SILESIAN UNIVERSITY OF TECHNOLOGY PUBLISHING HOUSE

SCIENTIFIC PAPERS OF SILESIAN UNIVERSITY OF TECHNOLOGY 2024
ORGANIZATION AND MANAGEMENT SERIES NO. 213

APPLICATION OF CHROMATOGRAPHIC DATA SEPARATION METHODS IN QUALITATIVE AND QUANTITATIVE DATA ANALYSIS

Mariusz ŚWIĘCICKI

Cracow University of Technology; mariusz.swiecicki@pk.edu.pl

Purpose: Currently, one of the significant limitations of problems related to data mining using classical classification methods, the results of which are used in the decision-making process, is the processing of large data sets. The first significant limitation in the use of classical classification methods is the need to ensure a constant data size. The second type of limitation is related to the dimension of the data. The last type of limitation that occurs when using classic classification algorithms is associated with the situation that a given input vector may contain data belonging to many classes simultaneously, then we are talking about the so-called multiclass vectors. On the other hand, as a result of processing large data sets, we want to obtain information that is not only qualitative but equally important in the decision-making process is quantitative information.
Design/methodology/approach: This work presents data classification methods based on the gas chromatography technique, which in issues related to the classification of large data sets are not subject to the above limitations and provide quantitative and qualitative information.
Findings: The article presents classification results for selected data sets. In the first case, the process of classifying sets was carried out, the individual vectors contained several tens of thousands of elements and several thousand attributes. Direct classification of vectors of such dimensions using commonly known methods without reducing the dimension of the data is practically impossible. The second type of data set is a heterogeneous set, i.e. a set containing various types of data, where, as in the first case, the input data vectors are suitably long. The third type of test data set is a multi-class set, during the classification of this set qualitative information is provided, as is the case with classic data mining methods, and quantitative information, which is a unique feature of unproposed data classification methods.
Originality/value: The article proposes an innovative method for the classification of multidimensional data based on the method of chromatographic separation of substances in gas chromatography. This method can be used in the classification of multi-class variable-length data vectors. This work shows that based on the chromatographic separation method, we obtain information also of a quantitative nature, and not only of a qualitative nature.
Keywords: natural computing algorithms, chromatographic separation, signal processing; data mining.
Category of the paper: Research paper.

1. Introduction

Nowadays, one of the important issues of machine learning is the processing of large data sets. Issues related to large data sets refer primarily to data sets that contain a large volume of data and data sets that are complex, i.e. they do not have a specific structure as in the case of data sets represented using a relational database, and also the spectrum of data types stored in this type of collections is wide ranging from text in natural language and a stream of numerical data through a set of graphic images to audio and video data (Hilbert, 2016; Reinsel et al., 2017).

The article presents an algorithm that can be used in issues related to the classification of large-scale data sets. The motivation to define this type of algorithm was the fact that currently the methods used to process this type of data are subject to several significant limitations.

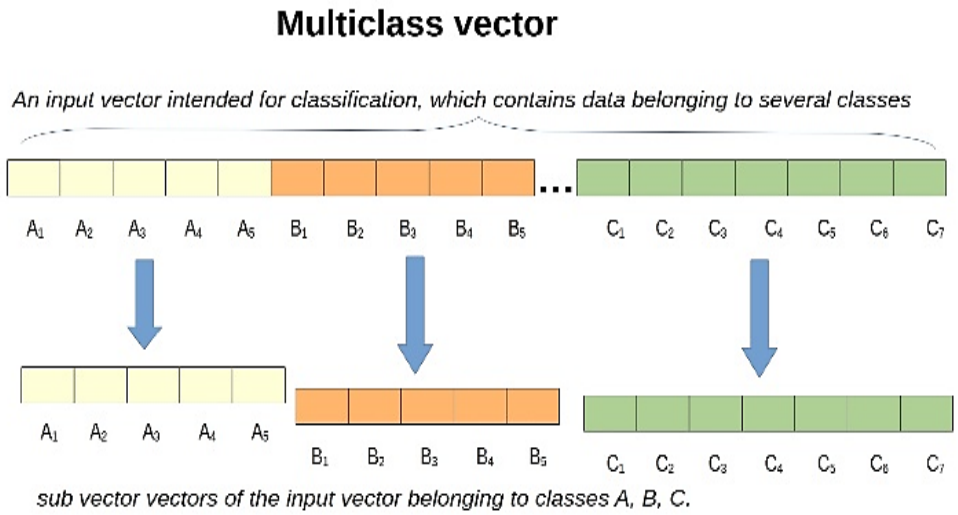

Figure 1. Multiclass vector structure.

The first significant limitation in the use of classical classification methods is the need to ensure a constant size of data - vectors that will be subject to the classification process. The second type of limitation is related to the dimension of the data. When we use classical methods for classifying large vectors, we always have to reduce the dimension of the input vectors using selected mathematical statistics methods (Hilbert, 2016; Reinsel et al., 2017).

Another limitation of currently used algorithms is that the classified data must be homogeneous, i.e. there can only be one type of data. If images are classified, non-image data whose data source is another phenomenon and which is in some way related to the classified images cannot also be classified as input. Finally, the last type of limitation that occurs when using classic classification algorithms is related to the situation that a given input vector may contain data belonging to many classes at the same time, and then in this article we are talking about the so-called called multi-class vectors (Brabazon et al., n.d.).

Moreover, it would be advisable that when processing the above type of data sets, the algorithms should provide both qualitative information, as is the case in most of the algorithms used, but also quantitative information.

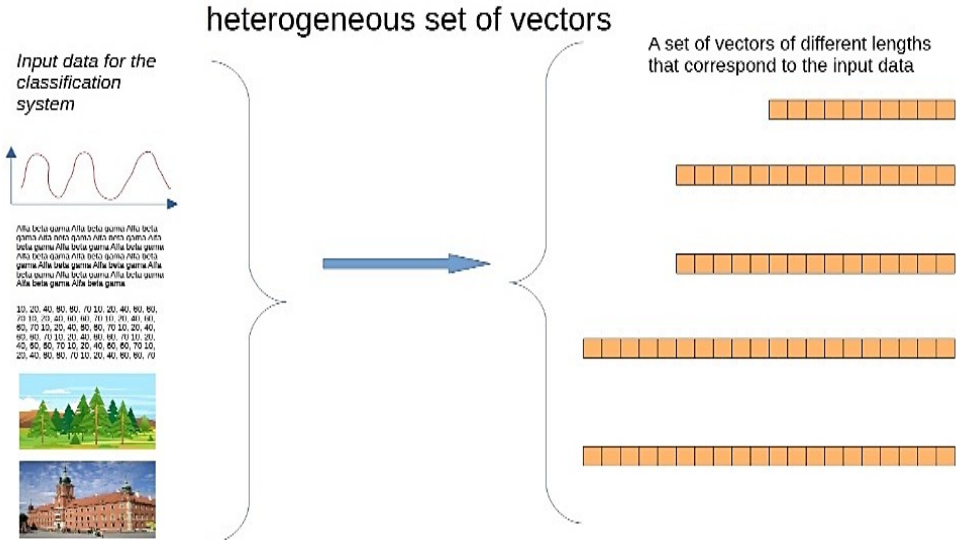

Figure 2. Vector structure of a heterogeneous set of vectors.

