Unsupervised Feature Selection Using Reversed Correlation for Improved Medical 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 inference has been usually used for medical data analysis, however in many cases it appears not to be efficient enough. Cluster analysis enables finding out groups of similar instances, for which statistical models can be built more effectively. In the paper a feature selection method for finding clustering attributes, which are supposed to improve performance of statistical analysis, is proposed. The method consists in selecting reversed correlated features as attributes of cluster analysis. The proposed technique has been evaluated by experiments done on real data sets of cardiovascular cases. Experiment results showed that the presented approach stimulates efficacy of statistical inference applied to medical diagnosis.**

***Keywords*—feature selection, reversed correlation, medical data, clustering**

I. Introduction

Statistical analysis of constantly growing amount of medical data is supposed to help practitioners in improving patient care, as well as proposing new therapies or developing the existing ones. However, in many cases, it turns out to be not effective enough. Such a situation takes place, when a considered group of patients is characterized by large values of standard deviations within parameters. As a result, the correlations between parameters, which seem to be useful for medical inference, are not possible to obtain [1].

To improve the performance of statistical models, a hybrid approach, which consisted of including clustering in the preprocessing phase of statistical analysis, has been presented in [2]. The investigations have shown that supporting statistical analysis by clustering provides significant benefits. Experiments conducted on the real data have demonstrated that the proposed hybrid method allowed to discover relationships which have not been identified previously. However the effects depended on the choice of attributes, which were used in the clustering process. Thus special attention should be drawn to feature selection process, which is usually conducted by experts with the necessary domain knowledge. Moreover, in many cases the results of an unsupervised feature selection cover or even improve the expert judgment [3], [4].

In the paper, we propose a new method of automatic unsupervised feature selection for clustering purpose. The proposed technique aims at choosing subsequent features as the least correlated with their predecessors. Consequently we build a subset of features strongly different from each other. Such a group of attributes is an initial set for clustering performed as a second step of our methodology. Building clusters enables to identify groups of similar instances, for which statistical models can be built effectively.

We focus on cardiovascular diseases, which are the leading causes of death in the majority of countries [5]. The proposed methodology was verified by experiments done on three different sets of real children cases. The results showed that automated feature selection based on reversed correlation analysis effectively supports clustering and consequently statistical inference for the diagnosis in a considered cardiovascular problem.

The rest of a paper is organized as follows. In Section II relevant work is presented. Next, in Section III, a new algorithm of feature selection is described. Section IV is dedicated to the experiments conducted on real data. Finally, in Section V, the results and some concluding remarks are discussed.

II. Related Work

Automatic feature selection methods were objects of interests of many researchers. Their main objective consists in providing generic introduction to variable elimination, which can be applied to a wide range of machine learning problems. A survey of feature selection methods was presented in [6]. Correlation based feature selection (CFS) has been proposed in [7]. The method considers correlation between a feature and a class and inter-correlation between features.

Several variants of CFS has been proposed so far. From among them one should mention the best-first-CFS method based on the best first search strategy to find the locally optimal subset of features by means of the CFS measure or genetic algorithm based CFS method [8]. Nguyen et al. proposed the method of finding the globally optimal subset of relevant features by means of CFS measure [9]. These algorithms were used for intrusion detection purpose. Deng et al. in turn investigated feature selection methods using both concepts of correlation and diversity. They applied feature combination methods to the selected features for detecting the stress levels [10]. A metric integrating the correlation and

978-1-5090-5795-5/17/$31.00

reliability information between each feature and each class obtained from Multiple Corresponcence Analysis was introduced in [11]. The proposed framework has been evaluated on highly inbalanced video concepts data. Application of CFS on bioinformatics datasets was presented in [12].

III. Methodology

The proposed feature selection method is one of the steps of the methodology of statistical analysis for medical diagnostics, that is based on clustering. The process starts with data preparation, which results in the initial dataset. Then feature selection based on statistical analysis of correlation coefficients, which enables choosing the set of attributes for building clusters is carried out. Next, groups of similar characteristics are distinguished, where the best clustering schema is chosen by cluster validation technique. Finally statistical analysis, which aims at finding new dependencies between all the collected parameters, is performed in clusters .The description of the main steps, excluding data preparation is presented with details in [2].

A. Feature Selection Method

The process of feature selection is crucial for the whole methodology of medical diagnosis as the future inference depends on groups of patients distinguished taking into account a particular subset of attributes. We propose an algorithm that uses correlation coefficients but in reversed order, i.e. we search for features least correlated with all their predecessors. The method is presented in Algorithm 1.

