ANALOGY-BASED REASONING IN CLASSIFIER CONSTRUCTION (WNIOSKOWANIE PRZEZ ANALOGIĘ W KONSTRUKCJI KLASYFIKATORÓW)

AUTOREFERAT ROZPRAWY DOKTORSKIEJ

ARKADIUSZ WOJNA

Przy rozwiązywaniu problemów człowiek często korzysta z wnioskowań przez analogię. Ich podstawą jest umiejętność korzystania z analogii przy kojarzeniu pojęć, faktów czy też sposobów postępowania. Stosowane w systemach uczących się [5, 10] metody wnioskowania przez analogię umożliwiają wnioskowanie o własnościach obiektów na podstawie podobieństw między obiektami. Rozprawa doktorska przynosi nowe wyniki o takich zaawansowanych metodach wnioskowania.

Przy wnioskowaniu przez analogię niezbędna jest informacja o zbiorze obiektów przykładowych. W rozprawie zakładamy, że obiekty opisane są przez wektory wartości cech. Przyjmujemy też, że każdemu obiektowi przypisana jest dodatkowo decyzja (wartość atrybutu decyzyjnego). Decyzje są jednak znane jedynie dla obiektów przykładowych. Problem klasyfikacji obiektów polega na wyuczeniu się ze zbioru obiektów przykładowych podejmowania właściwej decyzji dla obiektów z nieznaną decyzją. W rozprawie przedstawiono nowe metody wnioskowania przez analogię dla klasyfikacji obiektów.

Jedną z metodą wnioskowania przez analogię, stosowaną do problemu klasyfikacji, jest algorytm k najbliższych sąsiadów (k-nn) [4]. W metodzie tej decyzja dla nowego obiektu x jest podejmowana na podstawie ustalonej liczby k obiektów najbardziej podobnych do x, spośród obiektów przykładowych. Każdemu z tych k najbardziej podobnych obiektów algorytm przypisuje pewną wagę i dla obiektu x wybiera decyzję o największej sumie wag obiektów wyznaczonej dla obiektów podobnych do x.

Najważniejszym rezultatem rozprawy jest opracowanie dwóch nowych modeli klasyfikacji opartych na metodzie k-nn.

Pierwszy, z zaproponowanych w rozprawie modeli, otrzymano poprzez hybrydowe połączenie k-nn z indukcją reguł. Zaproponowane połączenie wykorzystuje minimalne reguły spójne, definiowane przez redukty lokalne [12, 23]. Redukt lokalny to minimalny podzbiór cech, który pozwala na odróżnienie obiektu przykładowego od innych przykładowych obiektów mających odmienną decyzję [12]. Takie minimalne reguły spójne mają dobre własności rozróżniania obiektów.

Zaproponowana w rozprawie kombinacja metody k-nn z indukcją reguł prowadzi w konsekwencji do uogólnienia modelu minimalnych reguł spójnych do postaci zależnej od metryki. Pokazujemy, że oryginalny model minimalnych reguł spójnych jest szczególnym przypadkiem modelu uogólnionego z metryką Hamminga.

W pracy [20] pokazano, że liczba wszystkich minimalnych reguł spójnych w oryginalnym modelu może być wykładnicza, zarówno względem liczby cech opisujących obiekty jak i liczby obiektów przykładowych. W rozprawie pokazujemy, że analogiczny fakt zachodzi dla modelu uogólnionego. W związku z tym generowanie wszystkich uogólnionych minimalnych reguł spójnych jest praktycznie często niewykonalne.

W pracy [1] zaproponowano efektywny, wielomianowy algorytm symulujący model klasyfikacji oparty o oryginalne minimalne reguły spójne. Wykorzystuje on pewne specyficzne własności minimalnych reguł spójnych, umożliwiające symulację ich działania bez jawnego generowania tych reguł. W rozprawie pokazujemy, że własności te są zachowane przez ugólnione minimalne reguły spójne, w związku z czym algorytm [1] działa poprawnie także dla modelu uogólnionego. Pokazaliśmy też pewną własność samego algorytmu, która pozwala zmodyfikować i przyspieszyć go do tego stopnia, że połączenie zmodyfikowanego algorytmu z metodą k-nn wydłuża jej czas działania tylko nieznacznie.

