Mach Learn (2015) 101:105–135
DOI 10.1007/s10994-015-5508-x

Probabilistic combination of classification rules and its application to medical diagnosis

Jakub M. Tomczak[1] · Maciej Zięba[1]

Received: 20 December 2013 / Accepted: 25 May 2015 / Published online: 20 June 2015

Abstract Application of machine learning to medical diagnosis entails facing two major issues, namely, a necessity of learning comprehensible models and a need of coping with imbalanced data phenomenon. The first one corresponds to a problem of implementing interpretable models, e.g., classification rules or decision trees. The second issue represents a situation in which the number of examples from one class (e.g., healthy patients) is significantly higher than the number of examples from the other class (e.g., ill patients). Learning algorithms which are prone to the imbalance data return biased models towards the majority class. In this paper, we propose a probabilistic combination of *soft rules*, which can be seen as a probabilistic version of the classification rules, by introducing new latent random variable called *conjunctive feature*. The conjunctive features represent conjunctions of values of attribute variables (features) and we assume that for given conjunctive feature the object and its label (class) become independent random variables. In order to deal with the between class imbalance problem, we present a new estimator which incorporates the knowledge about data imbalanceness into hyperparameters of initial probability of objects with fixed class labels. Additionally, we propose a method for aggregating sufficient statistics needed to estimate probabilities in a graph-based structure to speed up computations. At the end, we carry out two experiments: (1) using benchmark datasets, (2) using medical datasets. The results are discussed and the conclusions are drawn.

Electronic supplementary material The online version of this article (doi:10.1007/s10994-015-5508-x) contains supplementary material, which is available to authorized users.

Editors: Vadim Strijov, Richard Weber, Gerhard-Wilhelm Weber, and Süreyya Ozogur Akyüz.

✉ Jakub M. Tomczak
jakub.tomczak@pwr.edu.pl

Maciej Zięba
maciej.zieba@pwr.edu.pl

[1] Department of Computer Science, Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland

Keywords Probabilistic combination · Classification rules · Imbalanced data · Medical diagnosis

1 Introduction

Machine learning algorithms have been successfully applied to many complex problems in recent years, including intelligent analysis in medical diagnosis (Lavrač 1999; Kononenko 2001). There are different medical fields which have especially benefited from the machine learning methods, e.g., oncology diagnosis (Michalski et al. 1986), the diagnosis of breast cancer recurrence (Štrumbelj et al. 2010), lung cancer diagnosis (Zięba et al. 2014), cDNA microarray data analysis (Pearson et al. 2003), toxicology analysis (Blinova et al. 2003), supporting diabetes treatment (Tomczak and Gonczarek 2013). Among others, there are two crucial issues in extracting diagnostic models from medical data. First, there is a need of learning comprehensible models which provide interpretable knowledge to human doctors (Lavrač 1999). Second, medical data are recognized to be imbalanced (Mac Namee et al. 2002) which means that the number of examples from one class (e.g., healthy patients) is significantly higher than the number of examples from the other class (e.g., ill patients). In the imbalanced data phenomenon we can distinguish two subproblems (Japkowicz 2001; He and Garcia 2009): (i) the *between-class* imbalanced problem—data set exhibits an unequal distribution between its classes, (ii) the *within-class* imbalanced data problem—an unequal distribution of examples among subconcepts within a class. Here we discuss the imbalanced data phenomenon in medical domain only, however, this problem is widely encountered in other applications like credit scoring (Brown and Mues 2012) or fraud detection in telecommunication (Fawcett and Provost 1997).

In order to obtain comprehensible models classification rules or classification trees are typically used. There are different approaches to learning classification rules from data. One of the most traditional rules induction approaches is based on a separate-and-conquer strategy (see Fürnkranz 1999 for more details), e.g., AQ (Michalski et al. 1986), RIPPER (Cohen 1995) and OneR (Holte 1993). Another traditional approach to rules extraction uses searching in version (hypotheses) space which was applied in *Candidate Elimination Algorithm* (CEA) (Mitchell 1997) or in JSM method (Blinova et al. 2003).

A different approach to learning rules is based on taking advantage of association rules. The core of this approach exploits the idea of mining a special subset of associations rules whose right-hand-side are restricted to the class label (Liu et al. 1998). This line of research has been extended by applying fuzzy set theory (Chen and Chen 2008) or generalizing to multiple class problem and imbalanced data (Cerf et al. 2013).

The problem of the rules induction can be gently cast in the framework of rough sets (Pawlak et al. 1995). The general idea is to utilize rough set theory in order to obtain certain and approximate decision rules on the basis of approximations of decision classes, e.g., LEM1 and LEM2 (Stefanowski 1998) and their extension VC-DomLEM (Błaszczyński et al. 2011), or a dedicated method for imbalanced data (Stefanowski and Wilk 2006).

The accuracy of learning the classification rules can be increased by analyzing their statistical properties. The research in this direction has lead to many interesting methods, e.g., finding rules minimizing the difference between the rules margin and variance (Rückert and Kramer 2008), obtaining data dependent generalization bounds (Vorontsov and Ivahnenko 2011) or by applying Minimum Description Length paradigm (Vreeken et al. 2011).

In the machine learning community, it has been shown that the classification rules (trees) provide rather mediocre predictive performance in comparison to the strong classifiers such as Support Vector Machines (SVMs), Neural Networks or deep learning models (Kotsiantis 2007). One possible fashion of improving their accuracy and maintaining their comprehensibility at the same time is application of *ensemble learning* techniques. First approaches tried to combine classification rules with various decision making procedures, e.g., majority voting (Kononenko 1992) and bagging (Breiman 1996). A different technique aimed at probabilistic combination of decision trees by using tree averaging (Buntine 1992). More recently, Bayesian Model Averaging (BMA) of the classification rules was applied (Domingos 1997, 2000). The main idea of this approach was to combine several sets of classification rules which were induced using well-known rules induction algorithms (e.g., extracting rules from a decision tree C4.5). Nevertheless, it was noticed that the BMA of the classification rules suffered from overfitting (Domingos 2000). This result was further explained that the BMA cannot be applied as a model combination (Minka 2000) and thus a different approach should be used. We will propose a new fashion of a probabilistic combination for the classification rules.

Most of standard learning methods assume balanced datasets and/or equal misclassification costs. However, in the case of imbalanced datasets they fail in learning regularities within data which results in biased predictions across classes. Hitherto, a number of attempts to deal with the imbalanced data problem has been proposed (He and Garcia 2009; Japkowicz and Stephen 2002), i.e., sampling methods for imbalanced data (Kubat and Matwin 1997), cost-sensitive solutions (Elkan 2001), such as, cost-sensitive Naïve Bayes (Gama 2000), cost-sensitive SVMs (Masnadi-Shirazi and Vasconcelos 2010). Recently, a number of ensemble learning methods designed for imabalanced datasets has been proposed (Wang and Japkowicz 2010; Galar et al. 2012; Zięba et al. 2014).

In this paper, we present the probabilistic combination of classification rules for dealing with the both stated issues. The problem of interpretability of the model is solved by applying the classification rules. In order to increase its predictive performance we use probabilistic reasoning for combination of if-then rules. Next, we reduce the imbalanced data phenomenon by modifying the *Bayesian estimator* for categorical features, also known as m-estimate (Cestnik 1990; Džeroski et al. 1993; Fürnkranz and Flach 2003; Lavrač 1999; Zadrozny and Elkan 2001), with different misclassification costs.

Another issue we would like to consider in this paper is a method for aggregating the data for further reasoning. Typically, observations are stored with their number of occurrences. However, such approach implemented naively may be very cumbersome and requires a lot of computational resources. In order to increase the efficiency of data aggregation process, we propose a modification of a *graph-based data aggregation* (Tomczak and Gonczarek 2013).

The contribution of the paper is as follows:

- *Conjunctive features* as latent variables representing hidden relationships among features are presented.
- *Soft rules*, a probabilistic version of the classification rules, are proposed.
- The manner of the combination of the soft rules is outlined.
- The modification of the m-estimate for imbalanced data problem is introduced.
- The modification of graph-based data aggregation (later referred to as *graph-based memorization*) for reducing memory complexity of storing observations is outlined.
- The proposed approach is applied to the medical diagnosis.

The paper is organized as follows. In Sect. 2.1 a combination of the classification rules is outlined. In Sect. 2.2 a new approach to the probabilistic combination of the classification rules is described, including introduction of conjunctive features and soft rules. Next, Sect. 2.3 explains estimation of probabilities including the proposition of m-estimate for imbalanced data, object probability, and prior probability for conjunctive features. In Sect. 2.4 the graph-based memorization process is described and the fashion of using it in the estimation process is outlined. In Sect. 3 we study our model's performance in a simulation study (Sect. 3.1) and on benchmark datasets (Sect. 3.2), medical datasets (Sect. 3.3), and discuss the results (Sect. 3.4). Finally, conclusions are drawn in Sect. 4.

2 Methodology

2.1 Combination of crisp rules

Let $\mathbf{x} \in \mathcal{X}$ be an object described by D attributes, where each x_d, $d = 1, \dots, D$, can take only one of K_d possible values. We will write x_d^k if $x_d = k$. We denote the total number of all possible values of attributes by K, i.e., $\sum_d K_d = K$. Let $y \in \{-1, 1\}$ be a class label of $\mathbf{x}$. We refer to the examples with class label $y = -1$ as *negative*, and the ones with the label $y = 1$ are called *positive*. Additionally, we assume that the minority class (less frequent in the training data) is labeled with $y = 1$.

A *classification rule* r can be defined in two equivalent manners (Mitchell 1997). First fashion represents the classification rule as a binary-valued function $r : \mathcal{X} \to \{-1, 1\}$. On the other hand, the classification rule can be defined as an if-then rule in which the antecedent, a_r, is a conjunction of features' values, and the consequent, c_r, is a specific value of the class label. For example, for $\mathbf{x} \in \{1, 2, 3\} \times \{1, 2\}$, an exemplary if-then classification rule is IF $x_1^1 \wedge x_2^1$ THEN $y = 1$, where $a_r =$ "$x_1^1 \wedge x_2^1$" and $c_r = 1$. The key advantage of applying the classification rules to a prediction problem is that they are easily interpretable, i.e., they form a comprehensible model, and the final decision can be straightforwardly explained. Further, we assume a finite set of if-then rules or a space of all possible if-then rules for given feature space $\mathcal{X}$. In both cases we denote this set as $\mathcal{R}$.

In theory, the set of classification rules should cover the whole feature space and the rules should not contradict themselves, i.e., for the same features' values two (or more) different rules must return the same class label. However, in practice this condition may be false because of, e.g., several sets of rules are combined together (Kononenko 1992). Therefore, it is convenient to treat the set of classification rules as an *ensemble classifier* (sometimes called a *model combination*) (Dembczyński et al. 2008). Before introducing the combination of rules let us re-define the classification rule which we refer to as a *crisp rule*.[1] The crisp rule is a function $f : \mathcal{X} \times \{-1, 1\} \times \mathcal{R} \to \{0, 1\}$ in the following form:

$$f(\mathbf{x}, y, r) = \begin{cases} 1, & \text{if } \mathbf{x} \text{ is covered by } a_r \text{ and } y = c_r \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $r \in \mathcal{R}$ is a rule in the set of all possible if-then rules. For further simplicity we will denote $f(\mathbf{x}, y, r) \overset{\Delta}{=} f_r(\mathbf{x}, y)$.

In fact, the crisp rule is a standard boolean-valued function but it takes also the class label as an argument. This definition could be useful in multi-class problem, i.e., when there are more than two possible values of the class label. Moreover, the crisp rule represents a transformation of the if-then rule to a form which will be useful in the combined classifier.

