Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 303–310

DOI: 10.15439/2015F151
ACSIS, Vol. 5

On Integrating Clustering and Statistical Analysis for Supporting Cardiovascular Disease Diagnosis

Agnieszka Wosiak
Lodz University of Technology
Institute of Information Technology
ul. Wólczańska 215, 90-924 Łódź, Poland
Email: agnieszka.wosiak@p.lodz.pl

Danuta Zakrzewska
Lodz University of Technology
Institute of Information Technology
ul. Wólczańska 215, 90-924 Łódź, Poland
Email: danuta.zakrzewska@p.lodz.pl

***Abstract*—Statistical analysis of medical data plays significant role in medical diagnostics development. However in many cases the statistics is not effective enough. In the paper we consider combining statistical inference with clustering in the preprocessing phase of data analysis. The proposed methodology is checked on cardiovascular data and used for developing methods of early diagnosis of hypertension in children. Experiments, conducted on the real data, have demonstrated that the proposed hybrid approach allowed to discover relationships which have not been identified by using only the statistical methods. We have observed approximately 30% growth in the number of correlations between diagnosed attributes. Moreover all the obtained statistically significant dependencies were stronger in clusters rather than in the whole datasets.**

I. Introduction

IN RECENT years medical progress as well as the equipment development make possible collecting the increasing amount of data. One can expect that their analysis will help medical practitioners in improving patient care, proposing new therapies or developing the existing ones. However, statistical analysis, which is commonly used to support medical diagnosis, in many cases, turns out to be not effective enough. Such situation takes place, when the correlations between parameters, which seem to be useful for medical inference, are not possible to obtain. For example, it might happen, when standard deviation in the dataset takes on the large value [1].

To improve the performance of statistical models, we propose the new approach, which consists in including clustering in the preprocessing phase. Both of the techniques have already been broadly investigated in medical applications, but their combination has not been examined as supporting tools for medical diagnosis so far. The presented approach enables to identify groups of similar instances, for which statistical models can be built effectively. Special attention is drawn to feature selection process. We assume that the set of attributes used in clustering and statistical analysis phases should be different, not correlated and consistent with the process of medical diagnosis as well as the state of art of the approaches to statistical analysis of the medical data (see [2] for example).

In the paper we focus on cardiovascular diseases, which are the leading causes of death in the majority of countries [3]. The research aims at developing methods for early diagnosis of hypertension (high blood pressure) in children. The proposed methodology was verified by experiments done on three different sets of real children cases. Experiment results showed that application of the cluster analysis effectively supports statistical inference for the diagnosis in the considered cardiovascular problem.

The remainder of the paper is organized as follows. In Section II relevant work is presented. Next, the medical issues of hypertension problems in cardiovascular diagnosis are introduced, then the proposed methodology is described. In the following section, the experiments conducted on real data are depicted. Finally, the results are discussed and some concluding remarks are presented.

II. Related Work

In medicine most of the rules for diagnosis and treatment are based on statistical analysis. However, new challenges connected with medical data analysis impose application of more sophisticated methods such as data mining techniques which ensure the process improvement. Application of data mining techniques in biomedical and healthcare fields was discussed by Yoo et al. [4]. The authors stated that descriptive and predictive power of data mining could be widely used in these areas.

Cluster analysis has already been integrated with statistical methods for medical data in the research of Haldar [5]. However the goal of the study was not do discover new dependencies, but to define the phenotypes of clinical asthma. The research was proposed against other models of asthma classification and according to authors it might have played a supporting role for different phenotypes of heterogeneous asthma population. A survey of data mining methods that has been applied to traditional Chinese medicine (TCM) clinical data systems was provided by [6]. Cluster analysis, association rules, a latent structure model and a topic model were considered in the context of Chinese medicine.

In [1] different clinical decision support systems for heart disease prediction and diagnosis were compared. These systems were based on such data mining techniques as: a multiplayer perceptron, genetic algorithms, fuzzy rules, decision trees and Bayesian networks. As the result of investigations the authors stated that the considered techniques are not satisfactory and finally there is still lack of a solution for the

978-83-60810-66-8/$25.00

identification of treatment options for the patients with heart diseases.

In [7] a statistical inference of heart rate and blood pressure was examined. The authors considered three different approaches. The first one was based on examining correlation between raw data. Then since the measurements could be corrupted by noise, a filtration procedure was performed on data before correlating the signals. In the last approach least squares approximation was applied. The results of all of the techniques were similar. The obtained correlation coefficients seemed to be an unpredictable random numbers.

Meng et al. [8] compared the performance of logistic regression, artificial neural networks and decision tree models for predicting diabetes or prediabetes using common risk factors. The research proved the advantages of decision tree model comparing to the other considered techniques. The authors of [9] examined the performance of the alternate classification methods, such as bootstrap aggregation, boosting, random forests and support vector machines with conventional classification trees to classify patients with heart failure. Bashir et al. [10] proposed combination of three classifiers for intelligent heart disease diagnosis. Broad review of data mining techniques applied in this area was presented in [11].