Input: $F = f_1, f_2, f_3, ...f_n$ /* set of all 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
Initialize F_s with a randomly chosen feature $f_j \in F$;
do
1 Compute $C_{ij}(F_s, F \setminus F_s)$ as a vector of correlation coefficients between
F_s and each $f_i \in \{F \setminus F_s\}$;
2 Choose $f_j \in \{F \setminus F_s\}$ with the lowest value of correlation coefficient in a
vector $C_{ij}(F_s, F \setminus F_s)$;
3 Include f_j in F_s
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

First we start building a subset of features with a randomly chosen attribute taken from the whole set of parameters (see Algorithm 1, step 1). Then we compute the correlation coefficients between that feature and the rest of parameters (step 2). We choose the second feature as the one that resulted in the lowest value of correlation (step 3). Now our subset consists of two features. We perform another iteration of the algorithm by calculating vector of correlation coefficients between the subset of selected features and the rest of parameters. Once again we choose the attribute with the lowest value of correlation and append it to the subset. We repeat the steps unless all correlation coefficients indicate statistically significant dependencies (values of p and C_{ij} exceeded thresholds) or the number of features in the subset equals a stated percentage of the total set of attributes.

B. Clustering and Statistical Inference

Cluster analysis is one of the most commonly used data mining techniques, however in many cases, when the groups are detected, it is necessary to use other methods to discover the meaning of clustering [13]. Therefore, combination of cluster analysis and statistical inference seems to be the effective tool supporting medical diagnosis.

In this paper an expectation-maximization (EM) algorithm is considered as a clustering technique. EM uses the finite Gaussian mixtures model to generate probabilistic descriptions of clusters in terms of means and standard deviations [13]. Therefore such choice of the clustering algorithm for combination with statistical interference seems to be natural. What is more, a big advantage of EM algorithm is its possibility to select a number of clusters by cross validation techniques, what allows to obtain its optimal value [14]. Thus it is not necessary to determine the number of clusters at the beginning.

The results of clustering depend on input parameters and the structure of data sets, their evaluation can be done by computation of cluster validity indices. Many of them have been proposed in [16], [17]. One main type of the indices are based on sum-of-squares. According to variance analysis, the total sum-of-squares (SST) can be decomposed into two parts: within cluster sum-of-squares (SSW) and between cluster sum-of-squares (SSB). As it has been proved in [17], the trends of normalized SSW and SSW/SSB are almost the same, indicating that the factor of the SSW has a key effect in the ratio of SSW/SSB. Therefore only SSW will be further considered.

The value of SSW is defined as:

$$SSW(C, m) = \frac{1}{n} \sum_{i=1}^{m} \sum_{j \in C_i} \|x_j - C_{P(j)}\| \qquad (1)$$

where:

$X = x_1, x_2, \ldots, x_n$ is a set of data with n samples,
$C = C_1, C_2, \ldots, C_m$ are m non-overlapping clusters,
$P = P_1, P_2, \ldots, P_m$ is the optimal partition.

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). 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

TABLE I
THE CHARACTERISTICS OF DATASETS

Dataset	Instances	Main attributes	Supplementary attributes
HEART	30	14	35
ECHO	66	13	9
IUGR	47	6	40

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 $r_S(x, y)$ coefficient is used.

IV. 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. Each dataset was characterized by main parameters and supplementary attributes gathered for discovering new dependencies. There were no missing values within the attributes. The characteristics of datasets were presented in Table I.

B. Feature Selection

For each dataset, only parameters concerning main characteristics were considered as initial sets of attributes. Then we performed feature selection according to the proposed methodology presented in Section III-A. The thresholds for the algorithm were chosen as follows (see [19], [20]): $N = 50\%n$ for the maximal number of features; $R = 0.3$ for the maximal value of correlation coefficients; $P = 0.05$ for the maximal value of statistical significance p-value.

The first feature has been chosen randomly and the whole process has been initialized 5 times for each dataset. The higher number of runs would have been unnecessary and might have caused repeated subset of features as the initial datasets consisted of rather small number of attributes. Nevertheless, for larger datasets, the number of 10 can be considered for the best suited clusters.

