\lceil N/q \rceil$ vektorin erä alkaa vasta näiden jälkeen.

Lopputuloksena on, että syötevektoreiden reunajakaumat ovat lähes tasaisia välillä $[1, q']$ missä $q' \leq q$ – se, että reunajakaumista ei välttämättä tule tasaisia välillä $[1, q]$, johtuu mahdollisesta yhtäsuurten vektoreiden esiintymisestä katkaisukohdissa. On ehkä syytä todeta, että reunajakaumien tasaisuus ei suinkaan merkitse sitä, että vektorit $\mathbf{x}^i$ olisivat tasan jakautuneita syöteavaruuteen; yleensä histogrammin tasoitus komponenteittain kuitenkin tasoittaa myös itse jakaumaa.

3.2.2 SOMin muodostaman kuvauksen hyvyyden arviointi

SOM muodostaa kuvauksen syöteavaruudesta $\mathcal{X}$ karttahilalle $\mathcal{L}$. Syöteavaruuden $\mathcal{X} \subset \mathbb{R}^d$ geometrinen dimensio on aina d eli sama kuin syötevektoreiden pituus. Sen sijaan syöteavaruuden $\mathcal{X}$ sisäinen dimensio *(intrinsic dimension)* $d_{\mathcal{X}}$ on usein selvästi pienempi: tarkasteltavan ilmiön tila liikkuu jossakin monistossa, jonka kohinainen kuva syöteavaruus $\mathcal{X}$ on. Jättäen kohina huomioimatta voidaan syöteavaruuden $\mathcal{X}$ sisäiseksi dimensioksi määritellä tuon moniston dimensio.

Jos karttahilan $\mathcal{L}$ dimensio $d_{\mathcal{L}}$ on sama kuin $d_{\mathcal{X}}$, SOMin muodostama kuvaus $f : \mathcal{X} \rightarrow \mathcal{L}$ voi muistuttaa homeomorfismia: kuvaus on jatkuva,[1] ja sillä on jatkuva käänteiskuvaus. Jos sen sijaan $d_{\mathcal{L}} < d_{\mathcal{X}}$, näin ei tapahdu. Kun naapurustofunktio on riittävän kapea, kartta pyrkii approksimoimaan syöteavaruuden dimensiota laskostumalla [22, 24], jolloin kuvaukseen syntyy epäjatkuvuuskohtia. Leveällä naapurustofunktiolla puolestaan kuvaus kartalle pysyy jatkuvana, mutta käänteiskuvaus kartalta

[1] Tässä jätetään huomiotta karttahilan diskreettisyydestä väistämättä aiheutuvat epäjatkuvuuskohdat; tarkastellaan tilannetta ikäänkuin kartta olisi $d_{\mathcal{L}}$-dimensioinen yksikkökuutio.

syöteavaruuteen kärsii: kartan erottelukyky muodostuu huonoksi, karttayksiköiden alkukuvat eivät ole yksikäsitteisiä.

Jatkuvuus ja erottelukyky ovat siis toistensa kanssa ristiriitaisia ominaisuuksia silloin, kun $d_{\mathcal{L}} < d_{\mathcal{X}}$. Näiden ominaisuuksien painotus vaihtelee sovelluksesta toiseen, joten niiden mittaaminen erikseen on perusteltua. Toisaalta niiden tarkkailu antaa hyödyllistä tietoa kuvauksen ominaisuuksista, esimerkiksi kartan järjestymisestä ja sovittumisesta dataan, myös silloin, kun $d_{\mathcal{L}} = d_{\mathcal{X}}$. Kuvauksen jatkuvuutta ja erottelukykyä voidaan käyttää kriteerinä parhaan kartan valitsemiseen kokeillen useita erilaisia alustuksia, hilamuotoja sekä opetusparametreja.

Erottelukyvyn kvantifiointi

Kuvauksen erottelukyvyn mitaksi sopii *kvantisointivirhe* $\mathcal{E}_q$. Sillä tarkoitetaan syötevektoreiden keskimääräistä etäisyyttä lähimmästä karttayksiköstä:

$$\mathcal{E}_q = \frac{1}{N} \sum_{i=1}^{N} \min_j ||\mathbf{x}^i - \mathbf{m}_j|| \tag{3.9}$$

Jatkuvuuden kvantifiointi

Jatkuvuuden kvantifioimiseksi on esitetty useita erilaisia menetelmiä [25, 26, 27]. Tässä esitellään näistä vain yksi, *topografinen virhe* $\mathcal{E}_t$, jolla on tiettyjä etuja muihin menetelmiin nähden [27]. Määritellään ensin epäjatkuvuuskohtaa indikoiva funktio u seuraavasti:

$$u(\mathbf{x}^i) = \begin{cases} 1, & \text{kun } ||\mathbf{n}_{c_1} - \mathbf{n}_{c_2}|| > \sqrt{d_{\mathcal{L}}}, \\ & \text{missä } c_1 = \operatorname{argmin}_j ||\mathbf{m}_k - \mathbf{x}^i|| \text{ ja } c_2 = \operatorname{argmin}_{k \neq j} ||\mathbf{m}_k - \mathbf{x}^i|| \\ 0, & \text{muulloin} \end{cases} \tag{3.10}$$

Toisin sanoen, funktio $u(\mathbf{x}^i) = 1$ täsmälleen silloin, kun syötevektoria $\mathbf{x}^i$ lähimpänä ja toiseksi lähimpänä sijaitsevaa painovektoria vastaavat karttayksiköt eivät sijaitse vierekkäin, mikä osoittaa, että jossakin kohtaa näiden painovektoreiden välillä kuvauksessa on epäjatkuvuuskohta. Tässä vierekkäisiksi katsotaan myös "kulmittain vierekkäiset" eli kaksidimensioisessa tapauksessa 8-naapurustoon kuuluvat yksiköt, joiden välinen euklidinen etäisyys on suorakulmaisella kartalla juuri d_L; heksagonaalista karttahilaa käytettäessä d_L kaavassa (3.10) korvataan luvulla 1.

Topografinen virhe $\mathcal{E}_t$ määritellään nyt niiden näytevektoreiden suhteellisena osuutena, joilla on lokaalissa ympäristössään epäjatkuvuuskohta:

$$\mathcal{E}_t = \frac{1}{N} \sum_{i=1}^{N} u(\mathbf{x}^i) \tag{3.11}$$

3.3 SOMia täydentäviä työkaluja

SOM sopii perusmuodossaan ennenkaikkea datan visualisointiin ja ryhmittelyyn (klusterointiin) sekä vektorikvantisointiin. SOMin käyttöä tukemaan on Teknillisen korkeakoulun Neuroverkkojen tutkimusyksikössä kerätty ja kehitetty joukko erilaisia menetelmiä, jotka on koottu SOM_PAK-nimiseen ohjelmistopakettiin [28]. Joitakin näistä "SOM-perustyökaluista" esitellään lyhyesti seuraavassa.

Sammonin projektio

Kartan järjestymistä voidaan topografisen virheen $\mathcal{E}_t$ lisäksi tarkastella *Sammonin projektion* [29] avulla, joka on yksi versio moniulotteisesta skaalauksesta. Sammonin projektiossa kartan d-dimensioiset painovektorit kuvataan tasoon siten, että niiden väliset etäisyyssuhteet vastaavat mahdollisimman hyvin vektoreiden alkuperäisten euklidisten etäisyyksien välisiä suhteita d-dimensioisessa avaruudessa. Yhdistämällä Sammonin projektiossa karttanaapurit toisiinsa viivoilla nähdään, onko kartta hyvin järjestynyt: mitä enemmän viivat menevät ristiin, sitä heikommin kartta on järjestynyt.

Sammonin projektiolla voidaan myös tarkastella syöteavaruuden muotoa sekä mahdollista jakautumista erillisiin klustereihin, minkä perusteella pystytään etsimään syöteavaruuteen parhaiten sopivat kartan mittasuhteet. – SOM_PAKissa Sammonin projektio muodostetaan ohjelmalla "sammon".

Kartan kalibrointi

Opetetun kartan yksiköille voidaan antaa erilaisia attribuutteja sen perusteella, millaista dataa niille kuvautuu: kartta tavallaan kalibroidaan datan avulla. Käsillä olevassa sovelluksessa karttayksiköt voidaan esimerkiksi jakaa konkurssiyksiköihin ja terveisiin yksiköihin sen perusteella, kuvautuuko niille enemmän konkurssi- vai terveiden yritysten tilinpäätöksiä. – SOM_PAKissa kartan kalibrointi tehdään ohjelmalla "vcal".

Painovektoreiden välisten etäisyyksien visualisointi kartalla

Karttayksiköt on tyypillisesti sijoiteltu säännöllisen hilan muotoon. Sen sijaan niiden painovektorit syöteavaruudessa saattavat datasta riippuen sijaita hyvinkin epäsäännöllisesti. Jos syöteavaruus koostuu useasta ryhmästä (klusterista), jotka ovat selvästi erillisiä, hyvin järjestyneen kartan painovektorit jakautuvat myös näihin ryhmiin: samassa ryhmässä sijaitsevat painovektorit ovat lähellä toisiaan mutta kaukana muissa ryhmissä sijaitsevista painovektoreista. Ryhmittyminen saadaan näkymään kartalla ns. *umat*-esitysmuodon *(unified distance matrix)* [30, 31, 32] avulla, jossa karttapohja kuvataan sitä tummempana, mitä kauempana vierekkäisten karttayksiköiden painovektorit toisistaan ovat. – SOM_PAKissa kartan umat-esitys muodostetaan ohjelmalla "umat".

Kartan tarkastelu komponenteittain

Yksittäisten vektorikomponenttien vaikutusta kartan muotoutumiseen sekä komponenttien keskinäisiä riippuvuuksia voidaan tarkastella seuraavasti: värjätään mustaksi se karttayksikkö, jossa tarkasteltava komponentti on pienimmillään, valkoiseksi se yksikkö, jossa tarkasteltava vektorikomponentti on suurimmillaan, ja muut yksiköt saavat tältä väliltä komponentin arvoa vastaavan harmaasävyn. – SOM_PAKissa kuvat kartan komponenttitasoista muodostetaan ohjelmalla "planes".

3.4 SOM ohjatussa oppimisessa

Perusmuodossaan SOM on puhdas ohjaamattoman oppimisen menetelmä. Useissa sovelluksissa kuitenkin halutaan approksimoida jotakin tuntematonta funktiota $\mathbf{t}(\mathbf{x})$, josta kuitenkin tunnetaan joukko näytteitä eli $(\mathbf{x}^i, \mathbf{t}^i)$-pareja. Näiden avulla halutaan opettaa laadittava malli niin, että sen antama vaste pisteissä $\mathbf{x}^i$ eli $\mathbf{y}(\mathbf{x}^i)$ on mahdollisimman lähellä tavoitearvoa $\mathbf{t}^i$. Tällöin on kysymys ohjatusta oppimisesta.

SOMia voidaan soveltaa ohjattuun oppimiseen usealla eri tavalla; seuraavissa kappaleissa esitellään tässä tutkimuksessa käytetyt.

3.4.1 SOM-1

Suoraviivaisin tapa soveltaa SOMia funktion approksimointiin perustuu siihen, että SOM-algoritmilla syöteavaruus $\mathcal{X}$ saadaan jaettua erillisiin laareihin *(bins):* kullekin karttayksikölle $\mathbf{n}_j$ voidaan osoittaa se osa syöteavaruutta, joka kuuluu karttayksikön painovektorin $\mathbf{m}_j$ Voronoi-ympäristöön $\mathcal{V}_j$,

$$\mathcal{V}_j = \{\mathbf{v} \in \mathcal{X} \mid ||\mathbf{v} - \mathbf{m}_j|| < ||\mathbf{v} - \mathbf{m}_k||,\ k \neq j\}, \tag{3.12}$$

eli siihen alueeseen syöteavaruutta, jossa painovektori $\mathbf{m}_j$ on lähempänä kuin mikään toinen painovektori $\mathbf{m}_k$.

Tavoitefunktiota $\mathbf{t}(\mathbf{x})$ voidaan tällöin approksimoida antamalla $\mathbf{y}(\mathbf{x})$:n arvoksi kussakin Voronoi-ympäristössä yksinkertaisesti $\mathbf{t}$:n odotusarvo kyseisessä Voronoi-ympäristössä:

$$\mathbf{y}(\mathbf{x}) = \frac{\int_{\mathcal{V}_j} \mathbf{t}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}}{\int_{\mathcal{V}_j} p(\mathbf{x}) d\mathbf{x}}, \quad \text{kun } \mathbf{x} \in \mathcal{V}_j \tag{3.13}$$

Syöteavaruus siis kvantisoidaan kartan avulla, ja jokaiseen kvantisointiarvoon yhdistetään sopiva vastefunktion $\mathbf{y}$ arvo. Funktiosta $\mathbf{y}$ tulee näin menetellen väistämättä epäjatkuva, paitsi jos $\mathbf{t}$ on vakiofunktio, joten näyttäisi edulliselta käyttää mahdollisimman pientä kvantisointiaskelta. Kovin pieneksi askelta ei kuitenkaan voida valita, sillä käytännössä malli joudutaan rakentamaan äärellisestä määrästä näytevektoreita. Tällöin (3.13) tulee muotoon

$$\mathbf{y}(\mathbf{x}) = \frac{\sum_{\mathbf{x}^i \in \mathcal{V}_j} \mathbf{t}(\mathbf{x}^i)}{\#\{\mathbf{x}^i \mid \mathbf{x}^i \in \mathcal{V}_j\}}, \tag{3.14}$$

missä merkintä $\#\mathcal{J}$ tarkoittaa joukon $\mathcal{J}$ kardinaliteettia. Jotta $\mathbf{y}$ olisi luotettava estimaatti $\mathbf{t}$:lle, kuhunkin laariin tulee osua riittävän monta näytevektoria $\mathbf{x}^i$. Tämä rajoittaa käytettävän kartan kokoa ja siten mallin erottelukykyä.

3.4.2 SOM-2

Oletetaan, että $\mathbf{t}$ on suhteellisen sileä funktio. Tällöin $\mathbf{y}$:n arvo kussakin laarissa voidaan estimoida pienemmällä $(\mathbf{x}^i, \mathbf{t}^i)$-parien määrällä, jos vierekkäisten laarien $\mathbf{y}$-arvoja tasoitetaan lähemmäs toisiaan. Yhtä tapaa tämän tasoituksen tekemiseen kutsutaan tässä tutkimuksessa nimellä SOM-2. Se muistuttaa jossain määrin ohjattu SOM -algoritmia *(Supervised SOM)* [33, 22]. SOM-2:n kanssa identtistä lähestymistapaa on käytetty myös robotiikkasovelluksiin (mm. [34]).

SOM-2:ssa tunnetut syöte- ja vastevektoriparit – vektoreiden dimensiot olkoot p ja q – konkatenoidaan opetusvektoreiksi $\hat{\mathbf{x}}^i = [x_1^i \; x_2^i \; \cdots \; x_p^i \; t_1^i \; t_2^i \; \cdots \; t_q^i]^T$. Karttaa opetettaessa voittajan etsintään käytetään ainoastaan vektorin $\hat{\mathbf{x}}^i$ ensimmäistä p:tä komponenttia, mutta painovektoreiden kaikkia $(q + p)$:tä komponenttia päivitetään kaavan (3.2) mukaisesti. Kun opetetulle kartalle esitetään tuntematon syötevektori $\mathbf{x}^n$, etsitään sitä parhaiten vastaava karttayksikkö $\mathbf{n}_c$ painovektoreiden ensimmäisten p:n komponentin perusteella, jolloin vasteeksi saadaan $\mathbf{y}(\mathbf{x}) = [m_{c,p+1} \; m_{c,p+2} \; \cdots \; m_{c,q}]^T$.