III. Hypertension Problems in Cardiovascular Diagnosis

Hypertension is the cardiovascular disease, which may have their onset in the young [12]. Hence arterial hypertension is a significant problem in pediatric practice. It is estimated that this pathology affects 3-5% of the total children population, while for teenagers the percentage of hypertension cases increases up to 10%. Therefore finding effective methods which support early diagnosis of hypertension and thus help in implementing an appropriate management to prevent the disease is currently the matter of interests of many researchers.

Hypertension is mainly defined on the basis of blood pressure measurements. However the initial cardiac data can be characterized by over 50 attributes. All patients undergo physical examination, manual arterial blood pressure measurements (RR SBP, RR DBP), ambulatory blood pressure monitoring (ABPM-S, ABPM-D), echocardiographic examination to evaluate cardiac function using standard parameters (ejection fraction, shortening fraction and myocardial performance index) and tissue Doppler examination (systolic mitral annular velocity profile and regional function parameters: velocity, strain, strain rate). A selection of attributes, which cardiologists use to diagnose arterial hypertension (see [13], [14] and [15]) is shown in Table I. The first two columns of the table contain names and descriptions of selected parameters, the third one presents ranges of attribute values.

To improve early detection of hypertension in children, the researchers look for new factors, which may indicate the high blood pressure appearance [16]. Medical data analysis helps in evaluating the characteristics of the variables in the data sets of healthy and diagnosed children and discovering the relationships between all the parameters. The detailed medical descriptions and statistical analysis of these issues were subjects of the research presented in [13], [14] and [15]. The authors stated that the high value of standard deviation in the dataset disabled obtaining some of the correlations between parameters, which could have been useful for medical inference. That fact motivated us to build methodology described in the presented paper.

TABLE I
The selection of parameters collected for arterial hypertension evaluation

Name	Description	Range
HA	Arterial hypertension presence	Nom. (Yes/No)
BMI	Body mass index	11.9 - 31.6
BWT	Body weight	14.6 - 88.0
BSA	Body surface area	1.4 - 2.2
HC	Head circumference	29.0 - 36.0
PI	Ponderal index	14.58 - 26.20
RR SBP	Manual measurement of systolic blood pressure	86 - 150
RR DBP	Manual measurement of diastolic blood pressure	44 - 87
ABPM-S	Ambulatory systolic blood pressure monitoring	19 - 87
ABPM-D	Ambulatory diastolic blood pressure monitoring	4 - 78
IVSd	Interventricular septum	5 - 14
PWDs	Posterior wall thickness	11 - 19
TG	Triglyceride Level	24 - 236
E/A	Ratio of the early to the late mitral inflow velocities	1.15 - 1.77
DecT	Deceleration time - time interval of peak E-wave velocity to its extrapolation to the baseline	126 - 325
AEF	Atrial ejection force	4570 - 9158

IV. Materials and Methods

In medical research, an analysis of the results of observations plays the crucial role, as its effects are expected to be implemented into practical applications. The process of medical research is usually supported by statistical analysis but very often it is not effective enough. In many cases, dissimilarities or inconsistency within the data sets appear due to incorrect measurements or distortions. The presence of such deviations may lead to the rejection of true hypothesis in the case of small data sets. The use of clustering before conducting a statistical analysis allows to identify groups of similar cases, and thus to better evaluate respective parameters.

The proposed methodology of improving medical inference process from a medical data set consists of three main steps:

- feature selection, which enables choosing the set of attributes for building clusters,
- clustering based on the parameters indicated by the previous step to distinguish groups of similar characteristics,

- statistical analysis performed in clusters to find new dependencies between all the collected parameters.

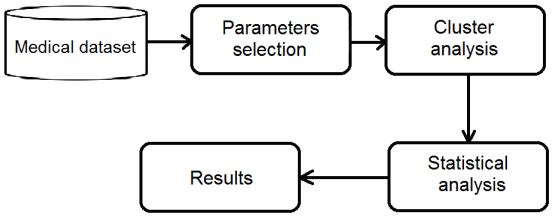

Fig. 1. System architecture for improved knowledge discovery using clustering techniques.

The general structure of our approach is shown on Fig. 1. We assume that clustering and statistical analysis are applied on the preprocessed and transformed data, and are preceded by feature selection process. The description of each step is presented in the following subsections.

A. Feature Selection

Feature selection is a technique of choosing an appropriate subset of the available attributes, which ensures building the model with high classification accuracy. In medical data analysis there exist two main feature selection approaches: using of automatic feature selection mechanisms or selecting parameters as a result arising from the process of medical diagnosis. The first one was considered in [17]. The authors tried to provide a generic introduction to variable elimination, which can be applied to a wide range of machine learning problems. They considered filter, wrapper and embedded methods. However, they found out that comparison of feature selection methods can only be done on the data of the same characteristics.

