FEATURE SELECTION IN THE EFFECTIVENESS RESEARCH OF A TRAINING PROGRAM FOR PATIENTS WITH THE ATRIAL FIBRILLATION

V.V. Kutikova[1], A.V. Gaidel[1,2], A.G. Khramov[1,2]

[1] Samara National Research University, Samara, Russia
[2] Image Processing Systems Institute, Russian Academy of Sciences, Samara, Russia

Abstract. We investigated training therapeutic program effectiveness of the school "Stop a Stroke", which aimed at reducing a risk of a stroke for patients with the atrial fibrillation. On the basis of two feature selection methods using a criterion of the discriminant analysis to determine the best feature subset, we concluded that patients, who trained at the school, in contrast to patients, who have not received training, take anticoagulants for a long time and have a higher level of knowledge about the atrial fibrillation.

Keywords: data mining, feature selection, discriminant analysis criterion.

Citation: Kutikova VV, Gaidel AV, Khramov AG. Feature selection in the effectiveness research of a training program for patients with the atrial fibrillation. CEUR Workshop Proceedings, 2016; 1638: 902-908. DOI: 10.18287/1613-0073-2016-1638-902-908

1 Introduction

Reducing the dimensionality of a feature space is one of the central issues in data mining. For the most classification and recovery regression problems it is necessary to select the best subset from a given feature set. This is because the use of a large feature number is not only computational expensive, but also affects the recognition accuracy, as irrelevant and redundant features, which complicate the decision-making process, can be used.

Feature selection methods are commonly used for the biomedical data analysis. In the work [1] using the method based on the ANCOVA an 80-gene biomarker of the smokers lung cancer was identified from 22216 features describing the expression levels of different genes. The accuracy, sensitivity and specificity of this biomarker were 83 %, 80% and 84%, respectively. For the selection of a small number of features one can use the brute force [2]. In biomedical data mining the sequential search algorithms [3] and genetic algorithms [4] are also used. Feature selection methods based on a criterion of the discriminant analysis have showed their effectiveness in [5, 6] for the analysis of biomedical images.

The aim of this work is the research of the effectiveness of the training therapeutic program of the school "Stop a Stroke", which aimed at reducing a risk of a stroke for patients with the atrial fibrillation. The used dataset contains observations of 12 features and classes of the observations are patient groups, namely: a main group and a comparison group. Patients of the main group have been trained at the school, and patients of the comparison group visited to a doctor, but have not received training. The data were obtained during two visits of patients to the doctor, namely: before (the first visit) and after (the second visit) the training course.

In order to evaluate the effectiveness of the training course, first of all, a feature subset, which distinguish the original data into two groups in the best way, is selected on the basis of data from the second visit; then, the performances of the obtained feature subset for two visits are compared; finally, conclusions about the effectiveness of the training course are draw.

As the main research tools two feature selection methods are used. The first method is based on the quality estimation of separate features using the discriminant analysis criterion and the second method allows to estimate the quality of feature subsets based on the same criterion.

2 Methods

2.1 Feature ordering in correspondence with the discriminant analysis criterion

According to the described in [4] method, features are ordered in correspondence with the discriminant analysis criterion [7]

$$J = \frac{\operatorname{tr} S_m}{\operatorname{tr} S_w}, \tag{1}$$

where $\operatorname{tr} S_w$ is the trace of a within-class scatter matrix, $\operatorname{tr} S_m$ is the trace of a mixture scatter matrix.

The within-class scatter matrix shows the scatter of samples around their respective class expected vectors:

$$S_w = \sum_{i=1}^{2} p_i M\left\{\left(X^{(i)} - M_i\right)\left(X^{(i)} - M_i\right)^T\right\},$$

where p_i is a class prior probability, $X^{(i)}$ is a feature vector from the *i*-th class, M_i is an expected vector of the *i*-th class.

The mixture scatter matrix is the correlation matrix of all feature vectors regardless of their class:

$$S_m = M\left\{\left(X - M_0\right)\left(X - M_0\right)^T\right\},$$

where $M_0 = p_1 M_1 + p_2 M_2$ is the expected vector of the mixture distribution.

The higher criterion value (1) is, the better a feature distinguishes the samples from different classes.

2.2 Sequential search of the best feature subset

The previously described approach to feature selection allows to assess the performance of each feature separately, but does not take into account dependencies between the features. Some of them can be useless by themself, but effective when combined with other features.

A sequential algorithm provides a search of the best features on the space of feature subsets. The basic idea is that starting from some initial subset on each step we move to the next state, in which one element is included into the current feature subset or excluded from it.