3.4.3 RBF-SOM

Sekä SOM-1 että SOM-2 kärsivät vasteen $\mathbf{y}(\mathbf{x})$ epäjatkuvuuskohdista Voronoi-naapurustojen reunoilla. Epäjatkuvuuskohtia on mahdotonta poistaa malleissa, joissa kukin syötevektori kuvautuu täsmälleen yhdelle karttayksikölle siten, että kaikki kyseiselle karttayksikölle kuvautuvat syötevektorit antavat saman vasteen. Tarvitaan jonkinlaista interpolointia niin, että vaste riippuu syötevektorin sijainnista useamman karttayksikön painovektoreihin nähden.

Yksi tapa toteuttaa tämä on yhdistää SOM ja RBF-neuroverkko. RBF- eli säteiskantafunktioneuroverkko *(Radial Basis Function network)* [35, 36, 37] on malli, jossa verkon vaste saadaan painotettuna summana syötevektorin $\mathbf{x}$ aiheuttamasta aktivaatiosta joukolle kantafunktioita ϕ_j. Tarkastellaan vastevektorin $\mathbf{y}$ komponenttia k:

$$y_k(\mathbf{x}) = \sum_{j=1}^{M} \mathbf{w}_{kj}\phi_j(\mathbf{x}) + \mathbf{w}_{k0} = \sum_{j=0}^{M} \mathbf{w}_{kj}\phi_j(\mathbf{x}) \tag{3.15}$$

Oikeanpuoleisimmassa muodossa kantafunktioihin on lisätty funktio $\phi_0 \equiv 1$, jolloin bias-tekijät $\mathbf{w}_{k0}$ on saatu kirjoitettua osaksi summaa. Kantafunktioina käytetään yleisesti Gaussin funktioita

$$\phi_j(\mathbf{x}) = \exp\left(-\frac{||\mathbf{x} - \boldsymbol{\mu}_j||^2}{2\beta_j^2}\right) \tag{3.16}$$

joiden keskipisteen määrää parametri $\boldsymbol{\mu}_j$ ja leveyden parametri β_j. Tämän kaltaisella RBF-verkolla pystytään tietyin varsin yleisin rajoituksin – edellytetään riittävän paljon sopivasti sijoiteltuja toisiaan tarpeeksi peittäviä kantafunktioita – approksimoimaan mitä hyvänsä jatkuvaa funktiota [38, 39, 40].

Tarkastellaan tilannetta, jossa RBF-verkon parametreille $\boldsymbol{\mu}_j$ ja β_j on valittu jotkin arvot, ja tavoitteena on määrittää painoille $\mathbf{w}_{kj}$ ne arvot, jotka minimoivat neliöllisen virhesumman

$$\mathcal{E}_{SSE} = \frac{1}{2}\sum_i \sum_k [y_k(\mathbf{x}^i) - t_k^i]^2 \tag{3.17}$$

Matriisimuodossa (3.15) voidaan kirjoittaa

$$\mathbf{y}(\mathbf{x}) = \mathbf{W}\boldsymbol{\phi}(\mathbf{x}) \tag{3.18}$$

missä $[\mathbf{W}]_{kj} = \mathbf{w}_{kj}$ ja $[\boldsymbol{\phi}]_j = \phi_j$. Tällöin virhefunktio 3.17 saadaan muotoon

$$\mathcal{E}_{SSE} = \frac{1}{2}\sum_i ||\mathbf{W}\boldsymbol{\phi}(\mathbf{x}^i) - \mathbf{t}^i||^2 \tag{3.19}$$

Kyseessä on neliöllinen painojen funktio, joten sen minimikohta löydetään asettamalla gradientti nollaksi. Ratkaisun on toteutettava ehto

$$(\boldsymbol{\Phi}^T\boldsymbol{\Phi})\mathbf{W}^T = \boldsymbol{\Phi}^T\mathbf{T} \tag{3.20}$$

missä on käytetty merkintöjä $[\boldsymbol{\Phi}]_{ij} = \phi_j(\mathbf{x}^i)$ ja $[\mathbf{T}]_{ik} = t_k^i$. Ehdon toteuttaa painomatriisi

$$\mathbf{W} = (\boldsymbol{\Phi}^\dagger\mathbf{T})^T \tag{3.21}$$

missä matriisi $\boldsymbol{\Phi}^\dagger$ on matriisin $\boldsymbol{\Phi}$ pseudoinverssi; jos inverssi $(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}$ on olemassa, on $\boldsymbol{\Phi}^\dagger = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T$. Käytännössä ratkaisu etsitään on tehokkainta etsiä singulaariarvohajoitelman *(Singular Value Decomposition, SVD)* avulla, jolloin matriisin $\boldsymbol{\Phi}^T\boldsymbol{\Phi}$ lähes-singulaarisuus ei aiheuta ongelmia; kts. esim. [41].

Painomatriisille $\mathbf{W}$ voidaan siis hakea annetuille kantafunktioiden parametreille optimaalinen ratkaisu suhteellisen helposti. Sen sijaan kantafunktioparametrien $\boldsymbol{\mu}_j$ ja β_j optimaalisten arvojen määrittäminen on vaikeampaa.

Usein käytetty ratkaisu on etsiä kantafunktioiden keskipisteet $\boldsymbol{\mu}_j$ jollakin ohjaamattoman oppimisen menetelmällä [37], kuten *K-means*-klusterointialgoritmilla [36] tai SOMilla [42]. SOMia käytettäessä kantafunktioiden keskipisteet sijoitetaan kartan painovektoreihin: $\boldsymbol{\mu}_j = \mathbf{m}_j$. Kartan opetuksessa on kiinnitettävä huomiota siihen, että kartan mittasuhteet ja dimensio vastaavat syöteavaruutta X, jolloin painovektorit pääsevät leviämään kaikkialle syöteavaruuteen [37, 42]. Samasta syystä naapurustofunktion on opetuksen loppuvaiheessa oltava riittävän pitkään tarpeeksi kapea.

Kantafunktioiden leveydet β_j voidaan SOMin avulla etsiä siten, että asetetaan

$$\beta_j = \rho \min_k ||\mathbf{m}_j - \mathbf{m}_k|| \tag{3.22}$$

missä ρ on kaikille kantafunktioille yhteinen parametri, jolle voidaan iteratiivisesti hakea optimiarvo [42]. Etäisyys lähimpään painovektoriin voidaan kaavassa 3.22 korvata keskimääräisellä etäisyydellä esimerkiksi neljään lähimpään painovektoriin.

Luku 4

Tilinpäätösanalyysin työkalut

Tilinpäätösanalyysissä pyritään muodostamaan mahdollisimman luotettava kuva yrityksen nykyisestä tilasta sekä tulevasta kehityksestä käyttäen pelkästään tilinpäätökseen sisältyviä tietoja. Tässä tutkimuksessa on keskitytty huomattavasti rajatumpaan aiheeseen: miten estimoida konkurssin todennäköisyyttä annetun tilinpäätöksen perusteella. Tavoitteeseen pääsemiseksi on tutkittu kahta asiaa:

- Konkurssiin ajautuminen ilmiönä
- Luokittelu terveisiin ja konkurssiyrityksiin

Ensimmäisessä kohdassa selvitetään, miten konkurssiin johtava kehitys näkyy tilinpäätöksissä. Samalla pyritään löytämään tekijöitä, joiden avulla konkurssiriskin kasvaminen pystyttäisiin tunnistamaan mahdollisimman luotettavasti jo varhaisessa vaiheessa. Työkaluna on itseorganisoiva kartta.

Toisessa kohdassa kokeillaan erilaisten luokittelumenetelmien soveltumista tilinpäätösaineistoon. Tehtävänä on arvioida konkurssiriskiä karkeimmalla mahdollisella tasolla: yritykset jaetaan terveisiin ja konkurssiyrityksiin sen perusteella, odotetaanko niiden toimivan vielä kahden vuoden kuluttua tilinpäätöksen antamisesta. Työkaluina on eräänlaisen standardin asemaan nousseen lineaarisen erotteluanalyysin (LDA) lisäksi joukko muita luokittimia: kvadraattinen erotteluanalyysi (QDA), k:n lähimmän naapurin luokitin (kNN), oppiva vektorikvantisointi (LVQ) sekä itseorganisoivaan karttaan ja säteiskantaverkkoon perustuva luokitin (RBF-SOM); lisäksi kokeillaan itseorganisoituvan kartan soveltumista luokitteluun kahdella eri tavalla käytettynä (perus-SOM, ohjattu SOM).

4.1 Konkurssi-ilmiön tutkiminen

Konkurssiin johtavaa kehitystä tutkitaan SOMin avulla. Kartan ja siihen liittyvien erilaisten työkalujen (kts. kappale 3.3) avulla saadaan data projisoitua tasoon, jolloin voidaan etsiä erilaisia tilinpäätösprofiileja sekä tarkastella sitä, miten nämä heijastuvat konkurssialttiuteen. Umat-esityksen ja Sammonin projektion avulla saadaan selville, löytyykö yritysaineistosta erillisiä ryppäitä, vai muodostuuko aineistosta yhtenäinen, multinormaalijakaumaa muistuttava möykky.

Visuaalinen tarkastelu on luotettavaa vain, jos kuvaus syöteavaruudesta kartalle on ainakin likimain jatkuva. Tätä voidaan arvioida Sammonin projektion ja topografisen virheen $\mathcal{E}_t$ avulla. Lisäksi kartan on oltava sopivan kokoinen aineiston laajuuteen nähden: liian pieni kartta ei ole tarpeeksi erottelukykyinen, liian suuressa kartassa taas tilinpäätösvektoreita kuvautuu karttayksiköille niin vähän, että yksiköitä ei saada luotettavasti kalibroitua.

4.2 Luokittelu terveisiin ja konkurssiyrityksiin

4.2.1 Luokitteluperusteet

Tehtävänä on luokitella annetut tilinpäätökset kahteen ryhmään, terveisiin ja konkurssiyrityksiin. Otetaan käyttöön seuraavat merkinnät: tilinpäätöstietovektori on $\mathbf{x}$, konkurssiyritysten antamien tilinpäätösten luokka $\mathcal{C}_1$ ja terveiden yritysten antamien tilinpäätösten luokka $\mathcal{C}_2$. Merkitään lisäksi tapahtuman q todennäköisyyttä $P(q)$. Tällöin todennäköisyys, että annettu tilinpäätös $\mathbf{x}$ kuuluisi luokkaan $\mathcal{C}_i$ voidaan kirjoittaa $P(\mathcal{C}_i|\mathbf{x})$. Jokainen tilinpäätös kuuluu joko terveiden tai konkurssiyritysten antamien tilinpäätösten luokkaan, joten $P(\mathcal{C}_1|\mathbf{x}) + P(\mathcal{C}_2|\mathbf{x}) = 1\ \forall\mathbf{x}$.

Luokitteluvirheiden kokonaismäärän minimointi

Luokitusvirheiden kokonaismäärän minimoi luokittelusääntö

$$\begin{cases} P(\mathcal{C}_1|\mathbf{x}) > P(\mathcal{C}_2|\mathbf{x}) & \Rightarrow \text{valitaan luokka } \mathcal{C}_1 \\ P(\mathcal{C}_1|\mathbf{x}) < P(\mathcal{C}_2|\mathbf{x}) & \Rightarrow \text{valitaan luokka } \mathcal{C}_2 \end{cases} \tag{4.1}$$

Toisin sanoen, valitaan se luokka, johon annettu syötevektori suurimmalla todennäköisyydellä kuuluu. Todennäköisyyksien ollessa yhtäsuuria sovitaan esimerkiksi, että valitaan aina luokka $\mathcal{C}_1$.

Jaetaan nyt syöteavaruus kahteen osaan $\mathcal{R}_1$ ja $\mathcal{R}_2$ seuraavasti:

$$\begin{aligned} \mathcal{R}_1 &= \{\mathbf{x} \mid P(\mathcal{C}_1|\mathbf{x}) > P(\mathcal{C}_2|\mathbf{x})\} \\ \mathcal{R}_2 &= \{\mathbf{x} \mid P(\mathcal{C}_1|\mathbf{x}) < P(\mathcal{C}_2|\mathbf{x})\} \end{aligned} \tag{4.2}$$

Luokittelusäännön (4.1) mukaisesti alueella $\mathcal{R}_1$ sijaitsevat vektorit $\mathbf{x}$ luokitellaan tällöin kuuluviksi luokkaan $\mathcal{C}_1$ ja vastaavasti alueella $\mathcal{R}_1$ sijaitsevat vektorit $\mathbf{x}$ luokitellaan kuuluviksi luokkaan $\mathcal{C}_2$. Tällöin virheluokitusten osuus kaikista luokituksista, ε, voidaan jakaa kahteen osaan, tyypin I virheisiin ε_1 ja tyypin II virheisiin ε_2:

$$\begin{aligned} \varepsilon &= \varepsilon_1 + \varepsilon_2 \\ &= P(\mathbf{x} \in \mathcal{R}_2|\mathcal{C}_1) + P(\mathbf{x} \in \mathcal{R}_1|\mathcal{C}_2) \\ &= \int_{\mathcal{R}_2} p(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1)d\mathbf{x} + \int_{\mathcal{R}_1} p(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2)d\mathbf{x} \end{aligned} \tag{4.3}$$

missä $p(\mathbf{x}|\mathcal{C}_i)$ on luokan $\mathcal{C}_i$ ehdollinen todennäköisyystiheys, ja $P(\mathcal{C}_i)$ on luokan $\mathcal{C}_i$ *a priori*-todennäköisyys, ts. luokan $\mathcal{C}_i$ tilinpäätösten osuus kaikista tilinpäätöksistä. Tyypin

I virheessä siis luokitellaan konkurssiyritys toimivaksi yritykseksi, tyypin II virheessä luokitellaan toimiva yritys konkurssiyritykseksi.

Kaavan (4.3) avulla nähdään, että luokittelusääntö (4.1) todella minimoi luokittelun kokonaisvirheen. Tarkastellaan aluetta $\mathcal{R}_1$, jolla sijaitsevat vektorit sijoitetaan luokittelusäännön (4.1) mukaisesti luokkaan $\mathcal{C}_1$. Bayesin säännön (pastori Thomas Bayes, 1702-61) mukaisesti pätee

$$P(\mathcal{C}_i|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_i)P(\mathcal{C}_i)}{p(\mathbf{x})} \tag{4.4}$$

joten kaavan 4.3 jälkimmäisen integraalin integrandi $p(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2)$ saadaan muotoon $P(\mathcal{C}_2|\mathbf{x})p(\mathbf{x})$. Tästä nähdään, että jos luokittelusääntöä muutettaisiin siten, että osa alueella $\mathcal{R}_1$ sijaitsevista vektoreista sijoitettaisiinkin luokkaan $\mathcal{C}_2$, jälkimmäisen integraalin ja siten myös kokonaisvirheen arvo kasvaisi, koska integrointialueessa $\mathcal{R}_1$ määritelmän (4.2) mukaisesti on $P(\mathcal{C}_1|\mathbf{x})p(\mathbf{x}) > P(\mathcal{C}_2|\mathbf{x})p(\mathbf{x})$. Vastaava tarkastelu pätee myös alueeseen $\mathcal{R}_2$.

