Hindawi
Complexity
Volume 2018, Article ID 2520706, 11 pages
https://doi.org/10.1155/2018/2520706

WILEY | Hindawi

Research Article

Integrating Correlation-Based Feature Selection and Clustering for Improved Cardiovascular Disease Diagnosis

Agnieszka Wosiak and Danuta Zakrzewska

Institute of Information Technology, Lodz University of Technology, 90-924, Poland

Correspondence should be addressed to Danuta Zakrzewska; danuta.zakrzewska@p.lodz.pl

Received 20 April 2018; Accepted 17 September 2018; Published 14 October 2018

Guest Editor: Ireneusz Czarnowski

Based on the growing problem of heart diseases, their efficient diagnosis is of great importance to the modern world. Statistical inference is the tool that most physicians use for diagnosis, though in many cases it does not appear powerful enough. Clustering of patient instances allows finding out groups for which statistical models can be built more efficiently. However, the performance of such an approach depends on the features used as clustering attributes. In this paper, the methodology that consists of combining unsupervised feature selection and grouping to improve the performance of statistical analysis is considered. We assume that the set of attributes used in clustering and statistical analysis phases should be different and not correlated. Thus, the method consisting of selecting reversed correlated features as attributes of cluster analysis is considered. The proposed methodology has been verified by experiments done on three real datasets of cardiovascular cases. The obtained effects have been evaluated regarding the number of detected dependencies between parameters. Experiment results showed the advantage of the presented approach compared to other feature selection methods and without using clustering to support statistical inference.

1. Introduction

Nowadays, data play a very important role in medical diagnostics since, due to equipment development, an increasing amount of data can be collected and thus, a huge volume of information concerning patient characteristics can be acquired. However, the possibilities of using data in medical diagnosis depend on the efficacy of the applied techniques. In practise, medical diagnostics are mainly supported by statistical inference, though in many cases it does not appear effective enough. It is worth emphasising that in medicine, the results of analysis are expected to be implemented in real life and thus the efficiency and usefulness of the methods should be taken into consideration. To obtain valuable recommendations for diagnostic statements, more sophisticated analytical methods are required. Including data mining algorithms to the process seems to be appropriate. Those techniques were recognized as efficient by Yoo et al. [1], who indicated that the application of descriptive and predictive methods are useful in biomedical as well as healthcare areas. In addition, stand-alone statistical analysis cannot be supportive in many cases, especially when correlations between attributes, considered as important by physicians, cannot be found. Such situation usually takes place for datasets of great standard deviation values [2]. What is more, dissimilarities or inconsistencies within the datasets can appear due to incorrect measurements or distortions. The presence of these kinds of deviations may lead to the rejection of true hypothesis; for example, such situation takes place when datasets are of small sizes. In these cases, supporting medical diagnosis becomes a complicated task, particularly when the number of attributes exceeds the number of records.

Integrating statistical analysis and data mining may not only improve the effectiveness of the obtained results, but also, by finding new dependencies between attributes, enable a multiperspective approach to medical diagnosis.

The research concerning the integration of cluster analysis and statistical methods on medical data, for defining the phenotypes of clinical asthma, has been presented in [3].

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 a heterogeneous asthma population. Data mining methods have been used in several clinical data systems. A survey of these systems and the applied techniques has been presented in [4]. Data mining techniques have been also considered in different clinical decision support systems for heart disease prediction and diagnosis in [2]. However, in the investigation results, the authors stated that the examined techniques are not satisfactory enough. Moreover, a solution for the identification of treatment options for patients with heart diseases is still lacking. Statistical inference of heart rate and blood pressure was investigated in [5]. The authors examined the correlation between raw data, then they examined the correlation between filtered data, and finally they applied the least squares approximation. In all the cases, the obtained correlation coefficients seemed to be unpredictable random numbers.

In this paper, we examine combining statistical inference and cluster analysis as a methodology supporting cardiovascular medical diagnosis. Including clustering in the preprocessing phase allows identifying groups of similar instances, for which respective parameters can be evaluated efficiently and thus statistical models of good quality can be created. Such an approach has been proposed in [6] to improve the performance of statistical models in hypertension problems in cardiovascular diagnosis. In the paper [7], a new reversed correlation algorithm (RCA) of an automatic unsupervised feature selection complemented the methodology. The RCA algorithm consisted of choosing subsequent features as the least correlated with their predecessors.

In the current research, we introduce a modification to the RCA that concerns the choice of the first attribute. Moreover, we extend the study [7] by comparing the performance of the considered algorithm with two other feature selection methods: correlation-based CFS and ReliefF. We also examine the effectiveness of the presented methodology regarding not only the statistical approach, but also the deterministic clustering algorithm with elbow criterion for determining the best number of clusters. Additionally, during the experiments we broaden the range of patients involved by changing the considered datasets. In the current research, instead of one of the three datasets gathered from children [7], we use a reference "CORONARY" dataset with a higher number of patient records. The dataset was derived from the UCI repository [8].

