

COMPUTER RECOGNITION OF DATA STRUCTURES USING CLUSTER ANALYSIS AND THE THEORY OF MATHEMATICAL RECORDS

KOMPUTEROWE ROZPOZNAWANIE STRUKTUR DANYCH Z WYKORZYSTANIEM ANALIZY SKUPIEŃ I TEORII EWIDENCJI MATEMATYCZNEJ

Topolski Mariusz, Topolska Katarzyna

WSB University in Wrocław

Abstract: *The paper proposes a new approach to the agglomeration of data in cluster analysis. The new approach assumes that sets of similar events are attributed to the cumulative probability of their occurrence at the same time. Such approaches will not be found in probability. Thanks to the mathematical theory of records fairly accurate classification of the object can be provided. This is the method which can be used in the cluster analysis by agglomeration. Figure 1 has been drawn for the purposes of better illustration of the problem. It shows the problem of classifying an object to one of the two classes: suitable or unsuitable for further use. Thanks to the merger of two classifiers: KNN algorithm (k nearest neighbours) and belief function a model was created, which is pretty strong as it seems to discriminate against space objects. It therefore seems reasonable to discriminate space of objects. The paper also shows a possibility of applying the proposed model to classification and the correlation between cytokines and features related to the occurrence of lymphocytic leukaemia. It therefore seems justified to carry out tests on this new method as regards various scientific problems.*

Keywords: *mathematical methods, cluster analysis*

Streszczenie: *W pracy zaproponowano nowe podejście do aglomeracji danych w analizie skupień. Nowe podejście zakłada, że zbiorom zdarzeń podobnych przypisywane jest łączne prawdopodobieństwo ich zaistnienia jednocześnie. Takiego podejścia nie znajdziemy w probabilistyce. Dzięki zastosowaniu teorii ewidencji matematycznej można dokonać dość trafnej klasyfikacji obiektu. Jest to metoda którą można stosować w analizie skupień metodą aglomeracji. Dla lepszego zobrazowania przedstawionego modelu wykonano rysunek 1. Przedstawia on problem klasyfikacji obiektu do jednej z klas: zdatny bądź nie zdatny do dalszej eksploatacji. Dzięki zastosowaniu fuzji dwóch klasyfikatorów: algorytmu kNN (k najbliższych sąsiadów) oraz funkcji przekonania został stworzony model, który dość silnie jak się wydaje dyskryminuje przestrzeń obiektów. Przedstawiono również możliwość zastosowania zaproponowanego modelu w zadaniu klasyfikacji i związku przesłanek białek cytokin i cech związanych z występowaniem białaczki limfatycznej. Wydaje się zatem zasadne przeprowadzenie testów nad nową metodą w różnych problemach dziedzin nauki.*

Słowa kluczowe: *metody matematyczne, analiza skupień*

COMPUTER RECOGNITION OF DATA STRUCTURES USING CLUSTER ANALYSIS AND THE THEORY OF MATHEMATICAL RECORDS

1. Introduction

A contemporary economy is based on knowledge. Every day new algorithms are created as well as methods aiming to support the solution of decision problems. Due to high complexity of the area of classified objects, the use of computer is necessary to make necessary calculations. One of the biggest problems which researchers face is the elaboration of methods which will precisely classify objects under uncertain conditions and in view of incomplete information. The theory of probability belongs to classical methods. However, it is limited to the estimation of probability of one incident and not a group of incidents happening at one time. Often in everyday life we deal with situations when various decisions have to be made on the basis of premises. This problem may be dealt with by the theory of mathematical records. Whereas the problem of classifying similar objects may be solved by the cluster analysis. The second point of this paper is devoted to this analysis. The research aims to look for a computer algorithm assigning an object to a certain class under uncertain conditions and uncertainty of measure information in particular with the same premises but different decisions.

2. Cluster analysis

One of the problems contemporary researchers face is organization and grouping data in sensible structures. The term cluster analysis was introduced by Tryon in 1939. The cluster analysis is also called as an exploratory data analysis aiming to assign data to specific structures of highest similarities inside [2]. At the same time combining objects with other objects should be the smallest. The above classification is very often used for classifying animals into specific species, transmitting sources of contamination, evaluating characteristic features of damages to mechanical elements etc. Apart from the cluster analysis used for evaluating the number of factors affecting the measurement of a specific feature, methods of factor analysis or multiple regression are used. These methods are predominantly used in the field of technical sciences.

We apply cluster analysis when we want to distinguish groups of similar objects if the objects are described by more than one feature.

Cluster analysis is a method distinguishing clusters of similar objects when new objects are described by more than one feature. The application of cluster analysis is very comprehensive. It is used by search engines for creating thematic groups, in psychology and medicine for aggregating psychological and somatic symptoms. So clustering or a group of a class shall be used to describe a set of objects where a similarity between a pair of objects is greater than a similarity of any object belonging to a class and an object not belonging to it [1].

There are numerous methods describing distances between clusters i.e. the nearest neighbour, the furthest neighbourhood, a median, a group mean, a centre of gravity or Ward[2].

In cluster analysis methods of clustering are very important and are divided into two categories: hierarchical and non-hierarchical. The hierarchical method groups objects in an iterative way into smaller and smaller clusters.

In non-hierarchical models you transfer objects between clusters looking for the best solution proceeding according to the established criteria. The most common cluster methods are an agglomeration method and K-mean method. Hierarchical models create a hierarchy of classifications for a set of objects starting with the set in which every object creates an independent cluster and ending with a division of objects making up a cluster.

