



NORGE

(12) PATENT

(19) NO

(51) Int Cl⁷

(11) 319838

G 06 N 5/00

(13) B1

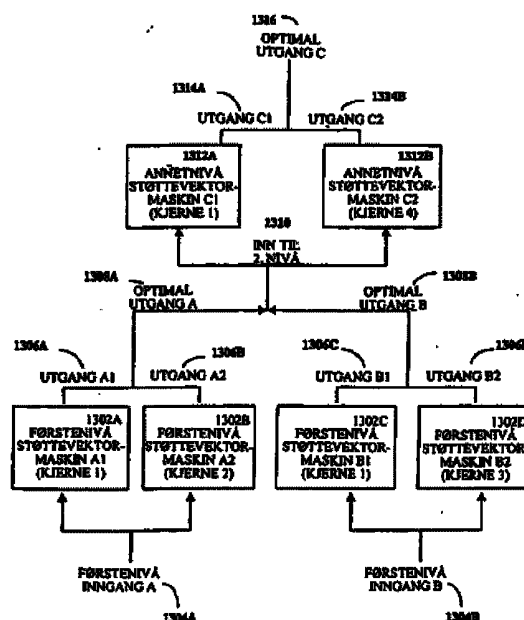
Patentstyret

(21)	Søknadsnr	20015723	(86)	Int.inng.dag og søknadsnr	2000.05.24 PCT/US00/14326
(22)	Inng.dag	2001.11.23	(85)	Videreføringsdag	2001.11.23
(24)	Løpedag	2000.05.24	(30)	Prioritet	1999.05.25, US, 135715
(41)	Alm.tilgj	2002.01.23			
(45)	Meddelt	2005.09.19			

(73)	Innehaver	Health Discovery Corp, 6709 Waters Avenue, GA31406 SAVANNAH, US
(72)	Oppfinner	Stephen D Barnhill, Savannah, GA, US
(74)	Fullmektig	Bryn Aarflot AS, Postboks 449 Sentrum, 0104 OSLO, NO

(54)	Benevnelse	Forbedring av kunnskapsoppdagelse fra multiple datasett ved å bruke flere støttevektormaskiner
(56)	Anførte publikasjoner	Osuna, E; Freund, R; Girosi, F; Support Vector Machines: Training and Applications; MIT C:BC:I Paper No 144; March 1997.
(57)	Sammendrag	

Det er beskrevet et system og en fremgangsmåte for å forbedre kunnskapsoppdagelse fra data ved å benytte flere læremaskiner generelt, og flere støttevektormaskiner spesielt. Treningsdata for en læremaskin blir forbehandlet for å tilføye mening til disse. Forbehandling av data kan innebære å transformere datapunktene og/eller å ekspandere datapunktene. Ved å tilføye mening til dataene blir læremaskinen forsynt med en større mengde informasjon for behandling. Med hensyn til støttevektormaskiner spesielt, er det slik at jo større den informasjonsmengde som skal behandles er, jo bedre generaliseringer omkring dataene kan oppnås. Flere støttevektormaskiner, som hver omfatter distinkte kjerner, blir trent med de forbehandlede treningsdata og blir testet med testdata som er forbehandlet på samme måte. Testutgangene fra flere støttevektormaskiner blir sammenlignet for å bestemme hvilken av testutgangene, om noen, som representerer en optimal løsning. Valg av én eller flere kjerner kan justeres, og én eller flere støttevektormaskiner kan trenes og testes på nytt. Optimale løsninger basert på distinkte inngangsdatasett, kan kombineres for å danne et nytt inndatasett som skal mates inn i én eller flere ytterligere støttevektormaskiner.



Foreliggende oppfinnelse gjelder bruk av læremaskiner til å oppdage kunnskap fra data. Mer spesielt vedrører foreliggende oppfinnelse optimaliseringer for læremaskiner og tilhørende inn- og ut-data for å forbedre den kunnskap som kan oppdages fra multiple datasett.

5 Kunnskapsoppdagelse er det mest ønskelige sluttprodukt fra datainnsamling. Nylige fremskritt i databaseteknologi har ført til en eksplosiv vekst i systemer og fremgangsmåter for generering, innsamling og lagring av uhyre store datamengder. Selv om databaseteknologi muliggjør effektiv innsamling og lagring av store datasett, blir utfordringen med å lette menneskelig forståelse av informasjonen i disse data enda vanskeligere. Med mange eksisterende teknikker har problemet blitt uløselig. Det er således fremdeles et behov for en ny generasjon automatiske verktøy for kunnskapsoppdagelse.

Som et spesielt eksempel befolker the Human Genome Project en multi-gigabyte database som beskriver den menneskelige genetiske kode. Før denne 15 kartleggingen av det menneskelige Genome er fullstendig (forventet i 2003), er størrelsen av databasen ventet å vokse betydelig. De uhyre store datamengder i en slik database overvelder tradisjonelle verktøy til dataanalyse, slik som regneark og ad hoc undersøkelser. Tradisjonelle metoder for dataanalyse kan brukes til å skape informative rapporter fra data, men har ikke evne til intelligent og automatisk å hjelpe mennesker til å analysere og finne mønstre for nyttig kunnskap i store 20 mengder data. Ved å bruke tradisjonelt aksepterte referanseområder og standarder til tolkning, er det likeledes ofte umulig for mennesker å identifisere mønsteret av nyttig kunnskap selv med meget små datamengder.

Et nylig fremskritt som har vist seg å være effektivt i noen eksempler på 25 maskinlæring, er det tilbakeforplantende neurale nettverk. Neurale tilbakeforplantningsnettverk er læremaskiner som kan trenes til å oppdage kunnskap i et datasett som ikke er umiddelbart synlig for et menneske. Det er imidlertid mange forskjellige problemer med neurale tilbakeforplantningsnettverk som hindrer neurale nettverk fra å være velregulerte læremaskiner. En betydelig ulempe med neurale 30 tilbakeforplantningsnettverk er f.eks. at den empiriske risikofunksjon kan ha mange lokale minima, et tilfelle som lett kan maskere den optimale løsning fra oppdagelse ved hjelp av denne teknikken. Vanlige optimaliseringsprosedyrer som

benyttes i forbindelse med neurale tilbakeforplantningsnettverk kan konvergere til et minimum, men den neurale nettverksmetoden kan ikke garantere at selv et lokalisert minimum blir oppnådd, enda mindre det ønskede globale minimum. Kvaliteten av den løsning som oppnås fra neurale nettverk, avhenger av mange faktorer. Spesielt bestemmer dyktigheten til den person som implementerer det neurale nettverk, det endelige resultat, men også faktorer så tilsynelatende gunstige som det tilfeldige utvalg av innledende vektorer kan føre til dårlige resultater. Konvergen-
 5 sen til den gradientbaserte metode som brukes i neural nettverkslæring, er videre iboende langsam. En ytterligere ulempe er at sigmond-funksjonen har en skaleringsfaktor som påvirker kvaliteten av tilnærmingen. Den største begrensende faktor med neurale nettverk relatert til kunnskapsoppdagelse, er kanskje den "dimen-
 10 sjonalitetsforbannelsen" som er tilknyttet den disproporsjonale vekst i nødvendig beregningstid og kraft for hvert ytterligere trekk eller dimensjon i treningsdataene.

Manglene ved neurale nettverk blir overvunnet ved å bruke støttevektormaskiner. Generelt uttrykt kartlegger en støttevektormaskin vektorer inn i et høydimensjonalt egenskapsrom gjennom en ikke-lineær kartleggingsfunksjon valgt på forhånd. I dette høydimensjonale egenskapsrommet blir det konstruert et optimalt separerende hyperplan. Det optimale hyperplan blir så brukt til å bestemme ting slik som klasseseparasjoner, regresjonstilpasning eller nøyaktighet ved densitets-
 15 estimering.

Inne i en støttevektormaskin kan dimensjonaliteten til egenskapsrommet være uhyre stort, en fjerde grads polynomkartleggingsfunksjon får f.eks et 200 dimensjonalt innmatingsrom til å bli kartlagt i et egenskapsrom med 1,6 milliarder dimensjoner. Kjernetrikket og Vapnik-Chervonenkis dimensjon gjør det mulig for støttevektormaskinen å trosse "dimensjonalitetsforbannelsen" som begrenser
 25 andre metoder, og effektivt å ulede generaliserbare svar fra dette meget høydimensjonale egenskapsrommet.

Hvis treningsvektorene er atskilt av det optimale hyperplan (eller det generaliserte optiske hyperplan), så er forventningsverdien for sannsynligheten for å begå en feil på et testeksempel begrenset av eksemplene i treningssettet. Denne begrensningen avhenger verken av dimensjonaliteten til egenskapsrommet eller av normen til vektoren av koeffisienter, heller ikke av begrensningen av antallet
 30

inngangsvektorer. Hvis det optimale hyperplan kan konstrueres av et lite antall støttevektorer i forhold til treningssettets størrelse, så vil generaliseringsevnen være høy selv i et uendelig dimensjonsrom.

Støttevektormaskiner som sådanne tilveiebringer en ønskelig løsning for
5 problemet med å oppdage kunnskap fra store mengder inngangsdata. En støttevektormaskins evne til å oppdage kunnskap fra et datasett er imidlertid begrenset i forhold til den informasjon som er innbefattet i treningsdatasettet. Det er følgelig et behov for et system og en fremgangsmåte for forbehandling av data slik at treningsdataene kan forøkes for å maksimalisere kunnskapsoppdagelsen ved hjelp av
10 støttevektormaskinen.

Råutgangen fra støttevektormaskinen kan videre ikke fullstendig avdekke kunnskapen i den lettest tolkbare form. Det er derfor fremdeles et behov for et system og en fremgangsmåte for etterbehandling av utdataene fra en støttevektormaskin for å maksimalisere verdien av den informasjon som leveres for menneskelig eller videre automatisk behandling.
15

Videre er en støttevektormaskins evne til å oppdage kunnskap fra data begrenset av valget av en kjerne. Følgelig er det fremdeles et behov for et forbedret system og en fremgangsmåte for å velge og/eller skape en ønsket kjerne for en støttevektormaskin.

20 Foreliggende oppfinnelse oppfyller de ovenfor beskrevne behov ved å tilveiebringe et system og en fremgangsmåte for å forbedre kunnskap oppdaget fra multiple datasett ved bruk av flere læremaskiner generelt, og flere støttevektormaskiner spesielt. Ett eller flere treningsdatasett blir forbehandlet for å muliggjøre den mest fordelaktige anvendelse av læremaskinen. Hvert treningsdatapunkt omfatter en vektor som har én eller flere koordinater. Forbehandling av treningsdatasettet kan omfatte å identifisere manglende eller feilaktige datapunkter og ta passende skritt for å korrigere de feilaktige data eller om nødvendig å fjerne observasjonen eller hele feltet fra problemets målområde. Forbehandling av treningsdatasettet kan også omfatte å tilføye dimensjonalitet til hvert treningsdatapunkt ved å
25 tilføye én eller flere nye koordinater til vektoren. De nye koordinater som er tildelt vektoren, kan være utledet ved å anvende en transformasjon på én eller flere av
30

de opprinnelige koordinater. Transformasjonen kan være basert på ekspertkunnskap, eller kan være beregningsmessig utledet. I en situasjon hvor treningsdatasettet omfatter en kontinuerlig variabel, kan transformasjonen omfatte optimal kategorisering av den kontinuerlige variable i treningsdatasettet.

5 På denne måten kan de ytterligere representasjoner av treningsdataene som tilveiebringes ved forbehandlingen, forbedre læremaskinens evne til å oppdage kunnskap fra disse. I den spesielle forbindelse med støttevektormaskiner er det slik at jo større dimensjonaliteten til treningssettet er, jo høyere er kvaliteten av de generaliseringer som kan utledes fra denne. Når den kunnskap som skal oppdages fra dataene, gjelder en regresjon eller densitetsestimering eller når treningsutgangen omfatter en kontinuerlig variabel, kan treningsutgangen etterbehandles ved optimal kategorisering av treningsutgangen for å utlede kategoriseringer fra den kontinuerlige variable.

Et testdatasett blir forbehandlet på samme måte som treningsdatasettet. 15 Den trenede læremaskinen blir så testet ved å bruke det forbehandlede testdatasett. En testutgang fra den trenede læremaskin kan etterbehandles for å bestemme om testutgangen er en optimal løsning. Etterbehandling av testutgangen kan omfatte å tolke testutgangen i et format som kan sammenlignes med testdatasettet. Alternative etterbehandlingstrinn kan forbedre den menneskelige tolknings- 20 evne eller egnethet for ytterligere behandling av utdataene.

I forbindelse med en støttevektormaskin sørger foreliggende oppfinnelse også for valget av en kjerne før trening av støttevektormaskinen. Valget av en kjerne kan være basert på tidligere kunnskap om det spesielle problem som skal undersøkes, eller analyse av egenskapene til eventuelle tilgjengelige data som skal brukes i forbindelse med læremaskinen, og er vanligvis avhengig av beskaf- 25 fenheten av den kunnskap som skal oppdages fra dataene. Eventuelt kan det anvendes en iterativ prosess som sammenligner etterbehandlede treningsutganger eller testutganger for å ta en bestemmelse om hvilken konfigurasjon som gir den optimale løsning. Hvis testutgangen ikke er den optimale løsning, kan valget av kjernen justeres og støttevektormaskinen kan trenes og testes om igjen. Når det 30 blir bestemt at den optimale løsning er blitt identifisert, kan et virkelig datasett

samles inn og forbehandles på samme måte som treningsdatasettet. Det forbehandlede ekte datasett blir matet inn i læremaskinen for behandling. Den ekte utgangen fra læremaskinen kan så etterbehandles ved å tolke den ekte utgang i en beregningsmessig utledet alfanumerisk sorterer.

5 I et eksempel på en utførelsesform er det tilveiebrakt et system for å forbedre kunnskap oppdaget fra data ved å bruke en støttevektormaskin. Systemeksempelet omfatter en lagringsanordning for lagring av et treningsdatasett og et testdatasett, og en prosessor for å kjøre en støttevektormaskin. Prosessoren er også opererbar for å samle inn treningsdatasettet fra databasen, forbehandle treningsdatasettet for å forbedre hvert av et antall treningsdatapunkter, trene støttevektormaskinen ved å bruke det forbehandlede treningsdatasett, samle inn testdatasettet fra databasen, forbehandle testdatasettet på samme måte som treningsdatasettet, teste den trenede støttevektormaskin ved å bruke det forbehandlede testdatasett, og som reaksjon på mottakelse av testutgangen fra den trenede støttevektormaskin, å etterbehandle testutgangen for å bestemme om testutgangen er en optimal løsning. Systemet kan også omfatte en kommunikasjonsanordning for å motta testdatasettet og treningsdatasettet fra en fjerntliggende kilde. I et slikt tilfelle kan prosessoren være i stand til å lagre treningsdatasettet i lagringsanordningen før forbehandlingssettet og å lagre testdatasettet i lagringsanordningen før forbehandling av testdatasettet. Eksempelet på systemet kan også omfatte en fremvisningsanordning for fremvisning av de etterbehandlede testdata. Prosessoren i eksempelet kan videre være innrettet for å utføre hver ytterligere funksjon som er beskrevet ovenfor. Kommunikasjonsanordningen kan videre være innrettet for å sende en beregningsmessig utledet, alfanumerisk sorterer til en fjerntliggende kilde.