Luokittelukustannusten minimointi

Luokittelulle voidaan asettaa muitakin tavoitteita kuin virheluokitusten kokonaismäärän minimointi. Oletetaan, että onnistuneen luokituksen arvo ja virheellisen luokituksen kustannus riippuu $\mathbf{x}$:n todellisesta luokasta. Merkitään luokkaan $\mathcal{C}_i$ kuuluvan vektorin luokittelusta luokkaan $\mathcal{C}_j$ aiheutuvia kustannuksia L_{ij}. Tällöin keskimääräiset luokituskustannukset ovat

$$\begin{aligned} R = \int_{\mathcal{R}_1} & [L_{11}p(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1) + L_{21}p(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2)]d\mathbf{x} + \\ & + \int_{\mathcal{R}_2} [L_{12}p(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1) + L_{22}p(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2)]d\mathbf{x} \end{aligned} \tag{4.5}$$

ja ne minimoi luokittelusääntö, jonka mukaan luokkaa $\mathcal{C}_1$ vastaavaksi alueeksi $\mathcal{R}_1$ valitaan niiden pisteiden $\mathbf{x}$ muodostama joukko, jossa

$$L_{11}p(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1) + L_{21}p(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2) < L_{12}p(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1) + L_{22}p(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2) \tag{4.6}$$

Loput pisteet kuuluvat alueeseen $\mathcal{R}_2$.

Neyman-Pearsonin luokittelukriteeri

Vaihtoehtona luokitteluvirheiden määrän tai luokittelukustannusten minimoinninnille voidaan käyttää ns. Neyman-Pearsonin luokittelukriteeriä [43], jossa kiinnitetään toinen virhetyyppi tietyn suuruiseksi; valitaan esimerkiksi tyypin I virheelle jokin ohjearvo α ja vaaditaan, että

$$\varepsilon_1 = \int_{\mathcal{R}_2} p(\mathbf{x}|\mathcal{C}_1)d\mathbf{x} = \alpha \tag{4.7}$$

Tämän rajoituksen puitteissa pyritään minimoimaan toinen virhetyyppi

$$\varepsilon_2 = \int_{\mathcal{R}_1} p(\mathbf{x}|\mathcal{C}_2)d\mathbf{x} \tag{4.8}$$

Neyman-Pearsonin kriteeri sopii varsinkin erilaisten hälytys- ja varoitusjärjestelmien luokittimiin, joiden toivotaan tyypillisesti antavan hälytyksen mahdollisimman luotettavasti aina kun siihen on aihetta, väärien hälytysten lisääntymisen kustannuksellakin. Hälytystilanteet ovat yleensä varsin harvinaisia, jolloin luokitteluvirheiden kokonaismäärän minimoiva luokitin ei ehkä antaisi koskaan hälytystä. Samoin suhteellisia luokituskustannuksia saattaa olla vaikea määrittää, esimerkiksi virheiden seurauksien yhteismitattomuuden vuoksi.

Uskottavuussuhdetesti-esitysmuoto luokittelukriteereille

On mielenkiintoista todeta, että edellä esitellyt kolme päätössääntöä voidaan kaikki esittää ns. uskottavuussuhdetestin *(likelihood ratio test)* muodossa, jossa verrataan uskottavuussuhdetta $\mathcal{L}$ eli luokkakohtaisten ehdollisten tiheyksien suhdetta johonkin ohjearvoon λ ja sen perusteella päätetään, kumpaan luokkaan annettu vektori kuuluu:

$$\mathcal{L}(\mathbf{x}) \equiv \frac{p(\mathbf{x}|\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)} \underset{\mathcal{C}_2}{\overset{\mathcal{C}_1}{\gtrless}} \lambda \tag{4.9}$$

Luokitteluvirheiden kokonaismäärän minimoiva päätössääntö (4.1) saadaan Bayesin sääntöä (4.4) käyttäen muotoon

$$\frac{p(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1)}{p(\mathbf{x})} \underset{\mathcal{C}_2}{\overset{\mathcal{C}_1}{\gtrless}} \frac{p(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2)}{p(\mathbf{x})} \tag{4.10}$$

$$\Leftrightarrow \frac{p(\mathbf{x}|\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)} \underset{\mathcal{C}_2}{\overset{\mathcal{C}_1}{\gtrless}} \frac{P(\mathcal{C}_2)}{P(\mathcal{C}_1)} = \lambda \tag{4.11}$$

Samoin luokittelukustannukset minimoiva päätössääntö (4.6) voidaan kirjoittaa muodossa

$$(L_{11} - L_{12})P(\mathcal{C}_1)p(\mathbf{x}|\mathcal{C}_1) \underset{\mathcal{C}_2}{\overset{\mathcal{C}_1}{\lessgtr}} (L_{21} - L_{22})P(\mathcal{C}_2)p(\mathbf{x}|\mathcal{C}_2) \tag{4.12}$$

$$\Leftrightarrow \frac{p(\mathbf{x}|\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)} \underset{\mathcal{C}_2}{\overset{\mathcal{C}_1}{\gtrless}} \frac{(L_{21} - L_{22})P(\mathcal{C}_2)}{(L_{11} - L_{12})P(\mathcal{C}_1)} = \lambda \tag{4.13}$$

Neyman-Pearsonin päätössääntö voidaan etsiä minimoimalla suuretta

$$\mathcal{J} = \lambda(\varepsilon_1 - \alpha) + \varepsilon_2 \tag{4.14}$$

missä λ on Lagrangen kerroin. Sijoittamalla tähän (4.7) ja (4.8) saadaan

$$\begin{aligned} \mathcal{J} &= \int_{\mathcal{R}_1} p(\mathbf{x}|\mathcal{C}_2)d\mathbf{x} + \lambda \left[\int_{\mathcal{R}_2} p(\mathbf{x}|\mathcal{C}_1)d\mathbf{x} - \alpha \right] \\ &= 1 - \lambda\alpha + \int_{\mathcal{R}_2} [\lambda p(\mathbf{x}|\mathcal{C}_1) - p(\mathbf{x}|\mathcal{C}_2)]d\mathbf{x} \end{aligned} \tag{4.15}$$

joka minimoituu, kun $\mathcal{R}_2$ valitaan siten, että viimeisen integraalin integrandi on kaikkialla negatiivinen. Ehto voidaan esittää muodossa

$$\frac{p(\mathbf{x}|\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)} \underset{\mathcal{C}_2}{\overset{\mathcal{C}_1}{\gtrless}} \lambda \tag{4.16}$$

Kahden ensimmäisen päätössäännön tapauksessa λ on helppo laskea suoraan, jos luokkien *a priori*-todennäköisyydet ja jälkimmäisen säännön kohdalla myös luokitusvirheiden kustannukset tunnetaan. Sen sijaan Neyman-Pearsonin säännön (4.16) λ on yleensä etsittävä empiirisesti iteroimalla siten, että ehto (4.7) täyttyy.[44]

– Tässä työssä luokittelu suoritetaan kahdella tavalla. Ensisijaisesti luokittelussa käytetään Neyman-Pearsonin virhekriteeriä; vertailun vuoksi joillakin luokittimilla kokeillaan lisäksi luokittelua virheiden kokonaismäärää minimoiden.

4.2.2 Luokittimet

Työssä kokeillaan kolmea erilaista SOMiin perustuvaa luokitinta, jotka perustuvat kappaleessa 3.4 esiteltyihin menetelmiin.

Perus-SOM- ja ohjattu SOM-luokittimissa approksimoidaan luokkakohtaista *a posteriori*-todennäköisyyttä erikseen kunkin karttayksikön Voronoi-ympäristössä (3.12). Syöteavaruus saadaan jaettua konkurssialueeseen ja terveeseen alueeseen siten, että tulkitaan konkurssiyksiköiksi ne karttayksiköt, joissa konkurssin todennäköisyys ylittää jonkin tietyn raja-arvon, ja luokitellaan näille yksiköille kuvautuvat tilinpäätökset konkurssitilinpäätöksiksi. RBF-SOM-luokittimessa saadaan kaavan (3.15) mukaisesti suoraan luokkakohtainen *a posteriori*-todennäköisyys; jos konkurssiluokkaan kuulumisen todennäköisyys ylittää raja-arvon, tilinpäätös luokitellaan konkurssiyrityksen antamaksi. Raja-arvo etsitään kaikissa luokittimissa haarukoimalla siten, että Neyman-Pearsonin kriteeri (4.7) toteutuu opetusaineistossa.

Lineaarinen ja kvadraattinen erotteluanalyysi

Lineaarinen ja kvadraattinen erotteluanalyysi *(linear discriminant analysis, LDA*, ja *quadratic discriminant analysis, QDA)* perustuvat oletukselle, että luokat $\mathcal{C}_j$ ovat multinormaalisti jakautuneita, eli että niiden tiheysfunktiot ovat muotoa

$$p(\mathbf{x}|\mathcal{C}_j) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}_j|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^T\mathbf{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\right] \tag{4.17}$$

missä $\mathbf{\Sigma}_j$ ja $\boldsymbol{\mu}_j$ ovat j:nnen luokan kovarianssimatriisi ja keskipiste, ja d on syöteavaruuden dimensio. Kahden luokan tapauksessa luokkien välisen rajapinnan (päätöspinnan) yhtälö saadaan säännöstä (4.10):

$$p(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1) = p(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2) \tag{4.18}$$

Sijoittamalla tähän (4.17) ja ottamalla puolittain logaritmit saadaan päätöspinnan yhtälö muotoon

$$\begin{aligned} \log P(\mathcal{C}_1) - \frac{1}{2}|\mathbf{\Sigma}_1| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\mathbf{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) = \\ = \log P(\mathcal{C}_2) - \frac{1}{2}|\mathbf{\Sigma}_2| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T\mathbf{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) \end{aligned} \tag{4.19}$$

Tällä neliöllisellä pinnalla luokkien *a posteriori*-todennäköisyydet ovat yhtäsuuret. Luokka, jolla on suurin *a posteriori*-todennäköisyys, saadaan siis selville QDA-päätössäännöllä

$$g_{\mathrm{QDA}} = \operatorname*{argmax}_{j=1,2} \left[\log P(\mathcal{C}_j) - \frac{1}{2}|\boldsymbol{\Sigma}_j| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) \right] \tag{4.20}$$

missä $g_{\mathrm{QDA}} : \mathbf{x} \to \{1, 2\}$ on luokitinfunktio, joka yhdistää kunkin vektorin jommankumman luokan indeksiin.

Jos oletetaan, että luokkien kovarianssimatriisit ovat yhtäsuuret, toisin sanoen $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, QDA-päätössääntö (4.20) sievenee LDA-päätössäännöksi

$$g_{\mathrm{LDA}} = \operatorname*{argmax}_{j=1,2} \left[\log P(\mathcal{C}_j) + \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_j) \right] \tag{4.21}$$

jossa luokkia erottaa hypertaso syöteavaruudessa.

Neyman-Pearsonin säännön soveltaminen tapahtuu yhtä suoraviivaisesti sijoittamalla luokkakohtaiset ehdolliset tiheydet (4.17) sääntöön (4.16) ja etsimällä se λ:n arvo, joka toteuttaa rajoitusehdon (4.7) valitulla α:n arvolla.

Lähimmän naapurin luokitin

k:n lähimmän naapurin luokitin *(k nearest neighbor classifier, kNN)* on ei-parametrinen luokitin, jossa kukin syötevektori $\mathbf{x}$ luokitellaan kuuluvaksi siihen luokkaan, johon enemmistö sitä lähimpänä olevista k:sta tunnetusta vektorista kuuluu. Tällöin kNN minimoi virheluokitusten kokonaismäärän. Menetelmä edellyttää toimiakseen suhteellisen suuren joukon tunnettuja näytevektoreita.

kNN-algoritmia voidaan myös muokata siten, että enemmistöäänestyksen sijasta vaaditaan vähintään j:n naapurin, $1 \leq j \leq k$, kuuluvan esimerkiksi luokkaan $\mathcal{C}_1$, jotta $\mathbf{x}$ luokiteltaisiin kuuluvaksi luokkaan $\mathcal{C}_1$; muuten $\mathbf{x}$ luokiteltaisiin luokkaan $\mathcal{C}_2$. Tällöin eri tyyppisten virheluokitusten suhteita voidaan muunnella, jolloin kNN-luokitinta voidaan käyttää myös Neyman-Pearsonin päätössääntöön perustuvassa luokittelussa.[45]

Oppiva vektorikvantisointi

Oppiva vektorikvantisointi *(Learning Vector Quantization, LVQ)* [22] on SOMille läheistä sukua oleva luokittelumenetelmä. Siinä luokat mallitetaan prototyyppivektoreilla $\mathbf{m}_j$, $j = 1 \ldots n$, jotka kukin kuuluvat johonkin luokista $\mathcal{C}_i$. Prototyyppivektorit voidaan jakaa tasan eri luokkien kesken tai suhteessa luokkien *a priori*-todennäköisyyksiin, ja niiden alkuarvot valitaan esimerkiksi kNN- tai SOM-algoritmeilla siten, että prototyyppivektorit sijaitsevat samaan luokkaan kuuluvien näytevektoreiden muodostamien klustereiden sisällä.

Prototyyppivektoreita opetetaan siten, että näytevektoreita, joiden luokka tunnetaan, käydään läpi yksi kerrallaan useita kertoja. Kullekin näytevektorille $\mathbf{x}^i$ etsitään sitä parhaiten vastaava yksikkö $\mathbf{m}_c$,

$$c = \operatorname*{argmin}_{j} \{||\mathbf{x}^i - \mathbf{m}_j||\} \tag{4.22}$$

jonka jälkeen kyseistä prototyyppivektoria opetetaan seuraavan säännön mukaisesti:

$$\mathbf{m}_c := \begin{cases} \mathbf{m}_c + \alpha(\mathbf{x}^i - \mathbf{m}_c) & \text{kun } \mathbf{x}^i \text{ ja } \mathbf{m}_c \text{ kuuluvat samaan luokkaan,} \\ \mathbf{m}_c - \alpha(\mathbf{x}^i - \mathbf{m}_c) & \text{muutoin} \end{cases} \tag{4.23}$$

Vuorossa olevaa näytevektoria lähinnä olevaa prototyyppivektoria siis siirretään kohti näytevektoria, jos vektorit kuuluvat samaan luokkaan, ja näytevektorista poispäin, jos ne kuuluvat eri luokkaan.

LVQ-algoritmista on olemassa useita eri versioita – LVQ1, OLVQ1, LVQ2, LVQ3 – joissa perusperiaate on kaikissa edellä esitetyn kaltainen; yksityiskohtaisesti algoritmit on esitelty esim. viitteessä [22].

