STUDIA INFORMATICA 2014
Volume 35 Number 2 (116)

Agnieszka NOWAK-BRZEZIŃSKA, Tomasz XIĘSKI
Uniwersytet Śląski, Instytut Informatyki

WYDOBYWANIE WIEDZY Z DANYCH ZŁOŻONYCH

Streszczenie. Artykuł przedstawia uniwersalną metodę wydobywania wiedzy z danych złożonych, uwzględniającą wykorzystanie technik opisu danych, algorytmów analizy skupień oraz efektywnych środków wizualizacji wydobytej wiedzy. Charakterystyczną cechą opisywanej metody jest zastosowanie dwuetapowego grupowania danych.

Słowa kluczowe: wizualizacja skupień, grupowanie, dane złożone, DBSCAN, AHC

DISCOVERING KNOWLEDGE FROM COMPLEX DATA

Summary. This work presents a universal knowledge discovery method from complex data, which takes into account the usage of data description techniques, cluster analysis algorithms and effective means of visualization of the discovered knowledge. A characteristic feature of this method is the usage of a two-stage clustering process.

Keywords: cluster visualization, clustering, complex data, DBSCAN, AHC

1. Wprowadzenie

Wydobywanie wiedzy ukrytej w danych stało się szczególnie istotne w ostatnich latach, gdy mamy do czynienia z nieustannie rosnącą liczbą informacji przechowywanych w bazach i hurtowniach danych. Dane te są gromadzone, ponieważ zakłada się, że mogą być źródłem nieznanych, potencjalnie użytecznych wzorców, korelacji i trendów. Odkryte wzorce mogą mieć skomplikowaną strukturę, przez co są trudne do dalszej analizy. Jednakże to nie tylko nadmierna ilość danych wpływa na trudności badawcze. Bardziej istotnym czynnikiem jest ich złożona struktura zarówno pod względem dużej liczby atrybutów opisujących każdy

obiekt danych, jak i użytych typów danych. Informacje zakodowane w bazie często opisane są atrybutami różnych typów, z wliczeniem w to wartości binarnych, dyskretnych, ciągłych, kategorycznych, tekstowych czy reprezentujących daty. Tego typu dane można nazwać złożonymi i będą one podstawą analizy w niniejszej pracy. Złożonym zbiorem danych będzie nazywana również struktura grup (i ich reprezentantów) utworzona w wyniku zastosowania kombinacji algorytmów analizy skupień opisywanej w niniejszym artykule.

Celem pracy jest przedstawienie uniwersalnej metody reprezentacji i wydobywania wiedzy ze złożonych zbiorów danych rzeczywistych o dużej liczebności, uwzględniającej wykorzystanie statystyki opisowej, algorytmów analizy skupień oraz efektywnych środków wizualizacji wydobytej wiedzy. Mianem wiedzy na potrzeby niniejszego artykułu określa się wyrażoną w postaci wzorców, trendów czy korelacji „informację odnośnie do otaczającego świata, która umożliwia ekspertowi podejmowanie decyzji" [12]. Niniejsza metoda stanowi bezpośrednie nawiązanie do procesu graficznej analizy eksploracyjnej (ang. *visual data mining*) opisywanego w pracach [13, 9], którego celem jest wykorzystanie naturalnych zdolności kognitywnych człowieka w procesie analizy danych. Skuteczność proponowanego podejścia zostanie przetestowana na zbiorze danych rzeczywistych opisującym funkcjonowanie urządzeń nadawczo-odbiorczych operatora telefonii komórkowej, choć należy nadmienić, że może ono zostać zastosowane do dowolnego zbioru danych.

2. Struktura zestawu danych rzeczywistych cell_loss

Zestaw danych *cell_loss*, stanowiący przedmiot analiz, agregował dane dotyczące urządzeń nadawczo-odbiorczych (tzw. komórek) rozlokowanych w aglomeracji śląskiej, pochodzące z okresu od kwietnia 2010 do stycznia 2011 roku. Składa się z 143 486 obiektów[1] zapisanych w jednej tabeli. Pomiar dostępności danej komórki był wykonywany w interwałach godzinnych. Struktura każdego rekordu danych została opisana w [7], jednakże najważniejsze atrybuty (w kontekście badań) to: identyfikator określonej komórki (*cellname*), identyfikator obszaru geograficznego, w którym zlokalizowana jest dana komórka (*obszarId*), identyfikator sektora kierunku, w którym komórka nadaje (*sektorId*), identyfikator kontrolera, który steruje pracą danej komórki (*kontrolerId*), identyfikator producenta danej komórki (*dostawcaId*), bezwzględny procent niedostępności określonej komórki w danej godzinie pomiarowej (*strata*), identyfikator zdarzenia[2] (*zdarzenieId*), data i godzina zajścia oraz końca danego zdarzenia (*start*, *koniec*), data wykonania pomiaru (*data*).