I et eksempel på en utførelsesform er det tilveiebrakt et system og en fremgangsmåte for å forbedre kunnskapsoppdagelse fra data ved å bruke flere læremaskiner generelt, og flere støttevektormaskiner spesielt. Treningsdata for en læremaskin blir forbehandlet for å tilføye mening til disse. Forbehandling av data kan innebære å transformere datapunkter og/eller ekspandere datapunktene. Ved 30 å tilføye mening til dataene blir læremaskinen forsynt med en større mengde informasjon til behandling. Med hensyn til støttevektormaskiner spesielt, er det slik at jo større informasjonsmengden som behandles er, jo bedre generaliseringer om

dataene kan utledes. Flere støttevektormaskiner som hver omfatter distinkte kjer-
 ner, blir trenet med de forbehandlede treningsdata og blir testet med testdata som
 er forbehandlet på samme måte. Testutgangene fra de flere støttevektormaskiner
 blir sammenlignet for å bestemme hvilken av testutgangene, om noen, som repre-
 5 senterer en optimal løsning. Valg av én eller flere kjerner kan justeres, og én eller
 flere støttevektormaskiner kan trenes og testes på nytt. Når det blir bestemt at en
 optimal løsning er oppnådd, blir ekte data forbehandlet og matet inn i støttevektor-
 maskinen som omfatter den kjerne som produserte den optimale løsning. Den
 ekte utgang fra læremaskinen kan så etterbehandles i en beregningsmessig utle-
 10 det, alfanumerisk sorterer for tolkning av et menneske eller en automatisk data-
 maskinprosess.

I et annet eksempel på en utførelsesform er det tilveiebrakt et system og en
 fremgangsmåte for optimal kategorisering av en kontinuerlig variabel. Et datasett
 som representerer en kontinuerlig variabel, omfatter datapunkter som hvert omfat-
 15 ter et sampel fra den kontinuerlige variable og en klasseidentifiserer. Et antall dis-
 tinkte klasseidentifiserere i datasettet blir bestemt, og et antall kandidatgrupper blir
 bestemt basert på området til samplene og et presisjonsnivå for samplene i data-
 settet. Hver kandidatgruppe representerer et delområde av samplene. For hver
 kandidatgruppe blir entropien til de datapunkter som faller innenfor kandidatgrup-
 20 pen beregnet. For hver sekvens med datagrupper som har en minimalisert kollek-
 tiv entropi blir så et grensepunkt i sampelområdet definert til å være ved grensen
 til den siste kandidatgruppe i sekvensen av kandidatgrupper. Som en iterativ pro-
 sess kan den kollektive entropi for forskjellige kombinasjoner av sekvensielle kan-
 didatgrupper beregnes. Også antall definerte grensepunkter kan justeres for å be-
 25 stemme det optimale antall grensepunkter, som er basert på en beregning av
 minimal entropi. Som nevnt kan eksempelet på systemet og fremgangsmåten for
 optimal kategorisering av en kontinuerlig variabel, brukes til forbehandling av data
 som skal mates inn i en læremaskin, og til etterbehandling av utgangen fra en
 læremaskin.

30 I nok et annet eksempel på en utførelsesform er det tilveiebrakt et system
 og en fremgangsmåte for å forbedre kunnskapsoppdagelse fra data ved å bruke
 en læremaskin generelt, og en støttevektormaskin spesielt, i et distribuert nett-
 miljø. En kunde kan overføre treningsdata, testdata og ekte data til en selgers ser-

ver fra en fjernliggende kilde, via et distribuert nett. Kunden kan også overføre til serveren identifikasjonsinformasjon slik som et brukernavn, et passord og en finansiell kontoidentifiserer. Treningsdataene, testdataene og de ekte data kan være lagret i en lagringsanordning. Treningsdata kan så forbehandles for å tilføye mening til disse. Forbehandling av data kan innebære å transformere datapunktene og/eller å ekspandere datapunktene. Ved å tilføye mening til dataene, blir læremaskinen forsynt med en større mengde informasjon til behandling. Når det gjelder støttevektormaskiner spesielt, er det slik at jo større mengde informasjon som behandles, jo bedre generalisering av dataene kan utledes. Læremaskinen blir derfor trent med det forbehandlede treningsdatasett og blir testet med testdataene som er forbehandlet på samme måte. Testutgangen fra læremaskinen blir etterbehandlet for å bestemme om kunnskapen som er oppdaget fra testdataene, er ønsket. Etterbehandling innebærer å tolke testutgangen i et format som kan sammenlignes med testdataene. Ekte data blir forbehandlet og matet inn i den trenede og testede læremaskin. Den ekte utgang fra læremaskinen kan så etterbehandles i en beregningsmessig utledet, alfanumerisk sorterer for tolkning av et menneske eller en automatisk datamaskinprosess. Før overføring av den alfanumeriske sorterer til kunden via det distribuerte nett, er serveren innrettet for å kommunisere med en finansinstitusjon med det formål å motta midler fra en finanskonto som tilhører kunden og som er identifisert av den finansiell kontoidentifiserer.

Ifølge nok et annet eksempel på en utførelsesform blir én eller flere støttevektormaskiner trent ved å bruke et første forbehandlet treningsdatasett og én eller flere andre støttevektormaskiner blir trent ved å bruke et annet forbehandlet treningsdatasett. De optimale utganger fra like støttevektormaskiner kan så kombineres for å danne et nytt inngangsdasett for én eller flere ytterligere støttevektormaskiner.

Oppfinnelsen er angitt i de vedføyde patentkrav.

Det vises til de vedføyde tegninger, hvor:

fig. 1 er et flytskjema som illustrerer et eksempel på en generell fremgangsmåte for å øke den kunnskap som kan oppdages fra data ved å benytte en læremaskin;

fig. 2 er et flytskjema som illustrerer et eksempel på en fremgangsmåte for å øke den kunnskap som kan oppdages fra data ved å bruke en støttevektormaskin;

fig. 3 er et flytskjema som illustrerer et eksempel på en fremgangsmåte for optimal kategorisering som kan brukes i en selvstendig konfigurasjon eller i forbindelse med en læremaskin for forbehandlings- eller etterbehandlings-teknikker i samsvar med en utførelsesform av foreliggende oppfinnelse;

fig. 4 illustrerer et eksempel på et uekspandert datasett som kan mates inn i en støttevektormaskin;

fig. 5 illustrerer et eksempel på en etterbehandlet utmating generert av en støttevektormaskin ved anvendelse av datasettet på fig. 4;

fig. 6 illustrerer et eksempel på et ekspandert datasett som kan mates inn i en støttevektormaskin;

fig. 7 illustrerer et eksempel på en etterbehandlet utmating generert av en støttevektormaskin ved bruk av datasettet på fig. 6;

fig. 8 illustrerer eksempler på inndata og utdata for en selvstendig anvendelse av den optimale kategoriseringsmetode på fig. 3;

fig. 9 er en sammenligning mellom eksempelet på den etterbehandlede utmating fra en første støttevektormaskin som omfatter en lineær kjerne, og en annen støttevektormaskin som omfatter en polynomkjerne;

fig. 10 er et funksjonsblokkskjema som illustrerer et eksempel på et driftsmiljø for en utførelsesform av foreliggende oppfinnelse;

fig. 11 er et funksjonsblokkskjema som illustrerer et alternativt driftsmiljø for en alternativ utførelsesform av foreliggende oppfinnelse;

fig. 12 er et funksjonsblokkskjema som illustrerer et eksempel på et nett-driftsmiljø for implementering av en ytterligere alternativ utførelsesform av foreliggende oppfinnelse; og

fig. 13 er et funksjonsblokkskjema som illustrerer et hierarkisk system med flere støttevektormaskiner.

30

Foreliggende oppfinnelse tilveiebringer forbedrede fremgangsmåter for å oppdage kunnskap fra data ved å bruke læremaskiner. Selv om flere eksempler på læremaskiner finnes og fremskritt er ventet på dette området, så setter eksem-

plene på utførelsesformer ifølge foreliggende oppfinnelse søkelyset på støttevektormaskinen. Som kjent på området omfatter læremaskiner algoritmer som kan trenes til å generalisere ved å bruke data med kjente resultater. Trenede lærer-maskinalgoritmer kan så anvendes på tilfeller med ukjent resultat til prediksjon eller
 5 forutsigelse. For eksempel kan en læremaskin trenes til å gjenkjenne mønstre i data, estimere regresjon i data eller estimere sannsynlighetsdensitet i data. Læremaskiner kan trenes til å løse en lang rekke problemer som kjent for fagkyndige på området. En trenet læremaskin kan eventuelt testes ved å bruke testdata for å sikre at dens utgang blir validert innenfor en aksepterbar feilmargin. Når en lære-
 10 maskin er trenet og testet, kan virkelige eller ekte data mates inn i denne. Den ekte utgang fra en læremaskin omfatter kunnskap oppdaget fra alle treningsdataene som er tilført de ekte data.

Et første aspekt ved foreliggende oppfinnelse har til hensikt å forbedre kunnskapsoppdagelse ved valgfri forbehandling av data forut for bruk av dataene
 15 til å trene en læremaskin og/eller valgfri etterbehandling av utgangen fra en læremaskin. Generelt sagt omfatter forbehandling av data å reformatere eller forbedre dataene for å gjøre det mulig for læremaskinen å bli anvendt på den mest fordelaktige måte. Etterbehandling innebærer likeledes tolkning av utgangen fra en læremaskin for å oppdage betydningsfulle kjennetegn ved disse. De betydnings-
 20 fulle kjennetegn som skal fastslås fra utgangen, kan være problem- eller data-spesifikke. Etterbehandling medfører tolkning av utgangen i en form som er forståelig av et menneske eller som er forståelig av en datamaskin.

Utførelseseksempler av foreliggende oppfinnelse vil heretter bli beskrevet under henvisning til tegningene, hvor like henvisningstall indikerer like elementer
 25 på de forskjellige figurer. Fig. 1 er et flytskjema som illustrerer en generell fremgangsmåte 100 for å forbedre kunnskapsoppdagelse ved bruk av læremaskiner. Fremgangsmåten 100 begynner ved en startblokk 101 og fortsetter til trinn 102 hvor et spesielt problem blir formalisert for anvendelse til kunnskapsoppdagelse gjennom maskinlæring. Spesielt viktig er en riktig formulering av den ønskede ut-
 30 gang fra læremaskinen. Ved prediksjon av fremtidig ytelse av et verdiinstrument

eller en markedsindeks, vil en læremaskin vanligvis oppnå bedre ytelse ved forutsigelse av den forventede fremtidige endring istedenfor å forutsi det fremtidige prisnivå. Den fremtidige prisforventning kan senere utledes i et etterbehandlings-trinn som diskutert senere i beskrivelsen.

5 Etter problemformalisering tar trinn 103 seg av innsamling av treningsdata. Treningsdata omfatter et sett med datapunkter med kjente karakteristikk. Treningsdata kan innsamles fra én eller flere lokale og/eller fjerntliggende kilder. Innsamlingen av treningsdata kan utføres manuelt eller ved hjelp av en automatisk prosess, slik som kjente elektroniske dataoverføringsmetoder. Et eksempel på en

10 utførelsesform av foreliggende oppfinnelse kan følgelig realiseres i et miljø med datamaskiner sammenkoplet i et nettverk. Eksempler på driftsmiljøer for realisering av forskjellige utførelsesformer av foreliggende oppfinnelse vil bli beskrevet i detalj i forbindelse med figurene 10-12.

I trinn 104 blir så de innsamlede treningsdata etter valg forbehandlet for å

15 gjøre det mulig for læremaskinen å bli anvendt så fordelaktig som mulig for å trekke ut den kunnskap som ligger i treningsdataene. I dette forbehandlingstrinnet kan treningsdataene etter valg utvides gjennom transformasjoner, kombinasjoner eller manipulasjon av individuelle eller flere mål innenfor registreringene av treningsdataene. Ekspandering eller utvidelse av data betyr slik det benyttes her, å endre

20 dimensjonaliteten til inngangsdataene ved å endre det antall observasjoner som er tilgjengelig for å bestemme hvert inngangspunkt (alternativt kan dette beskrives som å tilføye eller slette kolonner i en databasetabell). Som en illustrasjon kan et datapunkt omfatte koordinatene (1,4,9). En ekspandert versjon av dette datapunktet kan resultere i koordinatene (1,1,4,2,9,3). I dette eksempelet kan det ses at de

25 koordinater som er tilføyd det ekspanderte datapunkt, er basert på en kvadratrot-transformasjon av de opprinnelige koordinater. Ved å tilføye dimensjonalitet til datapunktet, gir dette ekspanderte datapunkt en variert representasjon av inndataene som er potensielt mer meningsfull for kunnskapsoppdagelse ved hjelp av en læremaskin. Dataekspansjon i denne betydningen gir muligheter for læremaskiner

30 til å oppdage kunnskap som ikke er lett synlig i de uekspanderte treningsdata.

Ekspandering av data kan omfatte å tilføre enhver type meningsfylt transformasjon til dataene og tilføye disse transformasjonene til de opprinnelige data. Kriteriene for å bestemme om en transformasjon er meningsfull kan avhenge av

inndataene selv og/eller den type kunnskap som søkes fra dataene. Illustrerende datatransformasjonstyper innbefatter: tillegg av ekspertinformasjon; merking; bi-nær omforming; sinus-, cosinus-, tangens-, cotangens- og andre trigonometriske transformasjoner; gruppering av objekter; skalering; probalistisk- og statistisk analyse; signifikanstesting; styrketesting; søking etter todimensjonal-regularitet; skjult Markov-modellering; identifikasjon av ekvivalensrelasjoner; anvendelse av kontingenstabeller; anvendelse av grafteori-prinsipper; frembringelse av vektorkart; addisjon, subtraksjon, multiplikasjon, divisjon, anvendelse av polynomligninger og andre algebraiske transformasjoner; identifikasjon av proporsjonalitet; bestemmelse av diskriminatoreffekt; osv. I forbindelse med medisinske data innbefatter potensielt meningsfylte transformasjoner: assosiering med kjente medisinske, standard referanseområder; fysiologisk trunkering; fysiologiske kombinasjoner; biokjemiske kombinasjoner; anvendelse av heuristiske regler; bestemmelse av diagnosekriterier; kliniske veiesystemer; diagnostiske transformasjoner; kliniske transformasjoner; anvendelse av ekspertkunnskap; merkingsteknikker; anvendelse av annen domenekunnskap; bayesisk nettverkskunnskap; osv. Disse og andre transformasjoner så vel som kombinasjoner av disse, vil kunne finnes av vanlige fagkyndige på området.

