STUDIA INFORMATICA 2014
Volume 35 Number 2 (116)

Agnieszka NOWAK-BRZEZIŃSKA, Artur TUROS
Uniwersytet Śląski, Instytut Informatyki

WYKRYWANIE DANYCH NIETYPOWYCH – NOWE PODEJŚCIE

Streszczenie. Celem artykułu jest przedstawienie idei wykrywania nietypowych danych w dużych zbiorach danych poddanych grupowaniu. Autorzy przedstawiają własne podejście do klasycznej wersji algorytmu *k*-średnich. Modyfikacja prezentuje także definicje skupień odstających i skupień wpływowych. W pracy przedstawiono również wyniki przeprowadzonych badań wraz z analizą ich rezultatów.

Słowa kluczowe: odchylenia w danych, dane nietypowe, analiza skupień, skupienia odstające, skupienia wpływowe

UNUSUAL DATA EXPLORATION – NEW APPROACH

Summary. The goal of the article is to present the idea of discovering unusual data in large datasets in which the many clusters were created. Authors presents the methods which is modification of classical version of k-means algorithm. The modification introduces the concept of an influential and outlier cluster. The paper consists also of the results of the experiments with the analysis of it.

Keywords: outliers in data, unusual data, cluster analysis, outlier cluster, influential cluster

1. Wprowadzenie

Dla człowieka proces odkrywania wiedzy z danych jest umiejętnością naturalną. Od najmłodszych lat uczymy się automatycznego podejmowania decyzji. Każda bezwarunkowa reakcja istoty ludzkiej jest poprzedzona wcześniej zdobytą wiedzą, np. sklasyfikowaniem pewnych informacji. Zarówno w „naturalnym procesie wnioskowania", jak i w informatyce, szczególnie w dziedzinie data mining, segregowanie danych ma kluczowe znaczenie. Klasyfikację informacji człowiek podświadomie wykorzystuje też do wynajdywania sytuacji czy

zjawisk niecodziennych. Tak wykryte odchylenia w danych przed usunięciem powinny być poprzedzone dodatkową analizą ze strony eksperta, gdyż ich usunięcie (bez wstępnej analizy przyczyn nietypowości) może nieść ze sobą utratę cennych informacji.

Istnieje wiele zastosowań algorytmów grupowania danych [1, 2, 4, 8]. Bardzo użyteczną metodą eksploracji wiedzy z danych jest analiza skupień (ang. *cluster analysis*), której celem jest sklasyfikowanie obiektów wejściowego zbioru danych (często wielowymiarowego), tak żeby obiekty znajdujące się razem w grupie były jak najbardziej do siebie podobne, jednocześnie będąc jak najmniej podobne do obiektów z pozostałych grup [10]. W środowisku naukowym tematyka ta cieszy się dużą popularnością. Spowodowała ona powstanie znacznej liczby algorytmów, które w literaturze są najczęściej dzielone następująco: niehierarchiczne (k-optymalizacyjne), hierarchiczne, rozmyte, probabilistyczne [2, 5, 10]. Autorzy będą rozważać jedynie podejście niehierarchiczne. Analiza skupień (a w zasadzie rezultaty jej zastosowania) podlega dalszym rozważaniom, mającym na celu wykrycie pewnych anomalii w utworzonych grupach danych.

Metody wykrywania odchyleń na podstawie grupowania danych opierają się generalnie na założeniu, że regularne dane należą do większych skupisk, a obserwacje odstające nie. Identyfikowanie obserwacji odstających za pomocą algorytmów niehierarchicznych (np. *k*-średnich) przebiega w sposób trójstopniowy. Odchyleniami w takim ujęciu będą – zgodnie z podstawowymi założeniami tej grupy metod – wszystkie obserwacje, które tworzą pojedyncze lub małoliczne skupienia [6]. Elementami odstającymi są również obiekty wpływowe, które w znaczny sposób oddziaływają na jakość tworzonych grup i ich reprezentantów. Ponadto nie bez znaczenia jest tutaj separowalność skupisk [6]. Grupa z obserwacjami znacznie odbiegająca od pozostałych grup powinna być przedmiotem głębszej analizy (np. może być nieczęstym trendem w obserwowanej dziedzinie i wcale nie wynikać z błędów wpisów do bazy danych).