[1] Pojęcie obiektu jest w tym przypadku utożsamiane z pojedynczym rekordem tabeli.

[2] Zdarzenie jest rejestrowane, gdy komórka jest poddawana naprawie lub jest niedostępna z innego powodu.

Celem analiz było wykrycie najbardziej problematycznych urządzeń nadawczo-odbiorczych (tj. charakteryzujących się wysokim średnim poziomem niedostępności i dużą częstością zdarzeń), z wykorzystaniem algorytmów analizy skupień, jak również identyfikacja ewentualnych przyczyn niedostępności (np. w wyniku wykrycia korelacji między kontrolerami a parametrami pracy sterowanych przez nie urządzeń). Na tej podstawie operator telefonii komórkowej może dokonać optymalizacji w strukturze sieci, co powinno się bezpośrednio przełożyć na poprawę jakości oferowanych usług. Należy również zaznaczyć, że wszystkie wyniki eksperymentów zostały zweryfikowane przez eksperta i mogą znaleźć zastosowanie w systemach monitorujących (np. pracę sieci komórkowej).

3. Proponowane podejście wydobywania wiedzy

Jednym z założeń proponowanego podejścia wydobywania wiedzy jest wykorzystanie technik analizy skupień. Algorytmy analizy skupień mają na celu pogrupowanie źródłowego zbioru danych w celu wykrycia nowych i potencjalnie użytecznych zależności między poszczególnymi obiektami w sposób nienadzorowany. Jednakże w kontekście analizy rzeczywistych danych złożonych problematyczna staje się interpretacja uzyskanych wyników (w postaci struktury grup), nawet przy uwzględnieniu wyłącznie przeglądu wygenerowanych reprezentantów skupień. Literatura przedmiotu [6] często zakłada określoną relację między liczbą obiektów ze zbioru danych (ozn. N) oraz liczbą stworzonych skupień (ozn. lSk), według której $lSk << N$. Dzięki temu utworzona liczba grup jest relatywnie niewielka, przez co możliwe jest ich porównanie w rozsądnym czasie z wykorzystaniem wrodzonych zdolności kognitywnych człowieka. Niestety w rzeczywistych zastosowaniach (np. podczas segmentacji klientów sieci hipermarketów) liczba obiektów jest mierzona w setkach tysięcy bądź milionach, przez co również algorytmy analizy skupień mogą wytworzyć dużą liczbę grup (rzędu kilku tysięcy bądź większą). Tego typu struktura jest bardzo trudna w analizie (szczególnie porównawczej) w rozsądnym czasie, dlatego też obecnie poszukuje się rozwiązań przedstawionego problemu, a jednym z nich może być zastosowanie, proponowanej przez autorów niniejszej pracy, metody wydobywania wiedzy, która składa się z następujących etapów:

1) selekcja cech na podstawie oceny eksperckiej,
2) identyfikacja rozkładu przyjmowanych wartości dla analizowanych cech w celu rozpoznania atrybutów dychotomicznych, wartości brakujących, zduplikowanych lub odstających; dyskretyzacja wartości atrybutów ilościowych przy konsultacji z ekspertem,
3) zastosowanie gęstościowego algorytmu analizy skupień w celu wygenerowania reprezentantów utworzonych grup,

4) zastosowanie hierarchicznego algorytmu grupowania, jednakże ograniczonego wyłącznie do reprezentantów skupień utworzonych w poprzednim kroku,
5) wizualizacja wyników analizy skupień za pomocą techniki map prostokątów.

Pierwszym etapem prac podczas analizy dużych, rzeczywistych zbiorów jest selekcja cech istotnych z punktu widzenia rozważanego problemu, dokonywana przez eksperta. Ze względu na złożoną strukturę takich zbiorów oraz konieczność posiadania często rozległej wiedzy dziedzinowej udział eksperta w tym procesie jest niezbędny.

Następnie rozpoczyna się etap przygotowania danych do analizy uwzględniający m.in. wykrycie wartości pustych, zduplikowanych, odstających czy ogólnego rozkładu dla danej cechy. W przeciwieństwie jednak do schematu klasycznego procesu odkrywania wiedzy wykorzystywanego w literaturze [6] etap ten ma pozwalać również na wykrycie wiedzy czy zależności z danych, przy wykorzystaniu metod opisu danych (zarówno opartych na statystyce opisowej, jak i technikach graficznych np. w postaci histogramów) [1]. Przykładowo w kontekście zbioru urządzeń sieciowych można powiązać liczbę błędów transmisji z konkretnymi urządzeniami, a następnie, dokonując wizualizacji na histogramie, sprawdzić, czy któreś z nich nie odstaje od reszty (pod względem tego parametru).

