Politechnika Białostocka
Rozprawy Naukowe Nr 158

Marek Krętowski

OBLICZENIA EWOLUCYJNE W EKSPLORACJI DANYCH

GLOBALNA INDUKCJA DRZEW DECYZYJNYCH

Wydawnictwo Politechniki Białostockiej
Białystok 2008

Wstęp

Współczesny świat uzależnia się coraz bardziej od systemów komputerowych. Coraz trudniej znaleźć firmę czy instytucję, w której nie byłyby wykorzystywane, w tej czy innej formie, rozwiązania informatyczne. Upowszechnienie się internetu powoduje rewolucyjne zmiany w sposobie ludzkiej pracy, komunikacji czy rozrywki.

W ogromnej liczbie funkcjonujących systemów komputerowych gromadzone i przetwarzane są informacje dotyczące procesów biznesowych, problemów badawczych czy spraw socjalnych. Rozwój technologii baz i hurtowni danych pozwala dziś przechowywać i udostępniać gigantyczną wprost ilość informacji, których liczba i dokładność nieustannie zresztą rośnie. W najbliższej przyszłości należy się spodziewać jeszcze bardziej intensywnego zalewu bardzo rozbudowanych danych pochodzących z nowych źródeł. Wystarczy wymienić coraz powszechniejsze dane genomiczne (genom każdego człowieka to około 2 miliardy 850 milionów par zasad) czy multimedialne (dane strumieniowe w postaci wideo).

Coraz częściej mamy do czynienia z sytuacją, gdy organizacje dysponują cennymi danymi, ale nie są w stanie odpowiednio ich przeanalizować. Problem efektywnego wykorzystania gromadzonych danych doprowadził na początku lat 90. ubiegłego stulecia do ukonstytuowania się nowej dziedziny nauki - pozyskiwania wiedzy z baz danych (ang. *knowledge discovery in databases*) [49]. Proces pozyskiwania wiedzy obejmuje wiele faz, począwszy od przygotowania danych, przez konstruowanie modeli czy też poszukiwanie ukrytych wzorców i zależności, aż do interpretacji i oceny odkrytych struktur. Najistotniejsza część tego procesu związana z analizą przygotowanych zbiorów danych nazywana jest eksploracją danych (ang. *data mining*) ([68], [34]). Wiele algorytmów rozwijanych w ramach tej dziedziny ma swoje korzenie w innych uznanych dziedzinach, takich jak statystyczne uczenie ([69], [82]) rozpoznawanie wzorców (ang. *pattern recognition*) [44] czy uczenie maszynowe (ang. *machine le-*

arning) ([126], [33]). W eksploracji danych klasyczne metody stworzone w innych dziedzinach adaptowane i stosowane są w kontekście dużych, rzeczywistych zbiorów danych. Wielkości zbiorów uczących, rzędu kilku tysięcy, kilkudziesięciu tysięcy czy nawet kilkuset tysięcy obiektów, nie są specjalnym zaskoczeniem. Ponadto, w odróżnieniu od typowych analiz statystycznych, dane, które są eksplorowane, zwykle nie były zbierane pod kątem tej konkretnej analizy. W wielu przypadkach są one gromadzone jako część codziennego funkcjonowania firm czy instytucji. Stąd też pojawia się określenie eksploracji danych jako *wtórnej analizy danych* [68].

Typowe zadania eksploracji danych, polegające na konstruowaniu efektywnych modeli czy poszukiwaniu interesujących wzorców, wymagają algorytmów charakteryzujących się wysoką złożonością obliczeniową. W połączeniu z wysoką wymiarowością analizowanych zbiorów danych, zarówno jeśli chodzi o liczbę obiektów jak i rozpatrywanych cech, wymusza to rozwijanie metod przybliżonych. Nie gwarantują one odnalezienia rozwiązania problemu najlepszego z możliwych, ale zwykle są w stanie zaproponować rozwiązanie bliskie optymalnego w rozsądnym, akceptowalnym przez użytkownika, czasie.

Jednym z najbardziej obiecujących kierunków rozwoju heurystycznych metod poszukiwania są algorytmy inspirowane mechanizmami rozwiązywania problemów, które można zaobserwować w naturze [122]. Ich wspólną cechą jest umiejętność unikania minimów lokalnych. Najbardziej znaną rodziną metod tego rodzaju są obliczenia ewolucyjne, które odwołują się do procesu ewolucji i selekcji naturalnej. W przeciwieństwie do metod klasycznych, w algorytmach ewolucyjnych nie są przetwarzane pojedyncze potencjalne rozwiązania problemu, ale populacje osobników reprezentujących te rozwiązania. W kolejnych generacjach osobniki podlegają różnicowaniu i rywalizują między sobą o przetrwanie opierając się na jakości odpowiadających im rozwiązań. Ocena jakości osobników mierzona jest poprzez wyliczanie wartości funkcji dopasowania.

