Hur relaterar det optimala valet av klassificeringsmetod till datamaterialets egenskaper?

En jämförande studie mellan logistisk regression, elastic net och boosting tillämpat på klassificeringsträd.

Blaise Ngendangenzwa
Jonathan Sundin

Student
Vt 2015
Kandidatuppsats, 15 hp
Statistik C, 30 hp

Abstract

Title: *"How does the optimal choice of classifier relate to the data set characteristics? - a comparative study between logistic regression, elastic net and boosting applied to classification tree"*

Lately, research in the field of statistical classification have been subject of increasing amounts of criticism due to ambiguous conclusions and that still no method that systematically outperforms any other method. As a result the choice of classifier, in many cases, rests on individual preferences rather than scientific findings. According to the literature this fact is driven by a relationship between classification accuracy and characteristics of the data set.

This thesis intend to examine the connection between classification accuracy and characteristics of the data set. This intention is achieved by applying logistic regression, elastic net and boosting applied to classification tree on six real world data sets with varying characteristics.

The main result shows that classification accuracy varies with the data set characteristics. Elastic net is preferable when the number of explanatory variables is larger than the number of observations, boosting applied to classification tree is, in turn, the optimal choice when the data suffers from multicollinearity whereas logistic regression is useful under the circumstance that there is a large number of observations. Hence the general conclusion from this thesis is that, overall, neither method systematically outperforms any other method. Thus classification accuracy seems to be a function of the data set characteristics.

Sammanfattning

På sistone har allt mer kritik riktats mot forskning inom klassificering. Trots att forskningen har resulterat i en uppsjö av klassificeringsmetoder finns det de som menar att den har varit ett misslyckande och pekar på det faktum att ingen klassificeringsmetod anses vara systematiskt bättre än den andra eller ens rena gissningar. Detta leder till att valet av klassificeringsmetod i många fall vilar på individuella preferenser snarare än på vetenskaplig grund. Enligt litteraturen bottnar detta faktum i ett underliggande samband mellan det optimala valet av klassificeringsmetod och egenskaperna som karaktäriserar datamaterialet.

Uppsatsen tar avstamp från denna problematik och syftar till att undersöka kopplingen mellan det optimala valet av klassificeringsmetod och datamaterialets egenskaper. Denna intention uppnår vi genom att tillämpa logistisk regression, elastic net och boosting tillämpat på klassificeringsträd på sex verkliga datamaterial med varierande statistiska egenskaper.

Resultatet visar att den relativa klassificeringsförmågan varierar med datamaterialen. Elastic net är att föredra antalet förklaringsvariabler är större än antalet observationer, boosting tillämpat på klassificeringsträd är i sin tur det optimala valet när det förekommer multikolinjäritet medan logistisk regression är användbar under förutsättningen att datamängden är stor. Den generella slutsatsen från uppsatsen är således att det optimala valet beror på datamaterialet. Därmed bekräftar uppsatsen stor del av tidigare forskning inom ämnet.

Populärvetenskaplig sammanfattning

Klassificering är ett samlingsnamn på statistiska metoder vilka syftar till att bestämma grupptillhörighet hos en observation, huruvida något är ett eller noll, sant eller falsk och svart eller vit. Klassificering har etablerats som ett av de mest användbara statistiska tillvägagångssätten och tillämpas idag inom en rad vitt skilda områden; alltifrån att sätta preliminära diagnoser utifrån en serie symptom till att avgöra kreditstatus baserat på finansiell historik. Genom att möta klassificeringsproblemen från statistiskt perspektiv adderas rationalitet och konsekvens till beslutsfattandet som annars inte vore möjlig att uppnå. I samband med att de första klassificeringsmetoderna introducerades initierades en jakt på att utveckla allt bättre klassificeringsmetoder och som resultat finns idag ett flertal sätt att tackla ett klassificeringsproblem. Detta till trots finns det de som menar att forskningen i ämnet har varit ett misslyckande och pekar på det faktum att ingen klassificeringsmetod systematiskt är bättre än den andra eller ens rena gissningar. Detta får konsekvenser som manifesteras i att valet av klassificeringsmetod i många fall vilar på individuella preferenser snarare än på vetenskaplig grund. Problematiken anses vara relaterad till att den optimala klassificeringsmetoden i stor utsträckning beror på vilka egenskaper som karaktäriserar det aktuella datamaterialet. Uppsatsen tar avstamp i denna problematik och fördjupar sig i kopplingen mellan datamaterial och det optimala valet av klassificeringsmetod. Genom att tillämpa tre vanliga klassificeringsmetoder

på sex verkliga datamaterial, med varierande statistiska egenskaper, hoppas vi utröna under vilka förutsättningar respektive tillvägagångssätt är att föredra. Resultatet visar att den relativa klassificeringsförmågan helt och hållet beror på datamaterialet.

1 Inledning

Det mänskliga beslutsfattandet utgår i stor utsträckning från den medfödda förmågan att identifiera mönster. Förmågan att urskilja relevanta likheter, skillnader och relationer skapar förutsättningar för att välgrundade beslut sedermera ska fattas (Jain et al., 2000). Inom den statistiska analysen har man länge kapitaliserat på detta faktum genom att återskapa processer som känner igen, identifierar och urskiljer mönster i ett datamaterial. Detta har framförallt kommit till användning i situationer som faller inom ramen av klassificering. Klassificering definieras som ett samlingsnamn av statistiska tillvägagångssätt vilka syftar till att prediktera kategorisk tillhörighet hos ett element, givet information om elementet. I korthet handlar klassificering om att prediktera det okända och avgöra huruvida något är sant eller falskt, ett eller noll och svart eller vitt. Inom den statistiska sfären insåg man tidigt fördelarna med att möta klassificeringsproblem utifrån ett statistiskt perspektiv och därmed eliminera inkonsekvens och irrationalitet som ofta präglar det mänskliga beslutsfattandet.

Idag har klassificering etablerats som ett av de mest användbara statistiska tillvägagångssätten och tillämpas inom en rad vetenskapliga discipliner såväl som mänskliga aktiviteter. Tekniken används exempelvis för att avgöra kreditstatus baserat på finansiell historik, automatiserad postsortering eller för att sätta preliminära diagnoser på patienter utifrån uppvisade symptom (James et al., 2013). Faktum är att vi ständigt ställs inför situationer i vår vardag vilka kan summeras som ett klassificeringsproblem vare sig det är skivsamlingar som ska sorteras, ett vinnande drag i ett schackparti som ska planeras eller aktier som ska köpas eller säljas.

Klassificering betraktas generellt från två perspektiv; maskininlärning och statistisk klassifikation. Den stora skillnaden mellan perspektiven går att härleda till att maskininlärning syftar till att skapa algoritmer och datadrivna processer som lär sig att klassificera av tidigare data medan statistisk klassificering baseras på statistiska modeller och därmed genererar sannolikheter för klasstillhörighet

snarare än direkta klassificeringar (Michie et al., 1994). Målet inom klassificering är ofta tudelat; dels vill man skapa en modell som med största säkerhet predikterar grupptillhörighet och dels vill man skapa en modell som belyser underliggande strukturer och relationer i datamaterialet (Breiman et al., 1984). Klassificeringsmodellerna kan utvärderas utifrån ett flertal kriterium, allmänt anses däremot en bra modell generera klassificeringar på testdata som i stor utsträckning överensstämmer med den faktiska grupptillhörigheten (Demšar, 2006). I samband med att de första klassificeringsmetoderna introducerades initierades även en strävan efter att utveckla alltmer tillförlitliga klassificeringsmetoder. Den statistiska forskningen har därför länge riktat sitt strålkastarljus mot ämnet och som resultat finns idag en uppsjö av tillvägagångssätt tillhörande en rad olika klassificeringsfamiljer; parametriska såväl som icke-parametriska metoder, metoder som baseras på normalfördelningen och metoder som utgår från klassificeringsträd. Tre vanligt förekommande tillvägagångssätt är logistisk regression (LR), elastic net (EN) och boosting tillämpat på klassificeringsträd (BTK). Inom LR skattas sannolikheter för att en observation ska tillhöra en specifik klass. Prediktionen baseras sedan på en förutbestämd beslutsregel. EN är en vidareutveckling av LR. Metoden tillämpar ett snarlikt tillvägagångssätt som LR fast inför restriktioner på parameterskattningar. BTK är ett sekventiellt tillvägagångssätt som syftar till att skapa mindre klassificeringsträd som succesivt lär sig att klassificera observationer.