The presented algorithm attempts to solve the problems defined above. The presented algorithm attempts to solve the problems defined above. The algorithm is inspired by the method of chromatographic separation of substances, which is used in analytical chemistry. The first chapter of the article will present the principle of chromatographic data separation, which is the basis of the defined algorithm. The second chapter will present an algorithm for data classification, which is inspired by the principles of chromatographic separation of substances using gas chromatography. The following parts of the article will present the results of calculations and classification for example data sets.

2. Properties of chromatographic separation of substances

Chromatography – an analytical technique used to separate or examine the composition of mixtures of chemical compounds. In each chromatographic technique, the mixture being tested is first separated and then the individual components are detected (Leweke, von Lieres, 2018; Pierce et al., 2021; Stilo et al., 2021). The substance is separated by passing the solution of the tested mixture through a specially prepared separation phase (bed), also called the stationary phase. The separation phase consists of substances that have sorption abilities or can otherwise affect the flowing substances. The intensity of this process varies for individual components of the mixture. Some components therefore remain in the phase longer and others for a shorter time, which allows them to be separated. The time a given component stays in the column is called the retention time (Blumberg, 2021; Pierce et al., 2021). Generally speaking,

chromatographic separation is a process in which a mixture of chemical compounds is separated into at least two fractions with different compositions. From a chemical point of view, the purpose of the substance separation process is to increase the concentration of one of the components of the initial mixture about the remaining components of the initial mixture. Separation takes place using physical methods and chemical reactions.

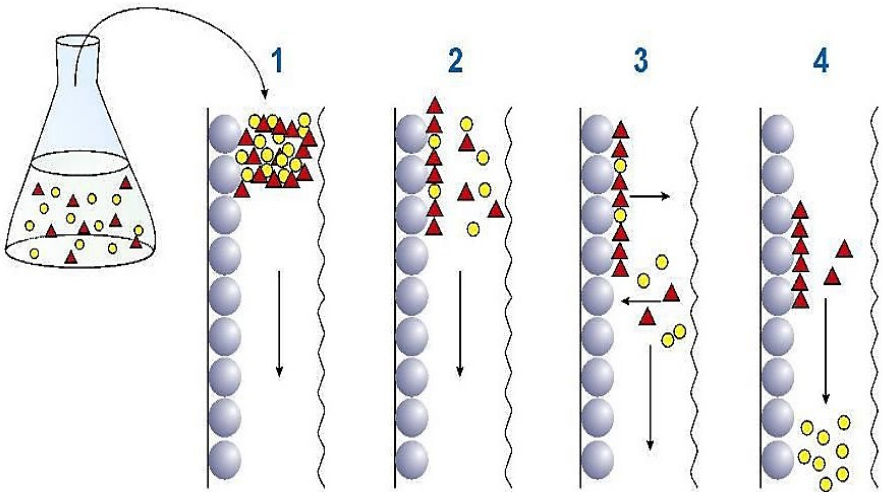

Figure 3. The idea of chromatographic separation of substances.

Source: https://chemia.ug.edu.pl/sites/default/files/_nodes/strona-chemia/17427/files/wyklad5_new.pdf

Figure 3 presents the chromatograph process, the idea of separating a mixture using chromatography we can see, the mixed substance is introduced at the entrance of the chromatographic column. The chromatographic column is filled with a substance or substances that have a different degree of affinity for the substances that are being separated - they have been introduced into the chromatographic column (Varhadi et al., 2020). Due to the above, the time it takes for each substance to leave the chromatographic column will be different and will depend on the degree of affinity of a given substance for substances that are in the stationary phase. The output data stream of the chromatograph is the relationship between the concentration of a given substance over time. This relationship is presented by a chromatograph, i.e. a graph showing the relationship between the concentration of a given substance and the time needed to leave the chromatographic column, i.e. the retention time (Giddings, 2017; Robards, Ryan, 2021).

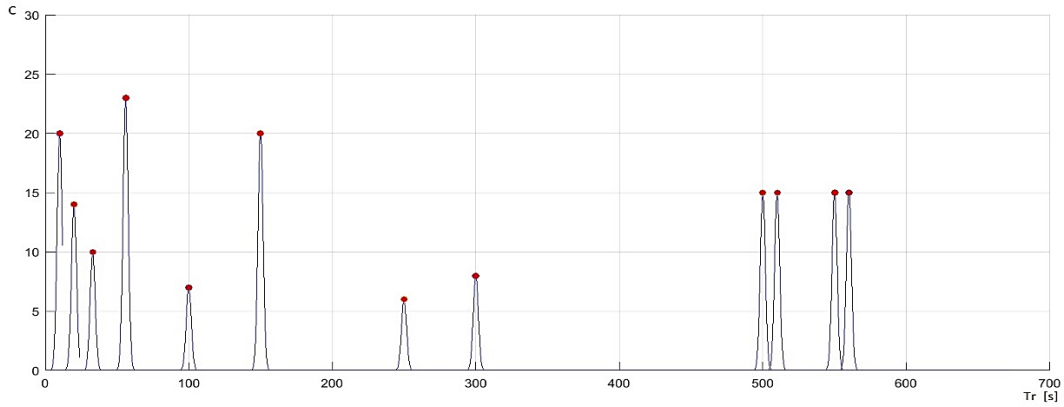

Figure 4. Graph of the substance concentration in the eluate flowing from the column as a function of the elution volume, and when the eluent flow rate is constant as a function of time.

As is known, the retention time of a given substance is characteristic and depends on the structure of the stationary phase and the structure of the substance being separated.

The chromatographic column is filled with a substance or substances that have a different degree of affinity for the substances that are being separated - they have been introduced into the chromatographic column (Hage, 1999; Urh et al., 2009; Varhadi et al., 2020). Due to the above, the time it takes for each substance to leave the chromatographic column will be different and will depend on the degree of affinity of a given substance for substances that are in the stationary phase. The output data stream of the chromatograph is the relationship between the concentration of a given substance over time. This relationship is presented by a chromatograph, i.e. a graph showing the relationship between the concentration of a given substance and the time needed to leave the chromatographic column, i.e. the retention time. As is known, the retention time of a given substance is characteristic and depends on the structure of the stationary phase and the structure of the substance being separated.

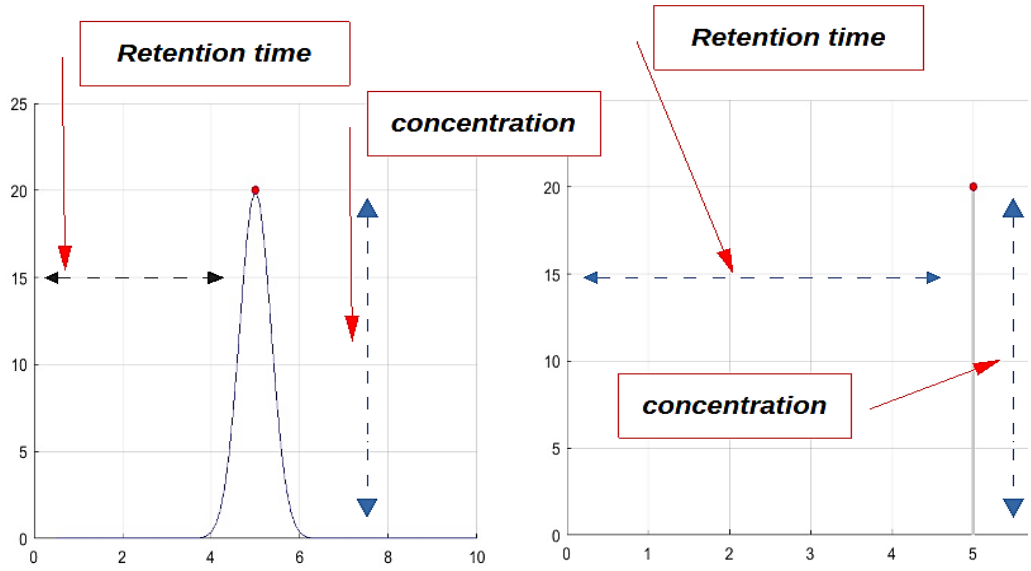

Figure 5. Substance identification: peak position, Peak height: quantification.