LVQ-algoritmit perustuvat sille havainnolle, että klassinen vektorikvantisointi – esim. *K-means*-algoritmi, jota SOM approksimoi, jos opetuksen lopussa naapurustofunktion leveysparametri $\sigma = 0$ riittävän kauan – pyrkii sijoittelemaan prototyyppivektorit siten, että niiden tiheys on jokin monotoninen näytevektoreiden tiheyden $p(\mathbf{x})$ funktio. Modifioimalla vektorikvantisointialgoritmia sopivasti saadaan jotakin muuta tiheysfunktiota $f(\mathbf{x})$ approksimoiva algoritmi. LVQ:n tapauksessa haluttu tiheysfunktio olisi

$$f(\mathbf{x}) = \begin{cases} p(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1) - p(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2) & \text{kun } \mathbf{x} \in \mathcal{R}_1 \\ p(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2) - p(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1) & \text{kun } \mathbf{x} \in \mathcal{R}_2 \end{cases} \tag{4.24}$$

missä $\mathcal{R}_1$ ja $\mathcal{R}_2$ ovat kaavan (4.2) mukaiset.

Kohosen viitteessä [22] esittämässä motivaatiossa LVQ:lle tarkastellaan edellisen yleistystä useammalle kuin kahdelle luokalle; tässä yhteydessä riittää kuitenkin tarkastella kahden luokan tapausta.

Vertaamalla kaavoja (4.1) ja (4.24) nähdään, että virheluokitusten määrän minimoivilla päätöspinnoilla $f(\mathbf{x}) = 0$; muualla $f(\mathbf{x}) > 0$. Alustetaan nyt kummankin luokan prototyyppivektorit satunnaisiin arvoihin alueelle, jossa kyseisen luokan *a posteriori*-todennäköisyys on suurempi kuin toisen luokan, mikä voidaan tehdä esimerkiksi kNN-algoritmilla. Tällöin luokittelutulosta voidaan parantaa, jos prototyyppivektoreita opetetaan siten, että ne kvantisoivat kyseisen alueen mahdollisimman hyvin. Opetuksen tavoitteena olisi siis saada kyseisen alueen kvantisointivirheen $\mathcal{E}_q$ odotusarvo tiheyden $f(\mathbf{x})$ suhteen pienenemään. Kvantisointivirheen odotusarvon gradientti prototyyppivektorille $\mathbf{m}_j \in \mathcal{C}_1$ saadaan kaavasta

$$\begin{aligned} \nabla_{\mathbf{m}_j}\mathcal{E}_q &= -2\int_{\mathcal{V}_j}(\mathbf{x} - \mathbf{m}_j)f(\mathbf{x})d\mathbf{x} \\ &= -2\int_{\mathcal{V}_j}(\mathbf{x} - \mathbf{m}_j)p(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1)d\mathbf{x} - 2\int_{\mathcal{V}_j}(\mathbf{x} - \mathbf{m}_j)p(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2)d\mathbf{x} \end{aligned} \tag{4.25}$$

sillä alustuksen mukaisesti $\mathbf{m}_j \in \mathcal{R}_1$; luokan $\mathcal{C}_2$ prototyyppivektoreille gradientti saadaan vastaavalla tavalla. Gradienttimenetelmällä kvantisointivirhettä voidaan nyt minimoida iteratiivisesti opetussäännöllä

$$\mathbf{m}_j := \mathbf{m}_j - \alpha'\nabla_{\mathbf{m}_j}\mathcal{E}_q = \begin{cases} \mathbf{m}_j + 2\alpha'(\mathbf{x}^i - \mathbf{m}_j) & \text{kun } \mathbf{x}^i\text{ja } \mathbf{m}_j\text{samasta luokasta,} \\ \mathbf{m}_j - 2\alpha'(\mathbf{x}^i - \mathbf{m}_j) & \text{kun } \mathbf{x}^i\text{ja } \mathbf{m}_j\text{eri luokista} \end{cases} \tag{4.26}$$

kun vuorossa oleva näytevektori $\mathbf{x}^i$ osuu prototyyppivektorin $\mathbf{m}_j$ Voronoi-ympäristöön $\mathcal{V}_j$; vain vuorossa olevaa näytevektoria lähimpänä olevaa prototyyppivektoria siis siirretään kullakin opetusaskeleella. Sijoittamalla $\alpha = 2\alpha'$ saadaan LVQn opetussääntö (4.23).

LVQ voidaan modifioida Neyman-Pearsonin kriteerille muuttamalla $f(\mathbf{x})$:n määritelmää (4.24) korvaamalla siinä esiintyvät *a priori*-todennäköisyydet $P(\mathcal{C}_1)$ ja $P(\mathcal{C}_2)$ luvuilla $\beta P(\mathcal{C}_1)$ ja $(1-\beta)P(\mathcal{C}_2)$, missä β on valittu siten, että Neyman-Pearsonin rajoitusehto (4.7) toteutuu; lisäksi vaaditaan, että $0 < \beta < 1$.

Nyt $f(\mathbf{x}) = 0$, kun

$$
\begin{aligned}
p(\mathbf{x}|\mathcal{C}_1)\beta P(\mathcal{C}_1) &= p(\mathbf{x}|\mathcal{C}_2)(1-\beta)P(\mathcal{C}_2) && (4.27)\\
\Leftrightarrow \frac{p(\mathbf{x}|\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)} &= \frac{(1-\beta)P(\mathcal{C}_2)}{\beta P(\mathcal{C}_1)} = \lambda && (4.28)
\end{aligned}
$$

eli päätöspinnat ovat kaavan (4.16) mukaiset. Modifioidun LVQn opetussääntö saadaan tästä samalla tavalla kuin tavallisenkin LVQn; lopputulokseksi tulee

$$
\mathbf{m}_c := \begin{cases} \mathbf{m}_c + \alpha\beta(\mathbf{x}^i - \mathbf{m}_c) & \text{kun } \mathbf{x}^i \text{ ja } \mathbf{m}_c \text{ kuuluvat samaan luokkaan,}\\ \mathbf{m}_c - \alpha(1-\beta)(\mathbf{x}^i - \mathbf{m}_c) & \text{muutoin} \end{cases} \tag{4.29}
$$

Opetussäännön lisäksi prototyyppivektoreiden alustusta on muutettava vastaavasti: enemmistöäänestyksen sijaan käytetään jotakin sopivasti painotettuun kNN-luokitukseen perustuvaa alustusta.

Luku 5

Aineisto

Tutkimuksessa yhteistyökumppanina oli Kera Oy, Suomessa toimiva pienten ja keskisuurten yritysten rahoittaja ja kehittäjä. Tutkimuksen otos ennen karsintaa koostui kaikkien Keran asiakkaiden tilinpäätöksistä vuosilta 1985-94[1], yhteensä 51 763 tilinpäätöstä kaikkiaan 11 471 eri yritykseltä. Aineistoon kuului 2 417 konkurssiyritystä, kun konkurssiksi lasketaan tuomittujen konkurssien lisäksi myös yrityksen toiminnan lopettaminen taloudellisten vaikeuksien vuoksi. Tästä aineistosta karsittiin alla kuvatulla tavalla suhteellisen homogeeninen, mutta edelleen varsin mittavan kokoinen lopullinen tutkimusaineisto.

Keran asiakkaat muodostavat hyvin valikoidun otoksen kaikista maassamme toimivista yrityksistä, joten tulosten yleistettävyydestä muihin kuin Keran asiakkaisiin – esimerkiksi Keran asiakkuutta hakeviin yrityksiin – ei ennen jatkotutkimuksia voi sanoa mitään varmaa. Sen sijaan Keran asiakaskunnan sitä osaa, joka vastaa tehtyjä rajauksia, lopullinen otos luonnollisesti kuvaa parhaalla mahdollisella tavalla.

Aineistossa esiintyvät tilinpäätökset olivat oikaistuja. Tämä on erittäin arvokas piirre: esimerkiksi tilinpäätösanalyytikot perustavat arvionsa käytännössä aina oikaistuihin tilinpäätöksiin niiden paremman luotettavuuden vuoksi. Joissakin suurehkoihin otoksiin perustuvissa tutkimuksissa (esim. [6]) on käytetty oikaisemattomia tilinpäätöksiä ja esitetty tälle myös joitakin perusteluja, mutta kyse lienee ollut lähinnä aineiston laajuuden ja oikaisemisen työläyden sanelemasta käytännön pakosta; jos vain on mahdollista käyttää oikaistuja tilinpäätöksiä, näin kannattaa aina myös tehdä.

5.1 Yritysten ja tilinpäätösten karsinta

Otoksen rajaaminen on tyypillisesti kompromissin tekoa tulosten yleistettävyyden ja tarkkuuden välillä. Mitä homogeenisempi on tutkittavien yritysten joukko, sitä pienempään osaan kaikista yrityksistä malli soveltuu, mutta sitä tarkemmin sen voidaan odottaa näitä yrityksiä kuvaavan. Tässä tutkimuksessa oli tavoitteena löytää yleisiä yritysten käyttäytymispiirteitä, jotka olisivat suhteellisen riippumattomia yrityksen toimialasta, koosta tai muista vastaavista seikoista. Siitä huolimatta aineistoa oli tarpeen

[1] Joukossa on muutama yritys, jonka tilinpäätöshistoria ulottuu kauemmas ajassa taaksepäin; varhaisin tilinpäätös on vuodelta 1977.

karsia varsin raskaasti, jotta päädyttäisiin edes jossain määrin yhdenmukaisesti käyttäytyvään yritysjoukkoon.

Ensimmäinen karsintakriteeri oli yritysmuoto. Osakeyhtiöiden tilinpäätökset eroavat muiden yhtiömuotojen tilinpäätöksistä ennen kaikkea siinä, miten yrityksen tuottamat voitot maksetaan omistajille. Ilman yritys yritykseltä tehtäviä oikaisuja eri yritysmuotojen tilinpäätökset eivät ole vertailukelpoisia. Siksi tutkimus rajoitettiin pelkkiin osakeyhtiöihin. Sama pitää paikkansa kahden erilaisen poistokäytännön suhteen. Vuoteen 1994 Kera Oy käytti yritystietokannassaan EVL-poistoja, siitä eteenpäin poistot on esitetty suunnitelman mukaisina (sumu-poistot), kuten nykyiset asetukset edellyttävät. Tutkimukseen otettiin mukaan vain EVL-poistoin laaditut tilinpäätökset, koska näitä oli huomattavasti enemmän; kts. taulukko 5.1, jossa on esitetty Kera Oy:n asiakasyritysten jakautuminen yhtiömuodon ja poistokäytännön perusteella.

Taulukko 5.1: Yritysten ja tilinpäätösten jakautuminen yhtiömuodon ja poistokäytännön perusteella (evl/sumu-tilinpäätökset ovat eri vuosilta, joten sama yritys saattaa esiintyä molemmissa ryhmissä)

	terveitä yrityksiä	konkurssi-yrityksiä	yrityksiä yhteensä	terveitä tilinp.	konkurssi-tilinp.	tilinp. yhteensä
oy / evl	3 928	1 264	5 192	18 171	4 210	22 381
oy / sumu	2 134	18	2 152	2 547	22	2 569
ei oy / evl	4 677	1 150	5 827	21 302	4 047	25 349
ei oy / sumu	1 292	9	1 301	1 454	10	1 464

Jatkossa aineistosta karsittiin pois "epätavalliset" yritykset. Koska aineisto oli alunperin erittäin laaja, karsinta voitiin tässä vaiheessa tehdä varsin raskaalla kädellä. Kriteerit aineistosta karsimiselle olivat lyhyesti kuvattuna seuraavanlaiset:

- *toimialat:* muut kuin teollisuusyritykset, sekä lisäksi ne teollisuusyritystoimialat, joilla liiketoiminta perustuu suhteellisen suurten, pitkäaikaisten yksittäisten projektien toimittamiseen; toimialaluokituskoodein ilmaistuna mukaan tulevat toimialaluokat 15000-37200 (pääluokka D: teollisuus), joista pois 20309 (muiden rakennuspuusepäntuotteitten kuin puutalojen valmistus), 26610-26669 (betoni-, sementti- ja kipsituotteitten valmistus) ja 28100-28120 (metallirakenteiden valmistus);

- *liian pienet ja suuret:* yritykset, jotka kokonsa puolesta kuuluvat aineiston ääripäihin; rajat vedettiin 0,5 Mmk ja 500 Mmk liikevaihtoon;

- *määrittymättömät:* yritykset, jotka eivät ole toistaiseksi menneet konkurssiin mutta jotka eivät silti välttämättä ole terveitä; kriteeriksi otettiin yrityksen omavaraisuusaste, jonka edellytettiin olevan positiivinen viimeisessä tunnetussa tilinpäätöksessä;

- *puutteita:* konkurssiin menneet yritykset, jotka eivät olleet toimittaneet tilinpäätöstietojaan yli kahteen vuoteen ennen konkurssia;
- *liian nuoret:* tilinpäätökset, jotka olivat alle viisivuotiaiden yritysten antamia.

Taulukko 5.2: Yritysten ja tilinpäätösten määrät karsinnan eri vaiheissa

kriteeri	terveitä yrityksiä	konkurssi-yrityksiä	yrityksiä yhteensä	terveitä tilinp.	konkurssi-tilinp.	tilinp. yhteensä
ennen kars.	3 928	1 264	5 192	18 171	4 210	22 381
toimialat	-1 662	-531	-2 193	-7 508	-2 702	-9 210
liian pienet	-440	-161	-601	-858	-259	-1 117
liian suuret	-23	0	-23	-65	0	-65
liian nuoret	-384	-238	-622	-1 684	-733	-2 417
määrittym.	-912		-912	-4 300		-4 300
puutteita		-133	-133		-356	-356
jäljellä	833	304	1 137	3 839	1 059	4 898

Taulukossa 5.2 on esitetty sekä yritysten että tilinpäätösten määrät kunkin karsintakriteerin kohdalla. Yritysten määrät on kirjattu taulukkoon sen mukaisesti, kuinka moni yritys jonakin tilivuotenaan vastasi kyseistä kriteeriä. Jos yritys vastasi useampaa karsintakriteeriä, se on laskettu mukaan useaan kertaan; toisaalta yrityksen tilinpäätökset kohdissa *liian pienet, liian suuret* ja *liian nuoret* on karsittu pois vain niiltä vuosilta, jolloin yritys vastasi jotakin karsintakriteeriä. Siksi jäljelle jää enemmän yrityksiä ja tilinpäätöksiä kuin vähennyslaskulla saisi.

Karsitussa aineistossa olevien konkurssitilinpäätösten määrät eri vuosina ennen konkurssia on esitetty taulukossa 5.3.

Taulukko 5.3: Konkurssitilinpäätösten määrät 0 ... 5 vuotta ennen konkurssia

vuotta konkurssiin	0	1	2	3	4	5	yhteensä
tilinpäätösten määrä	20	183	260	210	144	105	922

5.2 Käytetyt tunnusluvut

Tilinpäätösanalyysissä vallitsee runsaudenpula käyttökelpoisista tunnusluvuista. Rahoitusteoria ei ole kuitenkaan toistaiseksi löytänyt selkeää vastausta siihen, mitä lukuisat tilinpäätöksistä johdettavat eri tunnusluvut kertovat yrityksestä ja kuinka ne suhtautuvat toisiinsa. Tarkastellessaan useita aiheesta tehtyjä tutkimuksia Laitinen [2] toteaa, että teoreettiset kriteerit tunnuslukujen valinnalle puuttuvat, minkä vuoksi käytetyt tunnusluvut valitaan yleensä empiirisesti. Erot aineistoissa, vallitsevissa suhdanteissa jne. – mieltymyksissäkin – johtavat tällöin siihen, että käytetyt tunnusluvut vaihtelevat tutkimuksesta toiseen. Käytetyt aineistot ovat lisäksi usein niin pieniä, että vertailut eri tunnuslukujen tehokkuudesta eivät tuota tilastollisesti merkitseviä tuloksia.

