SZTUCZNA INTELIGENCJA JAKO NARZĘDZIE WSPOMAGAJĄCE BADANIA BIOMATERIAŁÓW

RYSZARD TADEUSIEWICZ

AKADEMIA GÓRNICZO-HUTNICZA W KRAKOWIE

Wprowadzenie

Tworzenie nowych biomateriałów, badanie ich właściwości, dobór optymalnego materiału do ustalonego zastosowania lub poszukiwanie technologii, pozwalającej uzyskać materiał o pożądanych właściwościach - to tylko niektóre z wielu typów badań, jakie są obecnie prowadzone w dziedzinie biomateriałów. Badania te są prowadzone z dużym nakładem środków a także wymagają dużego wysiłku, przeto jest racjonalne poszukiwanie metod, które pozwolą na maksymalnie efektywne wykorzystanie wyników tych badań.

Na pozór można 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, więc nagromadzenie wyników wielu takich eksperymentów - odpowiednio przebadanych i przeanalizowanych - może przynieść nowe spostrzeżenia (lub nawet nowe odkrycia), oparte wyłącznie na pogłębionej analizie nagromadzonych już danych. Analizy takiej nie można jednak prowadzić „ręcznie", gdyż zawartość użytecznej, ale jeszcze nie wyeksploatowanej wiedzy, w wynikach zarejestrowanych badań - jest stosunkowo niska. Potrzebne są do tego specjalne narzędzia informatyczne - i o tych właśnie narzędziach będzie mowa w tym referacie. Zanim jednak narzędzia te zostaną scharakteryzowane (i zarekomendowane do użycia przez badaczy parających się tworzeniem i aplikacjami biomateriałów) - warto dodać jeszcze kilka słów celem właściwego postawienia rozważanego tu problemu.

Zaczynając od stwierdzeń najbardziej oczywistych można odnotować fakt, że każdy badacz w momencie podejmowania określonego eksperymentu ma pewną hipotezę naukową, którą doświadczenia laboratoryjne potwierdzają - albo nie. Eksperyment jest tym lepszy, im precyzyjniej jego wynik nawiązuje do postawionej hipotezy i im precyzyjniej pozwala rozstrzygnąć o jej prawdziwości lub fałszywości. Jednak niezależnie od tego podstawowego celu i podstawowego efektu poznawczego - wynik przeprowadzonych badań wzbogaca zawsze wiedzę o problemie w znacznie szerszym zakresie. Wyniki pomiarów, uzyskane obrazy struktury materiału, pomiary jego właściwości, składu, interakcji z żywą tkanką - zawierają także odpowiedzi na pytania, których jawnie nie postawiono. Jeśli nagromadzimy dużo takich wyników i zaczniemy je przeszukiwać pod kątem takich odpowiedzi na nie postawione wcześniej pytania, to możemy uzyskać nowe informacje, wartościowe naukowe oraz użyteczne praktycznie - właściwie za darmo.

Nasuwa się tu nieodparcie analogia z górnictwem, które

ARTIFICIAL INTELLIGENCE AS A TOOL TO SUPPORT THE DEVELOPMENT AND TESTING OF BIOMATERIALS

RYSZARD TADEUSIEWICZ

ACADEMY OF MINING AND METALLURGY IN CRACOW

Introduction

Development of new materials, testing their properties, selection of the optimal material for the given application and searching for technologies to develop materials having the required properties are only few examples of research work in the field of biomaterial engineering. These studies are rather costly and labour-consuming, hence it seems worthwhile to seek new methods to effectively utilise the results.

At the first glance, this remark seems redundant - researchers do take good care to make use of the results of their work: they publish papers, search for new partners to implement new techniques, make use of earlier results in their subsequent work. It does not mean, however, that the results of studies are fully utilised because each experiment gives the answers to more questions than were actually posed. Accordingly, collecting, analysing and processing of the results of numerous experiments may lead to new observations (or even new discoveries) based only on thorough analysis of thus gathered data. However, this analysis cannot be performed „manually" as the fraction of useful though not utilised knowledge „contained" in the results of registered tests is relatively small. Such an analysis requires specialised IT tools, and that is the subject of the present study. Before these tools are described (and recommended for use in biomaterials engineering and applications), it is worthwhile to add a few words to provide the best formulation of the problem.

Let us start with an obvious, well-established fact: each researcher formulates a certain hypothesis before he begins the experiments and laboratory tests either verify or disprove it. The closer the relationship between tests results and the hypothesis and the more precise answers it gives, the more effective the experiment is considered to be. Yet regardless of this first objective and the basic cognitive effects, the results of tests contribute to the prior knowledge in a much wider sense. Measurements, obtained images of material structure, measurements of material properties, composition and interactions with the living tissues give also answers to questions that have not been explicitly asked. When one collects results of numerous tests and starts sorting through them with an eye to find the answers to unasked questions, one may come across new, useful and valuable information - practically for free.

Here comes the analogy to the mining sector where mining operations are followed by extraction of useful minerals, while the remaining materials are piled on the dumpsites. However, the dumps may also hide useful materials, too - it is necessary to locate them and find the suitable extraction techniques since extraction of a useful mineral or metal from the material mined to yield other products is usually much more difficult than extraction of the „first" mineral. Never-

wydobywa z trudem jakieś surowce, następnie wydziela z nich użyteczne produkty - a pozostały materiał odrzuca na hałdy. Jednak w tych hałdach mogą się także mieścić użyteczne produkty, tylko trzeba wykryć ich obecność i opracować technologie ich pozyskiwania, gdyż pozyskanie użytecznego produktu z surowca, który był wydobywany z myślą o innym produkcie - jest z reguły trudniejsze, niż ekstrakcja produktu pierwotnego. „Trudniejsze" - nie oznacza jednak „niemożliwe", trzeba tylko mieć odpowiednie narzędzia.

W zakresie rozważanych w tym referacie problemów badawczych związanych z biomateriałami, narzędziem zalecanym do wydobycia nowej wiedzy z pozornie całkowicie już wyeksploatowanych danych doświadczalnych jest sztuczna inteligencja. Tak nazywana jest dziedzina informatyki charakteryzująca się między innymi tym, że potrafi dostarczać **odpowiedzi na nie postawione jawnie pytania**. Mimo kontrowersyjnej nazwy dziedzina ta potwierdziła już w wielu obszarach swoją użyteczność, jest więc możliwe i celowe użycie jej także w związku z potrzebami badań nad biomateriałami, gdyż przy jej mądrym użyciu może dochodzić do wzbogacania wiedzy (między innymi w obszarach badań biomateriałów) - w istocie bez dodatkowych nakładów na nowe badania laboratoryjne [Tadeusiewicz, 2000].

Elementy sztucznej inteligencji

Celem tego rozdziału jest wprowadzenie Czytelnika w problematykę sztucznej inteligencji, a dokładniej pewnej jej części, tak zwanej eksploracyjnej analizy danych. Punktem wyjścia do rozważań jest definicja metod eksploracyjnej analizy danych (określanych również jako metody *data mining*). Następnie scharakteryzowano podstawowe typy problemów, które mogą być rozwiązywane za pomocą omawianej tu grupy metod. W końcowej części referatu zaprezentowano w sposób systematyczny i uporządkowany wszystkie etapy procesu eksploracyjnej analizy danych, co może stanowić podstawę dotworzenia 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,

theless „difficult" does not mean „impossible" - one has to provide specialised tools.

As far as research problems involved in biomaterial engineering are concerned, *artificial intelligence* seems to be a useful tool for uncovering the hidden knowledge from seemingly fully utilised experimental data. Artificial intelligence is a branch of information science, which is able to offer answers to questions that were not explicitly asked. Although the name is slightly controversial, artificial intelligence has already proved to be a useful tool in many areas, hence it is worthwhile to use in studies on biomaterials. When used in a prudent manner, it may enrich our existing knowledge (in the area of biomaterial testing) without any extra expenses on laboratory facilities [Tadeusiewicz, 2000].

Elements of Artificial Intelligence

The main aim of this section is to provide the Reader with the basic information about artificial intelligence, or to be more precise - about one of its branches called the exploratory data analysis. The starting point is the definition of the exploratory data analysis (also known as *data mining*). It is followed by presentation of the fundamental types of problems, which can be solved using the group of methods presented here. The final section of the study presents a systematic and ordered description of all stages in the exploratory data analysis process, which might become a starting point for developing practical applications of the presented method.

Artificial intelligence techniques are based on data processing, therefore it might be worthwhile to examine this notion in its most general sense. The meaning of the word *„information"* is very broad. It covers a written or aural information about our relatives as well as opinions relating to people's views and preferences published in various sources. Modern manufacturing systems are based on information, it helps individuals and companies to work more effectively and outdo the competitors. No matter whether the information is treated as *confidential and personal* or *an aid in fighting business rivals or as the strategic resource of the company,* still it will play a major role. The importance we attach to information is so great, that we have to seek and take care of it, store it properly and protect from certain dangers. Proper p*resentation*, *processing and transfer* of information are important, too. *Information shortages* or *delays in information delivery* have negative consequences. However, an overflow of information is a dangerous thing, too. Information ought to agree with reality, it must not be *ambiguous or contradictory*.