TABLE II
FEATURE SELECTION RESULTS

Dataset	Run	Size	Names of features		
HEART	1	6	RR_man_SBP	birth_weight	family_interview
			BMI	ABPM-D	ABPM-S
	2	6	weight	fundus	HR
			physical_activity	RR_man_DBP	ABPM-S
	3	6	height	RR_man_DBP	birth_weight
			HR	physical_activity	BMI
	4	6	RR_ABPM_SBP	RR_ABPM_DBP	physical_activity
			fundus	BMI	birth_weight
	5	5	BMI	height	fundus
			HR	physical_activity	
ECHO	1	3	BMI	24H_DBP	birth_weight
	2	2	day_SBP	birth_weight	
	3	2	24H_SBP_load	birth_weight	
	4	2	birth_weight	day_SBP	
	5	3	night_DBP	birth_weight	24H_SBP
IUGR	1	3	birth_weight	ponderal_index	Apgar_score
	2	4	head_circuit	5_percentile	Apgar_score
			ponderal_index		
	3	4	gestational_age	Apgar_score	5_percentile
			head_circuit		
	4	4	Apgar_score	ponderal_index	5_percentile
			head_circuit		
	5	3	ponderal_index	birth_weight	Apgar_score

The subsets of features presented in Table II were obtained as a result of the proposed feature selection process. The first column of the table represents names of datasets, the second - a number of runs, and the following columns contain the number and names of features selected in each run of the process ordered in accordance with our algorithm.

C. Cluster Analysis and Validation

In the next step of the experiments, clusters were created by using EM algorithm implemented by WEKA Open Source software [14]. Clusters were built taking into account attributes indicated by feature selection method. The results of clustering are presented in Table III, where the first column describes datasets, the second contains a number of the run, and the following ones present number of clusters, clustering schemes and the normalized values of SSW.

Each feature selection process produced different subset of attributes for cluster analysis and as a result different clustering schemes. The values of within cluster sum-of-squares were computed to choose the minimal total value as the best version of clustering. We excluded those runs, that produced only one cluster. Therefore, for ECHO dataset no validation was performed and results of the first run were directed to further analysis. The normalized values of SSW for HEART and IUGR datasets were presented in the last column of Table

TABLE III
CLUSTERING RESULTS

Dataset	Run	No of clusters	Clustering schema	SSW
HEART	1	3	6 10 14	0.86
	2	3	23 6 1	0.12
	3	3	17 5 8	1.00
	4	4	6 4 6 14	0.36
	5	3	5 5 20	0.55
ECHO	1	3	10 17 39	N/A
	2	1	66	N/A
	3	1	66	N/A
	4	1	66	N/A
	5	1	66	N/A
IUGR	1	1	47	N/A
	2	6	7 12 1 18 3 6	0.78
	3	2	22 25	0.25
	4	4	7 13 22 6	1.00
	5	1	47	N/A

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

Dataset	Whole dataset	Cluster	IncreaseA	IncreaseE
HEART	14	35	150%	107%
ECHO	13	16	23%	31%
IUGR	11	16	45%	36%

III. Runs with the lowest values of SSW were chosen for the next step of statistical analysis.

D. Statistical Analysis

Correlation values obtained for the clusters were compared with 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 $p - value <= 0.05$ ([19], [20]). The results of detected correlations are presented in Table IV, where the column (3) presents the numbers of discovered dependencies in clusters and the last two columns show the percentage increases of correlations in comparison to the numbers of correlations for the whole dataset (column (2)) respectively for automatic feature selection (column (4)) and for attributes indicated by experts (column (5)) (see [2]).

Moreover, in two cases automated feature selection approach gave better results in comparison to the situation, when feature selection was consistent with the process of medical diagnosis. For ECHO dataset, the obtained increase is lower, however the number of attributes is significantly smaller than for the other data sets, what makes expert feature selection much easier.

E. Choice of the First Attribute

As the first attribute of feature selection process has been chosen randomly during experiments, the performance of alternative approach to this step of the algorithm has been checked out. Taking into account that the presented method is based on reversed correlation, the attributes the least correlated with the others were chosen as the first attributes. The remained steps of the algorithm stayed the same. The experiments were carried out for all the considered datasets.

Obtained results showed the advantage of the approach, where the first feature is indicated randomly and clustering schema is chosen depending on cluster qualities. For the least correlated first attribute the following effects were obtained. In the case of HEART dataset the number of significant correlations equaled 28 (less in comparison with column 2 of Table IV). For ECHO dataset only one cluster was built by using attributes indicated by the algorithm. Finally, in the case of IUGR dataset, the algorithm indicated the same features as in the 4th run of the first approach (see Table II).

