Recenzja rozprawy doktorskiej

mgr. Kornela Chromińskiego pt. Zastosowanie procesów epigenetycznych w algorytmach genetycznych

Promotor: dr hab. Mariusz Boryczka, promotor pomocniczy: dr Magdalena Tkacz

1. Zagadnienia naukowe rozpatrywane w pracy.

Przedmiotem rozprawy jest opracowanie oryginalnych modyfikacji standardowego algorytmu genetycznego oraz analiza wpływu zastosowania tych modyfikacji na zwiększenie efektywności algorytmu na przykładzie kilku wybranych zadań. Biorąc pod uwagę, że algorytmy genetyczne są coraz powszechniej wykorzystywane do numerycznego rozwiązywania problemów optymalizacyjnych w różnych dziedzinach nauki, a podstawowym ich wadą jest duża liczba wykonywanych obliczeń, należy stwierdzić, że rozpatrywane w pracy problemy są aktualne i dobrze wpisują się w nurt obecnie prowadzonych prac naukowych. Główny nurt rozprawy mieści się w dyscyplinie informatyka, poruszając jednocześnie (w niewielkim stopniu) pewne zagadnienia z pogranicza matematyki i biologii. Praca łączy w sobie aspekty teoretyczne (opracowanie modyfikacji algorytmów na podstawie poznanych mechanizmów biologicznych), aplikacyjne (implementacja algorytmów) oraz analizy danych.

We wstępie do rozprawy została postawiona teza, mówiąca że „zastosowanie mechanizmów wzorowanych na procesach epigenetyki pozwala na poprawę efektywności działania algorytmów genetycznych”. Uważam, że wyniki przedstawione w rozprawie pozwalają stwierdzić, że teza została udowodniona. Można, co prawda, dyskutować, czy rzeczywiście wszystkie modyfikacje mają swoje odwzorowania w procesach epigenetycznych, o czym piszę w dalszej części recenzji, ale co najmniej dwie z nich ten warunek spełniają, a poprawa efektywności w wybranych przykładach jest wyraźna. Szczegółowe cele rozprawy, również wymienione jawnie we Wstępie, zostały osiągnięte.

2. Struktura pracy

Rozprawa składa się ze wstępu, czterech numerowanych rozdziałów, podsumowania oraz spisu literatury. Jak już wspomniałem powyżej, we Wstępie postawiono tezę i zdefiniowano cele szczegółowe, a także opisano układ rozprawy. W Rozdziale 1, będącym wprowadzeniem w tematykę pracy, Autor przedstawił uproszczoną charakterystykę problemów optymalizacyjnych oraz podstawowe informacje na temat algorytmów ewolucyjnych, a w szczególności algorytmów genetycznych. Rozdział 2 został poświęcony opisowi czterech problemów optymalizacji (tu powinno pojawić się pewne zastrzeżenie, dotyczące tego, kiedy poszukiwanie wartości odstających
jest rzeczywiście problemem optymalizacyjnym), jakie posłużyły do testowania opracowanych modyfikacji algorytmu genetycznego. Najważniejszą częścią pracy są Rozdziały 3 i 4. W pierwszym z nich Doktorant przedstawił opis trzech proponowanych modyfikacji podstawowego algorytmu genetycznego, próbując znaleźć dla nich odniesienie w procesach epigenetycznych. Należy podkreślić, że nie są to trzy warianty jednego pomysłu, ale każda z tych modyfikacji odnosi się do innego aspektu podstawowego algorytmu genetycznego. W Rozdziale 4 Doktorant przedstawił najpierw ogólny schemat przeprowadzonych eksperymentów numerycznych, a następnie, w uporządkowany sposób, wyniki tych eksperymentów, starając się pokazać, na ile poszczególne modyfikacje przynoszą pożądany efekt dla poszczególnych rozpatrywanych przykładów zadań optymalizacji. Rozdział „Podsumowanie i wnioski”, jak łatwo się domyślić, stanowi rekapitulację wyników przedstawionych w Rozdziale 4 oraz zawiera opis możliwych dalszych prac naukowych nad podobnymi zagadnieniami.