2. Wykrywanie odchyleń

Duże zbiory danych mają to do siebie, że trudno je scharakteryzować bez wstępnej analizy. Dopiero użycie pewnych metod statystyki opisowej, opartych na pomiarze średniej, wartości minimalnej czy maksymalnej, a także innych wartości charakteryzujących dane bądź użycie np. graficznych metod do wizualizacji danych pozwala stwierdzić, jaki jest rozkład danych w zbiorze. Jednak takie metody wstępnej (ogólnej) analizy zbiorów nie dają możliwości wykrywania np. odchyleń w danych. W przypadku dużych zbiorów wiadomo, że często mogą być one narażone na błędy przy pomiarze danych. Nietypowość danych nie musi jednak zawsze wynikać z błędu. Może wynikać także z faktycznej odmienności od reszty

danych zgromadzonych w zbiorze, który analizujemy. Bardzo istotne jest znalezienie metod efektywnie radzących sobie z problemem danych nietypowych. W niniejszym artykule zostaną przedstawione zarówno podejścia bardzo popularne, oparte na rozkładzie danych w zbiorze, jak i metody wywodzące się z analizy skupień. Ze względu na naukowe zainteresowania autorów związane z metodami grupowania danych właśnie tej drugiej części będzie w większości poświęcona ta praca.

Wśród metod opartych na rozkładzie danych było analizowanych wiele metod, jednak ostatecznie na potrzeby implementacji skupiono się na dwóch technikach: metodzie opartej na średniej i odchyleniu standardowych oraz metodzie opartej na rozstępie międzykwartylnym. Z metod wykrywania odchyleń w danych opartych na grupowaniu danych był analizowany algorytm *k*-średnich.

2.1. Metody opierające się na rozkładzie danych

Autorzy przyjęli następujące strategie identyfikowania obserwacji odstających przy uwzględnieniu kryterium odległości:

1) oparte na średniej i odchyleniu standardowych,
2) oparte na rozstępie międzykwartylnym.

Odpowiednio więc elementem odstającym będzie każda obserwacja V_i niemieszcząca się w przedziale: $\langle srednia(A) - p \cdot \sigma_A, srednia(A) + p \cdot \sigma_A \rangle$, gdzie σ_A to odchylenie standardowe dla A, *srednia(A)* zaś to wartość średnia wyznaczona dla zbioru wartości (A), z którego pochodzi obserwacja V_i. We wzorze tym p oznacza parametr sterujący wrażliwością na odchylenia. Widać bowiem, że im większa jest wartość p, tym większy (szerszy) jest przedział wartości uznawanych za typowe dane. Im mniejsza jest wartość p tym węższy jest przedział wartości uznawanych za typowe. Drugą analizowaną techniką była metoda oparta na rozstępie międzykwartylnym. Zgodnie z tą metodą elementem odstającym jest każda wartość atrybutu niemieszcząca się w przedziale: $\langle Q_1 - p \cdot IQR, Q_3 + p \cdot IQR \rangle$, czyli każda wartość, która jest położona przynajmniej o p razy IQR poniżej Q_1 lub p razy IQR powyżej Q_3, gdzie Q_1 to kwartyl pierwszy, Q_3 to kwartyl trzeci, p zaś to parametr skalujący wrażliwością na odchylenia w danych.

2.2. Algorytm *k*-średnich

Algorytm *k-średnich* jest jednym z najbardziej popularnych wśród algorytmów grupowania danych. Został zaproponowany w 1967 r. przez MacQueena [8]. Idea jego działania opiera się na powtarzaniu tego samego schematu czynności, do momentu kiedy w dwóch ostatnich iteracjach nie nastąpią żadne zmiany w przyporządkowaniu obiektów do poszczegól-

nych skupień. Parametrem wejściowym tej metody jest liczba grup (centroidów, punktów centralnych, stąd nazwa „średnich", środków) k, których początkowymi centrami są losowo wybrane obiekty ze zbioru. W wielu publikacjach można znaleźć informację, że już przy początkowym dobieraniu reprezentantów grup powinno się wybierać obiekty jak najbardziej oddalone od siebie. Takie „losowanie" zapewni optymalny przebieg procesu grupowania. W następnych iteracjach środki są wyznaczane przez obliczenie średniej odległości/średniego podobieństwa pomiędzy obiektami należącymi do poszczególnych grup, po czym następuje ponowne przyporządkowanie obiektów do skupień (przydzielane są do grupy, której środek znajduje się najbliżej/jest najbardziej podobny) [10]. Algorytm ten można przedstawić w postaci następującego pseudokodu [5, 2]:

1) podziel zbiór na k wstępnych skupień,
2) oblicz centroidy każdej z grup,
3) dokonaj ponownego podziału na skupienia, przyporządkowując obiekty do najbliżej leżącego/najbardziej podobnego centroidu,
4) powtarzaj kroki 2-3 do momentu zmian przyporządkowania.