De som er fagkyndige på området bør også innse at datatransformasjoner kan utføres uten å tilføye dimensjonalitet til datapunktene. Et datapunkt kan f.eks. omfatte koordinaten (A, B, C). En transformert versjon av dette datapunktet kan resultere i koordinatene (1, 2, 3), hvor koordinaten "1" har en viss kjent relasjon med koordinaten "A", koordinaten "2" har en viss kjent relasjon med koordinaten "B" og koordinaten "3" har en viss kjent relasjon med koordinaten "C". En transformasjon fra bokstaver til tall kan f.eks. være nødvendig hvis bokstavene ikke blir forstått av en læremaskin. Andre typer transformasjoner er mulige uten å tilføye dimensjonalitet til datapunktene, selv i forbindelse med data som opprinnelige er i numerisk form. Videre vil man forstå at forbehandling av data for å tilføye mening til disse, kan medføre å analysere ufullstendige, ødelagte eller på andre måter "skitne" data. En læremaskin kan ikke behandle "skitne" data på en meningsfull måte. Et forbehandlingstrinn kan således innebære rensing av et datasett for å fjerne, reparere eller erstatte "skitne" datapunkter.

Det vises igjen til fig. 1 hvor fremgangsmåten 100 fortsetter i trinn 106, hvor læremaskinen blir trenet ved å bruke de forbehandlede data. Som kjent på området blir en læremaskin trenet ved å justere dens driftsparametere inntil en ønsket treningsutgang blir oppnådd. Bestemmelsen av om en treningsutgang er ønsket, kan utføres enten manuelt eller automatisk ved å sammenligne treningsutgangen med de kjente karakteristikk for treningsdataene. En læremaskin anses å være trenet når dens treningsutgang er innenfor en forutbestemt feilterskel i forhold til treningsdataenes kjente karakteristikk. I visse situasjoner kan det være ønskelig, om ikke nødvendig, å etterbehandle treningsutgangen fra læremaskinen i trinn 107. Som nevnt innebærer etterbehandling av utgangen fra en læremaskin tolkning av utgangen til en meningsfylt form. I forbindelse med et regresjonsproblem kan det f.eks. være nødvendig å bestemme områdekategoriseringer for utgangen fra en læremaskin for å bestemme om inndatapunktene var korrekt kategorisert. I eksempelet med et mønstergjenkjenningsproblem er det ofte ikke nødvendig å etterbehandle treningsutgangen fra en læremaskin.

I trinn 108 blir testdata eventuelt samlet inn som forberedelse til testing av den trenede læremaskin. Testdataene kan være innsamlet fra én eller flere lokale og/eller fjerntliggende kilder. I praksis kan testdata og treningsdata være innsamlet fra den samme eller de samme kilder til samme tid. Testdata- og treningsdata-sett kan således skilles ut fra et felles datasett og lagres i et lokalt lagringsmedium for bruk som forskjellige inndatasett for en læremaskin. Uansett hvordan testdataene er innsamlet, må alle testdata som benyttes, være forbehandlet i trinn 110 på samme måte som treningsdataene ble. Som det bør være opplagt for de som er fagkyndige på området, kan en skikkelig test av læringen bare utføres ved å bruke testdata av samme format som treningsdataene. Ved trinn 112 blir så læremaskinen testet ved å bruke de eventuelt forbehandlede testdata. Testutgangen fra læremaskinen blir eventuelt etterbehandlet i trinn 114 for å bestemme om resultatene er ønskelige. Igjen innebærer etterbehandlingstrinnet tolkning av testutgangen til en meningsfylt form. Den meningsfylte form kan være én som er forståelig for et menneske eller én som er forståelig for en datamaskin. Uansett må testutgangen etterbehandles til en form som kan sammenlignes med testdataene for å bestemme om resultatene var ønskelige. Eksempler på etterbehandlingstrinn innbefatter, men er ikke begrenset til, følgende: optimale kategoriseringsbestemmelser, skale-

ringsteknikker (lineære og ikke-lineære), transformasjoner (lineære og ikke-lineære) og sannsynlighetsestimeringer. Fremgangsmåten 100 slutter ved trinn 116.

Fig. 2 er et flytskjema som illustrerer et eksempel på en fremgangsmåte 200 for å forbedre kunnskap som kan oppdages fra data, ved å bruke en spesiell type læremaskin kjent som en støttevektormaskin (SVM). En SVM implementerer en spesialisert algoritme for å tilveiebringe generalisering under estimering av en flerdimensjonal funksjon fra en begrenset samling med data. En SVM kan være spesielt nyttig til å løse problemer i forbindelse med estimering av avhengighet. Mer spesielt kan en SVM benyttes nøyaktig ved estimering av indikatorfunksjoner (f.eks. mønstergjenkjennelsesproblemer) og funksjoner med reelle verdier (f.eks. funksjonstilnærmelsesproblemer, regresjonsestimeringsproblemer, densitetsestimeringproblemer og løsning av inverse problemer). SVM ble opprinnelig utviklet av Vladimir N. Vapnik. Konseptene som ligger bak SVM er forklart i detalj i hans bok med tittel *Statistical Learning Theory* (John Wiley & Sons, Inc. 1988), som her ved inntas som referanse i sin helhet. En viss kjennskap til SVM'er og den terminologi som brukes i forbindelse med slike, er forutsatt i denne beskrivelsen.

Fremgangsmåteeksempelet 200 begynner ved startblokk 201 og går videre til trinn 202 hvor et problem blir formulert, og så til trinn 203 hvor et treningsdata-sett blir samlet inn. Som beskrevet i forbindelse med fig. 1 kan treningsdata innsamles fra én eller flere lokale og/eller fjernliggende kilder gjennom en manuell eller automatisk prosess. Ved trinn 204 blir treningsdataene forbehandlet etter valg. Igjen omfatter forbehandling av data å forbedre betydningen i treningsdataene ved å rense dataene, transformere dataene og/eller ekspandere dataene. Fagkyndige på området vil forstå at SVM'er er i stand til å behandle inndata som har uhyre stor dimensjonalitet. Jo større dimensjonaliteten til inndataene er, jo bedre generaliseringer er en SVM i virkeligheten i stand til å beregne. Selv om treningsdatatransformasjoner er mulige som ikke ekspanderer treningsdataene, er det derfor i den spesielle forbindelse med SVM'er å foretrekke at treningsdata ekspanderes ved å tilføye meningsfylt informasjon til disse.

Ved trinn 206 blir en kjerne valgt for SVM'en. Som kjent på området vil forskjellige kjerner få en SVM til å produsere varierende kvalitetsgrader i utgangen for et gitt sett med inndata. Valget av en riktig kjerne kan derfor være essensiell for den ønskede kvaliteten av utgangen fra SVM'en. I den utførelsesform av fore-

liggende oppfinnelse kan en kjerne velges basert på tidligere kunnskap om ytelse. Som kjent på området innbefatter eksempler på kjerner polynomiske kjerner, radialbasis-sorteringskjerner, lineære kjerner, osv. I en alternativ utførelsesform kan en kundetilpasset kjerne lages om er spesifikk for et spesielt problem eller en type datasett. I nok en annen utførelsesform kan de mange SVM'er trenes og testes samtidig, hver ved å bruke en forskjellig kjerne. Kvaliteten av utgangene for hver samtidig trenet og testet SVM kan sammenlignes ved å bruke en rekke velgbare eller veide matriser (se trinn 222) til å bestemme den mest ønskelige kjerne.

Så, ved trinn 208 blir det forbehandlede treningsdata satt inn i SVM'en. Ved trinn 210, blir SVM trenet ved å bruke de forbehandlede treningsdata til å generere et optimalt hyperplan. Etter valg kan treningsutgangen fra SVM så etterbehandles i trinn 211. Igjen kan etterbehandling av treningsutgangen være ønskelig eller endog nødvendig ved dette punkt for å beregne områder eller kategorier for utgangen på riktig måte. Ved trinn 212 blir testdata innsamlet på samme måte som tidligere beskrivelser av datainnsamling. Testdataene blir forbehandlet i trinn 214 på samme måte som treningsdataene ble. I trinn 216 blir de forbehandlede testdata matet inn i SVM for behandling for å bestemme om SVM'en var trenet på en ønsket måte. Testutgangen blir mottatt fra SVM'en ved trinn 218 og blir eventuelt etterbehandlet i trinn 220.

Basert på den etterbehandlede testutgang blir det i trinn 222 bestemt om et optimalt minimum ble oppnådd av SVM'en. Fagkyndige på området vil forstå at en SVM er innrettet for å sikre en utgang som har en global minimumsfeil. Som nevnt ovenfor vil imidlertid utgangsresultatene fra en SVM for et gitt datasett vanligvis variere i forhold til valget av en kjerne. Det er derfor i virkeligheten flere globale minima som kan fastslås av en gitt SVM for et gitt datasett. Uttrykket "optimalt minimum" eller "optimal løsning" refererer slik det benyttes her, til et valgt globalt minimum som antas å være optimalt (f.eks. den optimale løsning for et gitt sett med problemspesifikke, forutbestemte kriterier) sammenlignet med andre globale minima bestemt ved hjelp av en SVM. Bestemmelsen i trinn 222 om det optimale minimum er blitt fastslått, kan følgelig medføre å sammenligne utgangen fra en SVM med en historisk eller forutbestemt verdi. En slik forutbestemt verdi kan være avhengig av testdatasettet. I forbindelse med et mønstergjenkjennelsesproblem hvor et datapunkt er klassifisert av en SVM som enten å ha en viss karakteristikk

eller ikke ha karakteristikken, vil f.eks. en global minimumsfeil på 50% ikke være optimal. I dette eksempelet er et globalt minimum på 50% ikke bedre enn det resultat som ville bli oppnådd ved å kaste mynt og krone for å bestemme om datapunktet hadde den bestemte karakteristik. Som et annet eksempel, i det tilfelle hvor flere SVM'er blir trenet og testet samtidig med varierende kjerner, kan utgangene for hver SVM sammenlignes med hver av de andre SVM'ers utganger for å bestemme den praktiske, optimale løsning for det spesielle sett med kjerner. Bestemmelsen av om en optimal løsning er oppnådd, kan utføres manuelt eller ved hjelp av en automatisk sammenligningsprosess.

Hvis det blir bestemt at det optimale minimum ikke er blitt oppnådd ved hjelp av den trenede SVM, går fremgangsmåten videre til trinn 224 hvor kjernevalget blir justert. Justering av kjernevalget kan omfatte å velge én eller flere nye kjerner eller å justere kjerneparametere. I det tilfelle hvor flere SVM'er ble trenet og testet samtidig, kan videre valgte kjerner erstattes eller modifiseres mens andre kjerner kan brukes på nytt til kontrollformål. Etter at kjernevalget er justert, blir fremgangsmåten 200 gjentatt fra trinn 208 hvor de forbehandlede treningsdata blir matet inn i SVM for treningsformål. Når det ved trinn 222 blir bestemt at det optimale minimum er blitt oppnådd, går fremgangsmåten videre til trinn 226, hvor ekte data blir samlet inn på samme måte som beskrevet ovenfor. De ønskede utgangskarakteristikker som var kjent i forbindelse med treningsdataene og testdataene, er ikke kjent i forbindelse med de ekte data.

Ved trinn 228 blir de ekte data forbehandlet på samme måte som treningsdataene og testdataene. Ved trinn 230 blir de ekte forbehandlede data matet inn i SVM for behandling. Den ekte utgang fra SVM'en blir mottatt ved trinn 232 og blir etterbehandlet ved trinn 234. I én utførelsesform av foreliggende oppfinnelse omfatter etterbehandling å konvertere utgangen fra SVM'en til en beregningsmessig utledet alfanumerisk klassifiserer for tolkning av et menneske eller en datamaskin. Fortrinnsvis omfatter den alfanumeriske klassifiserer én enkelt verdi som lett kan forstås av mennesket eller datamaskinen. Fremgangsmåten 200 slutter ved trinn 236.

Fig. 3 er et flytskjema som illustrerer et eksempel på en optimal kategoriseringsmetode 300 som kan brukes til forbehandling av data eller etterbehandling av utgangen fra en læremaskin i samsvar med et eksempel på en utførelsesform av

foreliggende oppfinnelse. Som beskrevet nedenfor, vil i tillegg eksempelet på en optimal kategoriseringsmetode kunne brukes som en selvstendig kategoriserings-
 teknikk uavhengig av læremaskiner. Eksempelet på den optimale kategoriserings-
 metode 300 begynner ved startblokk 301 og fortsetter til trinn 302, hvor inndata-
 settet blir mottatt. Inndatasettet omfatter en sekvens av datasampler fra en konti-
 5 nuerlig variabel. Datasamplene faller innenfor to eller flere klassifiseringskatego-
 rier. I trinn 304 blir så gruppe- og klasse-sporingsvariable initialisert. Som kjent på
 området er gruppevariable relatert til oppløsning og klasse-sporingsvariable er rela-
 tert til antallet klassifikasjoner i datasettet. Bestemmelsen av verdiene for initialise-
 10 ring av gruppe- og klasse-sporingsvariable kan utføres manuelt eller gjennom en
 automatisk prosess, slik som et dataprogram for å analysere inndatasettet. Ved
 trinn 306 blir dataentropien for hver gruppe beregnet. Entropi er en matematisk
 størrelse som måler usikkerheten til en tilfeldig fordeling. I fremgangsmåten 300
 blir entropi benyttet til å måle graderingene av den inngangsvariable slik at maksi-
 15 mal klassifikasjonsevne blir oppnådd.

Fremgangsmåten 300 frembringer en rekke "kutt" av den kontinuerlig vari-
 able, slik at den kontinuerlig variable kan deles i diskrete kategorier. De kutt som
 velges i fremgangsmåten 300, er optimale i den forstand at den gjennomsnittlige
 entropi for hver resulterende, diskret kategori blir minimalisert. Ved trinn 308 blir
 20 det tatt en bestemmelse av om alle kutt er blitt plassert innenfor inndatasettet som
 omfatter den kontinuerlige variable. Hvis alle kutt ikke er blitt anbrakt, blir sekven-
 sielle gruppekombinasjoner testet for grensebestemmelse ved trinn 310. Fra trinn
 310 går fremgangsmåten 300 i sløyfe tilbake gjennom trinn 306 og vender tilbake
 til trinn 308 hvor det igjen blir bestemt om alle kutt er blitt anbrakt innenfor inndata-
 25 settet som omfatter den kontinuerlige variable. Når alle kutt er blitt anbrakt, blir en-
 tropien for hele systemet evaluert ved trinn 309 og sammenlignet med tidligere
 resultater fra testing av flere eller færre kutt. Hvis det ikke kan konkluderes med at
 en maksimal entropitilstand er blitt bestemt, så må andre mulige kuttvalg evalue-
 res, og fremgangsmåten fortsetter til trinn 311. Fra trinn 311 blir et hittil uprøvd valg
 30 av antall kutt valgt, og ovennevnte prosess blir gjentatt fra trinn 304. Når enten
 grensene for den oppløsning som bestemmes av gruppebredden er blitt testet

eller konvergensen til en minimumsløsning er blitt identifisert, blir de optimale klassifiseringskriterier matet ut ved trinn 312, og den optimale kategoriseringsmetode 300 slutter ved trinn 314.