Spośród innych metod zainspirowanych procesami zachodzącymi w naturze warto wymienić sztuczne systemy immunologiczne [161], poszukiwanie z tabu [62], systemy mrowiskowe (ang. *ant colony*) [21] czy roje cząstek (ang. *particle swarm*) [79].

W niniejszej monografii pragnę przedstawić możliwości wykorzystania metod ewolucyjnych w eksploracji danych. Skoncentrowałem się na indukcji drzew decyzyjnych, które należą do najbardziej rozpowszech-

nionych form reprezentacji wiedzy wydobywanej ze zbiorów danych. W typowych systemach eksploracji danych drzewa decyzyjne są konstruowane z wykorzystaniem algorytmu zstępującego, który jest bezpośrednią realizacją klasycznej zasady "dziel i zwyciężaj". W kolejnych tworzonych węzłach drzewa przeprowadzana jest lokalna optymalizacja testów według przyjętego kryterium, która jednak nie gwarantuje optymalności końcowej struktury. Proces generowania drzew przy użyciu tej zachłannej procedury jest szybki, ale uzyskiwane struktury decyzyjne nie zawsze są zadowalające. Zastosowanie specjalizowanych algorytmów ewolucyjnych umożliwia globalną indukcję, pozwalającą identyfikować wzajemne zależności w danych i uzyskiwać lepsze drzewa decyzyjne.

W monografii przedstawiona zostanie, zaproponowana przez autora, rodzina algorytmów umożliwiających indukcję zarówno typowych drzew jednowymiarowych, w których testy w węzłach analizują wartości pojedynczych atrybutów, jak i drzew skośnych, w których testy opierają się na liniowych kombinacjach wielu atrybutów, a także drzew mieszanych, które wykorzystują różne rodzaje testów. Autor zbadał również możliwość wykorzystania w procesie ewolucji drzew decyzyjnych algorytmów memetycznych, rozszerzających oryginalny algorytm ewolucyjny poprzez włączenie procedur przeszukiwania lokalnego. Ponadto zaprezentowany będzie wariant drzewa jednowymiarowego, który uwzględnia koszty podczas budowy klasyfikatora. Nieuniknioną ceną, którą płacimy za elastyczność i odporność metod ewolucyjnych, jest wydłużony czas obliczeń. Szczęśliwie algorytmy ewolucyjne są w sposób naturalny współbieżne. Zaprezentowana będzie zarówno implementacja równoległa na klastrze obliczeniowym jak i rozproszona wersja ewolucji drzew decyzyjnych.

Monografia składa się z ośmiu rozdziałów i kończy się podsumowaniem. W rozdziale pierwszym przywołano podstawowe informacje dotyczące obliczeń ewolucyjnych oraz zasad projektowania efektywnych algorytmów ewolucyjnych. Nie stanowi on pełnego przeglądu tej dynamicznie rozwijającej się dziedziny. W języku polskim opublikowano w ostatnim czasie kilka niezwykle interesujących podręczników i monografii dotyczących obliczeń ewolucyjnych (np.: [121], [137], [8], [122]), do których zgłębiania gorąco zachęcam. W związku z tym w rozdziale ograniczono się tylko do tych zagadnień, bez których zrozumienie dalszej części książki mogłoby być trudne.

W rozdziale drugim zaprezentowano krótkie wprowadzenie do zagadnień pozyskiwania wiedzy z baz danych, a w szczególności eksploracji danych. Omówiono typowe zadania algorytmów eksploracji danych oraz przedstawiono przebieg procesu poszukiwania modeli i wzorców. W dalszej części rozdziału skoncentrowano się na problemie klasyfikacji i przedstawiono drzewa decyzyjne oraz najbardziej z nimi spokrewnione reguły decyzyjne. Omówiono rodzaje drzew oraz klasyczne metody ich indukcji. W ostatniej części rozdziału przedstawiono istniejące systemy eksploracji wiedzy, bazujące na metodach ewolucyjnych.