Kluczowym elementem tej metody jest też dobór miary odległości/podobieństwa między obiektami. W podstawowym jej ujęciu wykorzystywana jest odległość euklidesowa, która jednak nie sprawdza się dla zbiorów mających dużą liczbę cech jakościowych. Większość istniejących narzędzi do analizy skupień zakłada w zbiorach danych wejściowych wielkości liczbowe i dla nich przewiduje jedynie implementację miar odległości (typowych dla danych liczbowych). Brak jest zaś narzędzi pozwalających szukać skupień w danych wielomiarowych reprezentowanych nie tylko numerycznie, lecz także w postaci cech typowo jakościowych. Z tego powodu jednym z celów pracy badawczej autorów stała się analiza skupień m.in. dla danych jakościowych, dlatego wybrano do grupowania, jako kryterium podobieństwa, miarę zaproponowaną przez Gowera [11]. Miara dzięki swojej elastyczności potrafi dostosować się zarówno do danych binarnych, ciągłych, jak i dyskretnych. Dla wektorów n-wymiarowych $x_i=[v_{i1},v_{i2},\ldots,v_{in}]$ oraz $x_j=[v_{j1},v_{j2},\ldots,v_{jn}]$ wyrażana jest następująco:

$$p(x_i, x_j) = \frac{\sum_{k=1}^{n} s_{ijk} \cdot w_{ijk}}{\sum_{k=1}^{n} w_{ijk}},$$ gdzie: w_{ijk} – waga, s_{ijk} – wartość podobieństwa obiektów x_i i x_j ze względu na k-tą. Dla danych jakościowych $s_{ijk} = \begin{cases} 1 & gdy \quad v_{ik} = v_{jk} \\ 0 & w.p.p. \end{cases}$. Dla danych ilościowych $s_{ijk} = 1 - \frac{|v_{ik} - v_{jk}|}{R_k}$, gdzie: R_k – rozstęp zmiennej k, v_{jk}, v_{ik} – wartości k-tej zmiennej dla obiektów x_i i x_j. Głównymi zaletami algorytmu k-średnich wydają się być jego prostota

i stosunkowo mała złożoność obliczeniowa $O(nki)$ (gdzie: i – liczba iteracji, k – liczba skupień, n – liczba obiektów).

2.3. Metoda wykrywania odchyleń opierająca się na algorytmie *k*-średnich

Po zakończonym procesie grupowania danych można rozpocząć eksplorację odchyleń. Wykrywanie obserwacji nietypowych za pomocą algorytmów grupowania przebiega trójetapowo (wyszczególnienie małolicznych skupień, znalezienie obiektów oraz grup obiektów wpływowych). Niezależnie od źródła i typu odchyleń często są one traktowane jednakowo – jako błędy w danych – i przez to często usuwane ze zbioru. Potrzebna jest więc metoda pozwalająca wykrywać obiekty nietypowe, ale również grupy danych odstających. Mogą być bowiem w dużych zbiorach danych małe grupy obiektów, stanowiące odchylenia w stosunku do całego zbioru, jednak niebędące błędami w danych. Nie powinno się ich usuwać, wręcz przeciwnie. Szczególnie warto podjąć próby oceny ich specyfiki czy źródła. W literaturze znane są prace, w których odchyleniami są tzw. małoliczne skupienia najczęściej utożsamiane ze skupieniami jednoelementowymi, co zdaniem autorów nie jest podejściem najwłaściwszym. Problem ten znalazł już swoje odzwierciedlenie w literaturze (np. [3, 6]), jednak brak jest jego matematycznego opisu. Autorzy proponują rozszerzenie klasycznego podejścia do wykrywania odchyleń pojedynczych przez zdefiniowanie tzw. skupień wpływowego oraz odstającego:

- skupienie wpływowe (ang. *influential cluster*), definiowane jako skupienie odchylone od średniego podobieństwa/średniej odległości pomiędzy skupieniami o wartość wyrażaną jako $p \cdot \left| \sqrt{j} \right|$ dla parametru p i liczby dziesiątek liczby n obiektów w zbiorze (j),
- skupienie odstające (ang. *cluster outlier*), definiowane jako skupienie, którego liczebność jest nie większa niż $\left\lceil \frac{2^j}{k} \right\rceil$, gdzie: j – liczba dziesiątek z liczby oznaczającej liczbę n obiektów w zbiorze, k – liczba skupień.

