



DATA MINING JAKO SZANSA NA RELATYWNIE TANIE DOKONYWANIE ODKRYĆ NAUKOWYCH POPRZEZ PRZEKOPYWANIE POZORNIE CAŁKOWICIE WYEKSPLOATOWANYCH DANYCH EMPIRYCZNYCH

Ryszard Tadeusiewicz, Akademia Górniczo-Hutnicza, Laboratorium Biocybernetyki

Wprowadzenie

Większość dyscyplin naukowych wchodzących w zakres nauk przyrodniczych (*science*) opiera się na badaniach eksperymentalnych. Teoria może oczywiście stymulować rozwój nowych kierunków dociekań i badań (oraz dostarczać ich poprawnej interpretacji), a także może ukierunkowywać eksperymenty tak, by dawały maksymalnie dużo wartościowych danych – jednak decydujące znaczenie ma zawsze **eksperyment**. We wszystkich przypadkach wątpliwych to właśnie weryfikacja eksperymentalna jest rozstrzygająca i tylko empirycznie potwierdzone fakty pozwalają na uzyskanie nowych i pewnych wiadomości, realnie poszerzających zrab posiadanej wiedzy o nowe informacje i nowe interpretacje. Dzięki takiemu podejściu, stosowanemu rygorystycznie od czasów Wielkich Przyrodników (Leonardo da Vinci, Newton, Galileusz), nasza wiedza o otaczającym świecie, a także o naturze życia i o rządzących nim prawach stale się wzbogaca. Ponieważ jest to wiedza empirycznie sprawdzona, a nie spekulatywna – z wiedzy tej coraz częściej możemy czerpać praktyczne korzyści.

Badania empiryczne są jednak – jak powszechnie wiadomo – bardzo kosztowne. Ich prowadzenie wiąże się z dużym nakładem środków, a także wymaga zawsze dużego wysiłku, zaś niekiedy wiążą się dodatkowo z koniecznością dokonywania trudnych wyborów natury etycznej (na przykład doświadczenia prowadzone na zwierzętach), przeto wysoce racjonalne jest poszukiwanie metod, które pozwolą na maksymalnie efektywne wykorzystanie **wszystkich** wyników tych badań.

Na pozór można by sądzić, że przytoczona wyżej uwaga jest zbędna – każdy badacz dba przecież o to, by odpowiednio spożytkować wyniki swych wysiłków: pisze publikacje, poszukuje partnerów dla wdrożeń, opiera się na wcześniejszych badaniach w kolejnych pracach. Nie jest to jednak wcale pełne wykorzystanie wyników badań, gdyż w istocie **każdy eksperyment przynosi więcej odpowiedzi, niż zawierało postawione przez badacza pytanie**.

Upraszczając nieco rzeczywiste sytuacje występujące w praktyce prowadzenia badań naukowych można przyjąć, że każdy badacz w momencie podejmowania określonego



eksperymentu ma zawsze najpierw pewną hipotezę naukową, którą doświadczenia laboratoryjne potwierdzają – albo nie. Eksperyment jest więc zawsze wstępnie „wycelowany”, co więcej, można przyjąć pogląd, że im lepszy badacz, tym jego eksperyment lepiej zogniskowany, a jego wynik precyzyjniej nawiązuje do postawionej hipotezy. Jeśli zatem dane empiryczne są tak zbierane, że pozwalają łatwo rozstrzygnąć prawdziwości lub fałszywości pewnej hipotezy, to wydobyć z nich innych użytecznych informacji może być bardzo trudne. Jednak nikt nie jest eksperymentatorem doskonałym i dlatego wynik przeprowadzonych badań wzbogaca zawsze wiedzę w znacznie szerszym zakresie, niż to badacz zaplanował, więc obok faktów wcześniej przewidywanych przez teorię rejestruje się (chcąc – nie chcąc) mnóstwo faktów dodatkowych. Te dodatkowe fakty, w szanującym się laboratorium także zebrane w precyzyjny sposób i w ściśle kontrolowanych warunkach, stanowią zasób, który nas tu interesuje. Zwykle nie są one analizowane ani interpretowane, bo badacza interesuje zwykle tylko jego własna teza i nic więcej, ale ślad tych spostrzeżeń w protokołach pozostaje. Niezależnie więc od intencji badacza i jego zainteresowań wyniki pomiarów, uzyskane obserwacje, zarejestrowane dane itd. – zawierają zawsze także odpowiedzi na pytania, których jawnie nie postawiono. Te dodatkowe fakty stanowią w typowych badaniach materiał odpadowy, którym nikt się nie interesuje, gdyż „koncentracja” nowych i niebanalnych spostrzeżeń w takich danych zarejestrowanych „przy okazji” lub przypadkowo – jest z oczywistych powodów bardzo niska.

Zasugerowany postawionym sobie celem badawczym eksperymentator wykorzystuje jedynie **część** prawdy, jaką odkryły i ujawniły jego eksperymenty. Każdy (bezwarunkowo **każdy!**) badacz kolekcjonuje i publikuje jedynie **część** wniosków, jakie można było z jego doświadczenia wyciągnąć – a resztę po prostu odrzuca, bo jest ona dla niego nieprzydatna. On miał swoją tezę, uzyskał dowód tej tezy i teraz skupia się nad tym, jak go wykorzystać. Tymczasem w tych nieprzeanalizowanych i niewykorzystanych rezultatach badań jest też sporo ciekawych i wartościowych **odpowiedzi**, niewykorzystanych wyłącznie z tego powodu, że nikomu nie przyszło na myśl, żeby postawić odpowiednie **pytania**.

W pojedynczym dobrze zaplanowanym i dobrze przeprowadzonym badaniu takich „odkryć ubocznych” może być niewiele, jednak nagromadzenie wyników wielu takich eksperymentów – odpowiednio przebadanych i przeanalizowanych – może przynieść nowe spostrzeżenia (lub nawet nowe odkrycia), tym cenniejsze, że uzyskiwane praktycznie bez żadnych kosztów, bo oparte wyłącznie na pogłębionej analizie nagromadzonych już danych [7, 8].

To stwierdzenie stanowi podstawową tezę prezentowanego referatu: badając i powtórnie analizując (bardziej wnikliwie, a jednocześnie wielokierunkowo) wyniki uzyskane w eksperymentach prowadzonych uprzednio w kierunku zupełnie innych rozstrzygnięć naukowych – można dotrzeć do zupełnie nowych i naukowo bardzo wartościowych konkluzji, których przesłanki są obecne w nagromadzonym materiale empirycznym, ale nie zostały jeszcze wykorzystane.



Przyczyny zjawiska i inspirująca analogia

Są dwa powody, przyczyniające się do tego, że takie „dziewicze” odkrycia mogą kryć się w materiale wcześniej pozornie „do cna” wyeksploatowanych raportów laboratoryjnych.

Pierwszy leży po stronie psychologii poznania. Jak wiadomo potrafimy dostrzec tylko to, na co jesteśmy w jakimś stopniu przygotowani. Prowadzący doświadczenia badacz ma zawsze pewną wizję tego, co spodziewa się znaleźć w wynikach eksperymentu. Jeśli wynik badań zawiera oczekiwany fenomen, to badacz skupia się na nim, analizuje go, interpretuje, zestawia z innymi wynikami (zarówno własnymi, jak i innych badaczy) – w wyniku czego jednak tak dalece zawęża swoje pole widzenia, że może wielokrotnie przejść obok danych lub sygnałów, które są zapowiedzią całkiem innego odkrycia, nie dostrzegając ich ani nawet nie domyślając się ich istnienia. Z kolei jeśli wynik eksperymentu nie zawiera oczekiwanych efektów, to badacz ulega frustracji – tym silniejszej, im bardziej zależało mu na sukcesie. Wynik „nieudanego” eksperymentu oglądany jest ze wstrętem, chciałoby się jak najszybciej o nim zapomnieć, dlatego jeśli w tym pozornym artefakcie znajduje się nawet ziarno doniosłego odkrycia, to zostanie ono także przeoczone. Trzeba geniuszu Fleminga, żeby w spleśniałym preparacie dostrzec zapowiedź penicyliny – większość badaczy takie nieudane doświadczenia wstydliwie sprząta i stara się o nich nie myśleć.

Drugi powód pomijania czy też zaniechania „odkryć wtórnych” wiąże się z faktem, że w dobrze przeprowadzonym doświadczeniu efekt podstawowy, to znaczy ten, który związany jest ze świadomie sformułowanym zagadnieniem badawczym jest wyrazisty i ewidentny. Natomiast wspomniane wyżej efekty uboczne, będące niejednokrotnie podstawą do dodatkowych odkryć, pojawiają się niewyraźnie, nieoczekiwanie i trudno je odróżnić od silnego tła przypadkowych „szumów”, które także występują w wynikach każdego eksperymentu.

Nasuwa się tu nieodparcie analogia z górnictwem, którą mimo jej trywialności warto przytoczyć, gdyż jest ona inspirująca. Pracę eksperymentatora można porównać do pracy górnika. Jeden z nich wydobywa wiedzę, drugi surowce, obaj jednak muszą zmagać się z różnymi przeszkodami, jakie zazdrośna Natura spiętrzyła na drodze do pożądaných skarbów. Praca górnika jak wiadomo na tym polega, że wydobywa on z trudem jakieś minerały za pomocą kosztownego procesu podziemnego ich urabiania. Następnie korzysta z różnych środków transportu, dzięki którym te minerały z wnętrza ziemi, jeszcze nieocenne i niewykorzystane, pojawiają się na jej powierzchni. Potem te wydobyte surowce podlegają dalszej obróbce polegającej – najprościej to ujmując – na wyselekcjonowaniu tego, co nam się przyda i odrzuceniu tego wszystkiego, co z punktu widzenia celu, dla którego wybudowano tę kopalnię, wydaje się bezwartościowe.

Dokładnie tak samo dzieje się w badaniach naukowych. Fakty odkryte w przemysłowych doświadczeniach albo w pomysłowych obserwacjach, wszystkie informacje i liczby najpierw trafiają do dzienników laboratoryjnych albo na dyski komputerów rejestrujących wyniki badań (to jest ten naukowy „urobek”), a dopiero potem badacz zajmie się ich interpretacją.



Wydobyty z kopalni urobek nie nadaje się z reguły do natychmiastowego wykorzystania, podobnie jak ograniczoną przydatność mają „gołe” wyniki laboratoryjne. Dlatego wyniki doświadczeń poddaje się analizie ukierunkowanej na określony cel, a wydobyte surowce poddaje się procesowi tak zwanego wzbogacania (na przykład flotacji), w wyniku którego wydziela z nich pewne a priori wybrane użyteczne produkty. Zarówno w kopalni jak i w laboratorium, a cały pozostały po tej „uszlachetniającej” obróbce materiał zostaje odrzucony. W kopalni tworzą się w wyniku tego tak zwane hałdy, na których spoczywają miliony ton skał, niegdyś pracowicie urobionych w kopalni, potem rozdrobnionych dla transportu, wydobytych na powierzchnię ziemi, obejrżanych – i porzuconych. W laboratoryjnych archiwach i na dyskach komputerów spoczywają miliony pozornie niepotrzebnych liczb. Są tam setki notatek, rysunków i rejestracji z różnych aparatów, a także klisz, zdjęć, wyników wykonanych obliczeń itp. – kiedyś uzyskanych z dużym nakładem pracy przy pomocy dobrej aparatury, przy użyciu cennych odczynników, a potem odrzuconych jako bezwartościowe.

