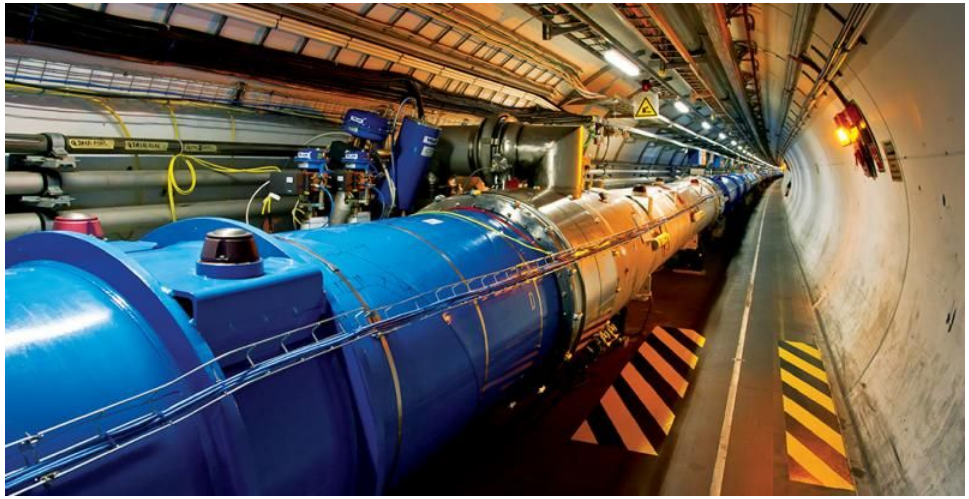
Andmekaeve osakestefüüsikas

Tudeng: Margus Pärt
Juhendaja: Mario Kadastik PhD
Retsentsent: Toomas Kirt PhD



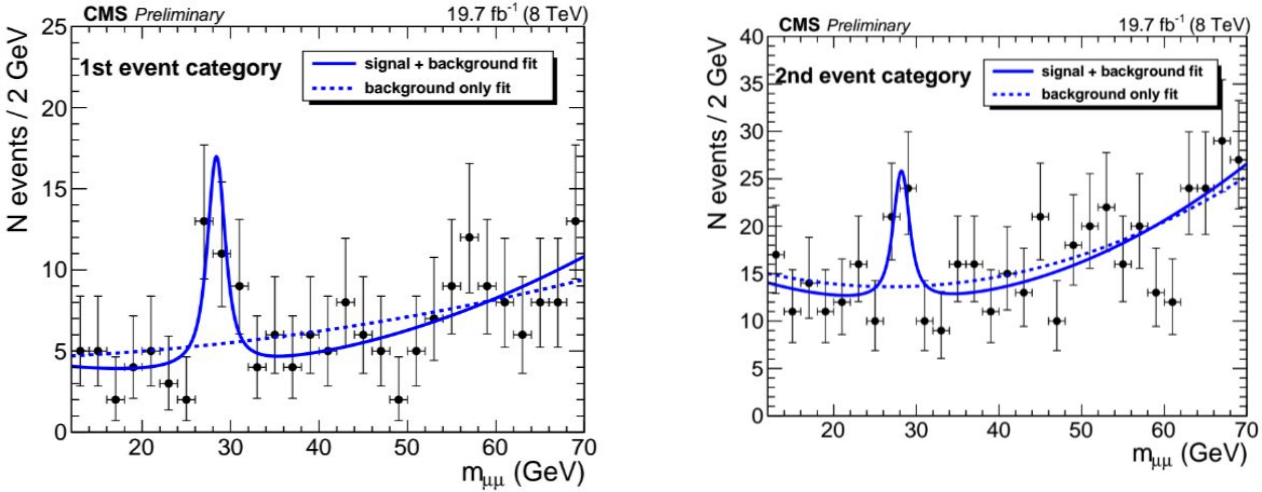
Diplomitöö eesmärgid

- Kokkupõrgete tulemusel tekkinud osakeste uurimine (juba klassifitseeritud andmete filtreerimine).
- 2017 a. CERN-i prootonikiirendi kokkupõrgete andmete analüüsi tarkvara parendamine.
- Võrrelda tehisliku närvivõrgu võimekust b-kvarki eristamisel ülejäänud kvarkidest kasutusel oleva otsustuspuuga.