An ensemble classifier is a weighted combination of base models. If we consider the crisp rules as the base models, then we can combine all crisp rules in $\mathcal{R}$ which yields the following ensemble classifier:

[1] We call the classification rules *crisp* in analogy to fuzzy sets and crisp sets.

$$g(\mathbf{x}, y) = \sum_{r \in \mathcal{R}} w_r f_r(\mathbf{x}, y), \tag{2}$$

where $w_r \in \mathbb{R}_+$ is a tunable parameter which denotes a weight of the rth rule. The final decision is the class label with the highest value of the combination $g(\mathbf{x}, y)$:

$$y^* = \arg\max_y g(\mathbf{x}, y). \tag{3}$$

Application of combination of crisp rules remains interpretability of the classification rules where the weights can be seen as *confidence levels* of rules. However, there are three problems associated with learning combination of the crisp rules. First, the crucial issue is how to determine the set of rules $\mathcal{R}$. The simplest approach is to apply a rule induction algorithm or several such algorithms in order to obtain $\mathcal{R}$. However, it has been shown that such a technique may give unsatisfactory results (Domingos 1997). On the other hand, summing over all possible crisp rules for all possible class labels is practically intractable. Second issue concerns learning the tunable parameters. There are different ensemble learning methods, e.g., bagging (Breiman 1996), boosting (Freund and Schapire 1997). Nonetheless, since the base learners corresponds to fixed, i.e., non-learnable, crisp rules it is more appropriate to apply other techniques for learning such as stacking (Wolpert 1992) or combinations based on statistical properties of the rules (Kononenko 1992). However, these learning procedures do not result in satisfactory predictive performance because in some cases they can decrease classification accuracy (Kononenko 1992) or even lead to overfitting (Domingos 2000). Third problem is that crisp rules assign only one class label to all objects they cover. This may be problematic in decision support systems in which human expert may would like to know the class label with its certainty level. These three issues are addressed in our probabilistic approach to the classification rules combination, which we refer to as *soft rules combination*.

2.2 Combination of soft rules

Let us consider the object $\mathbf{x}$ and the class label y as random variables. Additionally, we introduce a new random variable, which we refer to as *conjunctive feature*, φ, that corresponds to a conjunction of at least one feature with fixed value. Moreover, we assume that each of the features can take only one value within the conjunctive feature. The set of all possible conjunctive features is denoted by $\mathcal{F}$. For example, for $\mathbf{x} \in \{1, 2, 3\} \times \{1, 2\}$, a conjunctive feature can be $\varphi = x_1^1 \wedge x_2^1$, but the conjunction $(x_1^1 \vee x_1^2) \wedge x_2^1$ is not the conjunctive feature according to our definition. The conjunctive feature φ defines a set, that is:

$$\mathcal{X}_\varphi = \{\mathbf{x} \in \mathcal{X} : \mathbf{x} \text{ is covered by } \varphi\}. \tag{4}$$

For example, for $\mathbf{x} \in \{1, 2, 3\} \times \{1, 2\}$ and the conjunctive feature $\varphi = x_1^1$ one gets $\mathcal{X}_\varphi = \{(x_1^1, x_1^2), (x_1^1, x_2^2)\}$.

We assume that an object is described by the features and the class label which are observable random variables, while conjunctive features are treated as latent variables. The goal of our model is to use conjunctive features as a common structure that relates attributes and the class label, i.e., the conjunctive features consolidate two separate but related concepts. Since a conjunctive feature is a latent variable shared by the attributes and the class value, it generates both object $\mathbf{x}$ and its label y. As a consequence, for given φ random variables $\mathbf{x}$ and y become stochastically independent [see the model represented as a probabilistic graphical model (Cooper and Herskovits 1992) in Fig. 1], i.e., $p(y, \mathbf{x}|\varphi) = p(y|\varphi)\, p(\mathbf{x}|\varphi)$. The assumption about the independence allows to easier estimate the probabilities $p(y|\varphi)$ and $p(\mathbf{x}|\varphi)$ instead of the joint probability $p(y, \mathbf{x}|\varphi)$. Moreover, it is an unorthodox approach

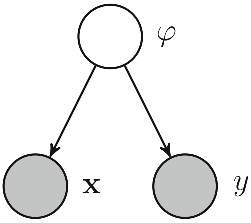


Fig. 1 Probabilistic graphical model for the considered model with latent conjunctive features. We assume that the conjunctive feature φ generates both object $\mathbf{x}$ and its label y. Latent variable is represented by *white node* and observable variables by *gray nodes*

comparing to the classification rules in which rules are deterministically associated with the class label. Here, we make a *soft* assumption about conjunctive feature's label. The proposed model is a specific kind of a *shared model* (Damianou et al. 2012) because a single hidden variable shares the information about both observable variables.

The idea of shared model is to introduce latent variables which capture common structure between two or more concepts. In the literature, there are different kinds of shared models. In one approach single shared latent structure is proposed to capture mutual information of observable variables (Shon et al. 2005). A different shared model introduces additional hidden variables which are typical for one of the concepts (Ek et al. 2008). Recently, fully Bayesian treatment of the shared model was proposed where the latent representation is marginalized out (Damianou et al. 2012). In our case, we aim at finding common hidden representation which allows reasoning about both the attributes and the class label. Therefore, the idea of our approach is very similar to the one presented in Shon et al. (2005) where there is one single common latent structure.

Hitherto, we have introduced new random variable, the conjunctive feature, that in fact corresponds to the antecedent of the if-then rule. However, we take advantage of probabilistic approach, and thus, the classification rules should be reformulated. We define a *soft rule* as a function which returns probability for the class label y and the features $\mathbf{x}$ conditioned on the conjunctive feature φ, $f : \mathcal{X} \times \{-1, 1\} \times \mathcal{F} \rightarrow [0, 1]$, that is:

$$f(\mathbf{x}, y, \varphi) = \begin{cases} p(y|\varphi)\ p(\mathbf{x}|\varphi), & \text{if } \mathbf{x} \text{ is covered by } \varphi \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $p(y|\varphi)$ is the probability of label y for all objects generated by φ, $p(\mathbf{x}|\varphi)$ is the probability of the object $\mathbf{x}$ given φ. In fact, in (5) we should write the joint distribution for $\mathbf{x}$ and y, $p(y, \mathbf{x})$, instead of $p(y|\varphi)\ p(\mathbf{x}|\varphi)$, but we have already applied the assumption about the conditional independence of $\mathbf{x}$ and y given φ. For further simplicity we will write $f(\mathbf{x}, y, \varphi) \overset{\Delta}{=} f_{\varphi}(\mathbf{x}, y)$.

In order to make decisions we need to obtain a predictive distribution $p(y|\mathbf{x})$ which enforces application of the summation rule, i.e., summation over all conjunctive features φ for the distribution $p(y, \varphi|\mathbf{x})$. However, we observe that there is no need to sum over all possible conjunctive features because only the ones covering $\mathbf{x}$ are important, and for all others we get $p(\mathbf{x}|\varphi) = 0$. Therefore, the predictive distribution takes the following form:

$$\begin{aligned} p(y|\mathbf{x}) &= \sum_{\mathcal{F}} p(y, \varphi|\mathbf{x}) \\ &= \sum_{\varphi:\mathbf{x}\in\mathcal{X}_{\varphi}} \frac{1}{p(\mathbf{x})}\ p(y, \mathbf{x}|\varphi)\ p(\varphi) \\ &\propto \sum_{\varphi:\mathbf{x}\in\mathcal{X}_{\varphi}} p(y|\varphi)\ p(\mathbf{x}|\varphi)\ p(\varphi), \end{aligned} \tag{6}$$

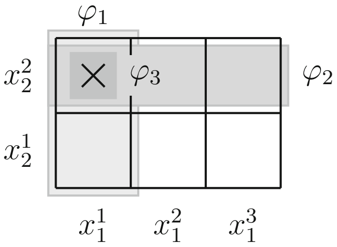


Fig. 2 An exemplary application of soft rules. For $\mathbf{x} \in \{1, 2, 3\} \times \{1, 2\}$, new object (x_1^1, x_2^2) (*black cross*) is covered by three conjunctive features: $\varphi_1 = x_1^1$ (*light gray rectangle*), $\varphi_2 = x_2^2$ (*gray rectangle*), and $\varphi_3 = x_1^1 \wedge x_2^2$ (*dark gray rectangle*). Other conjunctive features, e.g., $\varphi_4 = x_2^1$ or $\varphi_5 = x_1^2$, do not cover the object and thus are irrelevant in the prediction

where $p(\varphi)$ is *a priori* probability of a conjunctive feature. Notice that $p(\mathbf{x})$ is the same for all conjunctive features and hence can be omitted in further prediction. The final decision is the most probable class label. An application of exemplary conjunctive features is presented in Fig. 2.

Taking a closer look at the predictive distribution one can notice that it formulates the *combination of soft rules* (CSR) given by (5) with weights w_φ equal $p(\varphi)$, and the set of possible rules determined by the conjunctive features which cover $\mathbf{x}$, $\mathcal{R} = \{\varphi : \mathbf{x} \in \mathcal{X}_\varphi\}$. Below, we indicate which parts of the Eq. (6) correspond to elements in the ensemble classifier:

$$g(\mathbf{x}, y) = \sum_{\varphi:\mathbf{x}\in\mathcal{X}_\varphi} \underbrace{p(\varphi)}_{w_\varphi} \underbrace{p(y|\varphi)\ p(\mathbf{x}|\varphi)}_{f_\varphi(\mathbf{x},y)}. \tag{7}$$

The final decision is the class label with the highest value of the combination, i.e., $y^* = \arg\max_y g(y|\mathbf{x})$, which takes exactly the same form as for the crisp rules (3) but with the interpretation of unnormalized probabilities.

Note that the soft rules remain interpretable because it is a soft version of the crisp rule which in turn is the if-then rule. Hence, the soft rule can be represented as the if-then rule but with the *soft* consequent, i.e., the consequent consists the information about the class label and its probability p. For example, for $\mathbf{x} \in \{1, 2, 3\} \times \{1, 2\}$, the soft rule for $\varphi = x_1^1$ and the considered object $\mathbf{x}$ is as follows:

$$\begin{aligned} \text{IF } \mathbf{x}_1^1 \text{ THEN } & y = 1 \text{ with } p = p(y = 1|x_1^1)\ p(\mathbf{x}|x_1^1) \\ & \text{OR} \\ & y = -1 \text{ with } p = p(y = -1|x_1^1)\ p(\mathbf{x}|x_1^1) \end{aligned}$$

The application of the soft rules and their combination has the following advantages:

1. The set of soft rules used in the summation is determined automatically for given $\mathbf{x}$.
2. The weights of the soft rules in the combination are related to the prior for the conjunctive features. Hence, there is no need to propose an additional procedure for their determination.
3. The application of the probabilistic reasoning allows to utilize all information provided by the conjunctive features covering new object in contrary to the combination of the crisp rules which uses only a subset of all possible rules. Similar argument was also used in previous studies about the classification rules (Viswanathan and Webb 1998).

3. The soft rule can be represented as the crisp rule but with different consequent which returns probabilities of class labels instead of crisp assignment. In other words, the soft rule remains interpretability of the crisp rule and additionally assigns probabilities to the class labels.

The disadvantage associated with application of the conjunctive features is that their total number grows exponentially with the number of features D. Assuming for a while that the number of values of all features are equal, we can give an exact relationship between the number of the conjunctive features and the number of features:

Lemma 1 *Assuming $K_d = \kappa$ for $d = 1, \dots, D$, the number of all conjunctive features is equal* $(\kappa + 1)^D - 1$.

The justification of the relationship is trivial. Let us consider an object $\mathbf{x}$ which has D distinct features with number of values equal κ. Then, we need to consider all possible combinations of these distinct features except the empty conjunctive feature (i.e. $\varphi = \emptyset$) which results in:

$$\sum_{d=1}^{D} \binom{D}{d} \kappa^d - 1 = (\kappa + 1)^D - 1.$$

As we can see, the application of soft rules combination may be problematic when one deals with high-dimensional problems. Let us consider specifically the case of classifying new object. The object has D features and each attribute takes only one value (according to the Lemma 1 κ is equal 1). Therefore, in order to classify new object one needs to calculate the sum of $2^D - 1$ soft rules. For given N objects the general time complexity can be estimated by $O(N\, 2^D)$. Hence, application of (6) can be performed in an exact form for approximately up to 20 features. For higher-dimensional problems the time needed to perform the classification can be too long or the computational demands cannot meet computer requirements. In order to overcome these limitations a feature selection method can be applied or an approximate inference should be utilized. A different approach aims at faster computations via better coding schemes. In Sect. 2.4 we will show how to store sufficient statistics for calculating probabilities $p(y|\varphi)$ efficiently.

2.3 Probabilities calculation

In the following, we present the plug-in estimators of probabilities used in the classification rule (6) for given training data $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$. First, we propose the modification of the *m*-estimate for imbalanced data which is used in estimating $p(y|\varphi)$. Second, we define the probability of the object given the conjunctive feature $p(\mathbf{x}|\varphi)$. Next, we give the function for evaluating complexity of a conjunctive feature which is later applied to formulating *a priori* probability of conjunctive features $p(\varphi)$.