Skarby na hałdach

Badania geologów i mineralogów wielokrotnie już dowiodły, że w górniczych hałdach mogą się mieścić liczne użyteczne minerały, niedostępne w sposób naturalny na powierzchni Ziemi. Zamiast szukać ich w dziewiczych górach albo w tropikalnej dżungli – mamy je w samym centrum przemysłowych miast. W dodatku te surowce często można stosunkowo niskim kosztem przerobić na wartościowe produkty, ponieważ materiał skalny na hałdach został już wcześniej rozdrobniony, wstępnie obrobiony, a często także przetworzony chemicznie. Zamiast więc budować nową kopalnię ukierunkowaną na wydobycie nowego produktu – lepiej skorzystać z zasobów tego produktu w zwałach skalnych, które już zostały wydobyte. Opłaca się to nawet wtedy, gdy zawartość cennego składnika jest na hałdzie stosunkowo mała, znacznie mniejsza w stosunku do tego, co by się dało wydobyć z głębin ziemi w nowej kopalni. No bo ten surowiec na hałdzie już jest, gotowy i dostępny, a te podziemne skarby trzeba by było dopiero wydobywać, najczęściej ogromnym kosztem.

Takie wtórne wykorzystywanie hałd nie jest w ogólnym przypadku ani proste ani oczywiste, bo najpierw trzeba wykryć obecność jakichś cennych substancji w odpadowym materiale, a potem opracować technologie ich pozyskiwania. Co więcej, technologie te są zwykle trudne i skomplikowane, gdyż pozyskanie użytecznego produktu z surowca, który był wydobywany z myślą o innym produkcie – jest z reguły trudniejsze, niż ekstrakcja produktu pierwotnego z materiału specjalnie selekcjonowanego pod kątem jego wytwarzania. „Trudniejsze” – nie oznacza jednak „niemożliwe”, trzeba tylko mieć odpowiednie narzędzia.

Podobnie jest z każdą próbą wydobywania nowych prawd naukowych ze starych, pozornie całkowicie wyeksploatowanych, wyników badań i eksperymentów. Mamy tam ogromne zasoby faktów, zgromadzonych dużym nakładem wysiłku, myśli i pracy, rejestrowanych w dokładnie kontrolowanych warunkach, tyle tylko, że informujących „nie na temat” – więc po zarejestrowaniu po prostu porzuconych. Wśród tych informacji większość jest bez wartości, bo nie każda wiadomość, nawet stwierdzona naukowo, w istocie wzbogaca



wiedzę Ludzkości. Jednak wśród informacyjnych śmieci mogą się kryć prawdziwe „perełki”. Co więcej, nie tylko mogą – one tam na pewno są! Trzeba je jednak wydobyć, oddzielić od informacyjnego szumu, wyodrębnić i dokładnie zweryfikować – bo rygor odpowiedniego poziomu znamienności statystycznej wyniku musi dotyczyć wszystkich faktów, które chcemy potraktować jako naukowe, także i tych pozyskanych jako „produkty uboczne”.

Poszukiwanie nowych odkryć naukowych w wyeksploatowanych i porzuconych wynikach doświadczeń jest trudne z kilku względów, o których będzie niżej mowa. Ręczne przeszukiwanie stosów starych raportów laboratoryjnych w sytuacji, kiedy właściwie sami dobrze nie wiemy, czego w nich szukamy – jest pozbawione szans. Wszelkie nasuwające się tu porównania (szukanie igły w stogu siana, szukanie czarnego kota w ciemnym pokoju, w którym w dodatku może wcale nie być kota itd.) odwołują się w istocie do zadań o wiele łatwiejszych, chociaż też potocznie traktowanych jako beznadziejne. Rozważane zadanie jest trudne, ale nie niemożliwe, pod warunkiem, że będziemy mieli do dyspozycji odpowiednie narzędzia informatyczne. W dalszej części artykułu skupimy uwagę na tych właśnie narzędziach związanych z pojemnym hasłem *Data Mining*, pokazując, że instrumentarium współczesnej informatyki ma tu wyjątkowo dużo ciekawych propozycji. Zanim jednak do tego przejdziemy warto będzie przez chwilę skupić się na skali i na naturze trudności, które trzeba będzie pokonać.

Dlaczego jest to takie trudne?

Jeśli nawet w wyeksploatowanych wynikach badań naukowych ukryte są ziarna nowych odkryć, to musimy sobie zdawać sprawę, że są one tam obecne w postaci bardzo trudnej do odkrycia. Pierwsza i podstawowa trudność polega na tym, że użyteczne spostrzeżenia są w tych pozornie jałowych danych zawarte w bardzo małym „stężeniu”. Dobrze zaplanowane doświadczenie jest ukierunkowane głównie na zebranie danych, które potwierdzają hipotezę badacza – lub jej zaprzeczają. Wszelkie poboczne fakty, nawet jeśli się je rejestruje z powodu naukowej rzetelności, zawierają w sobie głównie przypadkowe szumy, a jeśli występują w nich dodatkowe wartościowe informacje, to ich koncentracja jest minimalna (w przeciwnym bowiem razie zwróciłyby na siebie uwagę badacza, który prowadził eksperymenty). Co więcej, musimy się liczyć z faktem, że te wszystkie dodatkowe dane są obecne w „odpadowych danych” w postaci zawikłanej i silnie zamaskowanej ogromną liczbą zupełnie nieistotnych oraz bezużytecznych informacji.

Jeśli jednak nagromadzimy bardzo dużo takich „odpadowych” wyników i zaczniemy je przeszukiwać pod kątem odpowiedzi na niepostawione wcześniej pytania, to możemy uzyskać nowe informacje, wartościowe naukowo oraz użyteczne praktycznie – właściwie za darmo. No, może nie całkiem za darmo: za cenę ogromnej pracy, jaką będą musiały wykonać komputery analizujące wskazane dane. No ale zasoby wolnych mocy obliczeniowych są coraz większe, więc dzisiaj można z powodzeniem wykonać mnóstwo nawet najbardziej wyrafinowanych i czasochłonnych analiz, o których jeszcze kilka lat temu można było zaledwie marzyć, więc szanse są!



Druga trudność, z jaką się tu zetkniemy, wiąże się z faktem, że my właściwie nie wiemy, czego szukamy. Twórca wyników, z których korzystamy, prowadząc badania miał jakiś dobrze zdefiniowany cel – ale w kontekście tego celu wyniki zostały już całkowicie wykorzystane. Natomiast my szukamy w jego wynikach „czegoś nowego i wartościowego”, zupełnie nie wiedząc, co by też to miało być. Żeby zrealizować takie poszukiwania trzeba sformułować setki różnych hipotez, zweryfikować je w oparciu o dane – i trafić (przy odrobinie szczęścia...) na tę jedną, która się w świetle zgromadzonych danych potwierdzi. Analizy takiej nie można jednak prowadzić „ręcznie”, gdyż – jak już wspomniano – zawartość użytecznej, ale jeszcze niewyeksplorowanej wiedzy w wynikach zarejestrowanych badań jest stosunkowo niska. Potrzebne są do tego specjalne narzędzia informatyczne – i o tych właśnie narzędziach będzie mowa w dalszej części tego referatu.

Sztuczna inteligencja – narzędzie informatyczne, którego nie ma, a jednak może się przydać

W zakresie rozważanych w tym referacie problemów badawczych narzędziem zalecanym do wydobywania nowej wiedzy z pozornie całkowicie już wyeksplorowanych danych doświadczalnych jest *sztuczna inteligencja*. Tak nazywana jest najnowsza (by nie powiedzieć – awangardowa) dziedzina informatyki, która jednak budzi wiele kontrowersji. Są tacy badacze (zwłaszcza filozofowie...), którzy pryncypialnie kwestionują możliwość istnienia czegoś takiego, jak sztuczna inteligencja, twierdzą bowiem, że już w samej nazwie tej dziedziny zawarta jest antynomia: inteligencja jest cechą specyficzną ludzką, więc żadne sztuczne narzędzie nie może wykazywać inteligencji. Zajmując się tą problematyką od wielu lat zwykle ignoruję te spory, uznając, że wolę jednak sztuczną inteligencję od naturalnej głupoty, a na ataki „purystów” odpowiadam: Dobrze, zgódźmy się, że sztucznej inteligencji nie ma, ale nawet jeśli ona nie istnieje – to i tak może się przydać, jest to bowiem część informatyki charakteryzująca się między innymi tym, że potrafi dostarczać **odpowiedzi na niepostawione jawnie pytania**.

Warto rozwinąć tę myśl, bo ma ona duże znaczenie dla rozważań przedstawianych w tym referacie. Otóż przyzwyczailiśmy się już do tego, że komputery na każde wyrażenie i poprawnie sformułowane pytania potrafią bardzo sprawnie i szybko odpowiadać, bo do tego je właśnie zbudowano. Wyznaczenie wartości matematycznej obliczanej ze skomplikowanego wzoru albo wyszukanie potrzebnej informacji wśród milionów innych danych – to dla nich „pestka”. Jeśli jednak nie wiadomo z góry, co i jak obliczać, jeśli nie podano, co trzeba wyszukać – to komputer z typowym programem jest bezsilny.

Nie dotyczy to jednak komputerów wyposażonych w programy sztucznej inteligencji – one potrafią rozpoznać wrogie zamiary nigdy wcześniej niewidzianego programu wirusa, one potrafią odgadnąć reguły, za pomocą których można przewidywać przyszłe zdarzenia (na przykład gospodarcze), one wreszcie wśród setek tysięcy pozornie bezwartościowych danych potrafią wykryć współzależność niosącą nowe, nigdy wcześniej nieprzewidywane odkrycie.



Przytoczmy może mało znany (mam nadzieję – wszystkich znających wcześniej tę historię przepraszam za jej powtórzenie...) **autentyczny** fakt z tego zakresu, pokazujący, jak data mining działa w obszarach związanych z gospodarką [1]. Otóż wiadomo, że każdy właściciel sklepu chce jak najlepiej poznać zwyczaje i preferencje swoich klientów, gdyż dysponując taką wiedzą, można „wcisnąć” klientowi dodatkowe towary i skłonić go do pozostawienia w sklepie dodatkowej porcji gotówki. Dlatego na zapleczu wielkich supermarketów siedzą całe sztaby fachowców, planujących rozmieszczenie towarów na regałach i lokalizację koszy z przecenionymi produktami z taką samą precyzją i pieczołowitością, z jaką generałowie rozmieszczają żołnierzy i czołgi na polu walki. Dla tych ludzi bardzo cenna jest na przykład wiadomość, że klient który przyszedł do sklepu po piwo, sięgnie także zapewne po puszkę orzeszków lub paczkę krakersów, jeśli spotka ją na swojej drodze do kasy – bo to może oznaczać dodatkowe krociowe zyski.

