Analiza zbioru danych "Patient-level dataset to study the effect of COVID-19 in people with Multiple Sclerosis" z wykorzystaniem nadzorowanego algorytmu uczenia maszynowego One-Class SVM(Support Vector Machines)

Anna Mrozek, Bartosz Panek i Przemysław Jura

ARTICLE INFO

Keywords:
One-Class SVM
COVID-19

STRESZCZENIE

Artykuł analizuje zbiór danych pacjentów ze stwardnieniem rozsianym (SM), którzy przeszli COVID-19, przy użyciu nadzorowanego algorytmu One-Class SVM. Celem jest identyfikacja przypadków o zwiększonym ryzyku ciężkiego przebiegu infekcji. Umożliwi to zidentyfikowanie charakterystycznych cech pacjentów bardziej narażonych na powikłania.

1. Wprowadzenie

Pandemia COVID-19 miała znaczący wpływ na osoby z chorobami przewlekłymi, w tym na pacjentów ze stwardnieniem rozsianym (SM). SM jest przewlekłą chorobą autoimmunologiczną, która wpływa na układ nerwowy i może prowadzić do trwałego uszkodzenia neuronów oraz ograniczenia funkcji motorycznych oraz poznawczych.[1][2] Ze względu na charakter choroby i stosowane terapie immunosupresyjne, pacjenci z SM mogą być bardziej narażeni na cięższy przebieg infekcji wirusowych, w tym COVID-19.[3][4] Analiza danych od pacjentów z SM, którzy przeszli COVID-19, może dostarczyć ważnych informacji na temat ryzyka powikłań i zidentyfikować czynniki przyczyniające się do poważniejszych objawów, co ostatecznie może pomóc w lepszym monitorowaniu i opiece nad tą grupą pacjentów.

2. Cel

Celem badania jest wykorzystanie nadzorowanego algorytmu One-Class SVM do analizy zbioru danych pacjentów ze stwardnieniem rozsianym, którzy przeszli COVID-19, w celu zidentyfikowania cech pacjentów o zwiększonym ryzyku ciężkiej infekcji. Oczekuje się, że wyniki analizy dostarczą wskazówek do opracowania bardziej ukierunkowanych strategii opieki i środków zapobiegawczych dla pacjentów ze stwardnieniem rozsianym w kontekście potencjalnych przyszłych pandemii lub innych zagrożeń wirusowych.

3. Przegląd literatry

W literaturze medycznej i naukowej, od początku pandemii COVID-19, wiele badań skupiało się na analizie wpływu wirusa SARS-CoV-2 na osoby cierpiące na choroby przewlekłe i autoimmunologiczne, takie jak stwardnienie rozsiane (SM). Wyniki tych badań sugerują, że pacjenci z SM, zwłaszcza ci stosujący terapie immunosupresyjne, mogą być bardziej narażeni na cięższy przebieg COVID-19 oraz związane z nim powikłania[5][6]. U pacjentów z SM często obserwuje się obniżoną odporność oraz większą podatność na infekcje, co wynika zarówno z samej choroby, jak i z efektów leków tłumiących układ odpornościowy[7]. Dodatkowo w badaniach zwraca się uwagę na takie czynniki ryzyka jak wiek, płeć, rodzaj stosowanego leczenia oraz inne choroby towarzyszące, które mogą zwiększać ryzyko ciężkiego przebiegu infekcji u tej grupy pacjentów.

W ostatnich latach coraz większą popularność zyskuje zastosowanie algorytmów uczenia maszynowego, takich jak Support Vector Machines (SVM), do analizy danych medycznych. Algorytmy te pozwalają na wykrywanie wzorców i anomalii w dużych, zróżnicowanych zbiorach danych[8]. W szczególności algorytm One-Class SVM, używany do wykrywania anomalii, jest przydatny do identyfikacji przypadków wysokiego ryzyka w danych medycznych, gdzie przypadki nietypowe (np. pacjenci bardziej narażeni na ciężki przebieg COVID-19) występują rzadko. Badania pokazują, że One-Class SVM dobrze sprawdza się w analizie populacji pacjentów, gdy dostęp do dużych zbiorów danych o przypadkach zdrowych jest ograniczony, co często stanowi wyzwanie w analizie medycznej[9].