2.3.1 Modified m*-estimate for class label*

It has been shown that estimating probabilities with relative frequencies is troublesome and results in unreliable estimators (Cestnik 1990; Mitchell 1997). In Cestnik (1990) it has been proposed to take advantage of conjugate distributions, that is, Categorical and Dirichlet distributions, which gives the *Bayesian estimator*, called also *m*-estimate. In the considered case of the binary classification we deal with Bernoulli distribution and its conjugate prior

Beta distribution. There are different possible choices of priors (Jaynes 1968), e.g., Jeffreys prior (Jeffreys 1946), however, we aim at reducing imbalanceness of data through prior. For this purpose Beta distribution suits well because it allows us to put more probabilistic mass on probability associated with minority class, which is not the case for any non-informative prior (e.g., Jeffreys prior). The application of m-estimate to $p(y|\varphi)$ yields the following estimate:

$$p(y|\varphi) \approx \frac{N_{y,\varphi} + m\,\pi_y}{\sum_y N_{y,\varphi} + m}, \tag{8}$$

where $N_{y,\varphi}$ is the number of occurrences of objects with the class label y covered by φ, m is the non-negative tunable parameter of the estimator, and π_y is the initial probability of an object in the class y.

In the context of the between-class imbalance problem the estimation of probability $p(y|\varphi)$ should not be burden by one class, i.e., the proper estimator should eliminate the influence of the majority class on the minority class. In the m-estimate we can modify π_y and m, which have the following interpretations. The former determines the initial probability of objects covered by φ in the class y and the latter is the number of objects that *should* be observed initially, i.e., before observing any data.

In order to eliminate the imbalanced data phenomenon we propose to weight each observation using the following proportion:

$$\tilde{\pi}_y = \frac{N}{2N_y}, \tag{9}$$

where N_y is the number of observations in the class y. The weight of an observation in the minority class is >1 while in the majority class—<1. Such proportion was used in learning SVMs in order to reduce imbalanced data phenomenon (Cawley 2006; Daemen and De Moor 2009). This proposition can be justified twofold:

1. It has been noted in (Cawley 2006) that the weighting (9) is asymptotically equivalent to re-sampling the data so that there is an equal number of positive and negative examples.
2. If we sum over all the observations weighted with (9) we get

$$\begin{aligned}\sum_n \tilde{\pi}_{y_n} &= N_1 \frac{N}{2N_1} + N_2 \frac{N}{2N_2} \\ &= N. \end{aligned} \tag{10}$$

In other words, such weighting preserves the number of training examples.

The initial probability can be obtained by normalizing the proportions:

$$\begin{aligned}\pi_y &= \frac{\tilde{\pi}_y}{\sum_{y'} \tilde{\pi}_{y'}} \\ &= \frac{N_{-y}}{N}, \end{aligned} \tag{11}$$

where N_{-y} is the number of occurrences of objects with the opposite class to y, i.e., $\pi_1 = \frac{N_{-1}}{N}$ and $\pi_{-1} = \frac{N_1}{N}$. Hence, the application of (11) can eliminate the imbalanced data phenomenon by increasing initial probability of sampling initial objects in the minority class.

Typically, the tunable parameter m is determined experimentally (Džeroski et al. 1993; Zadrozny and Elkan 2001). Later in the work the m-estimate with the prior (11) is referred to as *imbalanced* m-estimate (or shorter *im*-estimate).

2.3.2 Object probability

In the proposed model we assume that the conjunctive feature can generate both the class label and the object. In the simplest approach the probability of the object given the conjunctive feature can be 1 if the object is covered by the conjunctive feature and 0—otherwise. However, such fashion of assigning probabilities do not distinguish conjunctive features and their possible generative capabilities. Instead, we would prefer to assume that the object is sampled from a uniform distribution over the domain determined by φ:

$$p(\mathbf{x}|\varphi) = \begin{cases} \frac{1}{|\mathcal{X}_\varphi|}, & \text{if } \mathbf{x} \in \mathcal{X}_\varphi \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

where $|\mathcal{X}_\varphi|$ is the cardinality of the set determined by the conjunctive feature φ. Such approach is very similar to the *strong sampling* assumption (Tenenbaum and Griffiths 2001).

The object probability can be seen as a realization of a *semantic* principle of simplicity (*semantic* Occam's razor) because the larger the domain of the conjunctive feature is the more it is penalized. In other words, the larger domain determined by φ, the lower the probability of the object. The *semantic* comes from the interpretation of the probability, i.e., we consider the meaning of covering the object by the conjunctive feature. For example, for $\mathbf{x} \in \{1, 2, 3\} \times \{1, 2\}$, for the conjunctive feature $\varphi = \mathbf{x}_1^1$ one gets $\mathcal{X}_\varphi = \{(\mathbf{x}_1^1, \mathbf{x}_2^1), (\mathbf{x}_1^1, \mathbf{x}_2^2)\}$ and thus for $\mathbf{x} = (x_1^1, x_2^1)$ the object probability equals $p(\mathbf{x}|\varphi) = \frac{1}{2}$.

The philosophical considerations can be cast in a more formal justification by observing that the probability $p(\mathbf{x}|\varphi)$ defined as in (12) is *monotonic*. In the counting inference-based literature a rule descriptive measure $d : \mathcal{R} \to \mathbb{R}$ is said to be monotonic in rule r, if for any two rules r_1 and r_2, such that antecedent of r_1 has more formulae than antecedent of r_2, one obtains $d(r_1) \geq d(r_2)$ (Brzezinska et al. 2007; Ceglar and Roddick 2006).[2] In the considered case it is indeed true because the more formulae there are in the conjunctive feature, the higher value takes the probability (12).

Notice that the proposed fashion the object probability is calculated eliminates the within-class imbalanced data problem. Usually, the unequal distribution of examples within a class leads to biased estimates. Here, our assumption about the dependencies in the graphical model results in vanishing dependency between the class and the object for given conjunctive feature. Therefore, we calculate the object probability (12) using the cardinality of the set of objects $\mathcal{X}_\varphi$ determined by the conjunctive feature independent on the class label.

2.3.3 Conjunctive feature prior

The probability of the conjunctive features represents prior beliefs. One possible proposition of prior beliefs is the following: *The conjunctive features that contain less features are more probable*. In other words, the prior of the conjunctive features can be seen as a realization of a *syntactic* principle of simplicity (*syntactic* Occam's razor) which states that shorter (simpler) conjunctive features are *a priori* more probable than the longer ones. We say *syntactic* because we consider the meaning of the structure of the conjunctive feature. Hence, we propose the following function for measuring conjunctive feature's complexity:

$$h(\varphi) = \exp(-a D_\varphi), \tag{13}$$

where a is a free parameter, D_φ denotes the number of features in φ. Further in the paper we arbitrarily use $a = \frac{1}{D}$ which turned to work well in practice.

[2] The descriptive measure is said to be *anti-monotonic* if $d(r_1) \leq d(r_2)$ for the same assumptions.

Table 1 The summary of probabilities calculation used in the experiments

Probability	Estimate	Tunable parameters
$p(y\|\varphi)$	$\frac{N_{y,\varphi}+m\,\frac{N_{-y}}{N}}{\sum_y N_{y,\varphi}+m}$	m-determined experimentally
$p(\mathbf{x}\|\varphi)$	$\begin{cases}\frac{1}{\|\mathcal{X}_\varphi\|}, & \text{if } \mathbf{x}\in\mathcal{X}_\varphi\\ 0, & \text{otherwise}\end{cases}$	–
$p(\varphi)$	$\propto \exp(-aD_\varphi)$	$a=\frac{1}{D}$

We get the prior over conjunctive features by normalizing the complexity function which results in the Gibbs distribution:

$$p(\varphi)=\frac{h(\varphi)}{\sum_{\varphi'}h(\varphi')}. \tag{14}$$

On the contrary to the object probability, the conjunctive feature prior probability $p(\varphi)$ defined as in (14) is *anti-monotonic*. Extending the conjunctive feature by adding an appropriate formula (i.e. preserving it is still the conjunctive feature according to our definition) results in decreasing the probability in (13).

2.3.4 Probability calculation: summary

In Table 1 we give a summary of our considerations about how the probabilities used in the combination of the soft rules are calculated and indicate the final forms we use in the experiments. In the conjunctive feature prior we can omit calculating denominator since it is constant for all conjunctive features and does not influence final prediction. Hence, we use $w_\varphi = h(\varphi)$ instead of the probability (14) in the combination of the soft rules (7).

2.4 Graph-based memorization

As we have pointed out earlier, for equal number of features' values, there are $(\kappa+1)^D-1$ of all conjunctive features (see Lemma 1). In order to calculate probability estimators for $p(y|\varphi)$, we need to store the number of sufficient statistics (parameters) proportional to the number of all possible conjunctive features. In practice, such approach can be troublesome or even impossible to keep in computer's memory because of the exponential complexity. However, to speed-up calculations and limit the number of parameters, we can take advantage of the data aggregation method with the *graph-based representation* which we call *graph-based memorization*.

Let us define the graph $\mathcal{G}_y=(\mathcal{V}_y,\mathcal{E}_y)$ for the given class label y in the following manner. The set of vertices $\mathcal{V}_y$ consists of the nodes that represent considered features with all values, i.e., a node is a pair $v=(d,i)$ where d states for feature's index and i denotes value of the feature. Nodes corresponding to one feature form a layer. We add a *terminal* node $v_T=(D+1,1)$ to the set of vertices and it forms a separate layer. Moreover, the terminal node is added to every example as an additional feature which always takes value 1. The set of edges $\mathcal{E}_y$ consists of all connections between any two nodes (including the terminal node) from different layers, e.g., the edge in the class y connecting ith value in sth layer, $u=(s,i)$, and jth value in tth layer, $v=(t,j)$, is denoted by $e^y_{u,v}$. Additionally, all nodes

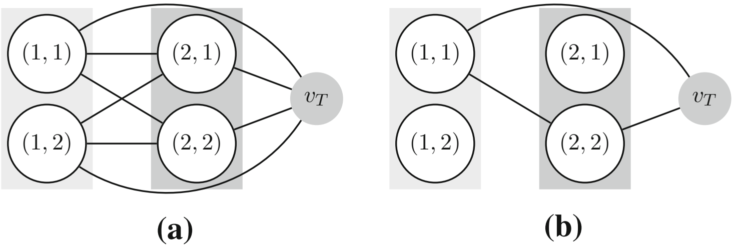


Fig. 3 **a** Exemplary graph for $\mathbf{x} \in \{1, 2\}^2$ in the class y. The *gray vertex* denotes the *terminal node*. The *light gray rectangle* represents first layer and the *darker one*—second layer. **b** Exemplary conjunctive feature $\varphi = x_1^1 \wedge x_2^2$ represented as a graph $\mathcal{G}_{y,\varphi}$ for $\mathbf{x} \in \{1, 2\}^2$

are connected with the terminal node. A weight of an edge represents the number of co-occurrences of two nodes in the training data, e.g., the weight of an edge in the class y connecting ith value in sth layer, $u = (s, i)$, and jth value in tth layer, $v = (t, j)$, is denoted by $w_{u,v}^{y}$. An exemplary graph is presented in Fig. 3a. All complete subgraphs which consist of at most one node from each layer and the terminal layer constitute conjunctive features in the class y, $\mathcal{G}_{y,\varphi} = (\mathcal{V}_y, \mathcal{E}_{y,\varphi})$. Note that the set of vertices is the same as in the graph $\mathcal{G}$ but the set of edges consists of only those edges which connect nodes included in the conjunctive feature φ. Additionally, we assume that the terminal node is included in each observation. An exemplary conjunctive feature as a graph is presented in Fig. 3b.

For given data $\mathcal{D}$ we can propose the following weights' updating procedure. If a pair of nodes $u = (s, i)$ and $v = (t, j)$ co-occurs in nth observation, $(\mathbf{x}_n, y_n)$, then

$$w_{u,v}^{y_n} := w_{u,v}^{y_n} + 1. \tag{15}$$

We need to perform updating for all co-occurrences of pairs of nodes in the observation $\mathbf{x}_n$ and proceed the procedure for all examples in $\mathcal{D}$. Remember that the terminal node is included in the observation, thus we always update $w_{u,v_T}^{y_n}$ for all features. Initially, all weights are set to zero. It is worth noting that the updating procedure is independent on the order of upcoming observations which means that the updating process is performed in an incremental manner and can be applied to a datastream. The procedure of the graph-based memorization is presented in Algorithm 1.