Otóż specjaliści od marketingu od dawna spoglądali z nadzieją na komputerowe rejestry kasowe. W końcu w kasie sklepowej każdy z nas musi się „wyspowiadać” z tego, co lubi i czego nie lubi, po prostu przedstawiając zawartość swojego koszyka. Miliony takich zarejestrowanych transakcji zawierają w sobie bezcenną wiedzę o zachowaniach tysięcy klientów, tylko jak ją wykorzystać? Komputer sklepowy zapytany o konkretne informacje w mgnieniu oka odpowie, ilu klientów kupiło czekoladę, a ilu szynkę, a także ilu było takich, którzy kupili zarówno szynkę, jak i czekoladę – ale z tego nic nie wynika. Wobec tego zastosowano techniki sztucznej inteligencji, żądając, żeby komputery wykryły wszelkie powtarzalne zachowania klientów, które do tej pory nie były znane. Wynikiem było między innymi wykrycie sporej grupy klientów, którzy wykazywali nieprzewidziane wcześniej zachowanie – otóż jeśli kupowali pieluszki dla dzieci, to sięgali również po piwo, przy czym zdarzało się to głównie w sobotnie wieczory. Było to w pierwszej chwili zadziwiające, jednak dokładniejsze obserwacje potwierdziły, że typowy młody tata wypędzony w sobotni wieczór przez żonę po pieluszki dla dziecka, odreagowuje stres i odbudowuje swoją męską godność, kupując sobie dodatkowo piwo. Uwzględniono ten fakt poprzez odpowiednie rozmieszczenie „pokus” na przewidywanej drodze klientów – i obroty sklepu wzrosły o dalszych kilka procent, przynosząc wielotysięczne zyski.

Przytaczam tę historię (absolutnie autentyczną!) po to, żeby wykazać, iż metodami sztucznej inteligencji rozwiązano już sporo praktycznych problemów, w których trzeba było znaleźć coś istotnie nowego w ogromnym zbiorze danych, z których większość była całkowicie bezwartościowa. W dodatku ta poszukiwana prawidłowość (z góry niemożliwa do przewidzenia, a więc taka, o którą nie można było jawnie zapytać) mogła się manifestować jedynie w ułamku promila obserwowanych danych, a jednak wykryta i potwierdzona komputerowo była bardzo istotna i użyteczna. Czyż nie czas by było na to, by wspomniane metody komputerowe zaczęły nam służyć w szlachetnym dziele poznawania i odkrywania wiedzy, zamiast tylko napędzać zyski rekinom handlowym?



Dane, informacje i wiedza w badaniach naukowych

Postulując wykorzystanie metod sztucznej inteligencji w badaniach naukowych, jestem świadom tego, jak płytko i powierzchowna jest w gruncie rzeczy analogia procesów wydobycia dodatkowych bogactw z hałd górniczych, procesu wykrywania zachowań oraz preferencji klientów i postulowanego w tej pracy procesu odkrywania nowych prawd naukowych w wynikach starych doświadczeń. Jednak po chwili zastanowienia trudno nie przyznać, że taka analogia może zachodzić, a skoro tak, to powinna być wykorzystana, gdyż przy jej mądrym użyciu może dochodzić do wzbogacania wiedzy (między innymi w obszarach badań biofarmakologicznych) w istocie bez dodatkowych nakładów na nowe badania laboratoryjne.

Dział sztucznej inteligencji, jaki może być w tym obszarze szczególnie przydatny, znany jest w literaturze pod nazwą *data mining* [3] (jak widać nie tylko mnie nasuwa się w tej sprawie analogia z górnictwem!). Nazwa jest niezbyt trafna, dlatego że celem poszukiwań w omawianej metodzie nie są **dane**, tylko zawarta w nich **wiedza**, zatem powinno się raczej mówić *knowledge mining* (podobnie jak poszukiwanie złota nazywa się *gold mining*), ale – jak to często bywa – nazwa ta już się przyjęła i funkcjonuje w całym piśmiennictwie fachowym, więc puryści językowi są znowu na straconej pozycji. Mimo kontrowersyjnej nazwy dziedzina ta potwierdziła już w wielu obszarach swoją użyteczność, jest więc możliwe i celowe użycie jej także w związku z potrzebami postulowanych tu badań. Poznamy ją teraz nieco bliżej.

Punktem wyjścia do rozważań jest definicja *data mining* jako grupy metod eksploracyjnej analizy danych [2]. Następnie scharakteryzowano podstawowe typy problemów, które mogą być rozwiązywane za pomocą omawianej tu grupy metod. Przedstawiono również przykładowe problemy, których rozwiązanie może wymagać stosowania metod eksploracyjnych. W końcowej części referatu zaprezentowano w sposób systematyczny i uporządkowany wszystkie etapy procesu eksploracyjnej analizy danych, co może stanowić podstawę do tworzenia praktycznych aplikacji omawianej tu metody.

Metody sztucznej inteligencji bazują na przetwarzaniu informacji, warto więc przez chwilę przyjrzeć się temu pojęciu w sposób maksymalnie ogólny. Zakres znaczeniowy słowa informacja jest bardzo szeroki. Obejmuje on zarówno usłyszaną bądź przeczytaną wiadomość o bliskiej osobie, jak i publikowane w wielu źródłach sądy dotyczące postaw czy preferencji społecznych. Na informacji opierają się nowoczesne systemy produkcyjne, ona pozwala jednostce czy organizacji sprawniej działać i zdobywać przewagę nad konkurencją. Niezależnie od tego, czy informacja traktowana będzie jako *poufna i osobista wiadomość*, czy też jako *źródło przewagi konkurencyjnej* lub *zasób strategiczny* firmy – nie ulega wątpliwości, że informacja jest bardzo ważna. Waga przypisywana informacji powoduje, że powinniśmy o nią *zabiegać* i o nią się *troszczyć*, prawidłowo ją *przechowując* i *chroniąc* przed różnorodnymi zagrożeniami. Duże znaczenie ma właściwa *prezentacja* informacji, właściwe jej *przetwarzanie* czy też *przesyłanie*. Bardzo groźny może się okazać *brak* informacji lub *opóźnienie* w jej dostarczeniu. Ale nie mniej groźny może być *zalew*



nadmiarem informacji. Informacje powinny być *zgodne z rzeczywistością*, nie mogą być *dwuznaczne* czy też *sprzeczne*.

Z pojęciem *informacji* mocno związane jest pojęcie *danych*. Definicji danych także można by było przytoczyć przynajmniej kilkanaście, jednak na użytek tej pracy umówmy się traktować dane jako informacje zarejestrowane i przetworzone do pewnej ustalonej, symbolicznej postaci. W definicji tej ważny jest fakt, że dane są informacją zarejestrowaną, zatem można w razie potrzeby wracać do nich wielokrotnie, zastając je każdorazowo w tej samej postaci, a także fakt, że danym nadano pewną ustaloną, powtarzalną formę, co upoważnia nas do mówienia o ich strukturalizacji. Na dane składają się informacje reprezentowane przez symbole – cyfry, litery, dźwięki, symbole graficzne, przy czym zwykle symbole te nie są nagromadzone w sposób przypadkowy, tylko tworzą pewne struktury podlegające pewnym prawidłowościom. Danymi są więc liczby (ciągi cyfr), wyrazy (ciągi znaków), wykresy czy obrazy (ciągi punktów lub innych symboli graficznych) – ale nie są to nigdy **dowolne** ciągi, tylko ciągi podlegające pewnym ograniczeniom. Na przykład w liczbie może wystąpić tylko jednorazowo separator oddzielający jej część ułamkową, sekwencja znaków staje się wyrazem, jeśli odpowiada zapisowi w słowniku jakiegoś określonego języka, a nagromadzenie symboli graficznych jest rysunkiem tylko wtedy, gdy są one odpowiednio dobrane i odpowiednio rozmieszczone.

Czy dane stanowią informacje? To zależy przede wszystkim od człowieka będącego ich odbiorcą. Jeżeli odbiorca jest w stanie **zinterpretować** otrzymane dane, to należy je traktować jako informację. Jeżeli dane pozostają dla odbiorcy tylko niezrozumiałymi ciągami symboli, to traktowanie ich w charakterze informacji nie jest uzasadnione. Warto zwrócić uwagę na czynnik subiektywizmu (czy może raczej relatywizmu – w odniesieniu do pewnej konkretnej osoby), zawierający się w przytoczonym stwierdzeniu. Zawartość laboratoryjnej bazy danych lub formuła nowego związku chemicznego będzie traktowana przed jednych jako skarbnica drogocennych informacji, ale dla innych pozostanie ciągiem nic nieznaczących znaków. Dane opisują jakieś fragmenty rzeczywistości – na przykład wyniki pewnego doświadczenia, ale zwykle wymagają odpowiedniej prezentacji, przetworzenia czy agregacji, aby stały się czytelne, zrozumiałe i użyteczne. Dopiero uzyskane rezultaty odpowiedniego przetworzenia danych oraz ich właściwej prezentacji noszą cechy użytecznej informacji. Same **dane** są więc w istocie wyłącznie **surowcem** informacyjnym, ponieważ **informację** jako taką trzeba dopiero wypracować, wykorzystując dane, ale dodając do nich niezbędny składnik inteligencji (własnej albo sztucznej) powiązanej ze świadomością celów rozważanego procesu informacyjnego, pozwalającej na ich właściwą selekcję, agregację i prezentację.

Zbiór posiadanych i powiązanych ze sobą informacji tworzy *wiedzę*. Miejscem, w którym wiedza powstaje, jest oczywiście głównie umysł odbiorcy informacji. Nowa informacja poprzez synergię z wiedzą wcześniej zgromadzoną bywa często źródłem zupełnie nowej wiedzy, pozornie niewynikającej bezpośrednio z samych dostarczonych informacji, gdyż często drobny na pozór kwant informacji może być czynnikiem decydującym o całościowym zrozumieniu jakiegoś zjawiska lub procesu. Jeśli do takiego całościowego zrozumienia dochodzi w jakimś sformalizowanym systemie odniesienia – to możemy mówić o tworzeniu teorii naukowej lub o budowie modelu. Jednak także doświadczenia codzienne



każdego człowieka obfitują w przykłady sytuacji, w których przyrost wiedzy (wewnętrznej) bywa całkiem niewspółmierny do ilości i jakości pozyskiwanej informacji. Często trzeba pracowicie zgromadzić bardzo duży zasób pozornie mało przydatnych wiadomości, uzyskując przez długi czas stosunkowo niewielki przyrost realnej wiedzy, by potem nagle, po pozyskaniu kolejnej, na pozór mało istotnej informacji, doznać wspaniałego uczucia olśnienia, kiedy nagle wszystkie fakty stają się jasne, związki i relacje widoczne, a efekt końcowy, w postaci przyrostu wiedzy, skokowo rośnie w następstwie swoistej krystalizacji informacji dokonywanej w odpowiednio zasilonym wiadomościami mózgu. Nie zmienia to jednak w żaden sposób faktu, że źródłem wiedzy są zawsze pracowicie gromadzone informacje, a źródłem informacji są (interpretowalne przez odbiorcę oraz zwykle odpowiednio przetworzone) dane.