Figure 5 shows a signal output generated by the chromatograph. This signal provides two important pieces of information: firstly, it provides information about the type of substance, which is responsible for the retention time - individual for each substance. The second type of information is quantitative information, how much of a specific substance was in the tested mixture. This is determined by the height of the peak, which informs about the concentration of the tested substance at the output of the chromatographic column (Hohrenk et al., 2020; Pezzatti et al., 2020; Witting, Böcker, 2020).

The figure on the left shows the output from a real chromatography system. The figure on the right shows the output signal generated by the presented algorithm. You can see that in this case we are dealing with an ideal peak, the output signal is not blurred due to, for example, detector imperfections.

3. Definition of chromatographic data separation algorithm

The chromatographic data separation algorithm is based on the basic paradigm that the processed data string is a complex chemical molecule with a chain-linear structure. This means that each data vector or set of vectors will be processed by the chromatographic algorithm in accordance with the rules that apply in the real chromatographic system (Święcicki, 2024).

The general principle of operation of the chromatographic data separation algorithm will be to treat the data vector as a mixture of chemical compounds and for each "chemical" compound the relationship between the concentration of a given substance at the output of the chromatographic system is calculated. In other words, it will involve calculating the spectrum as shown in the figure. In the first phase, we treat the vector of numbers as a polyatomic molecule with a linear structure. In the next phase, the molecule is divided into smaller molecules. In the last phase of the algorithm, each newly created molecule is processed by the chromatographic column, i.e. the retention time is calculated. As a result of these operations, a chromatogram is created, i.e. a graph describing the concentration of a given type of molecules as a function of time at the output of the "chromatographic" column. This relationship, i.e. the chromatogram, is later called the spectrum of a given starting substance.

The chromatographic data separation algorithm consists of the following sequence of operations, which are inspired by the functioning of a real chromatographic system:

1) Mixture creation phase for a given vector.
2) Retention time calculation phase.
3) Chromatogram creation phase.
4) Phase of the analysis chromatogram.

3.1. The phase of creating a mixture for a given

In the first phase of this algorithm, a set of vectors W consisting of any number of vectors of any length is transformed into a set of mixtures of substances through the process of dividing the fragmentation into smaller vectors of the same length. The fragmentation of the vector takes place in such a way that for each element from the set *W*, a mixture of substances is created that corresponds to this element of the set *W*.

Input data
$W=\{w_1,w_2,w_3....w_N\}$ – a set of data vectors that will be processed
Output data
set of substances $\boldsymbol{MS_i}$,that have been processed by a chromatographic column, i.e. they have a calculated retention time t_r
$\boldsymbol{MS_{i..M}=[];}$
$\boldsymbol{W=\{w_1, w_2, ...w_M\}}$
Foreach w ϵ W

1 For a given w_i data vector, create a mixture of substances - it will fragment the vector into sub-vectors of constant length
$\boldsymbol{MS_i=\{s_1, s_2, ...s_{M(i)}\}}$
$\boldsymbol{MS_i}$ *-a set of substances is created by dividing a vector into sub-vectors according to the adopted principle of division,*
$\boldsymbol{ms_i}$ *- the elements of this set is the set of substances resulting from the division of the vector* $\boldsymbol{w_i}$*,this means that the set will contain individual substances s which are not subject to further subdivision*
$\boldsymbol{ms_{M(i)}:=\{s_1, s_2, ...s_{M(i)}\}}$ *a substance that was created by splitting the wi vector.* $\boldsymbol{w_i}$.
2 ***Foreach s ϵ*** $\boldsymbol{ms_i}$

3 ***Calculate Retention Time*** *// t_r – the residence time of the substance in the stationary phase e.g.*
// according to the formula (5)
4 ***End***
End

Algorithm 1. Algorithm transforming a set of vectors into a set of chromatograms.

As shown in the algorithm presented above, the set of mixtures of substances that has been created is fed to the input of a "virtual chromatographic column" in which the process of migration of a given substance between the stationary phase and the mobile phase takes place.

3.2. Phase of calculating the retention time

The value of the retention time t_r depends on the affinity of the stationary phase for a given substance, which is an important value in the classification process. It is known that the value of the retention time depends on the affinity of the substance for the stationary phase that is filled in the chromatographic column. The final fragment of Algorithm 1**.** contains a sequence calculating the retention time value.

3.3. Chromatogram creation phase

The next stage of the presented algorithm is to create a chromatogram for a given mixture of substances that corresponds to the w_i element. The chromatogram is created as a result of the

registration of individual substances at the output of the chromatographic column. The moment at which a given substance will leave the chromatographic column depends on the retention time t_r. The purpose of the detector is to count the molecules of substances leaving the chromatographic column at a given moment of time.

Input data
For a given set of substances ***MS$_i$*** *,that have been processed by a chromatographic column, i.e. they have a calculated retention time* t_r
Output data
CH =$\{ch_1, ch_2, ch_3 \ldots ch_N\}$ *– a set of chromatograms, where each element of this set represents a chromatographic spectrum corresponding to a given element of the set W*
$ch_i = \{peak_1, peak_2, peak_3 \ldots peak_M\}$ *– Each chromatogram consists of a set of peaks*
ch_i***=[];***
Foreach s ϵ MS$_i$
peak$_i$[s.Tr]:=peak$_i$[s.Tr]+1 *// Calculation of "concentration" under a vector with a given retention time*
end

Algorithm 2. Algorithm for creating a ch_i chromatogram for a mixture belonging to the w_i vector.

To sum up, the operation of the two algorithms presented above, which model the processes occurring in a real chromatograph, can be presented below in a formalized notation that will later be used in the analysis of the algorithm. Let us assume that the stationary phase *FS* is an *m*-element vector as shown in equation (1) while the substance vector that was created as a result of the algorithm in the fragmentation process as a result of the operation of the first algorithm 1 is also an array with dimensions *NxM* presented in equation (2)

$$FS = \left(fs_1, fs_2, fs_3 \ldots fs_M\right) \tag{1}$$

$$S = \begin{bmatrix} s_{1,1}, s_{1,2}, s_{1,3} \ldots\ldots s_{1,M} \\ s_{2,1}, s_{2,2}, s_{3,1} \ldots\ldots s_{3,M} \\ \ldots\ldots\ldots \\ s_{N,1}, s_{N,2}, s_{N,1} \ldots\ldots s_{N,M} \end{bmatrix} \tag{2}$$

As we know, a chromatogram is made up of peaks, and a single peak is a pair of numbers, the first of which is the retention time t_r and the second is the concentration of substance C, formula (3), as the result of the operation of .

$$peak_i = \left(tr_i, C\right) \tag{3}$$

The retention time can be calculated using the F_{tr} function, which calculates the retention time value for a given substance and the vector describing the stationary phase (4).

$$tr_i = F_{tr}\left(S_{i,1\ldots M}, FS\right) \tag{4}$$

When calculating the retention time, the function calculates the retention time for a given substance taking into account the structure, i.e. the values of the stationary phase. For the purposes of further considerations, it can be assumed that the function calculating the retention time is expressed by formula (5).

$$tr_i = F_{tr}\left(S_{i,1\ldots M}, FS\right) = \sum_{k=1}^{M}\left(s_{i,k} \cdot fs_k\right) \tag{5}$$

As the presented formula shows, the scalar product of two vectors is calculated. The more similar the vectors are to each other, the greater the value of the calculated product, the greater the retention time for a given substance.