In this paper, we validate the performance of the investigated methodology applied to datasets of real patient records via numerical experiments. We consider three datasets of different proportions between the numbers of instances and attributes. The experimental results are evaluated by statistical inference performed on clusters. The results demonstrate that the statistical inference performed on clusters enable detection of new relationships, which have not been discovered in the whole datasets; thus, significant benefits of using the proposed hybrid approach for improving medical diagnosis can be recognized. The proposed feature selection algorithm outperforms the effects obtained by other considered techniques. As in all the analysed cases, we attained the best results regarding the numbers of discovered dependencies.

The remainder of the paper is organised as follows. In the next section, the cardiovascular disease diagnosis problem is introduced and the whole methodology is described including its overview, the RCA feature selection, and all the considered algorithms. Next, the experiments carried out for the methodology evaluation are presented regarding the dataset characteristics, and the results obtained at all the stages of the proposed method are discussed. The final section presents the study's conclusions and delineates future research.

2. Materials and Methods

2.1. Heart Disease Diagnosis Problem. The detection and diagnosis of heart diseases are of great importance due to their growing prevalence in the world population. Heart diseases result in severe disabilities and higher mortality than other diseases, including cancer. They cause more than 7 million deaths every year [9, 10].

Heart diseases include a diverse range of disorders: coronary artery diseases, stroke, heart failure, hypertensive heart disease, rheumatic heart disease, heart arrhythmia, and many others. Therefore, the detection of heart diseases from various factors is a complex issue, and the underlying mechanisms vary, depending on the considered problem and the conditions that affect the heart and the whole cardiovascular system. Moreover, there are many additional socioeconomic, demographic, and gestational factors that affect heart diseases, and are considered as their main reasons [11–13].

To improve early detection and diagnosis of heart abnormalities, new factors and dependencies that may indicate cardiovascular disorders are searched. Statistical data analysis supports the evaluation of the characteristics of the parameters in medical datasets and helps in discovering their mutual dependencies. However, in some situations the significance of statistical inference between medical attributes may be interfered by a wide range of values, subsets of relatively dissimilar instances, or outliers. Thus, there is a strong need for new techniques that will support statistical inference in finding parameter dependencies and thereby improve medical diagnosis.

2.2. The Method Overview. The considered methodology for supporting the process of medical diagnosis by patient dataset analysis consists of three main steps. They are preceded by data preparation, which aims at adjusting original datasets to analysis needs. The proposed steps can be presented as follows:

(1) Feature selection, based on statistical analysis of correlation coefficients, which enables appointing the set of attributes for clustering

(2) Finding groups of similar characteristics, including a validation technique used to determine the appropriate number of clusters

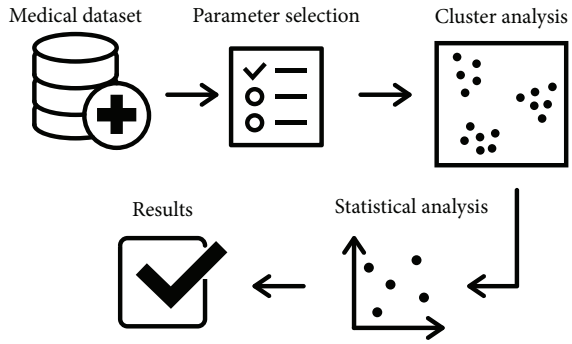

Figure 1: Overview of the methodology.

(3) Statistical analysis performed in clusters to find new dependencies between all the considered parameters

The general overview of the method is shown in Figure 1. We assume that clustering and statistical analysis are applied on the separate subsets of attributes. The descriptions of the main steps of the methodology are presented in Subsections 2.3, 2.4, and 2.5.

2.3. Feature Selection. Patient records usually contain many attributes that may be used for supporting medical diagnosis. However, the performance of the diagnostics process may depend on the choice of the attributes in all the phases of the considered methodology. The quality of results obtained in the final step depends not only on the choice of parameters used for finding correlations, but also depends on the quality of patient groups and thus on the subset of attributes used in the clustering process. Therefore, the process of feature selection for cluster analysis is crucial for the whole presented methodology of medical diagnosis.

Regarding the main supporting tool, which is a statistical inference according to physician preferences, we propose the reversed correlation algorithm (RCA) that uses correlation coefficients but in a reversed order. This means that we look for features that are the least correlated with all their predecessors.

First, we start building a subset of features with the attribute that is the least correlated with the others. Then, correlation coefficients between the chosen feature and the rest of the parameters are calculated. The attribute with the lowest correlation value is indicated as the second feature. The obtained subset of two features is further extended by adding the attribute of the correlation coefficient with the lowest value between the subset and the rest of the parameters. The process of appending the features of the lowest correlation values is repeated unless all the correlation coefficients indicate statistically significant dependencies (respective values exceed thresholds) or the number of features in the subset is equal to the determined percentage of the total number of attributes. The whole procedure is presented in the Algorithm 1.