Kolejnym krokiem jest zastosowanie algorytmu gęstościowego do wygenerowania struktury skupień i ich reprezentantów. Autorzy po analizie wielu możliwych algorytmów analizy skupień (co opublikowano m.in. w [8, 11]) jako optymalną wybrali technikę DBSCAN[3]. Charakteryzuje się ona bowiem możliwością odkrywania skupień o różnej strukturze, odpornością na występowanie wartości izolowanych czy relatywnie małą złożonością obliczeniową i zajętością pamięci [2]. Pierwszy krok algorytmu polega na wylosowaniu obiektu *p* oraz wyznaczeniu wszystkich obiektów, które są gęstościowo osiągalne z obiektu *p* (przy zadanych wartościach *Eps* – maksymalnego promienia sąsiedztwa i *MinPts* – minimalnej liczby obiektów wchodzących w skład grupy). Jeżeli *p* jest obiektem centralnym, to krok ten skutkuje powstaniem pierwszej grupy. Obiekt jest określany jako centralny, jeżeli co najmniej *MinPts* obiektów znajduje się w jego bezpośrednim sąsiedztwie (o promieniu *Eps*). Jeżeli *p* jest obiektem granicznym[4], to żaden obiekt nie jest gęstościowo osiągalny z *p*, więc algorytm wybiera kolejny obiekt ze zbioru danych [2]. Proces ten jest powtarzany, aż nie zostaną sprawdzone wszystkie obiekty ze zbioru danych. Obiekty niezaklasyfikowane do żadnego skupienia są oznaczane jako szum informacyjny.

Równie istotnym czynnikiem jak wybór właściwego algorytmu analizy skupień jest zdefiniowanie zrozumiałych i łatwych w dalszej interpretacji reprezentantów skupień.

[3] Jako miarę podobieństwa obiektów wybrano odległość Hamminga [6], pozwalającą operować na danych zarówno ilościowych, jak i jakościowych. Ponadto miara ta nie wprowadza znaczącego narzutu obliczeniowego.

[4] Obiekt nazywany jest granicznym, gdy w swoim sąsiedztwie ma mniej niż *MinPts* obiektów, ale jednocześnie znajduje się w sąsiedztwie obiektu centralnego.

W pracy [7] autorzy zaproponowali cztery koncepcje tworzenia reprezentantów grup oraz potwierdzili eksperymentalnie, że podejście wykorzystujące operator logiki boolowskiej AND jest najbardziej obiecujące (pod względem analizy i przeszukiwania utworzonej struktury). Główną zaletą zdefiniowania reprezentanta jako części wspólnej deskryptorów opisujących obiekty należące do jednej grupy jest fakt, że można w łatwy sposób dostrzec, jakie cechy są wspólne dla wszystkich obiektów danej grupy. Ponadto taka koncepcja umożliwia szybkie porównanie skupień ze sobą (pod kątem cech je odróżniających) ze względu na zwięzły opis takiego reprezentanta. Niestety, podczas analizy danych złożonych może zostać wygenerowanych wiele skupień, zatem podejście wydobywania wiedzy zakłada wykorzystanie drugiego algorytmu grupowania.

Powstałych reprezentantów skupień należy rozpatrywać jako uogólnienie wiedzy ukrytej w danych (wyrażonej w postaci związków i korelacji między obiektami). Przy wykorzystaniu reprezentantów jako nowych danych wejściowych możliwe jest zastosowanie innej techniki eksploracji danych – w tym przypadku proponuje się skorzystanie z hierarchicznego algorytmu grupowania (AHC [4]) – co nie było możliwe wcześniej ze względu na zbyt duży rozmiar zbioru danych[5]. Algorytm na początku przyporządkuje każdy obiekt do osobnego skupienia, a w każdej kolejnej iteracji łączone są ze sobą grupy charakteryzujące się największym współczynnikiem podobieństwa. Taka (ograniczona do danego poziomu) struktura jest przedstawiana użytkownikowi w formie czytelnej i prostej do interpretacji dzięki wykorzystaniu techniki map prostokątów (opisanej w pracach [10, 3, 5]).