Graph-based memorization allows us to aggregate data in graphs $\mathcal{G}_y$ for classes $y \in \{-1, 1\}$, and thus we can approximate the count of objects with the class label y which are covered by φ as follows:

$$N_{y,\varphi} \leq \min_{(u,v):e_{u,v}^{y} \in \mathcal{E}_{y,\varphi}} w_{u,v}^{y}. \tag{16}$$

The *im*-estimator can be determined using graph-based memorization by inserting (16) into (8).

The application of the graph-based memorization indeed allows to decrease the number of stored sufficient statistics according to the following Lemma:

Lemma 2 *Assuming $K_d = \kappa$ for $d = 1, \ldots, D$, $\kappa > 2$ and $D > 3$ or $\kappa = 2$ and $D > 4$, the number of sufficient statistics stored by the graph $\mathcal{G}$ is equal $(D\kappa + 1)^2$.*

Proof Let us use the adjacency matrix to represent the considered graph $\mathcal{G}$. Because there are $K + 1$ nodes, i.e., there are K nodes corresponding to features with values plus the terminal node, we need less than $(K + 1)^2$ weights in one class, because we do not allow edges within

Algorithm 1: Graph-based memorization.

```
Input  : training data consisting of $N$ examples, $D$ is the number of features, $y \in \{-1, 1\}$.
Output: For each class $y$ the graph $\mathcal{G}_y$ which contains aggregated training data.
1  Initialize: $\forall_y \ \forall_{u,v \in \mathcal{V}_y} \ w^y_{u,v} \longleftarrow 0$ ;
2  for $n = 1 \to N$ do
3      for $d$ from 1 to $D$ do
4          Set $i$ to be the value of the $d^{\text{th}}$ attribute of the $n^{\text{th}}$ observation;
5          $u \longleftarrow (d, i)$;
6          for $\delta$ from $d + 1$ to $D + 1$ do
7              Set $j$ to be the value of the $\delta^{\text{th}}$ attribute of the $n^{\text{th}}$ observation;
8              $v \longleftarrow (\delta, j)$;
9              $w^{y_n}_{u,v} \longleftarrow w^{y_n}_{u,v} + 1$;
10         end
11     end
12 end
```

one layer, i.e., among nodes representing values of the same feature. Moreover, we notice that weights are symmetric, i.e., for any two nodes u and v, $w^y_{u,v} = w^y_{v,u}$, because we calculate co-occurrence of two nodes. Hence, we need less than $(K + 1)^2/2$ weights in one class. Next, once we assume equality of features' domains and binary classification problem, we get the number of sufficient statistics to be equal $(D\kappa + 1)^2$. □

2.4.1 Example

Let us consider a toy example for graph-based memorization. The object is described by two variables which can take two values, i.e., $\mathbf{x} \in \{1, 2\} \times \{1, 2\}$, and $y \in \{-1, 1\}$. We assume there are three examples: i) (x^2_1, x^1_2) and $y = -1$, ii) (x^2_1, x^2_2) and $y = -1$, iii) (x^1_1, x^1_2) and $y = 1$.

According to the graph-based memorization, we begin with the first example and update each edge which is encountered in the example (line 9 in Algorithm 1). We have to remember that the terminal node is added to every example and that is why one needs to iterate to $D + 1$ instead of D (line 6 in Algorithm 1). The resulting graphs after including the first two and the last datum are presented in Fig. 4a, b, respectively.

For graphs as in Fig. 4b we are able to calculate probability of y for given counts $N_{y,\varphi}$ determined by (16), the conjunctive feature φ, initial probabilities and fixed value of m using (8). Let us assume that $m = 1$ and $\pi_1 = 0.5$. Then, for instance, for $\varphi = x^1_1 \wedge x^1_2$ we have $N_{1,\varphi} = \min\{1, 1\} = 1$ and $N_{-1,\varphi} = \min\{0, 1\} = 0$, and consequently $p(y = 1|\varphi) = \frac{1+0.5}{1+1} = 0.75$ and $p(y = -1|\varphi) = \frac{0+0.5}{1+1} = 0.25$.

2.5 Multi-class case

In our considerations we have assumed only two possible class labels. However, the presented approach can be straightforwardly generalized to the multi-class case. First of all, let us notice that in the presented equations for calculating the combination of soft rules, i.e., Eqs. (5) and (7), as well as in the equation for final prediction (3), there is no restriction for the number of classes. Similarly, in calculating the probability of class label [see Eq. (6)] the number of classes is not determined. We have only used the assumption of two classes in calculating weighting of observations in the Eq. (9). However, it is easy to generalize it to any number of classes by replacing 2 in the denominator with the number of classes. Finally, the graph-based memorization is also independent on the number of class labels because we build a graph

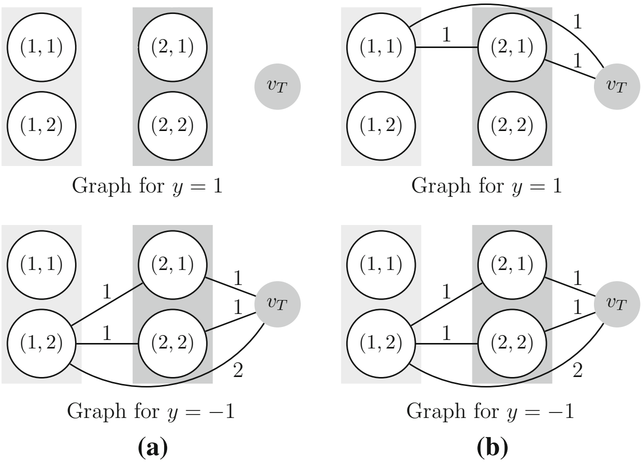


Fig. 4 An exemplary performance of the graph-based memorization for $\mathbf{x} \in \{1, 2\} \times \{1, 2\}$, $y \in \{-1, 1\}$ and three observations. Numbers above *edges* represent values of weights. More details can be found in text

for each class separately. Therefore, the whole process of storing sufficient statistics in the graph-based representation can be performed in the multi-class case.

3 Experiments

Data We carry out one simulation study on synthetic data and two experiments: Experiment 1 on synthetic benchmark datasets[3] (see Table 2), Experiment 2 on medical datasets[4] (see Table 3) including one real-life medical dataset provided by the Institute of Oncology, Ljubljana (Štrumbelj et al. 2010):

- *breast cancer*: the goal is the prediction of a recurrence of a breast cancer,
- *breast cancer Wisconsin*: the goal is the classification of a breast cancer as benign or malignant,
- *diabetes*: the goal is to classify the patient as tested positive for diabetes or not,
- *hepatitis*: the goal is to predict whether the patient suffering hepatitis will survive or die,
- *indian liver*: the goal is to classify the patient as healthy or with a liver issue,
- *liver disorders*: the goal is to classify the patient as healthy or with a liver disorder,
- *postoperative patient*: the goal is to classify the patient to be sent to hospital or home,
- *oncology*: the goal is to predict whether the patient will have a recurrence of a breast cancer or not.

The datasets are summarized in Tables 2 and 3 in which the number of features and the number of examples for each dataset are given. Additionally, we provide the imbalance ratio defined as the number of negative class examples divided by the number of positive class examples (Galar et al. 2012).

[3] Datasets are built-in the KEEL software (Alcalá et al. 2010).

[4] Datasets are taken from the UCI Machine Learning Repository (Frank and Asuncion 2010).

Table 2 The number of examples, the number of features and the imbalance ratio for benchmark datasets

Dataset	Number of examples	Number of features	Imbalance ratio
abalone19	4174	8	128.87
abalone9-18	731	8	16.68
car-2class	1728	5	2.25
ecoli-0-1-3-7vs2-6	281	7	39.15
ecoli-0vs1	220	7	1.86
ecoli1	336	7	3.36
ecoli2	336	7	5.46
ecoli3	336	7	8.77
ecoli4	336	7	13.84
glass-0123vs456	214	9	3.19
glass-016vs5	184	9	19.44
glass0	214	9	2.06
glass1	214	9	1.82
glass4	214	9	15.47
glass5	214	9	22.81
glass6	214	9	6.38
habermanImb	306	3	2.68
iris0	150	4	2.00
new-thyroid1	215	5	5.14
new-thyroid2	215	5	4.92
page-blocks13vs4	472	10	15.85
page-blocks0	5472	10	8.77
pimaImb	768	8	1.90
shuttle-c0-vs-c4	1829	9	13.87
shuttle-c2-vs-c4	129	9	20.5
vowel0	988	13	10.10
wisconsinImb	683	9	1.86
yeast-05679vs4	528	8	9.35
yeast-1289vs7	947	8	30.56
yeast-1vs7	459	8	13.87
yeast-2vs4	514	8	9.08
yeast-2vs8	482	8	23.10
yeast1	1484	8	2.46
yeast3	1484	8	8.11
yeast4	1484	8	28.41
yeast5	1484	8	32.78
yeast6	1484	8	39.15

Evaluation methodology. The proposed method, Combination of Soft Rules (**CSR**), was evaluated on the synthetic data and further was compared in two experiments with the following **non-rule-based methods**:

Table 3 The number of examples, the number of features and the imbalance ratio for medical datasets used in the experiments

Dataset	Number of examples	Number of features	Imbalance ratio
Breast cancer	286	9	2.62
Breast wisconsin	699	9	2
Diabetes	768	8	1.90
Hepatitis	155	19	2.78
Indian liver	583	10	1.87
Liver disorders	345	6	1.52
Post operative patient	90	8	5.43
Oncology	949	15	2.92

- AdaBoost (**AB**) (Freund and Schapire 1997),
- Bagging (**Bag**) (Breiman 1996),
- SMOTEBagging (**SBag**): Modified Bagging by learning base learners using SMOTE sampling technique (Chawla et al. 2002),
- SMOTEBoost (**SBoost**): Modified AdaBoost by learning base learners using SMOTE sampling technique (Chawla et al. 2002),
- Naïve Bayes classifier (**NB**),
- Cost-sensitive SVM (**CSVM**) with linear kernel (Cortes and Vapnik 1995),
- Neural Network (**NN**),

and **rule-based methods**:

- **C.45** tree learner[5] (Quinlan 1993),
- **RIPPER** classification rules learner (Cohen 1995),
- **OneR** classification rules learner (Holte 1993),
- **CFAR** classification rules learner based on fuzzy association rules (Chen and Chen 2008),
- **SGERD** fuzzy classification rules learner based on steady-state genetic algorithm (Mansoori et al. 2008),
- **ART** classification rules learner based on association rule tree (Berzal et al. 2004).

In order to evaluate the methods we applied the following assessment metrics:[6]

- *Gmean* (Geometric mean) which is defined as follows (Kubat and Matwin 1997; Kubat et al. 1997; He and Garcia 2009; Wang and Japkowicz 2010):

$$Gmean = \sqrt{\frac{TP}{TP+FN}\,\frac{TN}{TN+FP}}, \tag{17}$$

- *AUC* (Area Under Curve of the ROC curve) which is expressed in the following form (He and Garcia 2009):

$$AUC = \frac{1 + \frac{TP}{TP+FN} - \frac{FP}{TN+FP}}{2}, \tag{18}$$

[5] In the experiment we treat **C.45** as a rule-based method because all paths in the decision tree can be represented as a set of decision rules.

[6] *TP* is a number of positive examples classified as positive, *FN* is a number of positive examples classified as negative, *FP* is a number of negative examples classified as positive, *TN* is a number of negative examples classified as negative.

- *Precision* specifies how many examples from minority class were correctly classified comparing to the incorrectly labeled majority objects as minority ones (Fawcett 2006):

$$Precision = \frac{TP}{TP + FP}, \tag{19}$$

- *Recall* denotes the fraction of correctly classified objects from minority class to all examples labeled as minority class (Fawcett 2006):

$$Recall = \frac{TP}{TP + FN}. \tag{20}$$

It is advocated to use *Gmean* for imbalanced datasets because this metric punishes low classification accuracy of the minority class (Kubat and Matwin 1997; Kubat et al. 1997; He and Garcia 2009; Wang and Japkowicz 2010). Comparing to *AUC* the *Gmean* enforces high predictive accuracy on majority and minority classes. For further comparison of methods we calculated average ranks over benchmark and medical datasets according to *Gmean* and *AUC* which is a simple fashion of evaluating classification algorithms (Demšar 2006; Brazdil et al. 2003).