Rozszerzenie metody k-nn przy użyciu indukcji reguł jest pewnym rodzajem głosowania przez k najbliższych sąsiadów. Może ono być w naturalny sposób łączone z dowolnym innym systemem głosowania. Rozszerzenie to stanowi rodzaj weryfikacji i selekcji obiektów uznanych za podobne przez klasyczny algorytm k-nn. Wyniki eksperymentów pokazują, że zaproponowane rozszerzenie daje najlepszą skuteczność klasyfikacji przy połączeniu z modelem głosowania, w którym wagi głosów obiektów podobnych zależą od ich odległości do klasyfikowanego obiektu. Dla niektórych testowanych problemów klasyfikacyjnych model łączący k-nn i indukcję reguł zmniejszył błąd klasyfikacji względnie o kilkadziesiąt procent.

Dla wielu rzeczywistych problemów indukcja odpowiedniego globalnego modelu matematycznego z dostępnych zbiorów przykładów jest niewykonalna. Główną tego przyczyną jest fakt, że zjawiska opisywane przez rzeczywiste dane są złożone i w danych nie ma wystarczającej wiedzy do indukcji globalnego modelu z tych danych.

Dla takich rzeczywistych danych w rozprawie proponujemy metodę, nawiązującą do innego podejścia zwanego uczeniem transdukcyjnym [16]. W tym podejściu do konstrukcji modelu klasyfikacji proponuje się użycie nie tylko wiedzy zakodowanej w zbiorze obiektów przykładowych, ale również informacji o obiektach klasyfikowanych. W szczególności, podejście to umożliwia zastosowanie indywidualnego, lokalnego modelu klasyfikacji do każdego klasyfikowanego obiektu.

W rozprawie zaproponowano realizację podejścia transdukcyjnego [16] przyjmując jako punkt wyjścia metodę k-nn. W klasycznej metodzie k-nn definicja metryki jest ustalona dla całej przestrzeni obiektów, tymczasem gęstość i topologia obiektów w danych rzeczywistych są zwykle niejednorodne. Metoda zaproponowana w rozprawie pozwala, dla każdego klasyfikowanego obiektu, na indukcję oddzielnej metryki lokalnej, która jest następnie stosowana do klasyfikacji danego obiektu. Umożliwia to dopasowanie pojęcia podobieństwa do własności poszczególnych obiektów klasyfikowanych i podejmowanie właściwej decyzji również w specyficznych, odmiennych od pozostałych fragmentach przestrzeni obiektów.

Ważną cechą zaproponowanej metody jest jej pewnego rodzaju uniwersalność. Po pierwsze, do określenia metryki lokalnej można użyć dowolnej metryki, w szczególności wszystkich metryk z literatury używanych jako tzw. metryki globalne. Po drugie, do klasyfikacji obiektów można stosować wszystkie modele klasyfikacji, działające w oparciu o metrykę globalną, w szczególności zaproponowany wcześniej model łączący k-nn z indukcją reguł.

Wyniki eksperymentów opisanych w rozprawie potwierdzają, że model z indukcją metryk lokalnych może prowadzić do lepszych rezulatatów niż modele globalne. Okazał się on istotnie skuteczniejszy w trudnych problemach klasyfikacyjnych, w których błąd klasyfikacji modeli globalnych jest wysoki. W szczególności, dla jednego z trudniejszych testowanych problemów klasyfikacyjnych, polegającym na wykrywaniu miejsc w łańcuchach DNA, gdzie nastapiło połączenie dwóch fragmentów DNA przy operacji skrzyżowania, metoda z indukcją metryk lokalnych dała skuteczność klasyfikacji, która nigdy wcześniej nie była uzyskana i opisana w literaturze.

Ważnym zagadnienieniem dla zapewnienia wysokiej jakości klasyfikacji modeli opartych na k-nn jest dobór odpowiedniej miary podobieństwa pomiędzy obiektami. W rozprawie zbadano również własności różnych miar podobieństwa i zaproponowano istotne udoskonalenia dla niektórych miar znanych z literatury.

Dobrą metryką dla cech symbolicznych jest metryka VDM (ang. Value Difference Metric) [14]. W związku z tym wprowadzono analogiczne metryki [19, 18] dla cech numerycznych. Metryki te estymują rozkład prawdopodobieństwa wartości decyzji w zależności od poszczególnych wartości cech. Metryki zaproponowane dla cech numerycznych [19, 18] estymują ten rozkład na podstawie próbki obiektów przykładowych zależnej od wartości cechy, dla której wyznaczany jest rozkład. Jednakże wybór próbki w tych metrykach jest niezależny od gęstości obiektów przykładowych. Powoduje to, że próbka może być niereprezentatywna, tzn. może zawierać zbyt małą lub zbyt dużą liczba obiektów w próbce.