3. Analiza źródeł

Spis literature zawiera 186 pozycji, z czego dwie ([40] i [41]) są współautorstwa Doktoranta. W zdecydowanej większości są to prace bezpośrednio związane z tematyką rozprawy, zarówno klasyczne, jak i najnowsze, ale w bibliografii można też znaleźć kilka pozycji czysto biologicznych lub medycznych. Poza kilkoma (np. [35] – tekst nieopublikowany, [32] i kilka podobnych – prace w czasopismach o niewielkim znaczeniu, trudno weryfikowalne) są to publikacje dobrze wybrane. Tak więc można stwierdzić, że dobór bibliografii świadczy o dobrym rozeznaniu Doktoranta w literaturze światowej w tematyce, którą się zajmuje.

4. Oryginalność i silne strony rozprawy.

Wszystkie modyfikacje, zaproponowane przez Doktoranta są oryginalnymi pomysłami, mającymi swoje źródło w procesach biologicznych/biochemicznych. Silną stroną rozprawy jest nie tylko ich opracowanie, ale również przetestowanie na kilku znacząco różnych problemach optymalizacyjnych, dla którego z których wykorzystano dodatkowo różne zbiory danych/parametrów testowych. Tak więc wnioski końcowe, dotyczące prawdziwości postawionej hipotezy badawczej, zostały wynurzone na podstawie kompleksowych badań, a nie tylko pojedynczych przypadków (nawet jeśli samo opracowanie ich wyników pozostawia wiele do życzenia, o czym mowa w następnym punkcie recenzji).

5. Słabe strony rozprawy

Do słabych stron pracy należą:

- Brak formalnej analizy statystycznej wyników eksperymentów numerycznych oraz niekonsekwencja w wyborze porównywanych wielkości (raz jest to mediana z dziewięciu, innym razem średnia ze stu powtórzeń, w dodatku bez zwrócenia uwagi na przedziały ufności przy średniej)
- Brak pogłębionej analizy wyników – z czego mogą wynikać zaobserwowane różnice dla różnych zbiorów danych, jako jest ostateczna poprawa efektywności przy zastosowaniu poszczególnych modyfikacji (patrz uwaga 6.4)
- Bardzo uproszczone traktowanie problemu optymalizacji – począwszy od próby zdefiniowania po optymalizacji parametrów algorytmu
- Nie do końca trafnie dobrane analogie pomiędzy proponowanymi modyfikacjami algorytmu genetycznego a procesami epigenetycznymi

Dwie ostatnie uwagi nie są bezpośrednio związane z dyscypliną Informatyka, co umniejsza ich wagę. Brak formalnej, poprawnej analizy statystycznej otrzymanych wyników jest poważniejszym
zarzutem – jednakże na pierwszy rzut oka wyraźnie widać, że w zdecydowanej większości przypadków taka analiza potwierdziłaby słuszność wysunutych wniosków.

6. Szczegółowe uwagi merytoryczne i redakcyjne

W pracy pojawiały się pewne niejasności oraz niedomówienia, istotne w tym sensie, że Doktorant powinien się do nich odnieść w trakcie obrony (zwłaszcza 6.2-6.8). W szczególności:

6.1. Jedynie w jednym z czterech rozpatrywanych zadań (grupowanie) Doktorant pokusił się o porównanie samego rozwiązania zadania optymalizacji z rozwiązaniami otrzymanymi za pomocą innych metod. Powinno to być zrobione dla każdego z zadań.

6.2. O ile pierwsza z zaproponowanych modyfikacji niewątpliwie jest skuteczna, trudno mi się zgodzić z tym, że odwzorowuje dziedziczenie za pomocą prionu, a nawet w ogóle, że odzwierciedla ona procesy epigenetyczne.