In the experiment *Precision* and *Recall* were also examined because these two measures give a thorough insight into the classifier's predictive performance exclusively for minority class. For better understanding of the obtained results, we present graphically the Pareto frontier (Brzezinska et al. 2007; Vamplew et al. 2011) with respect to *Precision* and *Recall* for the considered methods.

In order to verify our claims about the time complexity of the proposed approach, we measured the average execution time of five folds, expressed in miliseconds. We examined the dependency between time and the number of attributes and the number of examples separately.

The presented approach uses categorical variables only, therefore, we applied discretizer which utilizes entropy minimization heuristics (Fayyad and Irani 1993). Additionally, we performed feature selection using correlation-based feature extraction with exhaustive search (Hall 1999) to the selected datasets. In the Experiment 2, the feature selection on *hepatitis* dataset resulted in ten features, and in the Experiment 3 (*oncology* dataset)—five features were selected.

The experiments were conducted in KEEL software[7] (Alcalá et al. 2010). In order to obtain results we applied fivefold cross validation. For each dataset the value of m in **CSF** was determined using a validation set.

3.1 Simulation study: synthetic data

In order to get better insight of the proposed approach, in the simulation study we want to state and verify two issues: (i) the behavior of **CSR** for different distributions and given examples, (ii) the time complexity of **CSR** with varying number of attributes or examples. The questions are verified with the following simulation set-ups:

(i) It is assumed that the considered phenomenon is described by 8 binary attributes, $\mathbf{x} \in \{0, 1\}^8$. The attributes are generated independently with the Bernoulli distribution with $p_d = 0.5$, for $d = 1, \ldots, 8$. Further, we consider four possible rule-based descriptions of the phenomenon (in the brackets we give the imbalance ratio of all possible configurations of $\mathbf{x}$), namely, a conjunction $x_1^1 \Rightarrow y = 1$ (1:1), a more specific

[7] http://www.keel.es/.

Table 4 Results of the simulation study in terms of *Gmean* and *AUC* for **CSR**

Rules	$N = 50$				$N = 200$			
	$\varepsilon = 0$	$\varepsilon = 0.01$	$\varepsilon = 0.05$	$\varepsilon = 0.1$	$\varepsilon = 0$	$\varepsilon = 0.01$	$\varepsilon = 0.05$	$\varepsilon = 0.1$
Gmean								
$x_1^1 \Rightarrow y = 1$	1	1	1	0.997	1	1	1	1
$x_1^1 \wedge x_8^1 \Rightarrow y = 1$	0.900	0.863	0.708	0.520	1	0.996	0.666	0.594
$x_1^1 \vee x_8^1 \Rightarrow y = 1$	0.914	0.808	0.758	0.521	1	1	0.803	0.542
$(x_1^1 \wedge x_5^1) \vee x_8^1 \Rightarrow y = 1$	0.965	0.924	0.902	0.837	1	1	0.986	0.980
AUC								
$x_1^1 \Rightarrow y = 1$	1	1	1	0.997	1	1	1	1
$x_1^1 \wedge x_8^1 \Rightarrow y = 1$	0.893	0.852	0.796	0.662	1	0.996	0.650	0.530
$x_1^1 \vee x_8^1 \Rightarrow y = 1$	0.924	0.866	0.817	0.665	1	1	0.843	0.713
$(x_1^1 \wedge x_5^1) \vee x_8^1 \Rightarrow y = 1$	0.966	0.927	0.908	0.879	1	1	0.986	0.981

The first column **Rules** indicates what is the underlying description of the phenomenon

conjunction $x_1^1 \wedge x_8^1 \Rightarrow y = 1$ (3:1), a disjunction $x_1^1 \vee x_8^1 \Rightarrow y = 1$ (3:1), and a more specific disjunction $(x_1^1 \wedge x_5^1) \vee x_8^1 \Rightarrow y = 1$ (5:3).[8] Additionally, during generating data we inject noise to the class label, i.e., we switch the class label with probability $\varepsilon \in \{0, 0.01, 0.05, 0.1\}$. Eventually, we evaluate **CSR** with all possible configurations without noise, while learning is performed with 50 and 200 examples. In learning the *im*-estimate was used with $m = 1$.

(ii) In order to check the time complexity we chose two datasets from the benchmark datasets, namely, vowel0 (13 attributes, 990 examples) and page-blocks0 (10 attributes, 5000 examples). The first of the mentioned datasets was used to evaluate **CSR** in the case of varying number of attributes, while the second one was applied to assess the time complexity for varying number of examples.

All simulations were repeated 5 times.

The averaged results for the first issue are presented in Table 4 and the averaged results for the second issue are given in Fig. 5.

We can draw the following conclusions from the simulation study:

(i) In general, we conclude that **CSR** can be used to learn different descriptions of the phenomenon but for the sufficient number of observations, see Table 4. A quite unexpected conclusion of the simulation experiment is that **CSR** can handle both conjunctions and disjunctions. A main disadvantage of the **CSR** is its sensitivity of the value of the parameter m. In the simulation runs we used fixed $m = 1$ which is an inappropriate value for imbalanced data. Therefore, it is needed to apply a model selection of this parameter.

(ii) As we presumed (see Lemma 1), **CSR** allows to handle up to 10 attributes in a reasonable time but for more than 10 attributes the computation time starts increasing drastically, see Fig. 5a. On the other hand, the time complexity in the case of the growing number of examples is linear, see Fig. 5b. This is an important fact which shows that **CSR** scales well for large datasets but with a moderate number of attributes.

[8] We consider the closed-world assumption, i.e., any $\mathbf{x}$ which is not covered by the given rule is presumed to belong to the class $y = -1$.

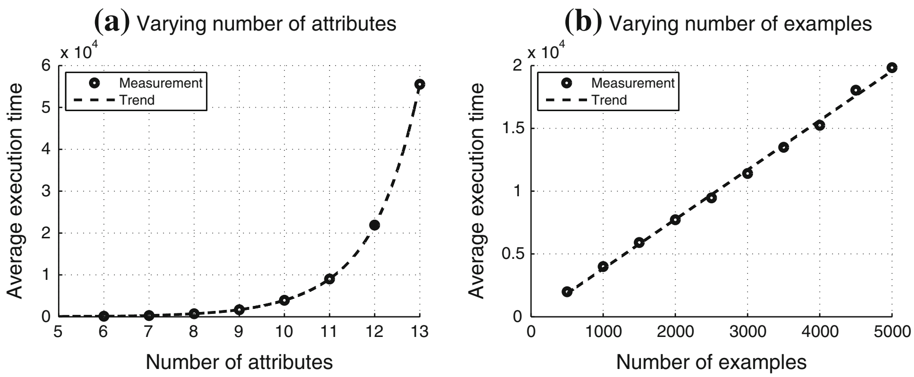


Fig. 5 Average execution time: **a** for varying number of attributes (vowel0 dataset), **b** for varying number of examples (page-blocks0 dataset)

3.2 Experiment 1: Benchmark data

In the first experiment we consider 37 benchmark datasets. These are generated from the benchmark datasets available in UCI ML Repository (Frank and Asuncion 2010) and are built-in the KEEL software. The classification problem for them is demanding because the datasets are highly imbalanced (see Table 2).

3.3 Experiment 2: Medical data

First, in the second experiment we evaluate our method on benchmark medical datasets (Frank and Asuncion 2010). It was noticed that medical problems associated with illnesses or post-operative prediction are typical examples of imbalanced data phenomenon (Mac Namee et al. 2002). In the considered medical datasets the issue of imbalanced data is observed, i.e., on average the imbalance ratio equals 2.6, varying from 1.52 to 5.43 (see Table 3).

Next, we took a closer look at the 949-case dataset provided by the Institute of Oncology Ljubljana (Štrumbelj et al. 2010). Each patient is described by 15 features:

- menopausal status;
- tumor stage;
- tumor grade;
- histological type of the tumor;
- level of progesterone receptors in tumor (in fmol per mg of protein);
- invasiveness of the tumor
- number of involved lymph nodes
- medical history
- lymphatic or vascular invasion;
- level of estrogen receptors in tumor (in fmol per mg of protein);
- diameter of the largest removed lymph node;
- ratio between involved and total lymph nodes removed;
- patient age group;
- application of a therapy (*cTherapy*);
- application of a therapy (*hTherapy*).

All the features were discretized by oncologists, based on how they use the features in everyday medical practice, see (Štrumbelj et al. 2010). The goal in the considered problem is the prognosis of a breast cancer recurrence within 10 years after surgery.

The *oncology* dataset, similarly to the benchmark medical datasets, suffers from the imbalance data phenomenon. The imbalance ratio for this dataset equals 2.92. It is a value which forces the application of techniques for preventing overfitting to the majority class.

For the *oncology* dataset we used an additional assessment metric, that is, the *classification accuracy* (*CA*) which is defined as follows:

$$CA = \frac{TP + TN}{TP + FN + FP + TN}. \tag{21}$$

The *CA* was applied because the authors of (Štrumbelj et al. 2010) have obtained results for randomly chosen 100 cases which were analyzed by two human doctors (*O1* and *O2* in Fig. 6). The oncologists were asked to predict the class value for these cases and then the *CA* value was calculated (Štrumbelj et al. 2010). The obtained quantities do not lead to a conclusion that the classifiers have significantly higher accuracy. However, they can give an insight in the usefulness of the application of machine learning methods in the medical domain.

3.4 Results of the experiments and discussion

The results of the Experiment 1 are presented in Table 5 (for ranks of *Gmean* and *AUC*), and Fig. 7a (for Pareto frontier with respect to *Precision* and *Recall*). The results of the Experiments 2 are presented in Table 5 (for ranks of *Gmean* and *AUC*), and Fig. 7b (for Pareto frontier with respect to *Precision* and *Recall*). Additionally, in Fig. 6 we provide a comparison of machine learning methods and human doctors in terms of *CA*. More detailed results of the experiments are given in Electronic Supplementary Material; Experiment 1: in Table A1 (for *Gmean*), Table A3 (for *AUC*)), Table A5 (for *Precision* and *Recall*), Experiment

Table 5 Detailed test results for **CSR** versus considered methods using ranks of *Gmean* and *AUC* for benchmark and medical datasets

Method	Gmean		AUC	
	Benchmark	Medical	Benchmark	Medical
CSR	***3.730***	**3.250**	***3.770***	***3.313***
AB	7.378	7.375	7.230	7.625
BAG	8.081	9.000	7.905	7.750
SBAG	**3.608**	4.500	**3.703**	5.500
SBoost	3.986	4.438	3.905	5.375
CSVM	6.878	6.438	6.689	5.813
C45	4.608	8.125	4.608	9.125
NN	8.284	6.500	8.973	8.375
NB	7.351	***3.938***	7.162	**3.188**
RIPPER	6.608	6.875	6.608	7.625
1R	11.068	10.500	11.027	9.563
CFAR	11.878	12.125	11.851	11.313
SGERD	11.432	9.625	11.365	9.375
ART	10.108	12.313	10.203	11.063

Best results in bold and second results in bold and italic

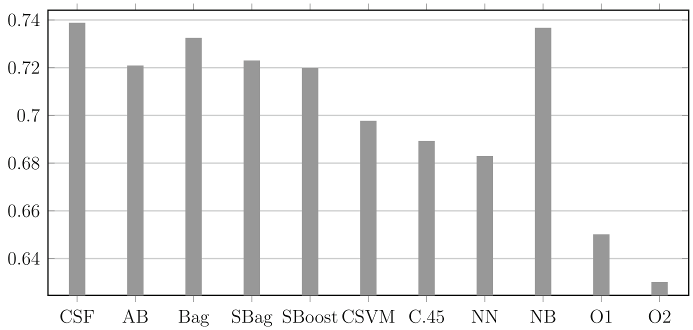


Fig. 6 Summary performance comparison of all algorithms and human oncologists, *O1* and *O2*. The performance is expressed by *classification accuracy* (21). Note both the human oncologists as well as the classifiers performed the classification for the same test set consisting of 100 examples