In order to compare the results of the proposed feature selection algorithms, two other techniques have been considered: the opposite approach represented by the correlation-based feature selection (CFS) and an extension of the relief algorithm called ReliefF.

Correlation-based feature selection (CFS) ranks attributes according to a heuristic evaluation function based on correlations [14]. The function evaluates subsets made of attribute vectors, which are correlated with the class label, but independent of each other. The CFS method assumes that irrelevant features show a low correlation with the class and therefore should be ignored by the algorithm. On the other hand, excess features should be examined, as they are usually strongly correlated with one or more of the other attributes. The criterion used to assess a subset of l features can be expressed as follows:

$$M_S = \frac{l\overline{t_{cf}}}{\sqrt{l + l(l-1)\overline{t_{ff}}}}, \quad (1)$$

where M_S is the evaluation of a subset of S consisting of l features, $\overline{t_{cf}}$ is the average correlation value between features and class labels, and $\overline{t_{ff}}$ is the average correlation value between two features.

There exist different variations of CFS that employ different attribute quality measures, such as Symmetrical Uncertainty, normalized symmetrical Minimum Description Length (MDL), or Relief.

Relief algorithm, described in [15], concerns the evaluation of attributes based on the similarity of the neighbouring examples in the set of analysed instances [16]. For the given set of training instances, sample size, and the relevancy threshold τ, Relief detects features that are statistically consistent with the target task. Relief picks an instance X from the set and its two nearest neighbours: one of the same class—called "near-hit" and one of the opposite class—called "near-miss". Then, it updates the feature weight vector W for every triplet and uses it to determine the average relevance feature vector. The algorithm selects those features for which the value of the average weight, called relevance level, exceeds the given threshold value τ.

ReliefF algorithm has been proposed in [16]. Contrary to Relief, it is not limited to two class problems, it is more effective and can deal with noisy or incomplete data, for missing values of attributes are treated probabilistically. Similarly to Relief, ReliefF randomly selects an instance X, but it searches for the determined number of the nearest neighbours from the same class, called "nearest hits," and the same number of the nearest neighbours from every different class ("nearest misses"). Then, it updates the vector W of estimations of the qualities for all the attributes depending on their values for X and sets of hits and misses.

2.4. Cluster Analysis. Cluster analysis is an unsupervised classification technique, which can be used for grouping complex multidimensional data. Opposite to supervised methods, the profiles of obtained groups cannot be obviously stated and using additional techniques for discovering the meaning of clustering is required in many cases [17]. On the other side, statistical analysis is the most popular tool used in the medical field. Therefore, in this area, combining clustering and

Input: $F = f_1, f_2, f_3, \ldots f_n$ */* set of all the features */;*
P/ statistical significance level */;*
R/ a threshold for correlation coefficient levels */;*
N/ the maximum of features for the subset/*;*
Output: F_s */* selected subset of features */;*
(1) Initialize F_s *with feature* $f_j \in F$ *that is the least correlated with other ones;*
(2) do
(3) Compute $C_{ij}(F_s, F \setminus F_s)$ *as a vector of correlation coefficients between* F_s *and each* $f_i \in \{F \setminus F_s\}$*;*
(4) Choose $f_j \in \{F \setminus F_s\}$ *with the lowest value of correlation coefficient in a vector* $C_{ij}(F_s, F \setminus F_s)$*;*
(5) Include f_j *in* F_s
(6) while $(s < N$ *AND* $p > P$ *AND* $C_{ij}(F_s, F \setminus F_s) < R)$.

Algorithm 1: Proposed feature selection algorithm using reversed correlations

statistical inference may not only enable patient grouping, but also finding dependencies between their characteristics and thus supporting medical diagnostics.

In further investigations, which aim at evaluating the presented technique regarding its efficiency on cardiovascular data, simple popular clustering algorithms will be considered, for such techniques are expected to be comprehensible for physicians.

We will examine two different clustering approaches: deterministic and probabilistic. The first approach will be represented by *k*-means algorithm, which in comparison to other techniques, demonstrated good performance for medical data regarding accuracy as well as lower root mean square error [18]. The *k*-means algorithm is one of the most popular partitioning methods, where clusters are built around *k* centers, by minimizing a distance function. The goal of the algorithm is to find the set of clusters for which the sum of the squared distance values between their points and respective centers is minimal. As the distance function, the Euclidean metric is used, which has been applied in most of the cases [19, 20]. The first *k* centers are usually chosen at random, which does not guarantee finding optimal clusters. To increase the chance of finding the optimum, the algorithm is usually launched several times with different initial choices and the result of the smallest total squared distance is indicated [20].

The goal of a statistical model is to find the most probable set of clusters on the basis of training data and prior expectations. As a representative of these techniques, EM (expectation-maximization) algorithm, based on the finite Gaussian mixtures model, has been investigated. EM generates probabilistic descriptions of clusters in terms of means and standard deviations [17]. The algorithm iteratively calculates the maximum likelihood estimated in parametric models in the presence of missing data [21]. EM enables using cross-validation for selecting the number of clusters and thus obtaining its optimal value [20]. That feature allows avoiding the determination of the number of clusters at the beginning of the algorithm.