Cheng, in turn, stated that in the case of cardiovascular diseases the feature subsets selected in the process of medical diagnosis improve the sensitivity of the analysis [18]. Such approach will be used in the proposed methodology, to obtain a set of attributes for further analysis.

B. Clustering

Cluster analysis is one of the most commonly used data mining techniques, as it may be applied to classify complex data of many variables and many dimensions. Unlike discriminant analysis, no classification variables are inserted to divide the original data. What is more, in many cases, when the groups are detected, it is necessary to use other methods to discover the meaning of clustering [19]. Therefore, combination of cluster analysis and statistical inference seems to be the effective tool supporting medical diagnosis.

During investigations of the effectiveness of the proposed methodology, two different clustering approaches: deterministic and probabilistic were considered. As the presented technique aims at supporting physicians in making medical diagnosis, there were chosen simple comprehensible algorithms, because doctors should understand the tools they use. In the first group, k-means algorithm has been considered. In the case of medical data, this technique gives higher accuracy and lower root mean square error (RMSE) in comparison with other clustering methods, such as fuzzy C-means clustering, mountain clustering or subtractive clustering [20]. As the probabilistic method, expectation-maximization (EM) algorithm has been investigated. EM uses the finite Gaussian mixtures model to generate probabilistic descriptions of clusters in terms of means and standard deviations [19].

1) The K-Means Algorithm: The k-means algorithm is one of the most popular clustering method. The clusters in data set are defined by minimizing a distance (dissimilarity) function. In most of the cases Euclidean metric is considered as distance function [2],[21].

Let us consider the set of n data $X = \{x_i; i = 1, ..., n\}$ and the set C of k cluster centers $C = \{c_j; j = 1, ..., k\}$. For a given k, the goal of clustering is to find C for which the function determined by (1) achieves its minimum.

$$min_j \left(\sum_{i=1}^{n} \|x_i - c_j\| \right) \tag{1}$$

The algorithm for k-means can be described as follows [21]:

1) Randomly choose k data points from X as the initial set C of cluster centers. Denote them by $c_j, j = 1, ..., k$.
2) Reassign all $x_i \in X$ to the closest cluster mean c_j
3) Update all $c_j \in C$ as means of the points assigned to the corresponding clusters.
4) Repeat steps 2 and 3 until cluster assignments do not change.

Choosing k initial centers at random, does not guarantee finding optimal clusters. To increase the chance of finding a global minimum in (1), it is usually suggested to run the algorithm several times with different initial choices and pick out the best final result - the one with the smallest total squared distance [21].

2) The EM Algorithm: The expectation - maximization (EM) algorithm is an iterative algorithm used to calculate maximum likelihood estimates in parametric models in the presence of missing data [22].

The goal of statistical models is to find the most likely set of clusters on the basis of training data and prior expectations. Expectation- Maximization algorithm (EM) uses the finite Gaussian mixtures model to generate probabilistic descriptions of clusters in terms of means and standard deviations [19]. The big advantage of EM algorithm is a possibility to select a number of clusters by cross validation techniques, what allows to obtain its optimal value [21]. That feature allows not to determine the number of clusters at the beginning. Similarly to k-means method, parameters are recomputed until the desired convergence value is achieved.

3) Optimal Number of Clusters: One of the most important issue connected with clustering is an identification of the optimal number of clusters. There exist different approaches to solve this problem.

In the case of k-means algorithm, the technique called elbow criterion has been considered. The elbow criterion says that one should choose a number of clusters, such that adding the next one does not increase quantity of information sufficiently [23]. When a graph for a validation measure calculated in the clusters is plotted against the number of clusters, at first amount of information is increasing, but at some point the gain starts decreasing, giving an angle in the graph, that is called the elbow.

In some cases, elbows cannot be unambiguously identified. Therefore it may be helpful to use another validation method for finding optimal number of clusters. As in the case of EM algorithm the optimal number of clusters can be determined by cross validation technique [21], the number of clusters indicated by elbow criterion can be confirmed by EM clustering.

It is worth mentioning that in medicine the number of clusters is very often equal to two as there exists common intention to split the whole data set into two groups [20]. Besides, when the number of considered instances is small, what very often takes place in medical applications, the bigger number of clusters would decrease the group sizes and as a consequence would make the medical inference less reliable as it is difficult to obtain sufficiently high power of statistical tests [2],[24].

C. Statistical Analysis

Statistical data analysis usually begins with an assessment of measures of descriptive statistics, which allows to detect errors that were not identified during data preparation phase. The basic descriptors, for which the evaluation is indicated, include measures of central tendency (arithmetic mean, median and modal), measures of dispersion (range and standard deviation). Next statistical inference using a suitable test is carried out. The selection of the test is made on the basis of the type and the structure of the analyzed data. It depends on attribute types, scale type, number of experimental groups and their dependency, as well as the test power. The test selection should be done in accordance with the requirements of the USMLE (The United States Medical Licensing Examination). In the current research, we will consider the tests commonly used in medical diagnosis problems [1]:

- Kolmogorov-Smirnov test, which is used to test for normality of distribution of the attributes,
- Unpaired two-sample Student's *t*-test for the significance of a difference between two normally distributed values of attributes,
- Mann-Whitney U test, which is a non-parametric test for significant differences determination, where attributes were in nominal scales.