Kolejne rozdziały monografii prezentują oryginalne rozwiązania autora w zakresie ewolucyjnej indukcji drzew decyzyjnych. Rozdział trzeci opisuje najbardziej podstawowy wariant globalnej indukcji drzew jednowymiarowych. Zaprezentowane w nim zostały m.in. specjalizowane operatory różnicowania wraz ze schematami ich stosowania, a także funkcja dopasowania. Rezultaty eksperymentalnej weryfikacji zaproponowanego rozwiązania na zestawach sztucznych, jak i rzeczywistych, zbiorów danych przedstawiono w końcowej części rozdziału.

W rozdziale czwartym podano możliwości rozszerzenia algorytmów ewolucyjnych poprzez wprowadzenie lokalnego przeszukiwania. Metody takie znane są pod wspólną nazwą algorytmów memetycznych. W zaproponowanym hybrydowym algorytmie indukcji drzew decyzyjnych modyfikacje czysto ewolucyjnego algorytmu obejmują zmianę sposobu inicjowania populacji początkowej oraz rozszerzenie operatora mutacji. W części eksperymentalnej rozdziału zbadano m.in wpływ częstości stosowania lokalnego poszukiwania na uzyskiwaną jakość klasyfikacji, rozmiary drzew oraz czas indukcji.

W rozdziale piątym omówiono specyfikę generowania drzew skośnych. Obejmuje ona głównie modyfikacje operatora mutacji oraz inny sposób wyliczania złożoności klasyfikatora w przypadku uwzględniania selekcji cech. Bezpośrednio wiąże się z tym zagadnieniem problem zbyt słabego dopasowania do danych, występujący zwłaszcza w dolnych partiach drzew, gdzie liczba obiektów uczących może być niewielka. Przedstawiono również mechanizm powiększania marginesów, rozumianych jako odległości przykładów uczących od granic decyzyjnych, którego zastosowanie poprawia właściwości generalizacyjne drzew. Podczas eksperymentalnej weryfikacji zaproponowanych rozwiązań zbadano m.in. wpływ różnych typów szumu na uzyskiwane klasyfikatory.

Rozdział szóstym dotyczy indukcji drzew mieszanych. Zadaniem algo-

rytmu ewolucyjnego jest nie tylko znalezienie najlepszej struktury drzewa i testów, ale również wybór reprezentacji testów w poszczególnych węzłach. Istotnym zagadnieniem jest ustalenie wzajemnej złożoności poszczególnych typów testów i zapewnienie mechanizmów zmiany rodzajów testów. W części eksperymentalnej pokazano m.in. możliwości automatycznego doboru reprezentacji do analizowanego problemu.

Zagadnienia dotyczące indukcji drzew uwzględniających koszty przedstawiono w rozdziale siódmym. Zaproponowane rozwiązanie umożliwia uwzględnianie podczas generowania drzew jednowymiarowych zarówno różnych kosztów błędnej klasyfikacji jak i kosztów testów. Najistotniejsza zmiana w stosunku do wersji nieczułych na koszty obejmuje zaproponowanie specyficznej funkcji dopasowania. Ponadto opracowano rozszerzenia operatorów różnicowania wykorzystujące informacje kosztowe. Przedstawione będą też wyniki eksperymentów, zarówno z jednym jak i dwoma rodzajami kosztów.

W rozdziale ósmym przedstawiono możliwości wykorzystania obliczeń równoległych i rozproszonych w ewolucyjnej indukcji drzew. Zaprezentowana będzie opracowana implementacja równoległa na klastrze obliczeniowym i przedstawione będzie uzyskiwane przyspieszenie obliczeń. Wąskim gardłem okazuje się migracja drzew pomiędzy procesorami i w tej sytuacji znaczące przyspieszenie może być uzyskane poprzez zredukowanie migracji. Prowadzi to w sposób nieuchronny do ewolucji rozproszonej i takie rozwiązanie zostało również zaprojektowane i eksperymentalnie zbadane.

Praca kończy się podsumowaniem, w którym podjęto próbę zaprezentowania zalet i wad zastosowania metod globalnej indukcji drzew decyzyjnych. Ponadto omówiono możliwe kierunki rozwoju zaproponowanych algorytmów.

Chciałbym bardzo serdecznie podziękować Panu Profesorowi Leonowi Bobrowskiemu, który zachęcił mnie do podjęcia trudu napisania tej monografii. Dziękuję również moim bardzo utalentowanym i pracowitym dyplomantom-asystentom Markowi Grzesiowi i Piotrowi Popczyńskiemu, z którymi miałem przyjemność pracować nad implementacją zaprezentowanych w książce rozwiązań. Specjalne podziękowania należą się również dr. Wojciechowi Kwedlo za bardzo uważne przeczytanie wstępnej wersji monografii i za wiele konstruktywnych uwag pod jej adresem.