Mathematica
Balkanica

New Series Vol. 24, 2010, Fasc.3-4

Classification of Smoking Cessation Status Using Various Data Mining Methods

Aleksandar Kartelj

This study examines different approaches of binary classification applied to the problem of making distinction between former and current smokers. Prediction is based on data collected in national survey performed by the National center for health statistics of America in 2000. The process consists of two essential parts. The first one determines which attributes are relevant to smokers status, by using methods like basic genetic algorithm and different evaluation functions [1]. The second part is a classification itself, performed by using methods like logistic regression, neural networks and others [2]. Solving these types of problems has its real contributions in decision support systems used by some health institutions.

AMS Subj. Classification: 62P10, 62H30, 68T01

Key Words: data mining, classification, induction learning

1. **Introduction**

Today data mining is one of the most popular and most exciting disciplines of applied informatics. It enables us to discover complex and hidden patterns in data, which can potentially bring to totally new conclusions in different disciplines, where sometimes even those disciplines experts cannot do better. One of especially interesting areas today in which data mining is often applied is certainly medicine. Decision support systems in health are developing more than ever, and their backbones are often theoretically founded on proved mathematical and physical principles.

1.1. **Some applications.** Classification problems are recognized in group of data mining methods. Beside, there are also methods of association, clustering techniques, regression etc. Classification can have exact mathematical models, heuristic models or random based models.

Some of the most important applications in real life are tumor classification, spam filters, decision making about good candidates for bank credits etc. Practically every real problem where the output is yes or no, can be solved using this technique. Naturally, the problem context must also be included in the process.

The smoking cessation status problem is a classification problem. It is highly dimensional, which means that it is dependent on large number of factors. Those factors are important in understanding of goal problem. One more potential gain of solving this problem is noticing new, until now not recognized correlated factors (in medical terminology dimension can be translated to symptom).

1.2. **Data.** Data collected by the National center for health statistics are presented in structured way (tabular form). In this study the main interest is put on only one of these surveys, the one oriented to questions related to adult persons. There were about 30000 respondents in it, and all answers are presented in one file (2000 NHIS sample adult file), which is publicly available at the site of this institution. Every person in this survey answered a set of 1429 questions, and some of these questions were related to smoking behaviour. Smoking status was cached in attribute SMKSTAT1. There are five possible statuses: `current´, `former´, `never´, `smoker, but currently unknown´and `unknown´. In this study we are solving the problem of binary classification, so we are interested only in instances which have values `current´or `former´. The idea is at first to filter only the relevant set of questions [1], and then based on their answers only produce binary classificator which will perform prediction of smoking status: former or current smoker.

2. **Preprocessing**

2.1. **Manual attribute elimination.** From more than 30000 instances, in the first step total of 14416 was selected (7421 current and 6995 former smoker). The ratio is well balanced, which contributes to the algorithm efficiency, as it will be seen later. The major of 1429 attributes is of nominal type, every with two or more possible values. A large part of the proposed set is irrelevant, so the first technique is a manual reduction of the attribute set, and then the automatic subset algorithm using data mining tool Weka v3.6 was used.

The manual attribute reduction is performed logically. So it is obvious on the first sight, that some attributes are redundant. For example, the prediction will not depend on rase of respondent, well not enough to make some important contribution to decision.

One other important heuristic, during the manual procedure is the value frequency on some question. Sometimes, there exists obviously correlated dependency, but still we should eliminate some feature (question). The best example are the questions about pregnancy. It is clear that most pregnant women will stop smoking, but only one small percent of women who were answering this inquiry were pregnant, so the relevance of these questions is not important in forming of general predictor [2].

An attribute can be relevant to the prediction, but sometimes we remove it because of redundancy. For example FRUITNO, FRUITTP, FRUITY and FRUITW all represent information whether and how much somebody eats fruits, with having in mind that FRUITY and FRUITW are generated using FRUITNO (mass of fruit) and FRUITTP (frequency of taking fruit). From this fact obvious is the redundancy of keeping attributes FRUITY and FRUITW, because they are already implicitly expressed through others.