The choice of the optimal number of clusters is one of the most important parts of the clustering process. In the case of the *k*-means algorithm, the elbow technique was used. It is based on the statement that the number of clusters should increase together with the increase of the quantity of information. The last number of clusters, for which a gain value was augmented, should be indicated as optimal. On the graph, where validation measure is plotted against the number of clusters, that point is presented as an angle, and called the elbow. There are cases, when angles cannot be unambiguously identified, and the number of clusters indicated by the elbow technique should be confirmed by other methods.

Thus, considering two clustering methods equipped with different techniques for choosing the optimal number of clusters may help in confirming the right choice. However, it is worth noticing that in medicine there exists the usual intent to split the whole dataset into two groups and thus the number of clusters is very often equal to two [18]. Besides, in medical applications, the number of collected instances is very small and the high number of clusters may result in small group sizes and in less reliable medical inference, as the consequence of the lack of statistical tests of high power [19, 22].

2.5. Statistical Analysis. Before carrying out statistical inference, the assessment of measures of descriptive statistics should be performed. Such an approach allows detecting errors that were not identified during the data preparation phase. As the main descriptors, for which the evaluation is indicated, one should mention central tendency measures (arithmetic mean, median, and modal) as well as dispersion measures (range and standard deviation). Next, an appropriate test is run as a part of the statistical analysis process. The test should be chosen according to the type and the structure of analysed data regarding such characteristics as attribute types, the scale type, the number of experimental groups, and their dependencies, as well as the test power. Additionally, the selection should be consistent with the requirements of the USMLE (The United States Medical Licensing Examination). In the presented research, these are considered the tests usually applied in medical diagnostics [2]:

(i) Kolmogorov–Smirnov test, which is used to check the normality of distribution of the attributes

(ii) Unpaired two-sample Student's *t*-test for the significance of a difference between two normally distributed values of attributes

(iii) Mann–Whitney U test, which is a nonparametric test for the determination of significant differences, where attributes are in nominal scales

Pearson's correlation coefficient $r_P(x, y)$ is used to express the impact of one variable measured in an interval or ratio scale to another variable in the same scale. Spearman's correlation $r_S(x, y)$ test is used, in the case when 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.

3. Results and Discussion

The performance of the proposed methodology has been examined by experiments conducted on the real datasets collected for supporting heart disease diagnosis. The statistical analysis results obtained for clusters have been compared with the ones taken for the whole datasets.

3.1. Data Description. The experiments were carried out on three datasets:

(i) "HEART"

(ii) "IUGR"

(iii) "CORONARY"

The "HEART" dataset consisted of 30 cases collected to discover dependencies between arterial hypertension and left ventricle systolic functions. The "IUGR" dataset includes 47 instances of children born with intrauterine growth restriction (IUGR), gathered to find out dependencies between abnormal blood pressure and being born as small for gestational age. The data of both of the datasets were collected in the Children's Cardiology and Rheumatology Department of the Second Chair of Paediatrics at the Medical University of Lodz.

Each dataset was characterized by two types of parameters: the main and the supplementary ones, all of them gathered for discovering new dependencies. The attributes correspond to high blood pressure and include echocardiography and blood pressure assessment, prenatal and neonatal history, risk factors for IUGR, and family survey of cardiovascular disease, as well as nutritional status. There were no missing values within the attributes. The full medical explanations of the data are given in [13, 23].

The "CORONARY" dataset also refers to cardiovascular problems. It comes from the UCI Machine Learning Repository [8]. The dataset contains the records of 303 patients, each of which is described by 54 features. The attributes were arranged in four groups of features: demographic, symptom and examination, and ECG, as well as laboratory and echo ones [24–26].

The summary of characteristics for all the datasets was presented in Table 1. The datasets have been chosen to ensure diversification of the mutual proportion between the number of instances and the number of attributes:

Table 1: The characteristics of datasets.

Dataset	Instances	Main attributes	Supplementary attributes
HEART	30	14	35
IUGR	47	6	40
CORONARY	303	10	44

(i) The number of instances in the "HEART" dataset is smaller than the number of parameters

(ii) The number of instances in the "IUGR" dataset is comparable with the number of attributes

(iii) In the "CORONARY" dataset, the number of instances is greater than the number of parameters

Tables 2–4 describe the selection of the parameters with the main statistical descriptors: the values of range, median or mean, and standard deviation (SD).

3.2. Selecting Relevant Features. For each dataset, only parameters concerning main characteristics were considered as initial attributes used for grouping. The selection of the appropriate features for building clusters has been performed by using three different techniques:

(1) The reversed correlation algorithm (RCA)

(2) CFS method

(3) ReliefF algorithm

The parameters necessary to run the RCA algorithm were chosen according to principles commonly approved in statistics (see [24, 28]):

(i) $N = 50\%\ n$ for the maximal number of features

(ii) $R = 0.3$ for the maximal value of correlation coefficients

(iii) $P = 0.05$ for the maximal value of statistical significance p value

In the case of the ReliefF algorithm, the threshold for the number of attributes included in the subset of selected features was set to $N = 50\%\ n$.

