äd.2viBeslutsträdfigure.2

Maskininlärnings Klassificerings Algoritmer inom Data Mining

Gustave Rousselet

September 2016

Innehåll

1 Introduktion

Då multimedia, sociala medier och internet anslutna enheter snabbt utvecklas, har människor och företag idag fler möjligheter än någonsin att samla in mer information. Detta är på grund av datans exponentiellt stora tillväxt. Under 2010 passerade mänskligheten barriären av en zettabyte skapad data i den digitala universum. År 2011 hade den siffran stigit till 1,8 ZB och 2015, 9 ZB. Med denna exponentiell modell av datatillväxt, kommer mänskligheten år 2020 över 44 ZB data (Gantz & Reinsel 2011). Data har blivit en viktig tillgång som kan vara jämföras med materiella tillgångar och humankapital. En studie genomförd av McKinsey 2011 fann att om big data kan kreativt och effektivt utnyttjas för att förbättra effektiviteten och kvaliteten, kan det potentiella värdet av den amerikanska medicinska industrin som vunnits genom data överträffa 300 miljarder dollar, vilket skulle minska den amerikanska sjukvårdskostnaderna med över 8 procent. Återförsäljare som helt utnyttjar deras data kan förbättra sin vinst med mer än 60 procent. Big data kan också användas för att förbättra effektiviteten i den statliga verksamheten, så att de utvecklade ekonomierna i Europa skulle kunna spara mer än 100 miljarder euro (Manyika et al. 2011). Värdet av datan ligger i de observationer, mönster och prognoser som kan skapas från dem, och eftersom datamängder blir allt större och komplexa, kan normala observationsmetoder inte producera dessa insikter. Vetenskapen av att använda datorer för att göra dessa observationer kallas "data mining"(Myatt 2007). Ett stort problem med den enorma ökningen av data är hur man ska gå tillväga för att klassificera data till meningsfulla klasser. Detta är anledningen till att företag som IBM, Facebook, Microsoft och Amazon satsar på teknik för att dra meningsfulla insikter och klassificeringar från stora datamängder (Fan & Bifet 2013).

2 Syfte

Denna vetenskapliga rapport kommer att fokusera på data mining, och omfattningen kommer att begränsas till klassificering av data med hjälp av maskininlärnings algoritmer. Denna vetenskapliga rapport syftar till att svara på frågan:

Kan maskininlärnings algoritmer som används i data mining producera meningsfulla klassificeringar från datamängder för stora för människors observationer?

Denna frågeställning kommer att besvaras genom att utföra ett experiment för att testa hur väl maskininlärnings algoritmer kan klassificera uppsättningar av osorterade data. Den bakgrundsinformation och teori som krävs för att förstå detta experiment kommer att beskrivas i teori delen av rapporten. Specifika detaljerna om experimentet kommer att redogöras i metoden.

3 Teori

Syftet med denna del av rapporten är att ge en kort översikt av begreppet data mining och maskininlärnings algoritmer som används i data mining. Den är inte avsedd att åstadkomma en omfattande förståelse av teorin, men för att helt enkelt bilda en grund som kan därefter användas för att bättre förstå den genomförda experimentet.

3.1 Data Mining

Data mining är vetenskapen av att använda datorkraft för att analysera data och dra meningsfulla insikter från den. Dessa insikter kan komma i form av klassifikationer, mönsterigenkänning , relationer och prognoser. Denna vetenskapliga rapport fokuserar på klassificering. Data mining genomförs vanligen med en startpunkt för att lösa ett problem. När det gäller klassificeringen, kommer detta problem i form av ”hur delar vi osorterade data in i olika meningsfulla klasser?”(Myatt 2007). Ett exempel från verkliga världen på detta skulle vara att man vill organisera bilder i klasser baserat på vilka objekt de innehåller, eller vill organisera sig kunder på ett mataffär in i olika klasser baserat på vilken mat de köper.

