

Recenzja

Wpłynęło dnia ...2020-07-03...

rozprawy doktorskiej mgr Michała Olszewskiego p.t.
„Machine learning approach to $H \rightarrow \tau\tau$ analysis in the CMS experiment”

Recenzowana rozprawa dotyczy analizy przydatności podejścia opartego o uczenie maszynowe („machine learning”) w analizie fizycznej dotyczącej rekonstrukcji bozonu Higgsa w kanałach rozpadu zawierających lepton τ w eksperymencie CMS, realizowanym w CERN. Zaprezentowano szczegółowo wybrane metody uczenia maszynowego oraz przeanalizowano, w jaki sposób mogą zostać włączone w proces wybranej analizy fizycznej. Zaproponowano kryteria oceny przydatności tego podejścia, polegające na wyznaczeniu poziomu istotności sygnału. Wynikiem pracy jest po pierwsze wybór najlepszego narzędzia z palety zaliczanej do metod uczenia maszynowego – sieci neuronowej, po drugie identyfikacja konkretnego etapu procesu analizy w którym to narzędzie jest najbardziej użyteczne – klasyfikacja zdarzeń („events”) oraz po trzecie jakościowe i ilościowe określenie zysków z zastosowania tej metody. Przedstawione badania są unikatowe i prezentują użyteczność nowatorskiego podejścia w rekonstrukcji bozonu Higgsa. W tym sensie rozprawa jest istotnym wkładem do badań w eksperymencie CMS.

Rozprawa napisana jest w języku angielskim i podzielona jest na 6 rozdziałów wraz z podsumowaniem oraz 5 załączników („Appendix”). Pierwsze trzy rozdziały to część wprowadzająca i opisowa pracy, w pozostałych trzech zawarto opis analizy danych wykonanych w ramach pracy i dyskusję wyników tej analizy.

Rozdział 1 zawiera krótkie wprowadzenie do tematyki produkcji bozonu Higgsa. Przedstawione są podstawy Modelu Standardowego oraz łamanie symetrii z mechanizmie Brouta-Englerta-Higgsa. Opisane są kluczowe właściwości bozonu Higgsa, w tym mechanizmy jego produkcji, w szczególności w zderzeniach protonów przy energiach osiągniętych w LHC oraz główne kanały jego rozpadu. Podkreślono znaczenie kanału rozpadu na parę leptonów τ , oraz wyjaśniono dlaczego identyfikacja bozonów Higgsa właśnie w tym kanale została wybrana jako zagadnienie analizowane w rozprawie. Jednocześnie zidentyfikowano główne wyzwanie jakie stoi przez fizykami w tej analizie, to znaczy konieczność znacznej redukcji tła pochodzącego od procesów QCD. Na zakończenie przedstawione są istotne informacje na temat leptonu tau oraz jego kanałów rozpadu oraz głównych wyzwań związanych z jego pomiarem w detektorach ogólnego przeznaczenia takich jak CMS. Rozdział pokazuje, że Autor bardzo starannie przestudiował fizyczne podstawy prowadzonej analizy. Ich przedstawienie jest kompletne i merytorycznie poprawne, oraz wystarcza do zrozumienia dalszych części rozprawy. Na pochwałę zasługuje staranność w przedstawieniu formalizmu zagadnienia.

Rozdział 2 opisuje narzędzie badawcze wykorzystywane w rozprawie, to znaczy kompleks Wielkiego Zderzacza Hadronów (Large Hadron Collider – LHC) znajdujący się w laboratorium