Zakres zastosowań metod data mining i klasyfikacja rozwiązywanych zadań

Dobór właściwej metody analizy danych uzależniony jest od charakteru rozpatrywanego problemu. Przed przystąpieniem do prezentacji dostępnych metod warto przedstawić krótką charakterystykę *typów rozpatrywanych problemów*, gdyż uporządkuje to wszystkie dalsze rozważania. Do podstawowych typów zadań rozwiązywanych z użyciem metod *data mining* zaliczymy:

- ◆ *Opis zależności*. Jest to najczęściej spotykane zadanie, dla którego rozwiązania stosujemy technikę *data mining*. Istota problemu polega w tym przypadku na tym, że mamy do dyspozycji dane opisujące fakty, a potrzebujemy informacji o tym, jakie są związki pomiędzy tymi faktami. Do tej klasy problemów zaliczać będziemy **dwie grupy** zagadnień:
 - Pierwsza z nich polegać będzie na podejmowaniu próby opisu *zależności istniejących pomiędzy wartościami zmiennych* (są to tzw. *problemy regresyjne*). Istnieje wiele metod badawczych przydatnych przy rozwiązywaniu tego typu zadań. Dokonując wyboru właściwego narzędzia, należy uwzględnić rodzaj zależności (liniowa, nieliniowa o znanym charakterze, zależność o nieznanym charakterze), problem skal pomiarowych użytych do wyrażenia wartości zmiennych, liczebności zbioru danych czy też dostępnego oprogramowania. Utworzone modele służą poznaniu analizowanych zjawisk, symulacji lub stanowią narzędzie prognozowaniu. Tworząc model opisujący zależności pomiędzy zmiennymi, badacz często wskazuje na jej kierunek, definiując, które zmienne mają charakter zmiennych objaśniających, a które zmiennych objaśnianych.
 - Drugą grupą zagadnień związanych z opisem zależności są *problemy asocjacyjne* – polegające na badaniu zależności pomiędzy faktami wystąpienia (bądź też braku wystąpienia) *pewnych zjawisk*, inaczej mówiąc, badaniu podlegać będzie *współwystępowanie zjawisk*. Oczywiście, problemy asocjacyjne stanowią szczególnie przypadek problemów regresyjnych, w których wartości analizowanych zmiennych mają charakter binarny. Jednakże z uwagi na duże znaczenie praktyczne i specyfikę



wykorzystywanych metod badawczych, warto niezależnie przyjrzeć się tej grupie problemów.

- ◆ *Klasyfikacja wzorcowa.* Przy tym zadaniu analizie poddawane są obiekty charakteryzowane przez wartości przyjętego zbioru zmiennych. Celem badań jest przypisanie poszczególnych obiektów (na podstawie wartości charakteryzujących je zmiennych) do wcześniej zdefiniowanych klas (właśnie z uwagi na **istnienie** wzorców klas ten rodzaj klasyfikacji określany jest mianem „wzorcowej”). Problemy klasyfikacji wzorcowej występują również w literaturze pod nazwą *problemów dyskryminacyjnych* bądź też problemów z zakresu *rozpoznawania obrazów* [5], przy czym słowo „obraz” w tym kontekście rozumiane jest właśnie jako synonim słowa „wzorec”, zatem metody rozpoznawania obrazów (ang. *pattern recognition*) mogą być stosowane do dowolnych danych, których rozpoznawanie i klasyfikacja może nas interesować.
- ◆ *Klasyfikacja bezwzorcowa.* W zadaniach omawianego tutaj typu w momencie przystępowania do badań nie są znane wzorce klas, na które należy rozbić dany zbiór danych, co więcej – zwykle brakuje nawet informacji o tym, jak duża jest przewidywana liczba klas [6]. A zatem w tych zadaniach celem analizy jest: rozpoznanie na podstawie samej tylko analizy danych struktury zbioru obiektów (występujących w postaci skupień danych, cechujących się pewnym poziomem wzajemnego podobieństwa i pewnym stopniem odrębności od danych należących do innych skupień), identyfikacja liczby i cech charakterystycznych występujących klas i przypisanie wszystkich lub przynajmniej znaczącej części obiektów do wyodrębnionych skupień. Często przy zadaniach klasyfikacji bezwzorcowej możliwe jest również badanie zależności występujących pomiędzy skupieniami i wnioskowanie na ich podstawie.
- ◆ *Analiza szeregów czasowych.* Ten rodzaj badań uwzględnia aspekt czasu [4]. Dane gromadzone w bazach i bankach danych zawsze związane są w jakiś sposób z czasem, gdyż określony fakt, którego rejestrację stanowi rozważana dana, miał miejsce w jakimś konkretnym momencie, który można bezpośrednio (np. na podstawie rejestracji chwili wprowadzenia danej do komputera) albo pośrednio (np. na podstawie opisu towarzyszącego rejestrowanym danym) ustalić i uwzględnić w bazie danych. Jednak nie zawsze czas jest istotnym faktorem przy analizie danych, gdyż wiele związków i relacji, których poszukuje się metodami *data mining*, z definicji powinno mieć charakter uniwersalny, niezależny od czasu. Są jednak takie problemy i takie zbiory danych, w których aspekt czasu ma zasadnicze znaczenie. Badacz próbuje wówczas zidentyfikować i opisać prawidłowości występujące pomiędzy wartościami zmiennych pochodzącymi z różnych okresów czasu, a techniki *data mining* mają go w tym wspomagać. W najprostszym przypadku analizy szeregów czasowych rozpatrywana jest pojedyncza zmienna i poszukiwane są zależności pomiędzy jej wartościami z danego okresu, a wartościami poprzedzającymi (mówimy wówczas o analizie szeregów jednowymiarowych). Ten schemat analizy może zostać rozszerzony na większą liczbę zmiennych – wówczas wartości jednej zmiennej uzależnione są od wcześniejszych wartości tej samej zmiennej, jak i od wcześniejszych wartości innych zmiennych. Cele stawiane metodom analizy danych uporządkowanych w czasie mogą być bardzo różne: czasami chcemy podać funkcję opisującą w precyzyjny sposób zależność pomiędzy



kolejnymi operacjami, a czasami przedmiotem naszych zainteresowań jest identyfikacja związku pomiędzy faktem aktualnego wystąpienia jakiegoś zjawiska a zaistnieniem pewnych zjawisk w dalszej bądź bliższej przeszłości. Takie zadanie, nazywane badaniem sekwencji zjawisk, ma szczególne znaczenie w zagadnieniach wspomagania procesów podejmowania decyzji (zwłaszcza gospodarczych), ale ma także bezspornie zastosowanie do zagadnień technologii materiałowych, w tym także biomateriałów. Podstawowym celem stosowania metod analizy szeregów czasowych jest możliwość wykorzystania skonstruowanych modeli do prognozowania dalszego przebiegu badanych zjawisk.

- ◆ *Problemy wyboru.* Z problemami tego typu spotykamy się wówczas, gdy spośród dostępnego zbioru elementów (np. obiektów albo zmiennych) musimy wybrać najlepsze (w sensie przyjętego sposobu oceniania). Algorytmy realizujące tego typu zadania w oparciu o techniki *data mining* są przydatne wówczas, gdy niemożliwe jest sprawdzenie i ocenienie **wszystkich** możliwych podzbiorów analizowanego zbioru elementów. Ma to miejsce w szczególności wtedy, gdy ich liczba jest na tyle duża, że czas potrzebny na wykonanie stosownych obliczeń niezbędnych do pełnego przeszukania zbioru wszystkich możliwości jest praktycznie nie do zaakceptowania dla badacza. Należy podkreślić, że techniki *data mining*, wykorzystywane w zadaniach sprowadzających się do problemów wyboru, prowadzą zwykle do znalezienia rozwiązań quasi-optimalnych, co oznacza, że rozwiązanie znalezione jest wystarczająco dobre, ale nie ma gwarancji, że jest możliwie najlepsze. Oznacza to, że techniki *data mining* należy stosować wyłącznie w takich problemach wyboru, w których z powodu stopnia złożoności użycie dokładnych metod optymalizacji jest niewykonalne, w przeciwnym przypadku najlepszym rozwiązaniem będzie jednak zaniechanie technik *data mining* i przeprowadzenie oceny wszystkich zestawów elementów. Problemy wyboru należą do znacznie szerszej klasy zagadnień z dziedziny optymalizacji, w których techniki *data mining* zaczynają odgrywać coraz większą rolę.

Dobór właściwej metody data mining do konkretnego zadania

Przystępując do badań, których elementem ma być użycie jednej z metod *data mining*, należy zawsze na wstępie dokonać identyfikacji problemu z wykorzystaniem podanych wyżej typów i kategorii. Po rozpoznaniu typu rozwiązywanego zagadnienia można dokonać wyboru właściwej metody analizy zgromadzonych danych. Przegląd przykładowych metod służących do rozwiązywania typowych problemów przedstawia tabela 1.

Po zidentyfikowaniu rodzaju rozwiązywanego problemu należy dokonać wyboru metod właściwych do jego rozwiązania. Wybór metody *data mining* stosowanej do analizy zgromadzonych danych powinien uwzględniać:

- ◆ aprioryczną wiedzę o badanym zjawisku (np. stopień znajomości ogólnych praw rządzących badanym zjawiskiem, liniowy bądź nieliniowy charakter zależności lub granic pomiędzy klasami, znajomość struktury szeregu czasowego);



- ◆ wielkość zbiorów danych (niektóre metody analizy wymagają dużej liczby zaobserwowanych przypadków, inne są preferowane przy małej liczbie danych);
- ◆ sposób wykorzystania wyników (w zależności od sytuacji mogą być wyżej oceniane metody modelujące sposób funkcjonowania badanego zjawiska albo metody mające budowę o charakterze *czarnej skrzynki*, albo metody dostarczające rezultatów w formie graficznej bądź metody dostarczające reguł decyzyjnych);
- ◆ dostępność oprogramowania.

Tabela 1. Rodzaje problemów i właściwe dla nich metody *Data Mining*.

Rodzaj problemu	Metody
Opis zależności	<ul style="list-style-type: none"> • statystyczne metody pomiaru zależności • sieci neuronowe typu MLP lub RBF • metody analizy współwystępowania • zbiory przybliżone
Klasyfikacja wzorcowa	<ul style="list-style-type: none"> • funkcje dyskryminacyjne • sieci neuronowe typu MLP • drzewa decyzyjne • systemy regułowe • zbiory przybliżone • metoda k-najbliższych sąsiadów
Klasyfikacja bezwzorcowa	<ul style="list-style-type: none"> • metody taksonomiczne • sieci neuronowe samouczące się • metody redukcji wymiaru przestrzeni danych • metody graficzne • algorytmy genetyczne
Analiza szeregów czasowych	<ul style="list-style-type: none"> • sieci neuronowe typu MLP lub RBF • metody analizy sygnałów • metody badania sekwencji
Problemy wyboru	<ul style="list-style-type: none"> • algorytmy genetyczne • sieci neuronowe typu Hopfielda • zbiory przybliżone

Jeśli to tylko możliwe, to należy zastosować więcej niż jedną metodę *data mining* do rozwiązania postawionego problemu. Za takim postępowaniem mogą przemawiać następujące przesłanki:

- ◆ Uzyskanie zbieżnych rozwiązań za pomocą różnych algorytmów *data mining* można traktować jako czynniki potwierdzające formułowane wnioski.