3.2 Maskininlärnings Algoritmer

Lärande är den mänskliga processen av att använda tidigare erfarenheter för att anpassa sig till de olika situationer som uppstår i våra liv. Denna inlärning kan komma i många olika sorter som motorik när ett barn lär sig gå, eller matematiska färdigheter när ett barn lär sig att lägga till. Den grundläggande principen bakom all inlärning är att det dras från tidigare erfarenheter och iakttagelser, och sedan användas för framtida situationer. Maskininlärning syftar till att tillämpa denna princip samma lärande datorer. Denna process är allmänt sammanfattas som en process av induktiv slutledning. Detta innebär att datorn observerar ett fenomen och sedan försök att karakterisera fenomenet genom att skapa en modell. Denna modell används sedan för att karakterisera framtida fenomen. Inom området för klassificering i data mining, är maskininlärning algoritmer som en datamängd som består av instanser med flera attribut där varje instans har tilldelats en klass, kallas detta övningsuppsättningen. Maskinen inlärningsalgoritm använder sedan denna utbildning som kommer att skapa en modell som så exakt som möjligt tar attribut instanserna, och ger en korrekt klassificering. Maskinen inlärningsalgoritm används sedan på en datamängd med instanser med samma attribut, men nu utan den tilldelade klassen. Maskinen inlärningsalgoritm använder sedan tidigare utvecklad metod att tilldela instanser till klasser (Japkowicz & Shah 2011). Förmågan hos maskininlärning algoritm för att korrekt klassificera fall bestäms sedan av mängden korrekt märkta instanser. De två maskininlärning algoritmer som kommer att fokusera på i detta dokument är: Naive Bayes och C4.5.

3.3 Naive Bayes

Naive Bayes är en maskininlärnings algoritm av typen Bayesian classifier. En Bayesian classifier är inom maskininlärning en klassificerings algoritm som använder Bayes teori.

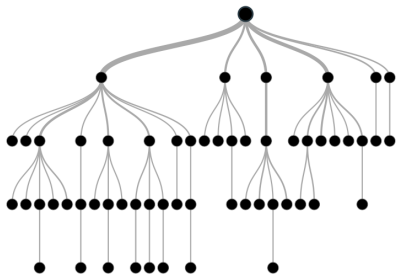
$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

Figur 1: Bayes teori

Bayes teori är en matematisk sats som beskriver sannolikheten av en händelse baserad på de villkor som har ett samband med händelsen(Augustin et al. 2014). I vårt fall är händelsen klassen, och villkoren är attributen av instanserna. Ett exempel på Bayes teori inom klassificering skulle vara om vi ville veta om en person hade cancer, och vi vet att deras ålder är 65 år. Här försöker vi klassificera instansen "person" som "har cancer"eller "inte har canceröch en Bayesian klassificerings algoritm skulle använda attributet för att förutsäga vilken klass personen skulle hamna i. En Naive Bayes algoritm förutsätter då helt enkelt att alla attribut är oberoende av varandra (Wu et al. 2008).

3.4 C4.5

C4.5 är en form av beslutsträd klassificerings algoritm inom maskininlärnings algoritmer som utvecklades av Ross Quinlan (Augustin et al. 2014). I grundläggande termer börjar denna algoritm med att studera träningsmängden och utveckla en trädstruktur baserad på entropin av de olika instanser. Roten av trädstrukturen är alla icke-klassificerade instanser, och bladen är klasserna som instanserna hamnar i. C4.5 algoritmer återanvänder därefter samma modell på datamängden som ska klassificeras, och instanser delas sedan in i olika klasser i botten av trädet (Japkowicz & Shah 2011).



Figur 2: Beslutsträd

4 Metod

Följande avsnitt fungerar som en översikt över den metod som användes för att samla information om den nödvändiga bakgrundsteorin, liksom den specifika metod som användes för att genomföra experimentet.

4.1 Literatur

Litteraturen som användes i denna studie har använts för att ge en kort men omfattande översikt av begreppen maskininlärningsalgoritmer, data mining, och big data. Den huvudsakliga källan till information om maskininlärningsalgoritmer har kommit från “Evaluating Learning Algorithms: A Classicfication Perspective”. Boken publicerades i 2011 av Cambridge University Press. Publiceringsdatumet gör detta till en värdefull informationskälla, på grund av den snabbt förändrandes landskap och utveckling av maskininlärningsalgoritmer. Cambridge University Press är också en mycket uppskattad utgivare, detta är en hänvisning till kvaliteten på informationen i boken. Boken har varit en värdefull resurs för studien eftersom den har en begränsad omfattning som specifikt motsvarar omfattningen av denna studie och därmed genomförda experimentet, nämligen utförandet av maskininlärning algoritmer inom klassificeringsproblem. “Introduction to Imprecise Probabiltiesöch Top 10 Algorithms in Data Mining”har fungerat som kompletterande källor för att ge mer information om specifika uppgifter om maskininlärningsalgoritmer. Den främsta källan till information om data mining har kommit från Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining”. Data mining är ett ämnesområde