Wizualizacja dokonuje rekurencyjnego podziału całej dostępnej przestrzeni roboczej na szereg prostokątów, układanych poziomo bądź pionowo względem siebie w zależności od reprezentowanego parametru ilościowego i dostępnego miejsca na ekranie. Dany prostokąt na wizualizacji symbolizuje konkretną grupę, a pole jego powierzchni jest tożsame z liczbą obiektów wchodzących w skład reprezentowanego skupienia. Dodatkowo z każdym prostokątem powiązane są statystyki opisowe, takie jak minimalne, maksymalne, średnie i najczęściej występujące wartości dla atrybutów obiektów wchodzących w skład określonej grupy. Należy jednak zaznaczyć, że pole powierzchni danego prostokąta może symbolizować dowolny parametr ilościowy istotny z punktu widzenia zadania eksploracji danych. Może to być średnia wartość np. zarejestrowanej liczby zdarzeń w ciągu dnia dla wszystkich obiektów skupienia (przez co możliwa jest identyfikacja najbardziej wadliwych urządzeń). Prostokątom można przydzielać kolory – przykładowo jasny kolor określa wysoką średnią wartość analizowanej cechy (wśród obiektów danego skupienia), a ciemny sytuację odwrotną. Jeżeli (dla analizowanego zbioru urządzeń nadawczo-odbiorczych) kolor symbolizuje stopień niedostępności konkretnego urządzenia, to bardzo jasne prostokąty

[5] Podejścia wykorzystywane przy analizie dużych zbiorów danych zostały przez autorów opisane w [8].

wskazują na urządzenia często niedostępne. Proponowana metoda wydobywania wiedzy mimo prostej koncepcji wydaje się dość skuteczna podczas analizy zbiorów danych złożonych (co zbadano w eksperymentach).

4. Eksperymenty obliczeniowe

Przeprowadzone w ramach niniejszej pracy eksperymenty mają na celu pokazanie przebiegu proponowanej metody odkrywania wiedzy, jak również potwierdzenie słuszności zastosowania w tym celu technik analizy skupień.

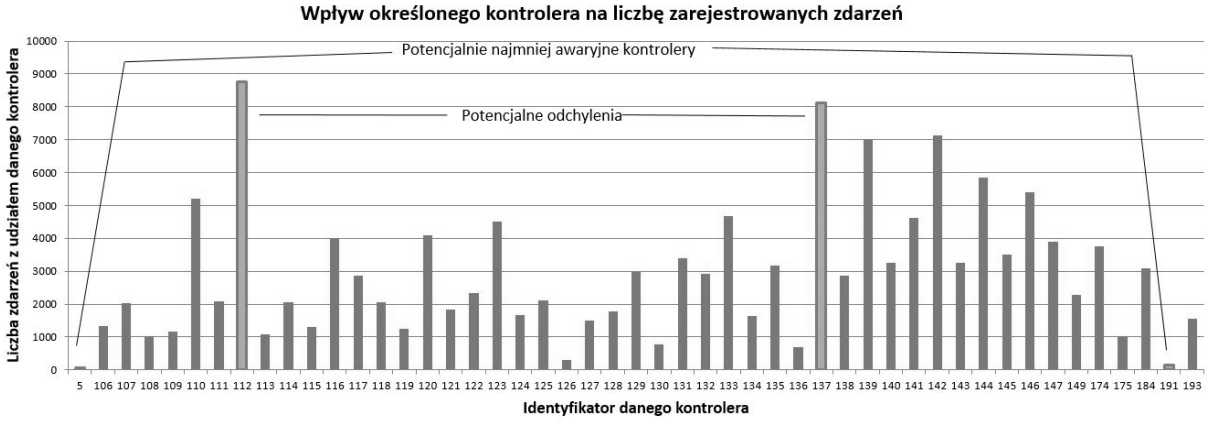
4.1. Eksperyment 1: analiza przydatności wykorzystania histogramów jako metody opisu danych w procesie odkrywania wiedzy

Pierwszym krokiem proponowanej przez autorów niniejszej pracy metody wydobywania wiedzy jest zapoznanie się ze strukturą danych. W tym celu autorzy wygenerowali szereg wykresów częstości występowania wartości unikalnych dla wszystkich atrybutów wchodzących w skład opisów obiektów analizowanego zbioru danych rzeczywistych[6]. Pozwoli to na identyfikację urządzeń nadawczo-odbiorczych odstających od innych pod względem określonych parametrów. Przykładowo wykres zilustrowany na rys. 1 pozwala ocenić wpływ zastosowania określonego kontrolera na liczbę zarejestrowanych zdarzeń[7].