- ◆ Wyniki uzyskane przez różne metody mogą naświetlać różne aspekty badanych zjawisk i przez to mogą na wiele sposobów wzbogacić pozyskaną wiedzę.
- ◆ Stosowane metody analizy różnią się znacznie postacią uzyskiwanych wyników i sposobami ich interpretacji i wykorzystania. Mając wyniki uzyskane z wykorzystaniem kilku podejść, można – zależnie od okoliczności – wykorzystywać taką postać wyników, która w danym kontekście okaże się najwłaściwsza.

Korzystanie z metod eksploracyjnej analizy danych

Realizacja badań z wykorzystaniem metod typu *data mining* jest procesem kilkietapowym. Do zasadniczych jego elementów należy zaliczyć:

- ◆ Zdefiniowanie celu badań i określenie typu (typów) problemu badawczego.
- ◆ Utworzenie zbioru danych.
- ◆ Wstępna analiza i wstępne przetworzenie danych.
- ◆ Wykonanie właściwych obliczeń.
- ◆ Weryfikacja poprawności uzyskanych wyników.
- ◆ Interpretacja uzyskanych rezultatów i ich wykorzystanie w procesie decyzyjnym.

W wielu przypadkach kolejne etapy badań są realizowane wielokrotnie, gdyż uzyskane wyniki mogą wskazywać na potrzebę powtórzenia jednego lub kilku kroków poprzedzających.

Zdefiniowanie celu badań i określenie typu problemu badawczego

Przystępując do badań pewnego zasobu danych z użyciem technik *data mining*, należy precyzyjnie **zdefiniować ich cel**. Dlatego zastosowanie techniki *data mining* w tych obszarach musi poprzedzać próba takiego sformułowania problemu decyzyjnego związanego z prowadzoną działalnością badawczą, aby wyniki pozyskane z użyciem *data mining* mogły nas przybliżać do tego celu.

Przyjęty cel badań determinuje całe dalsze postępowanie, stanowi uzasadnienie dla ponoszonych kosztów, pozwala na późniejszą ocenę sukcesu lub niepowodzenia przeprowadzonych badań.

Mając dobrze zdefiniowany cel badań, należy określić **typ problemu badawczego** (lub też typy problemów badawczych, gdyż w wielu przypadkach rozwiązywany problem ma charakter złożony i wieloaspektowy). Pomoże to w wyborze najbardziej odpowiedniej techniki *data mining*, którą zaangażujemy do rozwiązania problemu.

Utworzenie zbioru danych

Sprecyzowanie celu badań pozwala na podjęcie decyzji dotyczącej **zbioru danych**, stanowiącego podstawę do przeprowadzenia dalszych prac. Na tym etapie badań



podejmowanych jest kilka istotnych decyzji. Pierwsza z nich dotyczy *źródła danych*. Najbardziej naturalnym źródłem może być odpowiednio gromadzona i pielęgnowana baza danych laboratoryjnych. Czasami wykorzystywane są także informacje pochodzące z arkuszy kalkulacyjnych lub z plików o „płaskiej” strukturze. Z uwagi na powszechną komputeryzację systemów informacyjnych przedsiębiorstw coraz liczniejsze firmowe źródła danych mogą stanowić podstawę do przeprowadzenia coraz szerszego zakresu analiz. Zwykle do projektowanych badań przydatna będzie tylko pewna *część* zasobów informacyjnych przedsiębiorstwa. Jej dokładne zdefiniowanie jest bardzo ważne, ale trudno jest podać w tym zakresie jakieś dokładniejsze wskazówki, bo kryterium wstępnej selekcji danych silnie uzależnione jest od przyjętego celu badań. Brak tych dokładniejszych wskazówek nie powinien być jednak traktowany jako zachęta do zaniechania czynności selekcji danych – przeciwnie, zadanie to staje się jeszcze ważniejsze na skutek tego, że nie może być powierzone bezmyślnej maszynie i bezwarunkowo wymaga inteligentnej interwencji samego badacza.

Należy również odpowiedzieć na jedno techniczne pytanie: czy w trakcie obliczeń będziemy korzystać bezpośrednio z danych oryginalnych, czy też będziemy działać na *kopii* interesujących nas danych. Utworzenie kopii danych jest w większości przypadków lepszym rozwiązaniem – eliminujemy bowiem w ten sposób niebezpieczeństwo przypadkowego uszkodzenia (podczas działania algorytmów *data mining*) cennych dla przedsiębiorstwa informacji i stwarzamy możliwość bezpiecznego przekształcania danych (na przykład ich standaryzacji), jeśli tylko zachodzi tego potrzeba wynikająca z celów generowanych przez *data mining*. Ujemną stroną korzystania z kopii jest konieczność dysponowania odpowiednią ilością dodatkowej przestrzeni dyskowej, na której można będzie posadowić kopię danych sporządzoną na użytek *data mining*.

Podczas tworzenia kopii dla potrzeb *data mining* można ułatwić sobie zadanie poprzez kopiowanie wyłącznie wartości wcześniej wybranych zmiennych, uważanych za istotne dla rozważanego problemu, z pominięciem wszystkich tych, na ogół bardzo licznych danych identyfikacyjnych, potrzebnych do prowadzenia działalności biznesowej, ale zbytecznych w kontekście celów *data mining*. Należy się również zastanowić nad możliwością przeniesienia do plików kopii *wszystkich* zawartych w bazie wartości rozważanych danych (przypadków, wierszy) lub też tylko *niektórych* spośród nich (jest to problem określenia *zakresu badań*). Podstawową przesłanką jest w tym przypadku wpływ liczby przypadków na czas realizacji obliczeń. Zwykle wykorzystanie większej liczby danych prowadzi do uzyskania lepszych rezultatów, ale kosztem wydłużenia czasu obliczeń. Z tego powodu przyjęta wielkość roboczej kopii rozważanego zbioru danych powinna być kompromisem pomiędzy oczekiwaną jakością rezultatów a niezbędnym czasem obliczeń.

Decyzja dotycząca wykorzystania jedynie części przypadków pociąga za sobą problem *sposobu wyboru* przypadków uwzględnionych w trakcie obliczeń – oraz kryterium odrzucenia tych pozostałych. Bardzo często stosowana jest losowa metoda doboru, która ma tę zaletę, że w najmniejszym stopniu zniekształca prawidłowości występujące w źródłowym zbiorze danych, ale w pewnych sytuacjach mogą być preferowane inne rozwiązania. Dyskutując problematykę wyboru zmiennych i przypadków z bazy danych, nie można pominąć podstawowego narzędzia, jakim jest *strukturalny język zapytań* (język SQL), który pozwala



na operowanie danymi przechowywanymi w relacyjnych bazach danych. Struktury i własności tego języka mogą wyraźnie preferować jedne, a utrudniać inne metody wyboru zmiennych i przypadków z bazy danych, co należy mieć na uwadze planując metodykę prowadzenia badań – by nie narazić się na sytuację, w której dla uniknięcia pracochłonnego przetwarzania danych na etapie ich analizy planuje się nie mniej pracochłonną metodę selekcji ograniczonego podzbioru danych.

Wstępna analiza i wstępne przetworzenie danych

Kolejny etap badań obejmuje **wstępną analizę i wstępne przetworzenie wybranych danych**. W wielu przypadkach dane pochodzące z bazy danych wymagają przed uruchomieniem algorytmów *data mining* weryfikacji i (lub) przetworzenia do postaci dogodnej do dalszej analizy. Można wskazać na wiele przesłanek uzasadniających potrzebę realizacji tego etapu badań. Pierwszą z nich są *braki w danych*, które mogą wynikać z okresowej niedostępności pewnych informacji, z braku ujęcia potrzebnych informacji w dostępnych ewidencjach, z niewprowadzenia pewnych konkretnych danych do systemu informatycznego w pewnym okresie jego eksploatacji, czy też z wielu innych powodów. Braki w danych utrudniają dalszą procedurę badawczą: mniejsza ilość dostępnych informacji powoduje zwykle, że skonstruowany model jest gorszy i wnioski uzyskane przy jego pomocy są słabiej uzasadnione. Braki w danych uniemożliwiają także wykonanie pewnych wymaganych przez *data mining* procedur numerycznych.

W przypadku stwierdzenia braków w zasobach informacyjnych należy podjąć decyzję dotyczącą ich dalszego traktowania. Najczęściej wybiera się jedno z następujących rozwiązań:

- ◆ Podejmuje się próbę uzupełnienia zbioru danych na podstawie alternatywnych źródeł informacji (np. zapisów w dokumentach źródłowych). Jest to z pewnością najlepszy sposób rozwiązania problemu braków w danych. Nie zawsze jest on jednak możliwy do realizacji, gdyż często alternatywne źródło informacji nie istnieje.
- ◆ Przeprowadza się szacowanie brakujących informacji. Zwykle polega to na budowie statystycznego lub ekonometrycznego modelu, którego celem jest wyznaczenie wartości brakujących; taki sposób postępowania umożliwia wykonanie dalszych prac obliczeniowych, jednakże oszacowania brakujących danych prawie nigdy nie mają takiej wartości informacyjnej, jak rzeczywiste, prawidłowo zebrane dane.
- ◆ Ze zbioru danych usuwane są przypadki (wiersze) zawierające braki. Trzeba jednak mieć świadomość, że postępując w ten sposób, pozbywamy się wprawdzie kłopotów numerycznych i merytorycznych powodowanych przez brakujące dane, jednocześnie jednak tracimy bezpowrotnie również części istniejącej w bazie informacji, co najprawdopodobniej ujemnie wpłynie na jakość budowanego modelu.
- ◆ Ze zbioru danych usuwane są zmienne (kolumny), w których wystąpiły braki w danych. Takie działanie też prowadzi bardzo często do poważnego zmniejszenia dostępnych zasobów informacyjnych i jest zalecane jedynie w wyjątkowych przypadkach.



Wybór pomiędzy trzecią i czwartą propozycją jest uzależniony od sposobu rozlokowania braków w zbiorze danych. Jeśli pojawiają się one głównie w wartościach jednej zmiennej, to może to przemawiać za usunięciem tej zmiennej, chociaż oznacza to wyrzeczenie się całej informacji niesionej przez tę zmienną. Natomiast gdy występujące braki dotyczą wartości różnych zmiennych, to zdecydowanie lepszym rozwiązaniem może być pominięcie w dalszych obliczeniach odpowiednich przypadków zawierających te braki.