Wykrycie odchyleń będzie możliwe przy zastosowaniu na macierzy podobieństwa bądź odległości między obiektami a środkami skupień (centroidami) dla najlepszej iteracji (dostarczającej najmniejszej wartości sumy kwadratów różnic, TC). Wówczas algorytm wykrywania odchyleń będzie przebiegał w następujących 3 krokach.

1. Odchyleniem jest każde skupienie odstające $K_s = \{x_1, x_2, \ldots, x_i\}$, którego liczność ($i$) jest nie większa niż $\left\lceil \frac{2^j}{k} \right\rceil$, gdzie: K_s – s-te skupienie, j – liczba dziesiątek liczby n obiektów w zbiorze, i – liczba obiektów należących do skupienia, x_i – i-ty obiekt należący do skupienia K_s, k – liczba skupień.

2. Odchyleniem jest każdy obiekt wpływowy x_{is}, którego podobieństwo jest p razy mniejsze od średniego podobieństwa jego skupienia K_s: $p(x_{is}, K_s) < ((\overline{p(K_s)} - (p \cdot \overline{p(K_s)}))$, gdzie $p(x_{is}, K_s)$ – podobieństwo i-tego obiektu do centrum jego s-tego skupienia, $(\overline{p(K_s)}$ – średnie podobieństwo w s-tym skupieniu, p – parametr.
3. Odchyleniem jest każde skupienie wpływowe K_j, którego średnie podobieństwo do pozostałych skupień jest $p \cdot \sqrt{j}$ razy mniejsze od średniego podobieństwa pomiędzy skupieniami K: $\overline{p(K, K_s)} < ((\overline{p(K)} - (p \cdot \left|(\sqrt{j})\right| \cdot \overline{p(K)}))$, gdzie: $\overline{p(K, K_s)}$ – średnie podobieństwo skupienia K_s z osiąganych podobieństw do pozostałych skupień, $\overline{p(K)}$ – średnie podobieństwo pomiędzy skupieniami, p – parametr, j – liczba dziesiątek n liczby obiektów.

Dotychczasowe rozwiązania napotykają problemy np. z dużymi, symetrycznymi zbiorami danych. Tak jak sprawdza się przyjęcie progu p do wykrywania odchyleń wewnątrz skupień, tak często zastosowanie tego samego parametru pomiędzy grupami powoduje identyfikowanie skupień „poprawnych" jako odchylonych. Jest to spowodowane faktem, że podobieństwa pomiędzy grupami są przeważnie mniejsze i do tego bardziej zróżnicowane od podobieństw osiąganych wewnątrz grup. Rozwiązaniem mogłoby być zastosowanie dwóch odmiennych parametrów, jednak taka koncepcja wymaga wcześniejszego dokładnego przeanalizowania badanego zbioru, co powoduje utratę jej automatyzmu. Ponadto kłopotliwe byłoby określenie wielkości skupienia małolicznego. Rozwiązaniem tych problemów może być metoda wykrywania odchyleń zaproponowana przez autorów. W identyfikowaniu skupień wpływowych bierze ona pod uwagę wielkość zbioru oraz „naturalny" warunek stopu uzależniony od przyjętego parametru p (np. dla p = 0,1 jest to googol (10^{100}) obiektów). Zdaniem autorów w tak licznych zestawach danych nie powinno istnieć pojęcie grupy elementów nietypowych, ponieważ rzeczywiście może się w nim znaleźć wszystko. Wykrywanie skupień odstających poza wielkością zbioru bierze również pod uwagę liczbę grup. Taka definicja skupienia małolicznego pozwala na automatyczne dostosowywanie się zaproponowanej metody do różnych podziałów danych. Kluczową rolę odgrywa dobór odpowiedniej wielkości parametru p, który w eksperymentach ostatecznie ustawiono na 0,1 oraz 0,2.

