

Wykorzystanie metody Isolation k-means w kontroli jakości produkcji spożywczej

Współczesne systemy monitorujące procesy produkcyjne, zwłaszcza w przemyśle spożywczym, generują ogromne ilości danych w czasie rzeczywistym. Efektywne wykrywanie anomalii, czyli odchyleń od normy, które mogą wskazywać na awarie maszyn, błędy operatorskie, zanieczyszczenia lub inne zagrożenia jakościowe, stanowi jedno z kluczowych wyzwań w zapewnianiu bezpieczeństwa i ciągłości produkcji. W tym kontekście szczególnego znaczenia nabierają metody analizy danych, które potrafią identyfikować niestandardowe wzorce bez wcześniejszego oznaczania danych jako poprawnych lub błędnych.



Wojciech Moszczyński

Jedną z obiecujących metod jest Isolation k-means – podejście łączące zalety klasycznej klasteryzacji k-means z mechanizmem izolacji, charakterystycznym dla technik wykrywania anomalii, takich jak Isolation Forest.

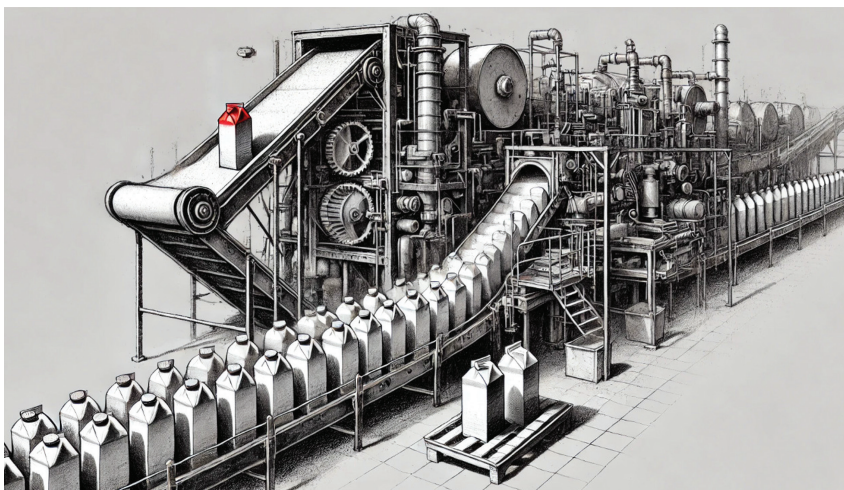
Isolation k-means to hybrydowa metoda analizy danych, która opiera się na dwóch podstawowych założeniach. Po pierwsze, wykorzystuje algorytm k-means do grupowania danych w klastry, czyli do identyfikacji naturalnych struktur i powtarzalnych wzorców w zbiorze danych. Po drugie, analizuje dystans pomiędzy punktami danych a centrami przypisanych im klastrów – zakładając, że dane leżące bardzo daleko od centrum klastra są potencjalnie anomaliami. Co więcej, algorytm może stosować dodatkowe kryteria izolacyjne, aby wyodrębnić punkty, które nie tylko są odległe, ale również rzadko współwystępują z innymi obserwacjami.

Metoda ta zyskuje szczególne znaczenie w kontekście przemysłu spożywczego, gdzie wykrywanie nieprawidłowości w czasie rzeczywistym może zapobiegać poważnym problemom, takim jak skażenia produktów, nieprawidłowe warunki temperaturowe czy niewłaściwe proporcje składników. Isolation k-means, dzięki swojej elastyczności i zdolności do działania na nieoznakowanych danych, może być skutecznie wykorzystywana do monitorowania parametrów produkcyjnych, identyfikowania odchyleń od standardowych profili procesów oraz wspomagania decyzji operatorów i inżynierów jakości.

W dalszych rozważaniach zostanie omówione, jak konkretnie można zaadaptować metodę Isolation k-means do danych szeregów czasowych w produkcji spożywczej oraz jakie korzyści może przynieść jej zastosowanie w rzeczywistych systemach kontroli jakości.

Nauka nienadzorowana

W nauce o danych w obszarze uczenia maszynowego wyróżniamy naukę nadzorowaną i nienadzorowaną. W nauce nadzorowanej maszyna jest instruowana przez człowieka, podczas gdy w nauce nienadzorowanej sama musi nauczyć się wszystkiego. W nauce nienadzorowanej tworzy się zdolność



do działania na nieoznakowanych danych. Oznacza to, że algorytm nie potrzebuje wcześniej przygotowanych etykiet, czyli informacji o tym, które dane są poprawne (normalne), a które są błędne (anomalie). Np. w klasycznym uczeniu nadzorowanym, dane muszą być wcześniej oznaczone, np. temperatura 60°C jest właściwą temperaturą, natomiast przekroczenie 95°C jest temperaturą zbyt wysoką.