Kolejnym problemem wymagającym rozważenia na etapie wstępnej analizy danych jest problem *wartości nietypowych*. Pojawia się on wtedy, gdy wprowadzicie w przeznaczonym do dalszego przetwarzania zbiorze danych występują wartości pochodzące z obserwacji lub z pomiaru, ale odbiegają one wyraźnie od wartości typowych. Wówczas badacz powinien odpowiedzieć sobie na pytania:

- ◆ Czy stwierdzone wartości nietypowe oddają stan rzeczywisty (czyli są tzw. anomaliami) czy też pojawiły się w wyniku błędu (na etapie pomiaru, ewidencjonowania, wprowadzania do bazy).
- ◆ Co zrobić z wartościami nietypowymi: czy pozostawić je w zbiorze danych czy też je usunąć i podjąć dalsze postępowanie analogiczne do tego, które realizowane jest w przypadku stwierdzenia braków w danych.

Odpowiedzi na powyższe pytania są trudne, powinny być podejmowane indywidualnie dla każdego rozważanego problemu badawczego. Poszukiwania odpowiedzi muszą być jednak prowadzone z dużą pieczołowitością, gdyż ich rezultaty mają istotny wpływ na efekty wszystkich dalszych prac.

Operacjonalizacja danych

Wstępne przetwarzanie danych to przede wszystkim ich *operacjonalizacja*. Polega ona na dokonaniu przekształcenia wartości analizowanych zmiennych za pomocą odpowiednio dobranych formuł matematycznych. Do najpopularniejszych metod operacjonalizacji należy skalowanie wartości rozważanych zmiennych, ich potęgowanie, logarytmowanie, wyznaczenie odwrotności lub wartości bezwzględnej, binaryzacja (na przykład realizowana poprzez uwzględnienie wyłącznie informacji o znaku wartości i zastąpienie wartości ujemnych przez „-1”, zaś wartości dodatnich przez „+1”). Do tej samej klasy przekształceń należą rozmaite filtracje wartości szeregu czasowego (polegające zwykle na uwzględnieniu wyłącznie zmian o pewnym, ściśle zdefiniowanym charakterze).

Operacjonalizacja może również polegać na ważeniu (poprzez specjalne współczynniki wagowe), normalizacji lub też standaryzacji wartości zmiennych. Sposób wykonania właściwych przekształceń zmiennych jest uzależniony od wielu czynników: może być podyktowany chęcią nadania szczególnego znaczenia pewnej zmiennej (poprzez jej ważenie), koniecznością uwzględnienia szczególnie interesującej informacji (np. znak liczby, filtracja danych), dążeniem do zmiany charakteru istniejących zależności nieliniowych na liniowe (np. poprzez logarytmowanie), czy też może wynikać z natury stosowanych algorytmów obliczeniowych (np. stosowanie sieci neuronowych wymusza wcześniejsze skalowanie danych do przedziału wartości akceptowanych w tych sieciach).



Reprezentacja danych

Wstępne przetworzenie danych obejmuje również przyjęcie właściwego sposobu *reprezentacji danych*. Ta operacja polega zwykle na przekształceniu wartości zmiennych (oryginalnych lub po wykonaniu operacjonalizacji) do postaci możliwej do dalszego przetworzenia za pomocą wybranych narzędzi badawczych. Sposób wykonania tej operacji jest bardzo mocno związany z dwoma elementami: *skalą pomiarową* wykorzystaną do wyrażenia wartości zmiennych oraz planowanymi do zastosowania *metodami analizy danych*. Nawiązując do problematyki skal pomiarowych należy zwrócić uwagę na najczęściej pojawiający się problem – w jaki sposób uwzględnić w trakcie obliczeń informacje o charakterze *jakościowym* (czyli wyrażone na nominalnej i porządkowej skali pomiarowej). Waga tego problemu jest znaczna, gdyż udział informacji jakościowych w danych przetwarzanych z użyciem technik *data mining* jest z reguły dość duży, zaś większość metod badawczych *data mining* przystosowana jest do operowania na wartościach numerycznych. Czasami występuje też problem odwrotny – wartości wyrażone na mocnych skalach pomiarowych należy przekształcić na wartości wyrażone na skalach słabych (ten kierunek zmian jest typowy na przykład przy stosowaniu drzew decyzyjnych).

Powyżej wspomniano również o wpływie metod analizy na przyjęcie sposobu reprezentacji danych – jest on rzeczywiście bardzo duży. Praktycznie każda metoda *data mining* preferuje określony rodzaj danych wejściowych i wszelkie pojawiające się na tym polu niezgodności należy starać się rozwiązywać poprzez zmianę sposobu reprezentacji danych. Wykonując tego typu przekształcenia należy postępować w sposób świadomy, kierując się wiedzą i doświadczeniem, gdyż nie każdy sposób zmiany reprezentacji jest uzasadniony, a błędy na tym etapie mogą mocno rzutować na wyniki uzyskane na dalszych etapach analizy i na efektywność prowadzonych prac (dobrym przykładem są tu algorytmy genetyczne, w których wszelkie obliczenia przeprowadzane są na właściwie zakodowanych – najczęściej binarnych – wartościach rozważanych zmiennych).

Forma prezentacji wyników obliczeń

Po zrealizowaniu przedstawionych powyżej wstępnych, bardzo ważnych i niezbędnych, etapów wstępnej analizy możemy przejść do **realizacji wybranych algorytmów obliczeniowych** *data mining*. Należy tylko stale mieć w pamięci zasadniczy paradygmat: podstawowym celem prowadzonych obliczeń jest *wzbogacenie posiadanej wiedzy decydenta o nowe informacje pozyskane z dostępnego zbioru danych*.

Uzyskane rezultaty obliczeń prezentują wydobyte z danych informacje – opisują istniejące prawidłowości, prezentują wyniki klasyfikacji, przybliżają strukturę zbiorowości, sugerują sposób dokonania wyboru. Jest rzeczą oczywistą, że rezultaty obliczeń mają stanowić odpowiedź na problem badawczy wyspecyfikowany jako cel badań. Należy jednak pamiętać, że w zależności od zastosowanej metody badawczej *forma (postać)* odpowiedzi może być całkowicie różna (nawet w przypadku rozwiązywania identycznego problemu badawczego). Stosując różne metody badawcze, możemy więc uzyskać zbliżone (lub nawet dokładnie te same) informacje, ale przedstawione w różnej postaci, a tym samym w różnym stopniu przydatne do analizy rozpatrywanych zjawisk, do prognozowania czy też do



wspomagania procesów decyzyjnych. Forma prezentacji nowych fragmentów wiedzy, uzyskanych za pomocą techniki *data mining* powinna być dostosowana również do potrzeb odbiorcy, przy czym bardziej zależy to od cech adresata informacji niż od cech samej informacji. Zwykle inna forma prezentacji wyników będzie preferowana przez doświadczonego analityka, inna przez menedżera, a jeszcze inna przez osobę po raz pierwszy zajmującą się analizowanym problemem. Do najczęściej spotykanych form prezentacji informacji pozyskanych z danych należy zaliczyć:

- ◆ *Formę graficzną* – pozwala na pogłądowe przedstawienie różnych typów problemów, ale czasami jest mało precyzyjna. Graficznie przedstawiona informacja jest z reguły łatwa do interpretacji przez człowieka, ale zwykle jest niezrozumiała dla maszyny, nie może więc być jedyną postacią danych wyjściowych w systemach, których wyniki mają być jeszcze dalej przetwarzane przez kolejne systemy wspomagające proces podejmowania decyzji. Wykresy i inne rysunki produkowane przez system *data mining* mogą charakteryzować badane obiekty (np. w postaci histogramów), prezentować istniejące zależności, strukturę zbioru (mapy percepcji, dendrogramy), sugerować sposób podejmowania decyzji (drzewa decyzyjne, schematy blokowe) itp. Dobrze dopracowany moduł produkujący graficzne prezentacje rozważanych problemów i ich rozwiązań jest wysoce użyteczny we wszystkich zastosowaniach techniki *data mining*, ale jest bardzo pracochłonny w wykonaniu i dlatego nie zawsze jest stosowany.
- ◆ *Statystyki opisowe* – charakteryzują badane aspekty rzeczywistości w postaci wartości wybranych mierników statystycznych; zwykle są proste do wyznaczenia, a jednocześnie doświadczonemu badaczowi nie sprawiają trudności w interpretacji, dzięki czemu pozwalają na szybką i dokładną ocenę analizowanego problemu.
- ◆ *Reguły decyzyjne* – prezentują pozyskane informacje w postaci stwierdzeń typu „jeżeli... to...”. Taka forma prezentacji jest również dogodna dla człowieka (oczywiście pod warunkiem, że liczba reguł nie jest zbyt duża), oraz – co ważne – może być również bezpośrednio wykorzystana przez system wspomagający proces podejmowania decyzji (np. reguły mogą zostać wprowadzone do bazy wiedzy systemu ekspertowego).
- ◆ *Równanie lub układ równań* – uzyskanie takiego opisu istniejących prawidłowości jest bardzo dogodne dla kogoś, kto chce badać metodami matematycznymi sposób zachowania się badanego fragmentu rzeczywistości. Taka postać wyniku pozwala na bardzo ogólne i daleko idące wnioski o fundamentalnych własnościach rozważanego problemu, a także pozwala na przeprowadzenie symulacji, umożliwia prognozowanie, daje wgląd w naturę problemu. Opis za pomocą równania jest bardzo zwarty, zwykle bez problemu może być wykorzystany przez człowieka i przez komputer, w prosty sposób może być przekształcony do postaci graficznej. Problem polega na tym, że uzyskanie na podstawie danych równania matematycznego jest albo bardzo trudne, albo może być obarczone szeregiem subiektywnych błędów z powodu arbitralnych założeń, jakie trzeba wprowadzić do systemu, aby w miarę łatwo zidentyfikować badany problem w formie równań matematycznych.
- ◆ *Sieć neuronową* – reprezentuje pozyskane informacje w postaci układu parametrów (zwykle wartości wag i progów) elementów współbieżnie przetwarzających dane



(tzw. sztucznych neuronów) [6]. Użycie sieci związane jest zawsze z możliwością uzyskiwania na jej wyjściu konkretnych odpowiedzi na konkretne pytania, oznacza to, że rozwiązaniem problemu, dostarczanym przez sieć, jest zawsze model problemu, a nie jego objaśnienie. Dzięki elastyczności neuronowego modelu sieć może opisywać zależności, reguły decyzyjne czy też strukturę zbioru. Może też być w prosty sposób przekształcona do postaci programu komputerowego modelującego problem, ale sposób jej działania trudno jest przekształcić do postaci łatwo interpretowalnych przez człowieka reguł decyzyjnych.

- ◆ *Graf* – pozwala na poglądowy opis zależności pomiędzy badanymi obiektami, w pewnym zakresie łączy zalety formy prezentacji graficznej i formy reguł decyzyjnych.
- ◆ *Program komputerowy* – ta forma prezentacji wyników nie jest generowana bezpośrednio przez żadną grupę metod *data mining* stosowanych do analizy danych, ale do tej postaci przekształcane są wszelkie inne formy prezentacji uzyskanych rezultatów zawsze wtedy, gdy mają być wykorzystane przez komputer – na przykład w celu bieżącego wspomagania procesu podejmowania decyzji.

