Prof. dr hab. inż. Jacek Koronacki
Instytut Podstaw Informatyki PAN
Jana Kazimierza 5, 01-248 Warszawa

Warszawa, 14 marca 2014

Recenzja rozprawy doktorskiej mgra Tomasza Xięskiego
Wydobywanie wiedzy z danych złożonych

Zakres tematyczny rozprawy

Autor przyjmuje, że dane złożone to dane o wielu atrybutach (cechach, zmiennych) mierzonych nie tylko na różnych skalach pomiarowych – od nominalnej do ilorazowej – ale także tekstowych czy reprezentujących daty. Wydobywanie wiedzy sprowadza w rozprawie do eksploracyjnej analizy danych, a ściślej – nie pomijając wstępnego przygotowania danych i wstępnej analizy eksploracyjnej – do analizy skupień. Tą drogą rzeczywiście możemy uzyskać bardzo istotną informację o strukturze ukrytej w danych – naturalnym pogrupowaniu danych w skupienia danych w pewnym sensie podobnych do siebie wewnątrz skupienia i do siebie niepodobnych, jeśli należą do różnych skupień.

Doktorant jest doskonale świadomy, z jakimi trudnościami należy się liczyć chcąc uzyskać prawdziwie reprezentatywne wyniki. Problemom tym, na przykład problemowi doboru parametrów analizowanych algorytmów, poświęca w rozprawie wiele uwagi.

Szczególny akcent kładzie na graficzną analizę eksploracyjną, ponieważ słusznie zauważa, iż wizualizacja danych – czy ich skupień – może być bardzo efektywnym i w pewnym sensie autonomicznym narzędziem analizy.

Swoje badania prowadzi w kontekście analizy dwóch rzeczywistych i w podanym już sensie złożonych zbiorów danych. Obydwa zbiory pochodzą z obszaru telekomunikacji, ale są bardzo odmienne pod względem ich struktury.

Na rozprawę składa się 9 rozdziałów, bibliografia, słownik pojęć, wykaz przeprowadzonych badań dodatkowych, spis rysunków i spis tabel.

Po krótkim Wprowadzeniu, Autor zwraca w drugim rozdziale uwagę na wagę wiedzy dziedzinowej w eksploracji danych i przy tej okazji przedstawia potrzebne elementy wiedzy telekomunikacyjnej oraz omawia zbiory danych, dla których będzie prowadzić analizę. Rozdział 3 zawiera kompetentne porównanie wybranych, znanych systemów analizy danych, tym samym uzasadniając zaproponowanie autorskiego systemu analizy, który nazwał DensGroup.

Rozdział 4 poświęcony jest zwartemu opisowi elementów wstępnej eksploracyjnej analizy danych (EDA), zaś rozdział 5 algorytmom analizy skupień opartym na analizie gęstości danych, a właściwie dwóm takim algorytmom, szczególnie popularnym, DBSCAN oraz OPTICS. W rozdziale 6 Autor omawia znane graficzne metody reprezentacji skupień.

Rozdział 7 ma charakter autorski – omówiony jest w nim projekt i implementacja systemu DensGroup. Rozdział 8 zawiera szczegółową analizę wydobywania wiedzy z dwóch wspomnianych rzeczywistych zbiorów danych z obszaru telekomunikacji. W rozdziale ostatnim znajdujemy podsumowanie całości, zwłaszcza wniosków dotyczących przeprowadzonych eksperymentów obliczeniowych.

Rozprawa liczy 167 stron. Bibliografia zawiera 91 pozycji.

Uwagi ogólne

Rozprawę należy ocenić jako bardzo dobrze skomponowaną – całość wywodu jest logiczna i spójna. Autor ma dużą erudycję, dobrze sformułował zadanie, jakie przed sobą postawił, oparłszy się na rzetelnym przeglądzie dostępnych rozwiązań. Przyjął, że potrzebne jest oprogramowanie niezbyt wymagające ze względu na złożoność obliczeniową i pamięciową, ale zarazem zdolne poradzić sobie z analizą skupień w przypadku złożonych danych (w sensie przezeń przyjętym) i przy stosunkowo licznych zbiorach takich danych. Dodatkowym wymaganiem uczynił prostotę posługiwania się oprogramowaniem – w domyśle przez niewprawnego użytkownika – oraz łatwym w obsłudze, ale dającym wiele informacji interfejsem graficznym.