3. Selecting best attribute subset

In learning algorithms often there is a problem with large input dimension. In data mining there are generally two techniques for solving this kind of problem [4]:

(1) Selecting subset of instances
(2) Selecting subset of features (attributes)

Selecting subset of instances is sometimes called sampling, and it represents one of the basic statistical tecniques. The goal is to select a representative sample, which is the sample that will contain the same or almost the same distribution, mathematical expectance, and dispersion as the original set. The easy way is certainly to select a random sample, but sometimes it is imperative to have a good distribution (the random sample can fail), so one possible approach is doing cluster analysis prior selection, and after that a proportional random selection from each cluster.

Selecting subset of features is a technique where we try to decrease the problem dimension. As it is described in 1.2 our problem is highly dimensional. This algorithm in the most general case tries to find the best subset of features from possible 2^N subsets, accordingly to evaluation function. The brute force algorithm is obviously time consuming even for sets with relativly small input dimension (N). So, there are different heuristical and random based principles, which can gain some performance. According to [1] there are 4 basic steps in the typical feature subset selection algorithm:

(1) generation procedure
(2) search procedure
(3) stopping criterion (eg. evaluation function)
(4) validation procedure

3.1. **Evaluation function.** The evaluation function is practically the condition of optimality in our case, but generally it can be statistical, heuristical or some other metric. The most common categorization of evaluation functions is on:

- Filter methods
- Wrapper methods

Evaluation function	Generality	Time complexity	Accuracy
Distance metric	Yes	Small	-
Info gain	Yes	Small	-
Dependency degree	Yes	Small	-
Consistency metric	Yes	Medium	-
Classifier error	No	Large	Very large'

Table 1. Evaluation functions

Method	Number of features	Correctly classified	Kappa statistic
Genetic algorithm (CfsSubsetEval)	24	71.52%	0.4289
Best first (CfsSubsetEval)	25	73.05%	0.4569
Ranker method (GainRatioAttributeEval)	25 (fixed)	71.31%	0.4244
Greedy algorithm (CfsSubsetEval)	25 (fixed)	73.00%	0.4587

Table 2. Selection results

Filter methods evaluate the subset quality accordingly to some prior criterion convention: distance measure, info gain, degree of dependence, consistency etc. **Wrapper methods** are not predefined, so they form criterion in dependence with learning algorithm. In most cases the criterion is the classification error itself. In table1 a comparison of evaluation functions using some key features is shown.

3.2. **Results comparison.** In this study there are few techniques used for feature subset selection algorithm 3, and they are all performed in Weka 3.6 data mining tool.

3.2.1. *Selected features.* Results 2 show that complete search (Best first) in combination with distance measure based evaluation function gave best results. Training was performed on random selected sample consisted of 66,6% of all instances. Afterward, the testing was performed using the rest of 33,3% instances. Selected attributes are shown in Table 3

4. **Classification**

Mathematically, the classification problem is defined as:
Let $\alpha = \{(x_1, y_1), \ldots, (x_n, y_n) | x_i \in R^n, y_i \in \{-1, 1\}\}$ be the training set. The function f, sometimes called predictor or classifier is formed using some learning algorithm and training set. After performing the learning algorithm, f maps

Feature	Description	Feature	Description
AGEP	Age	RATCAT	Income
HYPEV	Blood pressure	HEARAID	Hearing problems
RESTLESS	Restless	WRKLYR2	Had job in last 12m
ALCAMT	Alcohol	BMI	Body mass index
AUSUALPL	Where to go when sick	AHCAFYR1	Drug availability
ADNLONGR	Dental hygiene	FOBHAD	Blood analysis
AHCSYR2	Went to stomatologist	AHCSYR8	Went to doctor
SHTFLUYR	Flu vaccine	SHTPNUYR	Pneumonia vaccine
STD	Had infective illness	MILKKND	Milk kind
FRUITY	Fruit	VITEM	Vitamins in past 12m
CALC	Calcium use	MDTOB1	Asked about tobacco
SMHARM	Asked about tobacco risks	INCR150 INCR150	Opinion about tobacco increase
SKNX	Done complete head to toe check		