The term *„information"* is closely related to *„data"*. There are more than ten definitions of „data"; for the purpose of the present study data will mean information that is recorded and processed to get a predetermined, symbolic form. This definition emphasises that data means recorded information, hence one can refer to it many times while it would remain in the same form and that the data have a definite, repeatable representation, which implies data structuring. Data comprises information represented by symbols: digits, letters, sounds, and graphic symbols. Those symbols are not collected in a random, haphazard manner but form structures displaying certain regularities. The data include digits (digit sequences), words (sign sequences), diagrams or images (sequences of points or other graphic symbols). These are not **any** sequences, but they are subject to certain limitations. For example, a digit may have only one separator indicating the fraction, a sequence of signs becomes a word only when it corresponds to an entry in a dictionary of a specified language, while a collection of graphic symbols is regarded as a drawing only when the symbols are properly chosen and distributed.

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 nie znaczących 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 głównie umysł odbiorcy informacji. Nowa informacja poprzez synergię z wiedzą wcześniej zgromadzoną bywa często źródłem zupełnie nowej wiedzy, pozornie nie wynikającej 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 dokonywanych 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.

Do the data constitute information? That depends in the first place on the human reception. When the receiver is able to interpret the data, then they should be treated as information. When they seem to be only incomprehensible sequences of symbols, it is not justified to call them „information". We have to take into account subjectivism (or rather relativism - as interpretation is an ability of an individual person) of this statement. The contents of a laboratory database or a new chemical formula will be a valuable source of information for some people while for others it will be the sequence of meaningless symbols. Data describe some fragments of reality, such as experimental results, yet they usually require proper presentation, processing or aggregation to make them legible, understandable and useful. Duly processed and well-presented data possess the features of useful information. The **data** ought to be treated as a **raw material** and information will be worked out from the data by adding the indispensable element - intelligence (be it human or artificial) linked with the awareness of the information process which allows for suitable selection, segregation and representation of data.

The set of available, interrelated bits of information constitutes knowledge, which is generated mainly in the receiver's brain. By way of synergy with the prior knowledge, new information becomes the sources of new knowledge- apparently not directly related to the delivered information since a seemingly minor bit of information may decide about global understanding of a process or phenomenon. When this understanding is achieved through a formalised system of references, we can speak about development of a theory or creation of a model. In everyday experience of every human being we can find many situations when the buildup of knowledge (internal) seems inadequate in proportion to the quality and quantity of acquired information. Sometimes it is necessary to laboriously gather a vast number of seemingly irrelevant bits of information yet no buildup of real knowledge can be achieved for long time, when quite unexpectedly, after acquiring another seemingly unimportant bit of information, there comes a sudden revelation: all facts become clear and correlations obvious while the final effect - knowledge buildup, is rapid as a result of crystallisation of information in the human brain well-supplied with information. Still the fact remains that gathered bits of information are the source of knowledge while information has its source in data, interpreted by the receiver and accordingly processed.

The nature of information as well as methods for information collection, processing and transfer have been considered by scholars from various branches of science, including information science. It might seem that rapid development in this field at the end of the 20th and at the beginning of the 21st century would have allowed for solving the fundamental problems related to information, particularly those generated by academic and commercial practices. However, there are many cases when application of advanced computers connected by fast tele-information networks did not solve the existing problems, but even created new ones. This situation was well described by Gregory Piasetsky-Sharpio who remarked that *we expected a fountain of information from computers, what we really got was a flood of information.* That is why it is necessary to combine advanced information technologies (providing for effective data collection and transfer) with methods and tools allowing for **uncovering** of contained information, thus enhancing the knowledge resources.

Knowledge build-up through information acquired from data has a long history. We owe a lot to statistics and related areas, where the research have been involved for years in discovering and teaching others how to interpret data (especially when there are so numerous), how to present

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 w drugiej połowie XX wieku i w początkowej fazie bieżącego stulecia bardzo dynamiczny rozwój tej dziedziny pozwoli na rozwiązanie zasadniczych problemów informacyjnych, zwłaszcza tych generowanych przez praktykę naukową i gospodarczą. 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 scharakteryzował Gregory Piasetsky-Sharpio, który stwierdził: ***„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.

Wzbogacanie wiedzy na podstawie informacji pozyskanych z danych ma bardzo długą historię. Szczególne zasługi na tym polu ma statystyka i wywodzące się z niej dyscypliny, których przedstawiciele od dziesięcioleci odkrywają i uczą, jak można interpretować zgromadzone dane (zwłaszcza, gdy jest ich bardzo dużo), jak je prezentować, jak badać i jak opisywać występujące w nich prawidłowości. Począwszy od połowy XX wieku równolegle z rozwojem statystyki ma również miejsce szybki rozwój informatyki. Powiązanie zdobyczy obu dziedzin jest korzystne dla obu stron, ponieważ:

- systemy komputerowe pozwalają na wygodną realizację wszelkich, w tym także wyrafinowanych i złożonych obliczeniowo algorytmów analizy danych, wypracowanych na gruncie statystyki,
- metody statystyczne pozwalają na systematyczną i dobrze kontrolowaną analizę olbrzymich zasobów danych gromadzonych i przechowywanych w komputerowych bazach danych,
- następuje korzystna integracja metod wypracowanych na gruncie statystyki z metodami rozwijanymi przez badaczy działających na polu informatyki (głównie sztucznej inteligencji), sprzyjająca rozwojowi obydwu wymienionych dziedzin.

W rezultacie integracji osiągnięć statystyki z metodami wypracowanymi na gruncie sztucznej inteligencji i teorii baz danych pojawiła się między innymi grupa nowych metod analizy danych określana w literaturze światowej mianem metod *data mining*. Polski odpowiednik tego terminu nie został jeszcze skutecznie zaproponowany (chociaż w wielu pracach pojawiały się już różne pomysły na ten temat). Najbliższym znaczeniowo równoważnikiem anglojęzycznego *data mining* wydaje się termin „przeszukiwania danych", chociaż brakuje mu na pewno zawartej w angielskim oryginale, przemawiającej do wyobraźni wizji procesu drążenia szybów i korytarzy w ciemnych czeluściach masywów jałowych danych - w poszukiwaniu cennego kruszcu czystej wiedzy lub promienia światła na końcu tunelu.

Jakkolwiek byśmy jednak nie nazwali po polsku procesu *data mining* - ważniejsza od nazwy jest charakterystyka formalnych metod i komputerowych narzędzi, służących do tego celu. Definicja tego typu narzędzi podana została w pracy [Gatnar, 1997]:

„Data mining to określenie grupy metod szeroko rozumianej analizy danych mających na celu identyfikację nieznanych wcześniej prawidłowości występujących w dużych zbiorach danych. Powstałe wyniki mają postać łatwą do interpretacji przez prowadzącego badania."

and analyse those data and how to describe the existing regularities. Since the second half of the 20th century rapid development of statistics has been accompanied by tremendous advancements in information science. Combination of the achievements in these two areas is beneficial for all, because:

- computer systems allow for easy implementation of all, even sophisticated and complex data analysis algorithms developed using statistical methods
- statistical methods allow for systematic and controlled analysis of vast sources of data stored and backed in computer databases
- integration of statistical methods with those developed in information science (particularly artificial intelligence) is beneficial and promotes further development in these two areas

In the consequence of integration of recent advancements in statistics and information science, a new data analysis methods appeared, which are termed „*data mining*" (this name can be found in publications on the subject). No adequate Polish equivalent is available yet (though several proposals can be found). The expression „*przeszukiwanie danych*" (data searching) seems close in meaning, though it lacks the vision of shafts and driving tunnels through the darkness and depths of irrelevant data, us being in quest for valuable knowledge - seen as a light at the end of the dark tunnel.

Whatever the Polish translation of „*data mining*", the more important point is to characterise the formal methods and computer tools used for that purpose. These tools were defined in the work by Gatnar [Gatnar, 1997], who states that *data mining* is the description of a group of methods of broadly understood data analysis aimed to identify yet unknown regularities in large sets of data and the results are easy to interpret by the researchers.

In the work [StatSoft, 1997], the term „*data mining*" was translated as *„zgłębianie danych"* (data penetration), defined as an analytical process used for exploration of vast resources of data (...) in search for logical schemes and systematic correlations between variables, and then for evaluation of results through application of newly- discovered schemes for new sub-sets of data.

While discussing the fundamental principles of *data mining*, most important characteristics of these methods ought to be highlighted:

- The methods of *data mining* are diverse and the group of methods is continually enriched. They are derived from statistics, information science, signal analysis, mathematics and graphics.
- They are mostly induction methods. Conclusions are always based on the analysis of available data sources, not on the basis of abstract theories formulated a priori.
- The results are easy to interpret - and hence can be used in practical applications
- The methods considered here are independent of semantic content of examined information, which ensures uniformity of methods used to examine even very diverse problems
- The aim of these method is providing the description of the examined fragment of reality or forecasting.
- The methods are directed on practical applications, first of all supporting the decision- making processes

The scope of application of *data mining* methods and clasification of tasks

The choice of a suitable method of data analysis depends on the type of encountered problem. Before the available

Z kolei w pracy [StatSoft, 1997] termin *data mining* przetłumaczony został jako zgłębianie danych, które definiowane jest jako:

„proces analityczny, przeznaczony do eksploracji dużych zasobów danych (...) w poszukiwaniu logicznych schematów oraz systematycznych współzależności pomiędzy zmiennymi a następnie do oceny wyników poprzez zastosowanie wykrytych schematów dla nowych podzbiorów danych."