Ale w praktyce – zwłaszcza w produkcji spożywczej – często nie mamy takich progów krytycznych. Trudno jest z góry oznaczyć, które przypadki to anomalie, szczególnie jeśli linia produkcyjna często zmienia asortyment. Występują wtedy bardzo krótkie szarże produkcyjne. Właśnie dlatego algorytmy działające na nieoznakowanych danych, czyli zgodnie z zasadą nauki bez nadzoru, są elastyczne i odpowiednie do tego typu środowiska produkcyjnego. Algorytmy te uczą się wzorców z samych danych. Same próbują wykryć, co jest „typowe” i co od tego znacząco odbiega.

Metoda klastrowania k-means

Klastrowanie k-means to jedna z najpopularniejszych metod eksploracyjnej analizy danych¹ w dziedzinie uczenia maszynowego. Jest najczęściej stosowaną metodą klasteryzacji,

¹ Eksploracyjna analiza danych (Exploratory Data Analysis – EDA) to pierwszy etap analizy danych, robiony w celu zrozumienie struktury, właściwości i wzorców występujących w danych.

czyli grupowania obiektów pod względem cech, m.in. w analizie zachowań klientów, segmentacji obrazów, eksploracji danych genetycznych, a także w wykrywaniu anomalii.

To nieprawda, że klasteryzacja może być przeprowadzona na podstawie wyłącznie dwóch (w przestrzeni 2D) lub maksymalnie 3 wymiarach (w przestrzeni 3D). W praktyce, dane często mają kilkanaście, kilkadziesiąt, a nawet setki cech (np. dane sensoryczne z instalacji produkcyjnej) i wtedy k-means też działa – tylko bez wizualizacji (analiza odbywa się w przestrzeni wielowymiarowej). Dlaczego o tym piszę? Ponieważ prace badawcze prowadzone są na zlecenie klienta, a ten zazwyczaj musi „zobaczyć” podział co sprowadza się do wykresu klastrow. Gdy badane zjawisko ma np. 14 cech, stosuje się metody sprowadzające te 14 cech do 2-3 cech kluczowych stosując liniową metodę PCA (*Principal Component Analysis*) lub nieliniową metodę, np. t-SNE (*t-distributed Stochastic Neighbor Embedding*).

Omówieniu zagadnień związanych ze współpracą z klientami poświęciłem wiele innych moich publikacji. Skupmy się na głównym temacie tak, aby artykuł ten był przydatny w praktyce produkcji spożywczej.

W praktyce produkcyjnej stosując metodę klastrowania k-means jest np. mięso, które ma trochę niższą temperaturę oraz ma jeszcze 23 inne cechy. Te cechy umieszczają to mięso np. w klastrze A. Mamy kolejne mięso, które ma inne wartości wspomnianych 24 cech i znajduje się w klastrze B. Mamy też klastry: C, D i E. W praktyce produkcyjnej nic to nie oznacza, to znaczy, że istnieją pewne różnice pomiędzy produktami znajdującymi się na linii produkcyjnej lecz nie mają one znaczenia. Produkty mogą być przydzielane do klastrow z powodu minimalnych różnic w wielkościach cech. Proces przyporządkowania do klastrow może być prowadzony wyłącznie dla celów technicznych.

Metoda klastrowania k-means należy do grupy uczenia nienadzorowanego (*unsupervised learning*), co oznacza, że nie wymaga wcześniejszego oznaczenia danych – algorytm samodzielnie znajduje struktury (klastry) w zbiorze danych. Głównym celem k-means jest podział zbioru danych na k niepokrywających się grup (klastrow), w taki sposób, aby punkty w obrębie tego samego klastra były do siebie możliwie podobne, a różniły się od punktów, w naszym przypadku grup mięs w innych klastrowach.

Jakość klastrowania mierzy się używając metryki SSE (*Sum of Squared Errors*), zwanej też WCSS (*Within-Cluster Sum of Squares*) – suma kwadratów odległości każdego punktu od

swojego centroidu (czyli środka klastra). Im mniejsze SSE, tym punkty są bliżej swoich centroidów, czyli klastry są bardziej „zbite”. W praktyce oznacza to, że mięsa z jednej szarży produkcyjnej można istotnie odróżnić od siebie. Może to być ważne w przypadku badania jakiś zdarzeń związanych na przykład z zanieczyszczeniem bakteriologicznym, analizą szkodliwości lub zanieczyszczenia produktów. Dzięki prowadzonym badaniom dyskryminacyjnym można znaleźć odpowiedź na wiele pytań.