In cluster analysis the following agglomeration methods may be differentiated:

- single-linkage clustering,
- complete-linkage clustering,
- average linkage clustering,
- weighted average clustering,
- centroid linkage,
- medians,
- Ward's methods.

Distance measurements are as follows:

- Euclidean,
- city,
- Chebychev
- power,
- percent disagreement,
- 1-r Pearson's.

3. Elements of the theory of mathematical records

The theory of mathematical records is also called the theory of belief functions and Dempster Shafer. The theory allows us to create models of uncertainty dealing with accepting numerous values of a particular attribute at one time. It allows us to treat plausibility in a subjective way.

When classifying an object on the basis of features $x \in X$, it is classified to a certain class $j \in \theta$, where $\theta = \{1, 2, \dots, M\}$, and M is the number of classes [5].

Probably the occurrence of an event is a function of mass (1).

DEFINITION 1: *A prior function of probability allocation, determined on the basis of subjective judgment (e.g. a learning set) corresponds to prior probability of classes and can be written down as follows [6]:*

$$\begin{aligned} \sum m(\theta) &= 1 \\ m(\theta) &> 0, \end{aligned} \tag{1}$$

Where set θ is a focal element for Bel belief function meeting condition $m(\theta) > 0$.

In the Beys classifier knowledge is represented by conditional and unconditional probability. However, in reasoning by means of the theory of mathematical records we assign sentences with *Bel* value being a belief degree.

The probability allocation function (1) is a component of the *Bel* belief function according to the theory of Dempster-Shafer. The *Bel* belief function is the basis for reasoning and may be classified as follows:

DEFINITION 2: A belief function for a certain fuzzy set $Bel(Y)$ is referred to the function which results from base distribution function of probability allocation a priori class $m(\theta)$ with the function of probability allocation $m(Y^*)$ of the fuzzy set Y^* satisfying the dependency [5]:

$$Bel(Y) = \sum_{Y^{**} \in Y} m(Y^{**}) = \sum_{Y^{**} \in Y} m(\theta) \oplus m(Y^*) \quad (2)$$

$$\sum_{Y^{**} \in Y} m(\theta) \oplus m(Y^*) = \sum_{Y^{**} \in Y} \frac{\sum_{\theta \cap Y^* = Y^{**}} m(\theta) \cdot m(Y^*)}{1 - \sum_{\theta \cap Y^* = \emptyset} m(\theta) \cdot m(Y^*)} \quad (3)$$

where $Y, Y^*, Y^{**} = \theta$ for Beys belief function. In definition 2 there is formula 3 which is called the rule of Dempster's combination. This function enables to make independent convictions and updates.

4. Cluster analysis applying the theory of mathematical records

In cluster classification based on features of object $x \in X$, the object is classified into a certain class $j \in \theta$ (of the cluster), where $\theta = \{1, 2, \dots, M\}$, and M is the number of classes (focuses).

Using the nearest neighbour method according to the cluster analysis (distances between two objects x_i i x_k) we are going to use the Euclidean distance expressed by the pattern:

$$d(x_i, x_k) = d_{ik} = \sqrt{\sum_{j=1}^p (x_{ij} - x_{kj})^2} \quad (4)$$

where x_{ij} – a value of an object x_i featuring j , whereas p – the number of the features.

Having matrix n of objects and p variables we design a matrix of distances between particular objects:

$$D = [d_{ik}], \text{ where } i, k = 1, \dots, n \quad (5)$$

The total of all distances is calculated on the basis of matrix D .

$$d = \sum_{d_{ik}} \sqrt{\sum_{j=1}^p (x_{ij} - x_{kj})^2} \quad (6)$$

The probability allocation function for each distance is established on the basis of the formula as follows:

$$m(\{d_{ik}\}) = \frac{d_{ik}}{d}, \quad (7)$$

meeting the condition $\sum m(\{d_{ik}\}) = 1$ and $m(\emptyset) = 0$.

Recognition of similar objects is an essential stage of classification. For example when analysing object no. 6 we examine its similarity to 5 remaining objects. A situation may arise that the object being classified has 3 objects from cluster 1 and two objects from cluster 2. In the event the difference between objects from two clusters amounts 1, classification may be flawed. Then the distance between all those points is modified to the following set:

$$m(\{d_{ik1}, d_{ik2}, \dots, d_{ikN}\}) = \frac{\sum_{ik} (d_{ik1}, d_{ik2}, \dots, d_{ikN})}{d}. \quad (8)$$

Next, each element of matrix d is modified applying the rule. Finally, we define whether a particular feature belongs to the cluster by maximising the belief function expressed by the formula:

$$Bel(Y) = \min \left[\sum_{Y \in d} m(\Theta) \oplus m(d) \right] = \min \left[\sum_{Y \in d} \frac{\sum_{\Theta \in d=Y} m(\Theta) \cdot m(d)}{1 - \sum_{\Theta \in d=\emptyset} m(\Theta) \cdot m(d)} \right] \quad (9)$$

5. Practical application of the model

We theoretically consider a practical example. The examined mechanical element consists of two types of steel 38GSA and B27. Figure 1 presents measure points illustrating two events. Event A denotes that the mechanical element is suitable for further use. Whereas event B means it is not suitable.

