SYNDATA

Bilaga 1
Rapport från Syndata AB
Syntetisk data inom intensivvård Fas 1

Denna rapport beskriver utförande och sammanfattar lärdomar som Syndata förvärvat under projektet ”Syntetisk data inom intensivvård” som drivs i samarbete med Region Västerbotten (RVB) och AI Sweden. Dessutom presenteras ytterligare arbete och förbättringsmöjligheter samt förslag om fortsättning.

Bakgrund

RVB samarbetade med Syndata för att skapa syntetisering av 2 patientuppsättningar för att användas för:

- Förutsägelse av vårdbördan inom intensivvården
- Förutsägelse av tid för utskrivning på intensivvård

Genomförande

Syndata slutförde fas 1 framgångsrikt genom att leverera 4 modeller till RVB tillsammans med medel för att syntetisera olika datauppsättningsstorlekar. Syndatas beslut att dela 4 modeller (2 olika modeller för varje originaldatauppsättning) stöds av experiment som bedömde de två bästa presterande modellerna för varje datauppsättning (t.ex. urladdat och återställt). Baserat på Syndatas kvalitetsbedömning utvärderas båda modellerna som liknande kandidater för RVB-mål.

Den kompletta listan över leveranser inkluderar:

För urladdat dataset:

2 modeller (pickle filer)
2 syntetiska dataset, en för varje modell, 1x storlek
2 syntetiska dataset, en för varje modell, 5x storlek
1 Jupyter anteckningsbok, en för varje modell som innehåller
- Egenskapernas likhetsanalys
- Saknade värdeanalys
- Risk för re-identifikation

PDF med resultat av modell 1, 1x datauppsättningsstorlek

För återställd datauppsättning:

2 modeller (pickle filer)
2 syntetiska dataset, en för varje modell, 1x storlek
2 syntetiska dataset, en för varje modell, 5x storlek
1 Jupyter anteckningsbok, en för varje modell som innehåller
- Funktioner likhetsanalys
- Saknade värdeanalys
- Risk för re-identifikation

PDF med resultat av modell 1, 1x datauppsättningsstorlek

+ PDF-rapport med projektinlärningar och processdokumentation

Utvärdering av möjligheter med syntetisk data

Hur påverkas modellens prediktion av användningen av syntetiska data?

Prediktionen kommer att köras av RVB med deras överlevnadsanalysmodell. Inom fas 2 vore det fruktbart om Syndata har tillgång till den slutgiltiga modellen som ska använda det syntetiska data som skapas.

Lärdomar/förslag för fas 2: Att ha tillgång till modellen för överlevnadsanalys kunde ha varit användbart för att jämföra de två syntetiska datamängderna. Att ha tillgång till överlevnadsanalysen kommer att möjliggöra en förbättring av algoritmerna innan leverans.

Hur bra fungerar databalansering med syntetisk data?

För närvarande hålls proportionerna i syntetisk datauppsättning som i den ursprungliga datauppsättningen. En annan arbetsinriktning är att Syndata ska balansera de syntetiska datamängderna mellan exempelvis kvinnor och män.

Lärdomar/förslag för fas 2: som ett förslag bör balansering av datauppmängder undersökas.

Hur förhåller sig syntetisk data till originaldata?

Syndata utvecklade ett utvärderingsramverk för att systematiskt utvärdera likheter/skillnader mellan syntetisk data och originaldata. Ramverket uppskattar kvaliteten på syntetiska data genom att jämföra ett antal mätvärden. Dessa inkluderar: individuella distributioner, parvisa distributioner, korrelationen för saknade värden.

Lärdomar/förslag för fas 2: Syndatas utvärderingsram är under kontinuerlig utveckling. Vi planerar att utöka utvärderingsramverket med mått som sammanfattar robusthet över datauppsättningar i olika storlekar.

Hur kan man säkerställa att den syntetiska datan inte är för lik originaldatan?

Syndata arbetar med djupinlärning teknologi (Generative Adversarial Network, GAN) för syntetisering. Komplexiteten hos dessa algoritmer minskar risken för reverse engineering. Som det neurala nätverkets natur antyder representerar de syntetiskt producerade data syntetiska individer. Även om en syntetisk datapunkt har liknande värden som en datapunkt i originaldata, så betraktas datapunkten inte som en identifierbar person och därför blir troligen inte GDPR tillämpningsbart, det bör utredas vidare i Fas 2.

Som ett mått på likhet utvecklade Syndata ett mått som kallas Risk of Re-identification. Mätvärdet bedömer risknivån för att syntetisk datapunkt ska identifieras på nytt. Den tittar på hur lik den är jämfört med någon annan punkt i originaldata.