Wskazując na podstawowe cechy metod *data mining* należy zwrócić uwagę na następujące ich właściwości:

- Metody *data mining* są grupą bardzo zróżnicowaną i stale wzbogacaną. Wywodzą się one ze statystyki, informatyki, analizy sygnałów, matematyki, grafiki.
- Należą do metod o charakterze indukcyjnym. Formułowane wnioski wypływają zawsze z analizy dostępnych zbiorów danych, a nie z a priori przyjmowanych abstrakcyjnych teorii.
- Uzyskiwane rezultaty analizy są zwykle proste w interpretacji - i na tym polega ich praktyczna użyteczność.
- Rozważane metody są niezależne od semantycznej treści przeszukiwanych informacji, dlatego pozwalają na zunifikowane badanie bardzo zróżnicowanych grup zagadnień.
- Celem ich stosowania może być dostarczenie opisu badanego fragmentu rzeczywistości, bądź też prognozowanie.
- Są ukierunkowane na zastosowania praktyczne, przede wszystkim wspomaganie procesów decyzyjnych.

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*.

methods are presented, let us first concentrate on the types of problems to introduce some order in our further considerations. The main types of tasks that can be solved using *data mining* are as follows:

- *Defining functional relationships*. That is a frequent task that can be solved using *data mining*. The main problem in this case is that we have data describing facts while we need to find the correlations between these facts. This class of problems includes two groups:
  - Attempts to define the *correlations between the values of certain variables (regression problems)*. There are several methods for dealing with such tasks. While choosing the appropriate tool, one ought to take into account the type of relationship (whether it is linear, non-linear, of known or unknown character), measuring scales used to express the values, the size of data set and availability of software. Created models help to understand the processes and simulations or can be applied in forecasting. While creating a model describing the relationships between variable parameters, a researcher often has to provide some guiding lines and define which variables are explanatory and which are the sought ones.
  - Another group of tasks are *association problems,* which involve investigating the relationships between the facts of occurrence (or non-occurrence) of certain processes. In other words, we examine here co-occurrence of certain phenomena. An association problem is an example of a regression problem where the analysed variables are binary. Because of wide practical applications and specificity of research methods, this group should be treated independently.
- *Pattern classification* - which analyses objects characterised by parameters from the predetermined set of variables. The main aim is to assign the objects (on the basis of variables that characterise them) to predetermined classes (as we have here class patterns, this kind of classification is called „pattern classification"). In literature on the subject we can also find the terms „discrimination problem" or „pattern recognition". In this aspect „pattern" ought to the treated as a synonym of „image", hence pattern recognition methods can be applied to recognise and classify any data.
- *Non-pattern classification*. In case of such tasks, one does not know the pattern of classes into which the data set ought to be split, moreover there is no information how many classes one might get. Therefore, the aim of the data analysis is to recognise the structure of the set (data clusters revealing similarities and differences from other clusters), to identify the number and characteristic features of occurring classes and to assign all or nearly all objects to relevant clusters. It is often possible to investigate interrelations between the clusters and to draw certain conclusions.
- *Analysis of time series* - that is the analysis involving time aspects. Data stored in database or data banks are always somehow related with time since the given fact recorded at the given date took place at a definite moment which can be directly (i.e. by recording the moment the date is fed in the computer) or indirectly (from the description of the recorded data) determined and included in the database. However, time need not always be an important factor since many correlations sought with *data mining* methods should, by definition, be universal and independent of time. However, there are some problems and sets of data where time aspects are of major importance. The researchers then attempt to identify and describe the regularities displayed by variables obtained in various periods of time, and *data mining* techniques can be used as support. In the simplest case of time series analysis

Oczywiście, problemy asocjacyjne stanowią szczególny 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*, przy czym słowo „obraz" w tym kontekście rozumiane jest właśnie jako synonim słowa „wzorzec", 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 posiadany zbiór danych, co więcej - zwykle brakuje nawet informacji o tym, jak duża jest przewidywana liczba klas. 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. 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 poszukiwana 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 zawe examine a single variable and seek the relationship between its value at one instant and the preceding values (the analysis of one-dimensional series). This scheme may be extended to a greater number of variables while the values of one variable depend on the previous values of this and other variables. The aims of the analyses of time-ordered data are diverse: sometimes one seeks a function which would precisely define the relationship between subsequent operations, sometimes one means to identify the relationship between the fact of one process occurrence and occurrences of other processes in distant or recent past. This task, also termed „process sequence analysis" plays a major role in supporting the decision-making (particularly in research); furthermore, it can be well applied to material and biomaterial engineering [Tadeusiewicz, 2000]. The main objective of the time series analysis is application of created models to forecast further processes.

- *Problems of choice*. Such problems appear when one has to choose the best element (in terms of the accepted method of evaluation) from the available set of elements or variables. Algorithms implementing these tasks using *data mining* are useful when it is impossible to check and evaluate **all** subsets of the analysed set. It is so when the number of subsets is so great that the time spent on calculations to search all possibilities would be much too long and therefore unacceptable to the researcher. We have to emphasise that *data mining* techniques used in tasks involving choice mostly lead to quasi-optimal solutions, which means that the solution is sufficiently good though it need not be the best one. That is why *data mining* techniques ought to be used only to deal with most complex problems where more accurate methods of optimisation cannot be applied, otherwise the best solution would be to *abandon data mining* and evaluate all the subsets. Problems involving choice belong to a wider group of optimisation problems, where *data mining* techniques are growing in importance.

Selection of the suitable *data mining* technique for the given task

Prior to analyses with the use of *data mining* techniques, one has to identify the problem in terms of task types and categories, as explained in the previous section. Recognition of the problem is followed by selection of a suitable data analysis method. Methods for solving typical problems are summarised in TABLE 1.

On identifying the problem one can choose the methods for problem solving. The choice of the *data mining* method involve:

- a priori knowledge of the investigated phenomenon (general laws governing these processes, linearity or nonlinearity of relationships, class boundaries, time series structure)
- size of the data set (some methods require large numbers of elements while in others small numbers of data are prevalent)
- the manner of utilising results (depending on actual circumstances, the researcher might prefer the methods for simulation of process operation, black box methods, methods providing results in graphic form or in the shape of decision rules
- availability of software

If possible, it is recommended to apply more than one *data mining* method to solve a given problem. Such approach should be adopted for several reasons:

- Obtaining similar solutions using different *data mining* al-

istnieniem pewnych zjawisk w dalszej bądź bliższej przeszłości. Takie zadanie, nazywane badaniem sekwencji zjawisk, ma szczególne znaczenie w zagadnieniach wspomagania procesów podejmowania decyzji (zwłaszcza w badaniach naukowych), ale ma także bezspornie zastosowanie do szczegółowych zagadnień technologii materiałowych, w tym także biomateriałów [Tadeusiewicz, 2000]. 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-optymalnych, co oznacza, że rozwiązanie znalezione jest wystarczająco dobre, ale nie ma gwarancji, że jest najlepsze możliwe. 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 typów 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 charakter budowy *czarnej skrzynki*, albo metody dostarczające rezultatów w formie graficznej bądź metody dostarczające reguł decyzyjnych);
- dostępność oprogramowania.

gorithms may be treated as confirmation of formulated conclusions
- Results obtained in different methods may highlight several aspects of the phenomena, thus enriching our knowledge in many ways
- Data analysis methods offer different forms of data representation, interpretation and utilisation. Depending on the actual circumstances, one can choose the form of representation most appropriate in the given context.

Methods of exploratory data analysis

Data mining analyses involve several stages, the fundamental components being as follows [Heidsieck et al 2000]:

- defining the aim of the analysis and the type of task
- creating the set of data
- initial analysis and data pre-processing
- computations

Rodzaj problemu *Problem*	Metody *Methods*
Opis zależności *Description of relationships*	• Statystyczne metody pomiaru zależności *statistical methods of measuring relationships* • Sieci neuronowe typu MLP lub RBF *neural networks MLP or RBF* • Metody analizy współwystępowania *methods of co-occurrence analysis* • Zbiory przybliżone *approximated sets*
Klasyfikacja wzorcowa *Pattern classification*	• funkcje dyskryminacyjne *discriminant functions* • sieci neuronowe typu MLP *neural networks MLP* • drzewa decyzyjne *decision trees* • systemy regułowe *rule systems* • zbiory przybliżone *approximated sets* • metoda k-najbliższych sąsiadów *k-nearest neighbours*
Klasyfikacja bezwzorcowa *Non-pattern classification*	• metody taksonomiczne *taxonomic methods* • sieci neuronowe samouczące się *self-learning neural networks* • metody redukcji wymiaru przestrzeni danych *reduced dimensional space* • metody graficzne *graphical methods* • algorytmy genetyczne *genetic algorithms*
Analiza szeregów czasowych *Time series analysis*	• sieci neuronowe typu MLP lub RBF *neural networks MLP or RBF* • metody analizy sygnałów *methods of signal analysis* • metody badania sekwencji *methods of sequence analysis*
Problemy wyboru *Choice*	• algorytmy genetyczne *genetic algorithms* • sieci neuronowe typu Hopfielda *Hopfield neural networks* • zbiory przybliżone *approximated sets*

TABELA 1. Rodzaje problemów i właściwe dla nich metody *data mining*.