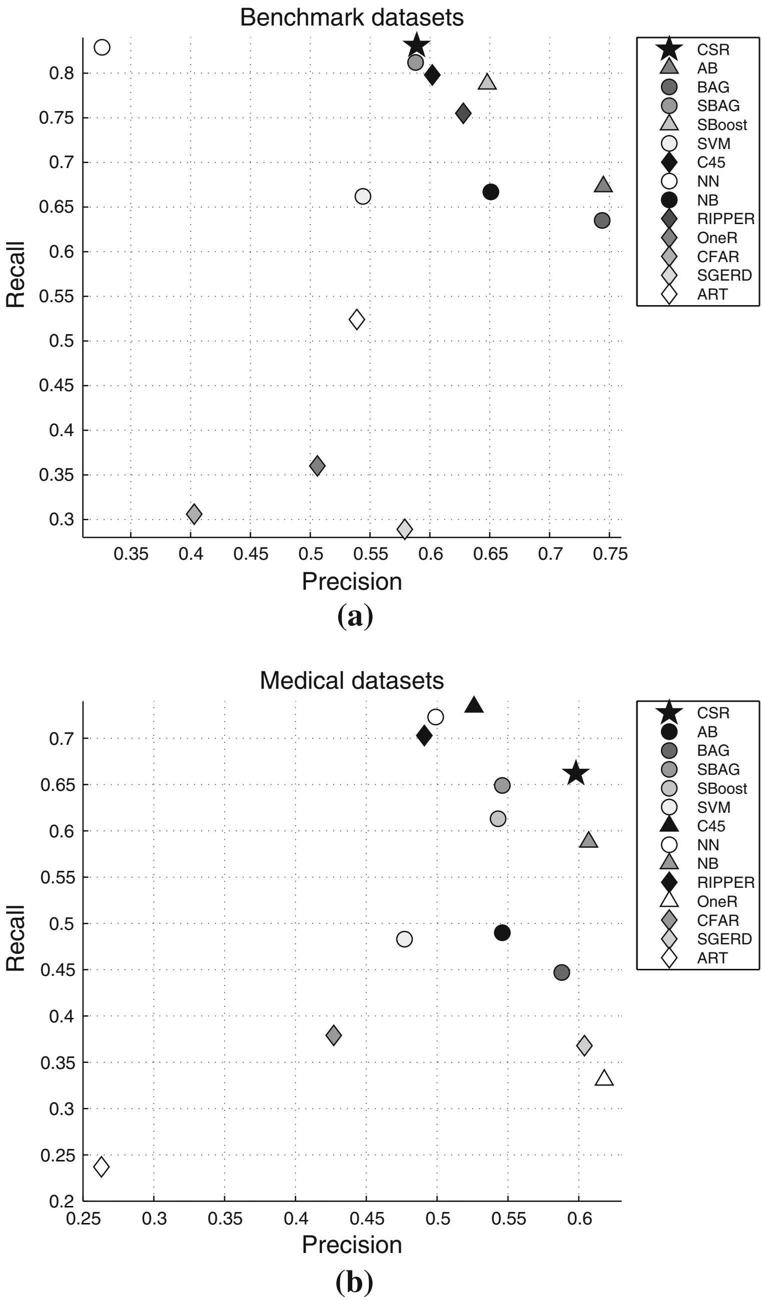
2: Table A2 (for *Gmean*), Table A4 (for *AUC*), Table A6 (for *Precision* and *Recall*). The results of time complexity analysis are gathered in Table A7 (see Electronic Supplementary Material) and more detailed results are presented in Figs. 8, 9, 10, and 11.

The results obtained within the carried out experiments show that according to the *Gmean* and the *AUC* metrics the proposed approach performs comparably with the best ensemble-based classifiers, that is, SMOTEBagging (**SBAG**) and SMOTEBoosting (**SBoost**), and slightly better than the best non-ensemble predictor, i.e., Naïve Bayes classifier (**NB**), see Tables A1, A2, A3, A4 in Electronic Supplementary Material, and the ranks in Table 5. However, it outperforms all rule-based methods where, for instance, the best rule-based methods **C45** and **RIPPER** achieved results worst by several ranks.

It is interesting that the combination of soft rules performed very good on datasets with small number of examples, e.g., ecoli-0-1-3-7vs2-6, glass-016vs5, new-thyroid2, hepatitis, postoperative patient. This effect can be explained twofold. First, in general the idea of combining models increases robustness to overfitting phenomenon. Second, it is possible that the application of the *im*-estimate allows to counteract the small size of the dataset. On the other hand, it can be noticed that the **CSR** achieved slightly worst results on datasets with higher number of attributes, e.g., page-blocks0, vowel0. The plausible explanation of this effect is due to the wrong assumption that the data are generated by one conjunctive feature. Perhaps, other kind of the shared model, e.g., with some kind of disjunction of the conjunctive features, would better represent hidden representation of data.

In terms of *Precision* and *Recall* it turned out that **CSR** is Pareto-optimal in case of both benchmark and medical datasets (see Fig. 7a, b), where methods forming Pareto frontier are denoted by triangular marks and our method is represented by a star). On medical datasets our method achieved balanced values of *Precision* and *Recall*. However, on benchmark datasets **CSR** has very high value of *Recall* but lower value of *Precision* which means that it is able to detect most of positive objects but is prone to label negatives as positives. In the case of medical domain such fact is less dangerous from the patient point of view where it is less harmful to classify a healthy patient as ill (*Recall*) than in the opposite manner (*Precision*). Nonetheless, comparing our approach to other rule-based methods it turns out that only **C45** and **RIPPER** are comparable (on benchmark datasets they are close the Pareto frontier, and on medical datasets **C45** is Pareto-optimal and **RIPPER** is close to the Pareto frontier). Additionally, the **OneR** is Pareto-optimal on medical datasets.

Quite surprising results were obtained by the Naïve Bayes which achieved second rank in terms of **Gmean** and first rank in terms of **AUC** on medical datasets (see Table 5).



(a)

(b)

Fig. 7 Pareto frontier for: **a** benchmark datasets, **b** medical datasets. Methods denoted by *triangular mark* constitutes the Pareto frontier. Rules-based methods are denoted by *diamonds*. Notice that in both cases the **CSR** (denoted by a *star*) is Pareto-optimal

Nevertheless, it is a well-known fact that this Bayesian classifier with features independence assumption behaves extremely good in the medical domain (Kononenko 2001). The general properties of the Naïve Bayes classifier has been thoroughly studied in (Domingos and

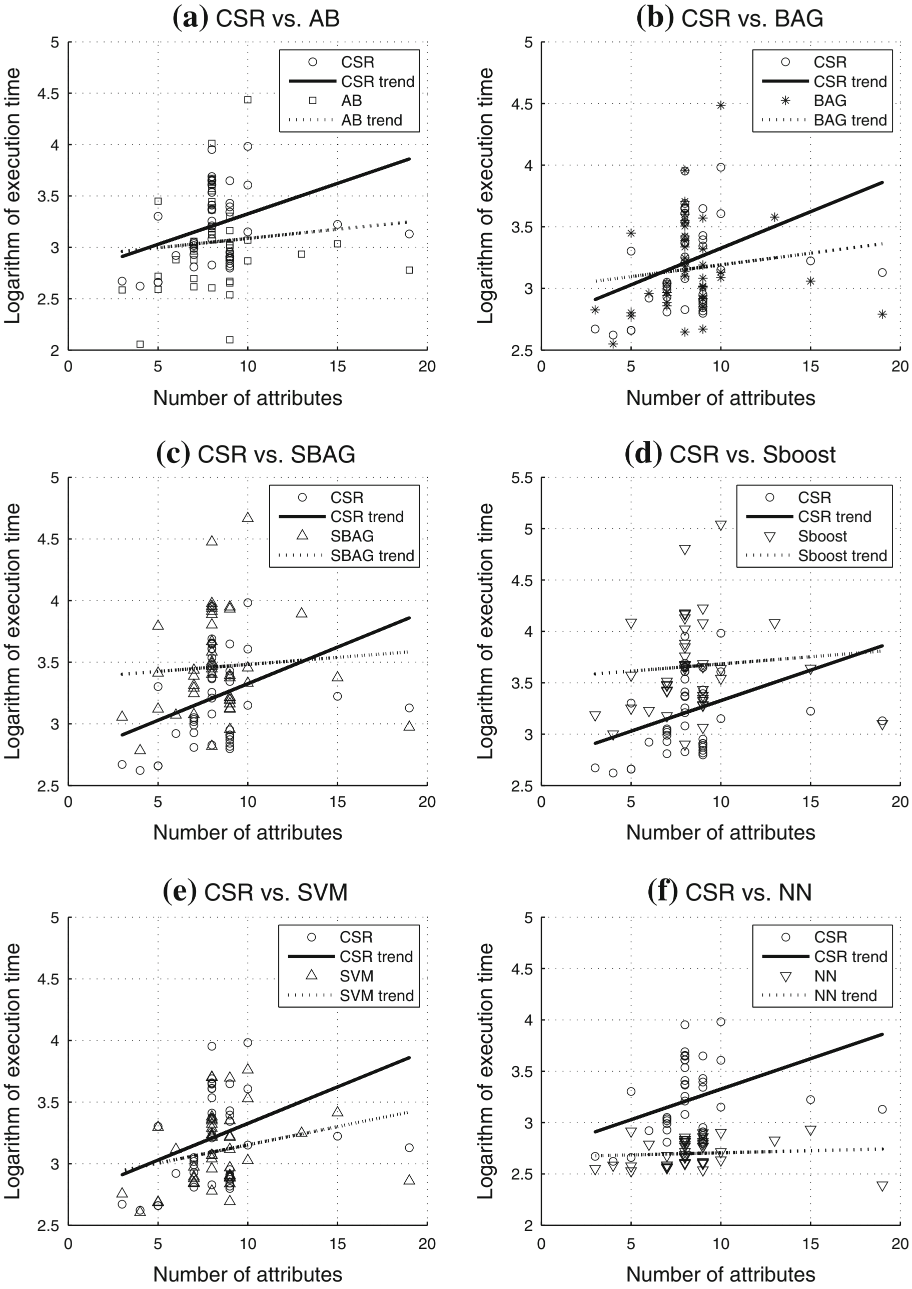


Fig. 8 Execution time comparison between **CSR** and non-rule-based methods with respect to the number of attributes

Pazzani 1997) and theoretical justification for its good performance has been given. However, we notice that our method performs better than **NB** classifier and obtains less varying values (see standard deviations Tables A1, A2, A3, A4 in Electronic Supplementary Material). It is