Den optimale kategoriseringsmetode 300 gjør bruk av dynamiske programmeringsteknikker. Som kjent på området kan dynamiske programmeringsteknikker benyttes til i betydelig grad å forbedre effektiviteten ved løsning av visse komplekse problemer ved omhyggelig å strukturere en algoritme for å redusere redundante beregninger. I det optimale kategoriseringsproblem vil den direkte løsning med uttømmende søking gjennom alle mulige datakutt i de kontinuerlige variable data, resultere i en algoritme med eksponensiell kompleksitet og vil gjøre problemet vanskelig selv med moderat dimensjonerte innmatninger. Ved å trekke fordel av den additive egenskapen til målfunksjonen, i dette problemet den gjennomsnittlige entropi, kan problemet inndeles i en rekke delproblemer. Ved riktig formulering av algoritmiske delstrukturer for løsning av hvert delproblem og lagring av løsningene på delproblemene, kan en stor mengde redundant beregning identifiseres og unngås. Som et resultat av å bruke den dynamiske programmeringsløsning, kan eksempelet på den optimale kategoriseringsmetode 300 implementeres som en algoritme med en polynomisk kompleksitet, som kan benyttes til å løse problemer med store dimensjoner.

Som nevnt ovenfor kan eksempelet på den optimale kategoriseringsmetode 300 benyttes ved forbehandling av data og/eller etterbehandling av utgangen fra en læremaskin. Som et forbehandlingstrinn kan f.eks. den optimale kategoriseringsmetode 300 benyttes til å trekke ut klassifiseringsinformasjon fra rådata. Som en etterbehandlingsteknikk kan eksempelet på den optimale kategoriseringsmetode benyttes til å bestemme de optimale grenseverdier for markører objektivt, basert på data, istedenfor å være bestemt av ad hoc løsninger. Det bør derfor være klart at eksempelet på den optimale kategoriseringsmetode 300 kan anvendes ved mønstergjenkjennelse, klassifisering, regresjonsproblemer, osv. Den optimale kategoriseringsmetode 300 kan også benyttes som en selvstendig kategoriseringsteknikk, uavhengig av SVM'er og andre læremaskiner. Et eksempel på en selvstendig anvendelse av den optimale kategoriseringsmetode 300, vil bli beskrevet under henvisning til fig. 8.

Fig. 4 illustrerer et eksempel på et uekspandert datasett 400 som kan benyttes som inngang til en støttevektormaskin. Dette datasettet 400 kalles "uekspandert" fordi ingen ytterligere informasjon er blitt tilføyd dette. Som vist omfatter det uekspanderte datasett et treningsdatasett 402 og et testdatasett 404. Både

5 det uekspanderte treningsdatasett 402 og det uekspanderte testdatasett 404 omfatter datapunkter, slik som f.eks. datapunkt 406, relatert til historiske kliniske data fra undersøkte, medisinske pasienter. Datasettet 400 kan benyttes til å trene en SVM til å bestemme om en brystkreftpasient vil oppleve en gjentakelse eller ikke.

Hvert datapunkt omfatter fire innkoordinater, eller dimensjoner, og en utgangsklassifisering vist som 406a-f som representerer medisinske data innsamlet

10 for hver pasient. Den første koordinaten 406a representerer spesielt "alder", den annen koordinat 406b representerer "østrogenreseptornivå", den tredje koordinaten 406c representerer "progesteronreseptornivå", den fjerde koordinaten 406d representerer "lymfenoder", den femte koordinaten 406e representerer "positive (kreftfarlige) ekstraherte lymfenoder", og utgangsklassifiseringen 406f representerer "gjentakelsesklassifiseringen". Den viktige kjente karakteristikk ved dataene

15 400 er utgangsklassifiseringen 406f (gjentakelsesklassifiseringen), som i dette tilfelle indikerer om den undersøkte medisinske pasient reagerte gunstig på behandling uten på nytt å få kreft ("-1") eller reagerte negativt på behandling med tilbakevending av kreft ("1"). Denne kjente karakteristikken vil bli brukt til læring ved behandling av treningsdataene i SVM'en, vil bli brukt på en evaluerende måte etter at testdataene er matet inn i SVM'en for derved å skape en "blindtest", og vil opp-

20 lagt være ukjent i de ekte data fra nåværende medisinske pasienter.

Fig. 5 illustrerer et eksempel på en testutgang 502 fra en SVM trenet med

25 det uekspanderte treningsdatasett 402 og testet med det uekspanderte datasett 404 som er vist på fig. 4. Testutgangen 502 er blitt etterbehandlet for å kunne forstås av et menneske eller en datamaskin. Som antydnet viser testutgangen 502 at totalt 24 sampler (datapunkter) ble undersøkt av SVM'en og at SVM'en uriktig identifiserte fire av åtte positive sampler (50%) og ukorrekt identifiserte seks av seksten negative sampler (37,5%).

30

Fig. 6 illustrerer et eksempel på et ekspandert datasett 600 som kan brukes som inngang i en støttevektormaskin. Dette datasettet 600 er referert til som "ekspandert" fordi ytterligere informasjon er blitt tilføyd settet. Legg merke til at bortsett

fra den tilføyde informasjon, er det ekspanderte datasett 600 identisk med det uekspanderte datasett 400 som er vist på fig. 4. Den ytterligere informasjon som er tilføyd det ekspanderte datasett, er blitt levert ved å bruke eksempelet på den optimale områdekategoriseringsmetode 300 som er beskrevet under henvisning til fig. 3. Som vist omfatter det ekspanderte datasett et treningsdatasett 602 og et testdatasett 604. Både det ekspanderte treningsdatasett 602 og det ekspanderte testdatasett 604 omfatter datapunkter, slik som f.eks. datapunkt 606, som er relatert til historiske data fra utvalgte medisinske pasienter. Igjen kan datasettet 600 benyttes til å trene en SVM til å lære om en brystkreftpasient vil oppleve en gjentakelse av lidelsen.

Ved anvendelse av den optimale kategoriseringsmetode 300 innbefatter hvert ekspandert datapunkt tjue koordinater (eller dimensjoner) 606a1-3 til 606e1-3, og en utgangsklassifikasjon 606f, som kollektivt representerer medisinske data og kategoriseringstransformasjoner av disse for hver pasient. Spesielt representerer den første koordinaten 606a "alder", den annen koordinat til den fjerde koordinat 606a1-606a3 er variable som kombineres for å representere en alderskategori. Et aldersområde kan f.eks. kategoriseres som "ung", "middelaldrende" og "gammel"-kategorier i forhold til det område med aldre som er tilstede i dataene. En streng med variable "0" (606a1), "0" (606a2), "1" (606a3) kan som vist benyttes til å indikere at en viss aldersverdi er kategorisert som "gammel". Likeledes kan en streng med variable "0" (606a1), "1" (606a2), "0" (606a3) brukes til å indikere at en viss aldersverdi er kategorisert som "middelaldrende". Også en streng med variable "1" (606a1), "0" (606a2), "0" (606a1) kan benyttes til å indikere at en viss aldersverdi er kategorisert som "ung". Fra en undersøkelse av fig. 6 kan det ses at en optimal kategorisering av området "alder"-verdier 606a, ved å bruke fremgangsmåten 300, ble bestemt å være 31-33 = "ung", 34 = "middelaldrende" og 35-49 = "gammel". De andre koordinatene, nemlig koordinaten 606b "østrogenreseptor-nivå", koordinat 606c "progeseteronreseptor-nivå", koordinat 606d "totalt ekstraherte lymfenoder", og koordinat 606e "positive (kreftfarlige) ekstraherte "lymfenoder", er blitt optimalt kategorisert på lignende måte.

Fig. 7 illustrerer et eksempel på en ekspandert testutgang 702 fra en SVM trenet med det ekspanderte treningsdatasett 602 og testet med det ekspanderte datasett 604 som er vist på fig. 6. Den ekspanderte testutgang 702 er blitt etter-

behandlet for å kunne forstås av et menneske eller en datamaskin. Som antyd-
 viser den ekspanderte testutgang 702 at 24 prøver totalt (datapunkter) ble under-
 søkt av SVM'en og at SVM'en ukorrekt identifiserte fire av åtte positive prøver
 (50%) og ukorrekt identifiserte fire av seksten negative prøver (25%). Ved å sam-
 5 menligne denne ekspanderte datautgangen 702 med den uekspanderte dataut-
 gangen 502 på fig. 5, kan det følgelig ses at ekspansjonen av datapunktene fører
 til forbedrede resultater (dvs. en lavere global minimumsfeil), spesielt et redusert
 antall pasienter som unødvendig vil bli underkastet etterfølgende kreftbe-
 handlinger.

10 Fig. 8 illustrerer et eksempel på en inngang og en utgang for en selvstendig
 anvendelse av den optimale kategoriseringsmetode 300 som er beskrevet på
 fig. 3. I eksempelet på fig. 8 omfatter inndatasettet 801 et "antall positive lymfeno-
 der" 802 og et tilsvarende "tilbakefallsklassifisering" 804. I dette eksempelet er den
 optimale kategoriseringsmetode 300 blitt anvendt på inndatasettet 801 for å lokali-
 15 sere det optimale grensepunkt for å bestemme behandling av krefttilbakefall, bas-
 ert ene og alene på antallet positive lymfenoder som er innsamlet i en etterfølg-
 ende, kirurgisk vevsprøve. Den velkjente kliniske standard er å foreskrive behand-
 ling for en pasient med minst tre positive noder. Den optimale kategoriseringsme-
 tode 300 demonstrerer imidlertid at den optimale grense 806, basert på inndata-
 20 ene 801, bør være ved den høyeste verdi på 5,5 lymfenoder, som svarer til en kli-
 nisk regel som foreskriver oppfølgingsbehandlinger av pasienter med minst seks
 positive lymfenoder.

Som vist i sammenligningstabellen 808, resulterte det ifølge teknikkens
 stand aksepterte kliniske grensepunkt ($\geq 3,0$) i 47% korrekt klassifiserte tilbakefall
 25 og 71% korrekt klassifiserte ikke-tilbakefall. 53% av tilbakefallene var følgelig urik-
 tig klassifisert (ytterligere behandling ble anbefalt på uriktig grunnlag) og 29% av
 ikke-tilbakefallene ble feilaktig klassifisert (ytterligere behandling ble uriktig anbe-
 falt). Det grensepunkt som bestemmes ved hjelp av den optimale kategoriserings-
 metode 300 ($\geq 5,5$) resulterte derimot i 33% korrekt klassifiserte tilbakefall og 97%
 30 korrekt klassifiserte ikke-tilbakefall. 67% av tilbakefallene ble følgelig uriktig klassi-
 fisert (ytterligere behandling ble feilaktig ikke anbefalt) og 3% av ikke-tilbakefall-
 ene ble uriktig klassifisert (ytterligere behandling ble uriktig anbefalt).

Som vist i dette eksempelet kan det være rimelig å oppnå en høyere grad av korrekt identifisering av de pasienter som kan unngå de etterkirurgiske kreftbehandlingsregimer, ved å benytte den optimale kategoriseringsmetode 300. Selv om grensepunktet som er bestemt ved hjelp av den optimale kategoriseringsmetode 300, gjelder en moderat høyere andel av ukorrekt klassifiserte tilbakefall, gir den en betydelig lavere prosentandel av ukorrekt klassifiserte ikke-tilbakefall. Ved å ta i betraktning kompromisset og innse at målet med optimaliseringsproblemet var å unngå unødvendig behandling, er således resultatene av det grensepunkt som er bestemt ved hjelp av den optimale kategoriseringsmetode 300, matematisk overlegne de ved det tidligere kjente kliniske grensepunkt. Denne type informasjon er potensielt uhyre nyttig for å tilveiebringe ytterligere innsikt til pasienter som avveier valget mellom å gjennomgå behandlinger slik som kjemoterapi, eller å risikere et tilbakefall av brystkreft.

Fig. 9 er en sammenligning mellom et eksempel på en etterbehandlet utgang fra en første støttevektormaskin som omfatter en lineær kjerne, og en annen støttevektormaskin som omfatter en polynomisk kjerne. Fig. 9 demonstrerer at en variasjon i valget av en kjerne kan påvirke kvalitetsnivået til utgangen fra en SVM. Som vist indikerer den etterbehandlede utgang fra en første SVM 902 som omfatter en lineær punktproduktkjerne, at for et gitt testsett med 24 prøver, pr. seks av åtte positive prøver ble uriktig identifisert og tre av seksten negative prøver ble uriktig identifisert. Til sammenligning indikerer den etterbehandlede utgang for en annen SVM 904 som omfatter en polynomisk kjerne, at for det samme testsett ble bare to av åtte positive prøver uriktig identifisert, og fire av seksten negative prøver ble identifisert. Til sammenligning ga den polynomiske kjerne betydelig forbedrede resultater vedrørende identifiseringen av positive prøver og ga bare litt dårligere resultater vedrørende identifiseringen av negative prøver. Som fagkyndige på området vil forstå, er derfor den globale minimumsfeil for den polynomiske kjerne lavere enn den globale minimumsfeil for den lineære kjerne for dette datasettet.

Fig. 10 og den følgende beskrivelse er ment å gi en kort og generell beskrivelse av et egnet beregningsmiljø for å implementere foreliggende oppfinnelse. Selv om det systemet som er vist på fig. 10, er en konvensjonell personlig datamaskin 1000, vil fagfolk på området forstå at oppfinnelsen også kan implementeres ved å bruke andre typer datasystem-konfigurasjoner. Datamaskinen 1000 inn-

befatter en sentralenhet 1022, et systemlager 1020 og en inn/ut-buss (I/O-buss) 1026. En systembuss 1021 kopler sentralenheten 1022 til systemlageret 1020. En busstyreenhet 1023 styrer flyten av data på I/O-bussen 1026 og mellom sentralenheten 1022 og en rekke interne og eksterne I/O-anordninger. I/O-anordningene som er koplet til I/O-bussen 1026, kan ha direkte tilgang til systemlageret 1020 ved å bruke en direkte lageraksess-styreenhet ("DMA"-enhet) 1024.

I/O-anordningene er koplet til I/O-bussen 1026 via et sett med anordningsgrensesnitt. Anordningsgrensesnittene kan innbefatte både maskinvarekomponenter og programvarekomponenter. For eksempel kan et hardplatedrev 1030 og et diskett-drev 1032 for lesing og skriving på fjernbare medier 1050, være tilkople
 10 I/O-bussen 1026 gjennom platedrev-styreenheter 1040. Et optisk platedrev 1034 for å lese eller skrive på optiske medier 1052 kan være tilkople
 I/O-bussen 1026 ved å bruke et lite datasystem-grensesnitt ("SCSI", Small Computer System Interface) 1041. Alternativt kan et IDE- (ATAPI) eller EIDE-grensesnitt være tilknyttet et
 15 optisk drev slik som tilfellet er med et CD-ROM-drev. Drevene og deres tilknyttede datamaskinlesbare medier gir ikke-flyktig lagring for datamaskinen 1000. I tillegg til det datamaskinlesbare medium som er beskrevet ovenfor, kan andre typer datamaskinlesbare medier også benyttes, slik som ZIP-drev eller lignende.