Lärdomar/förslag för fas 2: Funktionen Risk of Re-identification bör diskuteras och indata ställas in tillsammans med data ägaren. Vissa uppgifter, exempelvis unika eller extrem värden, skulle kunna innebära ett alltför högt värde på Risk of Re-identification. Beslut kan då behöva tas att eliminera eller baka in dessa.

Hur syntetiserar och använder man en regions känsliga originaldata på ett säkert sätt?

Syntetisering av regionens känsliga data kan utföras lokalt, inom regionens nätverk eller på Syndatas server. Förslagsvis sker detta innanför data ägarens brandväggar.

Lärdomar/förslag för fas 2: Varje dataägare kommer att ha den syntetiska motorn anpassad till deras krav/förhållanden och den syntetiska motorn kommer att skicka de syntetiska datamängderna till en databas. Vi ser tre möjliga sätt att arbeta för detta:

1) De syntetiska dataseten kommer att ha en form som alla är överens om.

2) De syntetiska dataseten produceras i samma format som original data. Det innebär att varje part som vill ta del av ett syntetisk dataset behöver, troligtvis, göra ändringar på formatet för att kunna ta in syntetiska datauppsättningar från andra dataägare.

3) De syntetiska dataseten är anpassade till varje regions krav. Det innebär att varje data ägares data syntetiseras och tillhandahålls i det format som passar varje part som ska ta del av det syntetiska data.

(3) är det mest fördelaktiga för data ägarna, som troligtvis också är samma som användarna av de syntetiska dataseten i Fas 2, men kommer att kräva ett arbete med att kartlägga de olika format som data ägarna använder.
Detta är också fallet med (2), skillnaden är dock att varje dataägare skulle behöva utföra det arbetet individuellt.
Vi rekommenderar därför att välja antingen alternativ (1) eller alternativ (3).

Vilka verktyg används för syntes?

Syndata använder GAN's, gaussiska copulas och andra algoritmer för syntetisering.

Hur kan syntetisk data utvärderas, kliniskt, juridiskt och tekniskt?

Lärdomar/förslag för fas 2:
Syndata föreslår ett utvärderingsramverk som bedömer de tekniska aspekterna av de syntetiska datamängderna. Detta är ett mycket viktigt ämne att hantera i fas 2 med tanke på att det är mycket känsliga uppgifter. Kräver behandlingen av data som syntetisering ett samtycke eller är informationsplikten tillräcklig? Även att utmana funktionen re-identifikation som Syndata tillhandahåller. I detta kommer även pre-processandet in i diskussionen – syftet med den syntetiska datamängden etc.

Vägledning för nationell spridning av metoder och möjligheter att arbeta med syntetisk data.

Beskrivning av användningsområden för syntetiska data (strukturerad data).

När behövs syntetisk data, vilka är de lämpligaste användningsområdena?

(1) Det är Syndatas ståndpunkt att det stora värdet av syntetisk data är att det bibehåller ett datasets värden, form och struktur men inte innehåller några personuppgifter. Därmed är GDPR inte applicerbart på det syntetiska datasetet (detta är dock något som bör verifieras i Fas 2) vilket möjliggör att det syntetiska datasetet kan

   (a) delas externt och internt
   (b) användas till mer än vad det finns legal bas för (GDPR)

(c) sparas för framtida bruk

(2) Med den teknik vi använder (GAN) för att syntetisera tillkommer också andra fördelar. Man kan
  (a) Förstora ett dataset
    (i) När man behöver mer data än vad som finns tillgängligt för att träna AI modeller
    (ii) Finna mindre svaga samband som man annars lätt kan missa
    (iii) Tidigare se samband i ett nytt skeende
  (b) Imputation, fylla ut luckor i ett dataset
  (c) Hitta och hantera skevheter i ett dataset
    (i) anpassa ett dataset till en annan population

När behöver man fler "syntetiska" patienter att träna AI-modellen på?

Fler "syntetiska" patienter kan behövas för att förbättra resultaten av en AI-modell. Ett sätt att bedöma när man ska lägga till fler "syntetiska" patienter är att jämföra förutsägelsekvaliteten för AI-modellerna med olika datamängder. Man kan lägga till nya syntetiska datapunkter för att modellens noggrannhet ska öka. Detta kan vara en uppgift att utreda inom fas 2.

Jämför kvaliteten på originaldata och syntetiska data från samma originaldata när de används i AI-modeller, vilka är skillnaderna?