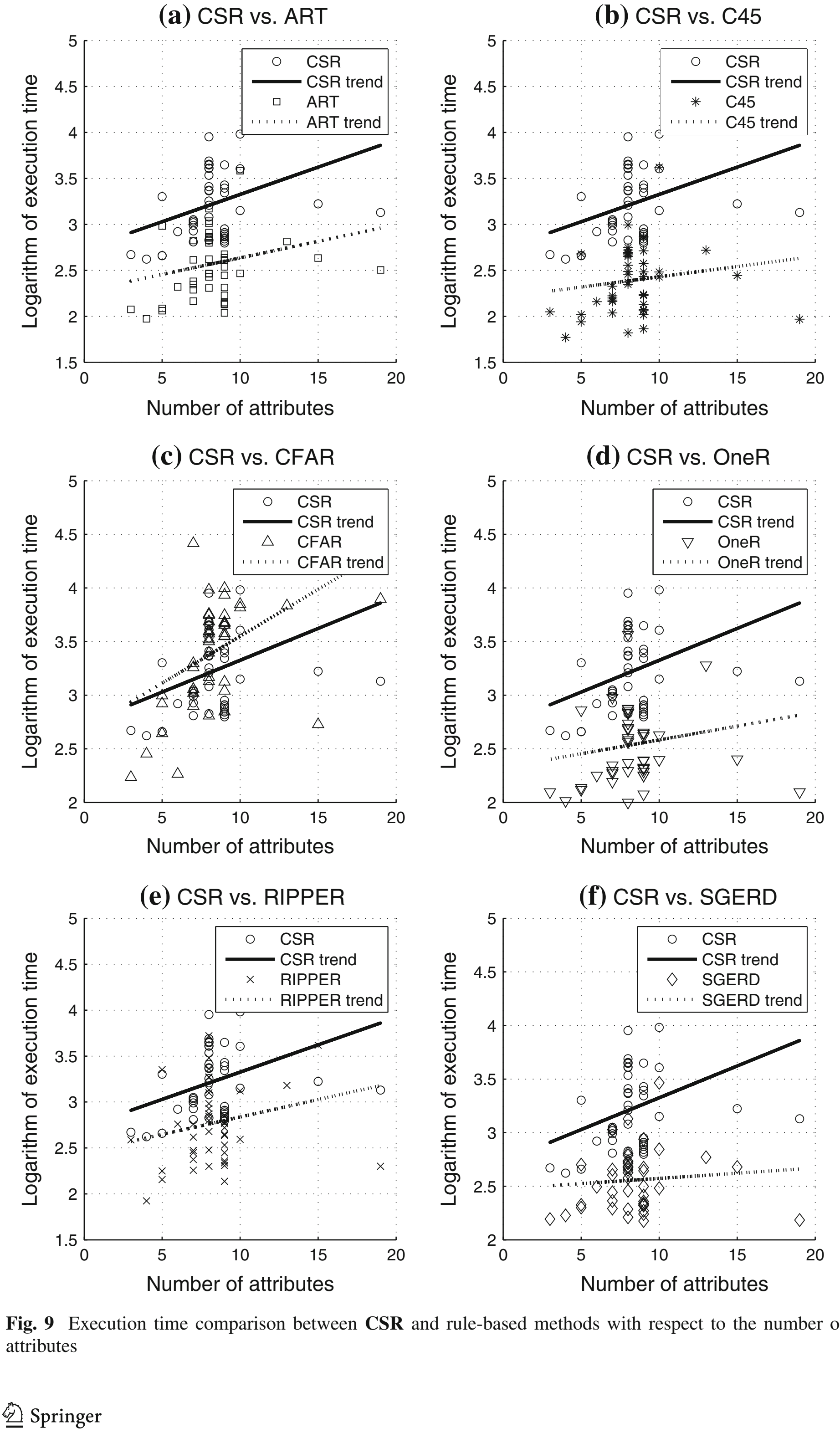


Fig. 9 Execution time comparison between **CSR** and rule-based methods with respect to the number of attributes

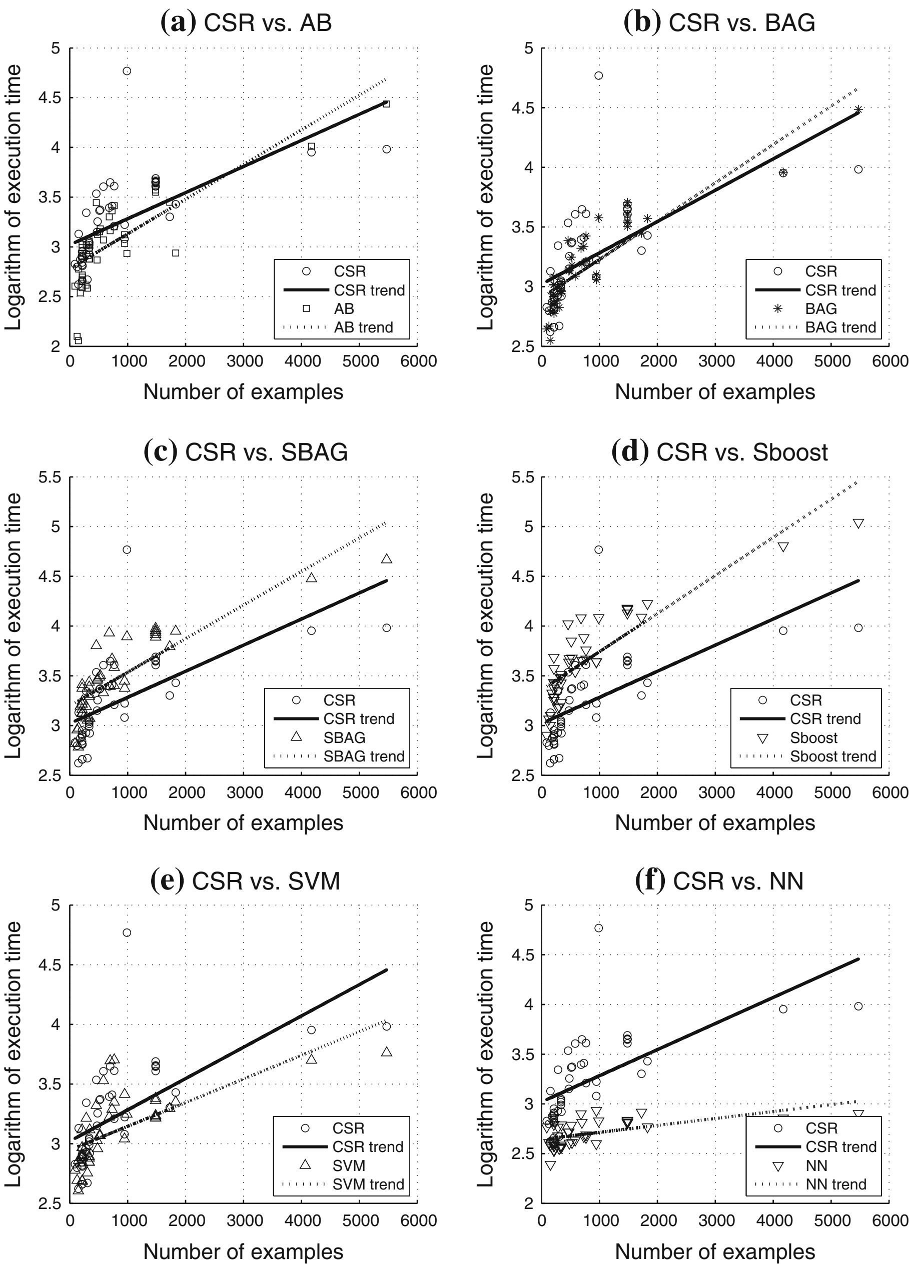


Fig. 10 Execution time comparison between **CSR** and non-rule-based methods with respect to the number of examples

an important result because both **NB** and our method try to approximate the optimal Bayes classifier.

In the paper, we have claimed that main disadvantage of our approach is the exponential growth of soft rules with increasing number of attributes (see Lemma 1). In the experiment

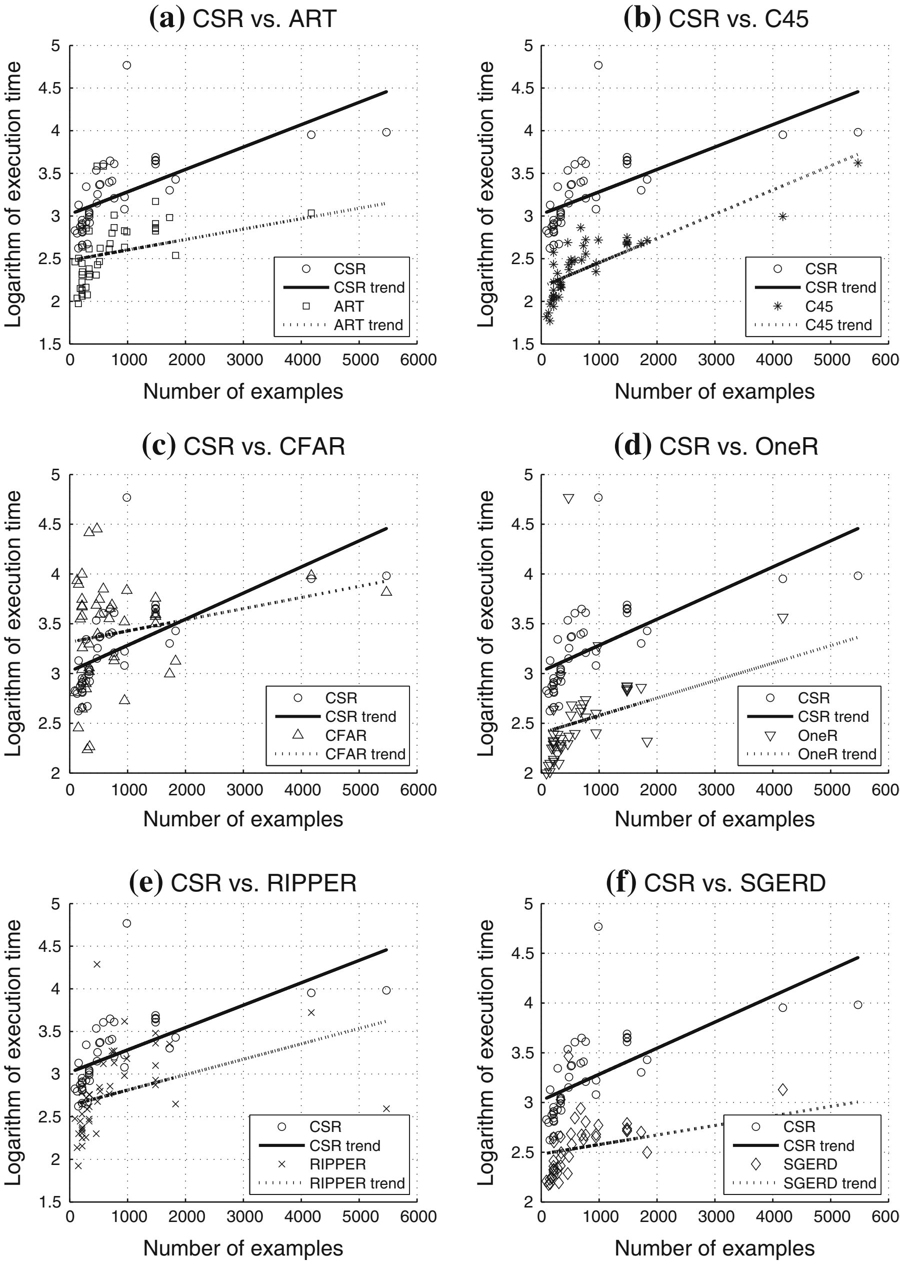


Fig. 11 Execution time comparison between **CSR** and rule-based methods with respect to the number of examples

we verified this claim empirically by calculating the average execution time of five cross-validation folds. In Figs. 8 and 9 the times with respect to the number of attributes for non-rules-based and rules-based methods, respectively, are presented. In Figs. 10 and 11 the times with respect to the growing number of examples for non-rules-based and rules-based methods, respectively, are given. We can notice that the execution time of **CSR** grows as the number of attributes increases, as expected. The same effect is observed in the case of the number of examples. However, we believe that this result can be a consequence of the testing procedure which is more time-consuming for any combination of models (see Fig. 10a–d) for ensemble classifiers). Comparing **CSR** to the non-rule-based methods we notice that our approach performs similarly or even takes less time than other combinations of models, i.e., **AB**, **BAG**, **SBAG**, and **SBoost**. It is especially evident for the ensemble classifiers which apply additional procedure of SMOTE (see Fig. 10c, d). However, **CSR** utilizes noticeably much more time than any of the rule-based methods. The only exception is **CFAR** which performs longer with respect to the number of attributes than our approach (see Fig. 9 c) and comparably with respect to the number of examples (see Fig. 11c).

Last but not least, we would like to address the results obtained by the classifiers in the *oncology* dataset to the ones received by oncologists. The human experts have been presented with 100 cases only and their relatively mediocre performance cannot be exaggerated. Nonetheless, it can be stated that the predictions of the machine learning methods are at least comparable with those of expert oncologists (see Fig. 6).

3.5 Exemplary knowledge for oncology data

At the end of the experimental section we would like to demonstrate exemplary knowledge in terms of soft rules. Let us consider the oncology data and one randomly chosen patient which is described as follows:

1. Menopausal status **false**.
2. Tumor stage **less than 20 mm**.
3. Tumor grade **medium**.
4. Histological type of the tumor **ductal**.
5. Level of progesterone receptors in tumor (in fmol per mg of protein) **more than 10**.
6. Invasiveness of the tumor **no**.
7. Number of involved lymph nodes **0**.
8. Application of a therapy (cTherapy) **false**.
9. Application of a therapy (hTherapy) **false**.
10. Medical history **1st generation breast, ovarian or prostate cancer**.
11. Lymphatic or vascular invasion **false**.
12. Level of estrogen receptors in tumor (in fmol per mg of protein) **more than 30**.
13. Diameter of the largest removed lymph node **less than 15 mm**.
14. Ratio between involved and total lymph nodes removed **0**.
15. Patient age group **under 40**.

