Prof. dr hab. inż. Ewa Skubalska-Rafajłowicz

Katedra Informatyki Technicznej

Wydział Informatyki i Telekomunikacji

Politechnika Wrocławska

Recenzja rozprawy doktorskiej

mgr inż Grzegorza Gołaszewskiego

“Wykrywanie śladów cząstek długożyciowych w eksperymencie LHCb (CERN, Genewa) metodami analizy danych i inteligencji obliczeniowej”

Informacje ogólne

Niniejsza recenzja została napisana na zlecenie Rady Naukowej Instytutu Badań Systemowych PAN w związku z toczącym się przewodem doktorskim Pana mgr inż. Grzegorza Gołaszewskiego.

Przewód doktorski otwarty został w trybie interdyscyplinarnym w dyscyplinach, których aktualne nazwy to: „Informatyka Techniczna i Telekomunikacja” oraz „Nauki Fizyczne”. Promotorami w tym przewodzie byli:

1) Prof. dr hab. inż. Piotr Kulczycki, Instytut Badań Systemowych, Polska Akademia Nauk, Akademia Górniczo-Hutnicza, Wydział Fizyki i Informatyki Stosowanej

2) Dr hab. inż. Tomasz Szumlak, prof. AGH, Akademia Górniczo-Hutnicza, Wydział Fizyki i Informatyki Stosowanej

3) Promotor pomocniczy: dr hab. inż. Szymon Łukasik, prof. AGH, Instytut Badań Systemowych, Polska Akademia Nauk, Akademia Górniczo-Hutnicza, Wydział Fizyki i Informatyki Stosowanej.

Ze względu na zakres moich kompetencji, skupię się głównie na aspektach informatycznych recenzowanej rozprawy, traktując opisane w niej problemy powstające w badaniu cząstek elementarnych jako źródło unikatowych danych i inspiracji do opracowywania

metod i algorytmów oraz jako obszar zastosowań, który jest fundamentalny dla rozwoju nauki.

Tematyka rozprawy

Recenzowana rozprawa dotyczy opracowania metod i algorytmów, których zadaniem jest zwiększenie precyzji przetwarzania danych pochodzących z eksperymentów, prowadzących do rozpadu cząstek o długim czasie życia. Jednym z istotnych zadań rozważanych w rozprawie jest redukcja liczby tak zwanych „fałszywych trajektorii", które Autor nazywa także „duchami". Powstają one na wcześniejszych etapach obróbki danych eksperymentalnych, których efekt jest traktowany w tej pracy jako dane wejściowe. Z drugiej strony, redukcja liczby „fałszywych trajektorii" ma istotne znaczenie dla wniosków o stanach materii, które formułują fizycy. Dodatkowym utrudnieniem jest fakt, że – z przyczyn zasadniczych – nie są dostępne dane uczące z podziałem na trajektorie właściwe i fałszywe. Jest to zatem wyzwanie zarówno na etapie uczenia jak i na etapie weryfikacji wyników.

Jako narzędzia do rozwiązania problemu mgr G. Gołaszewski wybrał metody statystycznej analizy dużych zbiorów danych i sztucznej inteligencji.

Tematykę rozprawy uważam za aktualną i ważną nie tylko dla fizyki cząstek elementarnych o długich czasach życia ale także dla informatyki, gdyż proponowane w rozprawie metody, po sformułowaniu ich w nieco ogólniejszej formie, mogą mieć szersze zastosowania.

Zawartość i kompozycja rozprawy

Rozprawa liczy 88 stron i składa się z Przedmowy, 7 rozdziałów oraz bibliografii, liczącej 66 pozycji. Przedmowa i rozdział wstępny zawierają sformułowanie celu rozprawy, szkic podejścia do rozwiązania problemu i osadzenie go w aktualnych realiach eksperymentów prowadzonych w CERNie. W szczególności, Doktorant wskazuje miejsce w całym, bardzo złożonym, oprogramowaniu CERNu, w którym opracowane przez Niego algorytmy mogą mieć zastosowanie.

