Dr inż. Michał Koziarski

Rozmowa z autorem pracy:
„Imbalanced data preprocessing techniques utilizing local data characteristics"

Czy może Pan w syntetyczny sposób scharakteryzować tematykę Pańskiej pracy?

Tematyką mojej pracy była klasyfikacja danych niezbalansowanych, a konkretnie opracowanie nowych metod przetwarzania wstępnego tego typu danych, które pomogłyby zredukować negatywny wpływ niezbalansowania na jakość klasyfikacji. O niezbalansowaniu danych możemy mówić wtedy, kiedy jedna z klas, które chcielibyśmy rozpoznawać wśród naszych danych, jest słabiej reprezentowana niż inne klasy – w rzeczywistych danych niezbalansowanie jest bardzo powszechne, a pytanie dotyczy raczej tego, jak silny jest jego wpływ, a nie czy w ogóle występuje.

Większość standardowych algorytmów klasyfikacji jest słabo przystosowana do operowania na danych niezbalansowanych i ma tendencję do preferowania klas silniej reprezentowanych kosztem jakości predykcji na klasach rzadziej reprezentowanych – co jest przeciwieństwem tego, co zwykle chcielibyśmy osiągnąć. Jednym z podejść do odwrócenia tego trendu jest zastosowanie metod wstępnego przetwarzania danych, to jest algorytmów, które manipulują danymi treningowymi: albo poprzez redukcję liczby obserwacji z klas silniej reprezentowanych, albo poprzez tworzenie syntetycznych obserwacji z klas rzadziej reprezentowanych – i właśnie tworzeniem nowych metod tego typu zajmowałem się w mojej pracy.

Proszę powiedzieć, jak zrodziło się zainteresowanie uczeniem maszynowym?

Uczeniem maszynowym (i szerzej sztuczną inteligencją) zacząłem się interesować na pierwszych latach moich studiów, zasadniczo od pierwszego zetknięcia z tą tematyką. Z mojego punktu widzenia był to problem, którego rozwiązanie umożliwiłoby rozwiązanie za jednym zamachem wszystkich innych problemów (tj. posiadając dostatecznie silne algorytmy sztucznej inteligencji bylibyśmy w stanie rozwiązać wszystkie pozostałe problemy).

Rozwój sztucznej inteligencji jest postrzegany jako niezwykle ważny trend w nauce. Czy spodziewa się Pan spektakularnego przełomu, czy też będziemy mieli raczej do czynienia z ewolucją liczoną w dekadach?

Spektakularny przełom już się zasadniczo dokonał, po prostu nie tyle w AGI (ang. *Artificial General Intelligence* – generalnej sztucznej inteligencji), czyli w uproszczeniu w tym, jak przedstawiana jest sztuczna inteligencja w filmach i literaturze, ale w zastosowaniach uczenia maszynowego do rozwiązywania konkretnych problemów praktycznych. A te są wszędzie wokół nas: poczynając od zawartości naszych telefonów, przez różnego rodzaju diagnozowanie medyczne i przyspieszanie procesu wynajdywania lekarstw, a na tak fundamentalnych tematach jak predykcja struktur białkowych kończąc – żeby podać tylko kilka przykładów.

Już na obecnym etapie Pańskie algorytmy wspomagają diagnostykę onkologiczną – prosiłbym o kilka zdań na temat tego typu współpracy.

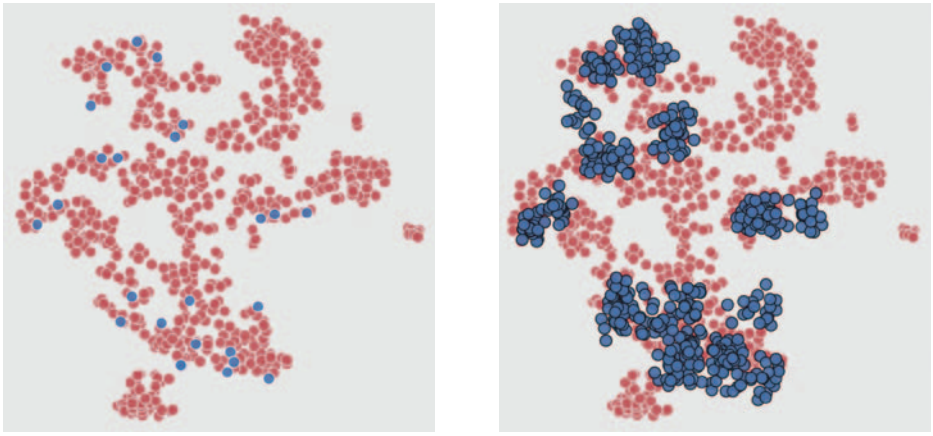
W części aplikacyjnej mojej pracy skupiłem się właśnie na automatycznej detekcji nowotworów na zdjęciach histopatologicznych, czyli mikroskopowych obrazach tkanki. Problem jest szczególnie istotny ze względu na fakt, że w Polsce brakuje nam aktualnie doświadczonych lekarzy histopatologów, więc jakiekolwiek przyspieszenie ich pracy miałoby duże znaczenie praktyczne.