Metoda wykrywania anomalii Isolation k-means

Isolation k-means to metoda wykrywania anomalii, która łączy dwa podejścia: klastrowanie (k-means) do identyfikacji typowych wzorców danych oraz izolację anomalii na wzór metod takich jak Isolation Forest, czyli skupienie się na identyfikacji punktów, które znacząco odstają od „normalnych” grup (klastrow). Na początku dane są grupowane przy pomocy klasycznego algorytmu k-means. Tworzone są centroidy, czyli środki klastrow, które reprezentują najczęstsze wzorce w danych.

Dla każdego punktu danych obliczana jest jego odległość od przypisanego centroidu. Im większa odległość, tym bardziej dany punkt nie pasuje do swojego klastra. Czyli np. system nadzorujący jakość na taśmie pakującej wykrył przykładowo 5 różnych grup (klastrow) kielbas. Jest to wciąż jeden asortyment (jeden index produktu), ale system wykrył podgrupy². Isolation k-means ocenia, jak bardzo punkt jest odizolowany od innych. Punkty bardzo odległe od jakichkolwiek podgrup są traktowane jako potencjalne anomalie.

Wprawdzie jest to metoda oparta na nauce bez nadzoru, ale można ustawić czułość wykrywania anomalii. Np. próg 95 percentyl największych odległości, powyżej którego punkty są uznawane za anomalie. Inna opcja to stosowanie skali znormalizowanej odległości (np. Z-score).

Jak przygotować dane do Isolation k-means?

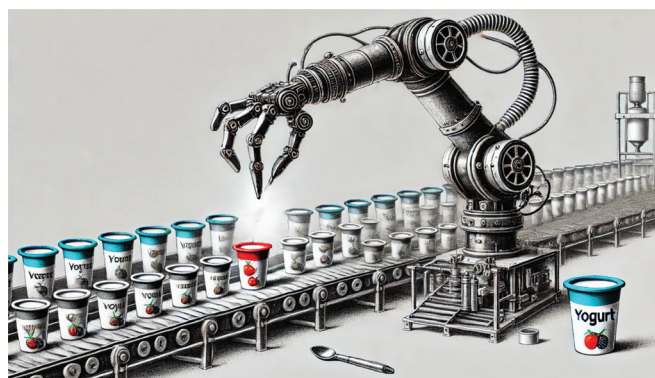
Ponieważ metoda bazuje na k-means, dane muszą być odpowiednio przygotowane – tak samo jak w klasycznym klastrowaniu.

Standaryzacja cech (skalowanie): wszystkie cechy powinny mieć podobną skalę (np. średnia 0, odchylenie 1), inaczej cechy z dużymi wartościami zdominują metrykę odległości. Alternatywnie można stosować normalizację cech, czyli transformować cechy do wartości z przedziału np. od -1 do 1.

Modele matematyczne działają zazwyczaj w ten sposób, że jeżeli coś ma dużą wartość nominalną to traktują tę cechę jako ważniejszą od cechy, która ma małą wartości nominalną. Tak więc w przypadku mięsa, np. temperatura 4°C byłaby mniej ważną cechą niż poziom wilgotności wynoszący 62%. W istocie obie wartości muszą występować w podobnym zakresie, aby model nie traktował ich w różny sposób.

Wybór cech: należy wybrać takie cechy, które mogą sensownie reprezentować „normalność” procesu. W przypadku produkcji spożywczej mogą to być: temperatura, czas produkcji, stężenia, ciśnienie, wilgotność itp. W produkcji przemysłowej występują rozmaite parametry, które niekoniecznie mają związek z jakością produkcji, a które mogą skutecznie zamulać cały

² W metodzie k-means liczba klastrow zostaje określona przez człowieka na początku analizy. Stosuje się tu tzw. „zasadę łokcia”.



proces analityczny. Np. sieci sensorów mogą mierzyć natężenie mocy silników, zużycie prądu w grzałkach lub mierzyć zużycie wody w wannach myjących. W istocie wartości te mogą nie mieć żadnego znaczenia w procesie analizy jakości mięsa.

Użycie danych historycznych: zbiór powinny zawierać jak najwięcej danych, aby k-means mógł nauczyć się typowych klastrów. W przypadku pomiarów streamingowych porcje pomiarowe powinny zawierać duże zbiory, co najmniej 100 produktów. Dane z systemów pomiarowych produkcji spożywczej mają postać szeregów czasowych. Warto przekształcić je w wektory cech, np. używając rolling window (okna czasowego), średnich, odchyłań, itp.

W przeciwieństwie do typowego klastrowania algorytmem k-means, który z zasady przypisuje każdy punkt (produkt) do jakiegoś klastra (podgrupy), Isolation k-means dopuszcza możliwość, że pewne punkty do żadnego klastra nie pasują – i wtedy są interpretowane jako anomalie.