3.4. Chromatogram formation phase

The next stage of the presented algorithm is to create a chromatogram for a given mixture of substances that corresponds to the w_i element. The chromatogram is created as a result of the registration of individual substances at the output of the chromatographic column. The moment at which a given substance will leave the chromatographic column depends on the retention time t_r. The purpose of the detector is to count the molecules of substances leaving the chromatographic column at a given moment of time.

3.5. Spectrum analysis phase

The last stage of recognizing substances that have been processed by the chromatographic system is the stage of classifying the output chromatographic spectrum and assigning it to the spectra of known substances.

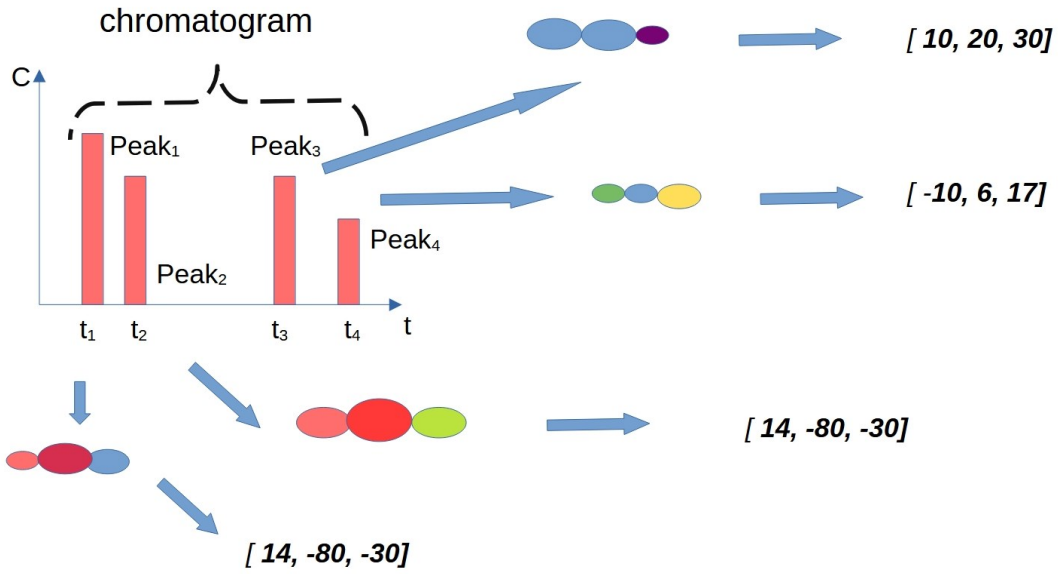

Figure 6. Structure of the chromatogram - the spectrum corresponding to the w_i vector.

The chromatogram of the tested mixture of substances describes the concentration of individual compounds of the fractions included in the tested substance or mixture, which were separated as a result of the chromatographic process, similarly to the presented algorithm. As shown in the figure above, as a result of the chromatography process, we obtain a chromatogram containing many peaks corresponding to the concentration of the "substance" that resulted from the operation of algorithm 1. The task of the classification algorithm will be

to assign the ch_x chromatogram to the chromatograms of known vectors, using the matching criterion, which is the peak retention time.

The last element of the data classification method discussed in this work is chromatogram classification. The previous points presented the structure of the chromatogram, which consists of peaks, and each peak has two attributes, namely its position on the level axis and the concentration of a given molecule, which is determined by the height of the peak.

This means that each chromatogram of the tested substance consists of a set of peaks, not necessarily in the same number. Having a set of chromatograms of reference substances, we can assign the chromatogram of the tested substance to the set of chromatograms of reference substances.

In this work, the assignment of the chromatogram of the tested substance to the set of reference chromatograms will be done on a different basis than as shown in work (Święcicki, 2024). Therefore, the time complexity of the previously proposed algorithm (Święcicki, 2024) was a significant problem in the case of a large number of chromatograms.

For this purpose, the following assumption was made that a given chromatogram is a set of points in n-dimensional space. The size of this space is determined by the design of the detector in the case of a physical chromatograph - that is, the detector is responsible for the sensitivity of the system, and this is also the case with the presented algorithm. A more sensitive detector will be able to identify the time moments in which particular fractions of the tested substances (i.e. vectors) appear at the output of the chromatographic column with greater accuracy. Individual moments of time refer to the retention time. It is known that in the case of the presented algorithm, the identification of retention time is important and determines the quality of the classification process.

For the needs of the presented algorithm, the time moments in which the detector will identify the concentration of a given fraction have been quantized, i.e. the retention time has been assumed to take a value that will correspond to the detector's operation, i.e. it is assumed that the detector in the presented algorithm is a discrete element when it comes to identifying the time moments in which the identified are the concentrations of individual fractions.

For such assumptions, the algorithm for assigning the chromatogram of an unknown "substance", i.e. a subvector that was created as a result of the fragmentation process, will take a simpler form compared to the algorithm presented in work (Święcicki, 2024)

Let us assume that the set of chromatograms is represented by a matrix of dimensions M by N. In this case, M will represent the number of reference chromatograms, while N will correspond to the maximum number of peaks in the chromatogram. N will depend on two factors, firstly, the accuracy of the detector - and in the case of the presented algorithm, the distance between time points, secondly, what retention time interval will be considered, and this depends on the classified data and the structure of the stationary phase. Formulas (6) and (7) show the proposed notation. If the chromatogram for a given column does not have a peak, the value zero is entered because the table columns clearly determine the retention time.

$$SetOfChrom=\begin{bmatrix} peak_{1,1}, peak_{1,2}, peak_{1,3}..... peak_{1,N} \\ peak_{2,1}, peak_{2,2}, peak_{3,1}..... peak_{3,N} \\ ..... \\ peak_{M,1}, peak_{M,2}, peak_{M,1}..... peak_{M,N} \end{bmatrix} \tag{6}$$

$$Ch_x=\left[peak_1, peak_2, peak_3..... peak_N \right] \tag{7}$$

On the other hand, it is known that each peak consists of two attributes (3): retention time and concentration of a given substance. Two matrices will then be used to represent the set of chromatograms. A matrix containing information about retention times, marked with the symbol *Tr*, and a matrix *C* containing information about the height of a given peak, i.e. the concentration of a given fraction. What is represented by the formulas (8) and (9).

$$SetOfTr=\begin{bmatrix} tr_{1,1}, tr_{1,2}, tr_{1,3}.....tr_{1,N} \\ tr_{2,1}, tr_{2,2}, tr_{3,1}.....tr_{3,N} \\ ..... \\ tr_{M,1}, tr_{M,2}, tr_{M,1}.....tr_{M,N} \end{bmatrix} \tag{8}$$

$$SetOfC=\begin{bmatrix} c_{1,1}, c_{1,2}, c_{1,3}..... c_{1,N} \\ c_{2,1}, c_{2,2}, c_{3,1}..... c_{3,N} \\ ..... \\ c_{M,1}, c_{M,2}, c_{M,1}..... c_{M,N} \end{bmatrix} \tag{9}$$

$$Tr_x=\left[tr_1, tr_2, tr_3.....tr_N \right] \tag{10}$$

$$C_x=\left[c_1, c_2, c_3..... c_N \right] \tag{11}$$

The T matrix will be able to provide qualitative information, while the C matrix will be used for calculations that will provide quantitative information.