Trots en gedigen arsenal av klassificeringsmetoder betraktas forskning inom området, till viss mån, som ett misslyckande. Forskningsresultaten har hittills varit splittrade, metodernas relativa för - och nackdelar är oklara och än idag finns inte en klassificeringsmetod som systematiskt är bättre än den andra eller ens rena gissningar (Van Der Walt and Barnard, 2006). Detta faktum anses vara relaterat till att klassificeringsförmågan hos en specifik klassificeringsmetod är kopplad till externa förhållanden. Klassificeringsförmågan påverkas bl.a. av problemets utformning, antaganden om fördelningar och mängden träningsdata. Mer än något annat anses dock klassificeringsförmågan vara en funktion av egenskaperna som karaktäriserar det aktuella datamaterialet. Vikten av datamaterialet i den statistiska analysen har bl.a. uppmärksammats av Wolpert and Macready (1995) som menar att valet av optimal klassificeringsmetod helt och hållet beror på egenskaperna som karaktäriserar datamaterialet. De konkretiserar slutsatsen i ”no free lunch-theorem” som i korthet hävdar att över ett stort antal varierande datamaterial är samtliga klassificeringsmetoder i stort sett lik-

värdiga. Ämnet har även hamnat i blickfånget hos Macià et al. (2013) som menar att om denna aspekt inte tas nog i beaktning, vid exempelvis empirisk validering av klassificeringsmetoder, riskerar slutsatserna inte bara att vara icke-generaliserbara utan även felaktiga. En försvårande omständighet är dock att inget datamaterial är det andra likt utan kan variera med avseende på en mängd egenskaper; förklaringsvariabler kan vara korrelerade med varandra, det kan innehålla extrema observationer, förklaringsvariabler kan vara irrelevanta och antalet förklaringsvariabler kan överstiga antalet observationer. Det finns alltså ett underliggande samband mellan klassificeringsförmågan och datamaterialets egenskaper som manifesteras i att det optimala valet av klassificeringsmetod varierar med datamaterialet. Den som ställs inför ett klassificeringsproblem ställs därför även inför problemet att identifiera den mest lämpliga klassificeringsmetoden. Ytterligare en omständighet som försvårar är dock att kopplingen mellan klassificeringsförmågan och datamaterialets egenskaper länge varit förpassad till den vetenskapliga periferin och därmed är höljt i dunkel (Van Der Walt and Barnard, 2006).

Problematiken som uppsatsämnet bottnar i är alltså triangulär där antalet klassificeringsmetoder, kopplingen mellan klassificeringsförmågan och datamaterialets egenskaper samt att kopplingen däremellan är relativt outforskad skapar en situation där den som ställs inför ett klassificeringsproblem har ett stort antal klassificeringsmetoder till sitt förfogande dock ofta utan kunskap hur metoderna tillämpas för bästa resultat. Svårigheten med att identifiera den optimala klassificeringsmetoden resulterar i att metodvalet ofta vilar på individuella preferenser och gammal vana snarare än på vetenskapliga grund (Fernández-Delgado et al., 2014). King et al. (1995) summerar problematiken på följande sätt; *"The only general conlusion that can be made from these studies, is that the algorithm in which the authors have most interest tend to do best."*

1.1 Syfte och problemformulering

Med utgångspunkt dels i det faktum att valet av klassificeringsmetod är förknippat med flertal frågetecken och dels i att datamaterial ständigt är föremål för utveckling både i termer av komplexitet och tillgänglighet finns ett outsinligt behov av studier som riktar uppmärksamhet mot kopplingen mellan datamaterialets egenskaper och klassificeringsförmågan. En ökad förståelse om samspelet däremellan skapar förutsättningar för att exempelvis prediktera klassificerings-

förmågan hos en given klassificeringsmetod på ett givet datamaterial. Valet av klassificeringsmetod skulle i sådana fall, i större utsträckning, kunna baseras på empiriska slutsatser snarare än gammal vana. Uppsatsen tar avstamp i den problematik som har beskrivits ovan och syftar till att undersöka hur ett datamaterials egenskaper relaterar till klassificeringsförmågan och följaktligen det optimala valet av klassificeringsmetod. Denna intention uppnår vi genom att tillämpa LR, EN och BTK på sex verkliga datamaterial med varierande statistiska egenskaper. Genom att undersöka om, hur och varför klassificeringsförmågan hos LR, EN och BTK varierar med valda datamaterial hoppas vi bidra med insikt om kopplingen mellan datamaterialets egenskaper och klassificeringsförmågan samt belysa relativa för- och nackdelar hos de aktuella klassificeringsmetoderna. Klassificeringsmetoderna utvärderas främst med avseende på andelen korrekta klassificeringar på testdata. Eftersom det kan vara missvisande att stirra alltför blint på detta mått har det kompletterats med andelen falsk-positiva och falsk-negativa klassificeringar.

Uppsatsens syfte uppnår vi genom att besvara följande frågeställning:

"Vilka egenskaper karaktäriserar det datamaterial som LR, EN respektive BTK presterar bäst på i termer av andelen korrekta klassificeringar samt andelen falsk-positiva och falsk-negativa klassificeringar?"

1.2 Disposition

Uppsatsen har följande disposition; I nästkommande del presenteras det teoretiska fundamentet för LR, EN och BTK, vilka vi sedan har som intention att jämföra med varandra. I den efterföljande delen presenteras datamaterialen som står till grund för den statistiska analysen. I den fjärde delen av uppsatsen ställer vi klassificeringsmetoderna sida-vid-sida och jämför dem med varandra med avseende på resultatet som den statistiska analysen genererat. I uppsatsens avslutande del diskuterar vi sedan det erhållna resultatet med avseende på teoretiska skillnader och tidigare empiriska slutsatser om klassificeringsmetoderna.

2 Metod

I detta avsnitt presenteras de teoretiska grunderna i klassificeringsmetoderna LR, EN och BTK.

2.1 Logistisk regression (LR)

Ett av de mest användbara statistiska verktygen för att tackla klassificeringsproblem är LR. LR är framförallt användbar under förutsättningen att responsvariabeln följer en binomialfördelning och kan anta ett av två värden. Metoden tar sig an klassificeringsproblemet genom att skatta ett antal betingade sannolikheter, $P(y \mid x_1, x_2, ..., x_p)$, för att den givna enheten tillhör den ena av grupperna. Förutom att generera prediktioner ger LR möjligheten att beskriva relationer mellan responsvariabeln och förklaringsvariablerna. Sannolikheterna beräknas i enlighet med den logistiska funktionen (1).

$$p(y_i = 1 \mid x_i) = \frac{e^{(\eta_i)}}{1 + e^{(\eta_i)}} \tag{1}$$

där

$$\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}.$$

Den logistiska funktionen är fundamentet i den LR och beskriver hur respons – och förklaringsvariabler relaterar till varandra. Som synes av (1) modelleras sannolikheten som en icke-linjär funktion av förklaringsvariablerna. Detta garanterar att skattade sannolikheter är bundna mellan 0 och 1 (Faraway, 2005). De skattade sannolikheterna utgör sedan basen för prediktionen, där en subjektivt bestämd beslutsregel, anger hur stor den skattade sannolikheten måste vara för att klassificeras till den ena av grupperna, vanligtvis så klassificeras dock en observation till grupp 1 om $P\left(y_i = 1|x_i\right) > \frac{1}{2}$. Parameterskattningarna är lösningen på ett optimeringsproblem där log likelihood-funktionen $l(\beta_0, \beta)$ i (2) maximeras med avseende på förklaringsvariablerna som ingår i modellen (Friedman, Hastie, and Tibshirani, 2010).

$$l(\beta_0, \beta) = \frac{1}{N} \sum_{i=1}^{N} y_i \left(\beta_0 + x_i^T \beta\right) - log\left(1 + e^{\left(\beta_0 + x_i^T \beta\right)}\right) \tag{2}$$

Som en direkt effekt av att LR skattas via maximum likelihood-metoden besitter parameterskattningarna asymptotiska egenskaper som konsistens, väntevärdesriktighet och approximativ normalfördelning. Den allmänna tumregeln säger att antalet observationer per förklaringsvariabel inte bör vara mindre än femtio för att LR ska fungera på ett tillfredsställande sätt. LR har idag etablerat sig som en de mest populära metoderna för klassificering. En popularitet som dels går att härleda till tradition men även från att den är lätt att tolka, bely-

ser strukturer i datamaterialet samt vilar på relativt få antaganden. Vid sidan om antagandet om diskret responsvariabel kräver LR varken normalfördelning, linjära samband mellan respons – och förklaringsvariablerna eller lika varians mellan förklaringsvariablerna och är därför tillämpningsbar inom många situationer. Vidare producerar LR linjära beslutsgränser och är därför framförallt tillämpningsbar när klasserna i det aktuella datamaterialet enklast separeras med en linjär beslutsgräns (Burns and Burns, 2008).