Usprawiedliwia to ograniczenie się w rozdziale 3 do przeanalizowania pakietu R jedynie w takim zakresie, jaki oferuje nakładka Rattle. Inna sprawa, że tym sposobem z pola widzenia zniknęły różne zaimplementowane w R algorytmy analizy skupień, także dla danych z różnych skal pomiarowych (np. ClustOfVar, nie mówiąc o programach kernlab). I nie zmienia to faktu, że materiał zawarty w rozdziałach 3 – 6 należy ocenić wysoko (poza – p. uwagi szczegółowe – podrozdziałem 4.1). Opis jest nie tylko trafny i jasny, ale podany bardzo atrakcyjnie – czyta się te rozdziały z dużą przyjemnością.

Niniejszy recenzent wolałby pracę o większym ładunku oryginalności, na przykład w obszarze badań nad nowymi własnymi algorytmami, może w nowy sposób wykorzystującymi idee spektralnej czy entropijnej analizy skupień, w taki przy tym sposób, by uczynić te algorytmy obliczeniowo i pamięciowo niezbyt wymagającymi.

Ale Doktorant miał prawo pójść w kierunku pracy bardziej inżynierskiej – pełnej reimplementacji algorytmów znanych oraz dodania własnego sposobu wykorzystania map prostokątów, do tego wzbogacenia całej procedury o dwustopniową analizę skupień (najpierw DBSCAN lub OPTICS i potem AHC na reprezentantach skupień wcześniej otrzymanych). Realizacja tych autorskich pomysłów wymagała ogromnej pracy i wysokiego rzemiosła programistycznego (wiem, że: Autor napisał zupełnie na nowo algorytm DBSCAN; w przeciwieństwie do implementacji w Wece tylko implementacja mgra Xięskiego algorytmu OPTICS radzi sobie z dużymi zbiorami danych; zaś implementując algorytm AHC co prawda wzorował się na znanej implementacji, ale w C, przepisał więc tę implementację na C++ i dostosował do struktury danych występujących w technologii Qt). System DensGroup jest systemem autorskim – całością łączącą w jedno gęstościowe podejście do analizy skupień z dobrą graficzną interpretacją wyników.

Mam jedynie pretensję do Doktoranta za ograniczenie się do badania podobieństwa (lub odmienności) między obserwacjami jedynie w oparciu o odległość Hamminga. Takie podejście nie może prowadzić – przynajmniej bezpośrednio – do sensownych wyników, gdy

atrybuty mają wiele możliwych wartości, na przykład gdy są (w zasadzie) ciągłe. To dlatego musiał w analizowanych przykładach otrzymać mgr Xięski ogromne liczby skupień. Takie podejście nie daje się usprawiedliwić potrzebą obniżenia złożoności obliczeniowej, w ogólności bowiem może okazać się błędem metodologicznym, którego nie da się naprawić wykonując powtórną analizę skupień na reprezentantach skupień otrzymanych w pierwszym kroku procedury. System DensGroup powinien mieć zaimplementowane inne miary odmienności.

Wszakże i bez tych innych możliwości pomiaru odmienności między obserwacjami system DensGroup jest dużym i ważnym osiągnięciem Doktoranta.

Ostatnia pochwała dotyczy rzetelności przeprowadzonych analiz rzeczywistych zbiorów danych. Chociaż – zdaniem niniejszego recenzenta – nie były to dane szczególnie wymagające, ponieważ wymiar (liczba atrybutów) obserwacji nie był duży.

Uwagi szczegółowe

Uwagi te są bardzo różnej wagi – po części są to uwagi bardzo drobne, nieistotne, po części mają nieco lub po prostu poważniejszy charakter. Podaję je zgodnie z kolejnością stron, których dotyczą. I jeszcze jeden potrzebny komentarz: zwykle poniższe uwagi mają charakter uwag krytycznych, ale w jednym przypadku jest to komentarz pozytywny.