Rezultaty uzyskane w trakcie analizy prowadzonej metodami *data mining* zawsze tworzą pewien *model* opisujący (mniej lub bardziej formalnie) wyodrębniony fragment rzeczywistości. Pojęcie modelu jest tu bardzo szerokie, gdyż obejmuje zarówno równanie matematyczne (lub ich układ), zestaw reguł decyzyjnych, wykres lub schemat, drzewo decyzyjne, graf czy też sieć neuronową. Modelem jest również program komputerowy implementujący odkryte prawidłowości.

Weryfikacja modelu

Wykonanie obliczeń nakazywanych przez wybraną metodę *data mining* i uzyskanie wyników w postaci wspomnianego wyżej modelu nie kończy bynajmniej procesu badawczego, gdyż uzyskane rezultaty wymagają jeszcze **weryfikacji**, która ma na celu udzielenie odpowiedzi na pytania:

- ◆ Czy uzyskany w wyniku obliczeń model działa poprawnie dla danych, które stanowiły bazę do jego utworzenia.
- ◆ Czy można oczekiwać, że utworzony model będzie działać poprawnie także dla innych danych, niż te, które były wykorzystane do jego skonstruowania.

Udzielenie odpowiedzi na pierwsze z postawionych pytań jest stosunkowo proste – wystarczy określić sposób działania modelu dla całego dostępnego zbioru danych i uzyskane wyniki skonfrontować z posiadanymi informacjami. Znacznie trudniejsza jest odpowiedź na pytanie drugie. Jak oszacować poprawność działania modelu dla danych, które są niedostępne w trakcie badań?

Mimo istniejących trudności drugie pytanie także nie powinno pozostać bez odpowiedzi. Znajomość sposobu działania modelu dla nowych danych, niedostępnych w trakcie badań, jest bardzo ważna, gdyż wskazuje na przydatność uzyskanych rozwiązań w przyszłości. Często zdolność prawidłowego działania modelu dla nowych danych (a więc dla takich, które nie były wykorzystywane w trakcie jego tworzenia) jest nazywana *zdolnością do*



generalizacji lub *zdolnością do uogólniania*. Ten typ właściwości pozwala na wykorzystanie modelu do prognozowania, wspomagania procesów decyzyjnych lub też klasyfikowania nieznanymi wcześniej obiektów. Istnieją różne sposoby szacowania posiadanej przez model zdolności do uogólniania. Do najczęściej spotykanych można zaliczyć:

- ◆ *Ocena zdolności do generalizacji za pomocą zbioru testowego*. Idea tej metody jest bardzo prosta – przed przeprowadzeniem obliczeń zbiór danych dzielony jest na dwie części, określane najczęściej jako *zbiór uczący* oraz *zbiór testowy*. Elementy wchodzące w skład zbioru uczącego zostaną wykorzystane do budowy modelu. Przypadki zaliczone do zbioru testowego nie są jednak przy tym w żaden sposób wykorzystywane, aż do chwili zakończenia prac nad modelem. Po zakończeniu obliczeń prowadzących do utworzenia modelu, jego działanie jest weryfikowane na zbiorze testowym. Przyjmuje się, że sposób funkcjonowania modelu dla zbioru testowego będzie analogiczne jak działanie modelu dla wszelkich w ogóle nowych danych (nie znanych podczas tworzenia modelu). Zachowanie modelu dla danych testowych odzwierciedla więc jego zdolność do generalizacji lub brak tej zdolności. Podstawą do przyjęcia takiego założenia jest fakt, że przypadki testowe nie uczestniczyły w tworzeniu modelu. Słabą stroną tej hipotezy badawczej jest jednak okoliczność, że zbiór testowy jest z reguły mało liczny (większą część posiadanych danych angażuje się raczej w proces uczenia modelu), więc stanowi on mało reprezentatywną próbkę dla potencjalnie nieskończonego licznego zbioru praktycznych zagadnień, które mają być rozwiązywane za pomocą modelu podczas jego normalnej eksploatacji.

Stosując przedstawioną metodę oceny modelu, należy podjąć dwie bardzo ważne decyzje:

- a) jaka powinna być liczebność zbioru uczącego, a jaka zbioru testowego;
- b) w jaki sposób dokonać podziału analizowanego zbioru na część uczącą i testową.

Na tak postawione pytania trudno udzielić jednoznacznych odpowiedzi. Próbując odpowiedzieć na pierwsze pytanie, można stwierdzić, że większość przypadków powinna zostać zaliczona do zbioru uczącego, zaś pozostałe do zbioru testowego. Ta „większość” oznacza zwykle 60–80 procent. Odpowiadając na drugie pytanie, można stwierdzić, że sposób podziału powinien zostać zaprojektowany w taki sposób, aby zarówno jeden, jak i drugi zbiór miały charakter reprezentatywny. W praktyce najczęściej dokonuje się przypisania przypadków do obu zbiorów w sposób losowy (dbając jednak o zachowanie ustalonych wcześniej proporcji w liczebności zbioru uczącego i testowego). Tylko w szczególnych przypadkach, gdy losowa procedura podziału przypadków nie zapewnia reprezentatywności, stosuje się inne metody.

Przedstawiona metoda szacowania jakości modelu posiada przykrą niedogodność – wydzielając zbiór testowy, zmniejszamy ilość informacji możliwej do wykorzystania na etapie konstruowania modelu, co może wpłynąć na pogorszenie jakości uzyskanych rozwiązań. Jest to szczególnie dotkliwe wtedy, gdy dysponujemy pierwotnym zbiorem danych o małej liczbie elementów. Przy małej liczebności zbiorów może pojawić się jeszcze jeden problem – ocena zdolności do generalizacji, dokonana na podstawie zbioru testowego, może nie odzwierciedlać w sposób prawidłowy rzeczywistego



poziomu tej cechy. Pewną próbą rozwiązania przedstawionych problemów są przedstawione poniżej dwie inne techniki szacowania jakości modelu.

- ◆ *Testowanie krzyżowe.* Ten sposób oceny modelu stanowi rozwinięcie przedstawionej powyżej metody wykorzystującej zbiór testowy. Sposób postępowania jest w tym przypadku następujący: dostępny zbiór danych dzieli się na n części (podział elementów do poszczególnych podzbiorów odbywa się zwykle w sposób losowy); następnie $n-1$ części wykorzystuje się w charakterze zbioru uczącego, zaś pozostała n -ta część spełnia funkcję zbioru testowego. Przedstawioną procedurę powtarza się n razy, przy czym przy każdej iteracji inny podzbiór wykorzystywany jest jako zbiór testowy. Postępując w ten sposób, otrzymujemy n ocen modelu dla różnych zbiorów testowych, pokrywających w sumie całość dostępnej informacji. Często uzyskane w testowaniu krzyżowym mierniki agreguje się do pojedynczej wartości (np. poprzez ich uśrednienie). Zaletą takiego sposobu przeprowadzania oceny zdolności do generalizacji jest przede wszystkim to, że wszystkie dostępne elementy danych wykorzystywane są zarówno do tworzenia modelu, jak i do jego oceny (oczywiście w kolejnych powtórzeniach dany przypadek występuje tylko jeden raz – albo jako element zbioru uczącego, albo zbioru testowego). Testowanie krzyżowe można polecić wówczas, gdy dysponujemy stosunkowo niewielkim zbiorem danych. Podstawową wadą jest wzrost (w przybliżeniu n -krotny) czasu obliczeń.
- ◆ *Zastosowanie metod bootstrapowych.* Na początku należy podkreślić, że technika bootstrapowa ma znacznie szerszy zakres zastosowań niż problematyka oceny jakości modeli. Ogólnie rzecz ujmując, jest ona zaawansowaną techniką symulacyjną, pozwalającą na podstawie dostępnego zbioru danych oszacować wartości i rozkłady pewnych statystyk. W naszym zastosowaniu będą to statystyki określające jakość modelu, ale podobne postępowanie można zastosować do określania wartości i rozkładów innych wielkości.

Punktem wyjścia jest n -elementowy zbiór danych. Prowadzone obliczenia mają charakter iteracyjny (przy czym liczba powtórzeń jest duża, zwykle większa od tysiąca) i obejmują następujące etapy:

- Na podstawie pierwotnego zbioru danych tworzony jest tzw. zbiór bootstrapowy; jest on konstruowany poprzez losowanie ze zwracaniem elementów ze zbioru pierwotnego i umieszczanie ich w zbiorze bootstrapowym (ponieważ stosuje się poprzez losowanie ze zwracaniem, więc w zbiorze bootstrapowym pewne elementy ze zbioru pierwotnego mogą wystąpić wielokrotnie, zaś inne mogą się wcale nie pojawić). Przy każdym powtórzeniu obliczeń procedura losowania jest powtarzana, więc skład zbioru bootstrapowego jest za każdym razem inny.
- Przeprowadzane są obliczenia przewidziane w używanej metodzie *data mining* przy wykorzystaniu zbioru bootstrapowego. W naszym przypadku, gdy celem prowadzonych obliczeń jest ocena zdolności do generalizacji modelu utworzonego techniką *data mining*, w każdej iteracji wykonane zostaną następujące czynności: zbiór bootstrapowy podzielony zostanie na zbiór uczący i zbiór testowy; zbiór



uczący posłuży do skonstruowania modelu, zaś na podstawie zbioru testowego obliczona zostanie pewna miara oceniająca jakość modelu.

Po zakończeniu obliczeń będziemy dysponować miernikami jakości modelu obliczonymi w trakcie kolejnych powtórzeń (czyli zwykle będziemy posiadać ponad 1000 wartości tych mierników). Ten ciąg wartości może zostać uśredniony (wyznaczona średnia stanowi oszacowanie rzeczywistej wartości miernika) lub może posłużyć do oszacowania rozkładu interesującej nas wartości (możliwe jest na przykład oszacowanie jego wariancji, która umożliwi wnioskowanie na temat wiarygodności oszacowania interesującego miernika).

Z uwagi na wymaganą dużą liczbę powtórzeń stosowanie techniki bootstrapowej jest bardzo kosztowne obliczeniowo (praktycznie niemożliwe jest przeprowadzenie obliczeń bez korzystania z techniki komputerowej), uzyskane rezultaty w pełni jednak wynagradzają poniesione nakłady.

Sposób pomiaru jakości modelu

Powyżej przedstawione zostały ogólnie różne sposoby przeprowadzania weryfikacji modeli uzyskiwanych w następstwie stosowania technik *data mining* do analizy dużych zbiorów danych. Omawiając poszczególne techniki, nie wskazano jednak sposobu pomiaru jakości modelu. Ten ważny problem wymaga również omówienia i zostanie to wykonane właśnie w tym podrozdziale.