Dokładniejszy opis zawartości oprogramowania znajdujemy w Rozdziale 2, który zawiera także wprowadzenie w tematykę eksperymentów prowadzonych w CERNie nad cząstkami o długich czasach życia. Równocześnie, w rozdziale tym Autor wskazuje te miejsca, w których zbierane są dane o trajektoriach cząstek. Dane te podlegają dalszemu przetwarzaniu, w tym potencjalnie za pomocą omawianych w rozprawie metod. Rozdział ten jest potrzebny, aby czytelnik mógł umiejscowić źródła trudności, które prowadzą do powstawania tzw. „fałszywych trajektorii".

W Rozdziale 3 mgr G. Gołaszewski opisuje najważniejsze narzędzia, które są dalej składowymi

proponowanych przez Niego metod. I tak, w podrozdziale 3.1 opisany został estymator gęstości rozkładu prawdopodobieństwa w wersji jednowymiarowej i oparty na jego całkowaniu estymator dystrybuanty. Przedstawiono także wielowymiarową wersję estymatora gęstości w wersji z diagonalnymi macierzami formy kwadratowej i jądrem gaussowskim, co prowadzi do algorytmu obliczeniowego o akceptowalnym nakładzie obliczeń. Pełny opis tego estymatora Autor podał w podrozdziale 4.3.1, natomiast opis dany wzorem (6) ma charakter poglądowy. Tak na marginesie, we wzorze (6) zapewne warto usunąć uproszczoną lewą stronę, bo niczego nie wyjaśnia, a formalnie nie jest poprawna.

Następnie, Doktorant skupia się na kluczowym problemie doboru parametru(-ów) wygładzania i opisuje wersję metody „wstawiania" (plug-in), dostosowaną do jąder typu gaussowskiego. Wybór ten jest trafny, gdyż podejście to w tej wersji cechuje relatywnie niewielki nakład obliczeń. Jest to istotne, gdyż dalej Autor opisuje próbę modyfikacji parametrów wygładzania poprzez zastosowanie mnożników, które czynią dobór parametrów wygładzania doborem lokalnym, tak by z różniącymi się parametrami wygładzania uwzględnić obszary gdzie gęstość jest bardziej „płaska" (np. w tzw „ogonach" rozkładów) oraz obszary gdzie wyjściowy estymator wykazuje większą zmienność. Szkoda, że w tym miejscu Autor nie rozważał także innych podejść stosowanych w lokalnym doborze parametrów wygładzania. Celem (i globalnym i lokalnym) jest tu zachowanie kompromisu pomiędzy obciążeniem, a wariancją estymatora. W tym kontekście, pożyteczne byłoby zbadanie zaproponowanej w pracy metody „mnożników".

Dyskusyjne jest proponowane w pracy ograniczenie nośników funkcji gaussowskich formujących estymator bez zastosowania normalizacji. Uzasadnieniem są względy obliczeniowe i nieduża odległość całki z estymatora od jedynki.

W podrozdziale 3.2 Doktorant opisuje sieć neuronową o strukturze enkodera i algorytm jej uczenia. Na uwagę zasługuje sposób wykorzystania nauczonego enkodera. Jednym z zastosowań autoenkoderów jest wykrywanie anomalii (czy też obiektów „odstających"). Zwykle niedopasowanie danych oceniane jest na podstawie błędu rekonstrukcji. W ocenianej rozprawie także zaproponowana sieć typu autoencoder ma służyć jako detektor do wykrywania elementów nietypowych w stosunku do tych, które pojawiały się, jako typowe, w ciągu uczącym. Mgr Gołaszewski proponuje stosować dalej w rozprawie tzw. rzadką (*sparse*) wersję enkodera.

Aby uzyskać cechę rzadkości w warstwie kodującej enkodera i jednocześnie zapewnić, by wagi sieci nie były nadmiernie duże, Autor wybrał rozwiązanie, w którym kryterium uczenia składa się z błędu średniokwadratowego rekonstrukcji wzorców przez enkoder , regularyzacji wag typu L_2 oraz dywergencji Kulbacka-Leiblera. W rozwiązaniach tych warstwa kodująca (wewnętrzna) zawiera dużo neuronów, a poprzez zastosowanie KL dywergencji wymusza się, by dla danego

wejścia aktywowane były tylko nieliczne z neuronów kodujących (aktywacja większości jest równa zero). W mojej ocenie wybór metody jest bardzo dobry, choć w pracy brakuje cytowań dotyczących zastosowania KL dywergencji w wymuszaniu rzadkości auto-enkoderów, a jest ich sporo. Podobnie, cytowanie pozycji [62] w odniesieniu do dywergencji Kulbacka-Leiblera, nie jest wnikliwie dobrane.