Lagunemiskanali uurimisprojektis osalemine

Eesmärk:

- Saada teada, kas 2012 a. andmetes leitud andmete liiasus 28.4 GeV piirkonnas oli statistiline fluksatsioon või avastati uus alg- või komposiitosake.

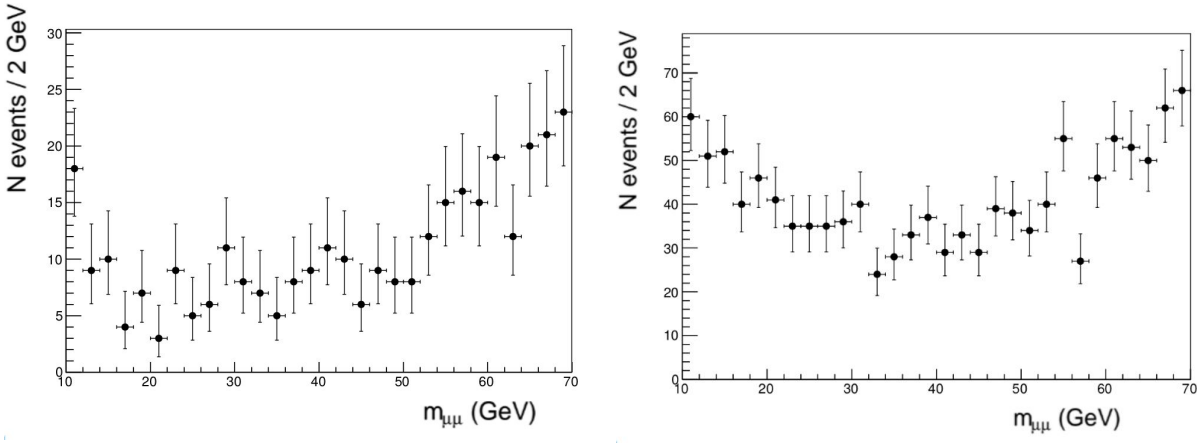


Joonised kirjeldavad kahe müüoni nelivektorite summa puhul liiasust 28.4 GeV piirkonnas.

Lagunemiskanali uurimisprojektis osalemine

Autori töö tulemusena:

- Valmis (C++-is kirjutatud) programm, mis kordas 2012 a. andmetele rakendatud filtrite tingimusi 2016 a. andmetele, salvestas filtrid läbinud andmeid ja joonistas filtrid läbinud sündmuste kohta käiva olulise teabe graafikutele.



Joonised ei kirjelda kahe müüoni nelivektorite summa liiasust 28.4 GeV resonantsi piirkonnas.

Parendused analüüsi toetavas tarkvaras

Probleem, mis ohustas uurimistulemuste õigeaegset avaldamist:

- Teadlaste käivitatud ööpäev läbi kestvad analüüsid ebaõnnestusid ilma vajaliku tulemuseta erinevatel juhuslikel ajahetkedel ilma selge veateateta.

Esimene samm lahenduse suunas:

- Olemasoleva tarkvara arhitektuuri ja protsessi mõistmine, logimise täiendamine.

Parendused analüüsi toetavas tarkvaras

Leitud probleem 1:

- Kõikide serverite vahel jaotatud ressurss, võrguketas ei tulnud toime vajaliku lugemiste ja kirjutamiste arvuga ajaühikus.

Lahendus 1:

- **Efektiivsem logimine**. Logimise salvestamine esmalt analüüsi teostavasse kobararvutisse ja alles peale analüüsi sammu lõppu logi tagasi kopeerimine analüüsi käivitanud serverisse.

Parendused analüüsi toetavas tarkvaras

Leitud probleem 2:

- Kobarserverites tehtud analüüsi alamosade tulemuste agregeerimine üheks andmefailiks ja hilisemaks töötlemiseks võttis kaua aega.

Lahendus 2:

- **Efektiivsem analüüsitulemuste agregeerimine**. Kui varasemalt agregeeriti kobararvutitel läbiviidud analüüsi tulemused ühe sammu käigus ühes serveris, siis diplomitöö autori programmeerimise tulemusena agregeeritakse nüüd andmeid rekursiivselt mitmes kobarserveris samaaegselt.

Parendused analüüsi toetavas tarkvaras

Leitud probleem 3:

- Andmekandjale ebaõnnestunult andmete kirjutamine ei pruukinud tähendada analüüsi katkemist ja selget veateadet.