W rozprawie zaproponowano metrykę DBVDM (ang. Density Based Value Difference Metric). Podstawą jej definicji jest obserwacja polegająca na tym, że wybór próbki wyznaczającej rozkład decyzji zależy od gęstości obiektów przykładowych. W ten sposób eliminowany jest problem zbył małego lub zbyt dużego rozmiaru próbki. To udoskonalenie dało istotnę poprawę skuteczności klasyfikacji w najtrudniejszym testowanym problemie klasyfikacyjnym.

Zarówno w metryce DBVDM jak i w metryce WVDM [18] estymacja rozkładu wartości decyzyjnych dla danej wartości numerycznej ma liniową złożoność czasową względem liczby obiektów przykładowych. W rozprawie proponujemy metodę, która w przypadku obu metryk wylicza rozkłady decyzyjne dla wszystkich wartości numerycznych danej cechy w czasie $O(n \log n)$, gdzie n jest liczbą obiektów przykładowych. Umożliwia to policzenie odległości pomiędzy dwoma obiektami w czasie logarytmicznym $O(\log n)$, a nawet w czasie stałym po uprzednim przekształceniu wszystkich obiektów w zbiorze przykładowym. Zaproponowane przyspieszenie jest w praktyce niezbędne przy zastosowaniu metryk DBVDM i WVDM do rzeczywistych zbiorów danych.

W rozprawie przedstawiono i przetestowano również dwa algorytmy ważenia atrybutów. Przedstawione w rozprawie algorytmy, w odróżnieniu od wielu algorytmów ważenia atrybutów opisanych w literaturze [17], wyróżnia ich ogólność. Można je stosować do każdej metryki będącej liniowym złożeniem dowolnych metryk dla poszczególnych atrybutów. Dla wszystkich metryk liniowych opisanych i przetestowanych w rozprawie zaproponowane metody ważenia atrybutów poprawiły skuteczność klasyfikacji algorytmów korzystających z tych metryk.

Rzeczywiste dane zawierają często tysiące lub miliony obiektów przykładowych. W związku z tym kluczowym czynnikiem dla uzyskania odpowiedniej wydajności metod opartych na k-nn jest czas wyszukiwania obiektów podobnych. W rozprawie zagadnienie to szczegółowo zbadano i zaproponowano metodę indeksowania i wyszukiwania obiektów podobnych udoskonalającą metody znane z literatury [3, 9, 15].

Metody te opierają się na strukturze indeksującej w postaci drzewa. Obiekty przykładowe przypisane są do liści drzewa indeksującego i szukanie najbardziej podobnych obiektów jest przyśpieszane poprzez wykluczenie z przeszukiwania niektórych fragmentów drzewa indeksującego, tzn. niektórych jego poddrzew.

W stosunku do metod znanych z literatury [3, 9, 15], metoda zaproponowana w rozprawie zawiera dwa usprawnienia. Po pierwsze, jednokrokowe procedury podziału obiektów w danym węźle pomiędzy jego synów, wykonywane przy konstrukcji drzewa indeksującego, zastąpione zostały procedurą iteracyjną, która w przypadku drzewa binarnego posiada własność zbieżności do pewnego optymalnego podziału obiektów. Po

drugie, każda z trzech metod z literatury [3, 9, 15] wykorzystuje inne matematyczne kryterium do wykluczania gałęzi drzewa indeksującego z przeszukiwania. W rozprawie proponujemy metodę stosującą wszystkie trzy kryteria jednocześnie. Eksperymenty przeprowadzone z danymi rzeczywistymi pokazują, że nowa metoda indeksowania i wyszukiwania może być nawet do kilku razy efektywniejsza od metod [3, 9, 15], przy czym oba dodane usprawnienia mają istotne znaczenie dla zaobserowowanej poprawy efektywności. Nowa metoda pozwala na zastosowanie algorytmu k-nn do danych zawierających kilkaset tysięcy obiektów przykładowych. W szczególności, dla największego testowanego zbioru danych wyszukiwanie najbliższego sąsiada przy użyciu nowej metody było średnio 4000 razy szybsze niż wyszukiwanie bez indeksowania.

Częściowe wyniki z rozprawy zostały opublikowane i zaprezentowane na międzynarodowych konferencjach RSCTC, ECML i ICDM [2, 7, 6, 13, 21] oraz w czasopiśmie Fundamenta Informaticae [8, 22]. Wszystkie nowe metody zaproponowane w rozprawie zostały zaimplementowane w języku Java i przetestowane. Część z nich jest już dołączona do bieżącej wersji systemu wnioskowania z danych Rough Set Exploration System (RSES) [2, 11].

Literatura

[1] J. G. Bazan. Discovery of decision rules by matching new objects against data tables. *Proceedings of the First International Conference on Rough Sets and Current Trends in Computing*, wolumen 1424 serii *Lectures Notes in Artificial Intelligence*, strony 521–528, Warsaw, Poland, 1998. Springer-Verlag.