W kontekście badań nad COVID-19 i SM, algorytmy nadzorowanego uczenia maszynowego, takie jak SVM, umożliwiają dokładniejszą analizę czynników ryzyka oraz lepsze prognozowanie ciężkiego przebiegu choroby. Wyniki takich analiz mogą być przydatne nie tylko dla lekarzy, ale także dla decydentów zajmujących się zdrowiem publicznym, ponieważ pozwalają na opracowanie bardziej ukierunkowanych strategii opieki dla osób z SM, które są bardziej narażone na zagrożenia związane z wirusami, takimi jak COVID-19.

4. Metodologia

W badaniu zastosowano algorytm nadzorowanego uczenia maszynowego One-Class SVM w celu identyfikacji pacjentów z podwyższonym ryzykiem ciężkiego przebiegu COVID-19 w grupie osób ze stwardnieniem rozsianym.

Analiza obejmowała cechy kliniczne pacjentów, takie jak wiek, płeć, typ leczenia oraz choroby współistniejące, aby zidentyfikować wzorce związane z większą podatnością na powikłania.

4.1. Dataset

Stwardnienie rozsiane (MS) to przewlekła choroba autoimmunologiczna, która wywołuje stan zapalny w obrębie ośrodkowego układu nerwowego. Choroba prowadzi do różnych stopni utraty funkcji przez uszkodzenia mieliny oraz włókien nerwowych [10]. Osoby z MS są bardziej podatne na infekcje z powodu złożonego działania samej choroby, jej leczenia i naturalnego przebiegu [11]. Z inicjatywy COVID-19 and MS Global Data Sharing Initiative (GDSI) zbadano, jak leki immunosupresyjne lub immunomodulujące wpływają na COVID-19 i jego przebieg u osób z MS. GDSI miała na celu zwiększenie skali zbierania danych dotyczących COVID-19 i dostarczenie społeczności związanej z MS informacji opartych na danych podczas pandemii [12]. W ramach GDSI wybrano kluczowe zmienne obejmujące informacje o COVID-19, stopniu jego ciężkości, leczeniu, dane demograficzne, historię i nasilenie MS, stosowanie leków modyfikujących przebieg choroby (DMT), choroby współistniejące i wybrane zachowania związane ze stylem życia, takie jak palenie tytoniu. Globalna społeczność MS współpracowała, przekazując dokumentację statusu COVID-19 u osób z MS za pośrednictwem centralnej platformy udostępnionej przez QMENTA [13].

Ten zbiór danych został zebrany za pomocą narzędzia do szybkiego wprowadzania danych, które umożliwiało klinicystom, osobom ze stwardnieniem rozsianym (PwMS) lub ich przedstawicielom bezpośrednie wprowadzanie informacji do centralnej platformy GDSI. Narzędzie to zawierało kwestionariusz oparty na wcześniej ustalonych zmiennych i nie gromadziło bezpośrednich danych osobowych, aby chronić prywatność użytkowników. Narzędzie zostało wyłączone 3 lutego 2022 roku.

Zbiór danych obejmuje informacje o 1141 osobach ze stwardnieniem rozsianym (PwMS). Aby zapewnić zgodność danych z wytycznymi HIPAA, przeprowadzono proces deidentyfikacji. Po zebraniu danych dokonano oceny ryzyka związanego z małymi komórkami (SCRA), klasyfikując zmienne na trzy kategorie: bezpośrednie identyfikatory, zmienne wrażliwe i identyfikatory pośrednie. Bezpośrednie identyfikatory to zmienne, które mogą jednoznacznie zidentyfikować osobę, zmienne wrażliwe to te, które respondent może chcieć zachować w tajemnicy, natomiast identyfikatory pośrednie mogą zidentyfikować osobę, jeśli są połączone z danymi z innych zbiorów.