LR är dock oförmögen till att utföra variabelselektion och bör därför användas med viss försiktighet när antalet förklaringsvariabler är stort i förhållande till antalet observationer eller när det förekommer starka korrelationer mellan förklaringsvariablerna. Under dessa förutsättningar tenderar den att överanpassa på träningsdata och generera missvisande parameterskattningar (Burns and Burns, 2008).

För att komma till svars med detta tillkortakommande har det utvecklats en familj av metoder, ”penalized” logistisk regression, som möter denna problematik genom att införa restriktioner i maximeringsproblemet. Det finns idag en handfull metoder som applicerar detta tillvägagångssätt, en av dessa är EN.

2.2 Elastic net (EN)

EN är en vidareutveckling av LR som arbetar genom att lägga till restriktioner på koefficienterna i modellen. Det har visat sig att EN klassificerar bättre än LR och den väljer bort irrelevanta förklaringsvariabler vilket underlättar tolkning av modellen. EN har en förmåga att reducera både bias och varians hos prediktionerna. Koefficienterna i modellen skattas i syfte att maximera funktion (3) d.v.s. differensen mellan log likelihood-funktionen (2) och strafftermen som läggs till på koefficienterna (Friedman et al., 2010).

$$\underset{(\beta_0,\beta)\epsilon R^{p+1}}{max} \left\{ l(\beta_0, \beta) - \lambda P_\alpha \left(\beta\right) \right\} \tag{3}$$

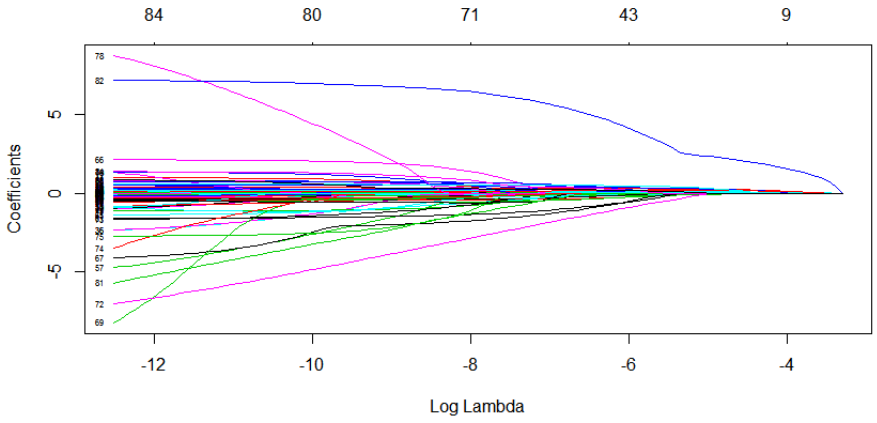
$$P_\alpha \left(\beta\right) = \sum_{j=1}^{p} \left[\frac{1}{2} \left(1 - \alpha\right) \beta_j^2 + \alpha \left|\beta_j\right| \right]$$

där $P_\alpha \left(\beta\right)$är strafftermen i EN.

Modellens intercept straffas inte och standardisering av förklaringsvariablerna

behövs för att strafftermen ska behandla dem likvärdigt. Innan EN kan användas måste dock två okända parametrar väljas:

1. Krympningsparametern λ ("shrinkage parameter") anger i vilken takt koefficienterna krymper mot 0. Om $\lambda = 0$ ger EN samma resultat som LR. Koefficienterna krymper snabbare mot 0 om större λ anges, om $\lambda = \infty$ är samtliga koefficienter noll (Se exempel i *figur* 1).



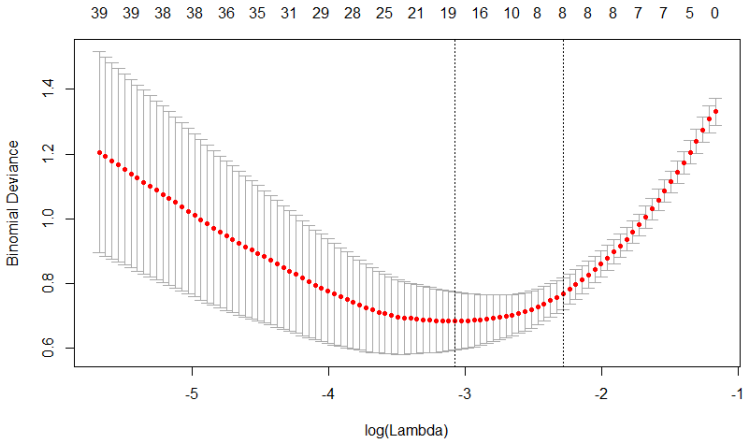
Figur 1: *Standariserade koefficienter för Caravan Dataset ((finns beskrivet i avsnittet datamaterial) plottas som en funktion av λ vid EN. Om $\lambda = 0$ skattas koefficienterna från LR. Koefficienterna krymper mot 0 i takt med att λ ökar mot ∞.*

2. Viktningsparametern α ("mixing parameter α") avgör hur l_1 och l_2-norm viktas .Om $\alpha = 1$ reduceras strafftermen till $\lambda \sum_{j=1}^{p} |\beta_j|$och funktionen omvandlas därmed till l_1-"penalized" logistisk regression (lasso), om $\alpha = 0$ så reduceras strafftermen istället till $\frac{1}{2}\lambda \sum_{j=1}^{p} \beta_j^2$ och funktionen omvandlas till ridge logistisk regression. α är oftast mellan 0 och 1 vilket innebär en kombination av lasso och ridge.

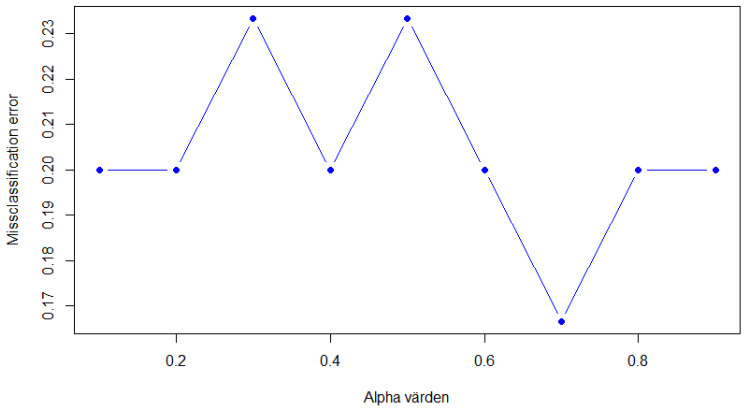
EN är alltså ofta en kompromiss mellan lasso och ridge eftersom den, i likhet med lasso, väljer ut vilka variabler som ingår i modellen och att den, i likhet med ridge, främjar en grupperingseffekt genom att tillåta korrelerade förklaringsvariabler inkluderas eller strykas bort från modellen tillsammans (Hastie et al., 2009). EN är i högsta grad användbar i situationer där det har visat sig att lasso

presterar sämre som t.ex. när antalet förklaringsvariabler är mycket större än antalet observationer ($p \gg n$) eller när korrelation mellan förklaringsvariabler är stark (Zou and Hastie, 2005).

I uppsatsen väljs värdet på krympningsparametern λ via korsvalidering (se exempel *i figur* 2). För att hitta den optimala viktningsparametern α används en loop där nio värden på α studeras, den som resulterat i minsta klassificeringsfel på testdata har sedan valts. Värden som studeras för α är 0,1 0,2 0,3 0,4 0,5 0,6 0,7 0,8 och 0,9 (*se figur* 3).



Figur 2: *Tiofaldig korsvalidering för att välja optimal λ på LSVT dataset (finns beskrivet i data avsnittet) vid EN. Devians (minus 2 log likelihood) plottas som en funktion av* $\log \lambda$*. Den vänstra vertikala linjen visar minsta devians medan den högra visar största värde på λ som ligger en standard-error från minimipunkten. Siffrorna som ligger längst upp indikerar storlekar på modeller.*



Figur 3: *Grafen visar klassificeringsfel på test data som en funktion av 9 olika värden på* α *som ligger mellan* $(0, 1)$ *på LSVT dataset (finns beskrivet i data avsnittet) vid EN. Optimal* α *i det här fallet blev* $\alpha = 0,7$.