Po pierwsze, dziedziczenie poprzez priony (w liczbie mnogiej) ma związek z podziałem zawartości komórki macierzystej (w tym prionów, które mogą zmieniać strukturę przestrzenną sąsiednich białek na nieprawidłową) pomiędzy komórki potomne. Rzeczywiście taki proces miałby charakter epigenetyczny. Jednakże w układzie biologicznym sekwencja zawarta w DNA nie ma zwiąiku z prawidłowym czy też nieprawidłowym faldoaniem – oba typy białek są kodowane przez tę samą sekwencję. W zaproponowanej modyfikacji wyraźnie wbudowana jest w chromosom dodatkowa sekwencja, mająca charakter dominujący. Wydaje się, że biologiczny odpowiednikiem przedstawionej modyfikacji jest uwzględnienie dodatkowego genu (który mógłby mieć charakter onkogenu), który, jeśli występuje w odpowiedniej formie (przeszedł pewną mutację), wpływa na funkcję przystosowania, a jeśli nie jest zmutowany, to takiego wpływu nie ma. W takim ujęciu nadal byłaby to nadal oryginalna modyfikacja kluczowego algorytmu, a jedynie odwołanie do natury byłoby inne.

6.3. Nie do końca jest jasne, czym jest jedna z operacji w trzeciej z proponowanych modyfikacji, zastąpienie sekwencji, różni się od modyfikacji zaproponowanej w drugim z algorytmów

6.4. Metodyka przeprowadzania eksperymentów – interpretacja zmiany w liczbie generacji i czasu potrzebnego na jedną generację

Badania były przeprowadzane w trzech etapach dla każdego z problemów. W pierwszym etapie Doktorant sprawdzał wpływ wartości parametru, określającego prawdopodobieństwo wystąpienia zdarzenia, uwzględnionego w modyfikacji algorytmu (uwaga – to nie jest prawdopodobieństwo modyfikacji, ponieważ pojęcie modyfikacji odnosi się do algorytmu, a algorytm zmodyfikowany był zawsze), a wskaźnikiem była liczba generacji potrzebna do znalezienia rozwiązania. Problem polega na tym, że liczba generacji może silnie zależeć od populacji początkowej i przeprowadzenie jedynie dziewięciu powtórzeń, oraz wybranie mediany jako wartości reprezentatywnej nie wydaje się uzasadnione (np. liczba generacji dla zbioru 4 w tabeli 4.4 w jako funkcja wartości parametru jest oscylacyjna) – dodatkowa uwaga na temat „optymalności” została zawarta w punkcie 6.16.

W trzecim etapie badań z kolei był mierzony czas potrzebny na przeprowadzenie obliczeń dla pojedynczej generacji – tu jako wskaźnik wybrano średnią ze 100 powtórzeń, przy czym podano również odchylenie standardowe. Wydaje się to brakiem konsekwencji. Ponieważ z praktycznego punktu widzenia istotny jest całkowity czas obliczeń, należałoby wyznaczyć średnią w obu przypadkach (etapy 1 i 3), a następnie wyznaczyć średnią i odchylenie standardowe całosci obliczeń (a najlepiej dystrybuant empiryczną tego czasu, po czym korzystając z odpowiednich testów statystycznych przetestować hipotezę o tym, czy różni się
ona od dystrybuanty, wyznaczonej dla standardowego algorytmu, w sposób istotny). Warto zauważyć, że w niektórych przypadkach wprowadzenie modyfikacji spowodowało zwiększenie odchylenia standardowego dla czasu pojedynczej generacji (Tab. 4.9, 4.17, 4.19), a w takim przypadku porównywanie wyłącznie wartości średniej może być niewystarczające. W Tab. 4.17 wielkość odchylenia standardowego jest większa od wartości średniej dla zbioru 5 przy modyfikacji Epi_1 – jak to rozumieć?

Na szczęście, wskaźniki procentowe pokazane w rozprawie w większości przypadków pozwalają stwierdzić z dużym prawdopodobieństwem, że taki test wskazałby istotna poprawę efektywności.