En fremvisningsanordning 1053, slik som en monitor, er koplet til I/O-bus-
 20 sen 1026 via et annet grensesnitt, slik som en videoadapter 1042. Et parallelt grensesnitt 1043 tilkople synkrone periferianordninger, slik som en laserskriver 1056, til I/O-bussen 1026. Et seriegrensesnitt 1044 forbinder kommunikasjonsanordninger med I/O-bussen 1026. En bruker kan innføre kommandoer og informasjon i datamaskinen 1000 via seriegrensesnittet 1044 eller ved å bruke en innmat-
 25 ingsanordning, slik som et tastatur 1038, en mus 1036 eller et modem 1057. Andre periferianordninger (ikke vist) kan også være tilkople
 datamaskinen 1000, slik som inn/ut-audioanordninger eller bilderegistreringsanordninger.

Et antall programmoduler kan være lagret på drevene og i systemlageret 1020. Systemlageret 1020 kan innbefatte både direkte- og lese-lageret (henholdsvis RAM og ROM). Programmodulene styrer hvordan datamaskinen 1000 funksjo-
 30 nerer og vekselvirker med brukeren, med I/O-anordningene eller med andre datamaskiner. Programmodulene innbefatter rutiner, operativsystemer 1065, brukerprogrammer, datastrukturer og andre programvare- eller fastvare-komponenter. I

en illustrerende utførelsesform kan foreliggende oppfinnelse omfatte én eller flere programmoduler 1075A for forbehandling, én eller flere programmoduler 1075B for etterbehandling og/eller én eller flere programmoduler 1077 for optimal kategorisering og én eller flere SVM-programmoduler 1070 lagret på drevene eller i systemlageret 1020 i datamaskinen 1000. Programmodulene 1075A for forbehandling, programmodulene 1075B for etterbehandling, sammen med SVM-programmodulene 1070 kan spesielt omfatte datamaskinutførbare instruksjoner for forbehandling av data og etterbehandling av utgangen fra en læremaskin, og for å implementere lærealgoritmen i henhold til de eksempler på fremgangsmåter som er beskrevet under henvisning til fig. 1 og 2. Programmoduler 1077 for optimal kategorisering kan videre omfatte datamaskinutførbare instruksjoner for optimal kategorisering av et datasett i henhold til de fremgangsmåter som er beskrevet under henvisning til fig. 3.

Datamaskinen 1000 kan operere i et nettverksmiljø ved å bruke logiske forbindelser til én eller flere fjerntliggende datamaskiner, slik som den fjerntliggende datamaskin 1060. Den fjerntliggende datamaskin 1060 kan være en tjener, en ruter, et nettverk med likeverdige stasjoner eller en annen vanlig nettnode, og innbefatter vanligvis mange av eller alle de elementer som er beskrevet i forbindelse med datamaskinen 1000. I et nettmiljø kan programmodulene og dataene være lagret på den fjerntliggende datamaskin 1060. De logiske forbindelser som er skissert på fig. 10, innbefatter et lokalt nett ("LAN") 1054 og et områdenett ("WAN") 1055. I et LAN-miljø kan et nettgrensesnitt 1045, slik som et eternett-adapterkort brukes til å forbinde datamaskinen 1000 med den fjerntliggende datamaskin 1060. I et WAN-miljø kan datamaskinen 1000 bruke en telekommunikasjonsanordning, slik som et modem 1057, til å opprette en forbindelse. Man vil forstå at nettforbindelsene som er vist, bare er illustrerende og at andre anordninger for å opprette en kommunikasjonsforbindelse mellom datamaskinene kan benyttes.

Fig. 11 er et funksjonsblokkskjema som illustrerer et alternativt eksempel på driftsmiljøet for implementering av foreliggende oppfinnelse. Foreliggende oppfinnelse kan implementeres i en spesialisert konfigurasjon med mange datasystemer. Et eksempel på en spesialisert konfigurasjon av mange datasystemer er her referert til som BLOWulfTM støttevektorprosessor (BSVP). BSVP'en kombinerer de siste fremskritt i maskinvareteknologi for parallell beregning med de siste matema-

tiske fremskritt i mønstergjenkjennelse, regresjonsestimering og densitetsestimering. Selv om kombinasjonene av disse teknologiene er en unik og ny implementering, er maskinvarekonfigurasjonen basert på Beowulf-superdatamaskinimplementeringer som først ble tatt i bruk av NASA Goddard Space Flight Center.

5 BSVP'en tilveiebringer den massive, parallelle beregningskraft som er nødvendig for å gjennomføre SVM-trening og evaluering på storskala-datasett. BSVP'en innbefatter en dobbeltparallel maskinvarearkitektur og kundetilpasset, parallelisert programvare for å muliggjør effektiv utnyttelse av både flerbruk og meldingsoverføring for effektivt å identifisere støttevektorer i praktiske anvendelser. 10 Optimalisering av både maskinvare og programvare gjør det mulig for BSVP å overgå vanlige SVM-implementeringer i betydelig grad. Etter hvert som varebehandlingsteknologi utvikles, blir videre oppgraderbarheten av BSVP sikret ved dens grunnlag i åpen kildeprogramvare og standardisert grensesnitteknologi. Fremtidige beregningsplattformer og netteknologi kan assimileres i BSVP'en når 15 de blir kostnadseffektive uten noen virkning på programvareimplementeringen.

Som vist i fig. 11, omfatter en Beowulf-klasse superberegningssklynge med tjue behandlingsmetoder 1104a-t og én vertsnode 1112. Behandlingsnodene 1104a-j er sammenkoplet via en sentral 1102a, mens behandlingsnodene 1104k-t er sammenkoplet via en sentral 1102b. Verstnoden 1112 er koplet til hver av nettsentralene 1102a eller 1102b (1102a er vist) via en passende eternett-kabel 1114. 20 Sentralen 1102a og sentralen 1102b er også koplet til hverandre via en passende eternett-kabel 1114 slik at alle tjue behandlingsnodene 1104a-t og vertsnoden 1112 effektivt er i kommunikasjon med hverandre. Sentralene 1102a og 1102b omfatter fortrinnsvis hurtige eternett-forbindelser. Den dobbeltparallelle arkitekturen til BSVP er utført ved implementering av Beowulf-superdatamaskinens parallelle konfigurasjon for meldingsoverføring med flere maskiner og under utnyttelse 25 av en dobbeltprossessor-datamaskin SMP med høy ytelse som vertsnoden 1112.

I dette eksempelet inneholder vertsnoden 1112 sømløs multiprosessorteknologi (SMP-teknologi) og består av en dobbelt 450MHz pentium II Xeon-basert maskin med 18GB med ultra SCSI-lager, 256MB hurtiglager, to 100Mbit/s NIC'er 30 og en 24GB DAT-reservebåndanordning for nettet. Vertskoden 1112 utfører NIS,

MPL og/eller PMV under Linux for å administrere aktiviteten til BSVP'en. Vertsnoden 1112 danner også portalen mellom BSVP og verdenen utenfor. Det interne nett i BSVP er således i og for seg isolert fra utenforliggende vekselvirkning, noe som gjør det mulig for hele klyngen å synes å funksjonere som en enkelt maskin.

5 De tjue behandlingsnodene 1104a-t er identisk konfigurerte datamaskiner som inneholder 150MHz pentium-prosessorer, 32MB RAM, 850MB HDD, 1,44MB FDD og en hurtig eternett mb100Mb/s NIC. Behandlingsnodene 1104a-t er sammenkoplet med hverandre og vertsnoden gjennom NFS-forbindelser over TCP/Ip. I tillegg til BSVP-beregninger er behandlingsnodene konfigurert for å tilveiebringe
10 demonstrasjonsmuligheter gjennom en tilkoplet monitorbank med hver nodes tastatur og mus rutet til en enkelt tastaturanordning og en enkel musanordning gjennom KVM-svitsjene 1108a og 1108b.

Tilpasning og utvikling av programvare gjør det mulig å optimalisere aktiviteter på BSVP. Samtidighet i seksjoner av SVM-prosesser blir utnyttet på den
15 mest fordelaktige måte gjennom den hybridparallelisering som tilveiebringes ved hjelp av BSVP-maskinvaren. Programvaren implementerer understøttelse gjennom hele syklusen fra rådata til implementert løsning. En databasemotor tilveiebringer den lagringskapasitet og den fleksibilitet som er nødvendig for forbehandling av rådata. Kundetilpassede rutiner automatiserer forbehandlingen av dataene
20 før SVM-trening. Flere transformasjoner og datamanipulasjoner blir utført i databasemiljøet for å generere kandidater til treningsdata.

Den teoretiske toppbehandlingskapasiteten til BSVP'en er 3,90GFLOPS. Basert på de vurderinger som er utført av NASA Goddard Space Flight Center på deres Beowulf-klasse maskiner, er den virkelig forventede ytelse antatt å være
25 omkring 1,56GFLOPS. Den ytelse som oppnås ved å bruke beregningskraften for varekomponenter i denne klyngemaskinen av Beowulf-klassen er på linje med den for superdatamaskiner slik som Cray J932/8. Ytterligere Beowulf-testing ved akademiske institusjoner og forskningssentre indikerer at en ytelse i størrelsesorden av 18 ganger en enkelt prosessor generelt kan oppnås på en Beowulf-klynge med
30 tjue noder. Et optimaliseringsproblem som f.eks. krever 17 minutter og 45 sekunder tid på en enkelt pentiumprosessor-datamaskin ble løst på 59 sekunder på en

Beowulf med 20 noder. Den høye ytelsen til BSVP'en muliggjør derfor praktisk analyse av datasett som nå er antatt å være for vanskelig å håndtere med konvensjonelle datasystemer.

Den massive beregningskraften til BSVP'en gjør den spesielt nyttig for implementering av flere SVM'er i parallell for å løse reelle problemer som innebærer et uhyre stort antall inndata. Eksempler på brukbarheten av SVM'er generelt og BSVP'en i særdeleshet, omfatter: genetisk forskning, spesielt the Human Genome Project; evaluering av forvaltet omsorgseffektivitet; terapeutiske beslutninger og oppfølginger; farmasøytiske utviklingsteknikker; oppdagelse av molekulære strukturer; prognostiske evalueringer; medisinsk informatikk; deteksjon av fakturasvin-
 10 del; inventarkontroll; lagerevalueringer og prediksjoner; vareevalueringer og prediksjoner; og sannsynlighetsestimater i forbindelse med forsikring.

Fagkyndige på området bør forstå at den BSVP-arkitektur som er beskrevet ovenfor, kun er illustrerende og ikke ment å begrense formålet med foreliggende oppfinnelse. Valget av tjue behandlingsnoder ble f.eks. basert på den velkjente Beowulf-arkitektur. BSVP'en kan imidlertid alternativ implementeres ved å bruke flere eller færre enn tjue behandlingsnoder. De spesielle maskinvare- og programvare-komponenter som er sitert ovenfor, er videre kun ment som eksempler. Som nevnt er BSVP-utførelsesformen av foreliggende oppfinnelse konfigurert for å
 20 være kompatibel med alternative og/eller fremtidige maskinvare- og programvare-komponenter.

Fig. 12 er et funksjonsblokkskjema som illustrerer et eksempel på et nettdriftsmiljø for implementering av en ytterligere alternativ utførelsesform av foreliggende oppfinnelse. I eksempelet på et nettdriftsmiljø kan en kunde 1202 eller en annen entitet overføre data via et fordelt datanett, slik som internett 1204, til en selger 1212. Fagkyndige på området vil forstå at kunden 1202 kan overføre data fra en datamaskin eller et laboratorieinstrument av enhver type som innbefatter eller er i kommunikasjon med en kommunikasjonsanordning og en datalagringsanordning. De data som overføres fra kunden 1202 kan være treningsdata, test-
 25 data og/eller virkelige data som skal behandles av en læremaskin. De data som overføres av kunden, blir mottatt på selgerens nettsjener 1206, som kan sende dataene til én eller flere læremaskiner via et internt nett 1214a-b. Som beskrevet tidligere kan læremaskiner omfatte SVM'er, BSVP'er 1100, neurale nettverk,

andre læremaskiner eller kombinasjoner av disse. Nettjeneren 1206 er fortrinnsvis isolert fra læremaskinen eller læremaskinene ved hjelp av en brannvegg 1208 eller et annet sikkerhetssystem. Selgeren 1212 kan også være i kommunikasjon med én eller flere finansinstitusjoner 1210, via internett 1204 eller en annen utpekt kommunikasjonsforbindelse eller en kommunikasjonsforbindelse som opprettes etter behov. Nettjeneren 1206 eller en annen kommunikasjonsanordning kan håndtere kommunikasjoner med én eller flere av finansinstitusjonene. Finansinstitusjonen eller -institusjonene kan omfatte banker, internettbanker, likvidasjonskontorer, kredit- eller debet-kortselskaper eller lignende.

Under drift kan selgeren tilby læremaskinbehandlingstjenester via et nettsted som befinner seg på nettjeneren 1206 eller en annen tjener i kommunikasjon med nettjeneren 1206. En kunde 1202 kan sende data til nettjeneren 1206 som skal behandles av en læremaskin. Kunden 1202 kan også overføre identifikasjonsinformasjon, slik som et brukernavn, et passord og/eller en finanskonto-identifiserer, til nettjeneren. Som svar på mottakelse av dataene og identifikasjonsinformasjonen, kan nettjeneren 1206 elektronisk trekke en forutbestemt mengde med verdier fra en finanskonto som holdes eller autoriseres av kunden 1202 i en finansinstitusjon 1210. I tillegg kan nettjeneren overføre kundens data til BSVP'en 1100 eller en annen læremaskin. Når BSVP'en 1100 har fullført behandlingen av data og etterbehandlingen av utgangen, blir den etterbehandlede utgang returnert til nettjeneren 1206. Som tidligere beskrevet kan utgangen fra en læremaskin bli etterbehandlet for å generere en beregningsmessig utledet alfanumerisk klassifiserer med en enkelt verdi eller flere verdier for tolkning av mennesker eller automatiske tolkninger. Nettjeneren 1206 kan så sikre at betaling fra kunden er blitt utført før den etterbehandlede utgang blir sendt tilbake til kunden 1202 via internettet 1204.

SVM'er kan brukes til å løse en lang rekke problemer i det virkelige liv. SVM'er kan f.eks. anvendes til å analysere konto- og inventar-data, lager- og varemarkedsdata, forsikringsdata, medisinske data, og så videre. Det ovenfor beskrevne nettmiljø har som sådant bred anvendelse over mange industrielle områder og markedssegmenter. I forbindelse med inventardataanalyse kan en kunde f.eks. være en selger. Selgeren kan levere inventar- og revisjons-data til nettjeneren 1206 til forutbestemte tider. Oppgjørs- og revisjons-dataene kan behandles av

BSVP'en og/eller én eller flere andre læremaskiner for å evaluere oppgjørskravene til selgeren. I forbindelse med medisinsk dataanalyse kan likeledes kundene være et medisinsk laboratorium og kan sende ekte data innsamlet fra en pasient til nettjeneren 1206 mens pasienten er tilstede i det medisinske laboratoriet. Det resultat som genereres ved behandling av de medisinske data med BSVP'en eller andre læremaskiner, kan sendes tilbake til det medisinske laboratorium og presenteres for pasienten.