The impact of one variable measured in an interval or ratio scale to another variable in the same scale can be expressed using the Pearson's correlation coefficient $r_P(x, y)$. In the case where one or both of the variables are measured with an ordinal scale, or variables are expressed as an interval scale, but the relationship is not a linear one, the Spearman's correlation $r_S(x, y)$ test is used.

V. Experimental Analysis and Results

The main objective of the experiments was to examine the performance of the proposed approach by comparing the results derived from statistical analysis carried out on clusters with the ones obtained for the whole datasets. The experiments were conducted on the real datasets, which were gathered for early diagnosis of arterial hypertension in children.

A. Data Description

There have been considered three different datasets ("HEART", "ECHO", "IUGR") collected from children hospitalized in the University Hospital No 4, Department of Cardiology and Rheumatology, Medical University of Lodz. Each of the dataset was examined for the particular cardiovascular problem:

- "HEART" - to discover dependencies between arterial hypertension and left ventricle systolic functions,
- "ECHO" - to evaluate correlations between arterial hypertension and myocardial functions using tissue Doppler echocardiography,
- "IUGR" - to discover dependencies between abnormal blood pressure and being born as small for gestational age.

The "HEART" dataset consisted of 30 cases, the "ECHO" dataset of 66 instances and the "IUGR" dataset contained 50 specimens. There were no missing values within attributes.

B. Cluster Analysis

In the first step of the experiments, the clusters for diagnosed children were created by using two clustering algorithms: k-means and EM implemented by WEKA Open Source software [21].

Clusters were built taking into account attributes according to feature selection method consistent with the process of medical diagnosis (see [2]). In the case of arterial hypertension, diagnosis performed by medical expert is mainly based on the blood pressure measurements (either manual or ambulatory monitored). The rest of the attributes are usually supportive for medical staff as each of them separately cannot indicate the disease and multivariate analysis is difficult to perform without any computer support. Therefore for "HEART" and "ECHO" datasets we considered 4 attributes: RR SBP, RR DBP, ABPM-S and ABPM-D in accordance with hypertension diagnosis (see Table I). In the case of "IUGR" dataset we used 3 attributes: birth weight, head circumference and ponderal index as consistent with the diagnosis of intrauterine growth restriction (being born as small for gestational age), and included 16 risk factors (i.a. hypertension in relatives, smoking during pregnancy) in a feature selection subset, which may have an impact on intrauterine growth restriction [28].

To choose the best number of clusters the elbow criterion has been applied. As validation measure, within cluster sum of squares has been considered. As the result, the charts for validation measures plotted against number of clusters indicated elbow points c=2 for "HEART" and "ECHO" datasets and c=3 for "IUGR" dataset (see Fig. 2).