2.3 Boosting tillämpat på klassificeringsträd (BTK)

Trädmetoder används både vid regressions - och klassificeringsproblem. Fokus i den här uppsatsen är att utnyttja klassificeringsträd för att prediktera en binär responsvariabel.

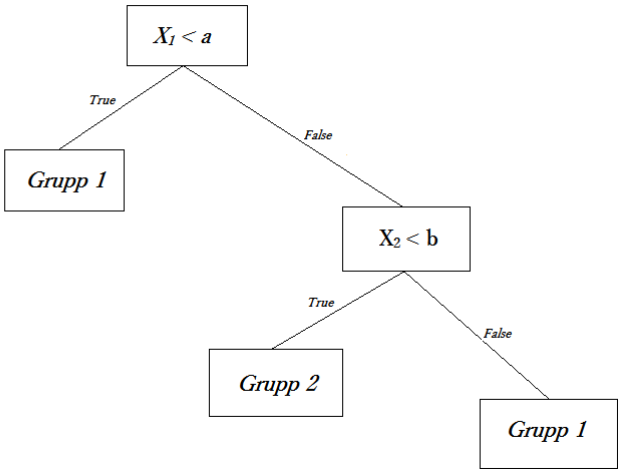
Ett klassificeringsträd arbetar genom att stegvis hitta binära delningar, med avseende på förklaringsvariablerna, som minimerar klassificeringsfelet i varje slutnod[1]. Förklaringsvariablerna kan vara både kontinuerliga och diskreta.

Ett träd börjar alltid med rotnoden[2] och söker sedan efter den binära delning som ger bästa information om klasserna. Trädet fortsätter med att hitta nya undernoder[3] och processen upprepas tills önskat resultat har uppnåtts, som t.ex. att antalet observationer är mindre eller lika med fem i varje slutnod. Det resulterande trädet innehåller ett stort antal delningar och korsvalidering används generellt sedan för att minska antalet delningar (James et al., 2013).

[1]En slutnod är en nod som saknar en binär delning.

[2]En rotnod ligger längst upp på träd och representerar den variabeln som ger den bästa delningen i data.

[3]En undernod är en nod som ligger mellan en rotnod och en slutnod.



Figur 4: *Exempel på klassificeringsträd med en binär responsvariabel (grupp 1 och grupp 2). Två binära delningar resulterar i två undernoder och tre slutnoder.*

I *figur 4* visas ett klassificeringsträd med två binära delningar. Först delas hela datamaterial med avseende på förklaringsvariabel X_1, om X_1 är mindre än en konstant a i rotnoden så klassas observationen till *grupp 1.* Sedan delas observationerna som inte uppfyllt kriteriet i rotnoden med avseende på förklaringsvariabel X_2. Om X_2 är mindre än en konstant b så klassas observationen till *grupp 2,* annars klassas den till *grupp 1.* Trädet innehåller två undernoder och tre slutnoder. Vid prediktion följs delningarna tills man hamnar i en specifik slutnod .

Gini index är ett bland flera mått för nodens renhet ("node purity")(James et al., 2013). Gini index mäter den totala variansen för nodernas renhet över samtliga klasser K och blir litet om alla $\hat{p}_{sk}$ är nära 0 eller 1 . Eftersom $\hat{p}_{sk}$ aldrig blir mindre än $\frac{1}{k}$, d.v.s. $\frac{1}{2}$ i vårt fall ($k = 2$ om det är en binär respons) så blir Gini index litet endast om $\hat{p}_{sk}$ är nära 1. Ett litet värde på Gini index indikerar att en slutnod domineras av observationer från en klass.

$$G(T) = \sum_{k=1}^{K} \hat{p}_{sk} (1 - \hat{p}_{sk}) \tag{4}$$

där

$$\hat{p}_{sk} = \frac{1}{N_s} \sum_{x_i \in R_s} I\left(y_i = k\right) \tag{5}$$

$\hat{p}_{sk}$ är proportionen av träningsobservationer som tillhör majoritetsklassen k i slutnod s , T är datamaterialet och N_s är antalet observationer som tillhör slutnoden s .

För att hitta den bästa delningen av varje nod väljs en förklaringsvariabel och en delningspunkt som minimerar följande ekvation(Muchai and Odongo, 2014):

$$G_{split}\left(T\right) = \frac{N_1}{N} G\left(T_1\right) + \frac{N_2}{N} G\left(T_2\right).$$

där datamaterial T är delat i 2 delmängder T_1 och T_2 med respektive storlek N_1 och N_2 .

Klassificeringsträd är kända för att vara lättolkade, det har dock visat sig att deras klassificeringsegenskaper är svaga (weak classifier[4]). Ett förslag till förbättring är att samla ihop ett större antal klassificeringsträd (James et al., 2013) .

Boosting är en av teknikerna som syftar till att förbättra klassificeringsegenskaper hos ett klassificeringsträd. Boosting anses vara en av de bättre metoderna som introducerats inom klassificering de senaste tjugo åren (Hastie et al., 2009). Idén bakom boosting är att skapa ett förbättrat klassificeringsträd genom att samla ihop ett flertal svaga klassificeringsträd(Hastie et al., 2009).

BTK har en förmåga att samtidigt reducera både bias och varians hos prediktionerna. BTK arbetar genom att tillämpa *algoritm 1* (Friedman, Hastie, Tibshirani, et al., 2000) på datamaterialet och därmed skatta många klassificeringsträd där varje klassificeringsträd sekventiellt har utvecklats från det föregående (se *figur* 5).

[4]En weak classifier klassificerar endast något bättre än slumpen.

Algoritm 1 *Real adaboost.*

1. Initialisera vikter för observationer $w_i = \frac{1}{N}$, $i = 1, 2, ..., N$.
2. För $m = 1$ till M :

   (a) Skatta ett träd från träningsdata med vikter w_i för att kunna räkna ut skattade sannolikheter $p_m(x) = \hat{P}_m(y_i = 1 \mid x_i) \in [0, 1]$.

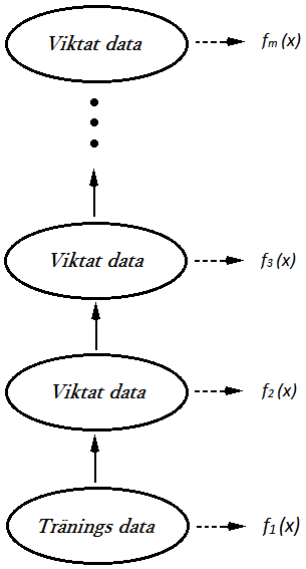
   (b) Sätt: $f_m(x) \leftarrow \frac{1}{2} \log \left[\frac{p_m(x)}{1-p_m(x)} \right] \in R$

   (c) Sätt $w_i \leftarrow w_i \exp\left[-y_i f_m(x_i)\right], i = 1, 2, ..., N$ och normalisera så att $\sum_i w_i = 1$.

3. Klassa observationen i majoritetsklassen $sign\left[\sum_{m=1}^{M} f_m(x)\right]$

För varje steg m i *algoritm 1* beräknas sannolikheten $p_m(X)$ för att en observation tillhör klass 1 (steg 2a). Efter det skattas $f_m(x)$ med ett klassificeringsträd som är ekvivalent med $\frac{1}{2}$ gånger logit-transformation av skattade sannolikheter i steg 2b. I steg 2c får alla observationer som klassades fel av ett klassificeringsträd $f_m(x)$ ökad vikt för att nästkommande träd $f_{m+1}(x)$ ska lägga fokus på dem. Denna procedur upprepas tills dess att alla M klassificeringsträd är skattade. Majoritetsklassen i samtliga klassificeringsträd avgör den slutgiltiga klassen för en observation (steg 3). Sannolikheter $P(y = 1 \mid x)$ bundna mellan 0 och 1 skattas på liknande sätt som vid LR (se *(1)*) (Friedman et al., 2000).

$$p(x) = P(y_i = 1 \mid x_i) = \frac{\exp\left[\sum_{m=1}^{M} f_m(x)\right]}{1 + \exp\left[\sum_{m=1}^{M} f_m(x)\right]}$$

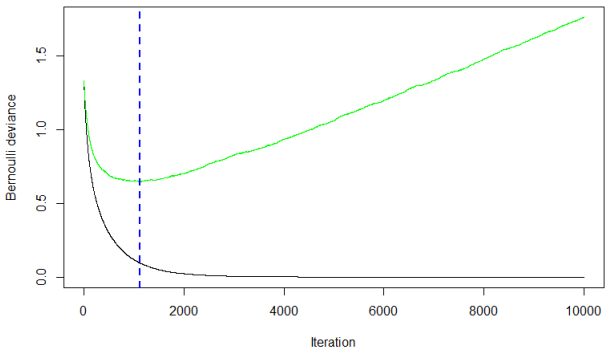


Figur 5: *Viktade versioner av träningsdata som används för skattning av M klassificeringsträd. Majoritetsklassen avgör den slutgiltiga klassen för en observation.*

Modellen för BTK har 3 "tuning parameters"(James et al., 2013):

1. Antalet klassificeringsträd M. BTK kan överanpassa om M är stort, även om överanpassningen tenderar att ske långsamt med ökande M. Korsvalidering används för att välja optimalt antal klassificeringsträd M (se exempel i *figur* 6). Bernoullideviansen används istället för klassificeringsfel vid korsvalidering (Ridgeway, 2007).

