Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/100405

Analysis of Questionnaire Results Using Metric Methods

Maria Rafalak[1,*], *Piotr Bilski*[2] *and Adam Wierzbicki*[1].

[1] Polish-Japanese Academy of Information Technology, ul. Koszykowa 86, 02-008 Warsaw, Poland.
[2] Institute of Radioelectronics, Warsaw University of Technology, ul. Nowowiejska 15/19, 00-665 Warsaw, Poland.

Received: 18 Mar. 2016, Revised: 4 May 2016, Accepted: 5 May 2016
Published online: 1 Jul. 2016

Abstract: The paper presents the application of the metric methods to the analysis of the questionnaires used in various fields. The generic methodology is presented, including particular modules, responsible for the subsequent operations. They include generation of category patterns based on the available data, application of envelopes, dataset complexity assessment and performing classification of questionnaire results. Metrics applied in the presented research are then introduced. The methodology is tested on three data sets from the psychological, sociological and educational domains. Results show the advantage of our approach compared to the standard classification and decision making methods. Also, it may be used for the results interpretation, finding relations in data, or evaluation the test discriminating power (regarding each question separately).Proposed methodology may be found beneficial in all areas where questionnaire data is used - from classical diagnosis to HCI and big-data applications.

Keywords: distance metrics, pattern generation, classification, questionnaire analysis

1 Introduction

Profiling people (users, patients, clients etc.) is nowadays a dynamically developing research area. It consists in observing human reactions on the presented events, or filling the questionnaire (or test) by the monitored person (further also called the testee). Premises used in this process are diversified and vary depending on two factors. The first one is the purpose of the analysis (for instance, personalization of the marketing strategy). The second one is the mode of the analysis: on-line (methods exploiting the Internet) or off-line (traditional approaches, like paper-pencil questionnaires). In general, profiling is based on behavioral data and/or declarations expressed by people during the examination.

Methodology of preparing the questionnaires is well established and multiple standard tools for this purpose exist. Their usage requires from the testee answering sequences of appropriately prepared questions (items). Further it allows for evaluating selected human characteristics. This includes psychological diagnoses, determining political views, professional skills and many other. Contemporary methods of evaluating answers from tests are of limited accuracy, considering mainly summed points, which are the main premise for making diagnostic decision. The structure of the test facilitates the deeper analysis, based on single answers.

The interpretation of answers is currently used in both scientific and practical applications. The rising interest in the Human-Computer Interaction (HCI) and the ability to make user profiles give the opportunity to learn about human habits and preferences. This knowledge may be exploited in the marketing, psychological or educational research and in every other discipline where questionnaires are used. Data gathered through questionnaires is very useful in creating adaptive interfaces, personalizing marketing offers or including in big-data algorithms. Classical approaches for the questionnaire data analysis have several drawbacks. For instance, in the qualitative analysis information about the distribution of answers in population is neglected. On the other hand, in the quantitative analysis the information about the certain answer pattern is lost.

This paper presents the novel methodology for the analysis of questionnaire results exploiting the metrics space concepts. The main operations include clustering of available data into profiles' patterns and calculating distances between them. The architecture is flexible enough to work with questionnaires of different origin. Its advantages include:

* Corresponding author e-mail: maria.rafalak@pjwstk.edu.pl

–The insight into the profiles of the respondents, determining not only the main profiles, but also sub-categories as well
–The ability to perform the classification based on the particular answers, assuming specific questions are of different importance
–Versatility enabling application of different metrics and clustering algorithms
–The ability to evaluate the questionnaire and its discriminating power (ability to differentiate between questionnaire respondents representing different groups)

The paper is organized as follows. In Section 2 the existing methodology for the questionnaires analysis and interpretation is presented. Section 3 contains the overview of the proposed generic architecture able to process the selected questionnaires. Its modules are briefly introduced, including the pattern analysis, distance calculation approaches and classification strategy selection. The mathematical apparatus used for specific calculations is described in Section 4. Section 5 contains details of the particular operations exploited within the architecture. In Section 6 datasets used for the experiments are described. In Section 7 the analysis and verification of the proposed approach is presented. Section 8 contains the experimental results and their discussion. In Section 9 conclusions and possible application of the proposed methodology are provided.

2 Existing methodology for the questionnaire analysis

Questionnaire data can be analyzed either in the qualitative or quantitative manner. To appreciate the power of the proposed approach, both are explained in the following subsections. The innovative approach to the questionnaire data analysis proposed in this paper constitutes their combination. This new method provides the more detailed analysis of obtained questionnaire results. Its benefits are twofold. Firstly, it provides an overview of how certain groups perform in every test item. Secondly, it enables the analysis of the single person's answers in relation to norms (results obtained by certain groups of people).

2.1 Quantitative analysis of questionnaire data

There are two main theoretical frameworks describing the quantitative analysis of questionnaire data: Classical Test Theory (CTT) and Item Response Theory (IRT), as described below.

2.1.1 Classical Test Theory (CTT)

This is one of the earliest conceptual frameworks referring to the questionnaire measurement, formulated by [1], [2]. It assumes that the test score, i.e. the sum of points assigned to answers given by the testee, reflects the intensity of measured trait with precision determined by the error of measurement. Interpretations of the test score can be formulated referring to norms defined as the scores obtained by the other members of population (interindividual perspective) or to other scores obtained by the same testee (intraindividual perspective).

Interindividual perspective

In CTT, there are two approaches to the analysis and interpretation of scores obtained by the testee. Both refer to results observed in the population to which the testee is compared to (interindividual perspective). The first approach is based on properties of the normal distribution (the ”68 - 95 - 99.7” rule) which assign obtained scores into distinct categories [3]. Results distant more than one standard deviation from the mean observed in the population are usually classified as high or low (depending on the direction of the difference). Other scores are classified as average. This approach is usually applied in tests dedicated to normal population when the test giver is interested in describing the testees performance. In CTT, there are two approaches to the analysis and interpretation of scores obtained by the testee. Both refer to results observed in the population to which the testee is compared to (interindividual perspective). The first approach is based on properties of the normal distribution (the ”68 - 95 - 99.7” rule) which are used to assign obtained scores into distinct categories [3]. Results distant more than one standard deviation from the mean observed in the population are usually classified as high or low (depending on the direction of the difference). Other scores are classified as average. This approach is usually applied in tests dedicated to normal population when the test giver is interested in describing the testees performance. The second approach popular in interpreting test scores requires setting a cutoff point that determines membership of the testee to a certain group. The exact cutoff value is usually established using the Receiver Operating Characteristics (ROC) curve or is determined arbitrarily. This approach is mostly applied in clinical or education testing when the test giver is interested if the testee meets necessary qualification criteria.

Intraindividual perspective

Adopting intarindividual perspective is to analyze test scores obtained by the testee in relation to his/her other results. It is sometimes called psychometric profiling aimed at determining strengths or weaknesses of the testee. Psychometric profiling is usually applied when a questionnaire measures construct with several dimensions depicted by scores in different questionnaire scales. Questionnaire constructors establish (using statistical

tests) what is the minimal difference between scale scores that can be interpreted as statistically significant.

2.1.2 Item Response Theory (IRT)

The IRT is frequently used in the computer administered tests.. It weights each question individually during the final score calculation, depending on the questions difficulty and discriminating power. Its disadvantage is considering only the final score in the result interpretation, disregarding responses to particular questions [4].

The general IRT model assumes [5] that the answer given to every item depends on a single latent trait (denoted by θ). In the two-parametric model (2PL) the probability of giving the correct answer to an item X_i is expressed by the following logistic function (so-called Item Characteristic Curve (ICC) Figure 1):

$$P(X_i|\theta) = \frac{1}{1+e^{-a_i(\theta-b_i)}} \quad (1)$$

where a_i is the item discrimination (the ability to differentiate between users showing various θ levels) and b_i is the item difficulty.