Zadanie to sprowadza się do zaklasyfikowania regionów tkanki jako zdrowych lub zmienionych nowotworowo, a w tym drugim przypadku – dodatkowo do oceny stopnia złośliwości nowotworu. Jest to problem mocno niezbalansowany, bo tkanka

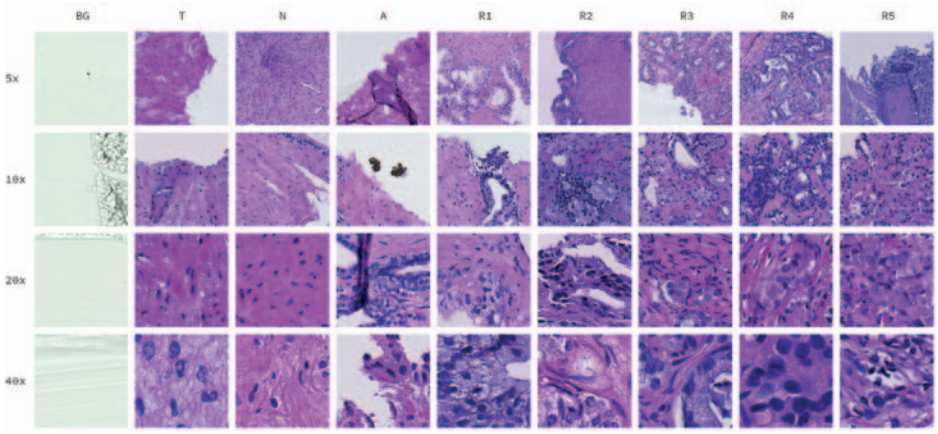
zdrowa jest zdecydowanie dominująca, a zwłaszcza niektóre stopnie złośliwości występują stosunkowo rzadko. Poza tym, od strony uczenia maszynowego zagadnienie jest ciekawe, bo zbiory danych jak i pojedyncze skany histopatologiczne są bardzo duże. Poza tym, mamy do czynienia z dużą niepewnością na poziomie etykiet, a cały proces klasyfikacji wymaga zastosowania sieci konwolucyjnych. Miałem więc możliwość zastosowania niektórych z opracowanych w mojej pracy algorytmów w finalnie opracowanym rozwiązaniu.

Czy Cyfronet spełnił Pańskie oczekiwania w zakresie dostępności oraz jakości zasobów obliczeniowych?

Bez dostępu do zasobów Cyfronetu moja praca zwyczajnie by nie powstała, a na pewno nie w swojej aktualnej formie – udostępnione zasoby umożliwiły mi opracowanie części bardziej złożonych obliczeniowo metod, a w przypadku tych mniej wymagających – przeprowadzenie dokładnych i obszernych prac eksperymentalnych.



Po lewej: przykład niezbalansowanego zbioru danych, z kolorem oznaczającym klasę obserwacji. Tradycyjne algorytmy klasyfikacji nie radzą sobie z silnie niezbalansowanymi danymi. Po prawej: zbiór po dodaniu syntetycznych obserwacji, wygenerowanych przy pomocy jednego z opracowanych algorytmów



Przykład fragmentów zdjęć histopatologicznych z różnych klas, klasyfikacja których była przedmiotem części aplikacyjnej doktoratu. Tego typu dane medyczne są często silnie niezbalansowane, ze znacząco mniejszą liczbą zdjęć zawierających zmiany nowotworowe