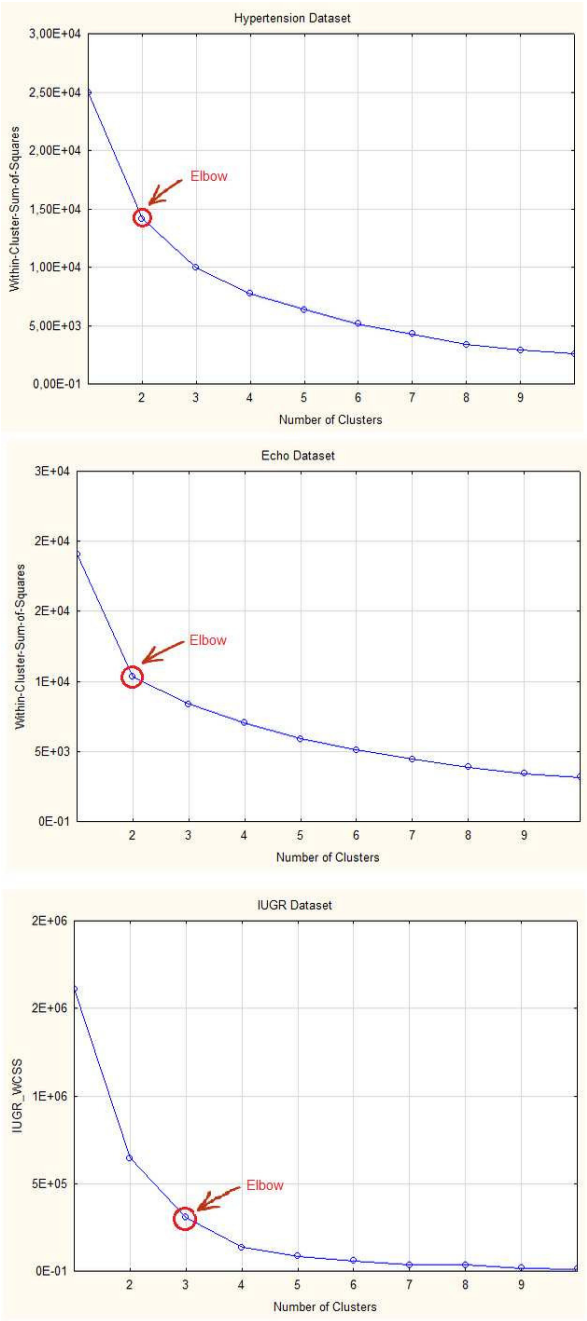

Fig. 2. Elbow charts for number of clusters determination.

Additionally to elbow criterion, EM algorithm which automatically generated number of clusters by using cross-validation [21] was implemented. The obtained results (c=2 for "HEART" and "ECHO" datasets and c=3 for "IUGR" dataset) confirmed the proper choice of the number of clusters.

It is worth mentioning that during clustering process, the subgroup characterized by higher mean values of parameters concerning arterial hypertension and lower standard deviations of those attributes or ,in the case of "IUGR" dataset, lower mean values of parameters concerning birth weight and size, was distinguished. In the case of "HEART" that subgroup consisted of 22 cases for k-means algorithm and 23 for EM method, "ECHO" subset included 44 and 35 instances respectively, "IUGR" subgroup contained 19 specimens for k-means algorithm and 25 for EM method. Sizes of the clusters for all the datasets are presented in Table II.

TABLE II
SIZES OF THE CLUSTERS

Dataset	Numbers of instances in clusters	
	k-means clustering	EM clustering
HEART	8, 22	7, 23
ECHO	22, 44	31, 35
IUGR	19, 14, 14	25, 15, 7

C. Statistical Analysis

Analysis of correlations among datasets, in order to support medical knowledge acquisition and decision making, is still one of the most popular techniques in medical research [25]. Therefore next step of the experiments concerned indication of the attributes, which are not significantly correlated with the ones used for building clusters, and thus can be used in statistical analysis process. By insignificant correlation we mean values with correlation coefficient r<0.3 and p-value>0.05 ([26], [27]). Tables III, IV, V and Fig. 3 present correlation values of the features used in clustering and the ones indicated for statistical analysis.

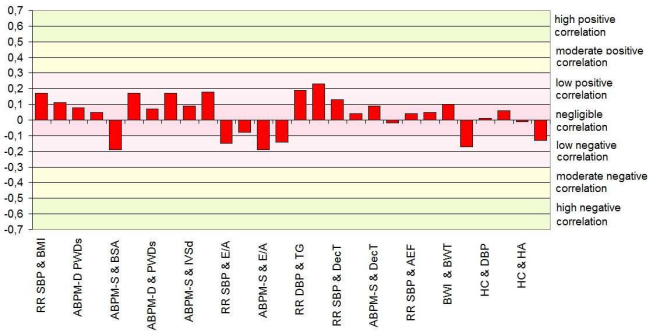

Fig. 3. Values of correlation coefficient between attributes used for clustering and statistical analysis.

Correlation values obtained for the clusters were compared to the ones got for the whole group of diagnosed children. Comparison of results confirmed effectiveness of the proposed methodology. For each dataset we obtained greater number of statistically significant correlations which may lead to improved medical diagnosis in the future. By significant correlations we mean values with correlation coefficient r>=0.3 and p-value<=0.05 ([26], [27]). The results of detected correlations are presented in Table VI, where the column (3) presents the numbers of discovered dependencies in clusters and the

TABLE III
VALUES OF CORRELATIONS BETWEEN FEATURES USED FOR CLUSTERING AND STATISTICAL ANALYSIS FOR "HEART" DATASET

Attribute 1	Attribute 2	Corr. coeff.	p-value
RR SBP	BMI	0.17	0.38
ABPM-S	BMI	0.11	0.55
ABPM-D	BMI	0.08	0.67
RR DBP	BSA	0.05	0.77
ABPM-S	BSA	-0.19	0.31
RR DBP	PWDs	0.17	0.36
ABPM-D	PWDs	0.07	0.67
RR SBP	IVSd	0.22	0.23
RR DBP	IVSd	0.17	0.35
ABPM-S	IVSd	0.09	0.61
ABPM-D	IVSd	0.18	0.32

TABLE IV
VALUES OF CORRELATIONS BETWEEN FEATURES USED FOR CLUSTERING AND STATISTICAL ANALYSIS FOR "ECHO" DATASET.