CERN oraz działający przy nim eksperyment CMS (Compact Muon Solenoid). CMS to ogromna międzynarodowa współpraca badawcza, w której skład wchodzi grupa z Uniwersytetu Warszawskiego, w tym Autor rozprawy. Rozdział opisuje szczegółowo najważniejsze elementy układu detekcyjnego, to znaczy układ magnesów opartych o cewki nadprzewodzące, a także detektory pozwalające na rekonstrukcję trajektorii cząstek naładowanych („Tracker”), kalorymetry hadronowe i elektromagnetyczne pozwalające na pomiar całkowitej depozycji energii przez odpowiednie rodzaje cząstek oraz system śledzenia mionów. Całość zakończona jest opisem systemu wyzwalania („trigger”). Z jednej strony kompleks LHC oraz detektor CMS są w stanie dostarczać i rejestrować ogromne ilości zderzeń, z drugiej strony badane procesy są relatywnie rzadkie. Ogromny strumień danych musi więc zostać zredukowany do rozmiarów pozwalających na realistyczne magazynowanie i analizę tych danych przy jednoczesnym zachowaniu informacji o interesujących procesach. Stąd konieczność użycia zaawansowanego systemu wyzwalania rejestracji i selekcji danych. Autor rozprawy w sposób zwięzły przedstawia najważniejsze informacje o układzie detekcyjnym, w tym o jego właściwościach istotnych z punktu widzenia analizy przeprowadzonej w rozprawie, takich jak rozdzielczość przestrzenna i energetyczna, wydajność rejestracji i szczegółowe kryteria selekcji danych w systemie wyzwalania wysokiego poziomu (High Level Trigger – HLT). Zawartość rozdziału pokazuje, że Autor w pełni rozumie zalety i ograniczenia narzędzia badawczego jakim się posługuje, t.j. detektora CMS. Informacje przedstawione w rozdziale znacząco ułatwiają zrozumienie dalszych części rozprawy.

Rozdział 3 to szczegółowy opis zaawansowanej procedury rekonstrukcji obiektów fizycznych w detektorze CMS – tzw. „Particle Flow” (w skrócie PF). Pozwala ona na identyfikację pojedynczych cząstek oraz bardziej skomplikowanych obiektów takich jak dżety oraz pomiar ich fundamentalnych parametrów fizycznych, czyli składowych pędu oraz całkowitej energii. Po pełnej rekonstrukcji całego zdarzenia możliwe jest też wyznaczenie jego parametrów globalnych, w tym brakującej energii poprzecznej oraz brakującego pędu poprzecznego. Procedura ta wykorzystuje pełnię możliwości całego układu detekcyjnego CMS i w zaawansowany sposób łączy unikatowe aspekty każdego z podsystemów. Na podstawie zrekonstruowanych obiektów możliwe jest przeprowadzenie analiz wyższego rzędu, t.j. rekonstrukcji kandydatów na leptony tau i ostatecznie rekonstrukcję masy niezmienniczej pary leptonów tau w celu poszukiwania nadwyżki produkcji par w obszarze masy bozonu Higgsa. Opis jest szczegółowy i zawiera kluczowe merytorycznie informacje na temat algorytmów rekonstrukcji obiektów fizycznych w zderzeniach w CMS. Autor udowadnia, że w szczegółach rozumie całą procedurę, co jest niezbędne przy poprawnym definiowaniu warunków brzegowych i doborze kryteriów selekcji w procesach uczenia maszynowego opisanych w dalszej części pracy.

Rozdział 4 rozpoczyna część rozprawy, która opisuje wyniki uzyskane bezpośrednio przez Autora. Wykonał on niezależną implementację analizy fizycznej $H \rightarrow \tau\tau$ w eksperymencie CMS, opartą o dedykowaną próbkę danych symulowanych, zawierających mechanizmy produkcji i rozpadu bozonu Higgsa na leptony tau, odpowiadające im procesy stanowiące tło oraz pełną propagację wygenerowanych cząstek przez model układu detekcyjnego CMS. Procedura składa się z wielu powiązanych kroków, takich jak przygotowanie danych, zastosowanie odpowiednich korekcji dotyczących pojedynczych cząstek, złożonych obiektów takich jak dżety oraz samych zderzeń. Izolacja sygnału odbywa się niezależnie w trzech kategoriach zdarzeń, które zostały odpowiednio zdefiniowane. Istotną kwestią jest tu odpowiedni dobór dodatkowej zmiennej, innej

dla każdej kategorii zderzeń, która pozwala w sposób bardziej efektywny dokonać izolacji sygnału i redukcji tła. Dodano również korekcje i normalizacje niezbędne przy analizowaniu zderzeń symulowanych. Ostatecznie oszacowano poziom tła i dokonano identyfikacji sygnału pochodzącego od bozonów Higgsa. Wyniki analizy porównano z analizą referencyjną, t.j. oficjalną analizą eksperymentu CMS, która była podstawą publikacji, oraz analizą weryfikującą poprawność tej pierwszej. Wyniki analizy Autora okazały się doskonale zgodne z obiema „oficjalnymi” analizami CMS, co pozwala mieć zaufanie do dalszych wyników przedstawionych w tej rozprawie. Opis analizy oddaje ogrom pracy, który musiał zostać włożony w jej przygotowanie. Nie ulega też wątpliwości, że Autor sprawnie posługuje się narzędziami programistycznymi oraz zasobami obliczeniowymi CMS, a także świetnie rozumie fizyczne uwarunkowania analizy.