The subsets of features presented in Table 5 were obtained as the results of the proposed feature selection process. The first column of the table represents names of datasets, the second column represents the names of the feature selection algorithms, and the following columns contain the number and names of selected features in the order indicated by the algorithms.

3.3. Data Clustering. In the next step of the experiments, the clusters for diagnosed patients were created by using two clustering algorithms: k-means and EM implemented by WEKA Open Source software [20].

Table 2: Characteristics of attributes for the "HEART" dataset.

Attribute(s)	Description	Range	Median/mean (mean range)	SD (SD range)
Main attributes				
BMI	Current body mass index	17.00 to 25.00	22.16	1.64
Birth_weight	Birth weight	2500 to 4000	3158	392.00
SBP, DBP, ABPM-SBP, ABPM-DBP	Average systolic/diastolic blood pressure taken manually and by ABPM	61 to 150	74.87 to 136.97	5.22 to 7.04
HR	Heart rate	44 to 91	75.97	11.20
Risk factors	Risk factors	True/false	—	—
Supplementary attributes				
IVSd, IVSs, PWDd, PWDs, LVDd, LVDs	Left ventricular dimensions	5.00 to 56.00	8.00 to 46.03	1.51 to 9.02
EF, SF	Systolic function	34 to 84	40 to 70	3 to 5
Sm, Sml, V/S/SR long/rad/circ	Tissue Doppler echocardiography parameters	−37 to 40.17	−27.25 to 29.64	0.42 to 6.35

Table 3: Characteristics of attributes for the "IUGR" dataset.

Attribute(s)	Description	Range	Median/mean (mean range)	Mode (N) or SD (SD range)
Main attributes				
Birth_weight	Birth weight	1980–2850	2556.70	2700 (7)
Head_circ	Head circumference	29–35	33	32 (16)
Gest_age	Gestational age	38–42	39	—
Apgar	Apgar score at 1 min	7–10	9	9 (23)
5_Percentile	Growth chart factor	True/false	—	False (25)
Supplementary attributes				
SBP, DBP	Average systolic/diastolic blood pressure	55–137	55–115	5.03–8.73
SBP load, DBP load	Blood pressure loads	0–96	9–20	10–21
LVm	Left ventricular mass (Simone, Devreux)	17.65–93.21	30.26–59.11	6.91–12.91
Risk factors	Risk factors	True/false	—	—

Clusters were built regarding the main characteristics and the parameters indicated by feature selection methods, namely RCA, CFS, and ReliefF.

In the case of the EM algorithm, the best number of clusters was indicated by using cross-validation. To choose the best number of clusters for k-means clustering, the elbow criterion has been applied and within cluster sum of squares has been considered as a validation measure. The charts of validation measures plotted against the number of clusters with marked elbow points for HEART, IUGR, and CORONARY datasets, respectively, are presented in Figures 2–4. For better result visualisation, the values of within cluster sum of squares were normalized.

The results of clustering are presented in Table 6, where the first column describes datasets, the second column contains the names of the feature selection methods, and the last two columns present the number of clusters and clustering schemes.

3.4. Statistical Inference. Correlation values obtained for the clusters were compared with the ones taken for the whole group of diagnosed patients in terms of different selection techniques. Comparison of results confirmed the effectiveness of the proposed methodology. For each dataset, we obtained a greater number of statistically significant correlations in clusters which may lead to improved medical diagnosis in the future. By significant correlations we mean values with correlation coefficient $r \geq 0.3$ and p value ≤ 0.05 ([27, 28]). The biggest growth of the number of correlations concerns the HEART dataset, where the number of instances is smaller than the number of parameters. The numbers of detected correlations are presented in Table 7.

One can easily notice that the results attained by the unsupervised RCA feature selection technique and supervised ReliefF algorithm were comparable; however, the first method outperforms the second one in the case of the IUGR dataset and k-means technique. As in many cases, the supervised technique of feature selection cannot be used due to the lack of information on labels; one can expect that the RCA method would be indicated as more often used than the ReliefF algorithm.

Table 4: Characteristics of attributes for the "CORONARY" dataset.