Przede wszystkim należy podkreślić, że sposób oceny jakości modelu jest uzależniony od wielu czynników, w tym przede wszystkim od rodzaju rozpatrywanego problemu – więc w takim układzie zostanie on niżej scharakteryzowany:

- ◆ W przypadku podejmowania prób budowy modeli służących do *opisu zależności* konstruowane mierniki jakości modelu podzielić możemy na dwie zasadnicze grupy: mierniki bezwzględne oraz mierniki względne.

Mierniki bezwzględne uwzględniają w sposób zagregowany zróżnicowanie pomiędzy *wartościami rzeczywistymi* (czyli tymi, które zostały zaobserwowane i wchodziły w skład zbioru danych) z *wartościami teoretycznymi* (czyli tymi, które zostały obliczone na podstawie modelu). W przypadku korzystania ze zmiennych o charakterze numerycznym takim najpopularniejszym sposobem agregacji jest obliczenie sumy kwadratów różnic pomiędzy wspomnianymi wartościami (czyli miary błędu zwanej SSE).

Oprócz bezwzględnych jakości modelu stosowane są również *mierniki względne*. Bazują one na porównaniu mierników bezwzględnych, wyznaczonych dla ocenianego modelu, z analogicznymi miernikami uzyskanymi dla innego modelu, stanowiącego punkt odniesienia. Mierniki względne służą więc przede wszystkim do porównywania jakości różnych modeli. W przypadku modeli opisujących zależności pomiędzy zmiennymi jako punkt odniesienia wykorzystuje się często liniową funkcję regresji.

- ◆ Budując modele rozwiązujące problemy *data mining* z zakresu *klasyfikacji wzorcowej* można przyjąć podobny schemat podziału mierników: mierniki bezwzględne będą



wtedy porównywały rzeczywisty i teoretyczny sposób zaklasyfikowania poszczególnych obiektów, zaś mierniki względne będą porównywały jakość różnych metod klasyfikujących. Punktem wyjścia do wyznaczania wartości mierników bezwzględnych będą (w przypadku problemów klasyfikacji wzorcowej) dwie wartości: liczba (lub ich odsetek) obiektów zaklasyfikowanych *prawidłowo* oraz liczba (odsetek) obiektów zaklasyfikowanych *błędnie*. Te podstawowe wartości można poddawać dalszej analizie – można je analizować w rozbiciu na poszczególne grupy obiektów lub też można badać, jakiego typu błędy popełniane są najczęściej.

- ◆ Ocena modeli rozwiązujących problemy z zakresu klasyfikacji bezwzorcowej jest znacznie trudniejsza. Podstawową przyczyną, utrudniającą ocenę jakości modelu, jest to, że musimy ocenić poprawność odkrytej przez model struktury zbioru obiektów, w sytuacji, gdy nie jest dostępna żadna informacja o *rzeczywistych* zależnościach występujących pomiędzy obiektami lub ich grupami. Podstawowa idea stosowanych mierników jest następująca: preferowany powinien być taki sposób podziału obiektów na grupy, który *minimalizuje* zróżnicowanie obiektów należących do tych samych grup, a jednocześnie *maksymalizuje* zróżnicowanie obiektów należących do różnych grup. To, bardzo ogólne, stwierdzenie znalazło swoje odzwierciedlenie w bardzo dużej liczbie mierników zaproponowanych i scharakteryzowanych w literaturze.
- ◆ Ocena modeli służących do *analizy szeregów czasowych* przebiega w sposób podobny jak ocena modeli regresyjnych (służących do opisu zależności) bądź też klasyfikacyjnych (rozwiązujących zagadnienia z zakresu klasyfikacji wzorcowej). Analogię z modelem regresyjnym względnie klasyfikacyjnym można przeprowadzić w zależności od ilościowego bądź jakościowego charakteru analizowanej zmiennej tworzącej szereg czasowy. W tym zadaniu stosować można również podobne mierniki jak w zadaniach oceny modeli regresyjnych bądź klasyfikacyjnych, ale w szczególnych przypadkach przydatne mogą być również inne miary, dostosowane do specyfiki szeregów czasowych.

I tak wśród mierników o charakterze bezwzględnym przydatny jest często miernik porównujący rzeczywisty i prognozowany *kierunek* zmian wartości zmiennej, zaś wśród miar względnych na uwagę zasługuje miernik porównujący jakość skonstruowanego modelu z tzw. *modelem naiwnym* (to jest takim, który zakłada, że w chwili $t+1$ wartość prognozowanej zmiennej utrzymywac się będzie dokładnie na takim samym poziomie jak w chwili t).

Ocena regresyjnych modeli szeregów czasowych może być również oparta na uzyskanym szeregu reszt, w którym testuje się obecność autokorelacji lub też które analizuje się za pomocą metod analizy widmowej;

- ◆ Do krótkiego omówienia pozostała jeszcze problematyka oceny poprawności wyników metod *data mining* dostarczających przesłanek do racjonalnego wyboru optymalnego (lub suboptymalnego) wariantu. Ocena jakości algorytmów tego typu opiera się na porównaniu efektów odpowiadających ocenianemu zbiorowi wybranych elementów z efektami możliwymi do uzyskania po zastosowaniu innych, alternatywnych metod wyboru elementów. To ogólne stwierdzenie może zostać w różny sposób



doprecyzowane, w zależności od rozważanego problemu. Często zarysowana procedura oceny wzbogacana jest o jeszcze jeden element, którym jest informacja o preferowanej *liczbie* elementów. W zastosowaniach praktycznych najczęściej preferowane są mniej liczne zbiory elementów (np. dokonując wyboru zmiennych objaśniających, zwykle dążymy do minimalizacji ich liczby), w związku z czym definicja miernika jakości wzbogacana jest o tzw. *człon kary*, który służy do pogarszania uzyskanej oceny wraz ze wzrostem różnicowania pomiędzy oczekiwaną i uzyskaną liczbą wybranych elementów.

Scharakteryzowane sposoby oceny modeli uzyskanych w wyniku eksploracji danych mają charakter ogólny i są wyłącznie uzależnione od typu rozpatrywanego problemu. Warto jednak pamiętać, że rozpatrywane problemy mają charakter biznesowy i, że do oceny uzyskanych rozwiązań należy również stosować (w razie potrzeby) mierniki o takim charakterze, które pozwolą na oszacowanie efektu ekonomicznego uzyskanego rozwiązania. Dokładny sposób oceny jest jednak bardzo mocno związany z charakterem rozpatrywanego problemu i nie może tu być omówiony w sposób ogólny.

Uwagi końcowe

Rozważania na temat natury, sposobów gromadzenia, przetwarzania i przesyłania informacji prowadzone są przez przedstawicieli różnych dyscyplin badawczych, w tym również na gruncie informatyki. Wydawałoby się, że obserwowany na początku bieżącego stulecia bardzo dynamiczny rozwój tej dziedziny pozwoli na rozwiązanie zasadniczych problemów informacyjnych, zwłaszcza tych generowanych przez problemy związane z rozwojem nauki i wspomaganie badań naukowych. Niestety, można wskazać na wiele przypadków, w których zastosowanie nowoczesnych komputerów, nawet połączonych ze sobą za pomocą szybkich sieci teleinformatycznych, nie tylko nie rozwiązało istniejących problemów informacyjnych, ale stało się źródłem nowych kłopotów. Sytuację tę bardzo trafnie charakteryzuje znane powiedzenie: „*od komputerów oczekiwaliśmy fontanny wiedzy, a dostaliśmy potop danych*”. Z tego powodu konieczne staje się sprzężenie nowoczesnych zdobyczy informatyki, związanych z możliwościami skutecznego gromadzenia i przesyłania danych, z metodami i narzędziami pozwalającymi na odkrycie zawartych w danych informacji, a tym samym na powiększenie posiadanego zasobu wiedzy. Przedstawiona w tej pracy propozycja wykorzystania metod sztucznej inteligencji do uzyskiwania nowych odkryć naukowych na podstawie pozornie całkowicie wyeksploatowanych wyników badań empirycznych, będąca istotą tej pracy, zdecydowanie wpisuje się w logikę rozważanych działań.

Przedstawiona w artykule koncepcja nie została jeszcze nigdzie praktycznie wykorzystana, nie ma więc możliwości przytoczenia przykładów jej zastosowania ani dyskusowania sensowności wyników uzyskanych przy jej pomocy. Być może przy próbie zastosowania tej koncepcji wyłonią się jakieś nieoczekiwane trudności, być może w wielu laboratoriach okaże się, że dokumentacja dawnych eksperymentów nie jest wystarczająco dokładna ani dostatecznie kompletna, żeby mogła stanowić podstawę sugerowanych tu poszukiwań, być może wreszcie wszystkie „odkrycia” naukowe, dokonywane przez algorytmy data mining



w starych danych empirycznych będą w istocie trywialne i mało użyteczne. Być może. Jednak jeśli nie podejmiemy próby, to się nie przekonamy, czy to jest, czy też nie jest możliwe, dlatego zachęcam czytelników tego artykułu do prób zastosowania naszkicowanej metodyki oraz do publikowania uzyskiwanych wyników. Pierwsza osoba, która nadeśle na adres autora tej pracy swoją publikację (wydrukowaną!) referującą oryginalny wynik naukowy, odkryty przy zastosowaniu opisanej w tym artykule metodyki wtórnego przekopywania pozornie całkowicie wyeksploatowanych wyników badań eksperymentalnych, otrzyma nagrodę w wysokości 2000 zł. Fundatorem nagrody jest firma StatSoft Polska.

Bibliografia

1. Berry M. J. A., Linoff G., *Data Mining Techniques For Marketing, Sales, and Customer Support*, Wiley Computer Publishing, 1997.
2. Deboeck G., Kohonen T. (Eds), *Visual Explorations in Finance with Self-Organizing Maps*, Springer-Verlag, London, 2000.
3. Heidsieck C, Uhr W, *Systematizing and Evaluating Data Mining Methods*, w: Decker R., Gaul W. (red.), *Classification and Information Processing at the Turn of the Millennium*, Springer-Verlag, Heidelberg, 2000.
4. Percival D. B., Walden A. T., *Wavelet Methods for Time Series Analysis*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2000.
5. Tadeusiewicz R., Flasiński M., *Rozpoznawanie obrazów*, Wydawnictwo Naukowe PWN, Warszawa, 1991.
6. Tadeusiewicz R., *Sieci neuronowe*, Akademicka Oficyna Wydawnicza RM, Warszawa, 1993.
7. Tadeusiewicz R.: *The Application of Neural Networks in Biotechnology and Biomaterials*, Prace Mineralogiczne, nr 89, Komisja Nauk Mineralogicznych PAN, 2000, pp. 9–17.
8. Tadeusiewicz R.: *Odkrycia bez próbki. Możliwość dokonywania dodatkowych odkryć naukowych poprzez drążenie (z użyciem sztucznej inteligencji) pozornie całkowicie wyeksploatowanych danych empirycznych*. Rozdział w książce: Nalepa I. (red.): *Szlaki przekazywania sygnałów komórkowych*, Instytut Farmakologii PAN, Kraków 2004, ss. 169-184.