2. Krympningsparametern λ ("Shrinkage parameter λ"). I regel sätts den till $0,01$ eller $0,001$ beroende på viket problem det gäller. Litet värde på λ kräver ett större antal klassificeringsträd M och leder till förbättrade prediktioner (Ridgeway, 2007). λ kommer in i *algoritm 1* i syftet att fördröja hela processen vilket leder till att minska överanpassning (James et al., 2013).



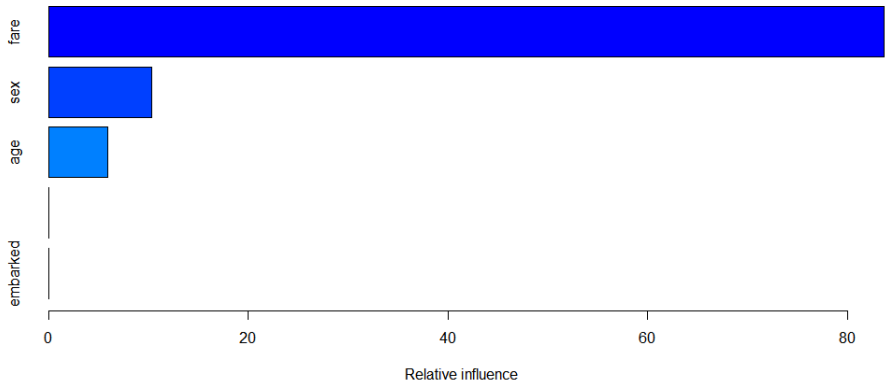
Figur 6: *10-faldig korsvalidering för att välja optimal "iterations " (antal klassificeringsträd M) på LSVT dataset (finns beskrivet i avsnittet datamaterial) vid BTK. 10.000 klassificeringsträd plottas mot Bernoullidevians. Den svarta linjen representerar deviansen på träningsdata, den gröna anger deviansen på testdata och den blåa vertikala streckad linjen visar minimipunkten för deviansen på testdata. Minsta deviansen uppnås vid 1126 träd.* $\lambda = 0,01$ *och* $d = 1$ *har använts för att skatta modellen.*

3. Antalet delningar d ("splits") som tillåts i varje enskilt träd. I princip fungerar $d = 1$ utmärkt. Om $d = 1$ kallas varje träd en stubbe ("stump") för att de bara innehåller en förklaringsvariabel (ingen interaktion tillåts här). Generellt sett kallas d interaktionsdjupet ("interaction depth") eftersom om $d > 1$ så tillåts interaktioner mellan förklaringsvariabler t.ex. om $d = 2$ så är interaktion begränsad mellan två förklaringsvariabler. En stubbe har i regel låg varians och ganska hög bias vid prediktioner.

I den här studien används $\lambda = 0,01$, $d = 1$ och totala antalet klassificeringsträd uppgår till antingen 100.000 eller 10.000 beroende på datamaterialets storlek, dels för att erhålla snabbare beräkningar och dels för att skydda mot överanpassning.

Genom att samla ihop många svaga klassificeringsträd skattas ett enda klassificeringsträd med bra egenskaper, det är dock omöjligt att få ett fint diagram likt *figur 1* som underlättar tolkning av resultatet. En samling av klassificeringsträd är inte lika tolkningsbart som ett enda klassificeringsträd, med hjälp av Gini index ges däremot möjlighet till en övergripande sammanfattning av hur inflytelserika förklaringsvariablerna är (se exempel i *figur* 7) (James et al.,

2013).



Figur 7: *Exempel på en plot som visar hur inflytelserika är förklaringsvariablerna som ingår i Titanic Dataset (finns beskrivet i avsnittet datamaterial). "Relative influence" är genomsnittet av förminskning i Gini index bland alla träd. I det här fallet har förklaringsvariabeln "fare" störst inflytande, sedan följer förklaringsvariablerna "sex" och "age", slutligen observeras att "embarked" och "pclass"inte har något inflytande alls i modellen.*

3 Datamaterial

För att besvara problemformuleringen och således uppnå uppsatsens syfte analyseras datamaterial som karaktäriseras av *i)* en stor datamängd, *ii)* förklaringsvariabler som mäts i samma skala, *iii)* att antalet observationer är större än antalet förklaringsvariabler *iv)* multikolinjäritet, *v)* att antalet förklaringsvariabler är större än antalet observationer och *vi)* att antalet observationer är större än antalet förklaringsvariabler samt att det är relativt hög-dimensionellt och präglas irrelevanta förklaringsvariabler. Datamaterialen har valts för att representera statistiska egenskaper vilka litteraturen har identifierat som betydelsefulla för klassificeringsförmågan. Litteraturen lyfter bl.a. fram fördelningen mellan kontinuerliga och kategoriska förklaringsvariabler samt antalet observationer i förhållande till antalet förklaringsvariabler som viktiga statistiska egenskaper (Tax and Duin, 2005). Datamaterialen har också valts för att representera vanliga situationer inom klassificering och för att uppnå en viss grad av mångfald på klassificeringsproblemen.

Den statistiska analysen syftar till att avgöra under vilka av förutsättningarna *i* - *vi* som LR, EN och BTK har främst klassificeringsförmåga. Den statistiska analysen försvåras dock av att verkliga datamaterial är komplexa och att det därför kan vara svårt att isolera effekten från den efterfrågade egenskapen. Ett datamaterial som präglas av multikolinjäritet kan exempelvis samtidigt definieras av att antalet förklaringsvariabler överstiger antalet observationer. För att motverka denna problematik och tydliggöra effekten från den eftersökta egenskapen har intentionen varit att välja datamaterial där den aktuella egenskapen är dominerande. Bortsett från att responsvariablerna har kodats om till att antingen anta 1 eller 0 och att kolumner med saknade värden exkluderats har datamaterialen inte bearbetats på förhand. Detta för att inte introducera förhållanden som riskerar att vara fördelaktiga för endera klassificeringsmetod. Datamaterialen har slumpmässigt delats in i ett tränings- och testdata, där ungefär 70% av observationerna tilldelats den tidigare och 30% av observationerna tilldelats den senare.

Datamaterialen har hämtats från UCI Machine Learning Repository (Lichman, 2013) och Kaggle (Anthony and Ben, 2010). *Titanic Data Set* har hämtats från Kaggle medan övriga datamaterial kommer från UCI Machine Learning Repository. UCI Machine Learning Repository har varit i drift sedan tidigt 1980-tal och har idag en databas bestående av 320 datamaterial. Databasen är vida använd och har främst etablerats som ett verktyg för att skänka empirisk validering till nyutvecklade klassificeringsmetoder och algoritmer. Kaggle är en plattform som framförallt verkar genom att anordna tävlingar inom statistisk modellering samt prediktion och som därför även tillhandahåller en stor mängd datamaterial. Genom att använda två erkända databaser erhåller vi hög trovärdighet och säkrar replikering. I tabellen nedan följer en sammanställning av datamaterialen. Tabellen visar antalet observationer, antalet förklaringsvariabler, fördelning mellan kontinuerliga och kategoriska förklaringsvariabler, den statistiska egenskap som datamaterialet representerar och slutligen andelen observationer som ingår i majoritetsklassen.