Lahendus 3:

- **Metaandmete loomine ja kontroll.** Diplomitöö autori programmeerimise tulemusena salvestatakse ja kontrollitakse automaatselt faili sisu metaandmetega, erinevuse korral kuvatakse selge veateade, mis võimaldab vea tekkekohta kiiremini leida ja põhjust lahendada.

Parendused analüüsi toetavas tarkvaras

Leitud probleem 4:

- Koodimuudatuste testimine võttis kaua aega ja tehtud vigasi oli raske avastada.

Lahendus 4:

- **Moodultestimise juurutamine.** Autori programeerimise tulemusena juurutati moodul-testimise raamistik, mis võimaldab automaatselt:
  1.) valideerida, et muudatustega ei ole rakenduse funktsionaalsus lõhutud;
  2.) dokumenteerib olemasoleva rakenduse funktsionaalsust;
  3.) kirjeldab programmilist liidestust komponentidega;

Detektori sensoreid tabanud algosakeste täpsem klassifitseerimine

Probleem:

- CMS sensorite poolt salvestatud kokkupõrgete andmed on ebapiisava täpsusega - salvestuseelsel rekonstrueerimisel kasutatakse lihtsaid algoritme, mis on kiired, aga ebatäpsed.

Lahendus:

- Klassifitseerida sündmuses osalenud algosakesed uuesti ja täpsemalt.

Detektori sensoreid tabanud algosakeste täpsem klassifitseerimine

Varasemalt kasutusel olnud klassifikaator tuvastamaks b-kvarke:

- Otsustusmetsaga leitud otsustuspuu.

Ülesanne:

- Testida, kui täpne on tehisnärvivõrk klassifitseerimisel võrreldes otsustuspuuga.

Parima tehisnärvivõrgu mudeli leidmine

Sammud:

1. Mõista, millist tüüpi tehisnärvivõrku kasutada sõltuvalt andmete sisust (pärilevivõrk, konvulitsiooniline võrk või rekurrentne võrk).
2. Valida sobiv raamistik ja programmeerimiskeel.
3. Leida võimalikult tehisnärvivõrgu mudel, mis oleks võimeline võimalikult efektiivselt (arvutusvõimsuse kasutus, mälukasutus).
4. Tehisnärvivõrgu täpsuse klassifitseerimisel võrdlemine seni kasutusel olnud otsustuspuu täpsusega.

Parima tehisnärvivõrgu mudeli leidmine

Tehisnärvivõrgu tüübi valik:

- Kuna iga kokkupõrke sündmust iseloomustavad parameetrid on ajast sõltumatud (eelmine sündmus ei mõjuta järgmist ja sündmust iseloomustavatel parameeteritel puudub ajaline järjestus), siis sobis pärilevivõrk (*feed forward network*)

Parima tehisnärvivõrgu mudeli leidmine

Sobiva raamistiku ja programmeerimiskeele valik.

- Kuna programmeerimiskeeled Python ja C++ on CERN-is ja KBFI-s kõige rohkem kasutusel olevad keeled, siis otsustati kasutada Python-it.
- Kuna TensorFlow oli kõige populaarsem avatud lähtekoodiga C++-s kirjutatud ja Pythoni liidestusega, siis otsustati kasutada TensorFlow-d.

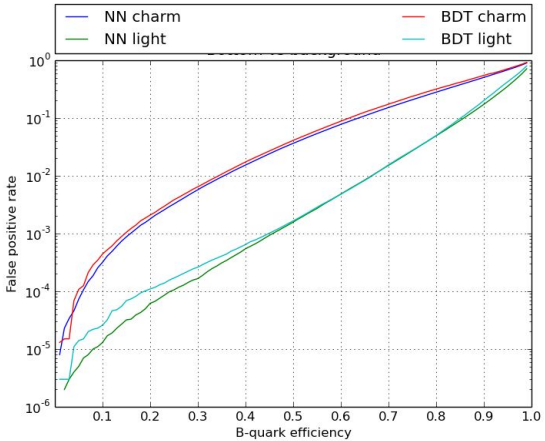
Parima tehisnärvivõrgu mudeli leidmine

Parima tehisnärvivõrgu mudeli topoloogia leidmine:

- Kasutati ammendavat otsingut (*exhaustive search*) üle tehisnärvivõrgu topoloogia võimalike konfiguratsioonide parameetriruumi (kombineerides peidetud kihtide arvu, närvide arvu peidetud kihtides, aktivatsioonifunktsioone, väljajätumeetodi poolt mittekasutud ühenduste hulka kihtide vahel, õpikiirust ja õpikiirusesammu vähendamist).