Yleinen valinta on ollut etsiä yksi tai useampia tunnuslukuja mittaamaan kutakin seuraavista yrityksen ominaisuuksista: (1) tulorahoituksen riittävyys eli kannattavuus, (2) maksuvalmius sekä (3) vakavaraisuus – kuten Prihti tutkimuksessaan suositteli (kts. luku 2.2.3). Tyypillisesti tunnuslukuja tarkastellaan jollakin tietyllä ajan hetkellä eli yhdessä tilinpäätöksessä, vaikka joissakin tutkimuksissa on myös tarkasteltu niiden muutosta jollakin aikajaksolla [2].

Tässä tutkimuksessa on otettu työhypoteesiksi, että ratkaisevaa on ominaisuuden (1) eli kannattavuuden käyttäytyminen kahtena perättäisenä vuotena[2] suhteessa ominaisuuteen (3), vakavaraisuuteen. Hypoteesi perustuu Kera Oy:llä aiemmin tehtyihin selvityksiin ja niiden pohjalta laadittuun, toimivaksi osoittautuneeseen mallin. Tarkastelun kohteeksi valittiin seuraavat neljä tunnuslukua:

1. käyttökateprosentti
2. rahoitustulosprosentti
3. rahoitustulosprosentti edellisessä tilinpäätöksessä
4. omavaraisuusaste

Näiden määritelmät on esitetty liitteessä A.

5.3 Datan esikäsittely

Työssä kokeiltiin kaikkia kolmea kappaleessa 3.2.1 kuvattua datan esikäsittelykeinoa perusmuodoissaan. Kaikille näytevektoreille tehtiin vain esikäsittelykeinojen mukaiset muunnokset; mitään vektoria ei karsittu pois tai typistetty lähimpään "tavalliseen" vektoriin poikkeavien arvojen vuoksi.

[2] Esimerkiksi Prihti laski omassa tutkimuksessaan [1] kannattavuuden keskiarvona kahdesta uusimmasta tilinpäätöksestä, mutta muut tunnusluvut suoraan tuoreimmasta tilinpäätöksestä

Luku 6

Tulokset

6.1 Konkurssoitumisilmiön tarkastelua

Tilinpäätösaineistolla opetettiin kartat kolmella eri tavoin esikäsitellyllä datajoukolla. Opetus tapahtui kahdessa vaiheessa; opetussekvenssin parametrit on esitetty taulukossa 6.1. Näin saatujen karttojen umat-kuvat ja Sammonin projektiot näkyvät kuvassa 6.1. Varianssi normalisoiden saadulla datalla opetetun kartan topografinen virhe oli 0,023; yksikköjananormalisoidulle datalla opetetulla kartalla se oli 0,017 ja histogrammi tasoittamalla saadulla datalla opetulla kartalla 0,019. Karttojen kvantisointivirheet eivät datajoukkojen erilaisten "mittakaavojen" vuoksi olleet vertailukelpoisia.

Taulukko 6.1: Karttojen opetuksessa käytetyt parametrit

			Vaihe I	Vaihe II
Koko:	18 x 12	Opetusaskeleita	10 000	25 000
Naapurustofunktio:	gaussinen	α_0	0,5	0,5
Topologia:	heksagonaalinen	σ_0	10	3

Vertailun vuoksi kokeiltiin myös karttojen opettamista kuplanaapurustoa käyttäen. Tällöin topografiseksi virheeksi tuli datan varianssi normalisoiden 0,068; yksikköjananormalisoinnilla virhe oli 0,047 ja histogrammin tasoituksella 0,065. Suurempi topografinen virhe näkyy myös karttojen komponenteittaisessa tarkastelussa: kuplanaapurustolla opetut kartat olivat hiukan kirjavampia, yhtenäiset vaaleat tai tummat alueet olivat niissä pienempiä kuin gaussisella naapurustolla opetetuissa kartoissa. Erot eivät kuitenkaan olleet suuria.

Umat-kuvissa kartta on tumma siellä, missä karttayksiköitten painovektorit ovat kaukana toisistaan, ja vaalea siellä, missä ne ovat lähellä toisiaan. Tässä "kaukana" ja "lähellä" on määritelty suhteellisesti, kullekin kartalle erikseen. Varianssi- ja yksikköjananormalisoitujen datajoukkojen kartoilla umat-kuvissa yksi nurkka on selvästi tummempi kuin muut. Kyseinen nurkka kurkottaa epätavallisen suuria negatiivisia tunnuslukuja kohti, valtaosa kartasta eli "normaalien" tunnuslukujen alue jää tasaisen vaaleaksi. Selvästi erillisiä ryppäitä datasta ei kuitenkaan löytynyt: tumma alue kuuluu samaan ryppääseen kuin muutkin, mutta sen häntäosaan.

Saman ilmiön havaittiin toistuvan myös varianssi- ja yksikköjananormalisoitujen karttojen komponenteittaisessa tarkastelussa: muu kartta on tasaisen vaaleaa, yksi nurkka tummaa. Tällaisella painovektoreiden välisien etäisyyksien tai vektorikomponenttien arvojen mukaan lineaarisesti muuttuvalla harmaasävykuvalla ei siksi pystytty kovin hyvin analysoimaan eri vektorikomponenttien – tunnuslukujen – välisiä riippuvuuksia.

Tilanne oli toinen histogrammin tasoituksella saadulla datalla opetetun kartan kohdalla. Umat-kuva on huomattavasti tasaisempi, samoin Sammonin projektio näyttää siistimmältä. Suurin ero muilla tavoilla normalisoiduilla datajoukoilla opetettuihin karttoihin oli kuitenkin komponteittaisessa tarkastelussa: histogrammin tasoituksella saadulla datalla opetetussa kartassa eri vektorikomponentit sijoittuivat kartalle selvästi eri tavoin. Siksi tästä eteenpäin kappaleen loppuun asti tarkastellaan vain tätä karttaa.

Kartan opetuksessa käytettyjen tunnuslukujen ryhmittyminen kartalle on esitetty kuvassa 6.2. Kuvassa näkyy lisäksi kullekin karttayksikölle kuvautuneiden konkurssiyritysten ja terveiden yritysten antamien tilinpäätösten antamien tilinpäätösten määrät – kriteerinä jaolle on pidetty sitä, onko yritys mennyt konkurssiin viiden vuoden kuluessa tilinpäätöksen antamisesta.

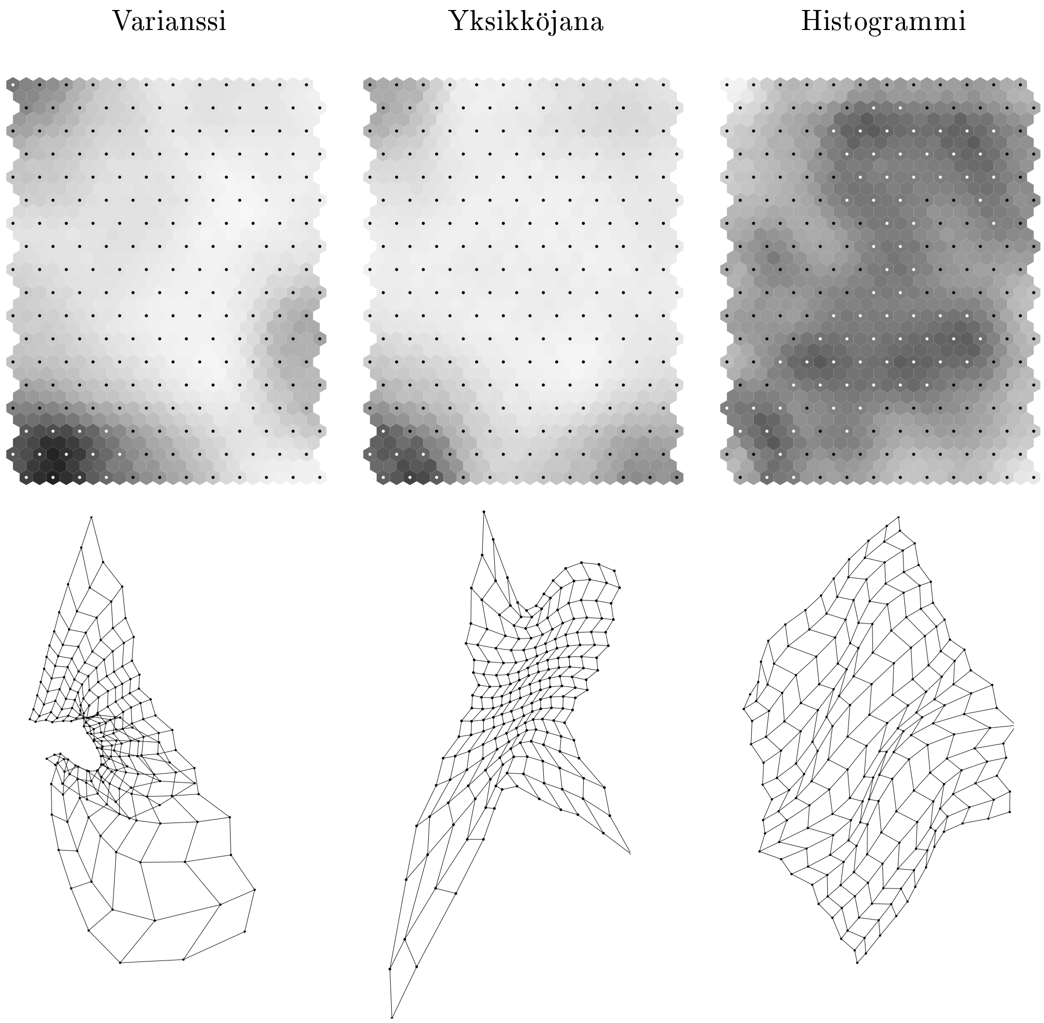
Kuvassa 6.3 on esitetty tunnuslukujen lisäksi eri vuosina ennen konkurssia annettujen tilinpäätösten sijoittuminen kartalle. Kartta on opetettu ohjattu SOM-menetelmällä siten, että tunnusluvut vastaavat kappaleen 3.4.2 merkintöjä käyttäen opetusvektorin $\mathbf{x}$-osaa, eli voittajayksikkö on etsitty pelkästään näiden komponenttien perusteella. Opetusvektoreissa on lisäksi ollut mukana $\mathbf{t}$-osa: kuusi binääristä komponenttia, jotka osoittavat, onko tilinpäätöksen antanut yritys mennyt konkursiin $0, 1, \ldots, 5$ vuoden kuluessa tilinpäätöksen antamisesta, ja seitsemäs, joka osoittaa, onko se ylipäätään mennyt konkursiin enintään viiden vuoden kuluessa tilinpäätöksen antamisesta, ts. $t_7 = t_1 \vee t_2 \vee t_3 \vee t_4 \vee t_5 \vee t_6$. Voittajan löydyttyä on kartan painovektoreiden $\mathbf{t}$-komponentteja on päivitetty samoin kuin $\mathbf{x}$-komponenttejakin opetussäännön (3.2) mukaisesti.

6.2 Luokittelu terveisiin ja konkurssiyrityksiin

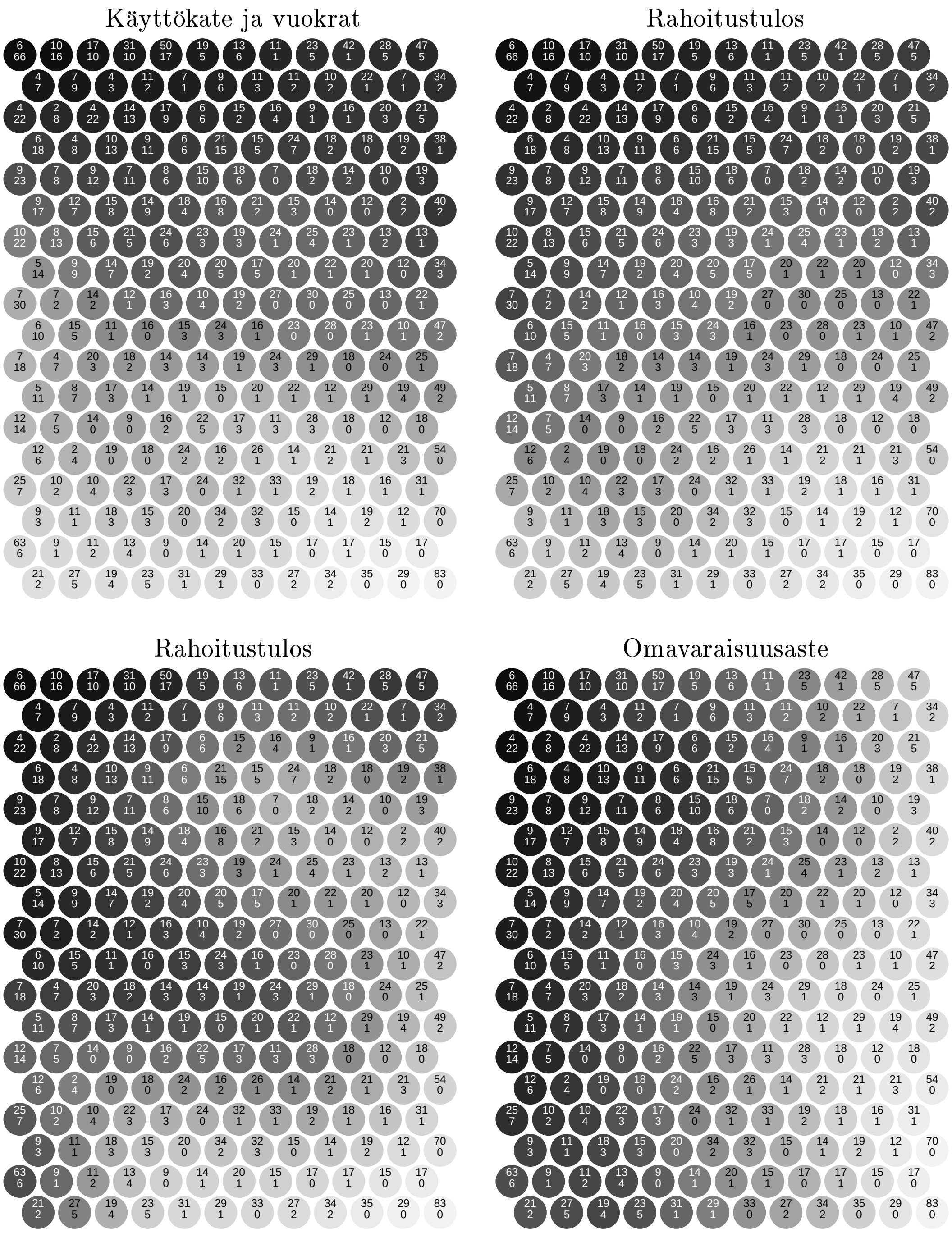
Luokittelutehtävässä pyrittiin löytämään ne yritykset, jotka ajautuivat konkurssiin, ennen kuin tilinpäätöksen kattaman tilivuoden päättymisestä oli kulunut kahta vuotta.