som snabbt förändras, och därmed, eftersom att boken nyligen publicerades 2007 så är den en värdefull informationskälla i den snabbt förändrandes landskap av data mining. Boken ger en inledande förståelse för teorin bakom data minig, och därför är det relevant för omfattningen av denna studie och experimentet som har begränsats till en grundlig förståelse av data mining. Kompletterande informationskällor på big data och data mining har kommit från rapporter och vetenskapliga artiklar. Detta beror på att big data är en aktuell fråga och flera vetenskapliga rapporter och fallstudier hat genomförts för att producera värdefulla insikter om ämnet. De rapporter och studier som är relevanta för denna studie måste vara heltäckande och ge en kort översikt till ämnet av big data och data mining, liksom framtidsfrågor inom området. De utvalda vetenskapliga rapporter Mining Big Data: Current Status och Future Forecasts", "Extracting Value from Chaosöch Big Data: The Next Frontier for Innovation, Competition and Productivitypassar de nämnda kriterier och samlar in information från olika studier för att dra insikter om nuvarande statusen och framtiden av big data och data mining.

4.2 Experiment

Experimentet bestod av att välja maskininlärningsalgoritmer för att prestations testa, hitta relevanta datamängder med tillräcklig variation och storlek, samla datamängderna och sedan genomföra experimentet av prestandan på maskininlärningsalgoritmerna.

4.2.1 Toolkit

Klassificerings algoritmer som har använts i detta experiment är alla tillgängliga i gratis programvaran WEKA Toolkit 3.8. WEKA Toolkit är ett open source maskininlärningsalgoritm verktygslåda skriven i Java. Den levereras med alla maskininlärningsalgoritmer tillgängliga och möjliggör för olika tester, såsom klassificering, regression, clustering etc. Omfattningen av denna studie är inom klassificering, och därmed var den enda testmiljön används.

4.2.2 Datamängder

Alla datamängder som har använts i denna experiment har samlats in från de fria och öppet tillgängliga arkivet UCI maskininlärning Repository. Anledningen till att denna arkiv valdes som en källa för datamängder är att

den är en allmän användning inom maskininlärning gemenskapen, med citerats i över 1000 vetenskapliga studier och rapporter. En viktig del för att se till att experimentens resultaten var vetenskapligt giltig var att välja datamängder som reflekterade ett brett och varierat utbud av data. Detta beror på att syftet med denna studie var att se om maskininlärningsalgoritmer kan producera värdefulla insikter från datamängder för stor för mänsklig observationskapacitet. För att kunna svara på denna fråga, skulle experimentet kräva ett brett spektrum av datamängder för att säkerställa att maskininlärningsalgoritmerna skulle användas och testas i olika problem och miljöer av klassificering. De valda datamängderna valdes för att passa ett kriterium för att ha olika egenskaper som spänner över ett brett spektrum av områden, antal instanser som sträcker sig från 150 till 2310, antal attribut som sträcker sig från 4 till 36, och vilken typ av attribut. Dessa val gjordes eftersom maskininlärningsalgoritmer måste testas på tillräckligt olika datamängder för att möjliggöra en jämförelse av resultaten från experimentet. Tabellen nedan summerar valen av datamängderna och deras egenskaper.