Autorskie podejście różni się od klasycznego podejścia wykrywania odchyleń przy użyciu metody k-średnich tym, że wykrywa nie tylko pojedyncze obiekty (wpływowe) wewnątrz skupień, lecz także skupienia wpływowe. W tym celu parametr oceniający potem stopień odchylenia opiera się na liczności zbioru z danymi. Dodatkową cechą charakteryzującą autorskie podejście jest wyróżnienie rozmiaru skupienia małolicznego. W literaturze jest wiele prac poświęconych temu zagadnieniu. Każda przyjmuje jakiś prób doświadczalnie, jednak żadna nie przedstawiała modelu matematycznego, który określałby dla każdego zbioru indy-

widualnie wymaganą liczność skupienia w zależności od liczności zbioru oraz liczności grup powstałych. Wtedy skupieniem odstającym będzie takie skupienie, którego rozmiar nie przekracza wartości $2^j/k$. Im większa była liczba grup, tym były one mniej liczne. To automatycznie wpływało na zmniejszenie minimalnej liczności skupienia małolicznego. Przykładowo dla zbioru liczącego 110 tys. obserwacji rozpatrzmy 3 przypadki:

- podział 1: dla $k = 3$ minimalna liczność skupienia, by nie uznać go za odstające, wynosi $(2^6)/3=21{,}33 \sim 22$,
- podział 2: dla $k = 33$ skupienie małoliczne to zawierające 2 obiekty lub 1 obiekt (bo $(2^6)/33=1{,}93$),
- podział 3: dla $k = 333$ skupienie małoliczne = jednoelementowe, bo pierwiastek z liczby obiektów $(2^6)/333 = 0{,}19$.

Złożoność obliczeniowa tej techniki jest rzędu $O(nk)$ (gdzie n – liczba obiektów w zbiorze, k – liczba skupień). Co ciekawe, przy dobieraniu odpowiednio „mocy" parametrów p oraz k metoda ta rokuje rozszerzenie funkcjonalności wykrywania odchyleń o identyfikacje lokalnych elementów odstających (znanych z metod opartych na gęstości danych). Temat ten będzie podstawą dalszych badań autorów.

3. Implementacja metody wykrywania odchyleń dla środowiska R

W celu zbadania skuteczności omawianego rozwiązania konieczna była implementacja proponowanych rozwiązań. Wybrano stworzenie pakietu dla środowiska R pozwalającego na następujące funkcje: weryfikacja typów cech badanego zbioru danych (ilościowe, jakościowe); uzupełnianie i usuwanie brakujących wartości; wykrycie, usunięcie i podsumowanie informacji o odchyleniach przy wykorzystaniu następujących metod: z wartości średniej i odchylenia standardowego (średnia arytmetyczna, mediana), z rozstępu międzykwartylowego (obie metody nadają się dla zbiorów o przewadze cech ilościowych), z autorskiej metody opartej na algorytmie k-średnich (nieistotny typ danych); wyliczenie tabeli podobieństwa pomiędzy obiektami za pomocą miary Gowera; przeprowadzenie i podsumowanie procesu grupowania algorytmem k-średnich na podstawie tej miary. Skryptowi została nadana nazwa TTmining [5, 7]. Wśród najbardziej istotnych dla autorów cech R'a można wymienić: jego wieloplatformowość (dostępne są jego implementacje na wiele systemów z rodziny Unix, Windows oraz MacOS), uniwersalność (przez co autorzy rozumieją możliwość wczytania danych zapisanych w plikach o różnych formatach) oraz mnogość dostępnych dla programistów złożonych struktur danych. W momencie oddania artykułu do recenzji trwały starania o włączenie pakietu w R-project publicznie.

4. Eksperymenty

Rozdział ma na celu prezentację wyników przeprowadzonych eksperymentów wraz z ich analizą. W ramach analizowanych zbiorów danych znalazły się: Iris, Car evaluation, Credit Approval oraz Movement Libras. Zostaną one krótko scharakteryzowane.

Zbiór danych Iris zawiera 150 pomiarów opisujących zmienność morfologiczną trzech spokrewnionych gatunków kwiatu Irys (Iris setosa, Iris virginica, Iris versicolor). Jest to najprawdopodobniej zbiór danych najczęściej wykorzystywany do testowania algorytmów uczenia maszynowego. Kwiaty dwóch z trzech gatunków zostały zebrane na półwyspie Gaspe z tego samego pastwiska i w tym samym czasie. Pomiary były wykonywane przez jedną osobę wykorzystującą jeden przyrząd pomiarowy. Zestaw danych składa się z 50 próbek każdego z trzech gatunków. Wszystkie obiekty zostały opisane przez cztery atrybuty mierzące długość i szerokość kielicha oraz długość i szerokość płatków. Dodatkowym atrybutem jest gatunek kwiatu.