Each sentence corresponds to medical history of the patient.

Further, let us assume that we have performed the graph-based memorization basing on the training data. Then we are able to generate soft rules for the given patient's description, for example:

1. IF application of a therapy (cTherapy) **false** AND lymphatic or vascular invasion **false**
THEN $y = 1$ with $p = 0.030$ or $y = -1$ with $p = 0.226$.
2. IF lymphatic or vascular invasion **false**
THEN $y = 1$ with $p = 0.107$ or $y = -1$ with $p = 0.353$.
3. IF menopausal status **false** AND tumor grade **medium**
THEN $y = 1$ with $p = 0.036$ or $y = -1$ with $p = 0.028$.

4. IF application of a therapy (hTherapy) **false**
   THEN $y = 1$ with $p = 0.125$ or $y = -1$ with $p = 0.316$.

We can notice that the second and fourth rule would significantly contribute to the final prediction (see (7)). Nonetheless, it is instructive to examine how class probabilities vary for given antecedent. Including both information about *cTheraphy* and *lymphatic or vascular invasions* drastically decreases probability of $y = 1$ in comparison to information about *lymphatic or vascular invasions* only (see rules 1 and 2). Therefore, one can easily generate the most important rules, i.e., the rules which are the most contributory to the final decision, and present them in form of an interpretable report to the physician or the patient. Moreover, in comparison to the crisp rules, the soft rules enrich the report by additional information about the probability of the class label in the consequent of the rule.

4 Conclusion

We have proposed the combination of soft rules in the application to the medical domain. The approach relies on probabilistic decision making with latent relationships among features represented by the conjunctive features and a new manner of estimating probabilities in case of imbalanced data problem, which is the modified *m*-estimator (*im*-estimator). Moreover, we have presented the graph-based memorization, a technique for aggregating data in the form of a graph. This technique enables an efficient fashion of memorizing observations. We would like also to emphasize that the combination of soft rules is the comprehensible model and can be useful in supporting medical diagnosis.

In the ongoing research we develop the Bayesian approach to the presented problem. Moreover, we focus on establishing a sampling method basing on Markov Chain Monte Carlo technique for approximate inference in high-dimensional spaces. In this paper, we have assumed that data is generated from one conjunctive feature which in fact approximates the true concept representation. Therefore, we leave investigating the inference with a set of conjunctive features as future research.

Acknowledgments The work conducted by Jakub M. Tomczak and Maciej Zięba is co-financed by the European Union within the Human Capital Strategy (*Project: Young Staff 2015. Developing of the University teaching offer and implementing Interdisciplinary Doctoral Studies*). Authors are grateful to Erik Štrumbelj for providing the oncology data. Authors are grateful to reviewers for fruitful comments.

References

Alcalá, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., et al. (2010). KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic and Soft Computing*, *17*, 255–287.

Berzal, F., Cubero, J., Sánchez, D., & Serrano, J. (2004). ART: A hybrid classification model. *Machine Learning*, *54*, 67–92.

Błaszczyński, J., Słowiński, R., & Szeląg, M. (2011). Sequential covering rule induction algorithm for variable consistency rough set approaches. *Information Sciences*, *181*(5), 987–1002.

Blinova, V., Dobrynin, D., Finn, V., Kuznetsov, S. O., & Pankratova, E. (2003). Toxicology analysis by means of the JSM-method. *Bioinformatics*, *19*(10), 1201–1207.

Brazdil, P. B., Soares, C., & Da Costa, J. P. (2003). Ranking learning algorithms: Using IBL and meta-learning on accuracy and time results. *Machine Learning*, *50*(3), 251–277.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, *39*(3), 3446–3453.

Brzezinska, I., Greco, S., & Slowinski, R. (2007). Mining Pareto-optimal rules with respect to support and confirmation or support and anti-support. *Engineering Applications of Artificial Intelligence*, *20*(5), 587–600.

Buntine, W. L. (1992). *A theory of learning classification rules*. PhD thesis, University of Technology, Sydney.

Cawley, G. C. (2006). Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs. In *International Joint Conference on Neural Networks (IJCNN'06)* (pp. 1661–1668). IEEE

Ceglar, A., & Roddick, J. F. (2006). Association mining. *ACM Computing Surveys (CSUR)*, *38*(2), 5.

Cerf, L., Gay, D., Selmaoui-Folcher, N., Crémilleux, B., & Boulicaut, J. F. (2013). Parameter-free classification in multi-class imbalanced data sets. *Data & Knowledge Engineering*, *87*, 109–129.

Cestnik, B. (1990). Estimating probabilities: A crucial task in machine learning. In *ECAI* (pp. 147–149).

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, Z., & Chen, G. (2008). Building an associative classifier based on fuzzy association rules. *International Journal of Computational Intelligence Systems*, *1*(3), 262–273.

Cohen, W. (1995). Fast effective rule induction. In *Machine learning: Proceedings of the Twelfth International Conference* (pp. 1–10).

Cooper, G. F., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine learning*, *9*(4), 309–347.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Daemen, A., & De Moor, B. (2009). Development of a kernel function for clinical data. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2009)* (pp. 5913–5917). IEEE

Damianou, A., Ek, C., Titsias, M., & Lawrence, N. (2012). Manifold relevance determination. In *ICML* (pp. 145–152).

Dembczyński, K., Kotłowski, W., & Słowiński, R. (2008). Maximum likelihood rule ensembles. In *ICML* (pp. 224–231). ACM

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, *7*, 1–30.

Domingos, P. (1997). Bayesian model averaging in rule induction. In *Preliminary papers of the sixth international workshop on artificial intelligence and statistics* (pp. 157–164).

Domingos, P. (2000). Bayesian averaging of classifiers and the overfitting problem. In *ICML* (pp. 223–230).

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, *29*(2), 103–130.

Džeroski, S., Cestnik, B., & Petrovski, I. (1993). Using the m-estimate in rule induction. *Journal of Computing and Information Technology*, *1*(1), 37–46.

Ek, C. H., Rihan, J., Torr, P. H., Rogez, G., & Lawrence, N. D. (2008). Ambiguity modeling in latent spaces. In *Machine learning for multimodal interaction* (pp. 62–73). Springer

Elkan, C. (2001). The foundations of cost-sensitive learning. In *IJCAI* (Vol. 17, pp. 973–978).

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874.

Fawcett, T., & Provost, F. (1997). Adaptive fraud detection. *Data Mining and Knowledge Discovery*, *1*(3), 291–316.

Fayyad, U., & Irani, K. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *IJCAI* (pp. 1022–1029).

Frank, A., & Asuncion, A. (2010). *UCI machine learning repository*. http://archive.ics.uci.edu/ml

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139.

Fürnkranz, J. (1999). Separate-and-conquer rule learning. *Artificial Intelligence Review*, *13*(1), 3–54.

Fürnkranz, J., & Flach, P. A. (2003). An analysis of rule evaluation metrics. In *Proceedings of the 20th international conference on machine learning (ICML-03)* (pp. 202–209).

Galar, M., Fernández, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, *42*(4), 463–484.

Gama, J. (2000). A cost-sensitive iterative Bayes. In *ICML*.

Hall, M. (1999). *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato.

He, H., & Garcia, E. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284.

Holte, R. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, *11*, 63–91.

Japkowicz, N. (2001). Concept-learning in the presence of between-class and within-class imbalances. In *Advances in artificial intelligence* (pp. 67–77). Springer

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, *6*(5), 429–449.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, *4*(3), 227–241.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London Series A Mathematical and Physical Sciences*, *186*(1007), 453–461.

Kononenko, I. (1992). Combining decisions of multiple rules. In *AIMSA* (pp. 87–96).

Kononenko, I. (2001). Machine learning for medical diagnosis: History, state of the art and perspective. *Artificial Intelligence in Medicine*, *23*(1), 89–109.

Kotsiantis, S. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, *31*, 249–268.

Kubat, M., & Matwin, S. (1997). Addressing the curse of imbalanced training sets: One-sided selection. In *ICML* (pp. 179–186).

Kubat, M., Holte, R., & Matwin, S. (1997). Learning when negative examples abound. In *ECML* (pp. 146–153).

Lavrač, N. (1999). Selected techniques for data mining in medicine. *Artificial Intelligence in Medicine*, *16*(1), 3–23.

Liu, B., Hsu, W., & Ma, Y. (1998). Integrating classification and association rule mining. In *4th International conference on knowledge discovery and data mining (KDD98)* (pp. 80–86).

Mac Namee, B., Cunningham, P., Byrne, S., & Corrigan, O. (2002). The problem of bias in training data in regression problems in medical decision support. *Artificial Intelligence in Medicine*, *24*(1), 51–70.

Mansoori, E., Zolghadri, M., & Katebi, S. (2008). SGERD: A steady-state genetic algorithm for extracting fuzzy classification rules from data. *IEEE Transactions on Fuzzy Systems*, *16*(4), 1061–1071.

Masnadi-Shirazi, H., & Vasconcelos, N. (2010). Risk minimization, probability elicitation, and cost-sensitive SVMs. In *ICML* (pp. 204–213).

Michalski, R. S., Mozetic, I., Hong, J., & Lavrac, N. (1986). The multi-purpose incremental learning system AQ15 and its testing application to three medical domains. In *Proc AAAI* (pp. 1–41).

Minka, T. P. (2000). *Bayesian model averaging is not model combination* (pp. 1–2). http://research.microsoft.com/en-us/um/people/minka/papers/minka-bma-isnt-mc

Mitchell, T. (1997). *Machine learning*. Boston: McGraw Hill.

Pawlak, Z., Grzymala-Busse, J., Slowinski, R., & Ziarko, W. (1995). Rough sets. *Communications of the ACM*, *38*(11), 88–95.

Pearson, R., Goney, G., & Shwaber, J. (2003). Imbalanced clustering for microarray time-series. In *ICML*.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Burlington: Morgan Kaufmann.

Rückert, U., & Kramer, S. (2008). Margin-based first-order rule learning. *Machine Learning*, *70*(2–3), 189–206.

Shon, A., Grochow, K., Hertzmann, A., & Rao, R. P. (2005). Learning shared latent structure for image synthesis and robotic imitation. In *Advances in neural information processing systems* (pp. 1233–1240).

Stefanowski, J. (1998). On rough set based approaches to induction of decision rules. *Rough Sets in Knowledge Discovery*, *1*(1), 500–529.

Stefanowski, J., & Wilk, S. (2006). Rough sets for handling imbalanced data: Combining filtering and rule-based classifiers. *Fundamenta Informaticae*, *72*(1), 379–391.

Štrumbelj, E., Bosnić, Z., Kononenko, I., Zakotnik, B., & Grašič Kuhar, C. (2010). Explanation and reliability of prediction models: The case of breast cancer recurrence. *Knowledge and Information Systems*, *24*(2), 305–324.

Tenenbaum, J., & Griffiths, T. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640.

Tomczak, J., & Gonczarek, A. (2013). Decision rules extraction from data stream in the presence of changing context for diabetes treatment. *Knowledge and Information Systems*, *34*, 521–546.

Vamplew, P., Dazeley, R., Berry, A., Issabekov, R., & Dekker, E. (2011). Empirical evaluation methods for multiobjective reinforcement learning algorithms. *Machine Learning*, *84*(1–2), 51–80.

Viswanathan, M., & Webb, G. I. (1998). Classification learning using all rules. In *Machine learning: ECML-98* (pp. 149–159).

Vorontsov, K., & Ivahnenko, A. (2011). Tight combinatorial generalization bounds for threshold conjunction rules. In *Pattern recognition and machine intelligence* (pp. 66–73). Springer

Vreeken, J., Van Leeuwen, M., & Siebes, A. (2011). Krimp: Mining itemsets that compress. *Data Mining and Knowledge Discovery*, *23*(1), 169–214.

Wang, B., & Japkowicz, N. (2010). Boosting support vector machines for imbalanced data sets. *Knowledge and Information Systems*, *25*(1), 1–20.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259.

Zadrozny, B., & Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *ICML* (pp. 609–616).

Zięba, M., Tomczak, J. M., Lubicz, M., & Świątek, J. (2014). Boosted SVM for extracting rules from imbalance data in application to prediction of the post-operative life expectancy in the lung cancer patients. *Applied Soft Computing*, *14*, 99–108.