Gliwice, 1 marca 2015

dr hab. inż. Krzysztof Fujarewicz
prof. nzw. Pol. Śl.
Instytut Automatyki
Politechnika Śląska
krzysztof.fujarewicz@polsl.pl

Recenzja Rozprawy Doktorskiej

Tytuł: **KNOWLEDGE DISCOVERY IN BIOLOGICAL DATA**

Autor: **mgr Indrajit Saha**

Promotor: **Dr hab. Dariusz Plewczyński**

Recenzowana rozprawa dotyczy metod analizy danych pozyskiwanych w naukach biologicznych i medycznych. W obszarach tych, ze względu na pojawienie się nowych technik pomiarowych, naukowcy pozyskują coraz większe zbiory danych, których analiza i interpretacja przestają być możliwe bez użycia komputerów i stworzonych specjalnie na ich potrzeby współczesnych narzędzi obliczeniowych. Stało się tak za sprawą pojawienia się, już pewien czas temu, mikromacierzy DNA, technik badań proteomicznych wykorzystujących spektrometrię mas czy też metod sekwencjonowania wysokiej przepustowości. Wszystko to sprawiło, że pojawiła się gwałtowna potrzeba utworzenia nowych narzędzi i metod obliczeniowych, uwzględniających specyfikę generowanych danych. Potrzebie tej wyszła naprzeciw społeczność naukowców reprezentujących inne dziedziny: informatykę, matematykę, statystykę, sztuczną inteligencję itp. Przykładem próby sprostania takiemu zapotrzebowaniu jest również niniejsza rozprawa doktorska.

Przedstawiona do recenzji praca liczy 161 stron, nie licząc stron zawierających streszczenia oraz spis treści. Napisana została w języku angielskim. Praca została napisana bardzo starannie i przejrzyście, a jej układ nie budzi większych zastrzeżeń.

Zakres rozprawy

Rozprawa swoim zakresem obejmuje wiele różnych zagadnień dotyczących zarówno nadzorowanej jak i nienadzorowanej analizy danych biomedycznych. Autor prezentuje

propozycje różnych algorytmów przetwarzania danych i z powodzeniem stosuje je do analizy danych rzeczywistych. Algorytmy te w głównej mierze dotyczą problemów uczenia nienadzorowanego polegających na grupowaniu (klasteryzacji) danych. Algorytmy te stanowią połączenie różnych narzędzi obliczeniowych, w tym: zbiorów rozmytych, algorytmów ewolucyjnych i genetycznych oraz bardzo popularnych w ostatnim czasie zbiorów, czy też komitetów (ensembles) wielu algorytmów pracujących niezależnie. Autor następnie każdy z opracowanych algorytmów stosuje do różnego typu danych, w tym: danych mikromacierzowych, obrazów z rezonansu magnetycznego (MR) mózgu, fizykochemicznych cech aminokwasów, oraz różnego typu danych proteomicznych takich jak potranslacyjne modyfikacje sekwencji białkowych i różnych interakcji pomiędzy białkami.

Na uwagę zwraca fakt, że *autor nie formułuje jednej jasno sprecyzowanej tezy pracy*, do czego jesteśmy przyzwyczajeni w tego typu dysertacjach, i co budzi również mój niedosyt w tej materii.

Zawartość Rozprawy

Recenzowana rozprawa składa się z 8 rozdziałów obejmujących 6 zasadniczych rozdziałów opisujących różne metody i zastosowania oraz rozdziały wstępu i zakończenia.

Autor rozprawy posiada bardzo duży dorobek publikacyjny i właściwie każdy z rozdziałów związany jest, mniej lub bardziej, z konkretnym artykułem wcześniej opublikowanym. Dodać należy, że mgr Indrajit Saha w większości przypadków jest pierwszym autorem tych publikacji oraz, że zostały one opublikowane w dobrych lub bardzo dobrych czasopismach naukowych, w tym: Expert Systems with Applications (IF=2,2), Amino Acids (3,91), Immunogenetics (2,49), Molecular Biosystems (3,18). Autor ma w swoim dorobku również inne bardzo dobre prace (np. w Pattern Recognition), które jednak nie dotyczą analizy danych biomedycznych.