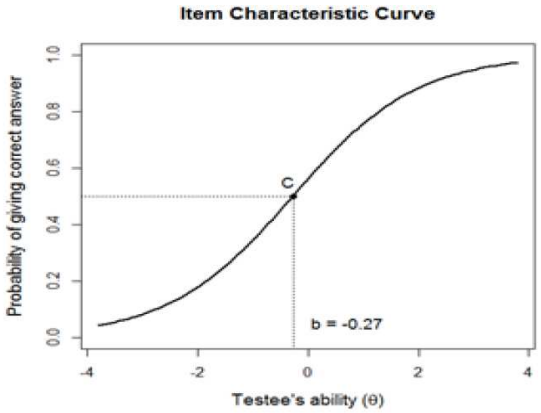

Fig. 1: Sample Item Characteristic Curve

Difficulty of an item in IRT (b_i) is defined as the ability level presented by the testee to have 50% chance for giving the correct answer to this item. In the example from Figure 1 the item difficulty is equal to -0.27. The testees ability depicted on the x-axis is expressed by a standardized scale (with zero mean and standard deviation equal to 1). The item discriminating power (a_i) in IRT is defined as the tangent of the curve slope measured in the point C.

More complicated models include additional parameters (like guessing 3PL or the upper limit asymptote 4PL; [6]) or allow for analyzing data with different scoring schemes (graded response model, partial credit model; [4] [7]. The test score in IRT is the estimation of the testee final θ level.

2.2 Qualitative analysis of questionnaire data

Qualitative research is used mainly in exploratory studies where the exact structure of the measured construct is not yet known. The data analysis (in general) is usually connected with applying specially designed methods like open-ended questions, interviews or observations. Collected answers given by the respondent are classified and coded by a trained professional. Interpretation of results obtained by an individual qualitative research highly depends on the theoretical framework adopted in certain research area (i.e. psychoanalysis) and rarely relies on the reference data.

The qualitative analysis of close-ended questionnaire items is definitely less common. This approach is used mainly in marketing research, therapy or education. The general idea behind it is to evaluate the specific answer selected by the testee and try to give them some additional meaning [8]. Testee's answers to particular items are treated as a starting point for further discussion with the researcher, therapist or teacher. Again, the final interpretation of the test result is often strongly determined by the theoretical framework adopted by the specialist. It is based rather on the interpreters professional experience than on research data. Therefore, such an approach to questionnaire data analysis often leads to subjectivity in final judgments.

2.3 Computer-aided analysis of questionnaire data

Computer-aided analysis of questionnaire data allows for applying advanced computational methods for the testee classification. Attempts to use decision trees [9] [10], fuzzy decision trees [11] or random forests [12] for this purpose were moderately successful. Other artificial intelligence methods like artificial neural networks [13] [14] were applied for the testee profiling and prove to give better classification results than the classical threshold-based procedure. There are also systems like Copernicus [15] that integrate several classification techniques to increase the classification accuracy based on the questionnaire data. Despite the fact that these solutions give accurate decisions, rules leading to this decision are usually too complex to interpret by the human. In practical applications (i.e giving psychological or educational diagnosis) methods that provide easily understandable criteria for decision making are preferable, even at the cost of the weaker classification accuracy. Teacher, psychologist, sociologist or any other decision making professional must be able (based on

questionnaire results) to justify their judgment and explain it to the testee or authorities especially when making high-stake decisions. Computer-aided analysis of questionnaire data allows for applying advanced computational methods for the testee classification. Attempts to use decision trees [9] [10], fuzzy decision trees [11] or random forests [12] for this purpose were moderately successful. Other artificial intelligence methods like artificial neural networks [13] [14] were applied for the testee profiling and prove to give better classification results than the classical threshold-based procedure. There are also systems like Copernicus [15] that integrate several classification techniques to increase the classification accuracy based on the questionnaire data. Despite the fact that these solutions give accurate decisions, rules leading to this decision are usually too complex to interpret by the human. In practical applications (i.e giving psychological or educational diagnosis) methods that provide easily understandable criteria for decision making are preferable, even at the cost of the weaker classification accuracy. Teacher, psychologist, sociologist or any other decision making professional must be able (based on questionnaire results) to justify their judgment and explain it to the testee or authorities especially when making high-stake decisions.

2.4 Problem statement

Both qualitative and quantitative approaches to questionnaire data analysis have their drawbacks. In the former, the test score (reflecting the sum of points assigned to answers given by the testee) is interpreted. Unfortunately, the information about the detailed answer pattern is lost. Figure 2 shows sample answers given by two testees. Both persons obtained the same test score (28 points) but gave different answers to every question. Using only CTT to data analysis, results of testee1 and testee2 are indistinguishable. Applying IRT algorithms may result in respondents differentiation (depending on item characteristics) regarding only the final test score. However, the answer patterns are still lost.

On the other hand, qualitative analysis is subjective and neglects the information about the average results obtained in the population. The methodology described in this paper tries to overcome weaknesses of each approach applied separately and give the new quality to the profiling of people based on the questionnaire data. Unlike other computer-aided methods for the questionnaire data analysis, it gives easily interpretable results and enables to understand the process underlying the final classification decision.

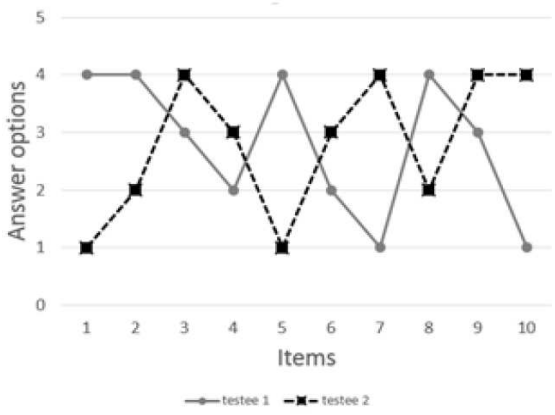

Fig. 2: Sample answers given by two testees in a questionnaire

3 Overview of the proposed generic architecture

The proposed methodology is the generic framework aimed at the analysis of questionnaire results. Current approaches to make decision about which category testee belongs to calculate the overall number of points obtained in the test. We believe such an approach in some cases may be too general and gives little space to interpretation based on the particular responses. Therefore the proposed method focuses both on the test score and the responses to subsequent test items. This allows for making decision about the category of the analyzed person and observing the pattern of his/her responses. The idea is to represent each questionnaire result as the vector of responses to particular questions. This way it is possible to calculate answer patterns characteristic to the particular group of persons. Additionally, the envelope for each pattern is generated, which considers not only the ”mean” responses, but also their distribution for the set of analyzed persons (reference groups). The expected benefits include the greater diagnosis accuracy, the ability to assess the quality of the questionnaire or the ability to track patterns in responses. The block scheme of the framework is presented in Figure 3.

The structure of the system contains the main implemented operations, required for the analysis and decision making. The architecture is generic enough to work with the unlabeled and labeled data. Knowledge extracted during the training stage may be exploited to the qualitative or quantitative analysis of the questionnaire itself or the respondents. The specific application of the system is the classification of the testee to one of the categories based on his responses to the questionnaire (which is represented by the ”decision making” module in Figure 3). The required input is the set of questionnaires filled by the respondents. Each questionnaire is the set of m questions $Q = \{\boldsymbol{q}_1, \cdots, \boldsymbol{q}_m\}$, where the specific ($j$-th) question is represented by the set of z discrete responses:

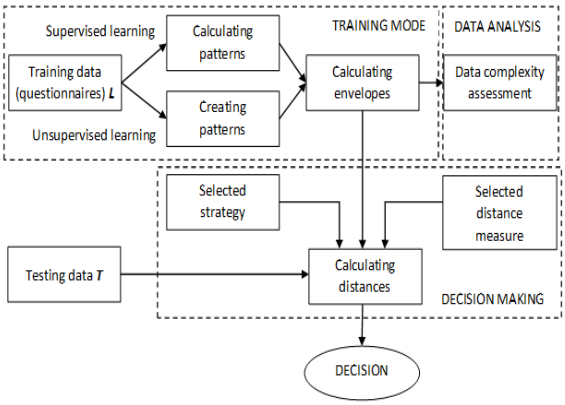

Fig. 3: Framework of the questionnaire interpretation system

$\boldsymbol{q}_j = \{q_j^1, ..., q_j^z\}$. All possible questionnaire results form the Carthesian product $c_p = q_1 \times q_2 \times \cdots \times q_m$. The execution of Q on the testee leads to the single combination from c_p. As knowledge about the discriminating power and characteristics of the test is collected from multiple executions, in the presented work Q is represented by the training set $\boldsymbol{L}$. It consists of n vectors (integer values representing responses to subsequent questions by the particular testee). In each column the number $v \in (1, \cdots, z)$ of the response option to the j-th question by i-th person $q_{ij}^v (i \in (1, \cdots, n), j \in (1, \cdots, m))$ is stored (starting from 1 up to the number of possible answers to the question z). For example, if the set of responses to the specific question contains the following values: ”very rarely”, ”rarely”, ”moderately”, ”often” or ”very often” (ordinal scale), the column corresponding to this question will contain the following values: ”1”, ”2”, ”3”, ”4” or ”5”. Assuming the questionnaire contains ten questions (each with five responses), the form of the training example is as follows (where the first occurrence of ”1” means that the analysed person provided the response option number 1 to the first question, while the value of ”5” is for the fifth response option provided for the second question):

$$\mathbf{e}_i = [1533212412]$$

$$\mathbf{L} = \begin{bmatrix} \mathbf{e}_1 \\ \vdots \\ \mathbf{e}_n \end{bmatrix} = \begin{bmatrix} q_v^{11} & \cdots & q_v^{1m} \\ \vdots & \ddots & \vdots \\ q_v^{n1} & \cdots & q_v^{nm} \end{bmatrix} \quad (2)$$

The training set $\boldsymbol{L}$ may be supplemented with the information about the category of each example. This additional column c allows for the supervised learning during the training mode. Otherwise, the unsupervised learning is only possible. In the following subsections the subsequent operations are described in detail.

Results of the typical questionnaire can be represented both as the time series, where the sequence of questions matters, or as the point in the m-dimensional space, where the sequence is not important (giving more freedom to build and interpret test results). In our approach we assume the sequence of questions is relevant. In this case the shape of the function created by the responses of the analyzed person (Figure 4) depends on the position of questions in the questionnaire. This allows for introducing multiple distance measures to determine the similarity between examples. Alternatively, the example may be represented as the vector in this space, starting in the beginning of the coordinate system and ending in the coordinates indicated by the subsequent responses to questions.

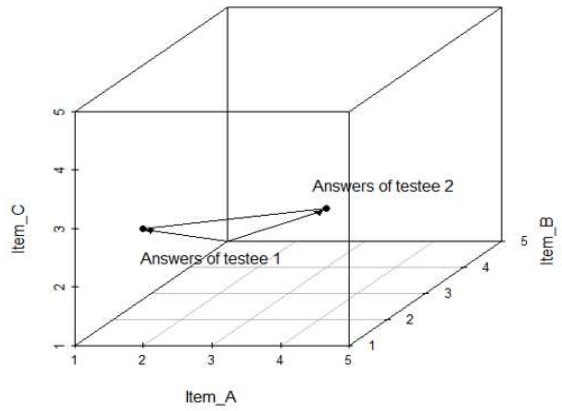

Fig. 4: Geometric interpretation of the questionnaire responses (three items)

The initial analysis of the training data consists in clustering examples similar to each other, which leads to the nominal patterns, representing subsequent groups (see: Section 5.1). This way determining the category of the actual example (a set of responses from a single person) requires calculating distances between the example and all group patterns. To increase the classification accuracy (measured as the overall number of correctly classified examples related to the cardinality of the testing set), the whole distribution of responses to the particular question by every group should be considered. Therefore, the envelopes for each question are generated, modifying the distance between the example and the category pattern by creating attraction areas and pulling the analyzed example towards the center of the answer distribution obtained for analyzed groups (see: Section 5.2). The use of envelopes is optional and can be introduced if there is the chance to improve the decision accuracy.

4 Mathematical apparatus

This section presents the metrics-based methodology used in the architecture from Figure 3. The particular distances are briefly introduced. Their application includes the classification of the selected testee and the questionnaire difficulty assessment. The distance can be calculated for the whole questionnaire (considering all dimensions of the point in Figure 4) or for the single question. In the first case, the overall similarity between the patterns is obtained. In the second case, the influence of each question on the classification can be individually modeled. For instance, during the classification of the sequence of responses to one of available patterns the overall distance may be calculated as the sum of distances between the particular coordinates. Alternatively, the distance may be expressed by the number of coordinates, for which the test result is closer to the selected pattern.

4.1 Histograms

In the analysis of questionnaire data histograms reflect the frequency of certain answers given by a group of testees to particular questionnaire items. The sample histogram of answers given to a questionnaire item having five response options is depicted in Figure 5.

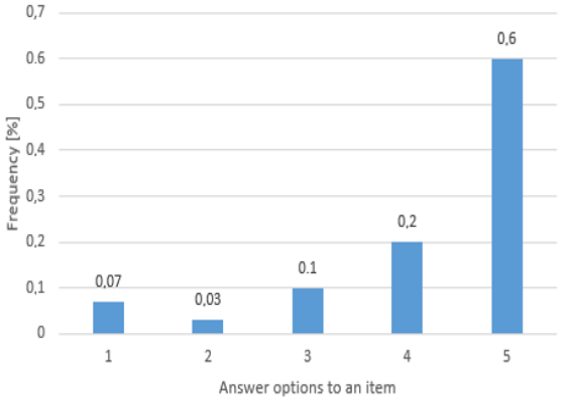

Fig. 5: Illustration of the response distribution for the selected question

The height of each (v-th) bar ($\widehat{q_{kj}^{v}}$) depends on the percentage of persons (assigned to the same k-th category) giving the particular response v (note that $\sum_v \widehat{q_{kj}^{v}} = 1$). Figure 6 shows that respondents from the considered group mostly give responses number v=4 and v=5, but other options are also present.

As the analyzed data sets contain the predefined number of object categories (for instance, ”depression”/”healthy”), histograms are created for every question and category separately. In the Psychology data set, for each question two histograms will be created: one for the healthy, the second for the individual with depression. For the Sociology data set, there will be three histograms for each question and so on.

In this paper histograms are used for creating envelopes that should improve testee classification in the case of the skewed distribution of responses given by the testees representing specific categories.

4.2 Implemented distance metrics

Among multiple options, the following metrics were proposed for calculating distances between the response patterns. The selected measures are well established in multiple domains and are easily interpreted in the geometric spaces.

– **Euclidean distance** [16] (3), which is the most popular approach in the geometrics spaces. All coordinates (i.e. responses to the particular questions) of the pattern $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ are treated equally and have the same impact on the overall distance. The weighting of the particular coordinates is possible, but in our approach we assume no information about the importance of the particular questions is known. If the distance is used to calculate the overall similarity between objects, its form is as follows:

$$d_E(\mathbf{e}_1, \mathbf{e}_2) = \sqrt{\sum_{j=1}^{m} (q_{1j} - q_{2j})^2} \qquad (3)$$

When each coordinate is treated separately, the distance is calculated for m=1.

– **Mahalanobis distance** [17] (4), which is the way to measure the distance between the response q_{ij} given by i-th person to item j the reference distribution y_j (Figure 6) of all obtained responses to the j-th item [17]. It is defined as:

$$d_{M1}(q_i j, y_j) = \sqrt{(q_i j - \bar{y_j})^T S_{y_j}^{-1} (q_i j - \bar{y_j})} \qquad (4)$$

where S_{y_j} stands for the covariance matrix of y; $\bar{y_j}$ stands for the mean of the distribution $\boldsymbol{y}$. Similarly to the Euclidean distance, (4) can be calculated for the single question, or for the whole questionnaire. In the latter case, partial distances (4) are summed. The Mahalanobis distance is popular in the cluster analysis. When considering N categories in the supervised learning, the minimal distance between the point and the distribution representing a class determines its membership.