Parima tehisnärvivõrgu mudeli leidmine

Tehisnärvivõrgu täpsuse klassifitseerimisel võrdlemine seni kasutusel olnud otsustuspuu täpsusega (*true positive rate vs false positive rate*):



Joonis: CERNi senise parima algoritmi võrdlus närvivõrguga. Võrdlus on koostatud 3 miljoni sündmuse põhjal jaotudes võrdselt B-kvarkide, C-kvarkide ja light-kvarkide vahel. B-kvark (signaal) X-teljel ja et UDSG- või C-kvark loetakse B-kvargiks (müra Y-teljel).

Parima tehisnärvivõrgu mudeli leidmine

Tehisnärvivõrgu täpsuse klassifitseerimisel võrdlemine seni kasutusel olnud otsustuspuu täpsusega (*true positive rate vs false positive rate*):

B-kvargi tuvastamise efektiivsus	**NN C-kvargi müra**	**BDT C-kvargi müra**	**NN light-kvarkide müra**	**BDT light-kvarkide müra**
0.1	0.00031	0.000419	0.000014	0.000034
0.2	0.001788	0.002039	0.00006	0.00012
0.3	0.005868	0.006496	0.00018	0.000289
0.4	0.015992	0.017776	0.000528	0.000658
0.5	0.03702	0.041414	0.001563	0.001657
0.6	0.077637	0.087768	0.004746	0.004863
0.7	0.151539	0.172835	0.014906	0.015401
0.8	0.280967	0.316567	0.049397	0.049774
0.9	0.507186	0.549237	0.175866	0.202589
0.99	0.899689	0.916134	0.712426	0.792432

Tabel kirjeldab, kuidas sõltuvalt sellest, kui suure tõenäosusega b-kvark peab olema tuvastatud ja võrdleb seda, kui palju C-kvarke või kergeid kvarke seetõttu b-kvarkina tuvastatakse.

Töö tulemused

- Lagunemiskanali andmete filtreerimine võtab keskmiselt 2 tundi (10%) vähem aega.
- Kui analüüs peaks katkema mõnes etapis mingil põhjusel, siis on võimalik vealogist selget teavet, kus ja miks viga tekkis.
- Analüüsitarkvaral on turvavõrk moodultestide näol.
- Ei leitud kahe müüoni nelivektorite summa liiasust 28.4 GeV resonantsi piirkonnas.
- Teatakse, et tehislik närvivõrk võib olla üle 10% täpsem kui seni kasutusel olnud otsustuspuu.
- Autor kirjutas üle aasta kestnud diplomitöö valmimise käigus üle 2000 rea C++-i, üle 5000 rea Python-it, üle 1000 rea Ruby-t, üle 300 rea Bash-i ja üle 500 rea HTML-i ja CSS-i.

Retsentsent PhD Toomas Kirt küsimused

- **Töö autor võiks kaitsmisel täpsemalt välja tuua, mis oli tema panus lõputöös toodud tulemuste saavutamisel?**

  Lõputöö autor programmeeris KBFI või CERN-i teadlaste poolt usaldatud ülesandeid, mis olid seotud andmekaevega osakestefüüsikas.

  Kõik töös loetletud tulemused olid saavutatud iseseisva programmeerimise tulemusel lähteülesande kirjelduse põhjal. KBFI või CERN-i teadlased kirjeldasid domeenivaldkonda ja valideerisid tulemusi.

Retsentsent PhD Toomas Kirt küsimused

- **Töö autor võiks selgitada, millise struktuuriga tehisnärvivõrku ta kasutas ja millised olid selle parameetrid?**

  Parima tulemuse andis pärilevivõrk (*feed forward network*), milles oli 4 peidetud kihti ja igas peidetud kihis 150 neuronit, ReLu aktivatsioonifunktsioon, puudus väljajätumeetodi kasutamine ja konstantne õpisamm oli 0.0005.

Eriti suur aitäh!

- Juhendaja ja kõik head inimesed KBFI-st ja CERN-ist.
- Õppeosakond.
- Õppejõud.
- Komisjon.