Attribute(s)	Description	Range	Median/mean (mean range)	Mode (N) or SD (SD range)
Main attributes				
Q wave, St elevation, St depression, Tinversion, LVH, poor R progression	ECG parameters	Yes/no	—	—
FBS	Fasting blood sugar	62–400	119	52
EF-TTE	Ejection fraction—transthoracic echocardiography	15–60	47	9
Region RWMA	Regional wall motion abnormalities	0–4	0 (217)	—
Supplementary attributes				
Age	Age	30–86	58.00	10.39
Weight	Weight	48–120	73.83	11.89
Sex	Sex	Male/female	—	Male (176)
BMI	BMI	18–41	27.25	4.10
DM, HTN, current smoker, ex-smoker, FH, obesity, CRF, airway disease, thyroid disease, CHF, DLP	Diabetes mellitus, hypertension, current smoker, ex-smoker, family history, obesity, chronic renal failure, cerebrovascular accident, airway disease, thyroid disease, congestive heart failure, dyslipidemia	Yes/no	—	—
Edema, weak peripheral pulse, lung rales, systolic murmur, diastolic murmur, typical chest pain, dyspnea	Symptom and examination parameters	Yes/no	—	—
Cr, TG, LDL, HDL, BUN, ESR, HB, K, Na, WBC, lymph, neut, PLT	Laboratory parameters (creatine, triglyceride, low density lipoprotein, high density lipoprotein, blood urea nitrogen, erythrocyte sedimentation rate, haemoglobin, potassium, sodium, white blood cell, lymphocyte, neutrophil, platelet)	0.5–18,000	1.05–7652.04	0.24–2413.74

Table 5: Feature selection results.

Dataset	FS algorithm	Size	Supplementary attributes
HEART	RCA	6	Physical_activity, fundus, BMI, HR, height, birth_weight
	CFS	1	Weight
	ReliefF	6	Physical_activity, family_interview, weight, fundus, height, BMI
IUGR	RCA	3	Apgar_score, ponderal_index, 5_percentile
	CFS	1	Birth_weight
	ReliefF	3	Head_circ, ponderal_index, birth_weight
CORONARY	RCA	4	FBS, EF-TTE, St depression, LVH
	CFS	5	Q wave, Tinversion, FBS, EF-TTE, region RWMA
	ReliefF	5	Region RWMA, Tinversion, St depression, St elevation, Q wave

4. Conclusions

The process of computer-aided medical studies is usually based on only one of the data analysis methods, most often a statistical approach. In this paper, we present an approach that integrates a feature selection technique and clustering with statistical inference, to improve medical diagnosis by finding out new dependencies between parameters. We consider using the new feature selection technique based on reversed correlations (RCA), combining it with two clustering algorithms: EM and k-means. We compare the RCA technique with two other feature selection methods: CFS and ReliefF. The comparison has been done by experiments carried out on real patient datasets. The experimental results are evaluated by a number of statistically significant correlations detected in clusters.

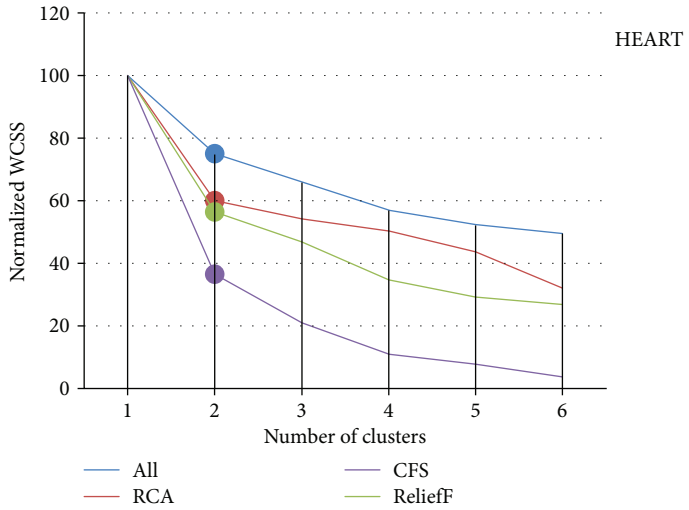

Figure 2: Validation of clustering for the HEART dataset.

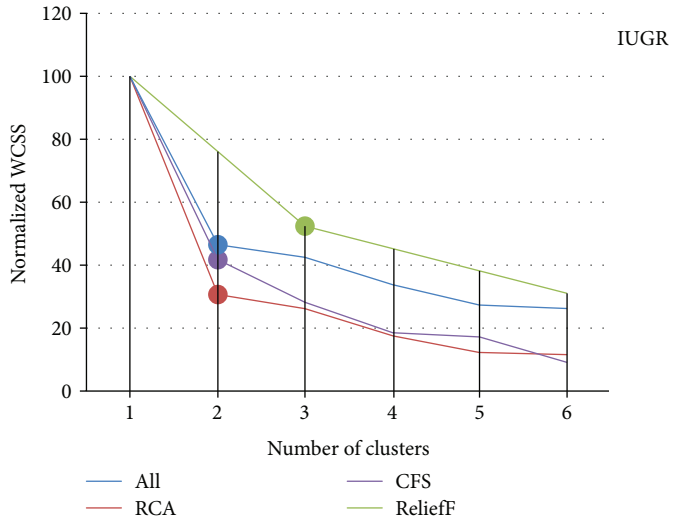

Figure 3: Validation of clustering for the IUGR dataset.