Attribute 1	Attribute 2	Corr. coeff.	p-value
RR SBP	E/A	-0.15	0.40
RR DBP	E/A	-0.08	0.48
ABPM-S	E/A	-0.19	0.12
ABPM-D	E/A	-0.14	0.26
RR SBP	TG	0.24	0.10
RR DBP	TG	0.19	0.13
ABPM-S	TG	0.23	0.10
RR SBP	DecT	0.13	0.27
RR DBP	DecT	0.04	0.76
ABPM-S	DecT	0.09	0.47
ABPM-D	DecT	-0.02	0.83
RR SBP	AEF	0.04	0.74
RR DBP	AEF	0.05	0.67

column (4) shows the percentage increase of correlations in comparison to the numbers of correlations for the whole dataset (column (2)).

Moreover the values of statistically significant correlations obtained after clustering were stronger than the corresponding

TABLE V
VALUES OF CORRELATIONS BETWEEN FEATURES USED FOR CLUSTERING AND STATISTICAL ANALYSIS FOR "IUGR" DATASET.

Attribute 1	Attribute 2	Corr. coeff.	p-value
BWI	BWT	0.10	0.49
HC	BWT	0.21	0.15
PI	BWT	-0.17	0.23
BWI	DBP	-0.25	0.10
HC	DBP	<0.01	0.98
PI	DBP	-0.23	0.11
BWI	HA	0.06	0.67
HC	HA	-0.01	0.93
PI	HA	-0.13	0.34

TABLE VI
NUMBER OF STATISTICALLY SIGNIFICANT CORRELATIONS DETECTED WITH AND WITHOUT CLUSTERING

Dataset name (1)	Whole dataset (2)	Cluster (3)	Increase $[in\%]$ (4)
Clustering method: k-means			
HEART	14	25	78%
ECHO	13	16	23%
IUGR	11	14	27%
Clustering method: EM			
HEART	14	29	107%
ECHO	13	17	31%
IUGR	11	15	36%

values for the whole diagnosed group. Some of these correlations - as an overview of obtained results - are presented on Fig. 4 and in Table VII, where the column (2)- *"Correlation type"* contains the names of correlation parameters, the column (3) presents values of correlation coefficient and the last column (4) shows the p-value of significance level.

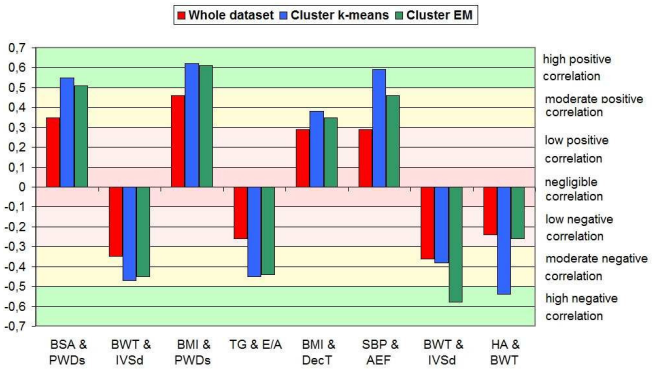

Fig. 4. Values of correlation coefficient for clusters in comparison to the whole dataset.

Concluding, the results of the experiments have shown that the proposed approach, which consists of supporting statistical inference by clustering, significantly improved the effectiveness of the medical data analysis for all the considered datasets.

VI. CONCLUSIONS

Typically, during the process of computer-aided clinical and epidemiological studies only one of selected data analysis method is involved. In spite of the mostly used statistical analysis, in the paper, a hybrid methodology of medical data analysis has been proposed. The presented method consists of combination of clustering and statistical inference, where the first technique is used as a data preprocessing tool for the second one. The investigations have shown that supporting statistical analysis by clustering provides significant benefits. Experiments conducted on the real data have demonstrated that

TABLE VII
SAMPLE VALUES OF CORRELATIONS DETECTED WITH AND WITHOUT CLUSTERING

Dataset name (1)	Correlation type (2)	Correlation coefficient (3)	Significance (p-value) (4)
Whole dataset			
HEART	BSA & PWDs	0.35	0.05
	BWT & IVSd	-0.35	0.05
	BMI & PWDs	0.46	0.01
ECHO	TG & E/A	-0.26	0.03
	BMI & DecT	0.29	0.02
	SBP & AEF	0.29	0.02
IUGR	BWT & DBP	-0.36	0.01
	HA & BWT	-0.24	0.08
Clustering method: k-means			
HEART	BSA & PWDs	0.55	0.01
	BWT & IVSd	-0.47	0.03
	BMI & PWDs	0.62	<0.01
ECHO	TG & E/A	-0.45	0.03
	BMI & DecT	0.38	0.01
	SBP & AEF	0.59	<0.01
IUGR	BWT & DBP	-0.38	0.01
	HA & BWT	-0.54	0.04
Clustering method: EM			
HEART	BSA & PWDs	0.51	0.01
	BWT & IVSd	-0.45	0.03
	BMI & PWDs	0.61	<0.01
ECHO	TG & E/A	-0.44	0.01
	BMI & DecT	0.35	0.05
	SBP & AEF	0.46	0.01
IUGR	BWT & DBP	-0.58	0.01
	HA & BWT	-0.26	0.05

the proposed hybrid method allowed to discover relationships which have not been identified previously.