Oryginalnym rozwiązaniem jest w mojej ocenie zaproponowana w pracy metoda wykrywania elementów nietypowych, która dalej stosowana jest do wykrywania trajektorii „duchów". Doktorant przedstawił podejście polegające na połączeniu obliczania wartości błędu średniokwadratowego enkodera z obliczaniem wartości estymatora gęstości rozkładu prawdopodobieństwa napotkania prawdziwej trajektorii jako podstawy do obliczenia współczynnika nietypowości. Współczynnik ten zdefiniowany jest jednak dopiero w Rozdziale 4. Jako sposób łącznego uwzględnienia obu ocen mgr G. Gołaszewski proponuje zastosowanie podejścia rozmytego do podejmowania decyzji o nietypowości obserwacji. W tym kontekście Autor omawia różne przykłady t-norm, a do dalszych obliczeń stosuje min(a,b). Podejście to jest istotnym wkładem Autora i w następnych rozdziałach jest ono badane oraz zastosowane do zagadnienia wykrywania fałszywych trajektorii cząstek elementarnych o długich czasach życia.

Rozdział 4 Doktorant rozpoczyna od bardzo przydatnego schematu (Rys. 18), który opisuje ogólny schemat obliczeń i proponowaną logikę wnioskowania, w oparciu o wskazane wyżej podejście. Na Rys. 18 pojawia się także blok parowania geometrycznego. Blok parowania jest już algorytmem dedykowanym do zastosowań w analizie danych w fizyce cząstek. Z punktu widzenia innych zastosowań wystarczy blok ten zastąpić właściwym algorytmem dedykowanym.

W podrozdziale 4.1 Autor opisuje proces konstrukcji pojedynczych śladów cząstek. Proces ten opiera się na znajomości praw fizyki i konkretnej aparatury pomiarowej. Jest to procedura opracowana wcześniej przez innych autorów i opisana w pracy [12]. Warto zaznaczyć, że algorytm ten pozwalał wstępnie usunąć takie ślady, które nie spełniają ograniczeń fizycznych. Ponadto, każdy ślad poddawany jest wstępnej weryfikacji za pomocą jednowarstwowego perceptronu o 15 neuronach w warstwie ukrytej.

Opisany w podrozdziale 4.2 proces tzw. parowania śladów prowadzi do dalszej, znacznej redukcji liczby fałszywych śladów. Także ten proces bazuje na prawach fizyki cząstek, ale jego opracowanie i oprogramowanie jest wkładem Autora.

Podrozdział 4.3 zawiera szczegółowy opis głównej koncepcji rozprawy w zastosowaniu do śladów par cząstek, które pierwotnie opisane są jako wektorami o 16 składowych. Jako etap wstępny Doktorant proponuje zastosowanie procedury analizy składowych głównych, co pozwala mu zredukować rozmiar tego opisu do 12 składowych. Następnie, szczegółowo opisany jest sposób przekształcania wartości estymat, otrzymanych za pomocą estymatora jądrowego, w rozmytą miarę przynależności. Podobnie, w miarę przynależności przekształcany jest wynik

zastosowania enkodera. Na tej podstawie tworzona jest funkcja decyzyjna, będąca t-normą powyższych wielkości, która porównywana jest z empirycznie wyznaczonym progiem.

W krótkim, ale ważnym Rozdziale 5 Doktorant przedstawia miary jakości, które następnie będą używane do oceny jakości proponowanego podejścia. Wobec fundamentalnej trudności braku sklasyfikowanych danych empiryczny, Autor proponuje badanie wydajności rekonstrukcji i ułamek udziału „duchów" za pomocą techniki Monte Carlo, która – w rym przypadku – wymagała opracowania algorytmu symulacji przypadkowych połączeń w pary obserwacji, które nie są ze sobą związane. Trzeci miernik, bazujący na rozkładach mas cząstek może być stosowany do danych rzeczywistych. W rozprawie jest on traktowany jako miernik specyficzny dla rozpatrywanego zastosowania. Warto jednak odnotować jego potencjalnie szersze zastosowania wszędzie tam, gdzie mamy do czynienia z obserwacjami obiektów o dającej się zmierzyć charakterystyce, która obserwowana jest na tle obiektów, dla których ta charakterystyka ma inną wartość. Dodatkowo, potrzebna jest znajomość rozkładów obiektów o różnych wartościach tej charakterystyki.