6.2.1 Luotettavuuskontrolli: Ristiinvalidointi

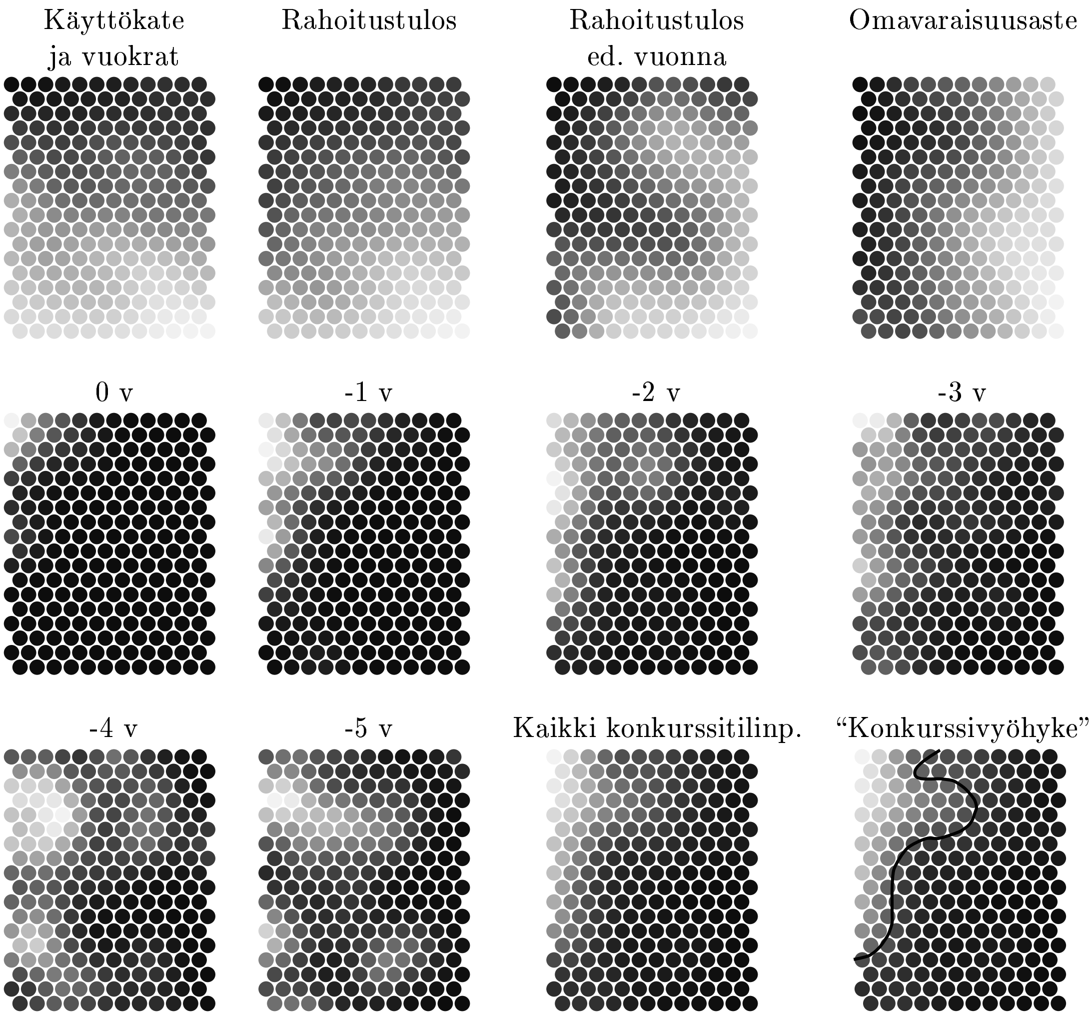
Luokittelu on tehty käyttäen viisinkertaista ristiinvalidointia: luokitin opetettiin käyttäen neljää viidesosaa aineistosta, ja luokitustulos evaluoitiin jäljellejääneellä viidesosalla. Tämä toistettiin viisi kertaa siten, että evaluoinnissa käytetty viidesosa vaihtui joka kerta. Tulokseksi ilmoitettiin saadun viiden luokitustuloksen keskiarvo; Neyman-Pearsonin kriteeriä käytettäessä laskettiin lisäksi tulosten keskihajonta, joka antaa jonkinlaisen käsityksen luokittelutulosten vaihtelusta.



Kuva 6.1: Eri tavoin normalisoitujen karttojen umat-kuvat ja Sammonin projektiot



Kuva 6.2: Tunnuslukujen sijoittuminen kartalle. Karttayksiköihin on merkitty niille kuvautuneiden terveiden (ylempi luku) ja konkurssiyritysten (alempi luku) antamien tilinpäätösten määrät, kun kriteerinä on pidetty sitä, onko yritys mennyt konkurssiin viiden vuoden kuluessa tilinpäätöksen antamisesta.



Kuva 6.3: Tunnuslukujen vs. konkurssitilinpäätösten eri vuosina ennen konkurssia kuvautuminen kartalle; mitä vaaleampi väri, sitä suurempi konkurssitilinpäätösten suhteellinen osuus. Alhaalla oikealla on erotettu viivalla se alue, jonka vasemmalla puolella oleville yksiköille kuvautuvista tilinpäätöksistä yli kolmannes on konkurssiyritysten antamia.

Tilinpäätösten jako viiteen ryhmään tehtiin siten, että ryhmät pyrittiin saamaan keskenään mahdollisimman samankaltaisiksi ja riippumattomiksi. Ryhmät täytettiin tilinpäätöksen antaneen yrityksen asiakasnumeron perusteella: aineiston pienimmän asiakasnumeron omistavan yrityksen tilinpäätökset sijoitettiin ensimmäiseen ryhmään, seuraavan toiseen ryhmään ja samoin jatkaen aina viidenteen ryhmään asti; kuudenneksi pienimmän asiakasnumeron omistavan yrityksen tilinpäätökset sijoitettiin jälleen ensimmäiseen ryhmään ja niin edelleen.

Tällä tavoin etsittiin parhaan tuloksen antavat parametrit luokittimille, ja tämä tulos on myös ilmoitettu kunkin luokittimen luokitustuloksena. On huomattava, että periaatteessa tämä ei vielä riitä kertomaan luokittimen toiminnasta riippumattomalla testiaineistolla: on mahdollista, että löydetyt parhaat parametrit toimivat hyvin nimenomaan näiden viiden opetus/testausparin kohdalla mutta huonosti jollakin toisella samasta jakaumasta peräisin olevalla aineistolla. Tällaisenaankin tuloksia voidaan kuitenkin pitää vähintäänkin suuntaa-antavina; tulosten pienet keskihajonnat ja aineiston laajuus antavat syytä uskoa, että riippumattomalla, riittävän suurella testiaineistolla saataisiin varsin samankaltaiset tulokset.

kNN-luokitin poikkesi muista luokittimista: algoritmin luonteen vuoksi se evaluoitiin *leave-one-out*-menetelmällä, toisin sanoen luokitustulos saatiin käymällä aineisto läpi tilinpäätös kerrallaan, jolloin luokittimen "opetukseen" käytettiin kaikkia muita paitsi kyseistä tilinpäätöstä.

6.2.2 Luokittelutulokset

Taulukossa 6.2 on esitetty muutamalla yleisesti käytetyllä luokittimella saadut luokittelutulokset, kun virhekriteerinä oli luokitteluvirheiden kokonaismäärän minimointi.

LVQ-algoritmin eri varianteista tässä käytettiin OLVQ1-algoritmia kahdella eri LVQ_PAKin alustusmahdollisuudella: eveninitillä (e), jossa kumpaankin luokkaan sijoitettiin yhtä monta prototyyppivektoria, sekä propinitillä (p), jossa prototyyppivektorien lukumäärä oli suhteessa luokan *a priori* -todennäköisyyteen. Muut LVQn parametrit, samoin kuin kNN-luokittimen k-arvo, haettiin kokeilemalla niin, että luokitteluvirhe saatiin mahdollisimman pieneksi. LVQ-algoritmien tulokset on ilmoitettu käyttäen dataa, jonka histogrammin oli tasoitettu, jolloin tulokset olivat niukasti parempia kuin datan varianssi normalisoiden saatiin. Muiden algoritmien kohdalla tilanne oli päinvastainen: varianssin normalisointi toimi hitusen paremmin, joten niiden tulokset on ilmoitettu sitä käyttäen – LDA:n ja QDA:n kohdalla samoihin tuloksiin päästäisiin tietysti myös yksikköjananormalisoinnilla.

Pelkkä virheluokitusten kokonaismäärän minimointi johtaa kestämättömän suureen tyypin I virheen suhteelliseen osuuteen, kuten taulukosta 6.2 nähdään. Yli puolet konkurssiin ajautuneista yrityksistä luokitellaan toimiviksi, vaikka juuri ne pitäisi tämänkaltaisen "hälytysjärjestelmän" löytää mahdollisimman luotettavasti. Tyypin II virheluokitukset, joita nyt tehdään suhteellisen harvoin, eivät sen sijaan ole yhtä haitallisia. Taulukossa 6.3 on esitetty Neyman-Pearsonin virhekriteerillä saatuja tuloksia eri virhetyypin I ohjearvoilla; mLVQ tarkoittaa Neyman-Pearsonin kriteerille modifioitua LVQta.

SOM-2-luokitin käyttäytyi mielenkiintoisesti kartan koon kasvaessa. SOM-1-luokitin

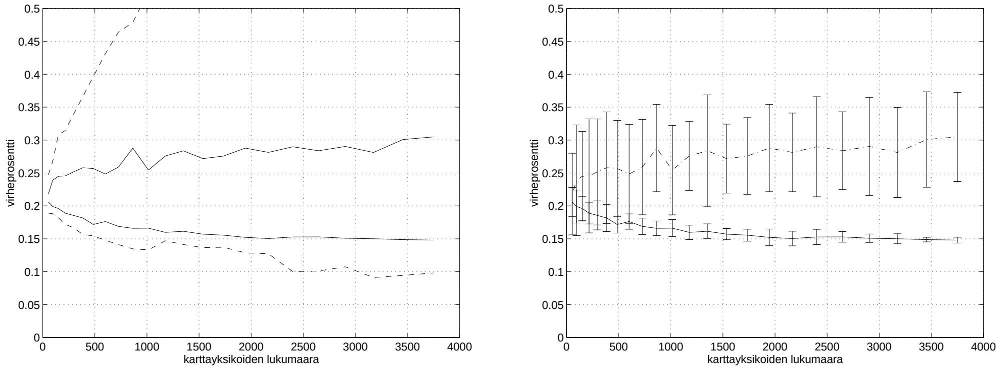
Taulukko 6.2: Luokittelutuloksia, kun päämääränä oli kokonaisvirheen minimointi (luvut prosentteina)

Luokitin	kokonaisvirhe	virhe I	virhe II
LVQ (e)	9,0	55,5	4,3
LVQ (p)	8,6	65,2	2,7
kNN (k=15)	8,5	75,2	1,5
LDA	10,5	47,1	6,6
QDA	11,1	55,9	6,5

Taulukko 6.3: Luokittelutuloksia käyttäen Neyman-Pearsonin virhekriteeriä eri virheen I ohjearvoilla (luvut prosentteina)

Luokitin	vI-ohje	kokonaisvirhe	kh	virhe I	kh	virhe II	kh
LDA	0,20	19,0	(1,6)	21,0	(6,2)	18,8	(2,3)
	0,25	15,7	(1,0)	25,7	(5,4)	14,6	(1,5)
	0,30	14,1	(1,0)	29,5	(5,1)	12,5	(1,4)
mLVQ	0,20	20,5	(2,0)	21,9	(4,1)	20,3	(2,4)
	0,25	15,9	(0,8)	25,7	(5,4)	14,9	(1,6)
	0,30	14,3	(1,0)	30,3	(4,5)	12,5	(1,5)
RBF-SOM	0,20	18,3	(1,2)	20,7	(5,6)	18,1	(1,7)
	0,25	15,8	(0,8)	26,4	(6,1)	14,7	(1,3)
	0,30	13,5	(1,0)	30,5	(6,4)	11,7	(1,6)
SOM-2	0,20	20,1	(1,9)	19,9	(6,3)	20,1	(2,6)
	0,25	16,6	(1,3)	25,4	(6,7)	15,7	(2,0)
	0,30	14,8	(0,4)	30,4	(6,8)	13,2	(0,9)
SOM-1	0,25	18,9	(2,8)	24,7	(6,3)	18,4	(3,4)
	0,30	18,2	(1,6)	30,8	(9,2)	16,9	(2,4)

vaatii useita näytevektoreita jokaiselle karttayksikölle, jotta se toimisi hyvin myös muulle kuin opetusaineistolle. Samantapaista ilmiötä voisi odottaa myös SOM-2-luokittimessa, mutta kartan koon kasvaessa luokittelutarkkuus pieneni yllättävän hitaasti, kuten kuvasta 6.4 ilmenee: luokitin toimi hyvin vielä silloinkin, kun kartassa oli kutakuinkin yhtä monta yksikköä kuin sen opettamiseen oli käytettävissä näytevektoreita.



Kuva 6.4: Vasemmalla: SOM-1-luokittimen (katkoviivalla) ja SOM-2-luokittimen kokonaisvirheet (alemmat viivat) ja tyypin I virheet; oikealla: SOM-2-luokittimen tulokset keskihajontoineen

Luku 7

Pohdintoja

7.1 Konkurssi-ilmiön tutkiminen

7.1.1 SOM tilinpäätösaineiston visualisoinnissa

SOM osoittautui hyväksi työkaluksi konkurssi-ilmiön tutkimisessa. Useista eri tavoin laadituista kartoista käyttökelpoisimpia olivat ne, joissa perustana olevan datan histogrammi oli tasoitettu ja joissa oli käytetty gaussista naapurustofunktiota. Muilla normalisointitavoilla kartan painovektoreiden arvojen tarkastelu komponenteittain ei SOM_PAKin planes-ohjelmalla onnistunut. Syynä tähän on se, että planes kuvaa komponentit harmaasävyiksi lineaarisesti, mutta varianssin normalisoinnin tai yksikköjananormalisoinnin jälkeen datassa olevat ääritapaukset saivat niitä vastaavat karttapisteet kuvautumaan mustiksi tai valkeiksi kartan valtaosan jäädessä tasaisen harmaaksi; komponenttien arvot olisi näissä tapauksissa pitänyt kuvata harmaasävyiksi epälineaarisesti, jotta visuaalinen tarkastelu olisi onnistunut.

Kuplanaapurustolla opetetut kartat olivat melko samankaltaisia gaussisella naapurustolla opetettujen kanssa. Erot eri naapurustofunktioilla opetettujen karttojen topografisessa virheessä ja ulkomuodossa johtunevat siitä, että samalla leveysparametrilla σ kuplanaapurusto on efektiivisesti kapeampi, jolloin kartan on helpompi laskostua, jos syöteavaruuden sisäinen dimensio on suurempi kuin kartan dimensio.