The classification aims to evaluate whether measurement x proves membership to group A or B . In Bayes theory we can establish the classifier which will differentiate with some flaws whether x belongs to A or B . The classifier proposed by this paper except for knowledge about probability of x belonging to group A or B also uses knowledge about the environment of the above point possessed by other objects. If those objects occur in the similar number we have a tool for creating groups and assigning them with overall probability.

According to the classical algorithm of three nearest neighbours, the object x would be classified to cluster A as two objects are nearest to x from set A . Pursuant to the new classification measure $Bel(Y)$ the examined object was classified to cluster B .

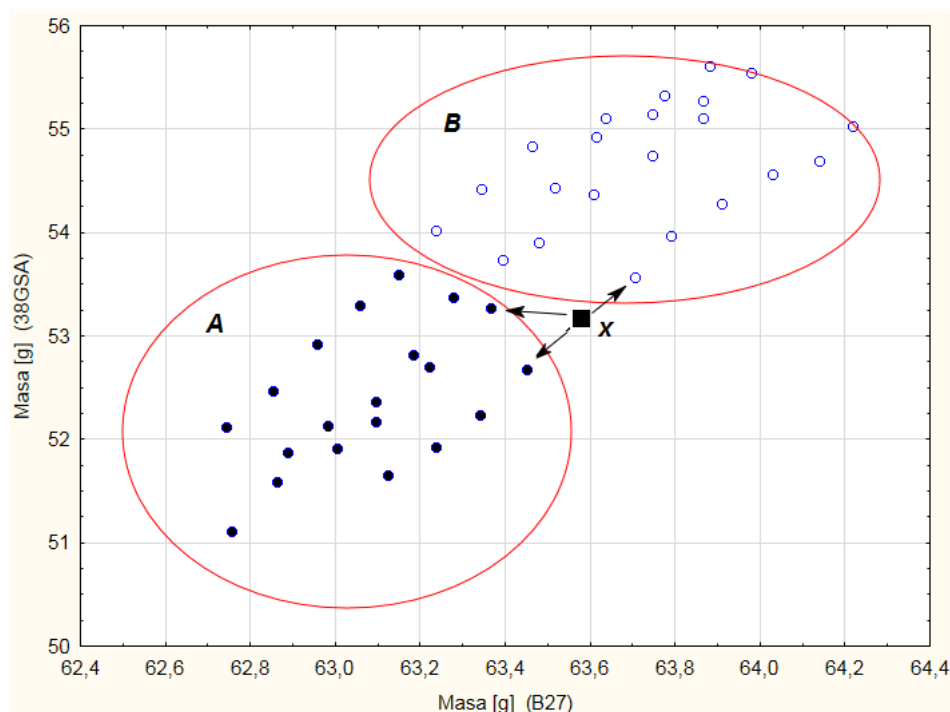


Fig. 1. Example of classifying point x to the set of clusters

The agglomeration method making use of the theory of mathematical records treats vicinity of points as an overall set of three distances (5). By means of a combination formula (6) probability density of the environment of the classified point was calculated. According to Figure 1 in vicinity of x there are more objects from cluster A than B but the density of the other is greater and this translates into the conviction that x belongs to cluster B objects.

The model proposed in this chapter may be also applied to detect association rules, for example to define specific products which customers most often buy together. This model may be used to evaluate a sequence of events for instance the order of places visited by tourists. Most importantly this method may be used for analysing customers' data and concluding about directions of possible purchases made by them.

The algorithm of cluster analysis together with a belief function originating from the theory of mathematical records may be applied to uncertain and imprecise data. In particular, in cases when we may have different classifications on the basis of premises. Such cases occur in medical data. For example the evaluation of cytokine parameters and the occurrence of leukaemia. Cytokines are proteins affecting the growth and proliferation. They stimulate cells taking an active part in immunological response of the organism. Numerous researchers deal with this subject [9]. The cluster analysis is used to evaluate the correlation between cytokines and occurrence of characteristics features of lymphocytic leukaemia.

Figure 2 presents the correlation between parameter of cytokines and leukaemia up to iI over 5 years.

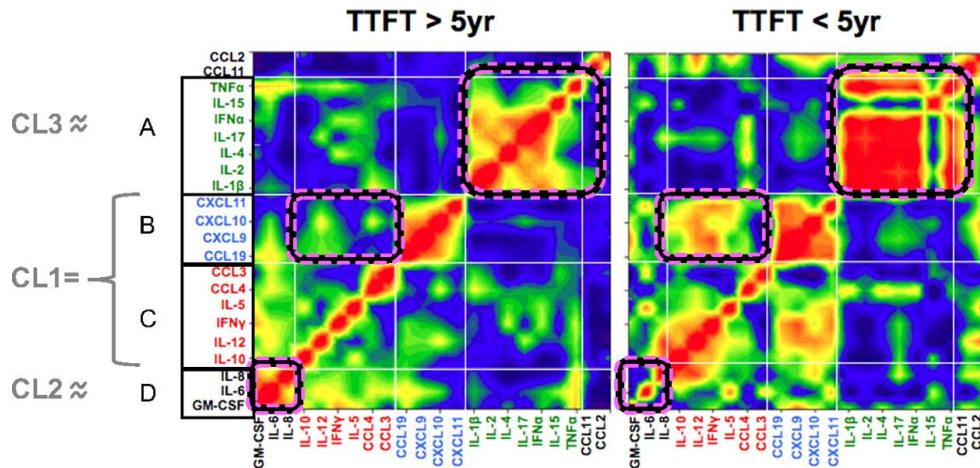


Fig. 2. Example classification of parameters of cytokines and leukaemia [9].