Rozdział 6 zawiera wyniki bardzo pracochłonnych badań skuteczności proponowanej metodyki w zastosowaniu do cząstek o długich czasach życia. Wyniki te są głównie interesujące dla fizyków, ale warto spojrzeć na nie także z punktu widzenia obliczeniowego.

Wspomniany przez mnie, przy omawianiu zawartości Rozdziału 5, trzeci miernik oceny jakości wymaga znajomości odpowiednich rozkładów prawdopodobieństw. W podrozdziale 6.1 mgr G. Gołaszewski omawia algorytm estymacji parametrów rodziny rozkładów dla danych empirycznych w postaci histogramów mas cząstek. Algorytm ten bazuje na doborze parametrów rozkładów poprzez minimalizację błędu średniokwadratowego między prawdopodobieństwami empirycznymi i teoretycznymi prawdopodobieństwami wpadania w hiperkostki wyznaczone przez wielowymiarową siatkę. Ze względu na dużą liczbę parametrów rodzin dopasowywanych rozkładów optymalizacja poprzez przegląd zupełny nie jest możliwa. Doktorant proponuje naturalną heurystykę, która polega na minimalizacji błędu najpierw na grubej siatce, której krok jest następnie redukowany, ale dalsze poszukiwania zawężane są tylko do okolic najlepszego dotąd znalezionego zestawu parametrów. Jasne jest, że taka zachłanna procedura nie gwarantuje znalezienia globalnego optimum w ogólnym przypadku. Jednakże, w omawianym zastosowaniu procedura ta dawała zadowalające rezultaty przy zachowaniu dopuszczalnych w praktyce czasów obliczeń.

W dalszej części Rozdziału 6 Autor przedstawia wyniki testowania procedur składających się na całość proponowanego podejścia, w podrozdziale 6.5 znajdujemy wyniki testów i porównań całościowych. Jako punkt odniesienia mgr G. Gołaszewski przyjmuje wyniki dostarczane przez algorytm *PrLongLivedTracking*, który jest stosowany w eksperymencie LHCb w CERN. Są one

porównywane z osobno stosowanymi algorytmami cząstkowymi:

a) odrzucania par, które nie spełniają ograniczeń fizycznych,

b) oceny nietypowości za pomocą enkodera,

c) oceny za pomocą wartości estymatora jądrowego

oraz z kombinacją w/w subalgorytmów. Do porównań stosowane są mierniki opisane w Rozdziale 5, a rezultaty przedstawiono w czterech tabelach. Autor podsumowuje uzyskane wyniki następująco: *„Zastosowanie całego kompletu procedur pozwala na uzyskanie największej redukcji śladów duchów, sięgającej 66 procentom śladów duchów identyfikowanych przez algorytm PrLongLivedTracking, przy utracie na wydajności rekonstrukcji około 24%. Należy podkreślić, że w analizach wymagających wysokiej czystości próby, czyli małego udziału duchów, polepszenie wyników w zakresie redukcji śladów duchów czyni taką utratę akceptowalną."* Analiza w/w tabel potwierdza tę, kluczową dla rozprawy, konkluzję.

Podobnie, w podrozdziale 6.6 Doktorant zastosował metodę porównań na podstawie rozkładów mas. Również to porównanie wskazuje na istotną poprawę, uzyskaną dzięki proponowanej przez Autora metodyce. Mianowicie, *„…już dla samego parowania poprawa wynosi 17,84%, natomiast przy* pełnej *filtracji 28,36%*." Odnotować trzeba rzetelność badań mgr G. Gołaszewskiego, który jednocześnie stwierdza, że dla barionów poprawa jest nieco mniejsza i uzasadnia ten fakt relatywnie mniejszym poziomem tła niż w poprzednim przypadku.