With this notation, performing calculations aimed at identifying an unknown chromatogram will involve performing relatively simple mathematical operations. Let us assume that the distance between two chromatograms will be measured using the cosine metric.

$$max\left(SetOfTr \cdot Tr_x^T \right)=max\left(\begin{bmatrix} tr_{1,1}, tr_{1,2}, tr_{1,3}.....tr_{1,N} \\ tr_{2,1}, tr_{2,2}, tr_{3,1}.....tr_{3,N} \\ ..... \\ tr_{M,1}, tr_{M,2}, tr_{M,1}.....tr_{M,N} \end{bmatrix} \cdot \left[trx_1, trx_2, trx_3.....trx_N \right]^T \right) \tag{12}$$

With this metric defined, finding the best match for an unknown chromatogram will come down to calculating formula (10) and finding the element with the highest value. The coordinates of this element will uniquely identify the chromatogram that best matches the identified chromatogram and thus the class to which this chromatogram belongs.

0 **Input data**
ch_x={$peak_1$,$peak_2$,$peak_3$....$peak_N$} – a chromatogram consisting of N peaks
Tr_x=[tr_1,tr_2,tr_3....tr_N] – according to the formula (10)
C_x=[c_1,c_2,c_3....c_N] – according to the formula (11)
N - determines how many classes it can be classified into at the same time

SetOfCH={ch_1,ch_2,ch_3....ch_M} according to the formula (6)
SetOfTr - according to the formula (8)
SetOfC - according to the formula (9)

Output data
SetNoClass – the class number to which the ch_x chromatogram was assigned
SetC - Concentrations of individual fragments occurring in the input vector
1 *R=SetOfTr*Tr_x // R={r_1,r_2,r_3.....r_N} according to the formula (12)*

2 *[RS, Index]=sort(R) // sorting by descending order*

3 *Index=Index(1:M) // obtaining M indices of the elements of the vector R with the largest value*

4 *Class=Index2NoClass(Index)// mapping the element's pattern from the R array to the class number*

5 *SetNoClass =SelectMostFrequentlyOccurringElements(N, Class)// N - determines how many classes it can be classified into at the same time*
6 ***Foreach s є SetNoClass***
SetC(s)= C_x(s) / SetOfC(s) // concentration calculation

End

Algorithm 3. Algorithm for classifying a vector w_x using its chromatogram ch_x where the chromatogram ch_x belongs to single-class or multi-class set.

The algorithm responsible for the classification of chromatograms is presented above. This algorithm consists of several important parts. In the first line 1, the degree of matching of the classified chromatogram to all reference chromatograms is calculated according to the cosine metric. In the next lines, *M* standard chromatograms that best match the tested chromatogram are selected.

In the following lines, the obtained indices to the best-fitting chromatograms are mapped to class numbers, and then the frequencies of occurrences of individual classes are counted.

According to the value of the *N* parameter, the N elements from the Class set that appear most frequently are selected. The *SetNoClass* variable will contain the class numbers that best matched the input chromatogram.

In line 6 of the presented algorithm, the concentration of individual fragments of the input vector is calculated in relation to the standard chromatograms

3.6. Problems of selecting the stationary phase

There are two significant problems when performing calculations using the algorithm presented above. The first problem, which was already indicated in the previous chapter, is related to the selection of the stationary phase in such a way that the chromatograms of vectors belonging to different classes are characterized by different retention times (Leweke, von Lieres, 2018; Pierce et al., 2021; Principle and Procedure…, n.d.). The second problem,

in a sense, is a derivative of the first problem, and is related to the fact that the chromatograms that are created in the process are complex, i.e. they contain a large number of peaks, which makes the classification process difficult by the presented algorithm classifying chromatograms.

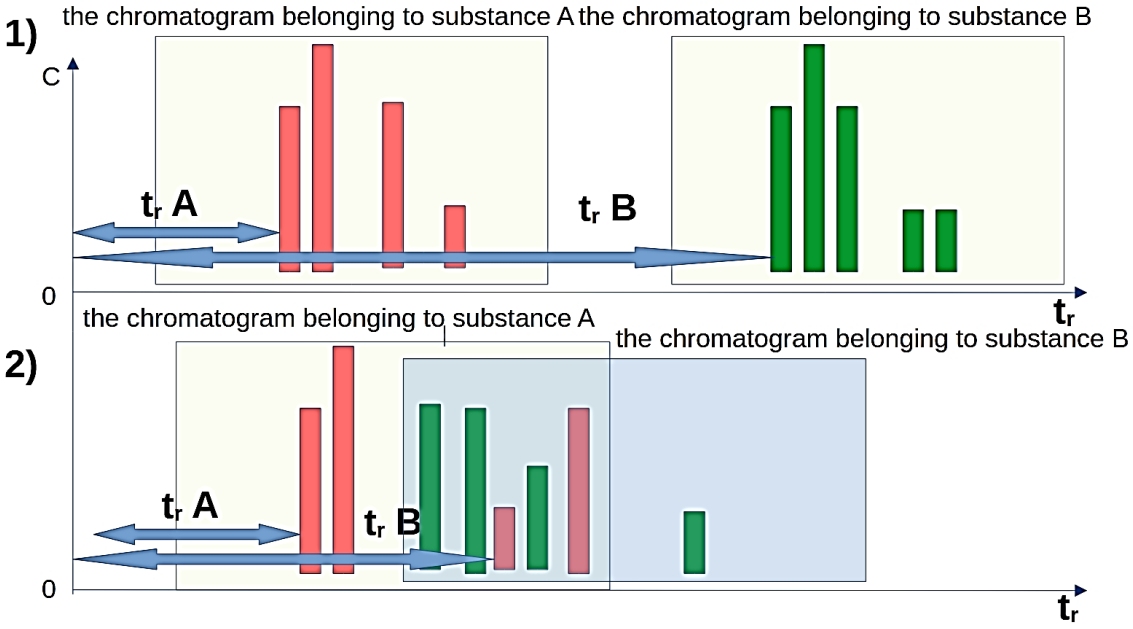

Figure 7. The phenomenon of overlapping retention times.

At this point, an analysis of the functioning of the algorithm will be carried out, taking into account the problem of selecting the stationary phase for a given set of input data vectors, for this purpose the following notations will be introduced (Learning by Simulations…, n.d.).