Drugi zbiór dotyczy rankingu samochodów (ang. *Car Evaluation Data Set*). Jest to typowo jakościowy zestaw danych o interesującym rozkładzie, który może być przydatny np. do testowania metod badających struktury danych. Zawiera ranking samochodów oparty na ocenie akceptowalności: ceny (kupno, utrzymanie), walorów technicznych (komfort, liczba drzwi, dozwolona liczba pasażerów, wielkość bagażnika) oraz bezpieczeństwa. Zbiór składa się z 1728 ocenionych samochodów opisanych sześcioma cechami: buying, maint, doors, persons, lug_boot, safety. Siódmym atrybutem jest klasa.

Trzecim zbiorem wybranym do analizy jest zbiór dotyczący decyzji kredytowych, a konkretnie wydania kart kredytowych (ang. *Credit Approval Data Set*). Charakteryzuje się dobrym podziałem na cechy jakościowe (o małej i dużej liczbie wartości) oraz cechy ilościowe. Zbiór składa się z 690 aplikacji kredytowych, z których każda opisana jest za pomocą piętnastu cech. Atrybut decyzyjny informuje o rodzaju decyzji kredytowej (przyjęta, odrzucona).

Ostatnim analizowanym zbiorem jest zbiór Movement Libras. Liczy 360 obserwacji będących zapisami 15 różnych znaków w brazylijskim języku migowym. Zbiór zawiera po 24 obserwacje dla każdej z 15 klas opisanych 90 atrybutami numerycznymi i 1 atrybutem etykietującym klasę.

W tabeli 1 przedstawiono wyniki osiągnięte dla metod opartych na odległości (są omówione w rozdziale 2.1).

Tabela 1

Wyniki eksperymentów dla metod opartych na odległości

Nazwa zbioru	p	Średnia arytmetyczna i odchylenie	Rozstęp międzykwatylny
Iris	1,5	46 / 30,67%	4 / 2,67%
	2	11 / 7,33%	1 / 0,67%
	3	1 / 0,67%	0 / 0%
Credit	1,5	207 / 30%	219 / 31,7%
	2	111 / 16%	182 / 26,3%
	3	52 / 7,54%	116 / 16,81%
Car	1	0 / 0%	0 / 0%
Movement Libras	1,5	283 / 78,6%	4 / 1,11%
	2	119 / 33,1%	0 / 0%
	3	2 / 0,6%	0 / 0,6%

Tabela 2 przedstawia z kolei wyniki uzyskane dla metod opartych na rezultatach grupowania algorytmem *k*-średnich. Pokazano w niej wyniki dla klasycznej metody *k*-średnich oraz autorskiej modyfikacji.

W tabeli 1 można zauważyć, iż w miarę wzrostu wartości parametru p zmniejsza się liczba znalezionych odchyleń pojedynczych, co jest sytuacją spodziewaną. Skoro rozszerzamy zakres możliwych wartości uznawanych za typowe, maleje liczba obserwacji niespełniających zadanego zakresu i określanych jako odstające. Generalnie metody statystyczne oparte na średniej i odchyleniu standardowym nie najlepiej sprostały postawionemu zadaniu. Dopiero test dla największej z badanych mocy ($p = 3$) dał satysfakcjonujące wyniki, tzn. wykryto rozsądną (zgodną z rzeczywistością) liczbę obiektów odstających, gdyż np. w zbiorze Iris jedynie jedna obserwacja różni się faktycznie od reszty obserwacji, tj. obiekt 16, który ma nadzwyczaj szeroki płatek kielicha. W drugim z analizowanych zbiorów znajdują się bardzo zróżnicowane obserwacje, w związku z tym liczba wykrytych obserwacji nietypowych jest tak duża. W przypadku zbioru Car – który jest opisany cechami jakościowymi – analizowano jedynie jeden przypadek dla wartości $p = 1$. Jak można zaobserwować w tabeli 1, działanie obu metod opartych na rozkładzie danych jest bardzo zróżnicowane. Dużo lepiej poradziły sobie metody oparte na grupowaniu danych metodą *k*-średnich. Wyniki tych badań przedsta-

wia tabela 2. Widzimy tam, że zaproponowana metoda wykrywa jako odstające rozsądną liczbę obserwacji w stosunku do liczności całego zbioru.