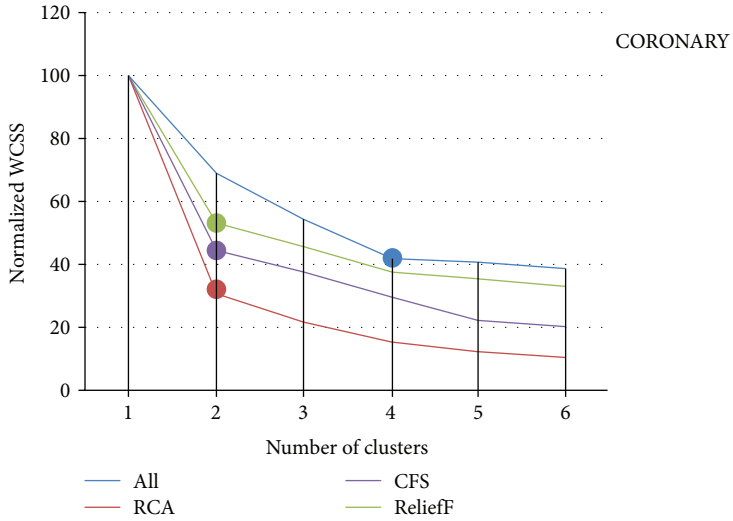

Figure 4: Validation of clustering for the CORONARY dataset.

Table 6: Clustering results.

Dataset (1)	FS algorithm (2)	Cluster algorithm (3)	No of clusters (4)	Clustering schema (5)			
HEART	Main attributes	EM	2	7	23		
		k-Means	2	8	22		
	RCA	EM	2	6	24		
		k-Means	2	6	24		
	CFS	EM	1	30			
		k-Means	2	11	19		
	ReliefF	EM	4	6	4	15	3
		EM	2	21	9		
		k-Means	2	6	24		
IUGR	Main attributes	EM	2	22	25		
		k-Means	2	22	25		
	RCA	EM	2	25	22		
		k-Means	2	25	22		
	CFS	EM	2	12	35		
		k-Means	2	16	31		
	ReliefF	EM	4	7	12	18	10
		k-Means	3	13	14	20	
CORONARY	Main attributes	EM	4	22	49	1	231
		k-Means	4	148	50	71	34
	RCA	EM	2	71	232		
		k-Means	2	232	71		
	CFS	EM	3	101	17	185	
		k-Means	2	213	90		
	ReliefF	EM	3	89	23	191	
		k-Means	2	213	90		

Table 7: Numbers of statistically significant correlations detected in the whole datasets and in clusters.

Dataset	Whole dataset	Main features		RCA		CFS		ReliefF	
		EM	*k*-Means	EM	*k*-Means	EM	*k*-Means	EM	*k*-Means
HEART	14	29	30	28	28	14	28	28	28
IUGR	11	15	15	16	16	11	11	16	15
CORONARY	14	15	20	16	16	15	16	16	16

The experiments have shown that the proposed hybrid approach provides significant benefits. The statistical inference performed in clusters enabled detection of new relationships, which have not been discovered in the whole datasets, regardless of the applied feature selection algorithm and the clustering technique. Moreover, the proposed RCA technique attained results at least as good as other considered feature selection methods, but as opposed to CFS and ReliefF, it belongs to unsupervised approaches, which implies a more flexible application. It is also worth emphasising that the presented approach has been checked using datasets of different mutual proportions between the number of instances and the number of attributes. The experimental results have shown that the proposed methodology performs well on datasets with the small number of instances and what is more, the biggest growth of the number of correlations concerns the dataset where the number of instances is smaller than the number of attributes. Such situations very often take place in the case of patient datasets.

Future research will focus on further investigations that aim at improving medical diagnostics by using hybrid approaches combining data mining and statistical inference. First, more datasets should be examined regarding different mutual proportions between the number of instances and the number of attributes. The research area should be broadened to diagnostics for the diseases of other types. Further research should also include indicating the effective integration of feature selection and clustering algorithms that will perform well combined with statistical inference.

Data Availability

The dataset "CORONARY" that supports the findings of this study is openly available at the UCI Machine Learning Repository at http://archive.ics.uci.edu/ml. The datasets "HEART" and "IUGR" are not publicly available due to ethical restrictions. The full medical description of the data can be found in [13, 23].

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

The authors received funding from the Institute of Information Technology, Lodz University of Technology.

References

[1] I. Yoo, P. Alafaireet, M. Marinov et al., "Data mining in healthcare and biomedicine: a survey of the literature," *Journal of Medical Systems*, vol. 36, no. 4, pp. 2431–2448, 2012.

[2] 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)*, vol. 2, no. 1, pp. 218–223, 2008.

[3] P. Haldar, I. D. Pavord, D. E. Shaw et al., "Cluster analysis and clinical asthma phenotypes," *American Journal of Respiratory and Critical Care Medicine*, vol. 178, no. 3, pp. 218–224, 2008.