Table 3. Selected features

arbitrary test instance t from R^n in ‟an appropriate″class c from $\{-1, 1\}$. Under ‟appropriate ″we mean the one that minimizes empirical risk

$$I_{emp}[f;n] = \frac{1}{n}\sum_{i=1}^{n} V(y_i - f(x_i)),$$

where V represents some task-specific loss function and y_i is the correct value for given x_i [3].

Classification methods. There are many techniques, for implementing above given preposition [6]. This study analyzed four different techniques:

(1) Logistic regression
(2) Multilayer perceptron
(3) SVM
(4) Decision tree C4.5

4.1. **Logistic regression.** This method tries to fit training samples under curve of sigmoid (logistic) function. Similarly, there is also linear regression, which tries to fit data under the linear function. Logistic regression represents referent politics in classification methods.

Method	Correctly classified	Kappa statistics	Running time (seconds)
Logistic regression	73.05%	0.4596	8
Multilayer perceptron	69.09%	0.3843	3021
SVM	55.07%	0.0822	603
J48 (C4.5)	70.90%	0.4166	6

Table 4. Classification results

4.2. **Multilayer perceptron.** Neural network fits highly dimensional data very well. It has good degree of generalization, which makes it possible to perform well on unseen (test) data.

4.3. **SVM - Support vector machine.** One of today most prominent technique. It performs extremely good in work with high dimensional data. This method tries to put margin between positive and negative instances [5].

4.4. **Decision tree C4.5.** In every step of the algorithm one attribute is chosen, and that attribute then represents new node in tree. By that node the tree branches on each possible value for that attribute. Then the algorithm is recursively called for each of newly created nodes. In Weka 3.6 this algorithm is called j48.

4.5. **Results comparison.** All algorithms were performed using Weka 3.6. Table 4 shows the results. Logistic regression made best result. Multilayer perceptron is still the best potential candidate, and the reason why it did not perform best here is probably because networks incorporated in Weka 3.6 have very general character, so they are not meant for working with this specific problem. Decision tree C4.5 is generally applicable on shallow data, and the proposed result is probably very close to the best possible one that can be gained using this technique. Even SVM matters today as the best and most popular technique, in this situation it failed. The potential reason is the large diversity of values on some attributes. Since SVM reduces the multiclass problem (more than 2 values per attribute) to binary, this probably comes out as a performance issue, and reduces the algorithm accuracy in some way.

5. **Conclusion and further work**

The selection of best classification method is almost always dependent on data itself. In future, different architectures of neural networks should be investigated, considering the fact that in this study only the simplest version of the multilayer perceptron was used.

References

[1] M. Dash, H. Liu, Feature Selection for Classification, *Jour. Intelligent Data Analysis,* **1**, 1997, 131–156.

[2] Mollie R. Poynton, Anna M. McDaniel, Classification of smoking cessation status with a backpropagation neural network, *Journal of Biomedical Informatics*, **39**(6), 2006, 680–686.

[3] Theodoros Evgeniou, Massimiliano Pontil and Tomaso Poggio, Statistical Learning Theory: A Primer, *International Journal of Computer Vision*, **38**, 1, 2000, 9–13.

[4] Pang-Ning Tan, Michael Steinbach and Vipin Kumar, Introduction to Data Mining, *Addison-Wesley*, 2005.

[5] N. Cristianini, J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, Cambridge 2000.

[6] S.B. Kotsiantis, *Supervised Machine Learning: A Review of Classification Techniques*, Informatica, **31**, 2007, 249–268.

Department of Informatics
Faculty of Mathematics
University of Belgrade
Studentski Trg 16
11000 Belgrade, Serbia
E-mail: kartelj@matf.bg.ac.rs