For the purpose of assessing data complexity (see: Section 5.5) the additional distance was introduced:

–**Earth-Mowers Distance (EMD)** (5) is a metric between two distributions. It is based on the solution to the transportation problem described in the linear optimization domain [18]. In general, EMD reflects the minimal cost required to transform one distribution into another, where $C_1 = \{(q^{\widehat{v}}_{jk_1}, v)\}_{k_1}^{n_1}$ and $C_2 = \{(q^{\widehat{v}}_{jk_2}, v)\}_{k_2}^{n_2}$ are two distributions of size n_1 and n_2 respectively, with $q^{z}_{jk_1}$ and $q_{(}jk_2)^z$ being the probabilities of observing answer options v for the j-th item in testee groups k_1 and k_2. If C_1 is treated as supplies and C_2 as demands, a flow $f_{k_1k_2}$ reflects the amount transported from supply k_1 to demand k_2. The EMD is defined by:

$$EMD(C_1, C_2) = \min \frac{\sum_v f_{k_1k_2} d_{k_1k_2}}{\sum_v f_{k_1k_2}} \quad (5)$$

where $d_{(k_1k_2)}$ stands for ground distance between location v_{jk_1} and v_{jk_2}. The general illustration of EMD is depicted in Figure 6.

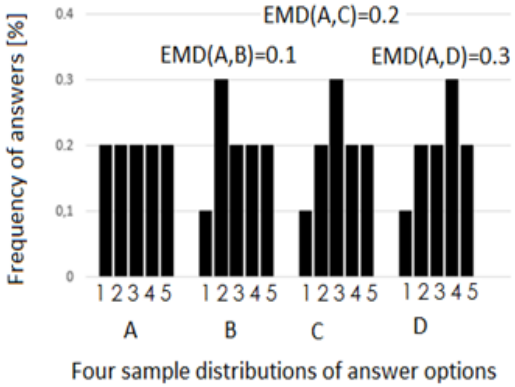

Fig. 6: Illustration of EMD distance for sample distribution of answers for a questionnaire having five answer options

In the example from Figure 7 there are four distributions (*A*, ..., *D*) reflecting proportions of answers to a questionnaire item with five response options $v = \{1,2,3,4,5\}$. To transform the distribution *A* into the distribution *C*, 0.1 of the distribution mass needs to be transported from $v = 1$ to $v = 3$. In this case $d_{k_Ak_C} = 2$, therefore EMD(*A*,*C*)$= 0.1 \cdot 2 = 0.2$.

4.3 Selected measure of central tendency

Every distribution can be characterized by the descriptive statistics. Information about the central position within the set of data is reflected by the measures of central

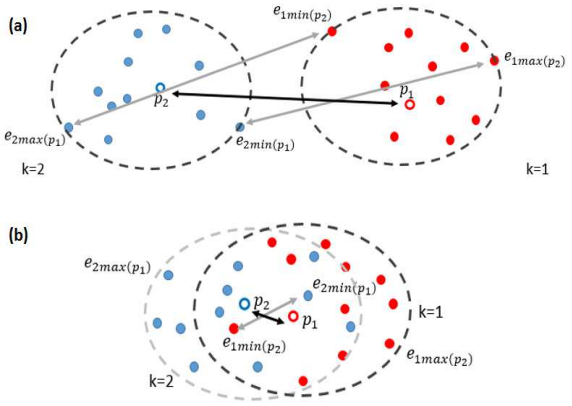

Fig. 7: Illustration of the data complexity assessment based on two categories where (a) subsets do not overlap (simple case) and (b) subsets overlap (difficult case).

tendency. The most common measures are *mean* and *median*. The former is defined by the sum of all values present in the dataset divided by the number of observations. The latter is the value that separates distribution of frequency (see: Figure 5) into two halves. Its main advantage is that it is takes the value from the set of possible answer options v and therefore is easier to interpret when analyzing discrete datasets. The median is also resilient to the extreme observations present in the dataset as long as they do not occur frequently [19]. However, in the testee classification and interpretation of questionnaire results, extreme observations are important. They give the interpreter or questionnaire constructor the general orientation about the maximal and minimal scores obtained in the analyzed testee group. Therefore, for the purpose of this paper the mean value of scores obtained by such a group was used for implementation of the proposed questionnaire data analysis framework.

5 Detailed questionnaire data analysis operations

This section presents the detailed operations executed in the architecture from Figure 3. It is specified, where the particular measures were applied.

5.1 Patterns extraction

The multiple examples in the set $\boldsymbol{L}$ are processed to extract the predefined number of response patterns, representing particular classification categories. The process depends on the training mode applied to the data. In both cases the operation of calculating the pattern $\boldsymbol{p}_k$ representative to the *k*-th category is performed. It is the

vector of means $\overline{q_{kj}}$ for particular responses to the subsequent questions (indexed by j):

$$\boldsymbol{p}_k = [\overline{q_{k1}} \cdots \overline{q_{km}}] \tag{6}$$

Note that while all data in $\boldsymbol{L}$ are discrete, the pattern may contain real values, which do not correspond to any response to the question, but represent the distribution of all possible responses.
For instance, if the following examples belong to the same category:

$$\begin{bmatrix} 1\ 2\ 3\ 4\ 1\ 2\ 1\ 2\ 1\ 5 \\ 1\ 1\ 3\ 1\ 3\ 2\ 5\ 3\ 1\ 4 \\ 2\ 2\ 3\ 1\ 2\ 2\ 4\ 1\ 3\ 5 \end{bmatrix}$$

the pattern for them is as follows:

$$p_k = \left[1.33\ 1.66\ 3\ 2\ 2\ 2\ 3.33\ 2\ 1.66\ 4.66\right]$$

The details of the pattern calculation depend on the training mode, i.e. the method of selecting the examples to the particular set, for which the pattern is calculated, as discussed below.

Supervised learning
In this mode all examples have already assigned categories [20] derived from non-questionnaire sources (for example, the political preference as the ”democrat”, ”republican” or ”independent”). This way every example has the following form, where positions in the subsequent columns include integer values:

$$\mathbf{e}_i = \left[q_{i1} \cdots q_{im}\ c_i\right] \tag{7}$$

In this context, training consists in finding the most characteristic pattern for each category. The examples for which the specific pattern is calculated, have known category (for example, because all respondents revealed their actual preferences).

Unsupervised learning
In this mode the set of examples, for which the pattern is calculated, is generated using the similarity between the particular vectors. This is the task of data clustering, i.e. generation of categories. In this case two problems must be solved:

– Selection of the clustering algorithm. Among multiple choices (such as the Nearest Neighbor, graph or conceptual clustering [21], fuzzy k-Means, etc.) the most suitable ones for the task should be compared and applied to the scheme. Because all sets presented in the paper are labeled, this part of the framework is not explained further and will be the aim of the future research.
– The number of generated categories. This is the typical problem in the clustering of the unlabeled data [22]. One of the approaches is predefining this parameter, as the algorithm works until all examples are assigned to any category. The more attractive approach generates the categories adaptively and analyses variability in data. This way the number of generated categories may be greater, but after the additional analysis by the operator, some categories may be merged into one.

After creating the categories, all sets of grouped examples have the pattern calculated, as explained in Section 5.1. Due to the character of the analyzed datasets, in this paper only supervised learning has been used. However, in other applications unsupervised learning can also be successfully applied.

5.2 Generation of envelopes

The patterns themselves may be enough to assign new data to any of available categories. In the general approach it is assumed that responses to every question form the normal distribution. In such a case, the mean value is a good representation of the most common answer. If the distribution is skewed, other measures are required to correctly classify the example. Therefore the additional envelope is created for each pattern. It is the area representing the distribution of all responses to the selected question. In the correctly designed test, the distribution of responses should be skewed towards the most common one for the particular category. Therefore the simple mean value calculation does not give the information about the distribution, being accurate only for the symmetric Gaussian one. In the correctly designed questionnaire, every question should maximally separate all considered categories. However, members of each group are usually diverse enough to give all possible responses to the particular question, though with varying frequency. Therefore the designed envelope should consider all of them. The proposed approach uses the envelope to construct the attraction area. The latter dislocates the j-th response q_{ij} of the particular (i-th) person by pulling it towards the particular category pattern.

$$q'_{ij} = |q^v_{ij} - \overline{q_{kj}}| \cdot (1 - \widehat{q^v_{kj}}) \tag{8}$$

where $\widehat{q^v_{kj}}$ is the percentage of the response option v to the question j observed in the category k and $\overline{q_{jk}}$ is the average response to this question j given in the k-th group.