The application of cluster analysis as illustrated in Figure 2 produces a picture of a group of parameters which are most coherent in lymphatic leukaemia. Adding mathematical records for uncertain boundaries between clusters may result in improvement of the quality of classification through more precise boundaries separated between them.

6. Conclusions

Very often researchers face the problem of classifying objects to specific groups. One of the method solving this problem is cluster analysis. The paper has presented a method of agglomerating measure data by means of the theory of mathematical records. This is the method which can be applied to the cluster analysis using the agglomeration method.

Figure 1 aims to illustrate the presented model. It presents the issue of classifying an object to one of the categories: fit or not fit for further usage. Thanks to the application of the merger of two classifiers: kNN algorithm (k nearest neighbours) and a belief function, a model has been drawn up which seems to discriminate the space of objects. The paper also shows a possibility of applying the proposed model to classification and the correlation between cytokines and features related to the occurrence of lymphocytic leukaemia.

7. References

- [1] Gatnar Eugeniusz, Walesiak Marek, *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*. Wrocław : Wydawnictwo Akademii Ekonomicznej, 2004, s. 317.
- [2] Grabiński Tadeusz *Metody taksonometrii*. Kraków: Wyd. AE 1988.

- [3] Kurzyński M., Woźniak M., Żołnierek A., Analiza przydatności wybranych metod rozpoznawania sekwencyjnego dla problemu z zakresu diagnostyki medycznej, Inteligentne Wydobywanie Informacji, Technologie Informacyjne: Diagnostyka. PWNT Gdańsk 2007.
- [4] Stanisław A., Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny, Tom 2. Modele liniowe i nieliniowe, Kraków 2007.
- [5] Topolski Mariusz, Komputerowe algorytmy rozpoznawania sekwencyjnego łączące teorię zbiorów rozmytych z teorią ewidencji matematycznej, Raport PRE 1/08, Politechnika Wrocławska (praca doktorska).
- [6] Trocki M., Zarządzanie Projektami, Polskie Wydawnictwo Ekonomiczne, Warszawa 2003 (cytat, str. 160).
- [7] Witkowski J., „Zarządzanie łańcuchem dostaw. Koncepcje, procedury, doświadczenia”, PWE, Warszawa 2003.
- [8] Woźniak M., Podstawy komputerowego rozpoznawania sterowanych łańcuchów Markowa z regułami eksperta i ciągiem uczącym - algorytmy i ich zastosowanie w diagnostyce medycznej, Raport PRE 2/96, Politechnika Wrocławska (praca doktorska).
- [9] Xiao-Jie Yan, Igor Dozmorov, Wentian Li, Sophia Yancopoulos, Cristina Sison Michael Centola, Preetesh Jain, Steven L. Allen, Jonathan E. Kolitz, Kanti R. Rai, Nicholas Chiorazzi, Barbara Sherry „Identification of outcome-correlated cytokine clusters in chronic lymphocytic leukemia” November 10, 2011.



Topolski Mariusz PhD in technical sciences awarded in 2008. Employed in WSB University in Wrocław at the position of Assistant Professor in the Institute of Logistics. The research conducted involves applying soft calculation methods to production processes supporting procurement logistics, production and distribution. The main specialist in charge of management matters, implementation, programming projects in CILS. A specialist in acquiring strategic customers and corporate marketing (Share 50%).



Topolska Katarzyna PhD in technical sciences awarded in 2007. Employed in WSB University in Wrocław at the position of Assistant Professor in the Institute of Logistics. Head of Department of Logistics engineering studies. Specialist in the design of logistics processes. Longtime coach in national and international companies. Experience in modeling logistic processes with the use of information systems - Anylogic, FlexSim (Share 50%).

KOMPUTEROWE ROZPOZNAWANIE STRUKTUR DANYCH Z WYKORZYSTANIEM ANALIZY SKUPIEŃ I TEORII EWIDENCJI MATEMATYCZNEJ

1. Wstęp

Współczesna gospodarka oparta jest na wiedzy. Każdego dnia powstają nowe algorytmy, metody czy modele których zasadniczym celem jest wspomaganie rozwiązywania problemów decyzyjnych. Ze względu na dużą złożoność przestrzeni klasyfikowanych obiektów aby dokonać obliczeń niezbędne jest wykorzystanie komputera. Jednym z największych problemów badaczy jest opracowanie takich metod, które będą jak najdokładniej klasyfikować obiekty w warunkach niepewności oraz niepełności informacji. Probabilistyka jest dziedziną, która należy do klasycznych metod. Jednakże ogranicza się ona do szacowania prawdopodobieństwa jednemu zdarzeniu a nie np. grupą zdarzeń zachodzących jednocześnie. Często w życiu codziennym mamy do czynienia z sytuacją, w której na podstawie tych samych przesłanek podejmowane są różne decyzje. Z tym problemem może poradzić sobie teoria ewidencji matematycznej. Zaś problem klasyfikacji obiektów podobnych może być rozwiązany za pomocą analizy skupień. To właśnie tej analizie poświęcony jest drugi punkt artykułu. Celem badań jest poszukiwanie komputerowego algorytmu klasyfikacji obiektu do pewnej klasy w warunkach niepewności i niepełności informacji pomiarowej, zwłaszcza w sytuacji tych samych przesłanej ale różnych decyzji.