TABLE 1. Types of problems and appropriate *data mining* methods.

- verification of results
- interpretation of results and application to decision- making

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 kilkuetapowym. Do zasadniczych jego elementów należy zaliczyć ([Heidsieck i in., 2000]):

- 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**. W rozważanych tu zastosowaniach metodyki *data mining* celem tym jest najczęściej potrzeba dowiedzenia się czegoś nowego o wybranych biomateriałach w celu uzyskania wzrostu efektywności ich produkcji lub zastosowania. 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 po-

In many cases subsequent stages are repeated several times to get the required results, one or more preceding steps can be iterated when necessary.

Defining the aim of the analysis and the type of task

In order to analyse a set of data using *data mining* techniques, one has to precisely define the aim of the analysis first. The aim of the analysis presented here is to learn more about certain biomaterials in order to improve the efficiency of their manufacturing or applications. That is why application of *data mining* in such field must be preceded by formulation of the decision problem relating to research activities such that the results of *data mining* analysis should land us as close to our goal as possible.

Defining the aim of the analysis determines the whole further work, justifies the costs, and allows for evaluation of performance (success or failure).

The aim being precisely defined, one has to determine the type of problem (or problems) to be tackled since in many cases the problems tend to be complex and involve many aspects. That helps to choose the most appropriate *data mining* technique, which will be use in problem solving.

Creating the set of data

Once the aim of the analysis is defined, one makes decisions as to sets of data, which is the basis for further work. Several important decisions have to be made at that stage. The first one concerns the sources of data. The most natural source of such data is well -supplied and supervised database of laboratory test results. One can also use the information from spreadsheets or from plane- structure files, such as data recorded by computerised laboratory equipment. The researchers planning to utilise the methods of artificial intelligence to obtain new information hidden in the „old" data may resort to their own data (from earlier tests) or may also include the results published by other authors (provided it is expressly indicated in the future work that such data are included). Such procedure is now becoming easier, thanks to computerisation of information systems (computerised libraries, the Internet as a source of information, conference materials available in electronic form). Digital-access data sources developed by companies are becoming increasingly popular and allow for performing a wider range of *data mining* analyses of biomaterial properties and applications, they also grow in importance in the context of newly developed technologies of biomaterial engineering.

The researchers will usually utilise only a part of available information resources. It is extremely difficult to define which part is that or to give any guidelines because the criterion of initial data selection is determined by the accepted aim. The fact that no guidelines are available should not discourage the researchers from data selection; it is just the opposite - this task must not be left to a thoughtless machine but requires the intelligent intervention of the researchers.

One technical question has to be answered now: shall we use original data for calculations or should we resort to copies? Making a copy is usually the better solution - it eliminates the risk of unintentional damage of valuable source information (while *data mining* algorithms are in operation) and facilitates safe transformation of data (such as standardisation), when required by *data mining* procedure. A major drawback of using a copy is that one needs extra disc space to store the copied data to be used for *data mining*.

dejmowanych 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 - na przykład danych rejestrowanych przez skomputeryzowaną aparaturę laboratoryjną. Badacz, który decyduje się na użycie technik sztucznej inteligencji dla pozyskania nowej informacji „ukrytej" w starych danych może bazować na własnych danych (pochodzących z wcześniej prowadzonych doświadczeń), ale może także wciągnąć do bazy danych opublikowane wyniki innych autorów - oczywiście wskazując w następnych pracach, że korzystał także i z takich właśnie danych. Postępowanie takie staje się technicznie coraz łatwiejsze, głównie z uwagi na powszechna komputeryzację systemów informacyjnych (skomputeryzowane biblioteki, Internet jako źródło informacji, materiały z konferencji dostępne w formie elektronicznej itp.). Również coraz liczniejsze cyfrowo dostępne firmowe źródła danych mogą stanowić podstawę do przeprowadzenia coraz szerszego zakresu analiz typu *data mining* zarówno w zakresie właściwości i zastosowań biomateriałów, jak i w kontekście nowo tworzonych technologii ich wytwarzania.

Zwykle do opisywanych tu badań przydatna będzie tylko pewna **część** zasobów informacyjnych, do jakich badacz ma dostęp. 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 - eliminuje bowiem niebezpieczeństwo przypadkowego uszkodzenia (podczas działania algorytmów *data mining*) cennych dla laboratorium informacji źródłowych i stwarza 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 badań, ale zbytecznych w kontekście celów *data mining*. Na przykład w każdym laboratorium prowadzi się ewidencję badań, rejestrując ich datę, numer próbki, nazwisko osoby prowadzącej pomiary itp., natomiast w procesie analiz *data mining* niepotrzebne są wszelkie te dane identyfikacyjne i porządkowe, ponieważ w celu wykrywania nowych współzależności i nowych związków - wystarczą same tylko merytoryczne wyniki badań. 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 wy-

While making copies to be used in *data mining*, one can resort to a simpler solution and copy only the values of predetermined variables, which are important for the given problem; while the multiple identification data, required for the analyses yet unnecessary in the context of *data mining*, can be omitted. For example, each laboratory has the record of tests, specifying the date, sample number, name of the person responsible for measurements and the like. In *data mining* analyses all these identification and ordering data are redundant because the merits of the test results are sufficient to establish new correlations and relationships. We also have to consider the possibility of filing the copies of *all* the values from the data base (elements, rows) or only some of these (it would be necessary to define the scope of the analysis). The basic consideration in this case is how the number of elements should affect the time spent on calculations. It is usually so that the greater the number of data, the better results, though the time spent on calculations is longer. That is why the acceptable size of the working copy is the compromise between the anticipated quality of test results and the time required for calculations.

When we decide to utilise only a fraction of cases , then we are faced with a problem how to choose the elements to be considered in calculations - and the criterion for rejecting others. A random method of selection is very often applied. Its main advantage is that it causes the least distortions of regularities existing in the source data set, yet in some cases other solutions are preferable. While discussing the selection of variables and elements from the database, we must not forget the fundamental tool, that is the SQL (structural question language) which allows for using the data stored in relational databases. The structure and properties of SQL might encourage the use of some and disregard other methods of data and element selection from the database. We have to bear that in mind while planning the test methodology so as to avoid the situation when, in an attempt to avoid painstaking data processing at the stage of data analysis, we plan none the less laborious method of selecting a limited data subspace.

Initial analysis and data pre-processing

The subsequent stage involves the initial analysis and pre-processing of selected data. In many cases the data from the database have to be verified or/and processed to a form suitable for further analysis before the *data mining* algorithms can be applied. There are several reasons why this stage is necessary. First of all, there are shortages of data as certain information may not be available for a time or it may not be included in available records, certain data may not be fed in the information system at some stage of system operations, or for many other reasons. Missing data make the test procedure more difficult. When the amount of available information is smaller, the developed model might be inferior in quality and drawn conclusions might not be adequately justified. The absence of certain data makes it impossible to complete certain numerical procedures required by *data mining*. When some information is missing, one has to make the decision how to deal with the problem. These are the most frequently applied solutions:

- One attempts to fill in the missing values basing on alternative sources of information (for example records in source documents). That is far and away the best method to deal with the problem of missing data, however it is not always practicable as alternative sources may not exist
- Missing information is assessed. That usually involves building a statistical model in order to determine the missing values. This approach allows for doing further calcula-

korzystanie 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 nie wprowadzenia 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 modelu, którego celem jest wyznaczenie wartości brakujących; taki sposób postępowania umożliwia wykonanie dalszych prac obliczeniowych, jednakże oszacowania brakujących danych prawie nigdy nie posiadają 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

tions, however assessed values never have the same status as the real, properly collected data

- The cases (rows) with missing information are removed from the database. One has to bear in mind, however, that in this way we get rid of essential numerical problems caused by missing data, yet at the same time we irretrievably lose a portion of existing information, which would negatively impact on quality of the developed model
- Variables (columns) with missing information are deleted from the data set. This often leads to major reduction of available information resources and is recommended in exceptional situations only

The choice between the third and the fourth option depends on missing data configuration within the set. When the missing data appear among the values of one variable, the reasonable solution would be to remove that variable even though all the information it carries gets lost. When some values of several variables are missing, the better solution would be to omit the relevant cases in further calculations.

Another problem that has to be considered at the stage of initial data analysis are atypical values, that is when the data set includes the values obtained from measurements or observations, yet those values significantly differ from the typical ones. In such circumstances the researcher has to answer the important questions:

- Do those atypical values reflect the real state of affairs (i.e. they are anomalies) or are they the result of errors (at the stage of measurements, data recording or feeding):
- What shall we do about those atypical values - shall they remain in the data set or perhaps we should remove them and proceed as in the case of missing data

The answers to these questions are difficult, each research problem requires and individual treatment. One has to work thoroughly to find these answers as they may significantly impact on the results of the whole research work.

Data operationisation

Data pre-processing means first of all data operationisation. It consists in transforming the analysed variables by means of selected mathematical formulas. The most popular operationisation techniques are: scaling of variables, raising to a power (with various power exponents), finding the logarithms, the inverse and absolute values, binary coding (for example taking into account only the information about the sign of the values and replacing the negative values with „-1" and positive values with „+1"). This class also covers time series filtering to separate the changes of predetermined character).