Analizując wspomniany wykres, można dojść do wniosku, że kontrolery o numerach 112 oraz 137 stanowią odchylenia, ponieważ jako jedyne biorą udział w ponad ośmiu tysiącach zdarzeń. Dodatkowo można by stwierdzić, że najmniej problemów związanych z dostępnością usług sieciowych sprawiają kontrolery o numerach 5 oraz 191, ponieważ odpowiednio 53 razy oraz 140 razy zostały zarejestrowane zdarzenia z ich udziałem, co dla ponad 140 tysięcy wpisów w badanym zbiorze jest pomijalną wartością. Niestety tego typu analiza jest błędna. Wynika to z faktu, że pod kontrolą danego kontrolera pracuje zwykle od kilku do kilkudziesięciu komórek. Oczywiste jest, że im więcej urządzeń nadawczo-odbiorczych przypisanych jest do określonego kontrolera, tym większe jest prawdopodobieństwo, że przynajmniej jedno z nich może być niedostępne, przez co identyfikator takiego kontrolera częściej będzie się pojawiał w statystykach. Dlatego też w kolejnym kroku stworzono wykres, na którym uwzględniono również liczbę sterowanych komórek. Potwierdziło to zależność, że choć zidentyfikowane wcześniej kontrolery o numerach 112 oraz 137 biorą udział w największej liczbie zarejestrowanych zdarzeń, to jednak odpowiadają one również za sterowanie znacznie

[6] W niniejszej pracy zamieszczono jedynie najistotniejsze z punktu widzenia prowadzonych badań wykresy i wnioski.
[7] Zarejestrowana liczba zdarzeń jest porównywalna z liczbą obiektów zapisanych w bazie.

większą liczbą komórek niż kontrolery 5 i 191, pojawiające się najrzadziej w zbiorze danych. Nie sposób zatem porównać pracę kontrolerów, analizując wyłącznie częstość ich występowania w zbiorze danych, dlatego też kolejny wykonany eksperyment polegał na stworzeniu zapytania SQL umożliwiającego prawidłową identyfikację przyczyny zarejestrowanych zdarzeń.



Rys. 1. Wpływ określonego kontrolera na liczbę zarejestrowanych zdarzeń
Fig. 1. Correlation between a particular controller and the number of registered events

4.2. Eksperyment 2: analiza możliwości zastosowania klauzul grupujących i funkcji agregujących języka SQL w celu znalezienia korelacji między obiektami analizowanego zbioru danych

Celem drugiego eksperymentu było zidentyfikowanie tych urządzeń nadawczo-odbiorczych (wraz z ich właściwościami), które w tym samym czasie miały dokładnie taką samą stratę (poziom dostępności). Pozwoli to na powiązanie urządzeń przypisanych przez system monitorujący do innego zdarzenia (o innym identyfikatorze), jednakże w rzeczywistości będących częścią tej samej awarii. Przykładowo, jeżeli grupa komórek (pozornie niezwiązanych ze sobą) charakteryzuje się w tym samym punkcie czasowym dokładnie takim samym stopniem dostępności, może to sugerować problem z kontrolerem, który nimi steruje, lub z łączem transmisyjnym. W tym celu posłużono się zapytaniem SQL, którego zadaniem miał być wybór tylko tych grup urządzeń, w ramach których występują dwa (lub więcej) kontrolery – ponieważ jest bardzo mało prawdopodobne, by dwa różne kontrolery były wadliwe. Kod opisywanego zapytania jest następujący:

```
select liczebnosc, cellname, obszarId, dostawcaId, kontrolerId, zdarzenieId,
round(strata, 7) as strata
from (
  select
  count(*) over (partition by start, koniec, strata) as liczebnosc,
  min(kontrolerId) over (partition by start, koniec, strata) as kontrolerID_min,
  max(kontrolerId) over (partition by start, koniec, strata) as kontrolerID_max,
  t.*
  from cell_loss t
  where t.[data] = '07-08-2010') t1
where t1.liczebnosc >= 15 and t1.kontrolerID_min <> kontrolerID_max
```

Przedstawione zapytanie SQL najpierw zlicza, ile komórek miało dokładnie taką samą dostępność w tym samym oknie czasowym 7 sierpnia 2010 roku. Następnie wyznacza tylko te grupy komórek, w których zawarte są co najmniej dwa kontrolery.

Warunek *liczebność >= 15* pozwala sterować wielkością odpowiedzi – przykładowo dla operatora sieci znaczące mogą być tylko zdarzenia, w których brała udział większa liczba komórek.