Kuvassa 6.3 käytetty esitysmuoto näyttäisi sopivan konkurssi-ilmiön laadulliseen tarkasteluun hyvin. Jos yksiköiden harmaasävy olisi laskettu suoraan yksikölle kuvautuneiden konkurssiyritysten ja terveiden yritysten antamien tilinpäätösten perusteella, kuvasta olisi tullut jonkin verran hajanaisempi ja vaikeampi tulkita. Nyt naapurustofunktio tasoittaa vierekkäisten yksiköiden arvoja lähemmäs toisiaan ja vähentää siten tilastollisen kohinan vaikutusta tuloksiin. Tasoitus ei ole kuitenkaan liian voimakasta: oikeassa alanurkassa olevassa kuvassa viivalla erotettu “konkurssialue”, jonka vasemmalla puolella oleville karttayksiköille kuvautuvista tilinpäätöksistä ainakin kolmannes on konkurssiyritysten antamia, käy melko hyvin yksiin naapurustofunktiolla tasoitettuun harmaasävykuvaan muodostuvan “konkurssialueen” kanssa.

7.1.2 Tilinpäätökarttojen tarkastelua

Tunnuslukujen erilaiset yhdistelmät muodostavat kartalle kuvissa 6.2 ja 6.3 selkeästi erottuvia alueita. Kannattavuutta mittaavat tunnusluvut, käyttökate ja vuokrat sekä rahoitustulos, kasvavat kartalla alaspäin siirryttäessä ja vakavaraisuutta mittaava tunnusluku, omavaraisuusaste, kartalla oikealle siirryttäessä. Käyttökatteen ja vuokrien heikommat arvot kartan oikeassa ja rahoitustulosten vasemmassa reunassa on helppo ymmärtää, kun niitä tarkastelee yhdessä omavaraisuusasteen kanssa. Yrityksillä, joilla on heikko omavaraisuusaste, rahoituskulut ovat tyypillisesti suuret, joten niillä hyväkään käyttökate ei saa aikaan hyvää rahoitustulosta; vastaavasti yrityksillä, joilla on hyvä omavaraisuusaste ja siten pienet rahoituskulut, käyttökate voi olla huonompikin ja rahoitustulos silti tyydyttävä.

Rahoitustulos tarkasteluvuotta edeltäneenä vuotena näkyy kartalla mielenkiintoisesti. Heikko edellisen vuoden rahoitustulos muodostaa alas oikealle kurkottavan kielekkeen; tällä alueella kuitenkin tarkasteluvuoden tunnusluvuilla on varsin hyvät arvot. Toisaalta kielekkeen yläpuolella on alue, jolla rahoitustulos edellisenä vuonna oli hyvä, mutta tarkasteluvuoden tunnusluvut ovat heikkoja. Olisi mielenkiintoista tietää, miten yritysten peräkkäisten vuosien tilinpäätökset liikkuvat tämän kielekkeen ympärillä – muodostuuko siihen alue, jolla osa yrityksistä pystyy toipumaan huonosta vuodesta ja siirtyy edelleen kohti kartan alaosaa, mutta osa ajautuu huonon vuoden jälkeen kohti vasemman yläreunan "konkurssinurkkaa"?

Konkurssiyritysten tilinpäätökset näyttäisivät kuvan 6.3 perusteella ryhmittyvän varsin selkeästi vielä kolme vuotta ennen konkurssia kartan vasempaan reunaan, sen ylempään puoliskoon. Tällä alueella sijaitsee suuri osa myös neljä ja viisi vuotta ennen konkurssia annetuista tilinpäätöksistä, mutta niitä on myös muissa osissa karttaa. Kuvasta voisi vetää sellaisen varovaisen johtopäätöksen, että konkurssin riski on yhteydessä huonoon omavaraisuusasteeseen, johon yhdistyy lisäksi heikko rahoitustulos kahtena peräkkäisenä vuotena.

7.2 Luokittelu terveisiin yrityksiin ja konkurssiyrityksiin

Yritysten luokittelu terveisiin ja konkurssikypsiin tilinpäätöstietojen perusteella tuotti käyttökelpoisia tuloksia vain silloin, kun luokittelussa käytettiin Neyman-Pearsonin virhekriteeriä. Taulukon 6.2 mukaisesti pelkkä virheluokitusten kokonaismäärän minimointi johti siihen, että konkurssiyritykset tunnistettiin aivan liian epäluotettavasti. Ero tyypin I ja II virheiden suhteellisissa määrissä johtunee suurimmalta osin siitä, että terveitä yrityksiä oli aineistossa huomattavasti enemmän, jolloin konkurssiyritykset tavallaan hautautuivat niiden alle. Tilanne on kuitenkin sama myös sovelluksissa vastaan tulevassa aineistossa: konkurssiyritykset ovat onneksi selvä vähemmistö kaikista yrityksistä. Onkin ehkä syytä suhtautua varauksella tutkimuksiin, joissa luokitin on rakennettu ja evaluoitu otoksella, jossa kumpaankin luokkaan kuuluvia yrityksiä on yhtä paljon; ainakaan niin saadut tulokset eivät ole suoraan vertailukelpoisia tässä työssä saatujen tulosten kanssa.

Erot eri luokittimien suorituskyvyssä olivat yllättävän pieniä. LDAlla saadut hyvät tulokset ovat linjassa aiempien tutkimusten [5, 46, 11] kanssa. Vaikka tilinpäätöstunnuslukujen jakauman on havaittu poikkeavan multinormaalisesta [8, 47], LDAn tehokkuus viittaisi siihen, että poikkeama on ainakin päätöspinnan läheisyydessä varsin pieni. QDAn LDAta heikompi tulos (taulukko 6.2) antaa lisäksi aihetta epäillä, että niin konkurssiyritysten kuin terveidenkin yritysten tilinpäätösvektoreiden kovarianssimatriisit ovat hyvin lähellä toisiaan. Tällöin QDA ei juurikaan hyödy siitä, että luokille estimoidaan kovarianssimatriisit erikseen, pikemminkin päinvastoin: QDAlla estimaatit tulevat näin epätarkemmiksi, koska ne perustuvat pienempään datajoukkoon.

Perusmuodossaan LVQ toimi selvästi paremmin kuin LDA (taulukko 6.2). Neyman-Pearsonin virhekriteerille sovellettuna LVQ sen sijaan hävisi sitä enemmän, mitä "epäsymmetrisemmäksi" opetus muodostui, ts. mitä enemmän kerroin β kaavassa (4.29) poikkesi arvosta 0,5. Syynä tähän saattaa olla se, että kaukana luokitusvirheiden kokonaismäärän minimoivista ("Bayes-optimaalisista") päätöspinnoista pienemmän luokan eli konkurssitilinpäätösten näytevektorit ovat niin harvassa verrattuna toisen luokan vektoreihin, että hyvää yleistyskykyä ei tämän tapaisella algoritmilla voida saavuttaa.

SOM1-luokitin jäi selvästi heikoimmaksi. Sen ongelmana oli, että liian pienellä kartalla syöteavaruutta ei saatu ositettua riittävän hienojakoisesti, liian suurella kartalla taas kullekin karttayksikölle kuvautui niin vähän näytevektoreita, että konkurssin todennäköisyyttä yksikön Voronoi-ympäristössä ei pystytty estimoimaan luotettavasti. Sen sijaan SOM2-luokitin toimi ehkä yllättävänkin hyvin: se hävisi vain noin prosenttiyksikön verran parhaille luokittimille. SOM2-luokittimessa naapurustofunktio tasoitti lähekkäin sijaitsevien karttayksiköiden konkurssin todennäköisyyden *a posteriori*-estimaatteja, jolloin mallin kompleksisuus pieneni huomattavasti siitä, mitä se muutoin olisi ollut – jopa kartta, jossa oli yksiköitä karkeasti ottaen yhtä paljon kuin sen opetukseen oli käytettävissä näytevektoreita, toimi luokittimena hyvin(kuva 6.4). Ilmoitetut tulokset saatiin kartalla, jonka opetuksessa käytettiin gaussista naapurustofunktiota, mutta myös kuplanaapurustolla opetettujen karttojen yleistyskyky oli kutakuinkin yhtä hyvä.

RBF-SOM toimi luokittimista parhaiten. Samoin kuin Neyman-Pearsonin kriteerille sovellettu LVQ, myös RBF-SOM toimi sitä paremmin, mitä lähempänä luokituskriteerissä oltiin virheiden kokonaismäärän minimointia, ja syykin saattaa olla sama: kovin epäsymmetrisessä luokituksessa konkurssitilinpäätösvektoreita oli päätöspintojen tuntumassa liian vähän.

Gaussisten kantafunktioiden leveyden määräävällä parametrilla ρ havaittiin olevan voimakas vaikutus RBF-SOMin suorituskykyyn – selvästi suurempi kuin esimerkiksi käytetyn kartan koolla tai muilla parametreilla. Tämä ei ole yllättävää, sillä parametrin arvosta riippuu, miten hyvin RBF-SOM pystyy yleistämään oppimaansa. Työssä kokeiltiin myös yleistämiskyvyn parantamista painomatriisin $\mathbf{W}$ regularisoinnilla [48], mutta siten ei päästy yhtä hyviin tuloksiin.

7.3 Jatkotutkimusmahdollisuuksia

Erilaisilla luokittimilla saadut samansuuntaiset tulokset antavat aihetta epäillä, että pelkästään luokittimia kehittämällä tämän aineiston käsittelyssä enää saavutettaisiin merkittävää parannusta. Sama pitänee paikkansa erilaisten tunnuslukuyhdistelmien käytön suhteen: työtä parhaiden tunnuslukujen löytämiseksi on tehty aiemmissa tutkimuksissa valtavasti, mutta erot tunnuslukuyhdistelmien tehokkuudessa ovat olleet varsin pieniä.

Näyttäisi siltä, että yksin tilinpäätöstunnuslukujen perusteella voidaan tunnistaa suuri osa konkurssiyrityksistä, mutta ei kaikkia. Usein on esitetty arvio, että 10 ... 15% konkursseista johtuu muista kuin tilinpäätöksistä ilmenevistä syistä [49]. Siksi kiintoisa aihe jatkotutkimukselle olisi tilinpäätöstunnuslukujen lisäksi myös muun informaation käyttäminen konkurssiennusteen pohjana. Perusteita sille, mitä tämä informaatio voisi olla, saataisiin vaikkapa tutkimalla niiden yritysten tekemiä konkursseja, jotka tässä työssä esitetyillä "konkurssikartoilla"(taulukko 6.3) tapahtuvat kaukana varsinaisesta "konkurssialueesta".

Konkurssi-ilmiön dynamiikkaa ei tässä tutkimuksessa tarkasteltu juuri lainkaan; ainoa linkki perättäisten vuosien tilinpäätöksien välillä oli, että tunnuslukujen joukkoon kuului myös edellisen vuoden rahoitustulos. On kuitenkin hyvin mahdollista, että yrityksen useampana peräkkäisenä vuotena antamista tilinpäätöksistä ilmenee sellaista tietoa yrityksen johdon toimintamalleista ja reagointikyvystä muuttuviin olosuhteisiin, jota ei voisi saada selville minkään yksittäisen tilinpäätöksen perusteella. Ilmiön dynamiikan tutkimiseen tässä työssä esitetyt "konkurssikartat" sopisivat hyvin.

Luokittimien välillä erot olivat tällä aineistolla pieniä. Kokeilemisen arvoinen asia voisi olla, miten kNN-luokitin toimii Neyman-Pearsonin kriteerillä. Jos oletus mLVQ:n ja RBF-SOMin suorituskyvyn heikkenemisen syystä pienillä tyypin I virheen ohjearvoilla pitää paikkansa, niin sama asia vaikuttaa myös kNN-luokittimeen.

Ongelmaan auttaisi laajemman opetusjoukon käyttäminen, mikä onnistuisi, jos aineistoa ei karsittaisi aivan niin raskaalla kädellä kuin nyt on tehty. Esimerkiksi muiden kuin osakeyhtiöiden tilinpäätökset voidaan muuntaa yhteismitallisiksi osakeyhtiöiden tilinpäätösten kanssa, jolloin aineisto kaksinkertaistuisi nykyisestään; samoin mukaan valittavien toimialojen karsinnassa voitaisiin käyttää väljempiä kriteerejä.

Luokittelun sijasta olisi kuitenkin arvokkaampaa pyrkiä estimoimaan suoraan konkurssin todennäköisyyttä annettujen tietojen perusteella. Työssä käytetyt luokittimet näyttäisivät sopivan tähän melko hyvin, lukuunottamatta kNN- ja LVQ-luokittimia. Ilmeisesti Neyman-Pearsonin kriteerissä pienillä tyypin I virheen ohjearvoilla viimeiset konkurssiyritykset joudutaan etsimään melko kovalla hinnalla syöteavaruuden alueilta, joilla konkurssin todennäköisyys ei ole suuri. Tällöin konkurssitilinpäätösten luokkaan sijoittuvien yritysten todellinen konkurssiriski vaihtelee huomattavasti.

– Viimeistään tässä vaiheessa työtä lienee paikallaan todeta, että sanoja konkurssiriski, konkurssin todennäköisyys ja muita vastaavia ilmaisuja on käytetty varsin väljässä mielessä. Todennäköisyys on tulkittu käytetyssä aineistossa esiintyneinä frekvensseinä – on oletettu, että jos tietty osa tietynkaltaiset tilinpäätökset antaneista yrityksistä on mennyt konkurssiin tietyn ajan kuluessa, näin tapahtuu vastaisuudessakin. Tulkinnassa ei sinänsä ole mitään vikaa. Sen perusteella voitaneen edelleen olettaa, että jos käytettävissä olevan aineiston koko kasvaa kohti ääretöntä, sen perusteella laadittavat mallit saadaan kuvaamaan aineiston kaltaisten yritysten keskimääräistä käyttäytymistä mielivaltaisen suurella tarkkuudella.

Ongelmia kuitenkin syntyy, jos todennäköisyydet halutaan tulkita taipumuksina (*propensities;* kts. esim. [50]), ts. että kaikilla tietynlaiset tilinpäätökset antaneilla yksitäisillä yrityksillä olisi yhtäläinen taipumus mennä konkurssiin. Oletukselle ei vielä tässä tutkimuksessa löytynyt näyttöä. On mahdollista, että yritykset ovat yksilöitä ja eroavat toisistaan monin tavoin, ja että niiden käyttäytymistä on mahdollista jossain määrin luotettavasti mallintaa ja ennustaa vain joukkona. Mutta toisaalta saattaa olla niinkin, että konkurssialttius on tietty yrityksen ominaisuus, joka indikoituu tilinpäätöstiedoissa aina samalla tavalla riippumatta yrityksen muista ominaisuuksista. Tämä on kiehtova mahdollisuus – mutta sen selvittely jääköön jatkotutkimuksiin, joille tässä työssä on rakennettu perustaa.