Tabela 2

Wyniki eksperymentów dla metod opartych na analizie skupień

Nazwa zbioru	p	*k*-średnich	
		autorska	klasyczna
		Najlepsza iteracja	Najlepsza iteracja
Iris	0,1	0 33,317%	101 67,33%
	0,2	0 0%	0 0%
Credit	0,1	39 5,65%	162 23,4%
	0,2	4 0,58%	8 1,16%
Car	0,1	463 26,79%	615 35,59%
	0,2	24 1,39%	344 19,91%
Movement Libras	0,1	4 1,11%	4 1,11%
	0,2	0 0%	0 0%

5. Podsumowanie

Problem wykrywania odchyleń w danych jest nietrywialny. Zbiory rzeczywiste najczęściej są wielowymiarowe i wielotypowe (numeryczne i nominalne). Fakt ten pociąga za sobą dodatkowe trudności w czasie eksploracji odchyleń. Kiedy analizowane są zbiory wielowymiarowe, to nie konkretna wartość, ale połączenie wymiarów rozumiane jest jako ekstremum. W danych opisanych za pomocą wielu cech istnieje możliwość, że cały obiekt będzie sklasyfikowany jako odchylenie z powodu jednej wartości (sytuacja możliwa przykładowo w metodach statystycznych). Usunięcie w tej sytuacji odchylenia, nawet w przypadku faktycznego błędu, może doprowadzić do utraty cennych informacji, które są przechowywane w pozostałych atrybutach. Ogromnie pomocne w takich sytuacjach są metody wykrywania odchyleń w danych oparte na wcześniejszym poszukiwaniu skupień w danych (np. algorytm k-Means). W pracy autorzy przedstawili opis podejść literaturowych oraz proponowanych dla nich modyfikacji. Implementacja własnych metod wykrywania odchyleń w postaci stworzonego dla środowiska R pakietu TTMining pozwoliła wykonać odpowiednią liczbę ekspery-

mentów. W rozdziale 4 przedstawiono analizę wyników uzyskanych w ramach prowadzonych badań.

BIBLIOGRAFIA

1. Larose D.: Odkrywanie wiedzy z danych, wprowadzenie do eksploracji danych. Wydawnictwo PWN, Warszawa 2006.
2. Han J., Kamber M., Pei J.: Data Mining: Concepts and Techniques. Elsevier, San Francisco 2012.
3. Nowak-Brzezińska A.: Eksploracja odchyleń w regułowych bazach wiedzy. Studia Informatica, ZN Pol. Śl., Vol. 33, No. 2A (105), Gliwice 2012.
4. Hawkins D.: Identification of Outliers. Chapman and Hall, London 1980.
5. Tomkowicz M.: Wpływ odchyleń na jakość grupowania danych wielowymiarowych. Praca magisterska, Uniwersytet Śląski, Katowice 2013.
6. Nowak-Brzezińska A.: Wykrywanie reguł nietypowych – metody oparte na analizie skupień. Studia Informatica, ZN Pol. Śl., Vol. 34, No. 2A (111), Gliwice 2013.
7. Turos A.: Analiza metod wykrywania odchyleń w danych wielowymiarowych. Praca magisterska, Uniwersytet Śląski, Katowice 2013.
8. MacQueen J.: Some Methods for classification and Analysis of Multivariate Observations. University of California, 1967.
9. Tryon R.: Cluster Analysis. McGraw Hill, New York 1939.
10. Xu R., Wunsch D.: Clustering. Wiley, New York 2008.
11. Myatt G., Johnson W.: Making sense of data. Wiley, New York 2009.

Wpłynęło do Redakcji 2 stycznia 2014 r.

Abstract

The aim of the study is to present the idea to detect unusual (outlier) data in large data sets subjected to clustering.

The authors present new approach to the classic version of k-means algorithm. The definitions of outlier cluster and influential cluster are included in the paper as well as the results of the experiments made for real data sets.

Adresy

Agnieszka NOWAK-BRZEZIŃSKA: Uniwersytet Śląski, Instytut Informatyki, ul. Będzińska 39, 41-200 Sosnowiec, Polska, agnieszka.nowak@us.edu.pl.

Artur TUROS: turos.artur@gmail.com.