Rozprawę zamyka krótki Rozdział 7, zawierający podsumowanie uzyskanych rezultatów i licząca 66 pozycji bibliografia. Autor informuje w nich, że proponowana metodyka została zaimplementowana jako oprogramowanie udostępnione w *gitlab* CERNu (poprzez modyfikację istniejącej procedury). Ponadto, *„Metoda ta została zrealizowana w ramach projektu OPUS NCBiR i obecnie jest wdrażana do środowiska oprogramowania eksperymentu LHCb w Genewie."* (str. 80) a bibliografia zawiera też dane dwóch rozdziałów, w monografiach wydanych przez wyd. Springer, autorstwa lub współautorstwa mgr G. Gołaszewskiego.

Najważniejsze osiągnięcia

Z punktu widzenia wkładu w informatykę, za najważniejsze osiągnięcie Doktoranta uważam zaproponowanie i symulacyjne oraz empiryczne zbadanie metodyki uczenia wykrywania nietypowych obiektów o strukturze wektorowej. Oryginalną cechą tej metodyki jest podwójne, równoległe, testowanie statystyczne w oparciu o uczenie enkodera i jądrowego estymatora gęstości rozkładu prawdopodobieństwa cech badanych obiektów. Istotnymi, oryginalnymi elementami tej metodyki jest opracowanie wskaźników niepodobieństwa na podstawie błędu rekonstrukcji enkodera i wartości estymaty prawdopodobieństwa „bycia prawdziwym śladem", a

następnie skomponowanie z nich decyzji, w oparciu o t-normę.

Na podstawie obiecujących wyników badań eksperymentalnych i faktu, że w CERN prowadzone są prace nad wdrożeniem proponowanej metodyki, spodziewać się można iż recenzowana rozprawa będzie mieć także wpływ na zwiększenie dokładności badań nad cząstkami elementarnymi o długich czasach życia.

Można zatem stwierdzić, że interdyscyplinarny charakter badań nad tą rozprawą spełnił swoją rolę gdyż zagadnienie z zakresu badania cząstek elementarnych stanowiło z jednej strony inspirację do stworzenia ciekawej metodyki wykrywania obiektów nietypowych, a z drugiej dostarczyło ważnych danych do oceny jej skuteczności.

Uwagi

1. Przed opublikowaniem ostatecznej wersji rozprawy warto by Autor przejrzał ją ponownie pod kątem ułatwienia czytania. Przykładowo, w dwóch miejscach Doktorant najpierw przedstawia wzór, a dopiero potem znajdujemy objaśnienia zawartych w nim symboli.

2. W badaniach symulacyjnych nad doborem progów dla wskaźników niepodobieństwa mgr G. Gołaszewski odwołuje się do poziomu istotności równego 0.05. Czy zdaniem Doktoranta przywołanie np. metody Bonferoniego doboru progów w przypadku testowania złożonych hipotez mogłoby stanowić wskazówkę do oceny poziomu istotności całej procedury ?

3. W rozprawie rozważany jest problem *„efektywności wykrywania i usuwania śladów duchów …. przy akceptowalnej utracie wydajności rekonstrukcji”.* Czy Autor planuje szersze badania nad tym zagadnieniem ?

4. W pracy brakuje jakiejkolwiek dyskusji definicji funkcji przynależności wyznaczanej wzorami (68 -69).

Cel postawiony we Wstępie do rozprawy został osiągnięty, a proponowane podejście zostało zweryfikowane na danych empirycznych.

Realizacja tego celu wymagała od Autora dużej wiedzy z zakresu badania cząstek elementarnych oraz z wielu obszarów informatyki, w tym: znajomości metod sztucznej inteligencji stosowanych w analizie danych oraz umiejętności ich implementacji.

KONKLUZJA

Podsumowując całokształt rozprawy doktorskiej mgr inż. Grzegorza Gołaszewskiego stwierdzam, że rozprawa ta jest wartościowa i wnosi wkład do badań nad metodami wykrywania obiektów nietypowych oraz zastosowań w badaniach śladów cząstek elementarnych o długich czasach życia.

W związku z tym stwierdzam, że rozprawa ta spełnia warunki i wymagania stawiane rozprawom doktorskim, określone w Ustawie z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki (Dz.U. Nr 65 z 2003 poz. 595 z późn. zm.) i wnoszę o dopuszczenie jej do publicznej obrony.

Prof. dr hab. inż. Ewa Skubalska-Rafajłowicz

Wrocław 10 pazdziernika 2022 roku

E. Skubalska-Rafajłowicz