2. Analiza skupień

Jednym z problemów współczesnych badaczy jest organizacja i grupowanie danych w sensowne struktury. Pojęcie analizy skupień zostało wprowadzone przez badacza Tryona w 1939 roku. Analizę skupień nazywana jest eksploracyjną analizą danych, której celem jest takie uporządkowanie danych do danej struktury aby ich podobieństwo wewnątrz grupy było największe [2]. Jednocześnie powiązanie tych obiektów z innymi obiektami innych grup ma być jak najmniejsze. Z powyższą klasyfikacją spotykamy się bardzo często np. do klasyfikacji zwierząt wewnątrz danego gatunku, emitowania źródeł zanieczyszczeń, oceny cech charakterystycznych uszkodzeń elementów mechanicznych itd. Oprócz analizy skupień do oceny ilości czynników wywierających wpływ na pomiar konkretnej cechy stosuje się metody analizy czynnikowej bądź regresji wielokrotnej. Te metody są przeważnie stosowane w dziedzinie nauk technicznych. Analizę skupień stosujemy gdy chcemy wyróżnić grupy obiektów podobnych, w przypadku gdy obiekty te są opisane przez więcej niż jedną cechę.

Analiza skupień jest metodą wyróżnienia grupy skupień obiektów podobnych, w przypadku kiedy nowe obiekty są opisane przez więcej niż jedną cechę. Zastosowanie analizy skupień jest bardzo obszerne.

Jest ona wykorzystywana w wyszukiwarkach internetowych do tworzenia grup tematycznych, psychologii imedycynie do agregacji objawów psychologicznych i somatycznych. Zatem skupieniem inaczej grupą czy klasą będziemy w dalszej części artykułu opisywać zbiór obiektów, w którym podobieństwo pomiędzy dowolną parą obiektów jest większe niż podobieństwo pomiędzy jakimkolwiek obiektem należącym do klasy, a dowolnym obiektem do niej nie należącym [1]. Istnieje wiele metod opisu odległości między skupieniami tj.: najbliższego sąsiada, najdalszego sąsiedztwa, mediany, średniej grupowej, środka ciężkości czy Warda[2].

W analizie skupień bardzo ważne są metody grupowania, które dzielą się na dwie kategorie: hierarchiczne i niehierarchiczne. W metodach hierarchicznych obiekty łączymy iteracyjnie w coraz mniejsze skupienia. W modelach niehierarchicznych przenosimy obiekty między skupieniami poszukując najlepszego rozwiązania postępując według pewnych ustalonych kryteriów. Najczęstszymi metodami skupień są: metoda aglomeracyjna i k-średnich. W modelach hierarchicznych tworzymy hierarchię klasyfikacji dla zbioru obiektów, zaczynając od zbioru, w którym każdy obiekt tworzy samodzielne skupienie, a kończąc na podziale obiektów, które tworzą skupienie.

W analizie skupień metodą aglomeracji można wyróżnić metody aglomeracji tj:

- pojedyncze wiązanie,
- pełne wiązanie,
- średnich połączeń,
- średnich połączeń ważonych,
- środków ciężkości,
- mediany,
- metody Warda.

Niatomiast miary odległości to:

- euklidesowa,
- miejska,
- Czebyszewa,
- potęgowa,
- niezgodność procentowa,
- 1-r Pearsona.

3. Elementy teorii ewidencji matematycznej

Teoria ewidencji matematycznej nazywana jest również teorią funkcji przekonania oraz Dempstera Shafera. Teoria ta pozwala tworzyć modele niepewności zajmujące się przyjmowanie wielu wartości danego atrybutu jednocześnie. Pozwala ona traktować prawdopodobieństwo w sposób subiektywny.

W klasyfikacji obiektu na podstawie cech $x \in X$ kwalifikuje się go do pewnej klasy $j \in \Theta$, gdzie $\Theta = \{1, 2, \dots, M\}$, a M jest liczbą klas [5].

Prawdopodobieństwo wystąpienia klas jest funkcją masy (1).

DEFINICJA 1 *Aprioryczną funkcją alokacji prawdopodobieństwa, wyznaczoną z subiektywnych sądów (np. zbiór uczący) odpowiada prawdopodobieństwu a priori klas i można je zapisać w formie [5]:*

$$\begin{aligned} \sum m(\theta) &= 1 \\ m(\theta) &> 0, \end{aligned} \quad (1)$$

gdzie zbiór θ jest elementem ogniskowym (fokalnym) dla funkcji przekonania Bel spełniającym warunek $m(\theta) > 0$. W klasyfikacji bayesowskiej wiedza jest reprezentowana przez prawdopodobieństwa warunkowe i bezwarunkowe. Natomiast we wnioskowaniu za pomocą teorii ewidencji matematycznej zdaniom przypisujemy wielkość Bel , będącą stopniem przekonania. Funkcja alokacji prawdopodobieństwa (1) jest składową funkcji przekonania Bel w sensie teorii Dempstera-Shafera. Funkcja przekonania Bel jest podstawą wnioskowania i można ją zdefiniować następująco:

DEFINICJA 2. *Funkcją przekonania dla pewnego zbioru rozmytego $Bel(Y)$ nazywamy taką funkcję, która jest wynikiem złożenia bazowych rozkładów funkcji alokacji prawdopodobieństwa a priori klas $m(\theta)$ z funkcją alokacji prawdopodobieństwa $m(Y^*)$ zbioru rozmytego Y^* spełniającą zależność [5]:*

$$Bel(Y) = \sum_{Y^{**} \in Y} m(Y^{**}) = \sum_{Y^{**} \in Y} m(\theta) \oplus m(Y^*) \quad (2)$$

$$\sum_{Y^{**} \in Y} m(\theta) \oplus m(Y^*) = \sum_{Y^{**} \in Y} \frac{\sum_{\theta \cap Y^* = Y^{**}} m(\theta) \cdot m(Y^*)}{1 - \sum_{\theta \cap Y^* = \emptyset} m(\theta) \cdot m(Y^*)} \quad (3)$$

gdzie $Y, Y^*, Y^{**} = \theta$ dla bayesowskiej funkcji przekonania. W definicji 2 znajduje się wzór 3, który jest nazywany regułą kombinacji Dempstera. Jest to funkcja umożliwiająca składanie niezależnych przekonań, jak również ich aktualizację.

4. Analiza skupień z zastosowaniem teorii ewidencji matematycznej

W klasyfikacji skupień na podstawie cech obiektu $x \in X$ kwalifikuje się obiekt do pewnej klasy $j \in \theta$ (skupienia), gdzie $\theta = \{1, 2, \dots, M\}$, a M jest liczbą klas (skupień).

Wykorzystując metodę najbliższego sąsiada zgodnie z teorią analizy skupień odległością (między dwoma obiektami x_i i x_k) będziemy używali odległości euklidesową, wyrażoną wzorem:

$$d(x_i, x_k) = d_{ik} = \sqrt{\sum_{j=1}^p (x_{ij} - x_{kj})^2} \quad (4)$$

gdzie x_{ij} – wartość obiektu x_i pod względem cechy j , natomiast p – liczba tych cech. Dysponując macierzą n obiektów i p zmiennych konstruujemy macierz odległości między poszczególnymi obiektami:

$$D = [d_{ik}], \text{ gdzie } i, k = 1, \dots, n \quad (5)$$

Z macierzy D liczymy sumę wszystkich odległości.

$$d = \sum_{i,k} \sqrt{\sum_{j=1}^p (x_{ij} - x_{kj})^2} \quad (6)$$

Funkcja alokacji prawdopodobieństwa dla każdej z odległości wyznaczamy ze wzoru:

$$m(\{d_{ik}\}) = \frac{d_{ik}}{d}, \quad (7)$$

spełniającym warunek $\sum m(\{d_{ik}\}) = 1$ oraz $m(\emptyset) = 0$.

Istotnym etapem klasyfikacji jest rozpoznanie obiektów podobnych. Przykładowo analizując obiekt nr 6, analizujemy jego podobieństwo do 5 pozostałych obiektów. Może zaistnieć taka sytuacja, że klasyfikowany obiekt w swoim najbliższym otoczeniu ma 3 obiekty ze skupienia 1 i dwa ze skupienia 2. W takiej sytuacji kiedy różnica między obiektami z dwóch skupień wynosi 1 klasyfikacja jest obarczona największym błędem. Zatem w tym przypadku odległość między tymi wszystkimi punktami jest modyfikowana do zbioru:

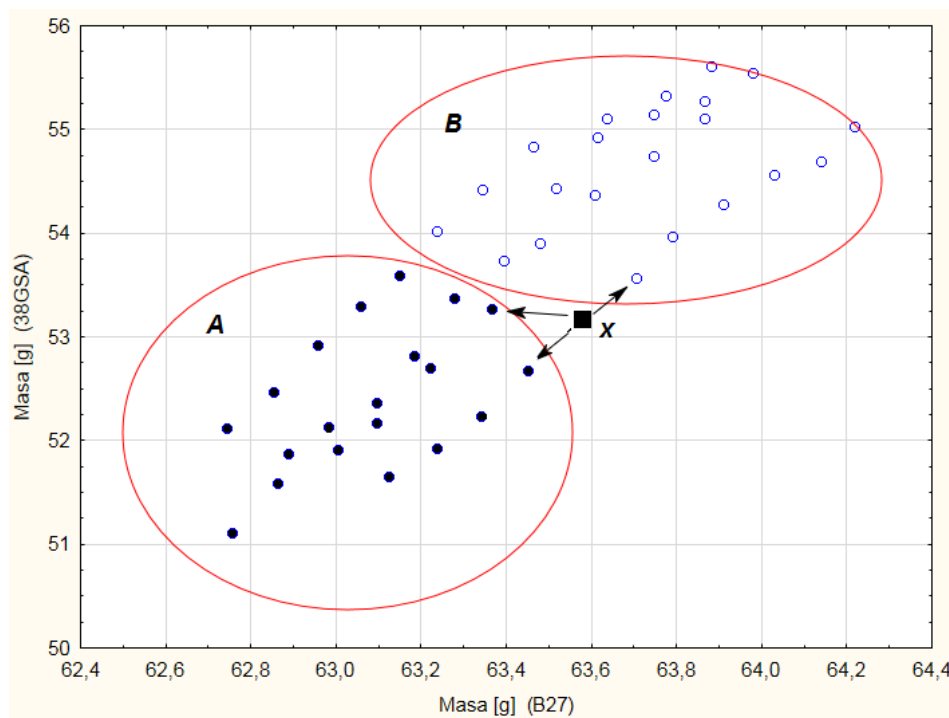
$$m(\{d_{ik1}, d_{ik2}, \dots, d_{ikN}\}) = \frac{\sum (d_{ik1}, d_{ik2}, \dots, d_{ikN})}{d}. \quad (8)$$