Let *F* be a set of all features, which participated in the selection, *X* be a current best feature subset of *F* and *Y* be a subset of the rest features that is $Y = F \setminus X$. The sequential algorithm consists of the following steps:

1. The choice of an evaluation function to measure the performance of a feature subset (in this work it is the criterion (1)), stopping criterion (the dimension of the current set *X* is equal to the dimension of the set *F*) and some initial subset (let it be the empty set).
2. The choice of “the search direction”. We step forward when we search the best subset among new subsets formed by including one feature in a current best subset:

$$X = X \cup \{y\},$$

where $y = \arg\max_{c \in Y} J(X \cup \{c\})$. We step back when we search the best subset among subsets which formed by excluding one feature from a current best subset:

$$X = X \setminus \{x\},$$

where $x = \arg\max_{c \in X} J(X \setminus \{c\})$.

3. If after finding a next subset the stopping criterion is fulfilled, then the search process is stopped, else we go to the step 2.

In this paper, we offer the "two steps forward, one step back" approach, and the best subset among other subsets with the same dimension is chosen as feature subset with the highest value of the criterion (1).

2.3 Evaluation of the training program effectiveness

Let J_0 be a value of the criterion (1) for some feature subset obtained on the basis of data from the first visit to the doctor, and J_1 be a criterion value calculated according to data from the second visit. An evaluation of the training program effectiveness for some feature subset is calculated as follows:

$$E = \frac{J_1}{J_0}. \tag{2}$$

The school has a “positive” effect on the feature subsets for which $E > 1$, a “negative” effect on the subsets for which $E < 1$, and the school has no effect on subsets with $E = 1$.

A conclusion about the effectiveness of the training program is done on the basis of obtained effectiveness values and within-group mean values of features which have more impact from the school.

3 Experimental results

The study of training program effectiveness were carried on the basis of 12 features, namely: answers to a questionnaire, which was filled by patients before and after the training course, blood pressure (systolic, diastolic), hemostasis parameters (prothrombin time, prothrombin, fibrinogen, partial thromboplastin time). The dataset included 69 observations (36 observations from the main group, and 33 from the comparison group) for two visits.

3.1 Effectiveness of the training course for separate features

Table 1 shows evaluation results of training program effectiveness for separate features. Each feature has a value of the criterion (1), which obtained on the basis of data from the first J_0 and second J_1 visits, as well as efficiency value E calculated according to the formula (2). Features are arranged in descending order of criterion value. Table 2 shows the within-class mean values for features presented in Table 1.

According to Table 1, after the training course (the second visit) features 1, 2 and 4 properly distinguish the groups of patients compared to the first visit that is the effectiveness values are quite large for these features. In addition, as shown in Table 2, mean values within the main group for these features have increased from the first to the second visits and have changed slightly within the comparison group. Hence, the training therapeutic program is effective for features 1, 2 and 4.

One also notices that for feature 3 the value J_0 is greater than the value J_1 that is this feature better distinguished the patient groups before the training course than after it. This effect is explained by the fact that within-class mean values have improved slightly from the first visit to the second (the patients began to realize importance of drug intake) and a within-class variance has increased.

Table 1. Effectiveness of the training course for separate features

№	Feature	J_1	J_0	*E*
1	Anticoagulant therapy (0 – don't take , 1 – take less than a year , 2 – from 1 to 5 years, 3 – more than 5 years)	9.35	0.97	9.64
2	How do you assess your level of knowledge about the atrial fibrillation? (1 – low, 5 – high)	5.21	0.98	5.33
3	How important is it to regularly to take a drug for stroke prevention in accordance with a prescription? (1 – no important, 5 – very important)	3.07	3.52	0.87
4	How do you assess your knowledge about the risk of stroke as the main complication of the atrial fibrillation? (1 – low, 5 – high)	2.58	1.03	2.50
5	Prothrombin (per cent)	1.19	1.08	1.10
6	Systolic blood pressure (mm Hg)	1.13	0.99	1.14
7	Prothrombin time (seconds)	1.13	0.99	1.14
8	Partial thromboplastin time (seconds)	1.07	1.03	1.03
9	Fibrinogen (g/l)	1.02	1.00	1.02
10	Aspirin (1 – take, 0 – don't take)	1.02	1.02	1.00
11	Diastolic blood pressure (mm Hg)	0.99	0.99	1.00
12	How much has the atrial fibrillation changed your daily life? (1 – hasn't changed, 5 – has changed greatly)	0.99	0.99	1.00

Table 2. Within-group mean values of features

№	**Main group**		**Comparison group**	
	Visit 1	Visit 2	Visit 1	Visit 2
1	0.06	2.06	0.12	0.12
2	1.47	4.31	1.55	1.48
3	4.08	4.15	1.39	1.61
4	1.49	4.28	1.89	2.08
5	88.58	84.02	97.34	95.18
6	166.94	140.78	168.48	144.3
7	13.13	13.78	13.06	12.98
8	31.70	31.90	30.19	30.01
9	4.66	4.26	4.58	4.37
10	0.94	0.94	0.79	0.79
11	98.17	87.47	97.70	88.24
12	3.02	3.56	3.06	3.30

3.2 Effectiveness of the training course for feature subsets

Table 3 shows evaluation results of training program effectiveness for subsets which obtained in accordance with the sequential feature selection method presented in section 2.2. In Table 3 for the first 10 best subsets of features from Table 1 the values J_0, J_1 and E are given.