Miejsce opublikowania wymienionych wyżej artykułów stanowi o sile zawartości merytorycznej rozprawy doktorskiej, z drugiej jednak strony owa odpowiedniość, czasem 1:1 artykuł/rozdział spowodowały, że Autor nie ustrzegł się pewnych błędów, braku spójności czy też niekonsekwencji, obniżając ogólny, bardzo dobry obraz pracy. Część tych błędów wymieniona będzie w sekcji „Uwagi" niniejszej recenzji.

Rozdział 1 – „Introduction and scope of the thesis" – stanowi wprowadzenie do całej pracy. Po krótkim wstępie Autor omawia w bardzo skrótowy sposób podstawy biologii molekularnej, która jest źródłem danych analizowanych w pracy, oraz różne klasyczne

metody nadzorowanej i nienadzorowanej analizy danych. Rozdział kończy przegląd zawartości pozostałych rozdziałów w kolejności ich występowania.

Rozdział 2 poświęcony jest metodom klasteryzacji danych z mikromacierzy DNA. Oparty jest na dwóch artykułach opublikowanych wcześniej w International Journal of Data Mining and Bioinformatics oraz w Expert Systems with Applications. Autor proponuje nową metodę klasteryzacji nazwaną przez niego Improved Differential Evolution based Automatic Fuzzy Clustering Technique (IDEAFC), której zaletą jest między innymi automatyczne wykrywanie liczby klastrów. Następnie metoda jest poddana pewnemu ulepszeniu, polegającemu na modyfikacji końcowego wyniku przy użyciu nadzorowanej metody uczenia jaką jest technika Maszyn Wektorów Nośnych (SVM). Efektem jest algorytm o nazwie jak wcześniej ale z dopiskiem SVM: IDEAFC-SVM. Rozdział zawiera następnie przedstawienie wyników otrzymanych podczas analizy tymi metodami rzeczywistych zbiorów danych ogólnodostępnych w istniejących repozytoriach danych. W dalszej części Autor przedstawia porównanie jakości działania jego metod z innymi znanymi metodami wykazując wyższość tych pierwszych.

Rozdział 3 stanowi propozycję kolejnego algorytmu klasteryzacji oraz wykorzystania go do analizy innego rodzaju danych jakimi są obrazy mózgu pozyskanego techniką rezonansu magnetycznego (MR). Algorytmem jest Multiobjective Differential Evolution based Fuzzy Clustring (MODEFC). Cechą odróżniającą go od innych metod jest wyposażenie go w mechanizm optymalizacji wielokryterialnej wykorzystującej dwa różne wskaźniki jakości oceniające jakość otrzymywanych klastrów. Podobnie jak w poprzednim rozdziale Autor porównuje jakość otrzymywanych wyników numerycznych w wynikami otrzymywanymi innymi metodami klasteryzacji wykazując przewagę metody MODEFC.

Rozdział 4 dotyczy klasteryzacji kilkuset cech opisujących własności fizykochemiczne aminokwasów zawartych w bazie AAindex. Oparty jest na artykule opublikowanym wcześniej w Amino Acids, którego autorzy: (i) zaproponowali Consensus Fuzzy Clustering Algorithm, będący „komitetem" złożonym z kilku innych metod grupowania danych, (ii) dokonali klasteryzacji nowej wersji bazy AAindex i w odróżnieniu od poprzedniego wyniku znanego z literatury (6 grup) dla poprzedniej wersji bazy danych otrzymali 8 grup cech, oraz (iii) wytypowali grupy cech reprezentujących poszczególne klastry charakteryzujące się wysoką jakością: HQI8, HQI24 i HQI40.

Rozdział 5 jest w pewnym sensie powiązany z rozdziałem poprzednim bo wykorzystuje otrzymane tam grupy cech HQIx. Rozdział ten jest ściśle związany z artykułem, którego współautorem jest mgr Saha i który został opublikowany również w Amino Acids. Rozdział

jak i artykuł opisuje kolejną wersję serwisu internetowego dokonującego predykcji modyfikacji potranslacyjnych białek AutoMotif Service (AMS 4.0), których wcześniejsze wersje były tworzone przez promotora doktoranta dr hab. Dariusza Plewczyńskiego. Podczas predykcji wykorzystanych zostało wiele klasyfikatorów (Consensus Approach) typu wielowarstwowy perceptron (MLP) .