W rozdziale 5 rozpoczyna się opis meritum rozprawy, to znaczy zastosowania w analizie metod uczenia maszynowego. Po pierwsze zidentyfikowano konkretny etap analizy, w którym uczenie maszynowe może przynieść największą korzyść. Wyodrębniono 14 klas zdarzeń, ze względu na procesy fizyczne w nich zachodzące. Są to zarówno klasy zderzeń, w których doszło do interesującego nas rozpadu bozonu Higgsa, jak i zdarzenia zawierające jedynie procesy będące tłem dla tych rozpadów. Klasyfikacja każdego zdarzenia jest znana, ponieważ pochodzi ono z próbki symulowanej. Następnie stosowane są procedury uczenia maszynowego, których zadaniem jest poprawne zakwalifikowanie danego zdarzenia do odpowiadającej mu klasy. Proces klasyfikacji musi zostać poprzedzony procedurą „uczenia”, w której wykorzystywana jest znajomość prawdziwej klasy zdarzenia. Nietrywialny jest też wybór poprawnego kryterium, które pozwoli ocenić efektywność metody. Wybrana została uśredniona wartość pola pod krzywą ROC (ROC AUC). Autor zaimplementował i przetestował cztery metody uczenia maszynowego: „Decision Tree”, „Random Forest”, „Gradient Boosted Decision Trees” oraz „Deep Learning” oparte o sieci neuronowe. Analiza wykazała, że metoda DL, czyli sieć neuronowa okazała się najlepsza, to znaczy osiągnęła najwyższy wskaźnik ROC AUC równy 0.875. Rozdział ten jest bardzo dobrym wprowadzeniem do tematyki uczenia maszynowego. Zawiera szczegółowy opis zastosowania czterech różnych technik do dobrze zdefiniowanego problemu fizycznego. Implementacja każdej z technik zależna jest od szeregu dodatkowych parametrów, których wpływ na ostateczny rezultat także został zbadany. Dzięki poprawnemu wyborowi kryterium poprawności wyniku te cztery bardzo różne metody mogły zostać poddane ilościowej ocenie. Rozdział daje przekonanie, że Autor w szczegółach opanował problem uczenia maszynowego, jego programistycznej i sprzętowej implementacji oraz jego adekwatności do konkretnej analizy fizycznej eksperymentu CMS.

Rozdział 6 kończy część poświęconą opisowi analiz wykonanych przez Autora. Metoda, która okazała się najbardziej skuteczna w rozdziale 5 zastosowana jest do wybranego zadania, w sposób który odpowiada analizie danych rzeczywistych. Deep Learning zastępuje tu standardową analizę opartą o kryteria wyboru (ang. „cuts”). Jej końcowy wynik (ang. „score”) jest używany zamiast tych właśnie kryteriów. Porównano wyniki poziomu istotności sygnału dla obu podejść do analizy. Analiza wykonana przez Autora, oparta o techniki uczenia maszynowego w każdej z trzech kategorii zderzeń pozwala na poprawę wyników od 20% do 35%, przy zachowaniu porównywalnej lub lepszej niepewności systematycznej.

Podsumowanie analiz, zawarte w rozdziale 7 pokazuje jasno, że cel postawiony w pracy

został osiągnięty. Przetestowano szereg technik uczenia maszynowego, co samo w sobie było ogromnym wyzwaniem technicznym wymagającym dogłębnej wiedzy programistycznej i fizycznej. Nie polegało to jedynie na schematycznym zaimplementowaniu wybranych algorytmów, ale na pracy koncepcyjnej, wymagającej dokonania nietrywialnej identyfikacji konkretnych elementów i etapów analizy, w których uczenie maszynowe może w pełni ujawnić swoje zalety. Zdecydowano się na użycie metody Deep Learning jako procedury zastępującej standardową technikę opartą o kryteria selekcji. Trafność wyboru potwierdzona jest wynikami ilościowymi, t.j. poprawą od 20% do 35% poziomu istotności wyniku w analizie rekonstrukcji bozonu Higgsa.