Data operationisation may also include weighing (using appropriate weighing factors), normalisation and standardisation of variables. The way the transformations are carried out depends on several factors: one may want to emphasise the importance of a given variable (through weighing), to take into account most interesting information (the sign of the numbers - data filtering) or to change the existing nonlinear relationships into linear ones (finding the logarithms). It also depends on the nature of computational algorithms (for example, application of neural networks requires pre-scaling of data so that they would fall in the acceptable intervals).

Data representation

Data pre-processing also involves the adoption of a suitable form of data representation. It usually consists of transforming the values of variable parameters (original values or following operationisation) to get the form suit-

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 wprawdzie 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 je 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. Odpowiedzi te muszą być jednak znalezione z dużą pieczołowitością, gdyż mają istotny wpływ na rezultaty 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 (z różnymi wyładnikami), 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 podyktowana chęcią nadania szczególnego znaczenia pewnej zmiennej (poprzez ważenie), uwzględnienie szczególnie interesującej informacji (np. znak liczby, filtracja danych), zmiany charakteru istniejących zależności nieliniowych na liniowe (np. poprzez logarytmowanie), czy tez może wynikać z natury stosowanych algorytmów obliczeniowych (np. stosowanie sieci neuronowych wymusza wcześniejsze skalowanie danych do przedziału wartości akceptowanych w tych sieciach).

able for further processing with the use of selected tools. The way this operation is carried out is closely related with the measuring scale used to express the values and the planned *data mining* techniques. As far as the measuring scales are concerned, one is faced here with a very common problem: how to include the qualitative information in calculations (i.e. information expressed in terms of nominal or ordered measuring scale). This problem is of primary importance since the portion of qualitative information in data processed with *data mining* techniques is usually rather large, while most *data mining* techniques are adapted to numerical values. Sometimes one is faced with the inverse problem: values expressed in strong measuring scales have to be transformed into weak-scale values (this operation is common while applying decision trees).

As it was already mentioned before, the choice of the data analysis method determines the way the data is represented. Actually, each *data mining* technique favours the definite type of input data and all discrepancies ought to be resolved through changing the data representation system. All these transformations ought to be done carefully by knowledgeable and experienced researchers, since not every change of the data representation system is justified while the errors committed at that stage may impact on the results obtained in subsequent stages, thus reducing work effectiveness (genetic algorithms are good examples - all calculations are run on binary, coded values of considered variables).

Forms of data representation

Having completed the first yet important and indispensable stages of the initial analysis, we proceed to **implementation of selected *data mining* algorithms.** There are quite a number of them, so let us focus on the group of most popular ones:
- statistical methods
- neural networks
- genetic algorithms
- classification trees
- nearest neighbours methods (searching for analogies)
- approximated sets
- association and sequence analysis

Since this paper is limited in scope, it is not possible to give even a brief description of those most popular data exploration techniques. The Reader will find the relevant information in bibliography, provided at the end of the paper. We have to remind just the fundamental paradigm: *„The basic aim of the computations is to enhance the knowledge by adding new information from the available data sets"*. No matter what methods are used, thus obtained results are information, (automatically!) deduced from data which describe the existing regularities, represent the results of classification, approximate the set structure and suggest how the choice should be made.

It is quite obvious that the results of computation should always give an answer to the question formulated as the specified aim of the research analysis. We have to bear in mind, however, that depending on the applied method, the *form* of this response may vary (even for the same research problem). Applying different methods, we receive similar or even the same information though in different form, thus the degree of their applicability to analyses, forecasting and decision-making support may vary, too. The forms of data representation should be also tailored to individual user's needs; the choice of particular form depends rather on the user's than on particular characteristics of the information itself. A beginner researcher (for example a PhD candidate), who deals with the problem for the first time, will usually prefer different forms of representation than an experi-

Reprezentacja danych

Wstępne przetworzenie danych obejmuje również przyjęcie właściwego sposobu *reprezentacji danych*. Ta operacja polega zwykle ze 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 (dobry 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***. Jest ich sporo, ale zestaw najczęściej opisywanych metod obejmuje:

- metody statystyczne;
- sieci neuronowe;
- algorytmy genetyczne;
- drzewa klasyfikacyjne;
- metody najbliższych sąsiadów (wnioskowanie przez analogię);
- zbiory przybliżone;
- metody badania asocjacji i sekwencji.

Niestety ograniczona objętość tego referatu nie pozwala na to, by dać chociażby skrótowy opis tych najpopularniejszych metod służących eksploracji danych, po bliższe informacje odsyłamy więc Czytelnika do informacji zawartych w bibliografii zestawionej na końcu pracy. Należy tylko przypomnieć zasadniczy paradygmat: Podstawowym celem prowadzonych obliczeń jest *wzbogacenie wiedzy o nowe informacje pozyskane z dostępnego zbioru danych*. Niezależnie więc od użytej metody uzyskane rezultaty obliczeń prezentują informacje, wydedukowane (automatycznie!) na podstawie danych, które opisują istniejące prawidłowości, prezentują wyniki klasyfikacji, przybliżają strukturę zbioro-

enced analyst engaged in biomaterial studies and engineering for years; still another forms will be chosen by physicians or veterinary doctors who deal with practical applications of biomaterials, and also by manufacturers. Most popular forms of representation of information obtained from data analysis are:

- *graphic* representations - provide pictorial representation of several types of problems, yet sometimes they may lack precision. Information in graphic form is easy to interpret by man though not interpretable for machines; hence that cannot be the only output in systems where the results have to be further processed by subsequent systems supporting the decision-making. Diagrams and graphs generated by *data mining* systems characterise the tested items (in form of histograms, for example), highlight the existing correlations and set structure (perception maps, tree diagrams) and suggest the course of further research or even support the decision-making process (decision trees, block diagrams).

  Well developed module generating graphic representations of analysed problems and solutions is very useful in all *data mining* applications, yet it tends to be time-consuming and that is why it is often abandoned.
- *descriptive statistics* - characterise the analysed aspects of reality in terms of selected statistical indices, they are usually easy to obtain and easy to interpret for experienced researchers and as such allow for quick and accurate evaluation of the analysed problem.
- *decision rules* - present the information in the form of statements: „If....., then.....". This form of data representation is also convenient, provided the number of rules in not too great, moreover it can be directly used by the system supporting the decision- making (for example, the rules can be fed into the database of an expert system).
- *equations or systems of equations* - such representation of existing rules is convenient for those who investigate the behaviour of the fragment of reality using mathematical tools. This representation allows for drawing very general and far-reaching conclusions relating to fundamental features of the investigated problem; furthermore, it helps in simulations, in forecasting and offers certain insight into the nature of the problem. Representation in the form of equations is very compact, may be easily utilised both by man and computer; it can be easily changed into the graphic form. However, it is extremely difficult to obtain information from data in the form of equations or the process may involve a series of subjective errors as one has to make certain arbitrary assumptions to easily identify the problem expressed in terms of mathematical equations.
- *neural networks* -represent the information as systems of parameters (usually weight and threshold values) of concurrent processing elements (so called artificial neurones). Application of neural networks always provides answers to given questions at the output, in other words - the solution supplied by the network will always be the model of the problem, not the explanation. Thanks to flexibility of the neurone model the network may determine functional relationships, decision rules and set structure. It can be easily transformed into a computer program which models the problem though the rules of network operations cannot be easily transformed into simple and interpretable decision rules.
- *graphs* - give a pictorial representation of relationships between the analysed items, in a certain sense graphs have all the advantages of graphic representation and decision rules.
- computer programs - this form is never generated directly by any *data mining* techniques used in data analysis, yet all other forms of data presentation are transformed into

wości, sugerują sposób dokonania wyboru.

Jest rzeczą oczywistą, że rezultaty obliczeń mają zawsze 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 technika *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 początkującego badacza (na przykład doktoranta), który po raz pierwszy zajmującą się analizowanym problemem inna przez doświadczonego analityka od lat wytwarzającego i badającego biomateriały, inna przez lekarza lub weterynarza, który je stosuje, a jeszcze inna przez producenta. Do najczęściej spotykanych form prezentacji informacji pozyskanych z danych należy zaliczyć:

- *formę graficzną* - pozwala na poglądowe 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 dalszych badań albo wręcz wspomagać proces 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 wnioskowanie 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

computer program whenever they are to be used by computers, for example to provide on-line support to the decision - making.

Results obtained from *data mining* analyses always generate a *model* which provides a more or less formal description of a fragment of reality. The concept of „model" is very broad here as it covers the mathematical equation (or system of equations), the set of decision rules, diagrams, schemes, decision trees, graphs or neural networks. A computer program implementing the uncovered rules is a model, too.

Model verification

Computations required in the selected *data mining* methods and obtaining the results in the form of a model (as explained in the previous section) does not mark the end of the research work since the results have to be verified so as to answer the following questions:

- does the obtained model function properly for its underlying data
- can the developed model be expected to function properly for other data than those used in model development

It is relatively easy to answer the first question- it is sufficient to observe how the model functions for the whole available set of data and compare thus obtained results with information at our disposal. The second question is much more difficult. How shall we assess the adequacy of the model for data not available during the tests?