6.5. Metodyka przeprowadzania eksperymentów – porównanie zmian wartości funkcji przystosowania

Wnioski, wysuwane na podstawie wykresów przedstawionych dla tych etapów dla wszystkich przykładów wydają się, przynajmniej w części, pochopne. Nie można porównywać „na oko” różnic pomiędzy przebiegami na wykresach, charakteryzujących się różnicą skalą. Należało oprzeć się na jakimś wskaźniku liczbowym, obiektywnym. Ponadto, w wielu przypadkach wygląda na to, że porównywane są wykresy startujące z różnych wartości początkowych, co sugeruje różne populacje początkowe (nawet gdyby startowały z tych samych wartości nie musiałyby te oznaczać tej samej populacji początkowej) – np. Rys. 4.9a,b 4.13b – na czym oparto w tych przypadkach wnioski?. Ponadto, w zadaniach grupowania danych i znajdowania wartości odstających oryginalne sformułowanie problemu polegało na minimalizacji wartości wybranego wskaźnika jakości – ale w klasycznym podejściu wykorzystując algorytmy genetyczne funkcja przystosowania zawsze powinna być maksymalizowana, co wymaga odpowiedniej transformacji zadania (czywiście nie ma to wpływu na jakość rozwiązania, ale na spójność opisu).

6.6. Zadanie grupowania danych.

Na str. 74 pojawiło się stwierdzenie, że uwzględnienie proponowanej modyfikacji w algorytmie genetycznym umożliwiło uniknięcie utknięcia w minimum lokalnym, w porównaniu do standardowego algorytmu. Mogł to jednak być problem wynikający ze zbyt małej wielkości populacji.

6.7. Zadanie dopasowania sekwencji

Czym doktorant kierował się, wybierając sekwencje do dopasowania lub, innymi słowy, czy taki wybór jest uzasadniony? (czywiście można poszukiwać dopasowania dla dowolnych sekwencji, ale nie zawsze ma to sens)

6.8. Zadanie dopasowania wielu sekwencji (podrozdział 2.4) wymaga dokładniejszego opisu, a w szczególności

- Czym jest gen w tym zadaniu? Czy chromosom składa się z jednego genu, czy też dopasowania symboli na jednej pozycji?
- W jaki sposób uwzględnia się przypadek, w którym jedna sekwencja ma wstawione delecje na początku, a inna na końcu (tzn. jaka jest długość dopasowania, która powinna być parametrem)?
- Rys. 2.2 jest niedobrą ilustracją, ponieważ przy takich trzech sekwencjach można byłoby szukać dopasowania lokalnego, a nie globalnego

-4-
• Opis funkcji przystosowania jest odpowiedni dla zadania dopasowania dwóch sekwencji; nie nadaje się do poszukiwania wielu sekwencji; ponadto podana funkcja nadaje się wyłącznie do poszukiwania dopasowania sekwencji zasad, a nie aminokwasów, a i w tym przypadku jest niestandardowa (identyczna kara za otwarcie i kontynuację przerwy).

• Na rysunkach zamieszczonych w podrozdziale 2.2 nie występuje dopasowanie dwóch różnych symboli (tzw. mismatch, czyli niezgodność). Czy to przypadk, czy też wynika to z założeń algorytmu (chyba nie, ponieważ w funkcji przystosowania taki przypadek został przewidziany).

• Prawdopodobieństwo wystąpienia mutacji zostało określone jako bardzo wysokie (50-100%). Jako to ma wpływ na zbieżność algorytmu? Najczęściej wysokie prawdopodobieństwo mutacji przeszkadza w znalezieniu rozwiązania, ponieważ kolejne generacje nie mają tendencji do gromadzenia się wokół rozwiązania.