Ponieważ w danych nie zbierano imion pacjentów, deidentyfikacja skupiła się na datach i wieku pacjentów. Daty w kolumnie „stop-or-end-date-combined" zostały przesunięte o losową liczbę dni (między -15 a 15), aby uniemożliwić identyfikację na podstawie dat. Wiek pacjentów sklasyfikowano w cztery grupy: 0 dla osób w wieku 0–17 lat, 1 dla osób między 18 a 50 lat, 2 dla osób między 51 a 70 lat, oraz 3 dla osób powyżej 71 lat. Dzięki temu żadne dokładne wartości wieku powyżej 90 lat nie zostały ujawnione. Po klasyfikacji zmiennych i wdrożeniu odpowiednich środków ostrożności dane zostały zdeidentyfikowane i spełniają standardy HIPAA, zachowując jednocześnie wartość badawczą. Ponadto, aby zapewnić ochronę prywatności, zastosowano techniki takie jak K-anonimizacja oraz różnorodność.

Zbiór danych obejmuje zestaw z góry określonych zmiennych (n=47), takich jak płeć, kategoria wiekowa, typ MS, wynik EDSS, status palenia oraz kategoria BMI. Te zmienne dostarczają informacji o demografii pacjentów, ich stanie klinicznym oraz symptomach związanych z COVID-19. Szczegółowy opis typów zmiennych i ich statystyki znajduje się w sekcji „Opis Danych".

4.2. Opis metody

Jednoklasowy SVM (One-Class Support Vector Machine, OCSVM) to algorytm przeznaczony do wykrywania anomalii w zbiorze danych. Zamiast klasyfikować dane do dwóch lub więcej klas, jak w klasycznym SVM, OCSVM koncentruje się wyłącznie na danych normalnych.[14][15] Jego celem jest zbudowanie granicy wokół większości przypadków normalnych, tworząc "strefę normalności", która odróżnia standardowe przypadki od anomalii.

Główna zasada działania OCSVM opiera się na maksymalizacji marginesu między przypadkami normalnymi a granicą, która oddziela normę od anomalii. Dzięki temu model lepiej rozpoznaje dane odstające, które znajdują się poza wyznaczoną strefą. W algorytmie znajduje się hiperparametr „nu", który pozwala kontrolować czułość modelu – decyduje on o maksymalnym odsetku błędów marginesowych oraz liczbie wektorów nośnych, wpływając na balans między surowością a tolerancją modelu.[16][17]

Kluczowym elementem działania OCSVM jest funkcja decyzyjna oparta na hiperpłaszczyźnie, która oddziela dane od początku układu współrzędnych. Wzór na hiperpłaszczyznę wyrażany jest jako:

$$f(x) = \mathbf{w} \cdot x - \rho \tag{1}$$

gdzie:

- w – wektor normalny do hiperpłaszczyzny, wyznaczony podczas procesu uczenia,
- x – próbka danych,
- ρ – jest wartością progową (bias), która definiuje margines.

OCSVM posiada szereg parametrów, dzięki którym możliwe jest dostosowywanie modelu do specyficznych potrzeb:

- **Parametr** ν: Kontroluje liczbę obserwacji uznawanych za anomalie. Parametr ten przyjmuje wartości w przedziale od 0 do 1 i determinuje udział anomalii, jakie model może tolerować w zbiorze treningowym (np. jeśli $\nu = 0.05$, oznacza to, że około 5% próbek zostanie sklasyfikowanych jako anomalie).

- **Parametr** γ: Wpływa na kształt granicy decyzyjnej poprzez dopasowywanie modelu do danych. Mniejsza wartość γ oznacza „szerszą" granicę, obejmującą więcej punktów, natomiast większa wartość γ powoduje, że granica jest bardziej dopasowana do danych, co może zwiększać ryzyko nadmiernego dopasowania (overfitting).