Liite A

Tilinpäätös ja keskeisimmät tunnusluvut

Tilinpäätöksessä esitetään yrityksen edellisen tilivuoden tuotot ja kulut summeerattuina tuloslaskelmassa sekä yrityksen varallisuus ja vastuut tilivuoden päättymishetkellä taseessa. Tuloslaskelmalle ja taseelle on lailla määrätty tietty muoto; virallinen tuloslaskelma ja tase kuitenkin aina oikaistaan ennenkuin niitä käytetään yritystutkimuksen tarpeisiin. Taulukoissa A.1 ja A.2 on esitetty oikaistu rahoitustuloksen sisältävä tuloslaskelma sekä oikaistu tase siinä muodossa, missä yritystutkimusneuvottelukunta suosittelee ne laadittavan [51]. Alle on koottu joitakin keskeisimpien tilinpäätöstunnuslukujen määritelmiä.
Kannattavuuden tunnuslukuja:

- Myyntikateprosentti = Myyntikate/Liikevaihto x 100%
- Käyttökateprosentti = Käyttökate/Liikevaihto x 100%
- Käyttökate+vuokrat,%= (Käyttökate + vuokrat + leasingvuokrat)/Liikevaihto x 100%
- Liiketulosprosentti = Liiketulos/Liikevaihto x 100%
- Rahoitustulosprosentti = Rahoitustulos/Liikevaihto x 100%
- Nettotulosprosentti = Nettotulos/Liikevaihto x 100%
- Nettorahoituskuluprosentti = Nettorahoituskulut/Liikevaihto x 100%

Pääoman tuoton tunnuslukuja:

- Kokonaispääoman tuottoprosentti = (Nettotulos + rahoituskulut + verot)/(Oikaistun taseen loppusumma keskimäärin tilikaudella) x 100%
- Sijoitetun pääoman tuottoprosentti = (Nettotulos + rahoituskulut + verot)/(Oikaistu oma pääoma + korollinen vieras pääoma keskimäärin tilikaudella) x 100%
- Oman pääoman tuottoprosentti = Nettotulos/(Oikaistu oma pääoma keskim.) x 100%

Rahoituksen tunnuslukuja:

- Omavaraisuusasteprosentti = Oikaistu oma pääoma / (Oikaistun taseen loppusumma - saadut ennakkomaksut) x 100%
- Suhteellinen velkaantuneisuusprosentti = (Oikaistun taseen velat - saadut ennakkomaksut)/Liikevaihto x 100%
- Velkaantumisaste = Korollinen vieras pääoma/(Oikaistu oma pääoma)

Taulukko A.1: Oikaistu tuloslaskelma

```
LIIKEVAIHTO                                _________
Muut liiketoiminnan tuotot                +_________

Aine ja tarvikekäyttö        _________
Ulkopuoliset palvelut        _________
Henkilöstökulut              _________
Muut muuttuvat kulut         _________
Valmistevaraston lis/väh  -/+_________    -_________

MYYNTIKATE
Henkilöstökulut              _________
Omistajan palkkakorjaus      _________
Vuokrat                      _________
Leasingvuokrat               _________
Muut kiinteät kulut          _________    -_________

KÄYTTÖKATE
Osinkotuotot                +_________
Rahoitustuotot              +_________
Korkokulut                  -_________
Muut rahoituskulut          -_________
Kurssierot                +/-_________
Kurssierojen oikais       +/-_________  +/-_________
Tilikauden verot                          -_________

RAHOITUSTULOS
Suunnitelman mukaiset poistot             -_________

NETTOTULOS
Poistot sijoituksista                     -_________
Satunnaiset tuotot/kulut                +/-_________

KOKONAISTULOS
Poistoeron lisäys/vähennys              -/+_________
Vapaaehtoisten varausten lisäys/vähennys-/+_________
Omistajan palkkakorjaus                 +/-_________
Eläkevastuuvajauksen muutos             +/-_________
Kuluiksi kirjatut aktivoinnit             +_________
Kurssierojen oikaisu                    +/-_________
Muut tuloksen oikaisut                  +/-_________

TILIKAUDEN TULOS
```

Taulukko A.2: Oikaistu tase

VASTAAVAA	VASTATTAVAA
KÄYTTÖOMAISUUS JA PITKÄAIKAISET SIJOITUKSET	OIKAISTU OMA PÄÄOMA
Aineettomat hyödykkeet	Sidottu oma pääoma
Liikearvo	Osake- ja muu peruspääoma
Muut aineettomat hyödykkeet	Vararahasto
Aineelliset hyödykkeet	Arvonkorotukset
Maa- ja vesialueet	Vapaa oma pääoma
Rakennukset ja rakennelmat	Kertyneet voittovarat
Koneet ja kalusto	Tilikauden tulos
Muut aineelliset hyödykkeet	
Arvopaperit ja pitkäaikaiset sijoitukset	Oman pääoman ehtoiset lainat
Tytär- ja osakkuusyhtiöomistukset	Oman pääoman oikaisut
Muut osakkeet ja vakuudet	
Sisäiset lainasaamiset	Varaukset
Muut lainasaamiset ja sijoitukset	Kertynyt poistoero
	Vapaaehtoiset varaukset
Leasingvastuukanta	VIERAS PÄÄOMA
	Pitkäaikainen
VAIHTO-OMAISUUS	Lainat rahoituslaitoksilta
Aineet ja tarvikkeet	Eläkelainat
Valmisteet ja puolivalmisteet	Sisäiset velat
Muu vaihto-omaisuus	Korottomat pitkäaikaiset velat
	Pakolliset varaukset
RAHOITUSOMAISUUS	Muut pitkäaikaiset velat
Saamiset	
Myyntisaamiset	Leasingvastuukanta
Sisäiset myyntisaamiset	Eläkevastuuvajaus
Lainasaamiset	
Sisäiset lainasaamiset	Lyhytaikainen
Muut saamiset	Saadut ennakot
	Ostovelat
Rahat ja muu rahoitusomaisuus	Sisäiset ostovelat
	Siirtovelat
	Sisäiset lyhytaikaiset velat
	Muut korolliset lyhytaik. velat

Kirjallisuutta

[1] Aatto Prihti, *Konkurssin ennustaminen taseinformaation avulla*, Kannattavuus- ja rahoitusanalyysit Ky, 2 edition, 1975.

[2] Erkki K. Laitinen, *Konkurssin ennustaminen*, Vaasan Yritysinformaatio Oy, 1990.

[3] William H. Beaver, "Financial ratios as predictors of failure", *Empirical research in accounting: selected studies. Supplement to Journal of Accounting Research*, vol. 4, pp. 71–127, 1966.

[4] Edward I. Altman, "Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy", *The Journal of Finance*, , no. 4, pp. 589–609, September 1968.

[5] Edward I. Altman, Robert G. Haldeman, and P. Narayanan, "Zeta-analysis. a new model to identify bankruptcy risk of corporations", *Journal of Banking and Finance*, , no. 1, pp. 29–54, 1977.

[6] Stefan Sundgren and Jan Wallin, "Assessing the probability of bankruptcy: an application of the Hosmer-Lemeshow modelling strategy", Tech. Rep. 272, Svenska Handelshögskolan, Vasa, Finland, December 1993.

[7] David Hosmer and Stanley Lemeshow, *Applied Logistic Regression*, John Wiley & Sons Inc., 1989.

[8] Mahmoud Ezzamel and Cecilio Mar-Molinero, "The distributional properties of financial ratios in UK manufacturing companies", *Journal of Business Finance & Accounting*, vol. 17, no. 1, Spring 1991.

[9] Bonifacio Martín del Brío and Carlos Serrano-Cinca, "Self-organizing neural networks for the analysis and representation of data: Some financial cases", *Neural Computing & Applications*, vol. 1, pp. 193–206, 1993.

[10] B. Back, G. Oosterom, K. Sere, and M. van Wezel, "A comparative study of neural networks in bankruptcy prediction", in *Multiple Paradigms for Artificial Intelligence (SteP94)*. Finnish Artificial Intelligence Society, 1994, Finnish Artificial Intelligence Society.

[11] K.Y. Tam and M.Y. Kiang, "Managerial applications of neural networks: the case of bank failure predictions", *Management Science*, vol. 38, no. 7, pp. 926–47, July 1992.

[12] D. Fletcher and E. Goss, "Forecasting with neural networks: an application using bankruptcy data", *Information and Management*, vol. 24, no. 3, pp. 159–67, March 1993.

[13] W. Raghupathi, L.L. Schkade, and B.S. Raju, "A neural network application for bankruptcy prediction", in *Proceedings of the Twenty-Fourth Annual Hawaii International Conference on System Sciences*, V. Milutinovic and B.D. Shriver, Eds., Los Alamitos, CA, USA, Jan 1991, IEEE, vol. 4, pp. 174–55, IEEE Comput. Soc. Press.

[14] Godwin Udo, "Neural network performance on the bankruptcy classification problem", *Computers and Industrial Engineering*, vol. 25, no. 1-4, pp. 377–80, 1993.

[15] J. Tsukuda and S.-I. Baba, "Predicting Japanese corporate bankruptcy in terms of financial data using neural networks", *Computers & Industrial Engineering*, vol. 27, pp. 445–8, September 1994.

[16] R.L. Wilson and R. Sharda, "Bankruptcy prediction using neural networks", *Decision Support Systems*, vol. 11, no. 5, pp. 545–57, June 1994.

[17] Youngohc Yoon, G. Swales, Jr., and T.M. Margavio, "A comparison of discriminant analysis versus artificial neural networks", *Journal of the Operational Research Society*, vol. 44, no. 1, pp. 51–60, January 1993.

[18] D.T. Cadden, "Neural networks and the mathematics of chaos – an investigation of these methodologies as accurate predictors of corporate bankruptcy", in *Proceedings. The First International Conference on Artificial Intelligence on Wall Street*, Los Alamitos, CA, USA, October 1991, IEEE, pp. 52–7, IEEE Comput. Soc Press.

[19] Simon Haykin, *Neural Networks*, MacMillan College Publishing Company, New York, 1994.

[20] Carlos Serrano-Cinca, "Beyond Z-analysis: Self-organizing neural networks for financial diagnosis", Discussion Papers in Acconting and Management Science 94-92, University of Southampton, December 1994, Hyväksytty Decision Support Systems'iin syksyllä 1995.

[21] Teuvo Kohonen, "Self-organized formation of topologically correct feature maps", *Biological Cybernetics*, vol. 43, pp. 59–69, 1982.

[22] Teuvo Kohonen, *Self-Organizing Maps*, Springer Series in Information Sciences 30. Springer, Berlin Heidelberg New York, 1995.

[23] Samuel Kaski and Teuvo Kohonen, "Structures of welfare and poverty in the world discovered by the self-organizing map", Tech. Rep. A24, Neuroverkkojen tutkimusyksikkö / Teknillinen korkeakoulu, February 1995.

[24] H. Ritter and K. Schulten, "Convergence properties of Kohonen's topology conserving maps: Fluctuations, stability, and dimension selection", *Biological Cybernetics*, vol. 60, no. 1, pp. 59–71, Nov. 1988.

[25] Hans-Ulrich Bauer and Klaus R. Pawelzik, "Quantifying the neighborhood preservation of self-organizing feature maps", *IEEE Transactions on Neural Networks*, vol. 3, no. 4, pp. 570–579, July 1992.

[26] Th. Villmann, R. Der, and Th. Martinetz, "A new quantitative measure of topology preservation in Kohonen's feature maps", in *Proc. ICNN'94, the IEEE Int. Conf. on Neural Networks*, Orlando, Florida, USA, June 1994, IEEE, pp. 645–648, IEEE.

[27] Kimmo Kiviluoto, "Topology preservation in self-organizing maps", Tech. Rep. A29, Neuroverkkojen tutkimusyksikkö / Teknillinen korkeakoulu, December 1995, Hyväksytty ICNN'96-konferenssijulkaisuun.

[28] T. Kohonen, J. Hynninen, J. Kangas, and J. Laaksonen, *The Self-Organizing Map Program Package*, Laboratory of Computer and Information Science, Helsinki University of Technology, Espoo, version 3.1 edition, 1995, available via ftp://cochlea.hut.fi/.

[29] John W. Sammon Jr., "A nonlinear mapping for data structure analysis", *IEEE Transactions on Computers*, vol. 5, no. C-18, pp. 401–9, May 1969.

[30] A. Ultsch, "Self-organized feature maps for monitoring and knowledge acquisition of a chemical process", in *International Conference on Artificial Neural Networks (ICANN93)*, S. Gielen and B. Kappen, Eds., London, 1993, pp. 864–7, Springer-Verlag.

[31] Jukka Iivarinen, Teuvo Kohonen, Jari Kangas, and Samuel Kaski, "Visualizing the cluster on the self-organizing map", in *Multiple Paradigms for Artificial Intelligence (SteP94)*. Finnish Artificial Intelligence Society, 1994, pp. 122–6, Finnish Artificial Intelligence Society.

[32] M. A. Kraaijveld, J. Mao, and A. K. Jain, "A non-linear projection method based on Kohonen's topology preserving maps", in *11th International Conference on Pattern Recognition (11 ICPR)*, Los Alamitos, CA, 1992, pp. 41–5, IEEE Comput. Soc. Press.

[33] Teuvo Kohonen, *Self-organization and Associative Memory*, Springer Series in Information Sciences 8. Springer, Berlin Heidelberg New York, 1984.

[34] J. Heikkonen, P. Koikkalainen, and E. Oja, "Self-organizing maps for collision-free navigation", in *World Congress on Neural Networks*. Neural Network Society, July 1993, vol. III of *International Neural Network Society Annual Meeting*, pp. 141–4.

[35] D. S. Broomhead and D. Lowe, "Multivariable functional interpolation and adaptive networks", *Complex Systems*, vol. 2, pp. 321–55, 1988.

[36] J. Moody and C. J. Darken, "Fast learning in networks of locally-tuned processing units", *Neural Computation*, vol. 1, no. 2, pp. 281–94, 1989.

[37] Christopher M. Bishop, *Neural Networks in Pattern Recognition*, Oxford University Press, Oxford, 1995.

[38] E. J. Hartman, J. D. Keeler, and J. M. Kowalski, "Layered neural networks with gaussian hidden units as universal approximations", *Neural Computation*, vol. 2, no. 2, pp. 210–5, 1990.

[39] J. Park and I. W. Sandberg, "Universal approximation using radial basis function networks", *Neural Computation*, vol. 3, no. 2, pp. 246–57, 1991.

[40] J. Park and I. W. Sandberg, "Approximation and radial basis function networks", *Neural Computation*, vol. 5, no. 2, pp. 305–16, 1993.

[41] Simon Haykin, *Modern Filters*, Macmillan Publishing Company, 1989.

[42] Ari Hämäläinen, *Self-organizing map and reduced kernel density estimation*, PhD thesis, Rolf Nevanlinna Instituutti, Helsingin yliopisto, September 1995.

[43] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses", *Phi. Trans. Roy. Soc. London*, vol. 231, pp. 289–337, 1933.

[44] Charles W. Therrien, *Decision, Estimation and Classification – an introduction to pattern recognition and related topics*, John Wiley & Sons, New York, 1989.

[45] Pierre Devijver and Josef Kittler, *Pattern recognition: a statistical approach*, Prentice-Hall International, London, 1982.

[46] J. A. Ohlson, "Financial ratios and the probabilistic prediction of bankruptcy", *Journal of Accounting Research*, vol. 18, no. 1, pp. 109–31, 1980.

[47] Gordon V. Karels and Arun J. Prakash, "Multivariate normality and forecasting of business bankruptcy", *Journal of Business Finance & Accounting*, pp. 573–92, Winter 1987.

[48] Christopher M. Bishop, "Improving the generalization properties of radial basis function neural networks", *Neural Computation*, vol. 3, no. 4, pp. 579–88, Winter 1991.

[49] Kera Oy Pentti Bergius, varatoimitusjohtaja, "Haastattelu 6.5.1996".

[50] James H. Fetzer, *Philosophy of Science*, Paragon House Issues of Philosophy. Paragon House, New York, 1993.

[51] Yritystutkimusneuvottelukunta, *Yritystutkimuksen tilinpäätösanalyysi*, Gaudeamus, 1995.