Pozostałe uwagi ogólne i komentarze, jakie nasuwają się podczas uważnej lektury recenzowanej rozprawy doktorskiej dotyczą następujących zagadnień:

6.9. Czy Doktorant spotkał się z implementacją algorytmu genetycznego, w którym kryterium stopu byłby czas obliczeń? (str. 27)
6.10. Jak rozumieć modyfikację, polegającą na zamianie reprezentacji z binarnej na (+,-) (str. 31) – przecież to nadal reprezentacja binarna.
6.11. Czym różni się problem optymalizacji załadunku, przedstawiony w rozprawie, od standardowego problemu plecakowego?
6.12. Czy (i ewentualnie jak) można powiązać zadanie wykrywania wartości odstających, przedstawione w podrozdziale 2.2 ze statystycznym testowaniem hipotez o wartościach odstających?

Ponadto nasuwają się następujące uwagi o charakterze dyskusyjnym:

6.13. W pracy przebadano wpływ każdego z algorytmów z osobna. Czy byłoby sensowne zastosowanie jakiejs ich kombinacji?
6.14. Grupowanie jest zadaniem stosunkowo prostym w małej liczbie wymiarów (w pracy, prawdopodobnie ze względu na możliwość graficznej prezentacji wyników, przykłady są 2- lub 3-wymiarowe. Jak jednak sprawdziłyby się modyfikacje w zadaniach, w których grupowane obiekty byłyby opisane wektorami cech o znacznie większej liczbie elementów?

Inne uwagi szczegółowe:

6.15. Próba jak najprostszej zdefiniowania problemów optymalizacyjnych w podrozdziale 1.1., niestety, nie została uwieńczona sukcesem. M.in. X w (1.1) wcale nie musi być zbiorem skończonym (poza tym zbiór i przestrzeń to jednak różne pojęcia), f(.) niekoniecznie musi być funkcją – w najogólniejszym przypadku jest to funkcja, natomiast podział i opis zadań optymalizacji przedstawiony w Tabeli 1.1. jest co najmniej dyskusyjny (m.in. trudno się zgodzić ze sposobem definiowania optymalizacji ciągłej/dyskretniej, czy też deterministycznej/stochastycznej).

6.16. Nie można mówić o znalezieniu optymalnej wartości parametru, jeśli może on przyjmować dowolną wartość z przedziału [0,1], a zostały sprawdzone wyniki tylko dla kilku wartości z tego przedziału. Można jedynie stwierdzić, że jest on najlepszy z wybranych do testów, ale to nie to samo, co optymalny.

6.17. Opis kodowania, przytoczony w Tabeli 1.4 jest standardowy dla popularnych opisów algorytmu genetycznego. Pomija jednak fakt, że zarówno liczby całkowite, jak i rzeczywiste są kodowane za pomocą zer i jedynek. Podobnie, nadmierne uproszczenie opisu prowadzi do
niepełnego przedstawienia np. operacji krzyżowania – najpierw mowa jest o losowaniu z rozkładu dwupunktowego (krzyżowanie zachodzi albo nie) a potem pojawia się stwierdzenie o losowaniu miejsca podziału genotypu, bez podania informacji, z jakiego rozkładu. Podobnie, brak jest w rozprawie informacji o tym, z jakich rozkładów losowane są inne wartości w algorytmach (np. w algorytmie 4 – długość zmiany, miejsce modyfikacji). Należy ze smutkiem stwierdzić, że taki brak dobałości o szczegóły jest charakterystyczny dla większości prac, poruszających problematykę algorytmów genetycznych

6.18.Dlaczego na rys. 1.4d znajduje się jeden potomek, a na rys. 1.4 a-c dwóch, przy czym jeden z nich to rodzic?