5.3 Classification strategy

The assignment of each example to the particular category is based on the comparison between it and the set of category patterns created according to the procedure from Section 5.1. The comparison is performed by calculating the distance between two answer patterns

using the particular measure. For this purpose the example should be treated as the point in the m-dimensional space (Figure 4), where each response (single dimension) has the same weight (influence on the overall distance calculation). Contrary to the traditional questionnaire analysis, the proposed system uses various strategies during the classification. They are as follows:

1. Counting the number of questions, for which the answer is closest to the particular pattern. This way the category represented by the pattern, for which the number of similar questions is greater wins and is returned as the decision of the system.
2. Calculating the overall distance for each dimension (questionnaire item). This way not only the number of answers closest to the pattern matters, but also their actual distance.
3. Other strategies, including the questionnaire scales (selecting only the subset of questions that measure the particular aspect of the category), may be used and will be considered in the future.

The most important parameter during this operation is the distance measure selection. Among various options, the most suitable measures for the task were selected and tested.

5.4 Distance calculation and decision making

This step is the application of knowledge extracted from the learning data and the selected measure to make a decision about the category of the analysed example e_i. In the presented research, the generalization ability of the system (the correct reaction on the examples not presented during the training stage) was measured by the Cross-Validation (CV). The subset of the original set L was selected as the testing set T and provided to the input of the system. The number of examples in T depends on the CV strategy and was selected to determine the minimum value allowing for the acceptable generalization. The classification accuracy rc of the proposed system was measured as the percentage of the correctly classified examples from the testing set:

$$r_c = \frac{|\mathbf{e}_i : c_i = h_i|}{|T|} \tag{9}$$

Among available CV methodologies the Repeated Random Sub-Sampling CV (RRSSCV) was selected [23]. It consists in randomly moving the predefined number of k examples from $\boldsymbol{L}$ to form the test set $\boldsymbol{T}$. This procedure was repeated 100 times. In each iteration, the system is trained on $\boldsymbol{L}$ and its accuracy tested on $\boldsymbol{T}$. The number of examples selected to $\boldsymbol{T}$ was set relatively to the size of $\boldsymbol{L}$ as $k = 0.2 \cdot |\boldsymbol{L}|$. The obtained results include the mean accuracy for the randomly selected examples and the standard deviation determining the variability in data.

5.5 Data complexity assessment

This is the auxiliary module in the decision making system, which allows for analyzing the available training data and determining their difficulty for the classification. In the supervised learning mode the labeled examples are separated and assigned to the particular categories based on their features (responses to the particular questions). It is statistically possible that even for the correctly designed questionnaire (i.e. consisting questions separating various categories of respondents with high accuracy), there are some persons giving responses not characteristic to their category. Therefore the additional calculation of complexity ratio enables predicting, what might be the error rate of the system, i.e. the relative number of incorrectly classified respondents. The complexity of training data was measured using one of two methods:

– Calculation of the scalar product between every two vectors constructed by the category patterns [24]. This allows for determining the angle between vectors (in radians), calculated as:

$$\beta = \arccos \frac{x \cdot y}{|x| \cdot |y|} \tag{10}$$

where $\boldsymbol{x} \cdot \boldsymbol{y}$ is the standard scalar product and $|\boldsymbol{x}|$ is the length of the vector $\boldsymbol{x}$. Patterns close to each other will have small value of the angle (going to zero). Categories easy to distinguish should have large value of the angle: close to 1.57 for diagonal examples and 3.14 for antipodal ones, respectively. This way it is possible to determine the difficulty to distinguish data based on their geometric features.

The scalar product was applied for the data complexity assessment as follows (see Figure 7 for the illustration, where $k = 1,2$). First, response patterns $\boldsymbol{p}_k$ for any two compared testee categories were calculated. Then, two answer vectors e were selected from the clusters forming each testee category. The first one ($\boldsymbol{e}_{1max}(p_2)$ and $\boldsymbol{e}_{2max}(p_1)$, respectively) represents the most distant vector from the pattern of the opposite category, while the second ($\boldsymbol{e}_{1min}(p_2)$ and $\boldsymbol{e}_{2min}(p_1)$, respectively) is the closest one from the pattern of the opposite category. The Euclidean distance is used to select these vectors. Finally, scalar products $\beta_{mean} = \beta(p_1, p_2)$ for patterns from selected clusters, $\beta_{min} = \beta(\boldsymbol{e}_{1min}(p_2), \boldsymbol{e}_{2min}(p_1))$ and $\beta_{max} = \beta(\boldsymbol{e}_{1max}(p_2), \boldsymbol{e}_{2max}(p_1))$ of the least distant and the most distant points from compared groups are calculated. It is assumed that if $\beta_{mean} < \beta_{min}$ or $\beta_{mean} > \beta_{max}$, then these two groups of vectors overlap and therefore are difficult to differentiate. Analogously, $\beta_{mean} \in \langle \beta_{min}, \beta_{max} \rangle$ indicates that the considered categories are easier to differentiate as their groups do not overlap significantly.

–Setting a threshold value dependent on the number of answer options in the analyzed questionnaire. The threshold φ was calculated using formula (11), reflecting mean EMD value resulting from questionnaire answer options:

$$\varphi = \frac{\max(v_j) - \min(v_j)}{\dim(v)} \tag{11}$$

where $\dim(v)$ stands for the number answer options in the questionnaire, v_j are numeric values assigned to answer options in the j-th question. For every question in the questionnaire, pairwise comparisons between answer distributions for distinguished category patterns were calculated using formula (12) and (13).

$$\eta(jk_1, jk_2) = \begin{cases} 1, EMD(jk_1, jk_2) \geq \phi \\ 0, EMD(jk_1, jk_2) < \phi \end{cases} \tag{12}$$

$$\delta_{k_1,k_2} = \frac{\psi}{m} = \frac{\sum_{j=1}^{m} \eta(jk_1, jk_2)}{m} \tag{13}$$

where m stands for number of items in the questionnaire and j,k stand for distribution of answers to item j obtained in testee group k. Greater values of δ suggest pattern categories that are relatively easy to distinguish.

6 Datasets description

Datasets used for demonstration analyses of the proposed methodology represent three various research disciplines - education study, sociological questionnaire and psychological test. They differ in the number of questions the testee had to answer, the number of distinguished categories and the resolution of the test, i.e. the ability to distinguish the categories based on the answers. The number of solved questionnaires is different in each case, therefore the presented results are relative to the size of the set. The aim of the classification depends on the specific domain, therefore the statistical measures (such as sensitivity or specificity) are not used, leaving the accuracy as the main quality measure. Because the obtained results depend on the interpretation method, standard approaches to analyze the questionnaire in each domain are also briefly discussed.

6.1 Education

This dataset comes from an international study on fourth-grade students' literacy achievements (PIRLS - Progress in International Reading Literacy Study)[1]. The latest edition of the study took place in 2011 and covered 48 countries. For the purpose of this paper only data concerning Polish students was selected ($N = 4130$). Eleven questions from the student survey concerning reading habits and attitudes towards reading were used as predictors. Answers were given on 1 to 4 Likert type scale. Student achievement data in literacy was used for classification. Authors of the study distinguished 5 groups, reflecting international benchmark score reached by every student. The categories referred to the literacy skills, ranging from low, basic level, to the high fluency and command of the language. Frequency of the students belonging to subsequent categories was 148, 587, 1544, 1465, 386 students, respectively. As the questionnaire results reflecting reading habits and attitudes towards reading were used mainly for descriptive study, no standard method for testee classification based on the questionnaire results is suggested by its constructors.

6.2 Sociology

This dataset[2] comes from the sociological study measuring political preferences of American population conducted by the Pew Research Center. The study took place in early 2014. For the analyses the answers given in 10 item Ideological Consistency Scale were selected as predictors. Every question consisted of two statements - each reflecting either liberal or conservative point of view. Respondents had to choose one of the provided options as consistent with their beliefs or select "don't know option". The answer to the additional question: "In politics TODAY, do you consider yourself a Republican, Democrat, or independent?" was used for respondents classification. The dataset consisted of 3064 Democrats, 2415 Republicans, and 3968 respondents declaring their political views as independent. The standard procedure for the testee classification based on questionnaire results refers to frequency of selected answer options. If the testee gives more statements reflecting the conservative point of view, he/she is assigned to the republican category. On the other hand, selection of more statements reflecting the liberal point of view leads to the assignment of the democrat category. If the set of responses does not lean towards any of the two mentioned category, the testee is labeled as independent.