Tabela 1

Rezultat opisywanego zapytania SQL

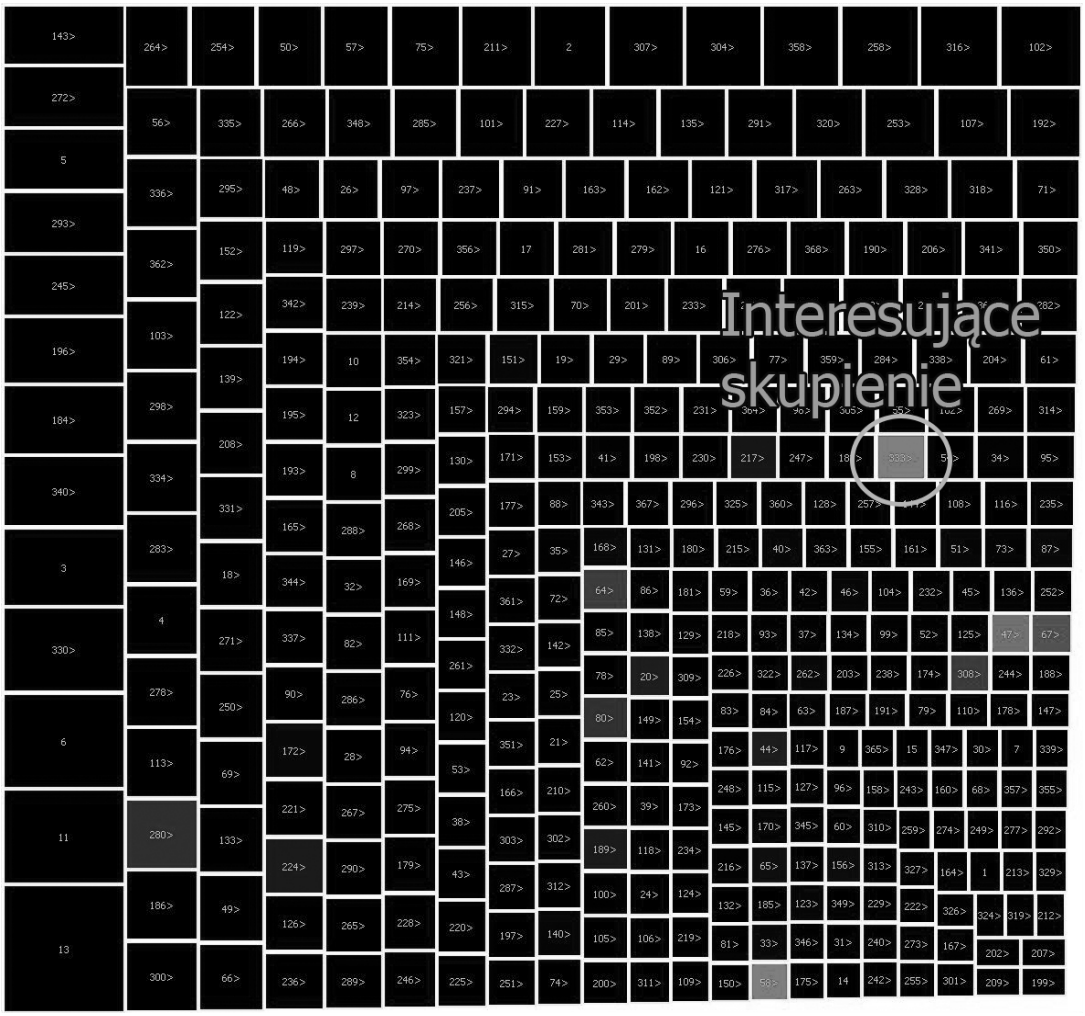
liczebność	cellname	obszarId	dostawcaId	kontrolerId	zdarzenieId	strata
15	50990B1	35	3	110	812947	0,0277778
15	55164A1	2	3	138	234199	0,0277778
15	55164A2	2	3	138	921761	0,0277778
15	55164B1	2	3	138	369401	0,0277778
15	50264B2	35	3	110	15825	0,0277778
15	50264B1	35	3	110	657815	0,0277778
15	50990A1	35	3	110	35467	0,0277778
15	50990A3	35	3	110	49839	0,0277778
15	55164A3	2	3	138	658199	0,0277778
15	50990B3	35	3	110	438378	0,0277778
15	50264B3	35	3	110	144339	0,0277778
15	55164B3	2	3	138	857316	0,0277778
15	50990B2	35	3	110	586977	0,0277778
15	55164B2	2	3	138	159097	0,0277778
15	50990A2	35	3	110	745688	0,0277778

Opisywane w tej sekcji zapytanie SQL zwraca w odpowiedzi jedno skupienie składające się z 15 obiektów (rekordów), co zostało zilustrowane w tabeli 1. Po dokładniejszej analizie zawartości wygenerowanej grupy można stwierdzić, że uwzględnia ona urządzenia tego samego producenta, sterowane przez kontrolery o identyfikatorach 110 i 138 oraz mające niemalże taki sam poziom (około 3%) niedostępności. Przyczyną wykrytego zdarzenia jest najprawdopodobniej tymczasowy brak zasilania (w godzinie pomiarowej), co zostało również potwierdzone przez eksperta dziedzinowego.