One can see that all 10 subsets better distinguish the patient groups after the training course, than before it. However, the first 3 subsets, which included features 1, 2 and 10, have large effectiveness value E compared to the remaining subsets. In addition, mean values of features 1 and 2 within the main group have increased from the first visit to the second and have not changed significantly within the comparison group. Taking into account these facts, we can conclude that the therapeutic training program is effective for features 1 and 2.

Table 3. Effectiveness of the training course for feature subsets

Features	J_1	J_0	***E***
1	9.35	0.97	9.64
1, 2	6.01	0.98	6.14
1, 2, 10	5.20	0.98	5.29
1, 2, 3, 10	4.08	2.16	1.89
1, 2, 3, 9, 10	3.59	1.95	1.84
1, 2, 3, 4, 9, 10	3.29	1.72	1.91
1, 2, 3, 4, 7, 9, 10	2.63	1.46	1.81
1, 2, 3, 4, 7, 9, 10, 12	2.18	1.32	1.66
1, 2, 3, 4, 7, 8, 9, 10, 12	1.42	1.10	1.29
1, 2, 3, 4, 7, 8, 9, 10, 11, 12	1.24	1.05	1.18

4 Conclusion

In this paper the research results of effectiveness of the training therapeutic program of the school "Stop a Stroke" are presented. Using two feature selection methods the feature subsets, which distinguish patients of the main and comparison groups in the best way, are founded on the basis of obtained after the training course patient data. Feature subsets, which have the large effectiveness values, are selected among the chosen best subsets. Taking into account the within-class mean values we concluded that, in general, this program turned out to be effective.

In particular, the research of the training program effectiveness for separate features has shown that training at school “Stop a stroke” is effective for features “Anticoagulant therapy” ($E = 9.62$), “How do you assess your level of knowledge about the atrial fibrillation?” ($E = 5.33$) и “How do you assess your knowledge about the risk of a stroke as the main complication of the atrial fibrillation?” ($E = 2.50$). Evaluating the effectiveness of the training course for feature subsets we obtained similar results. In that case the training program of the school turned out to be effective for the pair of features “Anticoagulant therapy” and “How do you assess your level of knowledge about the atrial fibrillation?” (*E = 6.14*). On the other hand, the patients, who trained at the school in contrast to patients, who have not received training, take anticoagulants for a long time, and have a higher level of knowledge about the atrial fibrillation, stroke risk, as the main complication of the atrial fibrillation.

Acknowledgements

The work was partially funded by Russian Science Foundation (RSF), grant No. 14-07-97040-р_поволжье_а, the Russian Federation Ministry of Education and Science and Fundamental Research Program NITD RAS «Bioinformatics, modern information technologies and mathematical methods in medicine».

References

1. Spira A, Beane JE, Shah V, Steiling K, Liu G, Schembri F, Gilman S, Dumas Y-M, Calner P, Sebastiani P, Sridhar S, Beamis J, Lamb C, Anderson T, Gerry, N, Keane J, Lenburg ME, Brody JS. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. Nature Medicine, 2007; 13(3): 361-366.
2. Ilyasova NY, Kupriyanov AV, Paringer RA. Formation of features for improving the quality of medical diagnosis based on discriminant analysis methods. Computer Optics, 2014; 38(4): 851-855. [In Russian]
3. Peng Y, Wu Z, Jiang J. A novel feature selection approach for biomedical data classification. Journal of Biomedical Informatics, 2010; 43: 15-23.
4. Tsai C-F, Eberle W, Chu CY. Genetic algorithms in feature and instance selection. Knowledge-Based Systems, 2013; 39: 240-247.
5. Gaidel AV, Pervushkin SS. Research of the textural features for the bony tissue diseases diagnostics using the roentgenograms. Computer Optics, 2013; 37(1): 113-119. [In Russian]
6. Kutikova VV, Gaidel AV. Study of informative feature selection approaches for the texture image recognition problem using the Laws’ masks. Computer Optics, 2015; 39(5): 744-750. DOI: 10.18287/0134-2452-2015-39-5-744-750.
7. Fukunaga K. Introduction to statistical pattern recognition. San Diego: Academic Press, 1990.