Despite those difficulties, the second question should not be left without an answer. The knowledge how the model will function for new data, inaccessible during the tests, is of major importance as it shows the adequacy of thus obtained solutions for future applications. The ability of the model to function correctly also for new data (the data not used for model development) is often called the *capability of generalisation*. A model having this property can be applied to forecasting, decision- making support or to classification of previously unknown objects. There are several methods for assessing the model's capability of generalisation. The most popular methods are:

- *evaluation of generalisation capability by means of a testing set.* The underlying principle is very simple - prior to computations the data set is divided in two parts, called the *learning set* and the *testing set*. Elements of the testing set are used in model development. The elements included in the testing set *are not utilised* in any way till the model is complete. After computations required to build the model, it is verified by means of the testing set. It is assumed that the model will function for the testing set data in the same way as for any new data (unknown when the model was developed). Model behaviour for the testing data reflects its capability of generalisation (or its absence). This assumption is based on the fact that testing elements were not used to create the model. This hypothesis has one major drawback, namely the testing set are usually small in size (the majority of data is involved in model learning processes), hence the sample is barely representative for potential infinitely large sets of practical problems to be solved by means of that model in its normal operation.

  Applying the presented evaluation method, one has to make two very important decisions:

  a) what should be the size of the learning set and of the testing set

  b) how to divide the set into the learning and testing sections

  It is difficult to give straightforward answers to these two questions. As far as the answer to the first question is concerned, it seems that a majority of elements should be

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). 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*. Posiadanie tego typu właściwości pozwala na wykorzystanie modelu do prognozowania, wspomagania procesów decyzyjnych lub też klasyfikowa-

included in the learning set while the remaining ones should go to the testing set. This „majority" usually covers 60-80 %. As far as the other question is concerned, the division should be designed in such a way that both sets be representative. In practice, the elements are assigned to these two sets randomly (utmost care should be taken to retain the predetermined proportions between the size of these two sets). Different methods are applied in exceptional cases only, when the ransom selection procedure does not ensure the required sample representativeness .

This methods for evaluation of model quality has one major drawback - while separating the testing set we reduce the amount of information to be used at the stage of model development, which may adversely impact on quality of thus obtained solutions. It is especially troublesome when the original data set has a small number of elements. When the sets are small in size, we might be faced with still another difficulty - the evaluation of the generalisation capability using the testing set need not reflect the real capability. Two techniques presented below are an attempt to overcome this problem.

- *Cross testing*. This evaluation method is an extension of the testing set method presented above. The procedure is as follows: the available data set is divided into n parts (the elements are usually assigned to the subsets at random), *n-1* parts are used as the learning set while the remaining *n*-th part acts as the testing set. The procedure is iterated *n* times, while each time different subset should be used as the testing set. Thus proceeding, we get n model evaluations for different testing sets, covering the whole available information. Measures obtained from cross testing are often aggregated to get a single indicator (for instance by averaging). A definite advantage of such evaluation procedure is that all available data elements are used both in model building and evaluation (any element appears only once in subsequent iterations - as the element of either the learning or the testing set). Cross testing is recommended when the data sets are relatively small in size. The main disadvantage is that the time spent on computations becomes nearly n times longer.
- *Bootstrap methods*. We have to bear in mind that the bootstrap method has wider applications than evaluation of model quality. Generally speaking, it is an advanced simulation method which assesses the parameter values and statistical distributions on the basis of an available data set. In our applications these statistics will determine model quality. Similar procedure can be applied to determine the values and distributions of other statistical parameters.

The starting point is the data set with n elements. Computations may be iterative (the number of iterations is rather large, usually more 1000), and involve the following stages:

- the bootstrap set is formed from the original data set by return sampling (elements from the original set are moved to the bootstrap set and returned (because of return sampling, certain elements in the bootstrap set may appear several times while other may be absent). In each iteration the sampling procedure is repeated , so each time the composition of the bootstrap set is different.
- computations required by the *data mining* method are performed using the bootstrap set.

In this case when the aim of these computations is to evaluate the generalisation capability of a model developed using the *data mining* technique, each iteration will involve the following operations: the bootstrap set is divided into the learning and testing sets, the learning set is used to build the model and a measure of model quality will be developed using the testing set.

When the computations are complete, we have at our disposal certain model quality measures obtained in sub-

nia nieznanych 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 posiadany 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 model. 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 posiadaną przez niego (lub nie posiadaną) zdolność do generalizacji. 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 nieskończenie licznego zbioru praktycznych zagadnień, które mają być rozwiązywane z 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 posiadanego zbioru na część uczącą i testową.

  Na tak postawione problemy trudno udzielić jednoznacznych odpowiedzi. Próbując odpowiedzieć na pierwsze pytanie można stwierdzić, że większość posiadanych 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 miał 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,

sequent iterations (usually we get more than 1000 values). This series might be averaged (the average is the assessment of the actual value of this indicator) or it may be used to assess the distributions of the relevant parameter (such as variance, which allows for assessing the reliability of the given quality measure evaluation).

Because of the large number of required iterations, the bootstrap technique is computationally very expensive (in practice calculations cannot be done without computer techniques), nevertheless the results fully justify the costs.

Methods of model quality measurements

Methods for verification of models obtained from analyses of large data sets using *data mining* techniques were presented in the previous paragraphs. Presentation of these techniques, however, did not provide the methods of model quality measurements. This issue is very important and will be addressed in this subsection.

We have to emphasise, that the way model quality is evaluated depends on several factors. First of all, it depends on the type of analysed problem and this particular aspect will be addressed.

- In the case of models, which are used to *describe functional relationships*, the developed quality indicators can be categorised into two groups: absolute and relative measures.

  Absolute measures take into account the aggregated differences between the real values (observed and included in the data set) and the theoretical ones (those calculated on the basis of the model). In case of numerical variables, the most popular aggregation technique is calculation of the sum of the square of the difference between those values - SSE (sum of square error).

$$SSE = \sum_{i=1}^{n} (y_i - d_i)^2$$

  where y_i stands for the i-th theoretical value, d_i - i-th real value, n- the number of elements in the set.

  Thus obtained value is then divided by n (the number of cases) to get the error mean- square. Next the root of the error mean- square is derived and thus obtained value will indicate how much the theoretical values differ from the real ones (on the average).

  Instead of SSE, one may also use its modulus (absolute value) while evaluating model quality, hence the atypical cases (where errors are significant) will impact on the indicator value in a lesser extent. The indicators presented constitute only a small fraction of quality measures used in practice (many other measures can be found in literature on the subject).

  Apart from absolute measures, there are also *relative measures* based on comparison between the absolute measures for the assessed model and analogous measures for other models, which act as reference points. Relative measures are used to compare the quality of several models. In case of models describing the relationships between variables, the linear regression function is often used as the reference point.
- while developing models that solve *pattern classification problems* within the framework of *data mining*, one may accept a similar division of measures: absolute measures would compare how the given object is classified in theory and in reality, while relative measures would compare the various classification methods in terms of their quality. The starting point for determining the values of absolute measures (in pattern classification) are two parameters: the number (or percentage) of *correctly classified* objects and the number (or percentage) of *incorrectly classified* ones.

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, ze 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 umożliwia wnioskowanie na temat wiarygodności oszacowania interesującego miernika).

Z uwagi na wymaganą duża 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.

These two basic values can be further analysed, for instance they can be split for the groups of objects. Furthermore, one can also analyse which types of errors are most frequent.

- evaluation of models used for solving the *non-pattern classification* problems is much more difficult. The main difficulty involved in model quality evaluation is that one has to evaluate the adequacy of the set structure discovered with the help of the model when no information about the *real* correlations between the objects or groups of objects is available. The underlying principle is as follows: such grouping of objects is preferred that the differences between the objects in the same group should be minimised while the differences between those in various groups be maximised. This general statement is reflected by a vast number of available measures described in literature on the subject.
- evaluation of models used for *time series analysis* is similar to evaluation of regression models (used to describe functional relationships) or classification models (used in pattern classification). The degree of analogy between the regression model or classifications depends on whether the analysed variable quantity in the series is qualitative or quantitative in character. In those cases similar measures can be used as in evaluation of regression or classification model, though in certain cases other measures may prove more useful, as they are better adapted to the specificity of time series.

  Among absolute measures, we often resort to the one comparing the real and forecasted directions of changes of the analysed variable, while among the relative measures it is worthwhile to mention the one comparing the quality of the developed model with that of so called „naive model" (i.e. the one assuming that at the instant $t+1$ the forecasted value of the variable parameter will remain on exactly the same level as at the instant t).

  Evaluation of regressive models of time series may also use the obtained series of residues to check whether autocorrelation is present, these series can be also analysed using spectral analysis techniques.
- Evaluation of adequacy of *data mining* results calls for a short explanation, too. *Data mining* techniques provide the assumptions for rational choice of the optimal (or sub-optimal) options - for example the optimal technology. Evaluating the quality of such algorithms involves comparison of results corresponding to the examined set of selected elements with the results obtained in alternative selection methods. This general statement can be made more precise, depending on the type of analysed problem. The applied evaluation procedure is often supported by one more element - i.e. information about the preferred number of elements. In practical applications smaller sets are preferred (while choosing the explanatory variables one often tends to minimise their number); accordingly the *penalty component* is added to the indicator definition. This component brings down the evaluation mark when the difference between the expected and real number of selected elements gets higher.