I en annen utførelsesform er det i foreliggende oppfinnelse tenkt at et antall støttevektormaskiner kan være konfigurert for hierarkisk å behandle flere datasett i parallell eller i sekvens. Spesielt kan én eller flere førstenivå-støttevektormaskiner trenes og testes for å behandle en første type data og én eller flere førstenivå-støttevektormaskiner kan være trenet og testet for å behandle en annen-type data. Ytterligere datatyper kan også behandles av andre førstenivå-støttevektormaskiner. Utgangen fra noen eller alle førstenivå-støttevektormaskinene kan kombineres på en logisk måte for å frembringe et inngangsdatasett for én eller flere annennivå-støttevektormaskiner. På lignende måte kan utgangen fra et antall annennivå-støttevektormaskiner kombineres på logisk måte for å frembringe inngangsdata for én eller flere tredjenivå-støttevektormaskiner. Hierarkiet av støttevektormaskiner kan utvides til et hvilket som helst antall nivåer som kan tenkes. På denne måten kan støttevektormaskiner ved et lavere hierarkisk nivå brukes til å forbehandle data som skal mates inn i det høyeste hierarkiske nivå av støttevektormaskiner. Også støttevektormaskiner i det høyeste hierarkiske nivå kan benyttes til å etterbehandle data som blir matet ut fra støttevektormaskiner ved det laveste hierarkiske nivå.

Hver støttevektormaskin i hierarkiet eller hvert hierarkisk nivå av støttevektormaskiner kan være konfigurert med en distinkt kjerne. For eksempel kan støttevektormaskiner som brukes til å behandle data av en første type, konfigureres med en første type kjerne, mens støttevektormaskiner som brukes til å behandle data av en annen type, kan være konfigurert med en annen type kjerne. I tillegg kan flere støttevektormaskiner i det samme eller andre hierarkiske nivåer være konfigurert for å behandle den samme type data ved å bruke distinkte kjerner.

Fig. 13 er gitt som et eksempel for å illustrere et hierarkisk system av støttevektormaskiner. Som vist kan én eller flere første nivå-støttevektormaskiner 1302A1 og 1302A2 trenes og testes for å behandle inngangsdata 1304A av en første type, slik som mammografidata, vedrørende et utvalg av medisinske pasienter. Én eller flere av disse støttevektormaskinene kan omfatte en distinkt kjerne (vist som kjerne 1 og kjerne 2). Én eller flere ytterligere første nivå-støttevektormaskiner 1302B1 og 1302B2 kan også være trenet og testet for å behandle data 1304B av en annen type; slik som genomiske data, for det samme eller et annet utvalg av medisinske pasienter. Igjen kan én eller flere av de ytterligere støttevektormaskiner omfatte en distinkt kjerne (vist som kjerne 1) og kjerne (3). Utgangen fra hver av de like første nivå-støttevektormaskiner kan sammenlignes med hverandre (dvs. utgang A1 1306A sammenlignet med utgang A2 1306B; utgang B1 1306C sammenlignet med utgang B2 1306D) for å bestemme optimale utmatninger (1308a og 1308B). De optimale utganger fra de to typene med første nivå-støttevektormaskiner 1308A og 1308B kan så kombineres for å danne et nytt multidimensjonalt inndatasett 1310, f.eks. vedrørende mammografi og genomiske data. Det nye datasett kan så behandles av én eller flere riktig trenede og testede annen nivå-støttevektormaskiner 1312A og 1312B. De resulterende utganger 1314A og 1314B fra annen nivå-støttevektormaskinene 1312A og 1312B kan sammenlignes for å bestemme en optimal utgang 1316. Den optimale utgang 1316 kan identifisere kausale forhold mellom mammografi- og genomiske data-punkter. Som det vil være klart for fagkyndige på området, kan det tenkte hierarki av støttevektormaskiner anvendes på et hvilket som helst område eller i en hvilken som helst industri hvor analyse av data ved hjelp av en læremaskin er ønsket.

Den hierarkiske behandling av flere datasett ved bruk av flere støttevektormaskiner kan brukes som en fremgangsmåte til forbehandling eller etterbehandling av data som skal mates inn i eller mates ut fra ytterligere andre støttevektormaskiner eller læremaskiner. I tillegg kan forbehandling eller etterbehandling av data utføres på inndataene og/eller utgangen fra den ovenfor beskrevne hierarkiske arkitektur av støttevektormaskiner.

Alternative utførelsesformer av foreliggende oppfinnelse vil lett kunne finnes av fagkyndige på det område som foreliggende oppfinnelse vedrører. Slike alternative utførelsesformer betraktes å være innbefattet i rammen for den foreliggende oppfinnelse. Rammen for den foreliggende oppfinnelse er følgelig beskrevet i de vedføyde patentkrav og er understøttet av den foregående beskrivelse.

PATENTKRAV

1. Fremgangsmåte for forbedring av kunnskapsopptagelse ved bruk av flere støttevektormaskiner,

5 k a r a k t e r i s e r t v e d :

å forbehandle et første treningsdatasett og et annet treningsdatasett for å tilføye dimensjonalitet til hvert av et antall treningsdatapunkter;

10 å trene én eller flere første støttevektormaskiner ved å bruke det første forbehandlede treningsdatasett, idet hver av de første støttevektormaskiner omfatter forskjellige kjerner;

å trene én eller flere andre støttevektormaskiner ved å bruke det annet forbehandlede treningsdatasett, idet hver av de andre støttevektormaskiner omfatter andre kjerner;

15 å forbehandle et første testdatasett på samme måte som det første treningsdatasett, og å forbehandle et annet testdatasett på samme måte som det annet treningsdatasett;

å teste hver av de første trenede støttevektormaskiner ved å bruke det første forbehandlede testdatasett, og å teste hver av de andre trenede støttevektormaskiner ved å bruke det annet forbehandlede testdatasett;

20 å sammenligne, som reaksjon på mottakelse av en første testutgang fra hver av de første trenede støttevektormaskiner, hver av de første testutganger med hverandre for å bestemme hvilken, om noen, av de første testutganger som er en første optimal løsning;

25 å sammenligne, som reaksjon på mottakelse av en annen testutgang fra hver av de andre trenede støttevektormaskiner, hver av de andre testutganger med hverandre for å bestemme hvilken, om noen, av de andre testutganger som er en annen optimal løsning; og

30 å kombinere den første optimale løsning med den annen optimale løsning for å skape et nytt inndatasett som skal mates inn i én eller flere ytterligere støttevektormaskiner.

2. Datamaskinlesbart medium på hvilket det er lagret datamaskinutførbare instruksjoner for å utføre fremgangsmåten ifølge krav 1.

3. Fremgangsmåte ifølge krav 1,

karakterisert ved at forbehandling av det første treningsdatasett og det annet treningsdatasett videre omfatter:

5 å bestemme at minst ett av treningsdatapunktene er "urent"; og

 å rense, som reaksjon på bestemmelse av at treningsdatapunktet er urent, det urene treningsdatapunkt.

4. Fremgangsmåte ifølge krav 3,

10 karakterisert ved at rensingen av det urene treningsdatapunkt omfatter å slette, reparere eller gjeninnsette datapunktet.

5. Fremgangsmåte ifølge krav 1,

15 karakterisert ved at hvert treningsdatapunkt omfatter en vektor som har én eller flere opprinnelige koordinater; og

 hvor forbehandling av treningsdatasettet omfatter å tilføye én eller flere nye koordinater til vektoren.

6. Datamaskinlesbart medium hvorpå det er lagret datamaskinlesbare instruksjoner for å utføre fremgangsmåten ifølge krav 5.

20

7. Fremgangsmåte ifølge krav 5,

karakterisert ved at én eller flere nye koordinater som tilføyes vektoren, blir utledet ved å anvende en transformasjon på én eller flere av de opprinnelige koordinater.

25

8. Fremgangsmåte ifølge krav 7,

karakterisert ved at transformasjonen er basert på ekspertkunnskap.

30 9. Fremgangsmåte ifølge krav 7,

karakterisert ved transformasjonen er beregningsmessig utledet.

10. Fremgangsmåte ifølge krav 7,
karakterisert ved at treningsdatasettet omfatter en kontinuerlig variabel;
og

at transformasjonen omfatter optimal kategorisering av den kontinuerlige
5 variable i treningsdatasettet.

11. Datamaskinlesbart medium hvorpå det er lagret datamaskinutførbare in-
struksjoner for å utføre fremgangsmåten ifølge krav 10.

10 12. Fremgangsmåte ifølge krav 1,
karakterisert ved at sammenligningen av hver av de første testutganger
med hverandre og sammenligningen av hver av de andre testutganger med hver-
andre omfatter:

å etterbehandle hver av testutgangene ved å tolke hver av testutgangene i
15 et felles format;

å sammenligne hver av de første etterbehandlede testutganger med hver-
andre for å bestemme hvilken av de første testutganger som representerer en
første laveste global minimumsfeil; og

å sammenligne hver av de andre etterbehandlede testutganger med hver-
20 andre for å bestemme hvilken av de andre testutganger som representerer en an-
nen, laveste global minimumsfeil.

13. Datamaskinlesbart medium hvorpå det er lagret datamaskinutførbare in-
struksjoner for å gjennomføre fremgangsmåten ifølge krav 12.

25

14. Fremgangsmåte ifølge krav 1,
karakterisert ved at den kunnskap som skal oppdages fra dataene,
vedrører en regresjons- eller densitets-estimering;

hvor hver støttevektormaskin frembringer en treningsutgang som omfatter
30 en kontinuerlig variabel; og

hvor fremgangsmåten videre omfatter det trinn å etterbehandle hver av
treningsutgangene ved optimal kategorisering av treningsutgangen for å utlede
grensepunkter i den kontinuerlige variable.

15. Fremgangsmåte ifølge krav 1,
karakterisert ved:

5 å bestemme, som reaksjon på sammenligning av hver av testutgangene
med hverandre, at ingen av testutgangene er den optimale løsning;

å justere de forskjellige kjerner i én eller flere av antallet støttevektor-
maskiner; og

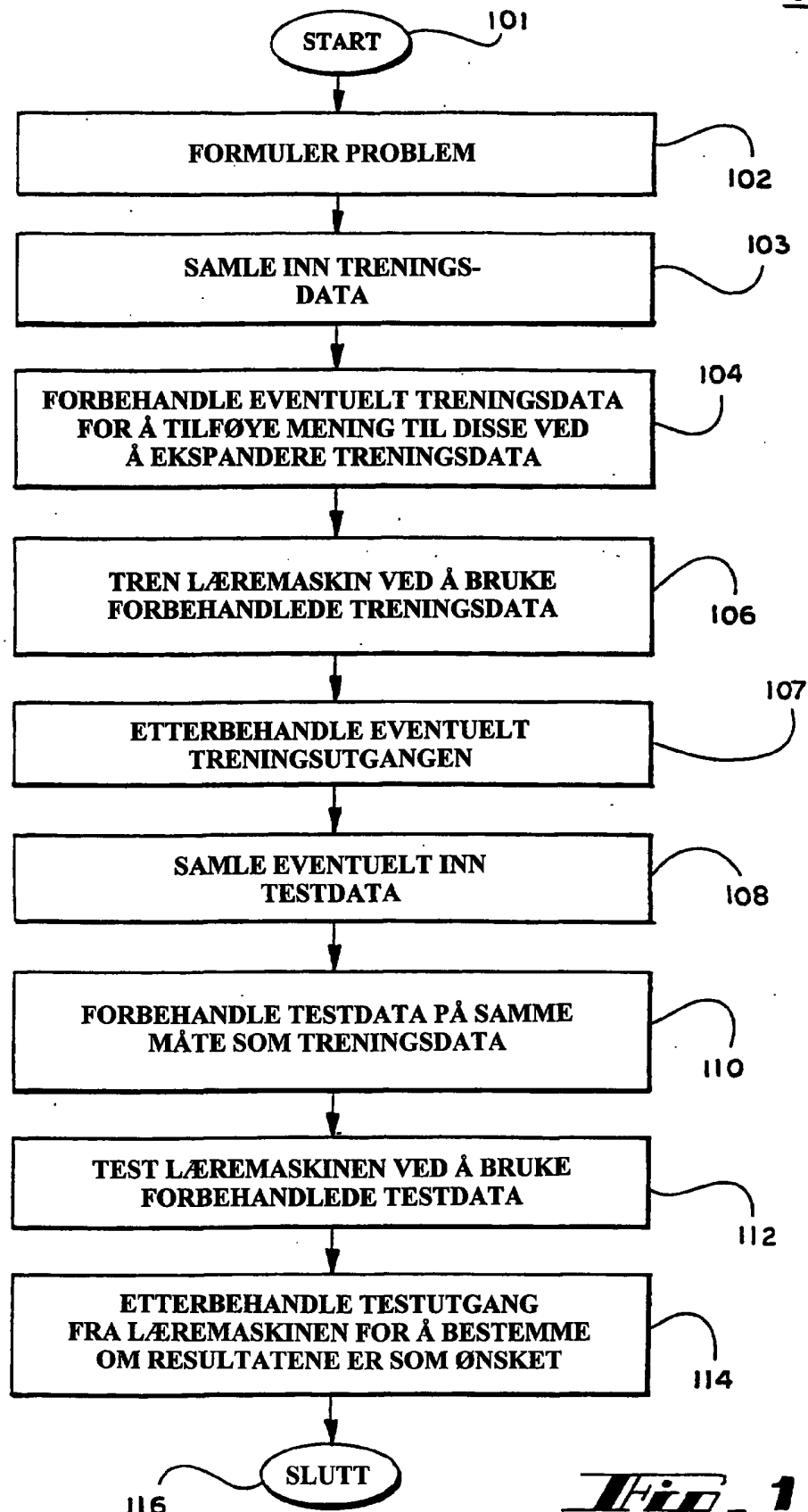
å trene og teste på nytt, som reaksjon på justering av valget av de andre
kjerner, antallet støttevektormaskiner.