Tabell 1: *Tabellen summerar datamaterialen med avseende på antalet observationer (n), antalet förklaringsvariabler (p), antalet kontinuerliga (Kontin) och kategoriska (Kateg) förklaringsvariabler, statistisk egenskap och andelen observationer i majoritetsklass (% i M.K).*

Data	n	p	Kontin	Kateg	Egenskap	% i M.K
Bank	45211	17	10	7	Stor datamängd	88%
Tic-Tac-Toe	958	9	0	9	Homogen skala	65%
Titanic	1043	5	3	2	$n > p$	59%
Parkinson	197	23	23	0	Multikolinjäritet	76%
LSVT	126	310	310	0	$p > n$	67%
Caravan	5822	85	42	43	$n > p$	94%

i) Bank Data Set innehåller information från en telemarketingkampanj utförd av en bank i syfte att locka till sig nya kunder. Datamaterialet innehåller 45 211 observationer och representerar en situation som karaktäriseras av en större datamängd. Datamaterialet innehåller 17 förklaringsvariabler, vilka fångar aspekter som personlig information, bankhistorik och socioekonomisk bakgrund hos den potentiella kunden. Förklaringsvariablerna mäts på en kontinuerlig såväl som diskret skala. Responsvariabeln fångar utfallet av telemarketingkampanjen och antar värde 1 om den uppringda individen ansluter sig till banken och antar värde 0 om så inte är fallet. Responsvariabeln antar värde 1 vid 5289 gånger av fallen medan den antar värde 0 vid 39 922 gånger av fallen, alltså resulterar telemarketingkampanjen att ca 12% av de uppringda individerna ansluter sig till banken.

ii) Tic-Tac-Toe Data Set innehåller information från 958 partier luffarschack och syftar till att analysera hur valet av spelstrategi, hos den som inleder partiet, påverkar dennes vinstchanser. Responsvariabeln fångar således huruvida den som inleder partiet vinner eller inte. Den antar värde 1 vid vinst och antar värde 0 vid förlust eller oavgjort. Förklaringsvariablerna syftar till att måla upp spelplanen och representerar en given position på spelbrädet. Eftersom luffarschack traditionellt spelas på en plan med dimensionen 3×3 finns det nio förklaringsvariabler. Samtliga förklaringsvariabler mäts på samma nominella skala, {x, o, b}, och visar huruvida en given position innehåller ”x”, ”o” eller ”b”, där den sistnämnda anger om ett fält är tomt. Datamaterialet har därför valts för representera en situation där förklaringsvariablerna graderas på samma sätt.

Responsvariabeln har en fördelning, där 623 observationer tillhör grupp 1 och 335 tillhör grupp 2, den som inleder partiet står alltså som vinnare i 65% av fallen.

iii) Titanic Survival Data Set innehåller information från den ödesdigra dagen Titanic kolliderade med ett isberg och förliste utanför Newfoundlands kustremsa. Datamaterialet används för att analysera faktorer som kan påverkat sannolikheten för en passagerare att överleva tragedin. Datamaterialet har valts för att representera ett scenario där antalet observationer är större än antalet förklaringsvariabler. Datamaterialet innehåller 1043 observationer och 14 förklaringsvariabler, vilka bl.a. fångar aspekter som passagerarens sociala status, kön och ålder. Majoriteten av förklaringsvariablerna mäts i kategorisk skala medan ålder och biljettpris mäts i kontinuerlig skala. Responsvariabeln visar huruvida en given passagerare överlevde eller inte och antar värde 1 om passageraren överlevde och värde 0 om passageraren omkom. Nämnas bör dock att ett antal förklaringsvariabler präglades av betydande bortfall och tvingades därför uteslutas från den statistiska analysen. Fördelningen hos responsvariabeln visar att den antar värde 1 vid 386 gånger av fallen medan den antar värde 0 vid 657 gånger av fallen, alltså överlevde 41% av individerna i datamaterialet katastrofen.

iv) Parkinson Data Set innehåller information kopplade till röstsignaler hos 31 individer, varav 23 diagnostiserats med Parkinson, i syfte att skilja friska från sjuka baserat på uttal. Datamaterialet innehåller 197 observationer och 23 förklaringsvariabler, som fångar aspekter kopplade avvikande röstsignaler. Datamaterialet präglas framförallt av starka korrelationer mellan förklaringsvariablerna, som inte sällan överstiger 0,99, och har därför valts för att representera en situation där multikolinjäritet förekommer. Vidare mäts samtliga förklaringsvariabler på en kontinuerlig skala. Responsvariabeln indikerar huruvida rösten avviker från det normala, något som i sådana fall skulle kunna vara ett tecken på Parkinson. Responsvariabeln antar värde 1 om röstsignalen inte anses vara normal, annars antar den värde 0. Den antar värde 1 vid 148 gånger av fallen och värde 0 vid 49 gånger av fallen, alltså pekar 76% av röstsignalerna på Parkinson.

v) LSVT Voice Rehabilitation Data Set liknar föregående datamaterial i aspekten att den innehåller information kopplade till röstsignaler hos individer diagnostiserade med Parkinson. Målet med datamaterialet skiljer sig dock åt och

syftar istället till att avgöra huruvida en specifik röstbehandling leder till ett förbättrat uttal eller inte. Datamaterialet innehåller 126 observationer samt 309 förklaringsvariabler och representerar därför ett scenario definierat av att antalet förklaringsvariabler överstiger antalet observationer. Förklaringsvariablerna mäts på en kontinuerlig skala och till skillnad från det föregående datamaterialet är korrelationerna dem emellan inte lika distinkta. Responsvariabeln visar huruvida behandlingen har varit framgångsrik eller inte och antar värde 1 om uttalet, efter behandling, har förbättrats och annars värde 0. Den antar värde 1 vid 84 gånger av fallen och värde 0 vid 42 gånger av fallen, alltså leder 67% av behandlingarna i datamaterialet till ett förbättrat uttal.

vi) Caravan Data Set innehåller information om ett antal individer för att analysera faktorer kopplade till tecknandet av bilförsäkring. Datamaterialet innehåller 5822 observationer och 85 förklaringsvariabler, där ena hälften fångar sociodemografiska faktorer och andra hälften beskriver i vilken utsträckning individen tidigare tecknat försäkringar. Noterbart är att samtliga förklaringsvariabler kopplade till sociodemografiska faktorer är insignifikanta. Responsvariabeln indikerar huruvida individen ansluter sig till försäkringsåtagandet eller inte och antar värde 1 vid försäkringsinköp och annars värde 0. Precis som *Titanic Data Set* har datamaterialet valts för att representera en situation där antalet observationer är större än antalet förklaringsvariabler. Detta datamaterial är däremot även är relativt hög-dimensionellt och innehåller irrelevant information. Den deskriptiva analysen visar att responsvariabeln antar värde 1 vid 348 gånger av fallen och värde 0 vid 5474 gånger av fallen, alltså väljer 12% av individerna i datamaterialet att teckna en bilförsäkring.

4 Resultat

I den statistiska analysen har vi placerat LR, EN och BTK sida-vid-sida och jämfört dem med avseende andelen korrekta klassificeringar (KK), andelen falska negativa-rapporteringar (FN) samt andelen falska positiva-rapporteringar (FP) på testdata i datamaterialen som presenterades i föregående avsnitt.

Tabell 2: *Andelen KK, andelen FN och andelen FP för LR, EN och BTK på Bank testdata*

	KK	FN	FP
LR	89,86	2,7	66,2
EN	89,89	2,6	66,4
BTK	90,1	3	61,7

Inledningsvis analyseras klassificeringsmetoderna med avseende på datamaterialet *Bank Data Set*. *Tabell 2* visar att KK är nästintill likvärdig mellan klassificeringsmetoderna. I tabellen kan vi notera att KK för BTK uppgår 90,1 % medan motsvarande siffra för EN och LR är 89,89% respektive 89,86%. Samtliga klassificeringsmetoder var bra på att hitta individer som inte anslöt sig till banken, som högst är FN 3%. Klassificeringsmetoderna var dock desto sämre på att identifiera individer som faktiskt anslöt sig till banken, där samtliga metoder rapporterade en FP högre än 60%.

Tabell 3: *Andelen KK, andelen FN och andelen FP för LR, EN och BTK på Tic-Tac-Toe testdata*

	KK	FN	FP
LR	97,8	6,3	0
EN	97,8	6,3	0
BTK	96,3	7,9	0,43

När klassificeringsmetoderna tillämpas på datamaterialet *Tic-Tac-Toe Data Set* noteras, i *tabell 3*, att de i stor utsträckning lyckas placera observationerna i rätt klass. EN och LR genererar korrekta klassificeringar vid 97,8 % av fallen medan motsvarande siffra för BTK uppgår till 96,3%. Klassificeringsmetoderna utförde även ett utmärkt jobb i att identifiera vunna såväl förlorade partier av den som initierade spelet, där FN inte är högre än 8% och FP inte är högre än 0,5%.