[4] X. Zhang, X. Zhou, R. Zhang, B. Liu, and Q. Xie, "Real-world clinical data mining on TCM clinical diagnosis and treatment: a survey," in *2012 IEEE 14th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pp. 88–93, Beijing, China, October 2012.

[5] A. Poliński, J. Kot, and 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)*, pp. 417–420, Szczecin, Poland, 2011.

[6] A. Wosiak and D. Zakrzewska, "On integrating clustering and statistical analysis for supporting cardiovascular disease diagnosis," in *Proceedings of the 2015 Federated Conference on Computer Science and Information Systems, IEEE 2015, Annals of Computer Science and Information Systems*, vol. 5, pp. 303–310, Lodz, Poland, 2015.

[7] A. Wosiak and D. Zakrzewska, "Unsupervised feature selection using reversed correlation for improved medical diagnosis," in *2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, P. Jędrzejowicz, T. Yildirim, and P. Czarnowski, Eds., pp. 18–22, IEEE, Gdynia Poland, 2017.

[8] M. Lichman, "UCI machine learning repository," 2017, http://archive.ics.uci.edu/ml.

[9] E. Claes, J. M. Atienza, G. V. Guinea et al., "Mechanical properties of human coronary arteries," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pp. 3792–3795, Buenos Aires, Argentina, August-September 2010.

[10] E. D. Grech, "Pathophysiology and investigation of coronary artery disease," *BMJ*, vol. 326, no. 7397, pp. 1027–1030, 2003.

[11] C. J. Murray and A. D. Lopez, *The Global Burden of Disease: A Comprehensive Assessment of Mortality and Disability from Diseases, Injuries, and Risk Factors in 1990 and Projected to 2020: Summary*, Global Burden of Disease And Injury Series, World Health Organization, 1996.

[12] K. Niewiadomska-Jarosik, J. Zamojska, A. Zamecznik, A. Wosiak, P. Jarosik, and J. Stańczyk, "Myocardial dysfunction in children with intrauterine growth restriction: an echocardiographic study," *Cardiovascular Journal of Africa*, vol. 28, no. 1, pp. 36–39, 2017.

[13] A. Zamecznik, K. Niewiadomska-Jarosik, A. Wosiak, J. Zamojska, J. Moll, and J. Stańczyk, "Intra-uterine growth restriction as a risk factor for hypertension in children six to 10 years old: cardiovascular topic," *Cardiovascular Journal Of Africa*, vol. 25, no. 2, pp. 73–77, 2014.

[14] A. M. Hall, "Correlation-based feature selection for machine learning," Doctoral Disertation, University Of Waikato, Department of Computer Science, 1999.

[15] K. Kira and L. A. Rendell, "A practical approach to feature selection," *Machine Learning Proceedings*, pp. 249–256, 1992.

[16] I. KononenkoF. Bergadano and L. De Raedt, "Estimating attributes: analysis and extensions of RELIEF," in *European Conference on Machine Learning: ECML 1994: Machine Learning: ECML-94*, vol. 784 of Lecture Notes in Computer Science, pp. 171–182, Springer, Berlin, Heidelberg, 1994.

[17] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, Elsevier, USA, 2011.

[18] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491–502, 2005.

[19] S. W. Looney and J. L. Hagan, "Statistical methods for assessing biomarkers and analyzing biomarker data," in *Essential Statistical Methods for Medical Statistics*, C. R. Rao, J. P. Miller, and D. C. Rao, Eds., pp. 27–65, Elsevier, 2011.

[20] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, San Francisco, USA, 2011.

[21] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[22] 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," *Journal of Medical Systems*, vol. 37, no. 2, p. 9899, 2013.

[23] J. Zamojska, K. Niewiadomska-Jarosik, A. Wosiak, P. Lipiec, and J. Stańczyk, *Myocardial Dysfunction Measured by Tissue Doppler Echocardiography in Children with Primary Arterial Hypertension*, Kardiologia Polska, 2015.

[24] R. Alizadehsani, J. Habibi, M. J. Hosseini et al., "A data mining approach for diagnosis of coronary artery disease," *Computer Methods and Programs in Biomedicine*, vol. 111, no. 1, pp. 52–61, 2013.

[25] R. Alizadehsani, M. H. Zangooei, M. J. Hosseini et al., "Coronary artery disease detection using computational intelligence methods," *Knowledge-Based Systems*, vol. 109, pp. 187–197, 2016.

[26] Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, and A. A. Yarifard, "Computer aided decision making for heart disease detection using hybrid neural network-genetic

algorithm," *Computer Methods and Programs in Biomedicine*, vol. 141, pp. 19–26, 2017.

[27] D. G. Altman and J. M. Bland, "Measurement in medicine: the analysis of method comparison studies," *The Statistician*, vol. 32, no. 3, pp. 307–317, 1983.

[28] D. E. Hinkle, W. Wiersma, and S. G. Jurs, *Applied Statistics for the Behavioral Sciences*, Houghton Mifflin, Boston, 5th Ed edition, 2003.