Tabell 1: Datamängder

Datamängd	Data Typ	Standard Uppgift	#Instanser	#Attribut
Soy Bean	Multivariate	Classification	638	36
Breast Cancer	Multivariate	Classifcation	569	32
German Credit	Multivariate	Classification	1000	20
Iris	Multivariate	Classification	150	4
Segment	Multivarite	Classifcation	2310	19

4.2.3 Maskininlärningsalgoritmer

Experimentet utfördes genom att först välja en lämplig algoritm för att verka som ett riktmärke för att kunna jämföra resultatet av maskininlärningsalgoritmerna. WEKA Toolkit ger en benchmarking algoritm, ZeroR algoritmen, som helt enkelt sorterar alla dessa fall i den vanligaste förekommande typen förekommer i träningsmängden. De maskininlärningsalgoritmer som prestanda testades, C.45 och Naive Bayes valdes eftersom de ger insikt in i prestandan av maskininlärningsalgoritmer av två olika typer: beslutsträd och Bayes algoritmer. Alla dessa tre maskininlärningsalgoritmer finns färdiga i WEKA Toolkit och behöver inte importeras eller ändras för varje enskilt test som skall genomföras. Däremot var träningsmängderna av maskininlärningsalgoritmerna

framställda genom ett "k-fold Cross Validation". Detta innebär att datamängder delades upp i k olika block, och sedan har k-1 av dessa block använts som träningsmängden för maskininlärningsalgoritmerna, och det sista blocket användes som en utvärderingsmängd för algoritmernas klassificerings prestanda. Denna beslut fattades eftersom alla datamängder hade olika storlekar, och därmed att ge en viss procent uppdelning av de som skulle användas i träningsmängden skulle ha orsakat variationer i de relativa storlekarna av träningsmängderna. "k-fold Cross Validationmöjliggör en jämnare fördelning av träningsmängder till prestanda testen av de testade maskininlärningsalgoritmer. De valda datamängder importerades till WEKA Toolkit för varje experiment.

5 Resultat

Den här delen av uppsatsen är till för att diskutera och dra insikter från resultaten från experimentet i förhållande till forskningsfrågan och medföljande bakgrunds teorin som presenterades i inledningen av detta dokument.

5.1 Experiment Resultat

Tabell 2: Experiment Resultat

Algoritm / Datamängd	Soy Bean	Breast Cancer	German Credit	Iris	Segment
Zero R	13.47%	70.27%	70.00%	33.33%	15.73%
Naive Bayes	92.97%	71.67%	75.40%	96.00%	81.06%
C4.5	91.50%	75.52%	70.50%	96.64%	95.73%

Resultaten av experimentet ger oss möjlighet att dra flera insikter om prestandan av maskininlärnings algoritmer. Den första observationen som kan göras är att maskininlärnings algoritmer överträffade prestandan benchmark i varje test som kördes. Av detta kan vi tydligt se att maskininlärnings algoritmer kan producera värdefulla klassificeringar på var och en av datamängder. Dessa resultat ger ett svar på uppsatsens frågeställning men kan dock ge ytterligare insikter och konsekvenser.Experimentet utfördes på ett sådant sätt att hastigheten av maskininlärnings algoritmerna inte spelade roll i resultaten. Men varje maskininlärnings algoritm hade en kompileringstid och körning under 2 sekunder för varje test. Syftet med denna studie var

att genomföra ett experiment för att se om maskininlärnings algoritmer kan producera värdefulla klassificeringar från data mängder för stora för human observationskapacitet. Denna frågeställning har en implicit fråga om maskininlärnings algoritmer kan överträffa människor i klassificeringsproblem i extremt stora datamängder. Forskningsfrågan formulerades med tanke på den nuvarande situationen av big data och data mining som det presenterades i bakgrundsinformationen. Resultaten av experimentet har demonstrerat att maskininlärnings algoritmer kan vara ett mycket användbart verktyg i behandlingen av datamängder för stora för mänsklig observations kapacitet. Med tanke på deras snabba kompileringstid och betydande noggrannhet som sett i testen Segment, Iris och Soy Bean kan vi se att maskininlärnings algoritmer har stor potential att lösa klassificeringsproblem inom extremt stora datamängder.

Vidare från experimentets resultat, kan vi se att resultatet för varje testad maskininlärnings algoritm inte följer en entydig bana. Naive Bayes algoritmen överträffade C4.5 algoritmen i Soy Bean och German Credit klassificeringstesten, men C4.5 var mer korrekt i Breast Cancer, Iris och Segment testen. Detta tyder på att maskininlärnings algoritmer har en varierande prestanda beroende på karaktären av den datamängd som testas. Denna insikt kan ligga till grund för ytterligare forskning om de varierande prestandan av maskininlärnings algoritmer inom klassificeringsproblem.