W wyniku opisywanego eksperymentu zidentyfikowano wiele urządzeń nadawczo-odbiorczych, które są częścią tego samego zdarzenia (utraty dostępności), mimo że system monitorujący przypisał je do zupełnie innych zdarzeń. Jednakże stworzenie zapytania SQL pomocnego w tym zadaniu było możliwe dopiero po uprzednim zapoznaniu się ze strukturą danych przez stworzenie wykresów analizy częstości, które zasugerowały istnienie interesujących korelacji między zdarzeniami – choć pierwotna teoria o wpływie kontrolera była błędna.

4.3. Eksperyment 3: analiza przydatności wizualizacji struktury skupień (powstałej przez zastosowanie kombinacji algorytmów gęstościowego i hierarchicznego) za pomocą techniki map prostokątów w procesie wydobywania wiedzy

Ostatni przeprowadzony eksperyment dotyczył dalszych etapów proponowanej metody odkrywania wiedzy, tj. zastosowania algorytmów analizy skupień i wizualizacji takiej struktury. Dokonano zatem grupowania algorytmem DBSCAN, w wyniku czego otrzymano strukturę niemalże ośmiu tysięcy grup[8], liczących od 1 do 6279 obiektów.



Rys. 2. Mapa prostokątów przestawiająca strukturę skupień komórek
Fig. 2. A rectangular treemap presenting the structure of cell clusters

Jak umotywowano wcześniej, jest to zbyt duża liczba do analizy w rozsądnym czasie, nawet mimo wykorzystywanej intuicyjnej techniki generowania opisów reprezentantów grup. Dlatego też następny krok uwzględniał zastosowanie aglomeracyjnego algorytmu AHC,

[8] Optymalne wartości parametrów startowych dla algorytmu DBSCAN wybrano, korzystając z heurystyki zaproponowanej przez autorów w pracy [7].

by otrzymać dwupoziomową hierarchię i znacznie zredukować liczbę powstałych skupień[9]. Jako kryterium łączenia skupień wykorzystano metodę średniego wiązania. W rezultacie uzyskano 368 skupień. Wygenerowane grupy zostały następnie poddane wizualizacji za pomocą techniki map prostokątów, co przedstawiono na rys. 2.

Biorąc pod uwagę strukturę zbioru komórek, w wizualizacji przedstawiającej rozkład grup skupiono się na parametrach częstości występowania zdarzeń (w ciągu danego dnia pomiarowego) oraz niedostępności urządzenia. Dlatego też w wizualizacji (przedstawionej na rys. 2) rozmiar prostokąta symbolizuje średnią częstość zarejestrowanych zdarzeń, natomiast kolor określa poziom straty (w ramach skupień). Im prostokąt miał większe pole, tym częściej notowane były zdarzenia (np. niedostępność urządzenia wynikająca z braku zasilania) związane z pracą urządzeń w danej grupie, natomiast im jaśniejszy kolor danego prostokąta, tym większą stratę (niedostępność) zarejestrowano. Priorytetem dla operatora sieci jest oczywiście dostępność usług, zatem to ten drugi parametr jest w tej analizie ważniejszy. Na tej podstawie udało się zlokalizować skupienie *331* (wyróżnione na rys. 2 określeniem „interesujące skupienie").

Skupienie *331* agreguje jedenaście rekordów ze zbioru danych, opisujących pracę czterech urządzeń o identyfikatorach *55171B2*, *58331A2*, *58331A3*, *58331A1*. Wszystkie wymienione komórki pochodzą od tego samego dostawcy i są sterowane przez kontroler 140. Trzy spośród wymienionych urządzeń (*58331A3*, *58331A2*, *58331A1*) pracują w pobliskich sektorach. Żadne z wyselekcjonowanych zdarzeń nie było planowane i nie było zleceń naprawczych. Pod względem czasu trwania zdarzenia niechlubnie wyróżnia się urządzenie *55171B2*, ponieważ było niedostępne aż 119 godzin. Dodatkowo jednokrotnie miało ono stratę poniżej 100%, co wykluczałoby celowe wyłączenie tego urządzenia (jak mogłaby sugerować nieustanna niedostępność). Udało się zatem zidentyfikować co najmniej jedno problematyczne urządzenie (potencjalnie cztery), którym należałoby się zająć w pierwszej kolejności i sprawdzić jego działanie na miejscu. Co istotne, urządzenie *55171B2* nie zostało wykryte innymi metodami, mimo przeprowadzenia przez autorów również innych zabiegów w celu jego identyfikacji.