W kolejnym kroku modyfikuje się każdy element macierzy d stosując regułę: Ostatecznie przynależność danej cechy do skupienia wyznacza się maksymalizując funkcję przekonania daną wzorem:

$$Bel(Y) = \min \left[\sum_{Y \in d} m(\Theta) \oplus m(d) \right] = \min \left[\sum_{Y \in d} \frac{\sum_{\Theta \in d=Y} m(\Theta) \cdot m(d)}{1 - \sum_{\Theta \in d=\emptyset} m(\Theta) \cdot m(d)} \right] \quad (9)$$

5. Praktyczne zastosowania modelu

Rozważmy tworetycznie przykład praktyczny. Badany element mechaniczny składa się z dwóch stali 38GSA i B27. Na rysunki 1 przedstawiono punkty pomiarowe obrazujące dwa zdarzenia. Zdarzenie A to takie w którym element mechaniczny jest zdalny do dalszej eksploatacji. Natomiast zdarzenie B świadczy o braku owej zdalności.



Rys. 1. Przykład klasyfikacji punktu x do zbioru skupień

Zadaniem klasyfikacji jest ocean czy pomiar x świadczy o przynależności do grupy A bądź B . W teorii Bayesa można wyznaczyć taki klasyfikator który będzie rozróżniał z pewnym błędem przynależność x do grupy A bądź B .

Klasyfikator zaproponowany w niniejszym artykule oprócz wiedzy o prawdopodobieństwie przynależności x do grupy A bądź B wykorzystujemy wiedzę o otoczeniu ww. punktu przez inne obiekty. Jeżeli te obiekty występują

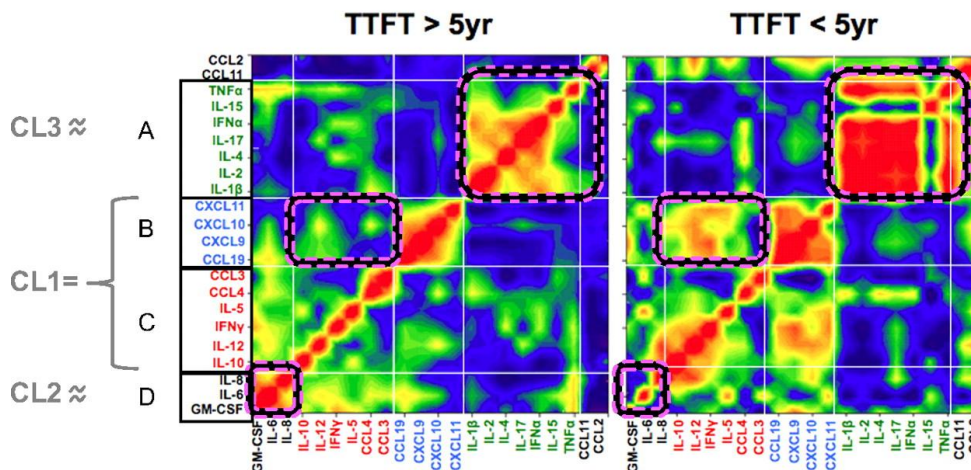
w podobnej liczebności mamy narzędzie pozwalające tworzyć grupy i przydzielać im całościowe prawdopodobieństwo.

Zgodnie z klasycznym algorytmem trzech najbliższych sąsiadów badany obiekt x zostałby zakwalifikowany do skupienia A , gdyż dwa obiekty są najbliżej x ze zbioru A . Na mocy nowej miary klasyfikacji $Bel(Y)$ badany obiekt został zakwalifikowany do skupienia B . W metodzie aglomeracji wykorzystującej teorię ewidencji matematycznej sąsiedztwo punktów zostało potraktowane jako łączny zbiór trzech odległości (5). Za pomocą reguły kombinacji (6) została obliczona gęstość prawdopodobieństwa otoczenia klasyfikowanego punktu. Z rysunku 1 można wywnioskować, że w otoczeniu x mimo, że jest więcej obiektów ze skupienia A niż B , to gęstość tych drugich jest większa, a to przekłada się na większe przekonanie o przynależności x do skupienia obiektów B .

Zaproponowany w niniejszym rozdziale model można zastosować również w zadaniu wykrywania reguł asocjacji, przykładowo do ustalenia konkretnych produktów jakie klienci najczęściej kupują razem. Model można zastosować do oceny kolejności pojawiających się zdarzeń, np. kolejność odwiedzanych przez turystów miejsc. Ważnym zastosowaniem zaproponowanej metody jest możliwość wykorzystania jej do analizy danych klientów oraz wnioskowaniu kierunku możliwości dokonywanych przez nich zakupów.