[2] J. G. Bazan, M. Szczuka, A. G. Wojna, M. Wojnarski. On the evolution of Rough Set Exploration System. *Proceedings of the Fourth International Conference on Rough Sets and Current Trends in Computing*, wolumen 3066 serii *Lectures Notes in Artificial Intelligence*, strony 592–601, Uppsala, Sweden, 2004. Springer-Verlag.

[3] S. Brin. Near neighbor search in large metric spaces. *Proceedings of the Twenty First International Conference on Very Large Databases*, strony 574–584, 1995.

[4] E. Fix, J. L. Hodges. Discriminatory analysis, non-parametric discrimination: Consistency properties. Raport instytutowy 4, USAF School of Aviation and Medicine, Randolph Air Field, 1951.

[5] J. Friedman, T. Hastie, R. Tibshirani. *The Elements of Statistical Learning*. Springer, New York, NY, 2001.

[6] G. Góra, A. G. Wojna. Local attribute value grouping for lazy rule induction. *Proceedings of the Third International Conference on Rough Sets and Current Trends in Computing*, wolumen 2475 serii *Lectures Notes in Artificial Intelligence*, strony 405–412, Penn State Great Valley, PA, 2002. Springer-Verlag.

[7] G. Góra, A. G. Wojna. RIONA: a classifier combining rule induction and k-nn method with automated selection of optimal neighbourhood. *Proceedings of the Thirteenth European Conference on Machine Learning*, wolumen 2430 serii *Lectures Notes in Artificial Intelligence*, strony 111–123, Helsinki, Finland, 2002. Springer-Verlag.

[8] G. Góra, A. G. Wojna. RIONA: a new classification system combining rule induction and instance-based learning. *Fundamenta Informaticae*, 51(4):369–390, 2002.

[9] I. Kalantari, G. McDonald. A data structure and an algorithm for the nearest point problem. *IEEE Transactions on Software Engineering*, 9(5):631–634, 1983.

[10] W. Klösgen, J. M. Żytkow, redaktorzy. *Handbook of Data Mining and Knowledge Discovery*. Oxford University Press, Inc., New York, NY, USA, 2002.

[11] A. Skowron, i in. Rough set exploration system. http://logic.mimuw.edu.pl/~rses, Institute of Mathematics, Warsaw University, Poland.

[12] A. Skowron, C. Rauszer. The discernibility matrices and functions in information systems. R. Slowinski, redaktor, *Intelligent Decision Support, Handbook of Applications and Advances of the Rough Sets Theory*, strony 331–362. Kluwer Academic Publishers, Dordrecht, 1992.

[13] A. Skowron, A. G. Wojna. K nearest neighbors classification with local induction of the simple value difference metric. *Proceedings of the Fourth International Conference on Rough Sets and Current Trends in Computing*, wolumen 3066 serii *Lectures Notes in Artificial Intelligence*, strony 229–234, Uppsala, Sweden, 2004. Springer-Verlag.

[14] C. Stanfill, D. Waltz. Toward memory-based reasoning. *Communications of the ACM*, 29(12):1213–1228, 1986.

[15] J. Uhlmann. Satisfying general proximity/similarity queries with metric trees. *Information Processing Letters*, 40(4):175–179, 1991.

[16] V. Vapnik. *Statistical Learning Theory*. Wiley, Chichester, GB, 1998.

[17] D. Wettschereck. *A Study of Distance-Based Machine Learning Algorithms*. Praca doktorska, Oregon State University, 1994.

[18] D. R. Wilson, T. R. Martinez. Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*, 6:1–34, 1997.

[19] D. R. Wilson, T. R. Martinez. An integrated instance-based learning algorithm. *Computational Intelligence*, 16(1):1–28, 2000.

[20] A. G. Wojna. Adaptacyjne definiowanie funkcji boolowskich z przykladow. Praca magisterska, Warsaw University, 2000.

[21] A. G. Wojna. Center-based indexing for nearest neighbors search. *Proceedings of the Third IEEE International Conference on Data Mining*, strony 681–684, Melbourne, Florida, USA, 2003. IEEE Computer Society Press.

[22] A. G. Wojna. Center-based indexing in vector and metric spaces. *Fundamenta Informaticae*, 56(3):285–310, 2003.

[23] J. Wróblewski. Covering with reducts - a fast algorithm for rule generation. *Proceedings of the First International Conference on Rough Sets and Current Trends in Computing*, wolumen 1424 serii *Lectures Notes in Artificial Intelligence*, strony 402–407, Warsaw, Poland, 1998. Springer-Verlag.