Depending on the dataset, the growth of 10% - 100 % in the number of correlations between diagnosed attributes was obtained. Moreover all the calculated statistically significant dependencies were stronger in clusters rather than in the whole datasets.

The results of the presented investigations can be further implemented in practical diagnostic applications and can constitute the basis for improved medical inference described in [13], [14] and [15].

Future research will consist in developing the proposed methodology by considering some additional problems connected with medical data analysis including data gathering, data quality assurance, feature selection and outliers detection. The last one is especially worth considering as statistical analysis results are deviation sensitive. Therefore, the data need to be checked for outlier instances before proceeding with the analysis. The problem of handling outliers can be considered separately, or can be included as part of clustering process, but in such a case cluster analysis algorithm which deals with outliers should be implemented - random sample consensus (RANSAC) algorithm is regarded as giving satisfactory results [29].

Feature selection plays the crucial role in classification analysis. Although the choice of clustering attributes was carefully examined by medical experts and was consistent with the research presented in [1], we cannot exclude the possibility that considering other attributes may produce new meaningful conclusions. Therefore in future investigations we intend to verify this approach with different automatic feature selection methods, including genetic algorithms [30].

The analysis carried out as part of the current study involved data from laboratory tests, medical observations and their interpretations. However, the data for the analysis can be acquired by imaging studies. For the specified cardiac system the future analysis will concern SPECT images, ECG and EEG signals. Additional further studies will focus on efficient mining in medical imaging data and binding them with any other numerical and text data.

REFERENCES

[1] S. U. Amin, K. Agarwal and R. Beg: "Data Mining in Clinical Decision Support Systems for Diagnosis", Prediction and Treatment of Heart Disease. Int J Adv Res Comput Eng Technol (IJARCET), 2008 vol. 2(1), pp. 218-223

[2] S.W. Looney and J.L. Hagan: "Statistical Methods for Assessing Biomarkers and Analyzing Biomarkers Data." In: C.R. Rao, J.P. Miller, D.C. Rao (eds): Essential Statistical Methods for Medical Statistics, Elsevier, 2011, pp. 27-65

[3] D.C. Davies, T. Moxham, K. Rees, S. Singh, A.J. Coats, S. Ebrahim, F. Lough and R.S. Taylor: "Exercise based rehabilitation for heart failure", Cochrane Database Syst Rev, 2010 vol. 4(1, pp. 1-57, DOI: 10.1002/14651858.CD001800.pub2

[4] I. Yoo, P. Alafaireet and M. Marinov: "Data Mining in Healthcare and Biomedicine: A Survey of the Literature", J Med Syst, 2012, vol. 36, pp. 2431-2448, DOI: 10.1007/s10916-011-9710-5

[5] P. Haldar, I.D. Pavord, D.E. Shaw, M.A. Berry, M. Thomas, C.E. Brightling, A.J. Wardlaw and R.H. Green: "Cluster Analysis and Clinical Asthma Phenotypes", Am J Resp Crit Care, 2008, vol. 178, pp. 218-224, DOI: 10.1164/rccm.200711-1754OC

[6] X. Zhang, X. Zhou, R. Zhang, B. Liu, and Q. Xie: "Real-world Clinical Data Mining on TCM Clinical Diagnosis and Treatment: A Survey", e-Health Networking. Applications and Services (Healthcom), 2012 IEEE 14th International Conference on, DOI: 10.1109/HealthCom.2012.6380072

[7] A. Poliński, J. Kot A. Meresta: "Analysis of Correlation Between Heart Rate and Blood Pressure", In: IEEE Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS), 2011, pp. 417-420

[8] X.H. Meng, Y.X. Huang, D.P. Rao, Q. Zhang, and Q. Liu: "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors", Kaohsiung J Med Sci, 2013, vol. 29, pp. 93-99, DOI: http://dx.doi.org/10.1016/j.kjms.2012.08.016

[9] P.C. Austin, J.V. Tu, J.E. Ho, D. Levy and D.S. Lee: "Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes", J Clin Epidemiol, 2013, vol. 66, pp. 398-407, DOI:10.1016/j.jclinepi.2012.11.008

[10] S. Bashir, U. Quamar and M.Y. Javed: "An Ensemble based Decision Support Framework for Intelligent Heart Disease Diagnosis", In: International Conference on Information Society (i-Society), 2014, pp. 259-264, DOI: 10.1109/i-Society.2014.7009056

[11] M. Shouman, T. Turner and R. Stocker: "Using Data Mining Techniques in Heart Disease Diagnosis and Treatment", In: Japan-Egypt Conference on Electronics, Communications and Computers, 2012, pp. 173-177, DOI: 10.1109/JEC-ECC.2012.6186978