Experimentets resultat visar också att ingen tydlig indikation av prestandan kan dras från de specifika egenskaperna hos de analyserade data mängderna. Man skulle kanske tro att eftersom maskininlärnings algoritmer byggs från träningsdata från datamängderna, att större datamängder med fler träningsfall skulle ge bättre klassificeringsprestanda. Denna trend återspeglas inte i resultatdatan. De maskininlärnings algoritmer överträffade ZeroR med 0,50 till 5,00% i German Credit testet med 1000 fall, men var bättre än ZeroR med 63,66-67% i Iris testet med 150 fall. Vidare kan man tro att fler attribut i datamängden skulle resultera i mer exakt klassificering. Denna trend saknar också stöd från experimentets resultatdata. Maskininlärnings algoritmerna visade bättre prestanda i Iris datamängden med 4 attribut, och sämre prestanda i German Credit testet med 20 attribut. Ytterligare forskning måste genomföras, kanske med ett större utbud av datamängder, för att studera den varierande prestandan av maskininlärnings algoritmer inom klassificeringsproblem med varierande karaktär av datamängder.

6 Diskussion

Den här delen av pappret kommer att diskutera litteraturen och experimentet som användes i uppsatsen för att försöka besvara frågeställningen. Diskussionen kommer att redogöra för de eventuella brister i litteraturen och experiment, samt förslag till eventuella ändringar som skall göras om experimentet skulle utföras på ett liknande sätt för framtida forskningsändamål.

6.1 Diskussion av Literatur

Litteraturen som användes för att genomföra denna studie fungerade som en värdefull källa till information för att bygga en grund av kunskap för att förstå genomförda experimentet. Detta är dock inte att säga att valet av litteraturen inte kunde förbättrats. Det viktigaste att ta hänsyn till när det handlar om teorin bakom data mining och maskininlärnings algoritmer, är det snabbt förändrande landskapet i studieområdet. Litteraturen som används måste vara nyligen publicerad och relevant för det nuvarande studielandskapet. Eventuellt i framtida studier, forskningsrapporterna borde nyligen ha publicerats (under det senaste året optimalt) och särskilt fokusera på endast en maskininlärnings algoritm och dess nuvarande status. Vidare kan rapporterna som fokuserar på data mining och big data ha en svag partiskhet som skulle leda författarna till att överbetona hur brådskande big data klassificeringsproblem egentligen är. Eftersom vikten av dessa frågor ligger i framtiden, måste författarna har tagit vissa friheter för att förutsäga. Detta skulle kunna leda till ogrundade antaganden som skapas inom ämnet. Som med alla förutsägelser om framtiden för ett område av studien/forskning, måste man ta hänsyn till de medförda komplikationer och felaktiga heuristik som kan uppstå under utformningen av förutsägelser. Framtida forskningsrapporter inom data mining och big data bör ta hänsyn till detta när framtida förutsägelser om ämnet citeras.

6.2 Diskussion av Experiment

Experimentet med denna studie var den huvudsakliga källan till information som användes för att besvara frågeställningen. Testerna producerade värdefull data som användes för att finna insikter om prestandan av maskininlärnings algoritmer på stora datamängder. Detta är dock inte att säga att experimentet utfördes utan fel. Tvärtom, det finns många förändringar och