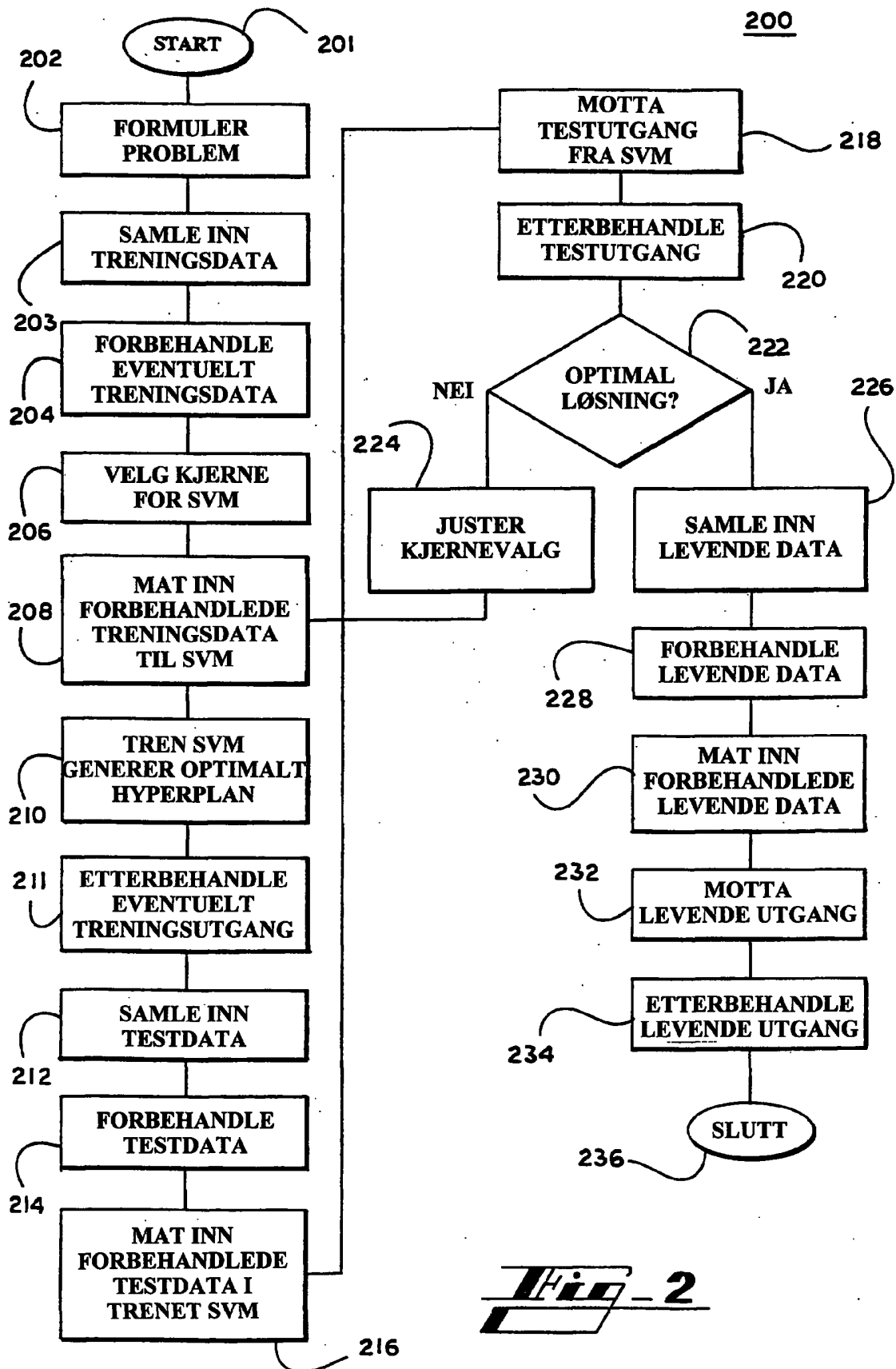
10

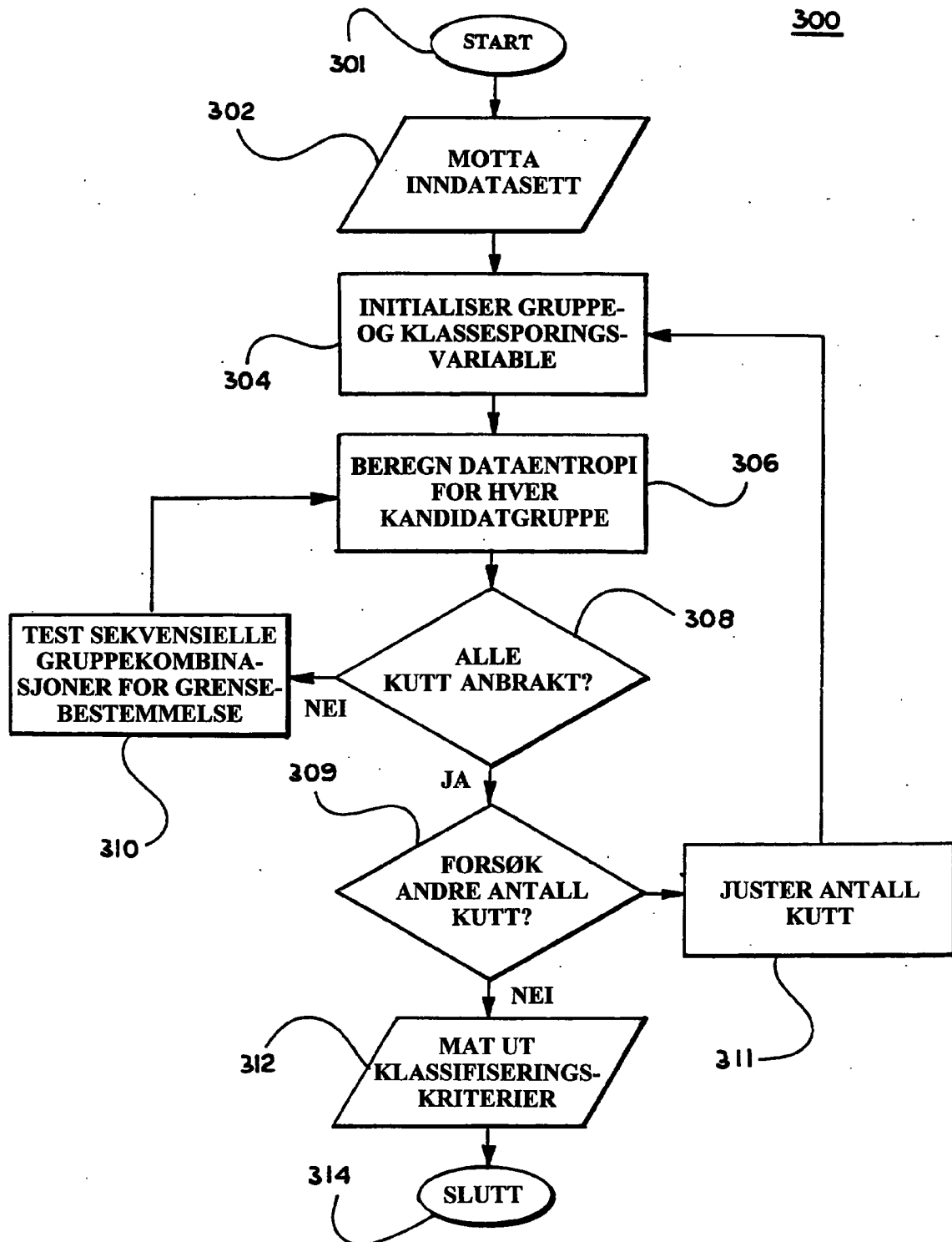
16. Datamaskinlesbart medium hvorpå det er lagret datamaskinutførbare in-
struksjoner for å gjennomføre fremgangsmåten ifølge krav 15.

17. Fremgangsmåte ifølge krav 15,

15 karakterisert ved at justering av de forskjellige kjerner blir utført basert
på tidligere ytelse eller historiske data og er avhengig av beskaffenheten av den
kunnskap som skal oppdages fra dataene eller dataenes beskaffenhets.







406a	406b	406c	406d	406e	406f
41	75	95	29	8	1
47	64	45	26	4	1
34	35	17	11	5	1
48	5	137	18	2	1
35	29	48	21	9	1
49	19	69	11	7	1
42	8	10	12	2	1
44	3	12	14	1	1
37	57	19	20	7	1
48	17	14	12	1	1
39	1	10	17	8	1
44	21	14	17	1	1
34	2	9	12	1	1
31	10	10	16	0	1
42	252	452	25	0	1
42	59	693	19	1	1
36	5	51	12	1	1
38	1	10	19	1	1
36	43	89	9	1	1
42	16	10	13	4	1
47	19	20	12	1	1
49	14	128	21	4	1
38	169	315	36	0	1
33	2	12	8	0	1
46	41	308	23	5	1
44	54	115	29	2	1
31	4	3	20	1	1
49	1	18	27	2	1
48	34	355	19	3	1
44	29	19	20	2	1
42	207	11	10	3	1
43	62	53	26	1	1
43	108	293	4	0	1
35	25	13	24	1	1
35	5	5	16	1	1
45	54	17	19	8	1
41	4	10	15	4	1
40	8	12	8	0	1
42	25	78	21	2	1
46	30	105	11	5	1
48	72	94	15	1	1
36	3	10	7	0	1
46	165	12	17	5	1
47	22	10	26	2	1
48	3	10	21	0	1
40	4	10	12	0	1
32	3	10	14	6	1
36	51	167	11	3	1
39	4	15	31	25	1
46	2	10	8	0	1
28	12	36	25	23	1
32	23	50	19	0	1
44	26	10	21	4	1
47	32	11	21	7	1
42	32	41	21	18	1
42	2	10	14	0	1
36	10	1	16	1	1
47	5	6	22	1	1
34	18	6	13	2	1
34	9	10	29	0	1
37	6	10	16	0	1
38	42	46	14	6	1
32	98	11	11	0	1
37	39	10	13	1	1
45	17	267	20	4	1
39	93	134	12	1	1
43	47	11	19	1	1
46	6	10	13	1	1
49	1	4	19	6	1
12	172	302	9	3	1

402

404

FIG.4

Vapniks polynom

Alfategn begrenset opp til 1000.

Innverdier vil bli skalert til å være mellom 0 og 1.

SV-nullterskel: 1e-16.

Marginterskel: 0,1

Objektiv nullterskel: 1e-07

Polynomgrad: 2.

5/13

Testsett:

Totalt antall prøver: 24

Positive prøver: 8

av hvilke er feil: 4

Negative prøver: 16

av hvilke er feil: 6

FIG.5

606a1 606a	606a3 606a2	606b1-3 606b	606c1-3 606c	6/13 606d1-3 606d	606e1-3 606e	606f
41	0	75	95	29	8	1
47	0	64	45	26	4	1
34	1	35	17	11	5	1
48	0	29	137	18	2	1
35	0	19	48	21	9	1
49	0	8	69	11	7	1
42	0	3	10	12	2	1
44	0	57	12	14	1	1
37	0	17	19	20	7	1
48	0	21	14	12	8	1
39	0	1	10	17	1	1
44	0	2	14	17	1	1
34	1	10	9	12	1	1
31	0	252	10	16	0	1
42	0	59	452	25	0	1
42	0	5	693	19	1	1
36	0	43	51	12	1	1
36	0	16	10	19	1	1
42	0	19	89	9	0	1
47	0	14	10	13	4	1
49	0	14	20	12	1	1
38	0	169	128	21	4	1
33	0	2	315	36	0	1
46	0	41	12	8	0	1
34	0	54	308	23	2	1
49	0	4	115	29	5	1
44	0	34	3	20	1	1
42	0	207	18	27	2	1
43	0	62	355	19	3	1
35	0	108	11	10	1	1
45	0	25	53	26	0	1
41	0	54	13	4	1	1
40	0	8	5	16	1	1
42	0	25	17	19	8	1
46	0	72	10	15	4	1
48	0	30	12	8	0	1
36	0	165	78	21	5	1
46	0	22	105	11	2	1
47	0	3	10	15	0	1
48	0	4	10	7	0	1
33	1	51	10	17	5	1
38	0	12	167	26	2	1
46	0	23	15	31	3	1
32	0	32	10	8	0	1
44	0	32	36	25	0	1
47	0	10	50	19	0	1
42	0	32	10	21	4	1
36	0	10	41	21	18	1
47	0	18	1	14	0	1
34	0	9	6	16	1	1
37	0	42	10	22	2	1
38	0	98	10	13	0	1
32	0	39	46	29	6	1
37	0	17	11	14	0	1
45	0	93	10	11	0	1
39	0	47	267	20	4	1
43	0	6	134	12	1	1
46	0	172	10	13	1	1
46	0	1	4	19	6	1

602

604

FIG. 6

702

Vapniks polynom

Alfategn begrenset opp til 1000.

Innverdier vil bli skalert til å være mellom 0 og 1.

SV-nullterskel: 1e-16.

Marginterskel: 0,1

Objektiv nullterskel: 1e-07

Polynomgrad: 2.

Testsett:

Totalt antall prøver: 24

Positive prøver:

8

av hvilke er feil:

4

Negative prøver:

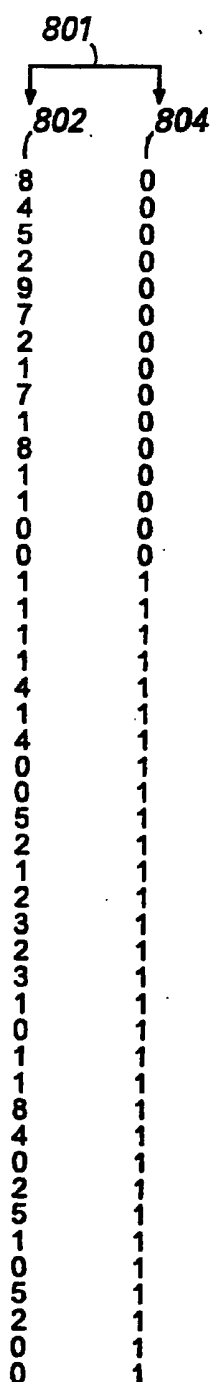
16

av hvilke er feil:

6

FIG.7

8/13



808

Kuttpunkt	Korrekt klassifiserte tilbakefall	Korrekt klassifiserte ikke-tilbakefall
Klinisk ($\geq 3,0$)	7 av 15 (47%)	22 av 31 (71%)
Optimalt ($\geq 5,5$)	5 av 15 (33%)	30 av 31 (97%)

Antall delintervaller: 2
 Antall klasser: 2
 Antall datapunkter: 46
 Nedre grense: -1
 Øvre grense: 10
 Antall grupper: 22

Reguleringskonstant: 1
 Datafil: = posnodes.prn
 Min entropi = -0,568342
 5,500000

806

FIG.8

902

Enkelt punktprodukt

Alfategn begrenset opp til 1000.

Innverdier vil ikke bli skalert.

SV-nullterskel: 1e-16.

Marginterskel: 0,1

Objektiv nullterskel: 1e-07

Testsett:

Totalt antall prøver: 24

Positive prøver: 8

av hvilke er feil: 6

Negative prøver: 16

av hvilke er feil: 3

904

Vapniks polynom

Alfategn begrenset opp til 1000.

Innverdier vil ikke bli skalert.

SV-nullterskel 1e-16.

Marginterskel: 0,1

Objektiv nullterskel: 1e-07

Polynomgrad: 2.