V. CONCLUSIONS

In the paper a new feature selection method for the process of improved statistical analysis is proposed. We considered integrated approach where statistical inference is supported by cluster analysis. The proposed automation of feature selection concerns clustering attributes and is based on reverse feature correlations. Experiments conducted on real datasets have shown a big potential of the proposed method. Clusters based on attributes selected by the considered technique are better fitted for discovering new dependencies. The meaningful increase of statistically significant correlations enables improvements of medical diagnosis.

Future research will consist in further evaluation of the method performance by experiments carried out on public datasets. The feature selection algorithm will be developed taking into account datasets of different number of instances and attributes as well as other clustering schemes obtained by using different algorithms and cluster validity indices.

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] A. Wosiak, D. Zakrzewska, *On Integrating Clustering and Statistical Analysis for Supporting Cardiovascular Disease Diagnosis*, Proceedings of the 2015 Federated Conference on Computer Science and Information Systems, IEEE 2015, Annals of Computer Science and Information Systems, vol. 5, eds.: M. Ganzha and L. Maciaszek and M. Paprzycki, pp. 303–310, DOI: 10.15439/2015F151

[3] N. Acir, O. Ozdamar and C. Guzelis, *Automatic classification of auditory brainstem responses using SVM-based feature selection algorithm for threshold detection*, Eng. Appl. of AI, Vol. 19, no. 2, 2006, pp. 209-218, DOI: 10.1016/j.engappai.2005.08.004

[4] Z. Xu, I. King and M. R.-T. Lyu, *Discriminative Semi-Supervised Feature Selection Via Manifold Regularization*, IEEE Transactions on Neural Networks, Vol. 21, No. 7, 2010, pp. 1033-1047, DOI: 10.1109/TNN.2010.2047114

[5] 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

[6] G. Chandrashekar, F. Sahin, *A survey on feature selection methods, Computers and Electrical Engineering*, Vol. 40, 2014, pp. 1628

[7] M. Hall, *Correlation Based Feature Selection for Machine Learning*, Doctoral Disertation, University of Waikato, Department of Computer Science, 1999

[8] Y. Chen, Y. Li, X.-Q. Cheng and L. Guo, *Survey and Taxonomy of Feature Selection Algorithms in Intrusion Detection System*, In: Proceedings of Inscrypt 2006, LNCS 4318, pp.153-167, 2006

[9] H. Nguyen, K. Franke and S. Petrovic, *Improving Effectiveness of Intrusion Detection by Correlation Feature Selection*, 2010 International Conference on Availability, Reliability and Security, 17-24, 2010

[10] Y. Deng, D.F. Hsu, Z. Wu and Ch.-H. Chu, *Feature selection and combination for stress identification using correlation and diversity*, 2012 International Symposium on Pervasive Systems, Algorithms and Networks, 37-43, 2012

[11] Q. Zhu, L. Lin, M.-L. Shuy and Sh.-Ch. Chen, *Feature Selection Using Correlation and Reliability Based Scoring Metric for Video Semantic Detection*, 2010 IEEE Fourth International Conference on Semantic Computing, 462-469, 2010

[12] R. Wald, T.M. Khoshgoftaar and A. Napolitano, *Using-Correlation Based Feature Selection for a Diverse Collection of Bioinformatics Datasets*, 2014 IEEE 14th International Conference on Bioinformatics and Bioengineering, 156-162, 2014

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

[14] I.H. Witten, E. Frank and M.A. Hall, *Data Mining. Practical machine learning tools and techniques*, Morgan Kaufmann, San Francisco, USA, 2011

[15] G. Gan, Cha. Ma and J. Wu, *Data clustering: theory, algorithms, and applications*, ASA-SIAM Series on Statistics and Applied Probability, Vol. 20, 2007.

[16] Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu, *Understanding of internal clustering validation measures*, Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE, 2010.

[17] Q. Zhao, X. Mantao, F. Pasi, *Sum-of-squares based cluster validity index and significance analysis*, Adaptive and Natural Computing Algorithms. Springer Berlin Heidelberg, 2009. 313-322.

[18] 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

[19] D.G. Altman and J.M. Bland: "Measurement in Medicine: the Analysis of Method Comparison Studies", The Statistician 32, 1983, pp. 307-317

[20] D.E. Hinkle, W. Wiersma and S.G. Jurs: Applied Statistics for the Behavioral Sciences. 5th ed. Boston: Houghton Mifflin, 2003