överväganden som bör beaktas om framtida forskningsrapporter eller studier skulle använda ett liknande experiment i sina metoder. Framtida experiment skulle kanske producera mer tillförlitliga resultat om maskininlärnings algoritmerna som testades hades utvecklats och implementerats av forskarna själva. Vid användningen av en toolkit, finns det en viss asymmetri av information associerad från det faktum att maskininlärnings algoritmerna inte utvecklats av forskaren själv. De kan programmeras på ett sådant sätt att fungera bättre med vissa datamängder än andra, eller kompromissa prestanda i ett försök att öka minneskapaciteten. Dessa frågor måste beaktas av forskaren vid användningen av toolkits. Dessutom skulle detta experiment har förbättrats med en större uppsättning av maskininlärnings algoritmer för att jämföra klassificerings prestandan. Detta skulle ha lett till mer omfattande insikter om den varierande prestandan maskininlärning algoritmer av olika typer i klassificeringsproblem. Men på grund av omfattningen av denna forskningspapper, var det ett aktivt val att bara fokusera på två maskininlärnings algoritmer. Ytterligare forskning och fallstudier skulle kanske överväga en annan forskningsfråga som använder sig av en bredare uppsättning av maskininlärnings algoritmer. En annan förbättring skulle vara att använda fler datamängder för att producera ett bredare utbud av karakteristiska skillnader mellan datamängderna. Som med alla vetenskapliga experiment, är det i intresset för forskarna att försöka upptäcka trender och dra slutsatser från experimentets resultatdata. Ett bredare utbud av egenskaper hos datamängderna skulle kanske ge en bättre grund som forskaren kan dra slutsatserna från. Detta beror på att variationen i datamängderna kan leda till mer distinkta och lätt igenkännbara mönster som bildas i resultatdatan. Framtida forskare som skulle välja att genomföra ett sådant experiment bör vara försiktiga för att inte välja alltför många datamängder, eftersom detta skulle kunna leda till ett överskott av information och identifiering av falska trender på grund av en felaktig analys av försöksdata.

6.3 Slusats

Syftet med denna studie som beskrevs i frågeställningen var att se om maskininlärnings algoritmer kan producera meningsfulla klassificeringar i datamängder för stora för mänsklig observations kapacitet. Med hjälp av bakgrunds teorin och resultatdatan från genomförda experiment, har svaret på frågeställningen visat sig vara att maskininlärnings algoritmer kan producera meningsfulla klassificeringar i datamängder för stora för mänsklig observa-

tions kapacitet. I diskussionsavsnittet har olika fel och brister i litteraturen samt genomförda experimentet framtagits. Ytterligare förslag av forskning med tanke på de resultat som uppnåtts i denna vetenskapliga rapporten kommer att beskrivas i nästföljande avsnittet.

7 Efterord

7.1 Vidare forskning

Såsom det beskrevs i början av denna uppsats, den främsta bidragande faktorn till utförandet av detta experiment var den aktuella diskussionen om big data och data mining. Med hänsyn till detta bör ytterligare forskning genomföras för att testa prestandan av maskininlärnings algoritmer på datamängder som inte kan beräknas från en enda processor, utan att beräkningskraften måste fördelas över ett stort antal processorer som arbetar parallellt för att bearbeta uppgifterna. Detta skulle kunna göras med hjälp av Hadoop MapReduce (Dean & Ghemawat 2010) till exempel, vilket skulle vara ett användbart verktyg i ett experiment som skulle närmare simulera de större datamängderna som diskuterats i bakgrundsinformationen. Dessutom kan forskningen bedrivas för att testa de varierande hastigheterna av maskininlärnings algoritmer inom klassificeringsproblem. Experimentet som genomfördes i denna forskning papper använde datamängder och maskininlärnings algoritmer som (på grund av den relativt mindre storlek och komplexitet) inte producerade långa kompileringstider. Men om framtida forskningen skulle undersöka hastigheterna hos maskininlärnings algoritmerna, bör forskarna överväga att använda betydligt större datamängder för att producera längre kompileringstider, vilket skulle möjliggöra en jämförelse av hastigheten av maskininlärnings algoritmerna.

Referenser

Augustin, T., Coolen, F. P., de Cooman, G. & Troffaes, M. C. (2014), *Introduction to imprecise probabilities*, John Wiley & Sons.

Dean, J. & Ghemawat, S. (2010), 'Mapreduce: a flexible data processing tool', *Communications of the ACM* **53**(1), 72–77.

Fan, W. & Bifet, A. (2013), 'Mining big data: current status, and forecast to the future', *ACM sIGKDD Explorations Newsletter* **14**(2), 1–5.

Gantz, J. & Reinsel, D. (2011), 'Extracting value from chaos', *IDC iview* **1142**, 1–12.

Japkowicz, N. & Shah, M. (2011), *Evaluating learning algorithms: a classification perspective*, Cambridge University Press.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. & Byers, A. H. (2011), 'Big data: The next frontier for innovation, competition, and productivity'.

Myatt, G. J. (2007), *Making sense of data: a practical guide to exploratory data analysis and data mining*, John Wiley & Sons.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y. et al. (2008), 'Top 10 algorithms in data mining', *Knowledge and information systems* **14**(1), 1–37.