Presented methods of evaluation of models obtained as the result of data exploration are general and depend exclusively on the type of analysed problems. We have to bear in mind, however, that the problems are mostly concerned with research; accordingly evaluation of solutions ought to apply the measures that would enable us to assess not only economic effects the solutions might bring, but also their value for research. Selection of the evaluation method , however, depends chiefly on the type of analysed problem and cannot be discussed here in more general terms.

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 wchodzą w skład zbioru danych) z *wartościami teoretycznymi* (czyli tymi, które zostały obliczone na podstawie modelu). W przypadku korzystania zmiennych o charakterze numerycznym takim najpopularniejszym sposobem agregacji jest obliczenie sumy kwadratów różnic pomiędzy wspomnianymi wartościami (czyli miary błędu zwanej SSE)

$$SSE = \sum_{i=1}^{n} (y_i - d_i)^2$$

  gdzie y_i reprezentuje i-tą wartość teoretyczną, d_i jest i-tą wartością rzeczywistą, zaś n jest liczbą elementów w zbiorze).

  Otrzymaną w ten sposób wartość możemy podzielić przez n (liczbę przypadków) otrzymując *błąd średniokwadratowy*. Z kolei z błędu średniokwadratowego możemy wyznaczyć pierwiastek - otrzymana wartość będzie nas informować, o ile (średnio) różnią się wartości teoretyczne od rzeczywistych.

  Czasami, określając jakość modelu, zamiast kwadratów różnic uwzględniamy ich moduły - w ten sposób przypadki nietypowe (dla których błędy są zwykle bardzo wysokie) będą w mniejszym stopniu wpływać na wartość miernika. Zaprezentowane mierniki stanowią tylko bardzo niewielki odsetek miar stosowanych w praktyce (w literaturze można znaleźć szereg innych propozycji).

  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;

Interpretation of results and application of results to decision-making processes

Positive verification of the created model justifies its **application to real-life situations**. The scope of applications is practically unlimited and covers:

- studies of isolated fragment of reality - in this case the model should be treated as the description of investigated phenomena
- forecasting - model may help to forecast how the given process will develop. It is worthwhile to mention that the term „forecasting" is given a very broad meaning. In time series analyses it means the determination of future values in the series, while in the case of cross-sectional data (relating to various objects, the time factor being neglected) forecasting might mean estimation of certain characteristics for new objects, unknown at the time when the model was created.
- simulation - in these applications the model is utilised to run substitute experiments in order to verify certain activities before they are commenced in real life. Using *data mining* results provides the answers to the question: *„What will happen if..?"* This group of applications is most useful in decision-making support as the model allows for checking the most likely outcomes of the decisions to be taken.
- development of expert systems- knowledge acquired from data analysis may create (or supplement) the base of the expert systems, i.e. programs able to answer the users' questions by way of automatic reasoning.
- development of decision - making systems - this application of *data mining* is brought down to an attempt to replace the decision- makers (i.e. humans) with artificial decision- making systems. Such attempts are made only in those areas where the time for decision- making is very short (for example in real-time process control) while it is necessary to analyse huge amounts of information.

Concluding remarks

Implementation of the model developed using *data mining* techniques requires further monitoring while in operation. Changes in reality may cause the model to become outdated or even useless after a while. If the discrepancies between the process and the model description get more frequent, the model ought to be updated. In some cases updating would involve the repetition of the whole procedure, in others- it would be sufficient to update model parameters only.

We have to emphasise that the proper use of these techniques brings significant benefits while indiscriminate usage of *data mining* techniques may bring losses and the responsibility for the decisions will always rest with the man, not the computer...

Pismiennictwo References

[Azoff, 1994] Azoff E. M., Neural Network Time Series Forecasting of Financial Markets, John Wiley & Sons, 1994

[Baestaens i in., 1994] Baestaens D. E., van den Bergh W. M., Wood D., Neural Network Solutions for Trading in Financial Markets, Pitman Publishing, London, 1994

[Bauer, 1994] Bauer R. J., Genetic Algorithms and Investment Strategies, John Wiley & Sons, Inc., 1994

[Bazarnik i in., 1992] Bazarnik J., Grabiński T., Kąciak E., Mynarski S., Sagan A., Badania marketingowe. Metody i oprogramowanie komputerowe, Canadian Consortium of Management Schools i Akademia Ekonomiczna w Krakowie, Kraków - Warszawa, 1992

[Berry i in., 1997] Berry M. J. A., Linoff G., *Data mining* Techniques

- 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 utrzymywać 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 postępowania - np. optymalnej technologii. 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 zróżnicowania 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 naukowy i, że do oceny uzyskanych rozwiązań należy również stosować (w razie potrzeby) mierniki o takim charakterze, które pozwolą na oszacowanie nie tylko efektu ekonomicznego uzyskanego rozwiązania, ale także jego wartości naukowej. Dokładny sposób oceny jest jed-

For Marketing, Sales, and Customer Support, Wiley Computer Publishing, 1997
[Białasiewicz, 2000] Białasiewicz J. T., Falki i aproksymacje, Wydawnictwa Naukowo-Techniczne, Warszawa, 2000
[Biela, 1992] Biela A., Skalowanie wielowymiarowe jako metoda badań naukowych, Towarzystwo Naukowe KUL, Lublin, 1992
[Deboeck i in., 2000] Deboeck G., Kohonen T. (Eds), Visual Explorations in Finance with Self-Organizing Maps, Springer-Verlag, London, 2000
[Biethahn i in., 1995] Biethahn J., Nissen V. (Eds.), Evolutionary Algorithms in Management Applications, Springer-Verlag, 1995
[Cieślak i in., 1997] Cieślak M. (red.) i inni, Prognozowanie gospodarcze. Metody i zastosowania, Wydawnictwo Naukowe PWN, Warszawa, 1997
[Gatnar, 1997] Gatnar E., *Data mining*: metody i zastosowania, Taksonomia, Sekcja Klasyfikacji i Analizy Danych PTS, Zeszyt 4, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, 1997
[Gatnar, 1998] Gatnar E., Symboliczne metody klasyfikacji danych, Wydawnictwo Naukowe PWN, Warszawa, 1998
[Goldberg, 1995] Goldberg D., Algorytmy genetyczne i ich zastosowania, WNT, Warszawa, 1995
[Grabiński i in., 1989] Grabiński T., Wydymus S., Zeliaś A. (red.), Metody taksonomii numerycznej w modelowaniu zjawisk społeczno - gospodarczych, PWN, Warszawa, 1989
[Grabiński, 1992] Grabiński T., Metody taksonometrii, Akademia Ekonomiczna w Krakowie, Kraków, 1992
[Grabowski, 1997] Grabowski M., Zastosowanie samoorganizujących się map cech Kohonena w analizie danych, Sekcja Klasyfikacji i Analizy Danych PTS, Zeszyt 4, 1997
[Grabowski, 1998] Grabowski M., Sieci neuronowe w analizie danych społeczno-ekonomicznych, maszynopis pracy doktorskiej, Akademia Ekonomiczna w Krakowie, Kraków, 1998
[Gwiazda, 1998] Gwiazda T., Algorytmy genetyczne. Zastosowania w finansach, Wydawnictwo Wyższej Szkoły Przedsiębiorczości i Zarządzania im. L. Koźmińskiego, Warszawa, 1998
[Heidsieck i in., 2000] 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
[Hellwig, 1981] Hellwig Z., Wielowymiarowa analiza porównawcza w badaniach wielocechowych obiektów gospodarczych, w: Metody i modele ekonomiczno-matematyczne w doskonaleniu zarządzania gospodarką socjalistyczną, PWE, Warszawa, 1981
[Hopfer, 1997] Hopfer A. (red.), Wycena nieruchomości i przedsiębiorstw, tom 1, Szacowanie nieruchomości, Twigger S.A., Warszawa, 1997
[Jajuga, 1990] Jajuga K., Statystyczna teoria rozpoznawania obrazów, PWN, Warszawa, 1990
[Jajuga, 1993] Jajuga K., Statystyczna analiza wielowymiarowa, Wydawnictwo Naukowe PWN, Warszawa, 1993
[Kucharczyk, 1982] Kucharczy J., Algorytmy analizy skupień w języku ALGOL 60, Państwowe Wydawnictwo Naukowe, Warszawa, 1982
[Kuratowki, 1975] Kuratowski, Mostowski, Wstęp do teorii mnogości i topologii, PWN, Warszawa, 1975
[Kurzydłowski, 2000] Kurzydłowski A., Zastosowanie drzew klasyfikacyjnych w segmentacji rynku, Sekcja Klasyfikacji i Analizy Danych PTS, Taksonomia. Zeszyt 7, Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław, 2000
[Kwaśnicka i in., 1995], Kwaśnicka H., Markowska - Kaczmar U., Zastosowanie algorytmów genetycznych w problemach optymalizacyjnych, Informatyka, nr 3, 1995
[Lula, 1999] Lula P., Jednokierunkowe sieci neuronowe w modelowaniu zjawisk ekonomicznych, Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków, 1999
[Lula, 1999a] Lula P., Metody eksploracyjnej analizy danych i możliwości ich zastosowań, Strategia Systemów Informacyjnych 1999. Materiały konferencyjne, Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków, 1999
[Lula, 199b] Lula P. Sieci neuronowe. Materiały na seminarium organizowane przez StatSoft Polska Sp. z o.o. 14 października 1999 r w Warszawie, StatSoft Polska, Kraków, 1999
[Martyniak, 2000] Martyniak Z. (red.), Zarządzanie informacją i komunikacją. Zagadnienia wybrane w świetle studiów i badań empirycznych, Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków, 2000
[Mirek, 1999] Mirek J., Skalowanie wielowymiarowe jako metoda segmentacji rynku, [w:] [Mynarski, 1999]
[Morajda, 1997] Morajda J., Algorytmy genetyczne oraz możliwo-

nak bardzo mocno związany z charakterem rozpatrywanego problemu i nie może tu być omówiony w sposób ogólny.