Rozdziały 6 i 7, mówiąc najogólniej dotyczą wykorzystania nadzorowanych metod analizy (klasyfikacji) predykcji różnego typu interakcji pomiędzy białkami. Bezpośrednio związane są z dwoma wcześniejszymi publikacjami doktoranta opublikowanymi w Immunogenetics i Molecular BioSystems. W rozdziale 6 jako klasyfikator wykorzystano komitet głosujący (ensemble) złożony techniki bootstrap, SVM i PCA. Klasyfikator taki został nazwany został przez autorów RotaSVM. Z kolei w rozdziale 7 porównano kilka klasycznych metod klasyfikacji użytych również do predykcji oddziaływań białko-białko.

Rozprawę kończy rozdział 8 podsumowujący całą pracę i wytyczający kierunki przyszłych prac. Po nim następuje spis literatury zawierający pokaźną liczbę 403 pozyjci literaturowych.

Istotne elementy rozprawy

Do istotnych i oryginalnych elementów rozprawy zaliczam:

1. Opracowanie algorytmu Improved Differential Evolution based Automatic Fuzzy Clustering Technique (IDEAFC) pozwalającego wykrywać liczbę klastrów w danych mikromacierzowych.
2. Technikę klasteryzacji wykorzystującą optymalizację wielokryterialną do analizy obrazów z rezonansu magnetycznego.
3. Wykonanie klasteryzacji kilkuset cech opisujących własności fizykochemiczne aminokwasów zawartych w bazie AAindex za pomocą autorskiego podejścia Consensus Fuzzy Clustering.
4. Wykorzystanie techniki rodzin klasyfikatorów do predykcji interakcji białko-białko.

Spośród wymienionych wyżej za najistotniejszy uważam pkt. 4. Wyniki klasteryzacji zostały poddane najwnikliwszej interpretacji od strony biologicznej, co znalazło również potwierdzenie w dużej liczbie cytowań odpowiedniej wcześniejszej publikacji.

Uwagi krytyczne

Poniżej wymieniłem główne uwagi krytyczne które nasunęły mi się podczas lektury rozprawy.

1. Brak jasno i wyraźnie sformułowanej tezy rozprawy.
2. Rozprawa zawiera wiele rysunków zaczerpniętych bezpośrednio z wcześniejszych artykułów. Powstaje pytanie czy nie powinny one być opatrzone odpowiednim komentarzem skąd pochodzą?
3. Równanie 1.2 zawiera błąd. Jego lewa strona powinna określać wielowymiarową funkcję gęstości prawdopodobieństwa warunkowego dla klasy C1 czyli drugi człon licznika wyrażenia (1.1). Ewentualnie można by zastąpić w (1.2) znak równości znakiem proporcjonalności i poczynić dodatkowe założenie o równości prawdopodobieństwa *a priori* wystąpienia klas.
4. Strona 12, linia 14. Problemy wieloklasowe można rozwiązywać za pomocą SVM bez dekompozycji One versus One czy też One versus Rest. Istnieją bowiem odpowiednie wersje wieloklasowe algorytmu SVM rozwiązujące jeden wspólny problem programowania kwadratowego.
5. Rozdział 3.4.2 i wyniki porównania metod klasteryzacji przedstawione w tabelach 3.1, 3.2.: Autor porównuje różne metody klasteryzacji za pomocą wskaźnika Minkowskiego (Minkowski Score, MS). Jednak „jego" metoda MODEFC w jawny sposób na pewnym etapie ten sam wskaźnik wykorzystuje do optymalizacji rozwiązania. W oczywisty sposób faworyzuje to tę metodę podczas porównania z innymi. Uwaga ta ma ogólniejszy charakter i dotyczy innych rozdziałów pracy, w których wykorzystuje się metody klasteryzacji. Proponowane metody dokonują optymalizacji (różnymi metodami) wskaźnika jakości „podobnego" do używanego następnie do oceny i porównania omawianej metody z metodami innymi. W sytuacji takiej „nierównej" rywalizacji warto byłoby porównać metody stosując „zewnętrzną" walidację (external validation) na nadzorowanym zbiorze danych, czyli taki dla którego znane są przynależności do klas. Niczego takiego jednak autor nigdzie w rozprawie nie robi.
6. W pracy jest bardzo dużo skrótów na określenie różnych metod. Czytelnik może się podczas lektury zgubić, czego sam doświadczałem. Według mnie praca powinna zostać uzupełniona o spis akronimów poprawiający czytelność rozprawy.