- Nieprawidłowe jest stwierdzenie, że „DNA zbudowana jest z czterech komplementarnych aminokwasów – po pierwsze, to jest zbudowana z czterech zasad, a po drugie są one parami komplementarne (str.14)
- Nie jest prawdą, że informacja o tym, czy białko powstanie, jest zapisane w DNA – powstanie (lub nie) białka jest procesem losowym, uzależnionym od wielu różnych czynników
- Pojedynczy gen nie zawiera pojedynczej informacji fenotypowej (jak autor rozumie pojęcie fenotypu i „informaty fenotypowej”? Gdzie w tym kontekście umieścić tzw. alternative splicing?) (str. 15)
- Nie należy mylić pojęć kodu i informacji zapisanej w danym kodzie, jak to ma miejsce np. na str. 16 („mutacja polega na samoistnej zmianie kodu genetycznego”), czy 47 („pozbicie się kodu DNA”).
- Trudno zgodzić się z potraktowaniem strategii ewolucyjnej jako algorytmu (Rys. 1.3)
- Dość często pojawia się w pracy określenie „oczekiwano wyniku” (np. str. 18, 52) – nie jest to najsłabsze sformułowanie;
- Nieprecyzyjne sformułowania znajdują się np. w Tabeli 1.3 (selekcja to nie jest ocena)
- Sformułowania w liście wypunktowanej na str. 21 są nierozumiałe (poza pierwszym, które jest oczywiste)
- Stwierdzenie, że „wadą metody selekcji proporcjonalnej jest możliwość wystąpienia sytuacji, gdy prawie całą powierzchnię koła zajmuje jeden osobnik” (str. 22) jest, moim zdaniem, nadużyciem – tak rzeczywiście może się stać, ale byłoby to efektem złego doboru parametrów algorytmu lub złego sposobu generowania populacji początkowej. Z kolei brak jest jakieśkołwiek uwagi na temat wad selekcji rankingowej
- Opis problemu grupowania danych na str. 36 zawiera wiele błędów redakcyjnych (m.in. w (2.3) pojawia się symbol Y, nigdzie nie zdefiniowany, pojedynczy indeks przy literze x raz oznacza numer obiektu, a innym razem numer jednej z cech, stwierdzenie „do najpopularniejszych należą algorytmy [...] oraz wiele innych” jest nonsensowne)
- Widać wyraźnie, że Doktorant jedynie powierzchownie zetknął się z genetyką i biologią. M.in. Stwierdzenie o wolnych cząstkach – pierwszy punkt w liście na str. 43 – jest co najmniej nieprecyzyjne; podobnie stwierdzenie, że „część zmian w genotypie nie jest bezpośrednio związana ze strukturą DNA [...]” (str. 43). Określenie, że eukarioty to organizmy wielokomórkowe (str. 49) jest już wstydliwym błędem, podobnie jak sformułowanie „rearanżacja genomu kodującego białko” (str. 50).
- Zamienne korzystanie z terminów „generacja” i „iteracja” w Rozdziale 4 może być mylące
- Brakuje podania wartości niektórych parametrów, jak np. wielkości początkowej populacji w 4.3.2, a określenie warunku stopu jako wartość „warunkowana długością zbioru liczba generacji bez poprawy wartości funkcji przystosowania” jest nieprecyzyjne – jak warunkowana?
6.20. Ponieważ na początku rozprawy Doktorant opisał konwencję edytorską (st. 9), to podawanie skrótu „ang.” W pracy jest zbędne.


7. Ocena końcowa rozprawy.

Podsumowując, pomimo licznych uwag szczegółowych, wymienionych powyżej, uważam, że silne strony pracy wymienione w punkcie 4 niniejszej recenzji przeważają nad słabymi. Wobec tego stwierdzam, że mgr Kornel Chromiński wykazał się wiedzą i umiejętności uprawniającymi go do ubiegania się o stopień doktora nauk technicznych w dyscyplinie Informatyka. Przedstawiona praca doktorska spełnia wymagania stawiane pracom doktorskim zarówno przez Ustawę o stopniach i tytułach naukowych oraz stopniach i tytułach w zakresie sztuki z dnia 14 marca 2003 roku (z późn. zmianami), jak i obowiązującą obecnie ustawę Prawo o szkolnictwie wyższym i nauce z dnia 3 lipca 2018 r. Wnioskuję o dopuszczenie mgra Kornela Chromińskiego do publicznej obrony rozprawy doktorskiej. Jednocześnie pragnę zaznaczyć, że w dostarczonej dokumentacji brakuje streszczenia rozprawy w języku angielskim (wymóg wskazany w punkcie 6 art. 13 Ustawy o stopniach i tytułach naukowych).