Testsett:

Totalt antall prøver: 24

Positive prøver: 8

av hvilke er feil: 2

Negative prøver: 16

av hvilke er feil: 4

9/13

FIG. 9

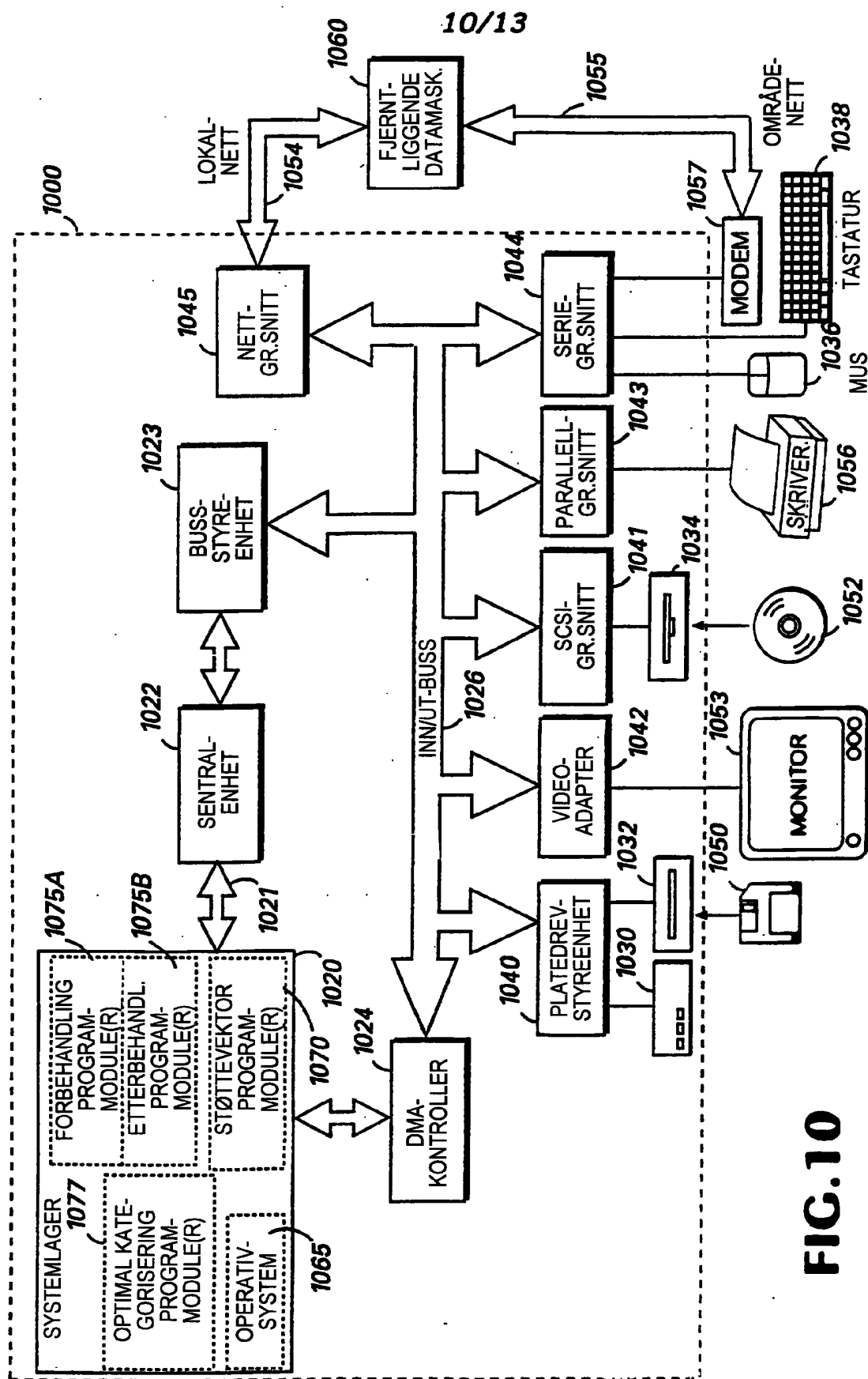
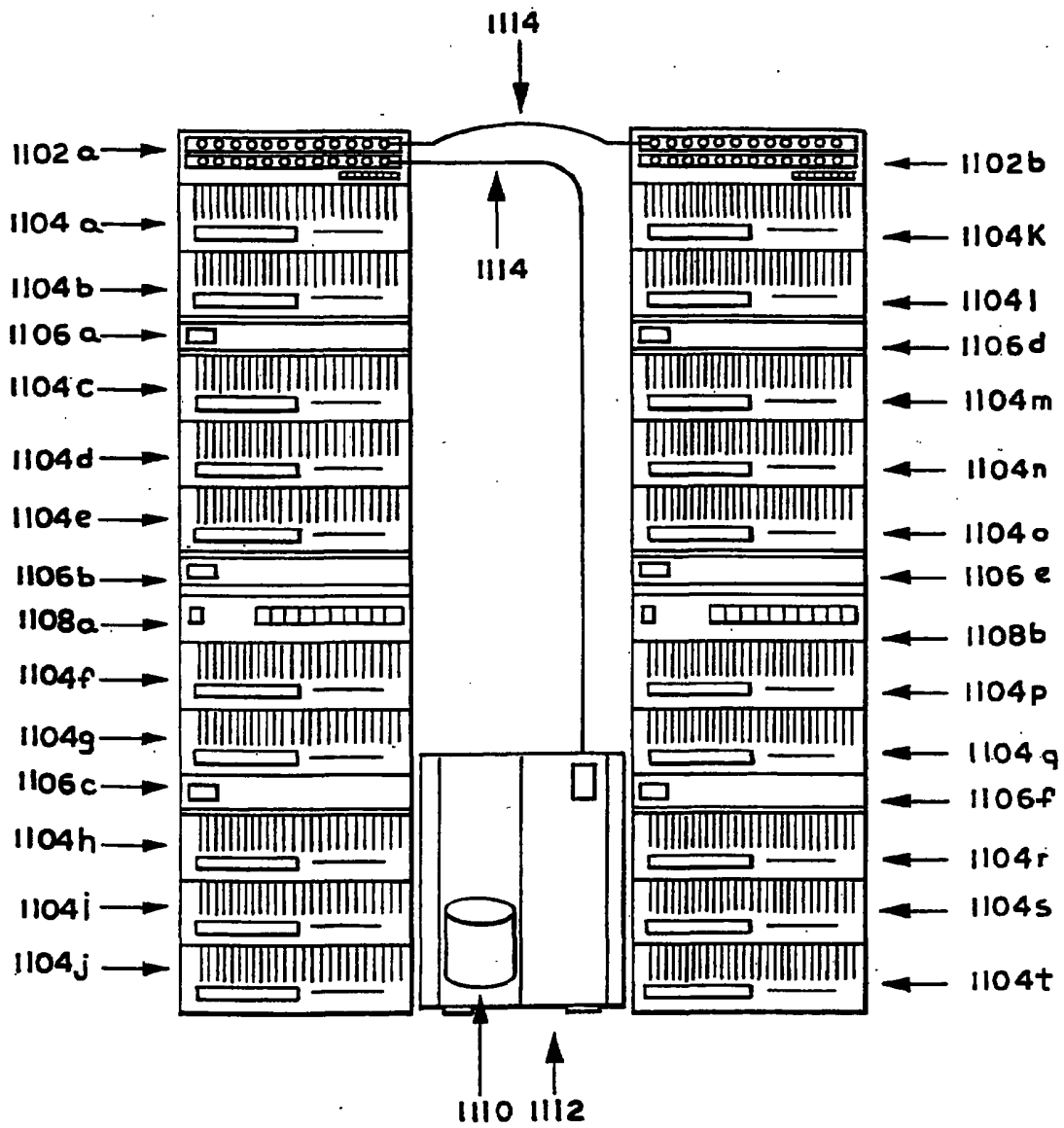


FIG.10

***Fig. 11***

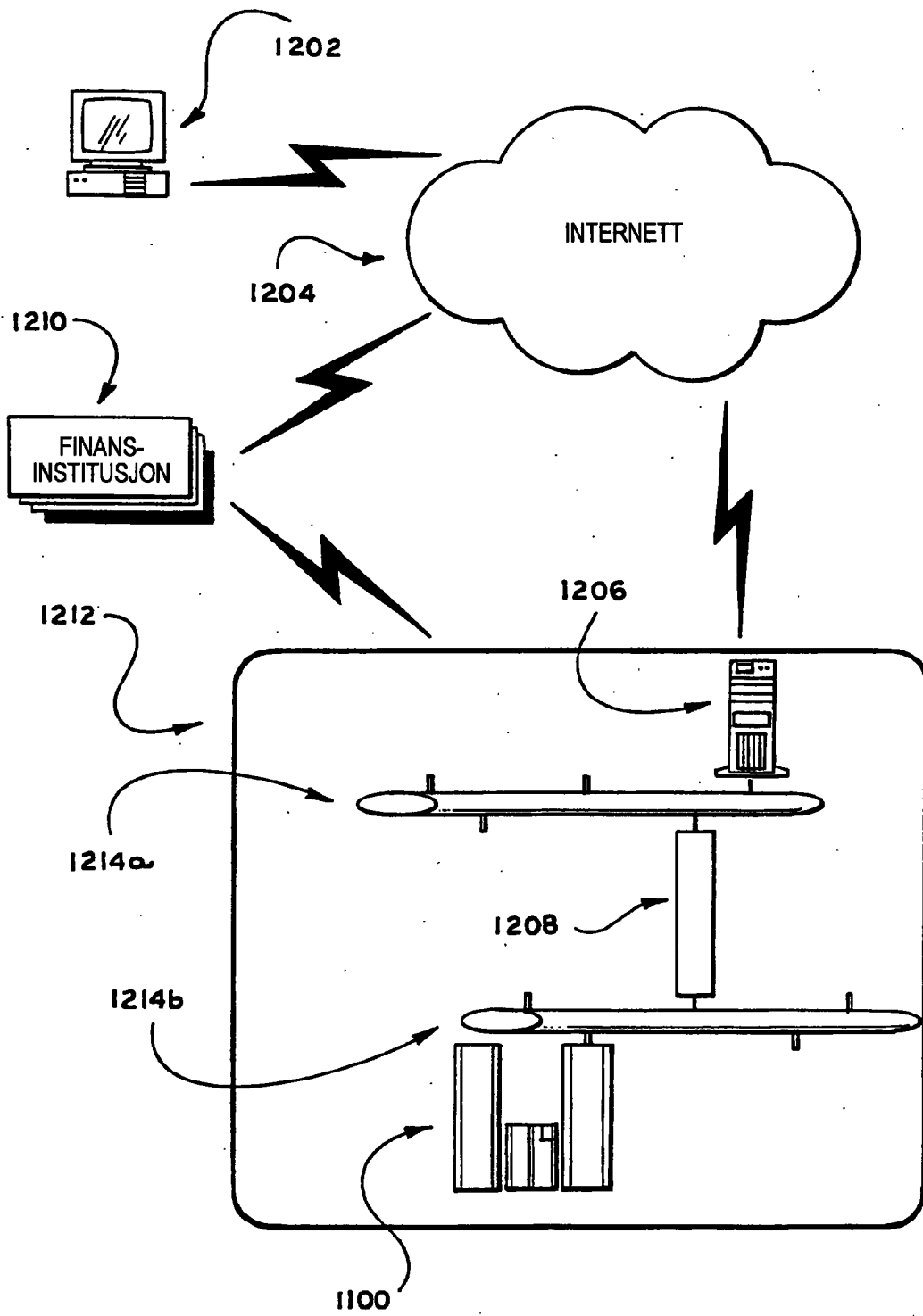


Fig. 12

13/13

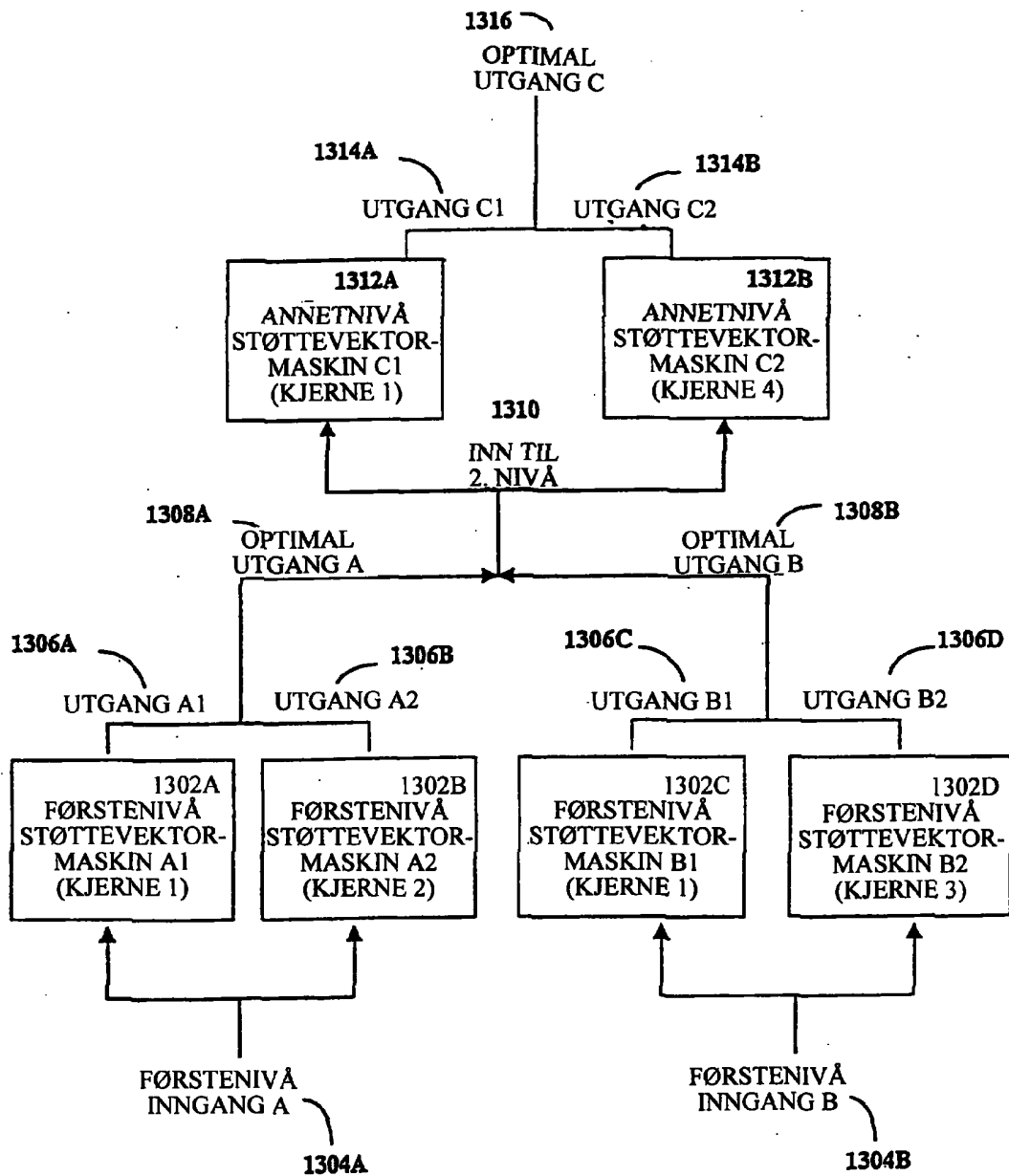


FIG. 13