Przedstawione powyżej uwagi krytyczne mają w dużej mierze charakter polemiczny i w niewielkim tylko stopniu zakłócają ogólną, bardzo dobrą ocenę Rozprawy.

Uwagi redakcyjne

Praca napisana jest starannie i trudno w niej doszukać się większych błędów natury edycyjnej. Poniżej wymieniono zauważone usterki redakcyjne.

1. W podrozdziale 1.4.1.5 pojawiają się wzory określające postaci jąder w SVM. Nie są one jednak dostatecznie opisane, nie wiadomo na przykład co to jest D, d, k. Ogólnie, rozdział 1 został potraktowany przez Autora bardzo ogólnikowo co spowodowało że nie ustrzegł się w nim pewnych usterek.
2. Wzór (1.5): nie wiadomo co oznaczają małe symbole c_i, c_j.
3. Niepotrzebnie zdefiniowano dwa razy medoid, na stronie 15 i na stronie 17.
4. Rozdział 2.4.2: nie wiadomo jakiego typu jądra maszyny SVM użyto i z jakimi parametrami.
5. Dane zaprezentowane w podrozdziale 2.5.1.3 zostały niedostatecznie opisane. Mówi się o 138 genach i 8 punktach czasowych. Nie wiadomo jednak nic o tym jakiemu czynnikowi poddano roślinę.
6. Podobna uwaga dotyczy danych przedstawionych w pkt. 2.5.1.4.
7. W pracy gdzieniegdzie występują „niezręczności" wynikające ze skopiowania wcześniejszych (dodajmy: swoich, czy też współautorskich) artykułów. Przykładowo, str. 84, linia 4 od dołu: „Our previously developed web-server …" miało sens w artykule, którego współautorem był dr hab. Plewczyński. W rozprawie doktorskiej, której autorem jest jedynie doktorant, sformułowanie to już nie ma prawa bytu bo mgr Saha nie był współautorem publikacji [316] i wcześniejszej wersji serwera. Podobnie następne zdanie „currently available version" odnosi się do wersji 3.0, podczas gdy obecnie „currently available version" jest wersja 4.0 bo artykuł już opublikowano.
8. Rozdział 6: nie wiadomo dlaczego wybrano takie a nie inne parametry metod klasyfikacji, z którymi porównywane są metody proponowane. Czy dokonano tego wybory na drodze walidacji?
9. W punkcie 6.3.2. podano jedynie definicje wskaźników oceny jakości klasyfikacji. Nie wiadomo natomiast jak testowano klasyfikatory. Czy był niezależny zbiór testowy? A jeśli nie to jaki schemat walidacji wybrano?

10. W rozdziale 7 natomiast napisano, że zastosowano 10-krotną walidację krzyżową ale z kolei nic nie napisano o wartościach parametrów wykorzystanych metod klasyfikacji oraz jak je dobrano. Szczególnie razi jedynie hasłowe potraktowanie sieci neuronowej (ANN). Nie podani jaka to sieć, o jakiej strukturze, jaki algorytm uczenia zastosowano i z jakimi parametrami.

Wnioski końcowe

Mgr Indrajit Saha wykazał się odpowiednią wiedzą z zakresu biologii komórki, bioinformatyki, uczenia maszynowego, statystycznej analizy danych, umiejętnością programowania, oraz zdolnością do twórczej analizy otrzymanych wyników.

Spora część uwag krytycznych wynikała z faktu, że rozprawę doktorską przygotowano na podstawie wcześniej opublikowanych prac. Trudno jednak robić zarzut z faktu posiadania przez doktoranta niemałego dorobku publikacyjnego.

Przedstawione w niniejszej recenzji uwagi krytyczne, w głównej mierze mające charakter marginalny lub polemiczny, w niewielkim jedynie stopniu obniżają bardzo wysoką ogólną ocenę przedstawionej rozprawy.

Stwierdzam, że praca p.t. *Knowledge discovery in biological data* spełnia wymagania stawiane rozprawom doktorskim przez Ustawę o Stopniach i Tytule Naukowym obowiązującą aktualnie w Polsce.

Stawiam wniosek o dopuszczenie jej do publicznej obrony. Jednocześnie, biorąc pod uwagę wysoki poziom recenzowanej rozprawy oraz dorobek naukowy Autora, wnioskuję o wyróżnienie rozprawy doktorskiej.