6.3 Psychology

This dataset comes from the validation study of the Depression Questionnaire [25] conducted in 2012 on the Polish population. The questionnaire consisted of 75 items with answers given on 1 to 4 Likert type scale (higher values suggest depression). This dataset includes

[1] dataset available online: http://timssandpirls.bc.edu/pirls2011/

[2] dataset available online: http://www.pewresearch.org/packages/political-polarization/

two major types of testees: people with clinically diagnosed depression ($N = 116$) and healthy individuals ($N = 518$). Answers for questionnaire items were treated as predictors while mental health condition was used for the classification. The Depression Questionnaire apart from total score distinguishes five specific scales reflecting different depression components [26]. Because this paper serves as an application example of proposed methodology in psychological testing, only the total score was used in further computations. The standard procedure for testee classification based on questionnaire result requires using a cutoff point. Testees who obtain total score higher than the cutoff, are classified as suffering from depression.

7 Experimental procedure

All three data sets (see: Section 4) were used to verify the proposed methodology. In each experiment the relation between the following parameters and the classification accuracy (measured as (9)) were verified:

–the distance measure
–the inclusion or exclusion of the envelope
–the number of distinguishable categories

The experimental procedure consists in the following three steps, repeated 100 times, according to the CV procedure. This way the mean accuracy and the standard deviation of results for the repeated experiments were obtained, representing the repeatability of outcomes, depending on the specific example selected to both subsets.

1. Dividing the original data set into two subsets, using the cross-validation procedure: learning ($\boldsymbol{L}$) and testing ($\boldsymbol{T}$) ones in the proportion 2:1.
2. Creating the patterns representing each considered category and generating the envelopes for them. The shape and the coordinates of the pattern depend on the examples, from which they are calculated.
3. Using the patterns (with envelopes, if needed) to classify all examples from the testing set $\boldsymbol{T}$. To maximize the accuracy, various parameters had to be tested to select the most promising values.

In the following subsections outcomes of the described operations are presented. The simulations were conducted using the R CRAN environment.

8 Results

This section covers experimental results from the application of the proposed methodology to the analysis of questionnaire data sets described above. The discussion covers the measurement of the dataset complexity and accuracy of the classification provided by our architecture.

8.1 Dataset complexity

The analyzed datasets differ regarding the questionnaire length, available response options and number of classification groups. Therefore, two measures of dataset complexity were used (see: Section 5.5) for the more detailed comparison. Computations were conducted using *emdist* and base packages. The proposed analysis allows for estimating, how easy it is to distinguish between two testee categories within the dataset. Secondly, it enables to compare different datasets regarding their complexity. In all questionnaires the extreme categories (the most distant from each other) have the greatest differences. The data complexity analysis shows how subsequent data sets are difficult to classify. The particular categories may be hardly distinguishable based on the responses given by the testee. This may be caused by the incorrectly designed questionnaire, or the difficulties in selecting candidates for the tests. If the data set is difficult (discernibility between categories is low), performance of any measure may be poor, which does not mean they are useless. The problem is within the data itself. Results in Table 1, 2 and 3 show different angles for the particular categories in the subsequent data sets. The angle between the patterns p_1, p_2 of the analyzed testee categories (”healthy” and ”ill”) is not fully informative without the measures between the closest and the farthest examples from the corresponding categories, respectively. In the Depression data set both categories are relatively easy to separate, therefore the distance between the patterns (β_{mean}) is close to the distance between the closest examples from both groups (β_{min}). The same is for the extreme categories (one and five) in the Education data set. All other categories will pose some problems for the analysis. For instance, it is difficult to distinguish between the category one and two in the Education set, as the angle between them is close to 0 and the distance between means is far from the distance between the closest patterns belonging to these categories. Similarly, the scalar product for people with different political views (Table 2) suggests there is the small difference between Independent and other voters. This is confirmed during the decision making about the new examples to the categories, as only half of them is classified correctly (see Section 6.2).

Table 1: Scalar product between groups in Psychology dataset.

		Healthy		
		β_{min}	β_{max}	β_{mean}
Depression	β_{min}	0.31		
	β_{max}		0.89	
	β_{mean}			0.31

Table 4 presents the alternative approach to the data complexity assessment, using the EMD measure. Results are similar to the cosine distance. The threshold ϕ was set

Table 2: Scalar product between groups in Sociology dataset (see: Section 5.5).

		Democrat			Independent			Republican		
		β_{min}	β_{max}	β_{mean}	β_{min}	β_{max}	β_{mean}	β_{min}	β_{max}	β_{mean}
Democrat	β_{min}				0.32			0.16		
	β_{max}					0.69			0.92	
	β_{mean}						0.2			0.39
Independent	β_{min}	0.32						0.23		
	β_{max}		0.69						0.65	
	β_{mean}			0.2						0.19
Republican	β_{min}	0.16			0.23					
	β_{max}		0.92			0.65				
	β_{mean}			0.39			0.19			

Table 3: Scalar product between groups in Education dataset (see: Section 5.5).

		One			Two			Three			Four			Five		
		β_{min}	β_{max}	β_{mean}	β_{min}	β_{max}	β_{mean}	β_{min}	β_{max}	β_{mean}	β_{min}	β_{max}	β_{mean}	β_{min}	β_{max}	β_{mean}
One	β_{min}				0.2			0.42			0.38			0.45		
	β_{max}					0.35			0.62			0.48			0.8	
	β_{mean}						0.11			0.21			0.34			0.46
Two	β_{min}	0.2						0.24			0.47			0.43		
	β_{max}		0.35						0.19			0.48			0.6	
	β_{mean}			0.11						0.1			0.28			0.32
Three	β_{min}	0.42			0.24						0.3			0.33		
	β_{max}		0.62			0.19						0.52			0.52	
	β_{mean}			0.21			0.1						0.13			0.23
Four	β_{min}	0.38			0.47			0.3						0.23		
	β_{max}		0.48			0.48			0.52						0.52	
	β_{mean}			0.34			0.28			0.13						0.1
Five	β_{min}	0.45			0.43			0.33			0.23					
	β_{max}		0.8			0.6			0.52			0.52				
	β_{mean}			0.46			0.32			0.23						

to the value depending on the number of available response options v and values assign to particular answer options. The values of ψ show, how many questions allow for distinguishing the selected categories in the particular questionnaire. Because the number of questions in each data set is different, the more informative is the relative ratio δ_{k_1,k_2}. As can be seen, again categories in the Depression data set are well distinguishable, as well as democrats from republicans or five from one from the Sociology and Education sets respectively. The tendency is especially well visible in the Education set, as the number of distinguishable questions decreases with the more similar categories (such as four and two, going to zero for neighboring ones, such as four and three or one and two).