Interpretacja uzyskanych rezultatów i ich wykorzystanie w procesie decyzyjnym

Pozytywna weryfikacja skonstruowanego modelu uzasadnia jego **wykorzystanie w rzeczywistej sytuacji**. Zakres zastosowań jest praktycznie nieograniczony i obejmuje między innymi:

- badanie wyodrębnionego fragmentu rzeczywistości - model należy wtedy traktować jako opis badanych zjawisk;
- prognozowanie - dysponując skonstruowanym modelem można prognozować sposób rozwoju badanego zjawiska. Warto podkreślić, że pojęcie prognozowania rozumiane jest tu bardzo szeroko - w trakcie analizy szeregów czasowych dotyczy ono określenia przyszłych wartości szeregu, zaś w przypadku danych o charakterze przekrojowym (czyli dotyczących różnych obiektów i nie uwzględniających czynnika czasu) prognozowanie może polegać na oszacowaniu pewnych charakterystyk dla nowych, nieznanych w trakcie konstruowania modelu, obiektów;
- symulację - to zastosowanie zakłada wykorzystanie modelu do wykonywania zastępczych eksperymentów, weryfikujących pewne działania przed ich podjęciem w rzeczywistości. Stosowanie wyników *data mining* w tym zakresie pozwala na uzyskanie odpowiedzi na pytanie: *co się stanie jeśli ...?* Ta grupa zastosowań jest szczególnie przydatna przy wspomaganiu procesów decyzyjnych, gdyż model pozwala na sprawdzenie prawdopodobnych skutków ewentualnych decyzji;
- tworzenie systemów ekspertowych - wiedza pozyskana w trakcie analizy danych może tworzyć (lub uzupełniać) bazę wiedzy systemu ekspertowego, czyli programu będącego w stanie odpowiadać - na drodze automatycznie prowadzonego wnioskowania - na pytania formułowane przez użytkowników;
- tworzenie systemów podejmujących decyzje - ten sposób użycia *data mining* sprowadza się do próby zastąpienia decydenta (człowieka) przez sztuczny system decyzyjny. Próby takie podejmowane są w tych dziedzinach, w których czas przewidziany na podjęcie decyzji jest bardzo krótki (na przykład sterowanie procesów w czasie rzeczywistym), a jednocześnie dla podjęcia decyzji konieczne jest przeanalizowanie bardzo dużej liczby informacji.

Uwagi końcowe

Wdrożenie modelu zbudowanego technikami *data mining* wymaga zawsze dalszego monitorowania sposobu jego funkcjonowania. Zmiany zachodzące w rzeczywistości mogą bowiem powodować dezaktualizację modelu i po pewnym czasie uczynić go nieprzydatnym. W przypadku stwierdzenia pojawiających się coraz częściej różnic pomiędzy przebiegiem badanego zjawiska a jego modelowym opisem - należy przeprowadzić aktualizację modelu, która w pewnych okolicznościach będzie polegała na powtórzeniu całej procedury badawczej, zaś w innych sprowadzać się będzie tylko do uaktualnienia parametrów modelu.

Należy bowiem podkreślić, że jakkolwiek mądre użycie omawianych tu technik przynosi znaczące korzyści, to bezkrytyczne stosowanie metod *data mining* może także przynieść duże szkody - zaś odpowiedzialność za podejmowane decyzje ponosi zawsze człowiek, a nie komputer...

ści ich zastosowań w systemach decyzyjnych, Materiały z XXXIII Konferencji Statystyków, Ekonometryków, Matematyków Polski Południowej oraz XV Seminarium Ekonometrycznego im. Profesora Zbigniewa Pawłowskiego, Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław, 1997
[Mrózek i in., 1999] Mrózek A., Płonka L., Analiza danych metodą zbiorów przybliżonych. Zastosowania w ekonomii, medycynie i sterowaniu, Akademicka Oficyna Wydawnicza PLJ, Warszawa, 1999
[Mynarski, 1999] Mynarski S. (red.), Zastosowanie metod wielowymiarowych w badaniach segmentacji i selektywności rynku, Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków, 1999
[Nałęcz, 2000] Nałęcz M. (red.), Biocybernetyka i inżynieria biomedyczna 2000. Tom 6:.Sieci neuronowe, Akademicka Oficyna Wydawnicza Exit, Warszawa, 2000
[Osowski, 1996] Osowski S., Sieci neuronowe w ujęciu algorytmicznym, WNT, Warszawa, 1996
[Pawełek i in., 1995] Pawełek B., Zeliaś A., Proste metody oceny ważności zmiennych diagnostycznych w badaniach taksonomicznych, Folia Oeconomica Cracoviensia, Vol. XXXVII - XXXVIII (1994 - 1995)
[Pawlak, 1982] Pawlak Z., Rough sets, International Journal of Information and Computer Science, vol. 11, No. 341, 1982
[Percival i in., 2000] Percival D. B., Walden A. T., Wavelet Methods for Time Series Analysis, Cambridge Series in Statistical and Probabilistic Methematics, Cambridge University Press, 2000
[Pociecha i in., 1988] Pociecha J., Podolec B., Sokołowski A., Zając K., Metody taksonomiczne w badaniach społeczno - ekonomicznych, Państwowe Wydawnictwo Naukowe, Warszawa, 1988
[Pring, 1998] Pring M. J., Podstawy analizy technicznej, WIG-Press, Warszawa, 1998
[Refenes, 1995] Refenes A.-P. (ed), Neural Networks in the Capital Markets, John Wiley & Sons, 1995
[Rozin, 1979] Rozin B. B., Teoria rozpoznawania obrazów w badaniach ekonomicznych, PWN, Warszawa, 1979
[Rojas, 1996] Rojas R., Neural Networks. A Systematic Introduction. Springer - Verlag, 1996
[Rutkowska, 1997] Rutkowska D., Piliński M., Rutkowski L., Sieci neuronowe, algorytmy genetyczne i systemy rozmyte, Wydawnictwo Naukowe PWN, Warszawa, 1997
[Rymarczyk, 1997] Rymarczyk M. (red.), Decyzje. Symulacje. Sieci neuronowe, Wydawnictwo Wyższej Szkoły Bankowej, Poznań, 1997
[StatSoft, 1997] StatSoft, Inc., STATISTICA for Windows [Computer program manual], Tulsa, 1997
[Szabatin, 2000] Szabatin J., Podstawy teorii sygnałów, Wydawnictwa Komunikacji i Łączności, Warszawa, 2000
[Tadeusiewicz i in., 1991] Tadeusiewicz R., Flasiński M., Rozpoznawanie obrazów, Wydawnictwo Naukowe PWN, Warszawa, 1991
[Tadeusiewicz, 1993] Tadeusiewicz R., Sieci neuronowe, Akademicka Oficyna Wydawnicza RM, Warszawa, 1993
[Tadeusiewicz, 2000] Tadeusiewicz R.: The Application of Neural Networks in Biotechnology and Biomaterials, Prace Mineralogiczne, nr 89, Komisja Nauk Mineralogicznych PAN, 2000, pp. 9 - 17
[Tadeusiewicz, 2001] Tadeusiewicz R.: Problem wyboru właściwej architektury sieci neuronowej. Informatyka w Technologii Materiałów, nr 1, tom 1, 2001, ss. 3-22
[Talaga i in., 1986] Talaga L., Zieliński Z., Analiza spektralna w modelowaniu ekonometrycznym, Państwowe Wydawnictwo Naukowe, Warszawa, 1986
[Theil, 1979] Theil H., Zasady ekonometrii, PWN, Warszawa, 1979
[Trippi i in., 1993] Trippi R., Turban E., (eds), Neural Networks in Finance and Investing, Probus Publishing Company, 1993
[Walesiak, 1996] Walesiak M., Metody analizy danych marketingowych, Wydawnictwo Naukowe PWN, Warszawa, 1996
[Walesiak i in., 2000] Walesiak M., Bąk A., Conjoint analysis w badaniach marketingowych, wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław, 2000
[Zaborski, 2001] Zaborski A., Skalowanie wielowymiarowe w badaniach marketingowych, Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław, 2001
[Zając i in., 1985] Zając K. (red.), Wykłady ze statystyki, Akademia Ekonomiczna w Krakowie, Kraków, 1985
[Zając, 1988] Zając K., Zarys metod statystycznych, PWE, Warszawa, 1988
[Zieliński, 1999] Zieliński T., Jak pokochać statystykę czyli STATISTICA do poduszki, StatSoft Polska Sp. z o. o., Kraków, 1999
[Zieliński, 2000] Zieliński J. (red.), Inteligentne systemy w zarządzaniu. Teoria i praktyka, Wydawnictwo Naukowe PWN, Warszawa, 2000