Algorytm analizy skupień z funkcją przekonania wywodzącą się z teorii ewidencji matematycznej może być stosowany w sytuacji danych niepewnych oraz nieprecyzyjnych. Szczególnie tam gdzie na podstawie tych samych przesłanek możemy mieć różne klasyfikacje. Z takimi przypadkami mamy do czynienia min w danych medycznych. Przykładem może być ocena parametrów cytokiny a występowania białaczki. Cytokiny są to białka wpływające na wzrost oraz proliferację. Pobudzają one komórki jakie biorą czynny udział w odpowiedzi immunologicznej organizmu. Tą tematyką zajmuje się wielu badaczy [9]. Analiza skupień jest wykorzystywana do oceny korelacji ww. cytoklin z występowaniem cech charakterystycznych białaczki limfatycznej. Przykładowo na rysunku 2 przedstawiono związek między parametrami cytokliny i białaczki.

Zastosowanie analizy skupień jak widać na przykładzie zobrazowanym na rysunku 2 daje obraz grup parametrów najsilniej spójnych w chorobie białaczki limfatycznej. Dołączenie do algorytmu ewidencji matematycznej dla niepewnych granic między skupieniami może spowodować polepszenie jakości klasyfikacji poprzez bardziej precyzyjnie wyodrębnionymi granicami między nimi.



Rys. 2. Przykład klasyfikacji parametrów cytokliny i białaczki. Źródło [9].

6. Wnioski

Badacze bardzo często stają przed problematyką klasyfikacji obiektów do określonych grup. Jedną z metod rozwiązywania owego problemu jest analiza skupień. W artykule zaprezentowano metodę aglomeracji danych pomiarowych z wykorzystaniem teorii ewidencji matematycznej. Jest to metoda którą można stosować w analizie skupień metodą aglomeracji. Dla lepszego zobrazowania przedstawionego modelu wykonano rysunek 1. Przedstawia on problem klasyfikacji obiektu do jednej z klas: zdatny bądź nie zdatny do dalszej eksploatacji. Dzięki zastosowaniu fuzji dwóch klasyfikatorów: algorytmu kNN (k najbliższych sąsiadów) oraz funkcji przekonania został stworzony model, który dość silnie jak się wydaje dyskryminuje przestrzeń obiektów. Przedstawiono również możliwość zastosowania zaproponowanego modelu w zadaniu klasyfikacji i związku przesłanek białek cytoklin i cech związanych z występowaniem białaczki limfatycznej. Wydaje się zatem zasadne przeprowadzenie testów nad nową metodą w różnych problemach dziedzin nauki.

7. Literatura

- [1] GATNAR Eugeniusz, WALESIAK Marek *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*. Wrocław : Wydawnictwo Akademii Ekonomicznej, 2004, s. 317
- [2] GRABIŃSKI Tadeusz *Metody taksonometrii*. Kraków: Wyd. AE 1988
- [3] Kurzyński M., Woźniak M., Żołnierek A., Analiza przydatności wybranych metod rozpoznawania sekwencyjnego dla problemu z zakresu diagnostyki

- medycznej, Inteligentne Wydobywanie Informacji, Technologie Informacyjne: Diagnostyka. PWNT Gdańsk 2007
- [4] Stanisław A., Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny, Tom 2. Modele liniowe i nieliniowe, Kraków 2007
- [5] Topolski Mariusz, Komputerowe algorytmy rozpoznawania sekwencyjnego łączące teorię zbiorów rozmytych z teorią ewidencji matematycznej, Raport PRE 1/08, Politechnika Wroclawska (praca doktorska)
- [6] Trocki M., Zarządzanie Projektami. Polskie Wydawnictwo Ekonomiczne, Warszawa 2003 (cytat, str. 160)
- [7] Witkowski J., „Zarządzanie łańcuchem dostaw. Koncepcje, procedury, doświadczenia”, PWE, Warszawa 2003.
- [8] Woźniak M., Podstawy komputerowego rozpoznawania sterowanych łańcuchów Markowa z regułami eksperta i ciągiem uczącym - algorytmy i ich zastosowanie w diagnostyce medycznej, Raport PRE 2/96, Politechnika Wroclawska (praca doktorska)
- [9] Xiao-Jie Yan, Igor Dozmorov, Wentian Li, Sophia Yancopoulos, Cristina SisonMichael Centola, Preetesh Jain, Steven L. Allen, Jonathan E. Kolitz, Kanti R. Rai, Nicholas Chiorazzi, Barbara Sherry „Identification of outcome-correlated cytokine clusters in chronic lymphocytic leukemia” November 10, 2011



Dr inż. Katarzyna Topolska - stopień doktora uzyskany w Instytucie Konstrukcji i Eksploatacji Maszyn Politechniki Wroclawskiej. Zatrudniona w Międzynarodowej Wyższej Szkole Logistyki i Transportu we Wrocławiu, obecne na stanowisku adiunkta. Współpracuje z Państwową Wyższą Szkołą w Wałbrzychu na stanowisku starszego wykładowcy. Prowadzenie szkoleń z zakresu logistyki, zarządzania, zarządzania logistycznego i systemów wspomagających procesy logistyczne (udział 50%).



Dr inż. Mariusz Topolski – stopień doktora uzyskany na Wydziale Elektroniki Politechniki Wroclawskiej. Zatrudniony w Międzynarodowej Wyższej Szkole Logistyki i Transportu we Wrocławiu, obecne na stanowisku adiunkta. Współpracuje z Państwową Wyższą Szkołą w Wałbrzychu na stanowisku starszego wykładowcy. Prowadzenie licznych szkoleń z zakresu sieci komputerowych, systemów operacyjnych, języków programowania C, inżynierii oprogramowania (prowadzenie kompleksowo projektu informatycznego), grafiki komputerowej (udział 50%).