Praktyczny przykład zastosowania Isolation k-means

Mamy linię produkcyjną w zakładzie mleczarskim, na której produkowany jest jogurt. Proces jest ściśle kontrolowany i obejmuje szereg kluczowych parametrów technologicznych, które muszą pozostawać w określonych granicach, aby zapewnić bezpieczeństwo i jakość końcowego produktu. Cechy te nie są sprecyzowane w procedurze pomiarowej.

Mamy 6 cech pomiarowych:

- Temperatura pasteryzacji mleka (°C)
- Czas fermentacji (minuty)
- Zawartość tłuszczu (%)
- Zawartość białka (%)
- pH produktu końcowego
- Liczba żywych kultur bakterii (log CFU/ml)

Dane z czujników i analizatorów są zbierane w czasie rzeczywistym w formie transakcyjnej. Każda pojedyncza partia 40 jogurtów ma swój unikatowy numer produkcyjny. Każda partia jogurtu może być reprezentowana jako punkt w sześciowymiarowej przestrzeni cech. Kluczowe jest, że nie analizujemy każdej cechy oddzielnie, ale traktujemy je jako jedną spójną obserwację, pełny profil procesowy danej partii. Algorytm k-means grupuje wiele wcześniejszych, poprawnych partii jogurtów w klastry, np. typowe ustawienia dla różnych wariantów cech. Dla każdego klastra tworzony jest centroid.

Gdy nowa partia zostaje wyprodukowana, jej profil (6 cech) jest porównywany z centroidami. Jeśli jest znacząco oddalona od najbliższego klastra (np. bardzo krótki czas fermentacji przy

jednoczesnym wysokim pH i niskiej zawartości kultur bakterii), traktowana jest jako potencjalna anomalia.

W trybie strumieniowym dane są analizowane na bieżąco. Dzięki możliwości szybkiego obliczenia odległości do centroidów, Isolation k-means może natychmiast identyfikować partie o nietypowych profilach, zanim produkt trafi do pakowania czy dystrybucji. W połączeniu z systemem alertów, operatorzy mogą zareagować natychmiast, zatrzymując linię lub wykonując dodatkowe testy laboratoryjne.

Korzyści z zastosowania metody Isolation k-means

Uwzględnia cały profil cech, a nie pojedyncze przekroczenie normy. To ważne, bo wiele problemów wynika z kombinacji cech, które osobno mieszczą się w granicach, ale razem tworzą nienaturalny przypadek.

Nie wymaga oznaczonych norm czy progów jakościowych, co jest bardzo przydatne w przypadku krótkich serii produkcyjnych. Anomalie produkcyjne są rzadkie i trudne do opisanego z góry.

Isolation k-means jest stosunkowo prostą metodą przez co dość szybką. Szybkość działania pozwala na użycie w czasie rzeczywistym (*real-time quality monitoring*).

Odporność na skalę danych po odpowiednim przeskalowaniu (np. standaryzacji) cech algorytm działa stabilnie nawet przy różnorodnych jednostkach miar. Jak wspomniałem wcześniej standaryzacja lub normalizacja danych jest tu obowiązkowa.

Podsumowanie

W niniejszych rozważaniach przedstawiono, jak metoda Isolation k-means może zostać z powodzeniem zastosowana do wykrywania anomalii w procesie produkcji spożywczej. Połączenie klasycznej klasteryzacji z analizą izolacji punktów umożliwia bardziej efektywne podejście do oceny jakości niż tradycyjne podejścia progowe.

Dzięki analizie całych profili cech procesowych, a nie tylko pojedynczych wartości, metoda ta potrafi identyfikować subtelne, ale potencjalnie groźne odstępstwa. Jej elastyczność, brak potrzeby definiowania progów i norm, a także zdolność działania w czasie rzeczywistym czynią ją szczególnie atrakcyjną w środowiskach produkcyjnych, gdzie niezawodność i szybkość reakcji są kluczowe.

W erze cyfrowej transformacji przemysłu spożywczego, metody takie jak Isolation k-means mogą stanowić fundament inteligentnych systemów kontroli jakości i predykcyjnego utrzymania ruchu wspierając bezpieczeństwo konsumenta i efektywność procesów.

Wojciech Moszczyński



Wojciech Moszczyński – absolwent Katedry Ekonometrii i Statystyki Uniwersytetu Mikołaja Kopernika w Toruniu; specjalista z zakresu ekonometrii, finansów, data science oraz rachunkowości zarządczej. Specjalizuje się w optymalizacji procesów produkcyjnych i logistycznych. Prowadzi badania w obszarze rozwoju i zastosowania sztucznej inteligencji. Od lat zaangażowany w popularyzację machine learning oraz data science w środowiskach biznesowych.