[12] B. Falkner, H. Kushner, G. Onesti and E.T. Angelakos: "Cardiovascular characteristics in adolescents who develop essential hypertension" Hypertension, 1981, vol. 3(5), pp. 521-527, DOI: 10.1161/01.HYP.3.5.521
[13] J. Zamojska, K. Niewiadomska-Jarosik, A. Wosiak, and J. Stanczyk: "Evaluation of left ventricular systolic function with the use of tissue Doppler echocardiography in children with primary arterial hypertension" ("Ocena funkcji skurczowej lewej komory z wykorzystaniem metody doplera tkankowego u dzieci z nadciśnieniem tetniczym pierwotnym"). Pol J Cardiol, 2012, vol. 4(2), pp.95-100 (in Polish)
[14] J. Zamojska, K. Niewiadomska-Jarosik, A. Wosiak, P. Lipiec, and J. Stanczyk: "Myocardial dysfunction measured by tissue Doppler echocardiography in children with primary arterial hypertension", Kardiol Pol (Polish Heart Journal), 2015, vol. 73(3), pp. 194-200, DOI: 10.5603/KP.a2014.0189
[15] A. Zamecznik, K. Niewiadomska-Jarosik, A. Wosiak, J. Zamojska, J. Moll and J. Stanczyk: "Intra-uterine growth restriction as a risk factor for hypertension in children six to 10 years old", Cardiovasc J Afr, vol. 25(2), 2014, pp. 73-77, DOI: dx.doi.org/10.5830/CVJA-2014-009
[16] J. Feber and M. Ahmed: "Hypertension in children: new trends and challenges", Clin Sci, 2010, vol. 119, pp. 151âĂ¿161, DOI: 10.1042/CS20090544
[17] G. Chandrashekar and F. Sahin: "A survey on feature selection methods", Computers and Electrical Engineering, vol. 40, 2014, pp. 16-28, DOI: dx.doi.org/10.1016/j.compeleceng.2013.11.024
[18] T.H. Cheng, C.P. Wei and V.S. Tseng: "Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches", Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems 2006, pp. 165 - 170, DOI: 10.1109/CBMS.2006.87
[19] J. Han, M. Kamber and J. Pei: "Data Mining: Concepts and Techniques", Elsevier, USA, 2011
[20] H. Liu, and L. Yu: Toward Integrating Feature Selection Algorithms for Classification and Clustering, IEEE T Knowl Data En, 2005, vol. 17, pp. 491-502, DOI: 10.1109/TKDE.2005.66
[21] I.H. Witten, E. Frank and M.A. Hall: "Data Mining. Practical machine learning tools and techniques", Morgan Kaufmann, San Francisco, USA, 2011
[22] A.P. Dempster, N.M. Laird and Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm". J R Stat Soc, 1977, vol. 39(1), pp. 1-38
[23] A.T. Azar, S.A. El-Said and A.E. Hassanien: "Fuzzy and hard clustering analysis for thyroid disease", Comput Meth Progr Bio, 2013, vol. 111(1), pp. 1-16, DOI: 10.1016/j.cmpb.2013.01.002
[24] Y.F. Wang, M.Y. Chang, R.D. Chiang, L.J. Hwang, C.M. Lee and Y.H. Wang: "Mining Medical Data: A Case Study of Endometriosis", J Med Syst, 2013, vol. 37:9899, DOI: 10.1007/s10916-012-9899-y
[25] N. Esfandiari, M.R. Babavalian, A.M.E. Moghadam and V.K. Tabar: "Knowledge discovery in medicine: Current issue and future trend", Expert Sys Appl, 2014, vol. 41(9), pp. 4434-4463, DOI: 10.1016/j.eswa.2014.01.011
[26] D.G. Altman and J.M. Bland: "Measurement in Medicine: the Analysis of Method Comparison Studies", The Statistician 32, 1983, pp. 307-317
[27] D.E. Hinkle, W. Wiersma and S.G. Jurs: Applied Statistics for the Behavioral Sciences. 5th ed. Boston: Houghton Mifflin, 2003
[28] F. Figueras and J. Gardosi: "Intrauterine growth restriction: new concepts in antenatal surveillance, diagnosis, and management", Am J Obstet Gynecol, 2011, vol. 204.4, pp. 288-300, DOI: 10.1016/j.ajog.2010.08.055
[29] M.T. El-Melegy: "Model-wise and point-wise random sample consensus for robust regression and outlier detection". Neural Netw, 2014, vol. 59, pp. 23-35, DOI:10.1016/j.neunet.2014.06.010
[30] M.A. Jabbar, P. Chandra and B.L. Deekshatulu: "Prediction of Risk Score for Heart Disease using Associative Classification and Hybrid Feature Subset Selection". 12th International Conference on Intelligent Systems Design and Applications (ISDA), 2012, DOI:10.1109/ISDA.2012.6416610