

Poznań 12 grudnia 2004

Jerzy Stefanowski

PROPOZYCJE TEMATÓW PRAC MAGISTERSKICH

Szanowni Państwo,

Zbliża się okres tworzenia listy tematów prac magisterskich. Państwo mogą poszukiwać tematów u każdego pracownika Instytutu posiadającego stopień przynajmniej doktora. Możecie także wcześniej **negocjować** z potencjalnymi promotorami sformułowanie tematu Was interesującego. Kierując się tą motywacją, informuję o zakresie moich zainteresowań (w ramach których możecie sami zaproponować swój temat) oraz o możliwych wstępnych tematach, z których w zamierzam w styczniu wybrać ostatecznie **około 3** tematów.

Można negocjować wybór i zakres samego tematu. Jeśli jesteś zainteresowana/-ny lub masz propozycje własnego tematu, którego mógłbym być promotorem, to proszę zgłoś się najpóźniej do końca roku (można także skorzystać z telefonu lub „korespondencji emaliowanej” Jerzy.Stefanowski@cs.put.poznan.pl). Proszę o jak najszybszą reakcję = kto będzie pierwszy, ten ma większe szanse na wybór i późniejszy sukces!

Moje główne obszary zainteresowań obejmują:

- Systemy uczące się (ang. *Machine Learning*)
- Eksploracja danych i odkrywanie wiedzy w bazach danych (ang. *Knowledge Discovery and Data Mining*)
- Web Intelligence (badanie zachowań użytkowników oraz analiza zawartości stron WWW; a także klasyfikacja przesyłek pocztowych)
- Analiza dokumentów tekstowych i systemy wyszukiwania informacji (patrz np. projekt Carrot2 – link www.cs.put.poznan.pl/dweiss/carrot)
- Teoria zbiorów przybliżonych
- Zastosowania praktyczne powyższych dziedzin oraz systemów wspomagania decyzji – zwłaszcza w medycynie, ekonomii, zarządzaniu i analizie danych finansowych oraz diagnostyce technicznej.

W ostatnich latach interesujące tematy moich prac mgr. dotyczyły między innymi: różnych podejść do indukcji reguł decyzyjnych, konstruowania złożonych systemów klasyfikujących (tzw. klasyfikatory bagging oraz n^2), grupowania wyników zapytań z internetowych systemów wyszukiwania informacji – rozwój systemu Carrot2 i metod wizualizacji wyników końcowych, modelowania niepewności w analizie danych z wykorzystaniem teorii zbiorów przybliżonych, rozszerzenia problemów poszukiwania reguł asocjacyjnych oraz realizacji rozszerzeń środowiska WEKA oraz różnych programów dydaktycznych.

Wstępna lista tematów, nad którymi się zastanawiam:

1. Klasyfikatory złożone – badania metod konstruowania odpowiedzi końcowej systemu na podstawie analizy działania klasyfikatorów składowych (w szczególności dotyczy rozwoju klasyfikatora n^2).
2. Algorytmy uczenia się reguł decyzyjnych z przykładów – rozwój metody opartej na podobieństwie nieprecyzyjnych opisów obiektów; implementacja właściwych algorytmów.
3. Metody budowy klasyfikatorów z danych, gdzie występują niezrównoważone licznosci klas decyzyjnych – także zastosowania medyczne przy rozpoznawaniu niezwykle ważnej diagnozy.
4. Wykorzystanie metod uczenia maszynowego do analizy danych pomiarowych charakteryzujących realizacje projektów programistycznych.
5. Wielokryterialna ocena reguł decyzyjnych dla zastosowań eksploracji baz danych.

6. Oprogramowanie podstawowych algorytmów grupowania danych dla zastosowań eksploracji danych.
7. Implementacja wybranych algorytmów dla potrzeb systemów wspomagających dydaktykę, np. system J.Żytkowa „49-ner” odkrywania praw naukowych i zależności funkcyjnych z danych lub wybrane metody wspomaganie decyzji..
8. Twoja własna propozycja tematu?