Tabell 4: Andelen *KK, andelen FN och andelen FP för LR, EN och BTK på Titanic testdata*

	KK	FN	FP
LR	79	14,93	28,88
EN	80	14,93	28,17
BTK	68	20,4	48,6

Resultatet från den statistiska analysen *av Titanic Data Set* visar att EN har högst KK. EN lyckas separera passagerare som klarade sig med livhanken i behåll från passagerare som följde Titanic till havets botten vid 80% av fallen. I *tabell 4* kan vi dock se att skillnaden mellan LR och EN är försumbar. LR tilldelade observationerna till korrekt klass i 79% av fallen. Vidare visar tabellen en större diskrepans mellan BTK på ena sidan samt EN och LR på andra sida. BTK är framgångsrik i sin klassificering vid 68% av fallen och är därmed 12 procentenheter sämre än EN. Andelen FN lika stora för LR och EN, ca 15%, medan BTK med sina 20,4 % var aningen sämre. I testdata tillhör 142 passagerare den klass som överlevde förlisningen, 69 passagerare, alltså nästan hälften av dem klassades fel av BTK, 40 passagerare klassades fel av EN och 41 passagerare klassades i fel av LR, alltså är FP ca 28% för båda metoder.

Tabell 5: *Andelen KK, andelen FN och andelen FP för LR, EN och BTK på Parkinson testdata*

	KK	FN	FP
LR	81	50	11,48
EN	87	42,86	6,6
BTK	88	28,57	6,6

I det efterföljande datamaterialet, *Parkinson Data Set*, ställs klassificeringsmetoderna inför en situation som präglas av multikolinjäritet. I *tabell 5* noterar vi att BTK lyckas klassificera observationerna till rätt klass vid 88% av fallen, vilket räcker till att etablera den som främsta tillvägagångssätt under den givna förutsättningen. Vidare är dock EN, med rapporterad KK på 87%, endast marginellt sämre. Slutligen genererar LR en KK som uppgår till 81%. LR och EN rapporterade en hög andel FN, 50% respektive 42,86%. BTK identifierade flest röstsignaler som inte ansågs avvika från normala och andelen FN uppgick till ca

29%. Samtliga klassificeringsmetoder var bättre att identifiera röstsignaler som faktiskt var normala, där andelen FP för EN och BTK var ca 7% medan den uppgick till 11 % för LR.

Tabell 6: *Andelen KK, andelen FN och andelen FP för LR, EN och BTK på LSVT testdata*

	KK	FN	FP
LR	43	50	54,17
EN	83	16,67	16,67
BTK	77	33,33	29,17

Genom att ögna *tabell 6* uppmärksammas en relativt stor differens mellan samtliga metoders rapporterade KK när analysen förflyttas till *LSVT Data Set*. EN fungerar bäst under förutsättningen och lyckas skilja ett acceptabelt uttal från ett icke-acceptabelt vid 83% av fallen. Motsvarande siffra för BTK uppgår till 77%. LR fungerar dock betydligt sämre än tidigare och genererar korrekta klassificeringar endast vid 43% av fallen. LR stryker automatiskt förklaringsvariabler i skattningsalgoritmen tills antalet är detsamma som antalet observationer. Vidare identifierade EN både flest korrekta acceptabla uttal och icke-acceptabla uttal. Rapporterad FN och FP är ca 17% i båda fallen. Motsvarande för BTK uppgick till 33% respektive 29%. LR är visar prov på sämre precision och lyckas varken idenfiera acceptabla eller icke-acceptable uttal hälften av alla gånger.

Tabell 7: *Andelen KK, andelen FN samt andelen FP av LR, EN och BTK på Caravan testdata*

	KK	FN	FP
LR	93,7	0,49	100
EN	94,1	0,23	98,79
BTK	94,1	0,15	100

I *tabell 7* noterar vi att samtliga klassificeringsmetoder i stort sett generat likvärdiga resultat när *Caravan Data Set* varit under lupp. KK hos både EN och BTK uppgår till 94,1 % medan motsvarande siffra för LR uppgår till 93,7 % och är därmed endast marginellt sämre än sina kombattanter.Vidare visar resultatet att klassificeringsmetoderna var bra på att identifiera individer som inte tecknade bilförsäkring, samtliga genererade under en FN som var lägre än 0.5

%. Klassificeringsmetoderna presterade dock under all kritik när fokus riktades mot att identifiera individer som faktiskt tecknat försäkring. Varken LR eller BTK lyckas identifiera en enda individ medan EN endast identifierde två av individerna som tecknat bilförsäkring.

5 Diskussion

I uppsatsens inledning ställde vi frågan; *"Vilka egenskaper karaktäriserar det datamaterial som LR, EN och BTK presterar bäst på i termer av andelen korrekta klassificeringar samt andelen falsk-negativa och falsk-positiva klassificeringar?"*. Uppsatsens syfte har från första stavelse varit att besvara frågeställningen och därmed komma till insikt om kopplingen mellan klassificeringsförmågan och ett datamaterials egenskaper samt belysa valda klassificeringsmetoders relativa för – och nackdelar. I och med den statistiska analysen, i föregående avsnitt, har frågeställningen besvarats och om man blickar tillbaka mot resultatdelen är det tydligt att klassificeringsmetoderna har visat prov på relativt likvärdig klassificeringsförmåga över merparten av datamaterialen. LR har generellt sett klassificerat sämre än både EN och BTK. När den tillämpades på datamaterialen *Bank Data Set* och *Caravan Data Set* var andelen korrekta klassificeringar däremot i paritet med både EN och BTK. Dessa är datamaterial vilka har en stor mängd observationer som gemensam nämnare. Det verkar således som att LR fungerar bra under förutsättningen att datamaterialet innehåller ett stort antal observationer i förhållande till antalet förklaringsvariabler. Resultatet är i linje med tidigare empiriska slutsatser från bl.a. King et al. (1995) som rekommenderar LR givet en stor datamängd. Resultatet går också hand i hand med teori. LR har, som tidigare nämnt, goda asymptotiska egenskaper och hamnar därmed i sitt rätta element när datamängden är stor och antalet observationer är stort relativt antalet förklaringsvariabler. LR hanterar däremot datamaterialen *LSVT Data Set* och *Parkinson Data Set* desto sämre. I kontrast till föregående datamaterial är detta datamaterialen med minst antal observationer. Lika bra som LR fungerar vid stor datamängd lika dåligt tenderar den alltså att fungera under en mindre datamängd. Enligt Jordan (2002) kan detta härledas till att LR är asymptotiskt ineffektiv och behöver relativt många observationer för att komma till rätta. Vidare karaktäriseras datamaterialen av att antalet förklaringsvariabler är större än antalet observationer och multikolinjäritet. Eftersom LR inte genomför variabelselektion tenderar den dels att överanpassa på träningsdata och dels vara föremål för skenande varians och missvisande

parameterskattningar under denna förutsättning. Det är också tydligt att LR, överlag, har aningen högre FP samt FN och därmed även är aningen sämre än EN och BTK på att idenfiera faktiska positiva och negativa utfall.

I resultatdelen ser vi att EN genomgående genererat en hög andel korrekta klassificeringar, än om ofta i paritet med antingen LR eller BTK. EN sticker däremot ut från mängden när *LSVT Data Set* är under lupp. Det verkar alltså som att EN är att föredra framför LR och BTK i situationer där antalet förklaringsvariabler är många i förhållande till antalet observationer. Resultatet kommer dock inte som en överraskning. Ett av ENs främsta försäljningsargument är att den är kapabel till variabelselektion och därför är användbar när förklaringsvariabler behöver uteslutas från analysen. Som nämndes i teoridelen uppmuntrar EN även en grupperingseffekt. EN är därför även lämplig när multikolinjäritet är ett faktum. Detta är något som reflekteras i den höga andel korrekta klassificeringar i *Parkinson Data Set*. Generellt sett är EN det mest konsekventa tillvägagångssättet i analysen och hanterar, med bravur, även datamaterial som genomsyras av ett stort antal observationer samt innehållande både kontinuerliga och kategoriska förklaringsvariabler. EN har dessutom lägst rapporterad FP samt FN och är därmed främsta metoden att urskilja faktiska positiva och negativa utfall. Den låga andelen FN är ett faktum som gör att den kommer till användning exempelvis i en medicinsk kontext där en felaktig negativ klassificering kan vara skillnad mellan liv och död.