As the presented formula shows, the scalar product of two vectors is calculated. The more similar the vectors are to each other, the greater the value of the calculated product, the greater the retention time for a given substance. The description of the algorithm and the drawing above show that the correctness of classification is significantly influenced by the distribution of peaks in the chromatogram of the reference substance as well as in the chromatogram of the classified substance. The optimal situation occurs when the distances between individual chromatograms are large or, in other words, the peaks of individual substances do not overlap. The formula describing the distance between the peaks of the chromatogram is presented in formula (13). This formula describes the distance between the i-th and j-th peak.

$$d_{i,j}=\left(tr_i-tr_j\right)^2 \tag{13}$$

Based on the above-mentioned considerations, a criterion for selecting the stationary layer for a given data set can be defined. The structure of the stationary phase - elements of the FS vector should be selected so that for a given data vector the sum of the distances between peaks is the largest, this relationship is expressed by the formula (14).

$$E\left(fs_1, fs_2, fs_3 \ldots fs_M\right) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \square\, d_{i,j} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \square \left(tr_i - tr_j\right)^2 \tag{14}$$

In other words, the elements of the stationary phase should be selected so that the expression described in formula (15) representing the sum of the distances between peaks has the largest value.

$$max\left(E\left(fs_1, fs_2, fs_3 \ldots fs_M\right)\right) \tag{15}$$

To find the maximum of the function, the conditions presented in formulas (16), (17) must be met.

$$\begin{gathered} \frac{\partial E\left(fs_1, fs_2, fs_3 \ldots fs_M\right)}{\partial fs_1} = 0 \\ \frac{\partial E\left(fs_1, fs_2, fs_3 \ldots fs_M\right)}{\partial fs_2} = 0 \\ \ldots \\ \frac{\partial E\left(fs_1, fs_2, fs_3 \ldots fs_M\right)}{\partial fs_M} = 0 \end{gathered} \tag{16}$$

$$\begin{gathered} \frac{\partial^2 E\left(fs_1, fs_2, fs_3 \ldots fs_M\right)}{\partial fs_1^2} < 0 \\ \frac{\partial^2 E\left(fs_1, fs_2, fs_3 \ldots fs_M\right)}{\partial fs_2^2} < 0 \\ \ldots \\ \frac{\partial^2 E\left(fs_1, fs_2, fs_3 \ldots fs_M\right)}{\partial fs_M^2} < 0 \end{gathered} \tag{17}$$

To simplify further considerations and without losing the generality of the conclusions drawn, suppose the stationary phase consists of two elements M = 2 and the number of substances for which we want to calculate the chromatogram is four N = 4, then the expressions presented above will take the following form:

$$FS = \left(fs_1, fs_2\right) \tag{18}$$

$$S = \begin{bmatrix} s_{1,1}, s_{1,2}, \\ s_{2,1}, s_{2,2} \\ s_{3,1}, s_{3,2} \\ s_{4,1}, s_{4,2} \end{bmatrix} \tag{19}$$

$$E\left(fs_1, fs_2\right) = \sum_{i=1}^{4} \sum_{j=i+1}^{4} \square\, d_{i,j} \; d_{1,2} + d_{1,3} + d_{1,4} + d_{2,3} + d_{2,4} + d_{3,4} \sum_{i=1}^{4} \sum_{j=i+1}^{4} \square \left\{F_{tr}\left(s_{i,1..M}, FS\right) - F_{tr}\left(s_{j,1..M}, FS\right)\right\}^2 \tag{20}$$

The function that we maximize for a given input set does not have a maximum. A graph of this function for an example data set is shown below.

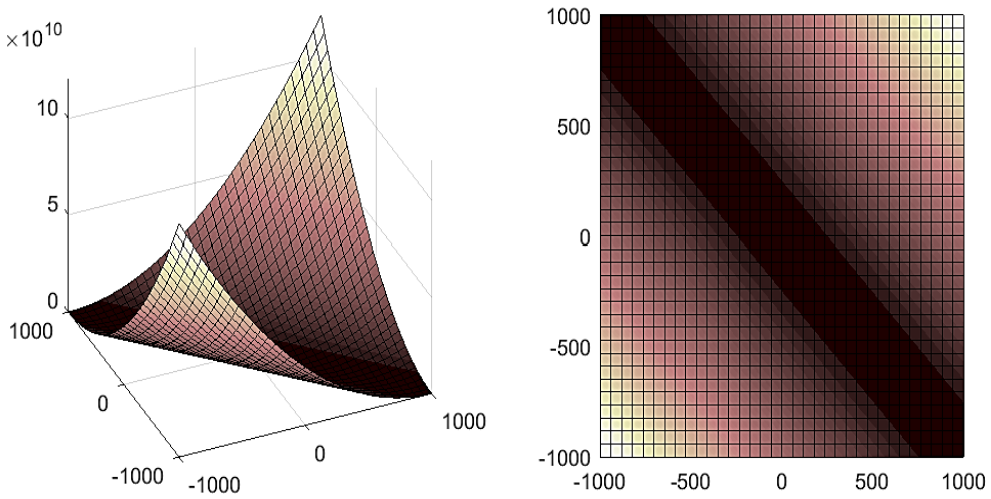

Figure 8. Graph of the maximized function from formula (13) depending on the values of the elements of the stationary phase.

The presented graph shows that the function defined by formula (14) or (20) does not have a maximum in the function responsible for determining the retention time (4), but if the elements of the stationary phase vector have the same sign, the value of function (20) is not limited. This means that the distances between individual peaks will increase proportionally as long as the values of the stationary phase elements increase and if all the stationary phase elements have the same sign. The existence of such a relationship is beneficial, but if we classify highly distorted vectors, the distances between the peaks of the classified chromatogram may differ significantly from the peaks of the chromatogram of the reference vector, which will result in incorrect classification. In this case, replace functions (5) ith a non-linear function.

4. Classification of selected data sets

This chapter will present the classification results for two types of data sets, namely for single-class sets and the second type of classification whose results will be presented is the classification of a multi-class set. Both in the case of the first and the second type of classification, the classification will be performed on files that are in a generally available repository and that were used in the process of testing other classification algorithms. It seems that the above conditions are met by the data sets made available on the UCI Repository website.

The following data sets were selected for the tests. The first set of Thyroid Diseases is related to the medicine and diagnosis of diseases related to the thyroid gland. The remaining two datasets – Landsat and List satellite data – concern image recognition. These datasets were

downloaded from the publicly available UCI Machine Learning Repository (Home – UCI…, n.d.). Detailed information about these datasets is provided. Statistics for the selected datasets are presented in the table below.

Table 1.
Statistics of the test datasets

Data Set	Dimensionality	Number of classes	No of train samples	No of test samples
Thyroid	21	3	3772	3428
Landsat Satellite	36	6	4435	2000
Letters	16	26	15000	5000

All calculations and implementation of individual algorithms were carried out using the MATLAB computing environment.

4.1. Classification of a homogeneous and one-class data set

As can be seen from the presented calculation results, the algorithm presented on sample single-class data sets does not differ significantly from other algorithms. The results achieved are average, but it should be emphasized that these are single-class sets with a very small number of attributes. The classified vectors have 16, 36 and 21 attributes respectively.

Table 2.
Percentage of correct classifications for the Thyroid Disease Data Set (Swiecicki, n.d.)