W wszystkim przypadkach wymagana jest dobra motywacja i chęć osiągnięcia sukcesu u zainteresowanej osoby, znajomość wybranych języków programowania i/lub gotowość nauczania się nowych narzędzi, znajomość j. angielskiego wystarczająca do sprawnego i szybkiego czytania literatury i dokumentacji, chęci do samodzielnych badań (szczególnie poszukiwań literatury w Internecie).

Opis wybranych pomysłów.

Tematy: Klasyfikatory złożone.

Opis: Praca nawiązuje do problemu budowania w sposób automatyczny systemów klasyfikujących. Większość dotychczasowych badań w uczeniu maszynowym i odkrywaniu wiedzy w bazach danych było prowadzonych nad tworzeniem i eksperymentalną oceną pojedynczych algorytmów uczących się (patrz np. wykład z uczenia maszynowego). W ostatnich latach obserwuje się rosnące zainteresowanie tworzeniem złożonych systemów klasyfikujących. Idea ta polega na integracji wielu pojedynczych algorytmów uczenia się w jeden system klasyfikujący. Celem takiej integracji jest przede wszystkim osiągnięcie lepszej trafności klasyfikacji niż otrzymanej w rezultacie użycia oddzielnie pojedynczych klasyfikatorów wchodzących w skład systemu. Praca magisterska wymagałaby rozwijania i implementacji jednego z wybranych sposobów budowania klasyfikatorów złożonych, np. *boosting* lub n^2 , a także przeprowadzenie jego eksperymentalnej oceny. Można zapoznać się z wcześniej prowadzonymi pracami mgr. w tym zakresie.

Tematy: Klasyfikatory dla nierównoważonych klas decyzyjnych..

Opis: Praca nawiązuje do problemu budowania w sposób automatyczny systemów klasyfikujących. W wielu przypadkach nie rozważano dostatecznie dobrze zadań, w których konieczne jest rozpoznanie przypadków z pewnej klasy mniejszościowej, która ma szczególne znaczenie w danym zastosowaniu. Problem polega na tym, że liczba przypadków z tej klasy jest znacznie mniejsza niż przypadków z innych klas, co bardzo utrudnia skonstruowania efektywnego klasyfikatora. Sytuacje takie występują w problem diagnostyki medycznej niebezpiecznej choroby oraz problemach analizy finansowej, np. w zakresie rozpoznawania przyczyn źle udzielonych kredytów lub bankructw. Celem pracy jest współpraca z promotorem nad nowymi algorytmami indukcji klasyfikatorów dla rozpoznawania wybranej klasy decyzyjnej o szczególnym znaczeniu. Planuje się wprowadzenie nowej metody tworzenia klasyfikatora regułowego, w którym zbiór reguł dla klasy mniejszościowej o szczególnym znaczeniu, będzie odpowiednio modyfikowany poprzez zmiany oceny tych reguł lub degenerowanie dodatkowych reguł tak, aby zwiększyć skuteczność rozpoznawania przypadków w tej klasy. Proces sterowania algorytmem indukcji reguł będzie wykorzystywał kryterium związane z miarami wrażliwości i specyficzności stosowane w analizie krzywej ROC (ang. *Receiver Operating Characteristic*) stosowanej do analizy klasyfikacji binarnej lub zastosowaniach medycznych. Metody te będą weryfikowane na danych z zakresu medycyny lub bankowości.

Temat: Algorytmy uczenia się z przykładów – środowisko WEKA.

Opis: Dla celów edukacyjnych i badawczych przygotowuje się i rozpowszechnia bezpłatnie oprogramowanie zawierające implementacje różnego rodzaju metod maszynowego uczenia się z przykładu. Przykładem takiego środowiska jest projekt „open-source” napisany w języku Java znany pod nazwa WEKA – szczegóły patrz strona <http://www.cs.waikato.ac.nz/ml/weka/>. Celem tego projektu jest wprowadzenie do tego środowiska nowych klas zawierających implementacje nowych lub znanych, a niezaimplementowanych, algorytmów uczenia maszynowego i eksploracji danych. Przykładowo w zeszłym roku z sukcesem wprowadzono implementacje algorytmu indukcji reguł o nazwie MODLEM. Wymagana jest dobra znajomość języka Java.