- **Jądro (*ang. kernel*)**: One-Class SVM często wykorzystuje funkcje jądrowe, aby modelować nieliniowe granice decyzyjne i lepiej dopasować się do danych, które nie są liniowo rozdzielne w przestrzeni cech. Wybór jądra, zwany również „sztuczką jądra", wpływa na charakter dopasowania modelu i jego wydajność.

Podczas treningu OCSVM analizuje wyłącznie dane normalne, co sprawia, że jest szczególnie użyteczny w sytuacjach, gdy anomalie są rzadkie lub trudne do zidentyfikowania. Model staje się wtedy bardziej niezawodny w rzeczywistych zastosowaniach, takich jak wykrywanie oszustw, monitorowanie awarii czy zabezpieczanie sieci komputerowych. OCSVM może wykorzystać różne funkcje jądra, co umożliwia mu wykrywanie zarówno prostych, jak i bardziej złożonych odchyleń.[18]

4.3. Opis przeprowadzonych obliczeń
4.4. Wykorzystane metryki oceny

5. Wyniki

6. Opis wyników

7. Podsumowanie

8. Bibliografia

[1] https://diag.pl/pacjent/artykuly/jakie-sa-pierwsze-objawy-i-sposoby-leczenia-stwardnienia-rozsianego/

[2] https://www.medicover.pl/o-zdrowiu/stwardnienie-rozsiane-objawy-przyczyny-i-leczenie,6619,n,192

[3] https://pl.wikipedia.org/wiki/Stwardnienie_rozsiane

[4] https://ptsr.org.pl/strona/133,covid-19-a-sm

[5] Sormani MP, De Rossi N, Schiavetti I, Carmisciano L, Cordioli C, Radaelli M, et al. Disease modifying therapies and COVID-19 severity in multiple sclerosis. Ann Neurol. 2021 Apr

[6] Louapre C, Collongues N, Stankoff B, Giannesini C, Papeix C, Bensa C, et al. Clinical characteristics and outcomes in patients with coronavirus disease 2019 and multiple sclerosis. JAMA Neurol. 2020 Sep

[7] Simpson-Yap S, De Brouwer E, Kalincik T, Rijke N, Hillert JA, Walton C, et al. Associations of Disease-Modifying Therapies With COVID-19 Severity in Multiple Sclerosis. Neurology. 2021 Nov

[8] Schiff MA, Rae-Grant A, Gilden D, Franklin GM. Practice guideline: Disease-modifying therapies for adults with multiple sclerosis: Report of the guideline development, dissemination, and implementation subcommittee of the American Academy of Neurology. Neurology. 2019 Jan

[9] Erfani P, Mitchell AJ, Hameed S, Heydarpour P, Ghaffaripour R, Sahraian MA. Systematic review of health-related quality of life in multiple sclerosis patients: The impact of pharmacological treatments and lifestyle. J Neurol Sci. 2016 Dec

[10] Calabresi PA. Diagnosis and management of multiple sclerosis. Am Fam Physician. 2004 Nov

[11] Montgomery S, Hillert J, Bahmanyar S. Hospital admission due to infections in multiple sclerosis patients. Eur J Neurol. 2013 Aug

[12] Peeters LM, Parciak T, Walton C, Geys L, Moreau Y, De Brouwer E, et al. COVID-19 in people with multiple sclerosis: A global data sharing initiative. Mult Scler Houndmills Basingstoke Engl. 2020 Sep

[13] Simpson-Yap S, De Brouwer E, Kalincik T, Rijke N, Hillert JA, Walton C, et al. Associations of Disease-Modifying Therapies With COVID-19 Severity in Multiple Sclerosis. Neurology. 2021 Nov

[14] https://physionet.org/content/patient-level-data-covid-ms/1.0.1/

[15] https://www.geeksforgeeks.org/understanding-one-class-support-vector-machines/

[16] https://scikit-learn.org/dev/modules/generated/sklearn.svm.OneC

[17] https://www.baeldung.com/cs/one-class-svm

[18] https://medium.com/@roshmitadey/anomaly-detection-using-support-vectors-2c1b842213ed