Algorithm	**%Test**
CART tree	99.36
SSV tree	99.33
MLP+SCG, 4 neurons	99.24
SVM Minkovsky kernel	99.18
MLP+SCG, 4 neurons, 45 SV	98.92
FSM 10 rules	98.90
MLP+SCG, 12 neurons	98.83
Cascade correlation	98.5
MLP+backprop	98.5
SVM Gaussian kernel	98.4
k-NN, k = 1, 8 features	97.3
Naive Bayes	96.1
SVM Gauss, C = 1 s = 0.1	94.7
Chrom	**94.6**
1-NN Manhattan,	93.8
SVM lin, C = 1	93.3

Table 2 shows the classification results of the Thyroid Disease set using the chromatographic algorithm. As shown in the table, the presented algorithm does not differ from other algorithms in terms of classification quality.

Table 3.
Percentage of correct classifications for the Landsat Satellite data Set (Swiecicki, n.d.)

Algorithm	**%Test**
MLP, 36 nodes, +SVNT	91.3
MLP, 36 nodes,	91.0
kNN, k = 3, Manhattan	90.9
FSMneurofuzzy, learn 0.95	89.7
kNN, k = 1, Euclidean	89.4
SVM Gaussian kernel	88.4
RBF, Statlog result	87.9
Chrom	**87.8**
MLP, Statlog result	86.1
Bayesian Tree	85.3
C4.5 tree	85.0
SSV tree	84.3
Cascade	83.7
LDA Discrim	82.9
Kohonen	82.1
Bayes	71.3

Table 3 shows the classification results of the Landsat Satellite data Set using the chromatographic algorithm. Similarly to the previous case. As can be seen from the table presented, the presented algorithm does not differ from other algorithms in terms of classification quality. In order to improve the quality of classification in this case, it would be necessary to consider changing the level of fragmentation (Algorithm 1**.**).

Table 4.
Percentage of correct classifications for the Letter Recognition Data Set (Swiecicki, n.d.)

Algorithm	**%Test**
Chrom	**94.10**
ALLOC80	93.60
K-NN	93.20
LVQ	92.10
Quadisc	88.70
CN2	88.50
Bayesian Tree	87.60
NewId	87.20
IndCART	87.00
C4.5	86.80
DIPOL92	82.40
RBF	76.60
Logdisc	76.60
Kohonen	74.80
Backprop	67.30

In the case of classification using the Letter, the presented algorithm turned out to be the best.

The sets from the point of view of the presented algorithm are not well selected - because they have relatively few attributes, but there is a large set of publications presenting the results of the classification of these sets using various classification algorithms.

4.2. Principle of multiclass vector classification

The basic property of the presented algorithm is the ability to classify data containing data from several classes at the same time, i.e. the so-called multi-class vectors. To illustrate this case, assume that we have a vector containing data belonging to two classes, e.g. data belonging to class 1 and class 2. Then the chromatogram of such a vector will look as shown in Figure 9.

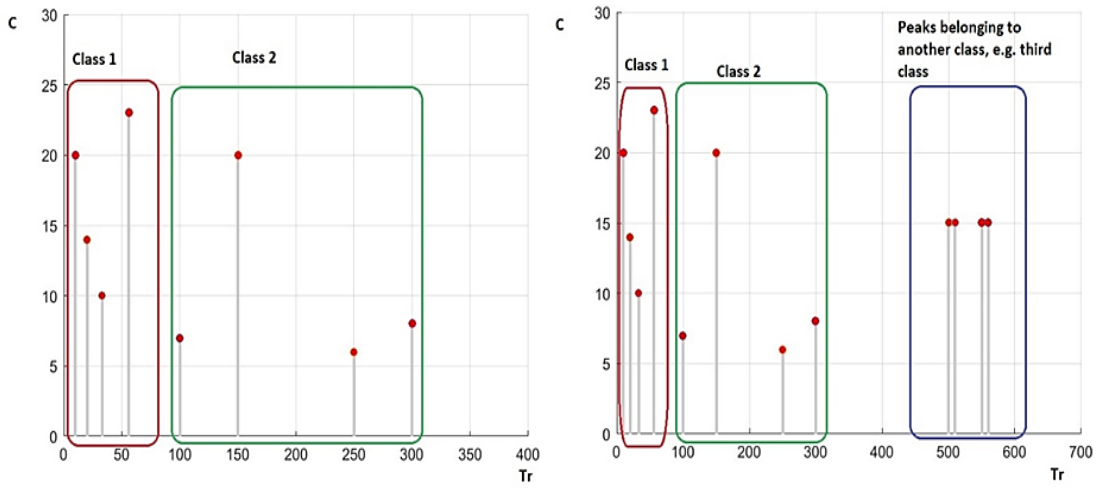

Figure 9. Chromatograms of vectors belonging to a) two classes: b) to three classes.

As can be seen from the figures presented, the chromatograms contain peaks specific to chromatograms belonging to class 1 and class 2. After adding peaks that belong to class, e.g. class 3, we do not lose information about peaks belonging to classes 1 and 2. All peaks should be correctly identified.

Thanks to the proposed data processing technique, it is possible to classify cases in which the classified vector belongs to several classes at the same time.

4.3. Classification of multi class vectors

Using the letter data set, a multi-class set was created - one contains fragments belonging to five classes. The task of the algorithm was to answer the question into what classes a given input vector could be classified.

This means that a given input vector may contain any combination of data for classification, which may belong to a given number of classes. For a single class, it contains pieces of data that may belong to a single class. However, if the number of classes is equal to two, then the input vector that was processed by algorithm 1 may contain data that may belong to at most two classes.

Table 5.
Classification results using a multi-class set chromatographic algorithm

Fragmen-tation	Number of Classes									
	1		2		3		4		5	
	NoClass	C	NoClass	C	NoClass	C	NoClass	C	NoClass	C
16	92.4	92.4	73.3	70,2	47.9	44,2	26.7	24.7	11.5	10.1
8	90.4	90.4	70.3	68.4	45.9	40,4	25.7	21.7	10.0	8.2
4	89.4	89.4	69.3	65.1	44.9	40.0	24.7	20.4	9.8	7.3

Table 5 presents the classification results of a multi-class set created from the Letter set. The presented results show the percentage of correct classifications. Each column related to the number of classes has been divided into two columns - the first column called *NoClass* - gives the percentage correctness of the classification in terms of quality, i.e. whether fragments belonging to the appropriate classes have been correctly identified in a given input vector. The second column called *C* - concentration informs how many of these fragments have been correctly identified, i.e. it provides quantitative information

$$C_{CorrectPercent} = 100 \cdot \left(1 - \frac{\left| \sum_{i=1}^{NumberOfClass} C^{out}{}_i - C^{pattern}{}_i \right|}{\sum_{i=1}^{NumberOfClass} C^{pattern}{}_i}\right) \tag{21}$$

In order to assess the correctness of the classification in quantitative terms, an indicator was defined that was used to evaluate this type of classification. This indicator is presented using a formula. The $C_{pattern}$ vector contains information about the actual number of vector fragments in the vector that is subjected to the classification process. Whereas the C_{out} vector contains information about the number of vector fragments identified by the classifier. The $C_{CorectPercent}$ indicator gives the percentage of correct classifications for a given input vector out of the number of fragments of subvectors belonging to particular classes - i.e. concentrations.

As the table shows, providing quantitative information is less accurate compared to the situation when we evaluate the classifier in qualitative terms.

As the results presented in the table show, the presented algorithm correctly identified three classes - the percentage of correct answers was over fifty percent. However, when indicating the remaining classes to which fragments of the input vector belong, the number of correct answers was no longer satisfactory.