8.2 Classification accuracy

All presented computations were conducted using packages stats and base in R CRAN. The overall results of the classification outcomes for various configurations of the proposed methodology are summarized in Table 5. The Euclidean and Mahalanobis distance were used to calculate the similarity measures between the analyzed examples from the testing set and the patterns of each category. Two classification strategies are present here: the maximum number of questions with responses closest to the particular category pattern (”item”) and the shortest distance from the pattern (”sum”). The application of envelope is determined (”YES” in the ”Envelope” column if included, ”NO” otherwise). Values for the particular data sets represent the mean value of the relative accuracy (μ) and the standard deviation (σ) obtained in repeating the experiment 100 times. The ”Education” set was tested twice: for all categories and only for examples belonging to two extreme categories to verify the distinguishability between the most distant classes. Values in bold font indicate the configurations producing the optimal results. The proposed methodology was confronted against the standard classification procedure, comparing the obtained score with the threshold value and making the decision based on the comparison result. The threshold value must be usually adjusted to maximize the accuracy, which is not the case for our approach. Also, the obtained results were compared to the random category assignment. The categories were assigned with the equal probability for each category (”random distribution”) and the subsequent

Table 4: Data complexity assessment using EMD threshold approach (see: Section 5.5)

Dataset name	Dataset details	Group 1	Group 2	ψ	$\delta(k_1,k_2)$
Psychology	$m=75$ $v=1,2,3,4$ $\phi=0.75$	Depression	Healthy	54	0.72
Sociology	$m=10$ $v=0,1,2$ $\phi=0.66$	Democrat Democrat Independent	Independent Republican Republican	0 8 0	0 0.8 0
Education	$m=11$ $v=1,2,3,4$ $\phi=0.75$	Five Five Five Five Four Four Four One One Three	Four One Three Two One Three Two Three Two Two	0 9 1 4 5 0 4 4 0 0	0 0.82 0.09 0.36 0.45 0 0.36 0.36 0 0

probabilities proportional to the frequency of categories in the set (”proportional distribution”). In both cases results obtained with our methodology are better (although sometimes slightly) than reference approaches.

In most cases introduction of the envelope allows for maximizing the accuracy, proving its usefulness. The Euclidean distance (where each question is equally important) is better than its Mahalanobis counterpart. The percentages of accuracies differ significantly between the data sets, which can be explained by the number of considered categories (as proves the Education dataset set analysis) and the difficulties in distinguishing between them (see Section 6.1). In the former case the correct classification percentage deteriorates with the number of various classes (which was expected) from over 90 percent for only two categories to below thirty for five classes. The difficulty of assigning examples to the particular pattern can be analyzed using their graphical representation, as shown in Figures 8 to 10. On the x-axis, the question number are present with the distinguished categories. The y-axis represents numbers of answers to the questions in the questionnaire. For each question, values for all categories are present. The vertical bar represents the distribution of answers with the color intensity proportional to the number of the specific answers to the question. Black horizontal stripes positioned in each bar are mean values. The best separation of categories is visible in the Sociology set, where three categories are distinguished (with the symbol ”D” for ”Democrat”, ”I” for ”Independent” and ”R” for ”Republican”), based on three envelope bars for each question. The most important is the separation between the ”D” and ”R” categories. The mean values of answers are usually distant for these categories, which is also confirmed by the distribution bars intensity. For instance, the answers for the first question is usually ”3” for the Republican and ”1” for the Democrat. The middle answer (”I dont know”) is selected rarely, therefore it could be easily eliminated from the test.

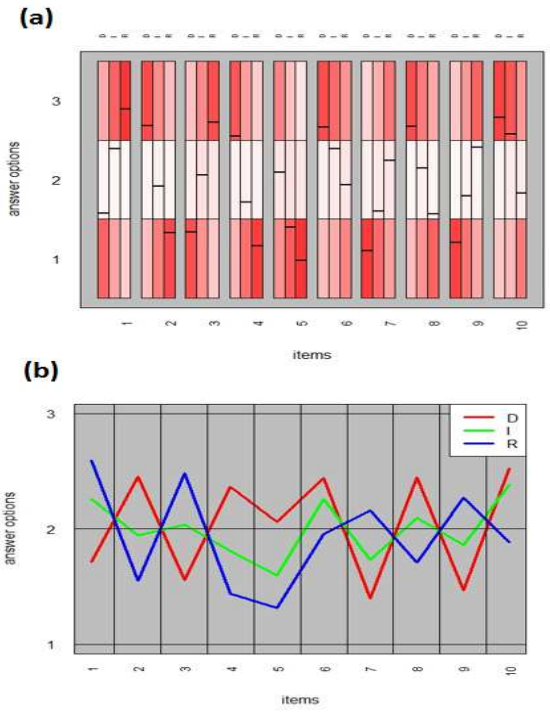

Fig. 8: Illustration of Sociology dataset.

The Education dataset is the most difficult, as it contains five categories to distinguish. However, proposed methodology also provides the acceptable separation. Figure 10 shows that in items 4-8 and item 11 the most popular response option in all of the analyzed groups is ”1”. Items 3, 9 and 11 relatively well distinguish testee

Table 5 Accuracy of analyzed methods

Dataset			Sociology		Psychology		Education (5 categories)		Education (2 categories)	
Distance measure	Classification category	Envelope	μ [%]	σ [%]	μ [%]	σ [%]	μ [%]	σ [%]	μ [%]	σ [%]
Euclidean	Item	NO	0.47	0.009	**0.9**	**0.023**	0.12	0.011	0.90	0.031
Euclidean	Sum	NO	0.51	0.011	0.89	0.020	0.23	0.013	0.92	0.021
Euclidean	Item	YES	0.47	0.011	**0.9**	**0.024**	0.12	0.010	0.89	0.033
Euclidean	Sum	YES	**0.52**	**0.009**	0.88	0.026	**0.29**	**0.016**	**0.93**	**0.021**
Mahalanobis	Item	NO	0.47	0.011	0.89	0.022	0.27	0.015	0.89	0.032
Standard procedure	Sum	NO	0.42	0.024	0.42	0.054				
Random (proportional distribution)		NO	0.35		0.30		0.30		0.59	
Random (random distribution)		NO	0.33		0.50		0.50		0.50	

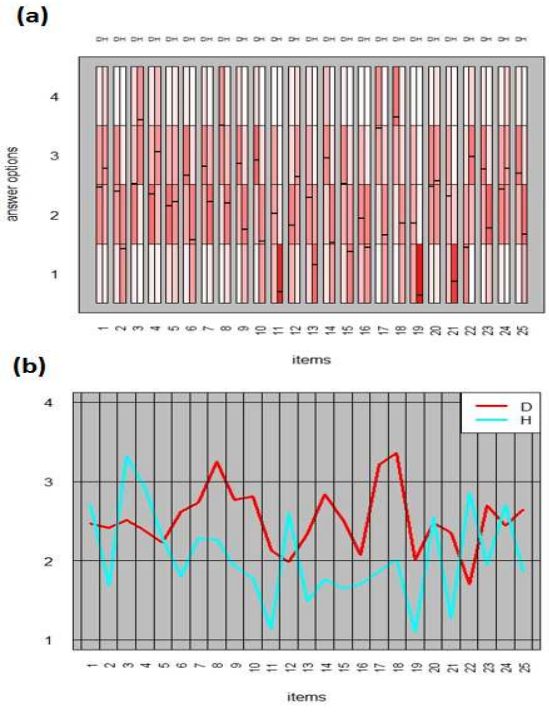

Fig. 9: Illustration of the Psychology dataset.

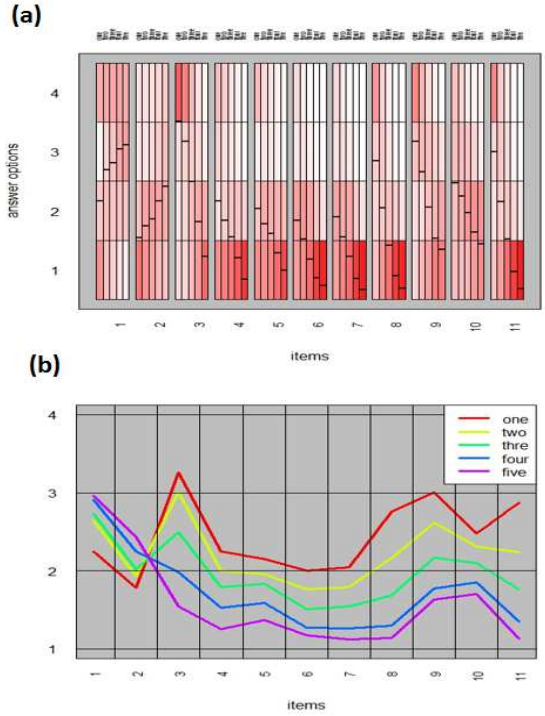

Fig. 10: Illustration of Education dataset.

groups. As can be expected, to increase the classification accuracy the number of categories should be decreased. The question is whether all original five categories are required, as the three (one, three and five) would be sufficient to describe the variability in the ability to read among pupils of the primary schools.