5. Podsumowanie

Celem niniejszego artykułu było przedstawienie uniwersalnej metody reprezentacji i wydobywania wiedzy ze złożonych zbiorów danych rzeczywistych o dużej liczebności,

[9] Hierarchia została zbudowana na podstawie progu podobieństwa równego pięć. Oznacza to, że jeżeli na danym poziomie dwa węzły hierarchii (obiekty) różniły się więcej niż pięcioma cechami, struktura była przycinana do tego poziomu. Wartość pięć została dobrana arbitralnie.

uwzględniającej wykorzystanie statystyki opisowej, algorytmów analizy skupień oraz efektywnych środków wizualizacji wydobytej wiedzy. Proponowana metoda zakłada zastosowanie dwuetapowej techniki generowania skupień (łączącej hierarchiczne i gęstościowe algorytmy grupowania), jak również wizualizację wyników w postaci map prostokątów. Wyniki przeprowadzonych eksperymentów sugerują, że w przypadku analizy dużych zbiorów danych złożonych dopiero przy zastosowaniu kilku metod eksploracji danych wspomaganych przez czytelną wizualizację wyników możliwe jest odkrycie nowej, poprawnej i użytecznej wiedzy.

Niniejsza praca jest częścią projektu „Eksploracja regułowych baz wiedzy" sfinansowanego ze środków Narodowego Centrum Nauki (*NCN: 2011/03/D/ST6/03027*).

BIBLIOGRAFIA

1. Dasu T., Johnson T.: Exploratory Data Mining and Data Cleaning. John Wiley & Sons, Inc., USA 2003.
2. Ester M., Kriegel H. P., Sander J., Xu X.: A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining, USA 1996.
3. Bruls M., Huizing K., van Wijk J.: Squarified Treemaps. Proceedings of the Joint Eurographics and IEEE TCVG Symposium on Visualization, Springer, Vienna 2000.
4. Berry M. W., Browne M.: Lecture notes in Data Mining. World Scientific Publishing Co. Pte. Ltd., Singapur 2009.
5. Shneiderman B., Wattenberg M.: Ordered Treemap Layouts. Proceedings of the IEEE Symposium on Information Visualization 2001, INFOVIS '01, IEEE Computer Society, USA 2001.
6. Han J., Kamber M., Pei J.: Data Mining. Concepts and Techniques. Elsevier Inc, USA 2012.
7. Wakulicz-Deja A., Nowak-Brzezińska A., Xięski T.: Efficiency of complex data clustering. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg 2011.
8. Wakulicz-Deja A., Nowak-Brzezińska A., Xięski T.: Density-based method for clustering and visualization of complex data. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg 2012.
9. Simoff S. J., Böhlen M. H., Mazeika A.: Visual Data Mining. Theory, Techniques and Tools for Visual Analytics. Springer, Berlin, Heidelberg 2008.

10. Xięski T., Nowak-Brzezińska A.: Gęstościowa metoda grupowania i wizualizacji danych złożonych. Studia Informatica, ZN Pol. Śl., Vol. 33, No. 2A (105), Gliwice 2012, s. 453÷464.
11. Nowak-Brzezińska A., Jach T., Xięski T.: Wybór algorytmu grupowania a efektywność wyszukiwania dokumentów. Studia Informatica, ZN Pol. Śl., Vol. 31, No. 2A (89), Gliwice 2010, s. 147÷162.
12. Berka P., Rauch J., Zighed D. A.: Data Mining and Medical Knowledge Management: Cases and Applications. Information Science Reference. IGI Publishing, USA 2009.
13. Xięski T.: Metody reprezentacji danych złożonych. [w:] Wakulicz-Deja A. (red.): Systemy wspomagania decyzji. Wydawnictwo Uniwersytetu Śląskiego, Katowice 2011.

Wpłynęło do Redakcji 6 lutego 2014 r.

Abstract

In this paper the topic of discovering knowledge from complex data is discussed. Authors propose a universal knowledge discovery approach, which takes into account the usage of data description techniques, cluster analysis algorithms and effective means of visualization. A characteristic feature of this method is the usage of a two-stage clustering process (which combines the results of an agglomerative and a density based algorithm). This work also refers directly to the visual data mining process [9], by taking advantage of a rectangular treemap visualization technique and human cognitive abilities in the result evaluation stage.

Adresy

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