- s. 1 i wiele razy dalej: Być może Autor znalazł gdzieś nazwę „kategoryczne" na zmienne jakościowe, ale jest to termin z oczywistych względów zły.
- s. 2 i dalej: Zamiast „kryterium stopu" powinno się pisać (mówić) o kryterium zatrzymania procedury.
- s. 15: Nie „półtorej", a półtora miliona.
- s. 16 i dalej: Autor używa wymiennie terminów analiza skupień i segmentacja. Nie jest to poprawne, choć wielu autorów czyni tak samo. Segmentacja to podział na z góry zadaną liczbę grup, niekoniecznie związaną z liczbą rzeczywiście obecnych w danych skupień.
- s. 18: Autor ma rację: z każdej fazy procesu wydobywania wiedzy powinien być możliwy powrót do każdej wcześniejszej fazy.
- s. 25: Nie „procesów", a procesorów (to zapewne Word chciał być mądrzejszy od Autora).
- s. 42: Autor zdaje się nie znać pojęcia skal pomiarowych, w tym rozróżnienia między skalą ilorazową i przedziałową. Zdaje się także lekceważyć wagę rozróżnienia między skalą porządkową i nominalną.
- podrozdz. 4.1: Autor nie wymienia zasadności/konieczności pomiaru wyostrzenia (kurtozy) rozkładu. Nie dość mocno odnosi się do zasadności/konieczności pomiaru skośności rozkładu (nie zauważa np., że w przypadku rozkładu skośnego pomiar odchylenia standardowego nie bardzo ma sens). Modę definiuje wyłącznie dla rozkładu dyskretnego, co już jest poważnym błędem, a następnie podaje ją na rys. 4.1 dla rozkładu ciągłego. Podana teza dotycząca przedziału plus minus dwa sigma jest fałszywa, jeśli dane nie mają rozkładu normalnego!
- podrozdz. 4.2: Polskojęzyczne odpowiedniki terminu *bins* są obecne w literaturze od dziesiątków lat. I chyba nikt nigdy nie użył nazwy kubły. Na s. 49 Autor pisze, że „nie istnieją jasne wytyczne, pozwalające wybrać optymalną liczbę przedziałów dla każdego rodzaju danych". Nie jest to sformułowanie szczęśliwe. Pomijając dlaczego, dodam tylko, że Autor nie dość uważnie wczytał się w odpowiednie fragmenty monografii Davida Scotta. Gdyby był się wczytał, znacznie lepiej opisałby problem doboru tej liczby i jednocześnie(!) doboru brzegów owych przedziałów. Na pewno wspomniałby wówczas także uśrednionych

histogramach przesuwanych (ASH). To nieprawda, że wykres kwantyl-kwantyl musi być nieczytelny, gdy liczba danych jest duża. Na pewno nie jest tak, gdy chodzi o normalny wykres kwantylowy – wręcz przeciwnie, im więcej danych, tym pewniejsza informacja o rozkładzie danych.
- s. 81, koniec podrozdz. 6.1: Co to znaczy „abstrakcyjny zestaw danych"?
- s.103: Weka nie jest programem.
- s. 120, rys. 8.1: Nie jest dla mnie jasne, dlaczego kontroler 137 stanowi odchylenie (a jeśli on nie byłby odchyleniem, to i 112 też nie; z kolei jeśli kontroler 137 jest odchyleniem, to może 139 i 142 też?).
- s. 129: Heurystyka typu „Utworzona liczba skupień powinna stanowić około 6% liczności danego zbioru" wydaje mi się nie mieć sensu.
- s. 138: Czy listing 8.5 rzeczywiście służy wyznaczeniu korelacji między temperaturą a obciążeniem procesora?
- w różnych miejscach rozprawy: Nie nazywałbym podrozdziałów sekcjami – nikt nie powinien tego czynić, ponieważ jest to niepotrzebny anglicyzm.

Jakkolwiek wymienione uwagi szczegółowe dotykają niekiedy kwestii niebanalnych i przez to obniżają wartość pewnych fragmentów rozprawy, to w oczywisty sposób są zbyt drobne, by mieć istotny wpływ na ogólnie pozytywną jej ocenę.

Konkluzja

Rozprawę oceniam jako spełniającą wymagania stawiane pracom doktorskim w dziedzinie nauk technicznych w dyscyplinie informatyka. Wnoszę o dopuszczenie mgra Tomasza Xięskiego do dalszych etapów przewodu doktorskiego.