Streszczenie rozprawy doktorskiej

autorstwa mgr inż. Tomasza Orczyka

Klasyfikacja danych niekompletnych w oparciu o komitet klasyfikatorów

Teza:

Możliwe jest utrzymanie dokładności klasyfikacji na danych niepełnych poprzez wyłonienie komitetu klasyfikatorów działających w oparciu o wstępnie wyselekcjonowane cechy.

Celem pracy było opracowanie komitetu klasyfikatorów, przeznaczonego do klasyfikacji danych w których występują cechy nie posiadające zdefiniowanych wartości. Klasyfikator miałby być zdolny do przetwarzania niekompletnych wektorów cech, bez konieczności ich wstępnego uzupełniania, a klasyfikacja miałaby się odbywać w oparciu o wstępnie wyselekcjonowane cechy. W pracy zostały wyszczególnione **cele cząstkowe**:

1. Oszacowanie wpływu brakujących lub usuniętych cech obiektu na jakość klasyfikacji.
2. Opracowanie struktury komitetu klasyfikatorów.
3. Wybór klasyfikatorów działających w komitecie.
4. Opracowanie algorytmu podejmowania decyzji (fusera) komitetu klasyfikatorów.
5. Wybór cech dystynktywnych dla poszczególnych klas obiektów.
6. Testowanie opracowanego systemu na danych rzeczywistych.
7. Weryfikacja przydatności opracowanego klasyfikatora do budowy systemu oceny stopnia włóknienia wątroby u pacjentów z wirusowym zapaleniem wątroby typu C w oparciu o analizę parametrów krwi obwodowej.

W Rozprawie zbadano wpływ obecności wartości *null* w danych na powstawanie niepełnych wektorów referencyjnych (uczących) w zależności od rozmiaru podprzestrzeni cech, na jakiej pracują klasyfikatory składowe komitetu. Wpływ brakujących wartości cech na jakość klasyfikacji został również potwierdzony eksperymentalnie.

Na podstawie wniosków dotyczących rozkładu brakujących wartości cech wśród wektorów referencyjnych, zaproponowana została struktura komitetu klasyfikatorów, oparta na podziale przestrzeni cech na wektory jednoelementowe.

Dla zaproponowanej struktury komitetu klasyfikatorów, przetestowany został szereg konwencjonalnych klasyfikatorów w roli klasyfikatorów składowych komitetu. Jako klasyfikator składowy opracowywanego komitetu wybrany został jedyny klasyfikator, który odniósł korzyść z takiej struktury komitetu – klasyfikator *k*-NN.

Jako metodę wyłaniania decyzji komitetu klasyfikatorów zaproponowano uśrednianie bayesowskie, uzupełnione o współczynnik wagi dla poszczególnych klas obiektów referencyjnych, mający poprawić jakość klasyfikacji w odniesieniu do obiektów mało licznie reprezentowanych w zbiorze referencyjnym.

Komitet o takiej strukturze dokonuje wstępnego, dynamicznego filtrowania cech, w oparciu o wektor danych klasyfikowanych. Cechy nie posiadające zdefiniowanej wartości w tym wektorze są ignorowane w procesie klasyfikacji. W celu poprawy jakości klasyfikacji, zaproponowano metodę wstępnej selekcji cech, opartą o klasyfikator składowy proponowanego komitetu. Metoda ta, wykorzystuje ranking cech dystynktywnych dla poszczególnych klas ze zbioru referencyjnego do wskazania suboptymalnego podzbioru cech, w oparciu o które ma być prowadzona klasyfikacja.

Zaproponowany klasyfikator komitetowy SF*k*-NN/C został przetestowany na szeregu benchmarkowych baz danych, zawierających pełne dane rzeczywiste. W celu określenia wpływu braku wartości losowych cech na jakość klasyfikacji do danych, zarówno klasyfikowanych jak i referencyjnych, sztucznie wprowadzano wartości *null*, zastępując nimi istniejące wartości losowo wybranych cech. Badania zostały przeprowadzone zarówno bez jak i z wstępną selekcją cech.

Ostatecznie, klasyfikator został użyty do klasyfikacji rzeczywistych danych medycznych – analizy krwi pacjentów zakażonych WZW-C. Niezdefiniowane wartości w tym zbiorze danych występowały w sposób naturalny. Wyniki testów były spójne z uzyskanymi wcześniej wynikami na danych z których sztucznie usunięto niektóre wartości. Tym samym potwierdzona została przydatność proponowanego klasyfikatora SF*k*-NN/C do budowy systemu oceny stopnia włóknienia wątroby u pacjentów z wirusowym zapaleniem wątroby typu C.

Realizacja celów cząstkowych umożliwiła potwierdzenie tezy pracy. Potwierdzenie to ma charakter eksperymentalny, wsparty wynikami testów statystycznych.