Rozprawę uzupełnia kompletna biografia, składająca się ze 168 pozycji oraz 5 dodatków („Appendices”). W Dodatku A podano zestawy symulacji Monte-Carlo użyte w rozprawie, chociaż ze względu na to, że w wykazie Autor posługuje się wewnętrzną nomenklaturą kolaboracji CMS są to dane w dużej mierze nieczytelne dla postronnego czytelnika. Dodatek B podaje techniczne informacje dotyczące rozdziału 4, ale niestety cierpi na podobną nieczytelność jak dodatek pierwszy. Oficjalne wyniki analizy CMS $H \rightarrow \tau\tau$ podaje Dodatek C. Ich załączenie istotnie uzupełnia dyskusję w rozprawie. Dodatek D podaje rozwinięcie wybranych skrótów w rozprawie, choć wydaje się, że zawiera pewne nieścisłości, np. w definicji „mjj” pojawia się wyrażenie „njets<0”, choć zapewne powinno być „njets<1”, lub nawet „njets<2”, a wyrażenie „mat_pt” zapewne powinno brzmieć „met_pt”. Dodatek E podaje bardzo użyteczne formalne definicje statystyczne.

Rozprawa doktorska opisuje bardzo skomplikowaną i zaawansowaną analizę, zarówno pod względem znaczenia fizycznego jak i zastosowanych technik uczenia maszynowego. Otrzymany wynik jest zachęcający, jednak z samej rozprawy, nie wynika bezpośrednio jakie są perspektywy zastosowania tej techniki w oficjalnej publikacji eksperymentu CMS. Wspomniana jest planowana publikacja publicznej noty, nie jest jednak wyjaśnione, czy analiza prezentowana w tej rozprawie będzie wchodzić w jej skład. Pewnym ograniczeniem prezentowanej rozprawy jest fakt, że ostateczna analiza nie była przeprowadzona na rzeczywistych danych eksperymentalnych. Dopiero taka procedura byłaby ostatecznym testem na prawidłowość tego podejścia. Szczegółowy opis techniczny procedur niezbędnych w analizie pokazuje, że do jej przeprowadzenia konieczne jest ustalenie wielu wolnych parametrów i jest ona zależna w nietrywialny sposób od aktualnej konfiguracji detektora, szczegółów implementacji skomplikowanej procedury rekonstrukcji „Particle Flow”, a także od zewnętrznych uwarunkowań, takich jak energia i świetlna wiązka w LHC. Uczenie maszynowe jest często czułe na konkretny zestaw danych użytych do „uczenia” czy też „trenowania” procedury. Biorąc pod uwagę, że parametry o których mowa powyżej na pewno będą się dynamicznie zmieniać przez lata eksploatacji detektora CMS, pojawia się pytanie o to, jak odporna na te zmiany, czyli jak uniwersalna będzie strategia analizy oparta o uczenie maszynowe. Użyteczne w oszacowaniu takiej uniwersalności byłoby zastosowanie algorytmu „nauczonego” na danych Monte-Carlo z wybranego okresu działalności CMS (np. roku zbierania danych, czy konkretnego Runu LHC) do danych Monte-Carlo z innego okresu i/lub do danych rzeczywistych.

Praca jest przygotowana starannie od strony edytorskiej. Rysunki są duże i czytelne oraz jasno opisane. Pewne zastrzeżenie budzi niekonsekwencja w stosowaniu numeracji równań (niektóre równania nie są opatrzone numerami, co utrudnia odniesienie się do nich). W pracy pojawiło się kilka drobnych błędów językowych, jednak wszystkie te uwagi nie wpływają

negatywnie na odbiór pracy.

Rozprawa przedstawia unikatowe, nietrywialne wyniki analizy rozpadu bozonu Higgsa na leptony tau, czyli zagadnienia niezwykle istotnego i aktualnego w fizyce. Odważnie proponuje zastosowanie nowatorskich metod analizy opartych o najnowsze osiągnięcia w dziedzinie uczenia maszynowego, które samo w sobie jest zagadnieniem aktualnym i interesującym. Całość prac wykonana została we współpracy międzynarodowej w wymagającym środowisku naukowym, jakim jest kolaboracja CMS. Podsumowując uważam, że recenzowana rozprawa mgr Michała Olszewskiego spełnia niezbędne wymagania stawiane przed rozprawami doktorskimi i wnoszę o dopuszczenie Autora do dalszych etapów przewodu doktorskiego.



Adam Kisiel