Nevertheless, it should be noted that the difference between the correct answers of the quantitative classifier and the qualitative answers of the classifier was not that large – it was several percentage points, and the classifier provided quantitative information, which is quite important in various decision-making processes.

5. Conclusions

The article presents an innovative method of data processing similar to chromatographic data separation in analytical chemistry. Three algorithms have been proposed, the aim of which is to convert the data vector into a set of mixtures and classifications of individual chemical substances, which was created in the process of transforming the input vector. The paper presents results for well-known sets such as Thyroid, Landsat Satellite and

Letters. As shown, this data classification technique will be suitable for multi-class sets, i.e. those containing fragments in which vectors contain fragments of data belonging to several classes at the same time. The work also shows that using the proposed algorithms it is possible to obtain information about the classified vector not only of a qualitative nature, but also that the presented classification technique provides quantitative information.

The article presents an algorithm for the separation of chromatographic data, which was inspired by one of the methods of analytical chemistry, which is resolution chromatography. Three algorithms were proposed for chromatographic data separation, which constitute the chromatographic data separation process. Algorithm 1, whose task is to transform a set of input vectors into a set of mixtures of substances. Algorithm 2, the algorithm responsible for calculating the retention time. The third algorithm is responsible for assigning the spectrum of the unknown substance, i.e. the input vector, to the chromatograms of the reference vectors. As shown in all the above-mentioned types of datasets, the proposed classification mechanism performed relatively well. In order to improve the classification efficiency of the presented mechanism, it would be necessary, first of all, to algorithmize the problem of stationary phase selection taking into account nonlinear functions.

Based on the presented results, it can be assumed that the chromatographic data separation technique can be successfully used in the processing of large data sets, where the data do not always have such features as a constant vector length, a relatively small number of elements in the vectors and are heterogeneous.

References

1. Blumberg, L.M. (2021). Theory of gas chromatography. *Gas Chromatography*, pp. 19-97. Retrieved from: https://doi.org/10.1016/B978-0-12-820675-1.00026-5.
2. Brabazon, A., O'Neill, M., Mcgarraghy, S. (n.d.). *Natural Computing Series Natural Computing Algorithms*. Retrieved from: www.springer.com/series/, 6.02.2024.
3. Giddings, J.C. (2017). Dynamics of chromatography: Principles and theory. *Dynamics of Chromatography: Principles and Theory*, pp. 1-323. Retrieved from:

https://doi.org/10.1201/9781315275871/DYNAMICS-CHROMATOGRAPHY-CALVIN-GIDDINGS.

4. Hage, D.S. (1999). *Affinity Chromatography: A Review of Clinical Applications*. Retrieved from: https://academic.oup.com/clinchem/article/45/5/593/5643177.
5. Hilbert, M. (2016). Big Data for Development: A Review of Promises and Challenges. *Development Policy Review*, *34(1)*, pp. 135-174. Retrieved from: https://doi.org/10.1111/DPR.12142.
6. Hohrenk, L.L., Itzel, F., Baetz, N., Tuerk, J., Vosough, M., Schmidt, T.C. (2020). Comparison of Software Tools for Liquid Chromatography-High-Resolution Mass Spectrometry Data Processing in Nontarget Screening of Environmental Samples. *Analytical Chemistry*, *92(2),* pp. 1898-1907. Retrieved from: https://doi.org/10.1021/ACS.ANALCHEM.9B04095/ SUPPL_FILE/AC9B04095_SI_001.PDF.
7. *Home – UCI Machine Learning Repository* (n.d.). Retrieved from: https://archive.ics.uci.edu/, 3.06.2024.
8. *Learning by Simulations: Overlapping Peaks* (n.d.). Retrieved from: https://www.vias.org/simulations/simusoft_peakoverlap.html, 6.02.2024.
9. Leweke, S., von Lieres, E. (2018). Chromatography Analysis and Design Toolkit (CADET). *Computers & Chemical Engineering*, *113*, pp. 274-294. Retrieved from: https://doi.org/10.1016/J.COMPCHEMENG.2018.02.025.
10. Pezzatti, J., Boccard, J., Codesido, S., Gagnebin, Y., Joshi, A., Picard, D., González-Ruiz, V., Rudaz, S. (2020). Implementation of liquid chromatography–high resolution mass spectrometry methods for untargeted metabolomic analyses of biological samples: A tutorial. *Analytica Chimica Acta*, *1105*, pp. 28-44. Retrieved from: https://doi.org/10.1016/J.ACA.2019.12.062.
11. Pierce, K.M., Trinklein, T.J., Nadeau, J.S., Synovec, R.E. (2021). Data analysis methods for gas chromatography. *Gas Chromatography*, pp. 525-546. Retrieved from: https://doi.org/10.1016/B978-0-12-820675-1.00007-1.
12. *Principle and Procedure of Affinity Chromatography* (n.d.). Retrieved from: https://whatishplc.com/hplc-basics/affinity-chromatography/, 6.02.2024.
13. Reinsel, D., Gantz, J., Rydning, J. (2017). *Data Age 2025: The Evolution of Data to Life-Critical Don't Focus on Big Data; Focus on the Data That's Big.* Sponsored by Seagate. Retrieved from: www.idc.com.
14. Robards, K., Ryan, D. (2021). Principles and Practice of Modern Chromatographic Methods. *Principles and Practice of Modern Chromatographic Methods*, pp. 1-518. Retrieved from: https://doi.org/10.1016/B978-0-12-822096-2.09993-X.
15. Stilo, F., Bicchi, C., Jimenez-Carvelo, A.M., Cuadros-Rodriguez, L., Reichenbach, S.E., Cordero, C. (2021). Chromatographic fingerprinting by comprehensive two-dimensional chromatography: Fundamentals and tools. *TrAC Trends in Analytical Chemistry*, *134*, *116133*. Retrieved from: https://doi.org/10.1016/J.TRAC.2020.116133.

16. Swiecicki, M. (2009). An algorithm based on the construction of Braun's cathode ray tube as a novel technique for data classification. *Neutral Information Processing, Vol. 5864 LNCS, Iss. PART 2,* pp. *710-719*, *5864 LNCS* (PART 2). Retrieved from: https://doi.org/10.1007/978-3-642-10684-2_79.
17. Święcicki, M. (2024). A Classification Algorithm Inspired by the Chromatographic Separation Mechanism Dedicated to the Classification of Variable-length and Multi-class Vectors. *JOIV: International Journal on Informatics Visualization*, *8(1)*. Retrieved from: https://doi.org/10.62527/JOIV.8.1.2324.
18. Urh, M., Simpson, D., Zhao, K. (2009). Affinity chromatography: general methods. *Methods in Enzymology*, *463*(C), pp. 417-438. Retrieved from: https://doi.org/ 10.1016/S0076-6879(09)63026-3
19. Varhadi, S.D., Gaikwad, V.A., Sali, R.R., Chambalwar, K., Kandekar, V. (2020). *A Short Review on: Definition, Principle and Applications of High Performance Liquid Chromatography. Introduction, 19*(2), pp. 628-634. Retrieved from: www.ijppr.humanjournals.com
20. Witting, M., Böcker, S. (2020). Current status of retention time prediction in metabolite identification. *Journal of Separation Science*, *43(9-10)*, pp. 1746-1754. Retrieved from: https://doi.org/10.1002/JSSC.202000060.