I fas 1 använde vi på Syndatas utvärderingsramverk. Den övergripande kvaliteten på syntetiska datauppsättningar bedömdes med hjälp av detta ramverk. Dessutom körde Syndata modellutvärderingar med linjära modeller. Resultaten avslutades med de två bästa modellerna för var och en av de ursprungliga datamängderna.

Lärdomar/förslag för fas 2: Även om vi har generella utvärderingsmått, så är det bästa tillvägagångssättet att utvärdera syntetisk data genom att att utvärdera inom det sammanhang den ska användas för. I detta projekt hade det varit att testa den med överlevnadsanalysmodellen som RVB använder.

Lärdomar

Följande lärdomar har förvärvats under projektperioden:

1) Individuell distributionskvalitet påverkas i hög grad av den korrekta definitionen av kategoriska kontra icke-kategoriska värden.

En viss variation i distributioner förväntas, men den bekväma tröskeln mellan distributioner bör diskuteras med dataägaren.

I vissa fall kan användaren korrigera distributionerna med lägsta likhet. För närvarande är detta inte ett skalbart alternativ, eftersom korrigeringen för en mer lämplig distribution måste göras manuellt. I framtiden kan Syndata titta på ett automatiserat sätt att korrigera dessa distributioner.

2) Parvisa fördelningar kan bedömas med olika koefficienter. Valet av koefficient ska göras med hänsyn till funktionernas datatyper.

3) Det rekommenderas att undvika upprepad information med syntetiseringsmodeller (och maskininlärningsmodeller i allmänhet). Även om precisionen kan förbli densamma, kommer modellerna att ta längre tid att träna och förmodligen vara instabila. Med upprepad information hänvisar vi till kolumner som innehåller samma information (dvs de duplicerar data i ett annat format/datatyp).

4) Även om generella utvärderingsmått för en syntetiserad datauppsättning kan tillhandahållas, är den bästa utvärderingen för en syntetiserad datauppsättning att utvärdera inom det sammanhang som den ska användas för. I projektets fall är den bästa utvärderingen som rekommenderas att testa den med överlevnadsanalysmodellen som RVB använder. Testande av den syntetiska data med den slutgiltiga modellen som datat ska användas i ger också möjligheter till att anpassa algoritmerna än mer.

5) Domänexpertis är viktig för att förbättra syntetiseringen. En domänexpert ger insikter om betydelsen av data (till exempel för att bedöma om funktioner är kategoriska eller inte), och kan även ge ytterligare värde för att identifiera orsakssamband i data. Så en nära koppling mellan SynData och en domänexpert vore att föredra i fas 2.

6) Utvärderingen av saknade värden (missing values) visar potential för förbättring. Syndata skulle vilja undersöka ett bättre sätt att bevara saknade värdens proportioner och korrelationer. Detta kan göras genom att göra ytterligare arbete med att förbättra nuvarande modeller.

Leverabler i Fas 1 från Syndata

Visuella verktyg/mått för Syndatas ramverk:
- 6 visuella utvärderingar av syntetisk datakvalitet
- Har likheter (4 visuella utvärderingar)
- Saknade värden (2 visuella utvärderingar)
- 6 syntetiska datakvalitetsutvärderingsmått.
- WD, medelvärde och standardavvikelse för WD
- Euklidiskt avstånd för ömsesidig information (för parvisa distributioner)
- Euklidiskt avstånd för Pearsons korrelation (för parvisa distributioner)
- Euklidiskt avstånd för ömsesidig information (för saknade värden)
- Euklidiskt avstånd för Pearsons korrelation (för saknade värden)
- Risk för re-identification

Diskussionspunkter inför Fas 2

Vilka typer av leveranser bör skapas i Fas 2?

- Förbättringar av syntetiska datauppsättningar genom att ta bort upprepad information och korrekt identifiera kategoriska egenskaper
- Bedömning av robustheten hos syntetiska dataset: utveckling av statistik med ökningen av syntetiska datasets storlek
- Förbättrad mätning av parvisa distributioner. Anpassning baserat på datasetets egenskaper
- Juridisk utvärdering av det syntetiska datasetet och utvärderingsramverket.
- Förbättra funktionen ”risk of re-identification” med anpassade nivåer av personlig information.
- Simulera specifikt use case scenario för RVB.

Vilka kompetenser behövs i Fas 2?

Utöver det som var med i Fas 1 skulle det främsta vara

1) Domänexpertis med datakunskap.
2) Jurister
3) IMY, Integritetskyddsmyndigheten
4) KTH, utvärdering av dataset som skapats genom en GAN.