9 Discussion and future work

The proposed methodology improves testee classification accuracy. However, big discrepancies in classification accuracy between standard procedure recommended by the questionnaire constructors and the methodology proposed in this paper observed in psychological dataset are at first glance puzzling. The standard procedure recommends using cutoff equal to 130 points testees obtaining higher scores are classified as suffering from depression. However, [25] notes that the results obtained in the questionnaire strongly differ depending on age adolescents tend to obtain significantly higher scores than adults. Therefore using the cutoff point is recommended only for the adult population. In this case, questionnaire constructors report 90 percent sensitivity and 87 percent specificity [25]. The dataset used in this paper contained questionnaire results from both adolescents (16-17 y/o; $N = 83$) and adults (18-81 y/o; $N = 436$). It is suspected that the dataset structure might have influenced accuracy

score reported for standard procedure (see: Table 5). However, it is worth stressing out that methodology proposed in this paper led to nearly 90 percent accuracy in the testee classification irrespective of selected distance metrics or classification strategy.

The introduction of envelopes (see: Section 5.2) improved the accuracy of testee classification. However, its impact was weaker than expected. One reason for this may be that the distribution of testee answers in the analyzed groups were not highly skewed. In normal Gaussian distribution mean is equal to median[19]. The average absolute difference between mean and median in the analyzed datasets was equal: 0.26 for the Psychology, 0.40 for the Education and 0.63 for the Sociology dataset. A shown in Table 1 the greatest improvement in classification accuracy was observed for Education dataset while the weakest for Psychology dataset. This result confirms the assumption that greater improvement in classification using envelopes is observed for distributions more differing from the normal Gaussian distribution. Therefore, it can be concluded that the introduction of envelopes is advantageous in questionnaire data analysis it can boost classification accuracy when questionnaire scores distribution is skewed and does not deteriorate accuracy when the distribution is normal. Furthermore, we postulate that proposed method for determining dataset complexity may be treated as a measure of questionnaire quality (δ reflects discriminating power of whole questionnaire). Also, combined with the visualization methods presented in Figures 8 - 10 it can be useful for questionnaire users and questionnaire constructors. The former would treat it as an indicator of questionnaire quality while making decision about questionnaire purchase. The latter may find it helpful in the process of questionnaire construction. Knowing which items poorly differentiate analyzed groups, allows for replacing them with other, more useful items. The proposed methodology allows for the interpretation of questionnaire results. Visualization methods from Figures 8-10 enable the easy comparison between answers given by a testee and the frequency distributions obtained for the representatives of certain population (i.e. people suffering from depression or showing certain political preferences). Hence, it combines qualitative and quantitative approach in the questionnaire data analysis. The planned research includes examination of how different threshold values for the data complexity assessment procedure (δ) change the results of data complexity assessment. Additional distance metrics may also be tested and their efficiency compared to the ones presented in this paper. This includes both the testee classification and data complexity assessment. Another interesting step in the methodology development may be the introduction of fuzzy logic for the classification or modifying proposed methodology by including the interaction between the results obtained by the testees in different questionnaire scales.. We believe the approach described in this paper may be adapted for profiling testees and detecting their subcategories. For this purpose, results obtained by the testee in particular questionnaire scales (smaller subsets of items) can be analyzed and interpreted.

References

[1] M. R. Novick, Journal of Mathematical Psychology **3**, 1-18 (1966).
[2] F. M. Lord, M. R. Novick, A. Birnbaum, Statistical theories of mental test scores, Addison-Wesley, 1968.
[3] R. M. Groves, F.J. Fowler Jr, M.P. Couper, J.M. Lepkowski, E. Singer, R. Tourangeau, Survey methodology, Volume 561, John Wiley & Sons, 2011.
[4] S. E. Embretson, S.P. Reise, Item response theory, Psychology Press, Mahwah, 2011.
[5] F.M. Lord, Applications of item response theory to practical testing problems, Erlbaum, Mahwah, 1980.
[6] D. Magis, Applied Psychological Measurement **37**, 304-315 (2013).
[7] D. Thissen, L. Steinberg, Psychometrika **51**, 567-577 (1986).
[8] D. L. Altheide, C.J. Schneider, Qualitative media analysis, Volume **38**, Sage, 2012.
[9] D. Kelley-Winstead, New Directions in Education Research: Using Data Mining Techniques to Explore Predictors of Grade Retention, Doctoral dissertation, George Mason University, 2010.
[10] M. Jekel, S. Fiedler, A. Glckner, Judgment and Decision Making **6**, 782-799 (2011).
[11] V. Levashenko, E. Zaitseva, K. Pancerz, J.Gomua, Fuzzy Decision Tree Based Classification of Psychometric Data, Position paper, Federated Conference on Computer Science and Information Systems, 3741 (2014).
[12] M. Bacauskiene, A. Verikas, A. Gelzinis, A. Vegiene, Expert Systems with Applications **39**, 5506-5512 (2012).
[13] G. P. Zhang, IEEE Transactions On Systems, man, and Cybernetics **30**, 451-462 (2000).
[14] A. G. Di Nuovo, S. Di Nuovo, S.Buono, Artificial intelligence in medicine **54**, 135-145 (2012).
[15] O. Mich, A. Burda, K. Pancerz, J. Gomula, Digital Technologies, 255-261 (2014).
[16] M. M. Deza, E. Deza, Encyclopedia of distances, Springer, Berlin Heidelberg, 2009.
[17] P. C. Mahalanobis, Proceedings of the National Institute of Sciences **2**, 49-55 (1936).
[18] Y. Rubner, C. Tomasi, L. Guibas, Proceedings ICCV, 5966 (1998).
[19] F.J. Gravetter, L.B. Wallnau, Statistics for the behavioral sciences, Belmont, Wadsworth Thomson Learning, 2000.
[20] R.S. Michalski, J.G. Carbonell, T.M. Mitchell, Machine learning: An artificial intelligence approach, Springer Science & Business Media, 2013.
[21] P. Bilski S. Rabarijoely, Proceedings of AICS 2014, 28-37 (2014).
[22] J. Vesanto and E. Alhoniemi, IEEE Transations on Neural Networks, **11**, No. 3, 586-600 (2000).
[23] R. Ward, Information Theory IEEE Transactions **55**, 5773-5782 (2009).
[24] J.W. Dettman, Mathematical methods in physics and engineering, Courier Corporation, 2013.

[25] E. Lojek, J. Stanczak, A. Wojcik, Kwestionariusz do Pomiaru Depresji KPD test manual, Pracownia Testow Psychologicznych Polskiego Towarzystwa Psychologicznego, Warsaw, 2015.
[26] E. Lojek, J. Stanczak, A. Wojcik, International Neuropsychological Society (INS) Mid-year Meeting, Jerusalem, 2014.

Maria Rafalak is a PhD candidate at the Polish-Japanese Academy of Information Technology in Warsaw. She has reached a Msc in psychology from the University of Warsaw and BSc in computer science and econometrics from the Warsaw University of Life Sciences. In 2012 she was an intern at the University of Cambridge psychometric center. Her scientific interests focus on developing new algorithms for psychometric purposes.

Piotr Bilski was born in 1977 in Olsztyn, Poland. He graduated from Warsaw University of Technology, Institute of Radioelectronics, obtaining MSc degree in 2001 (with honors), PhD degree in 2006 (with honors) and DSc degree in 2014. Currently he is an Assistant Professor in the Institute of Radioelectronics, Warsaw University of Technology. His main scientific interests include diagnostics of analog systems, design and analysis of virtual instrumentation, application of artificial intelligence and machine learning methods to the environmental sciences. He is the member of IEEE, IMEKO TC10 and POLSPAR and reviewer for such journals like Measurement, IEEE Transactions on Instrumentation and Measurement, Expert Systems with Applications.

Adam Wierzbicki received his Ph.D. degree from the Warsaw University of Technology and a habilitation title from the Institute of Systems Research of the Polish Academy of Sciences. He is currently employed at the Polish-Japanese Institute for Information Technology, where he has the position Full Professor and of Vice-Dean of the Department of Informatics. He is in expert in Peer-to-Peer computing. His current research interests focus on social informatics, in particular on trust management and fairness in distributed systems.