I resultatdelen ser vi att BTK levererar en relativt hög andel korrekta klassificeringar på samtliga datamaterial förutom *Titanic Data Set*. Om än marginellt placerar den sig framför EN i *Parkinson Data Set* såväl som *Bank Data Set*. Dessa är datamaterial som karaktäriseras av multikolinjäritet och en större datamängd. Den höga andelen korrekta klassificeringar på det tidigare datamaterialet är väntat och går i linje med teori. Detta faktum går att härleda till att BTK är konstruerad för att stegvis inkludera en förklaringsvariabel i taget och därmed duckar problematiken med multikolinjäritet. Eftersom BTK på detta sätt tillämpar variabelselektion och inkluderar förklaringsvariabler efter deras relativa relevans används den, enligt teori, även med fördel när antalet förklaringsvariabler är större än antalet observationer. Det observerade resultatet går dock till viss del stick-i-stäv med teorin när andelen korrekta klassificeringar är lägre än motsvarande hos EN. Detta kan möjligtvis förklaras utifrån krympningsparametern, λ. Ett lägre värde på λ leder till att modellen kräver fler klas-

sificeringsträd vilket i slutändan mynnar ut i förbättrad prediktion. Det man vinner i prediktion förlorar man dock i tid i form av accelererande beräkningskostnader. Det är därför ofta praktiskt omöjligt att välja alltför låga värden på λ (Ridgeway, 2007). I denna uppsats används uteslutandeλ=0,01. Vidare noteras att BTK fungerar sämre än både EN och LR på *Titanic Data Set*. Vad detta resultat beror på kan vi endast spekulera i. Datamaterialet karaktäriseras av att antalet observationer är stort i förhållande till antalet förklaringsvariabler och bör således inte utgöra ett hinder för BTK. En anledning kan dock vara att datamaterialet innehåller extrema observationer och betydande inslag av slump. Enligt Maclin and Opitz (1997) är BTK känslig för dessa situationer i och med att den vid den sekventiella utvecklingen kan lägga för mycket vikt vid onödig information och därmed överanpassa på träningsdata. En annan anledning kan vara att klasserna i datamaterialet bäst skiljs åt med en linjär beslutsgräns. BTK är en icke-parametrisk metod och kan därför producera en obefogat komplex beslutsgräns. BTK är således ömsom effektiv, när den tillämpas på rätt sätt, och ömsom undermålig, när den tillämpas på fel sätt. Den uppmärksamme noterade att BTK under fallet av multikolinjäritet genererade låg andel FN och lyckades identifiera normala röstsignaler i betydligt större utsträckning än EN och LR. En aspekt att ta i beaktning är dock kopplad till rapporterad FP. I *Caravan Data Set* lyckas BTK inte identifiera en enda tecknad försäkring. Detta kan vara ett resultat att endast 6% av observationerna tillhör denna klass och att träningsdata för klassen inte var tillräckligt stor. Det bör dock nämnas att detta tillkortakommande kan appliceras på LR och till viss grad EN. Samtliga klassificeringsmetoder tenderar alltså att fungera dåligt när majoritetsklassen är stor i förhållande till minoritetsklassen. En stor majoritetsklass i förhållande till minoritetsklass ställer även större krav på klassificeringsmetoderna i och med att det krävs mer för att förbättra den naiva prediktionen.

För att återkoppla till problemformuleringen karaktäriseras datamaterialen som LR fungerar bäst på av stor datamängd, datamaterialet som EN fungerar bäst på karaktäriseras av att antalet förklaringsvariabler är större än antalet observationer och datamaterialet som BTK fungerar bäst på präglas av multikolinjäritet. Vidare har den statistiska analysen exponerat mångsidigheten hos EN, som konstant har en hög klassificeringsförmåga. LR bör undvikas när datamängden är liten, antalet förklaringsvariabler är större än antalet observationer och multikolinjäritet förekommer. Slutligen avslöjas avigsidorna hos BTK när det aktuella datamaterialet innehåller extrema observationer och bjuder in till över-

anpassning. Det förekommer inte heller någon tydlig avvägning mellan andelen korrekta klassificeringar på ena sidan och FN samt FP på andra sidan utan dessa tenderar snarare att variera tillsammans. Den generella slutsatsen från uppsatsen är således att klassificeringsförmågan hos respektive klassificeringsmetod varierar med datamaterialet. Det optimala valet av klassificeringsmetod beror alltså på de statistiska egenskaper som karaktäriserar det aktuella datamaterialet.

I tabell 2-9 är det uppenbart att diskrepansen i KK, FN och FP, som slutsatserna baseras på, i några fall är försumbara och kan vara resultat av slumpmässig variation. Ett sätt att skänka ytterligare tillförlitlighet till slutsatserna hade varit att testa om de observerade skillnaderna är signifikanta. Ytterligare en aspekt att ta hänsyn till är det faktum att resultatet gäller för datamaterialen som använts i uppsatsen. Verkliga datamaterial är komplexa och förankrar den statistiska analysen i verkligheten. Komplexiteten i datamaterialen har dock även effekten att det är inte självklart att slutsatserna går att generalisera på andra datamaterial med liknande statistiska egenskaper. Ett sätt att undkomma denna problematik hade varit att simulera datamaterial med önskade statistiska egenskaper och därmed säkra att effekten isoleras. Ett annat sätt hade varit att utöka den statistiska analysen med fler verkliga datamaterial för att därmed stärka slutsatserna och erhålla mer generaliserbara resultat.

Avslutningsvis, är det viktigt att komma ihåg att klassificeringsmetoderna begränsas av kvaliteten på datamaterialet. Det vill säga att även om klassificeringsmetoderna har sina relativa för – och nackdelar så spelar dessa ofta mindre roll om datamaterialet är av hög kvalité, i den mån att det exempelvis innehåller ett stort antal observationer eller saknar brister som multikolinjäritet. Ett målande exempel för detta är att klassificeringsförmågan mellan klassificeringsmetoderna är i stort sett likvärdig i datamaterialen som präglas av ett stort antal observationer. Den som ställs inför ett klassificeringsproblem bör därför ha i bakhuvudet att det, i vissa fall, kan vara klokare att bearbeta och förbättra datamaterialet snarare än att ge sig ut i den snåriga djungeln av klassificeringsmetoder.

Referenser

Goldbloom Anthony and Hamner Ben. kaggle.com, 2010. URL `https://www.kaggle.com/`.

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, Belmont, Calif., 1984.

Robert P Burns and Richard Burns. *Business research methods and statistics using SPSS*. Sage, California, 2008.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7, 2006.

Julian J Faraway. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. CRC press, Boca Raton, FL, 2005.

Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1), 2014.

Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2), 2000.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1), 2010.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*, volume 2. Springer, New York, 2009.

Anil K Jain, Robert P. W. Duin, and Jianchang Mao. Statistical pattern recognition: A review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(1), 2000.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*. Springer, New York, 2013.

A Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14, 2002.

Ross D. King, Cao Feng, and Alistair Sutherland. Statlog: comparison of classification algorithms on large real-world problems. *Applied Artificial Intelligence an International Journal*, 9(3), 1995.

M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

Núria Macià, Ester Bernadó-Mansilla, Albert Orriols-Puig, and Tin Kam Ho. Learner excellence biased by data set selection: A case for data characterisation and artificial data sets. *Pattern Recognition*, 46(3), 2013.

Richard Maclin and David Opitz. An empirical evaluation of bagging and boosting. *AAAI/IAAI*, 1997, 1997.

Donald Michie, David J Spiegelhalter, and Charles C Taylor. Machine learning, neural and statistical classification. 1994.

Eunice Muchai and Leo Odongo. Comparison of crisp and fuzzy classification trees using gini index impurity measure on simulated data. *European Scientific Journal*, 10(18), 2014.

Greg Ridgeway. Generalized boosted models: A guide to the gbm package. *Update*, 1(1), 2007.

David MJ Tax and Robert PW Duin. Characterizing one-class datasets. 2005.

Christiaan Van Der Walt and Etienne Barnard. Data characteristics that determine classifier performance. 2006.

David H Wolpert and William G Macready. No free lunch theorems for search. Technical report, Technical Report SFI-TR-95-02-010, Santa Fe Institute, 1995.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 2005.