Published in final edited form as:

IEEE Trans Nanobioscience. 2012 September ; 11(3): 228–236. doi:10.1109/TNB.2012.2213264.

Combined Rule Extraction and Feature Elimination in Supervised Classification

Sheng Liu,

Department of Computer and Information Science, University of Mississippi, University, MS 38677, USA

Ronak Y. Patel,

Department of Medicinal Chemistry, University of Mississippi, University, MS 38677, USA

Department of Genetics, Washington University School of Medicine, Saint Louis, MO 63108, USA

Pankaj R. Daga,

Department of Medicinal Chemistry, University of Mississippi, University, MS 38677, USA

Haining Liu,

Department of Medicinal Chemistry, University of Mississippi, University, MS 38677, USA

Gang Fu,

Department of Medicinal Chemistry, University of Mississippi, University, MS 38677, USA

Robert J. Doerksen,

Department of Medicinal Chemistry, University of Mississippi, University, MS 38677, USA

Yixin Chen, and

Department of Computer and Information Science, University of Mississippi, University, MS 38677, USA

Dawn E. Wilkins

Department of Computer and Information Science, University of Mississippi, University, MS 38677, USA

Abstract

There are a vast number of biology related research problems involving a combination of multiple sources of data to achieve a better understanding of the underlying problems. It is important to select and interpret the most important information from these sources. Thus it will be beneficial to have a good algorithm to simultaneously extract rules and select features for better interpretation of the predictive model. We propose an efficient algorithm, Combined Rule Extraction and Feature Elimination (CRF), based on 1-norm regularized random forests. CRF simultaneously extracts a small number of rules generated by random forests and selects important features. We applied CRF to several drug activity prediction and microarray data sets. CRF is capable of producing performance comparable with state-of-the-art prediction algorithms using a small number of decision rules. Some of the decision rules are biologically significant.

Index Terms

Rule extraction; Feature selection; Multi-Class classification; Random forests

I. Introduction

Many computational methods are developed to analyze biological data from different viewpoints, e.g., prediction, feature subset selection, and rule-based interpretation. As these methods are geared towards specific purposes, their domains of application usually do not overlap: classifiers with high prediction performance, e.g., support vector machines (SVM) [53] or random forests (RF) [7], may be a “black box” model lacking interpretability; while an easy-to-interpret model, such as a decision tree [8], does not compete well with SVM or RF on classification accuracy. In many biological applications, interpretability of a model per se (i.e., relevant features and predictive rules) is sometimes as important as the prediction performance. Finding a right tradeoff between prediction performance and model interpretability is thus important. This is the problem that we attempt to tackle here. There is a rich resource of prior work on rule-based learning and feature selection in the fields of bioinformatics and statistical learning. It is beyond the scope of this article to supply a complete survey of the respective areas. Below we review some of the main findings most closely related to this article.

A. Rule-Based Learning

Rule-based classifiers are well known for their capability of shedding light on the decision process in addition to making a prediction. Rule extraction techniques roughly fall into two categories: direct approaches and indirect approaches.

In a direct approach, rules are generated by solving a single optimization problem. For example, Gopalakrishnan et al. [21] adopted a Bayesian network approach to rule learning and evaluated the rules using Bayesian scores. In [26], Jiang et al. used a simulated annealing bump hunting strategy to derive interpretable rules. Li et al. [30] proposed a search algorithm to discover significant rules from bio-medical data based on emerging patterns. Among the direct approaches, decision trees are arguably the most popular technique [8], [11], [40], [41]. Although the rules of a decision tree are easy-to-interpret, the prediction performance is, in general, not competitive against many “black box” classifiers such as SVM and artificial neural networks. Cohen [12] proposed a rule extraction algorithm based on decision trees. Its prediction performance is comparable to decision tree methods. In [31], a fast algorithm based on decision tree committees was developed that scales well with high dimensional data.

In an indirect approach, a predictive model is built from training data. Rules are then extracted from the model. Artificial neural networks (ANN) and SVM are two of the most popular algorithms used to build predictive models. Rule extraction from ANNs has been investigated by many researchers [4], [14], [15], [19], [24], [27], [34], [42], [44], [46], [49]–[52]. Towell and Shavlik [50] divided the process into three steps and provided an efficient

solution for the most challenging step, knowledge extraction. Huysmans et al. [23] and Jacobsson [24] surveyed ANN-based rule extraction techniques.

SVM-based rule extraction has also been explored extensively due to the high performance of SVM. Núñez et al. [37] proposed a clustering method to interpret an SVM classifier. Martens et al. [35] used an active learning approach to extract rules from the support vectors and decision boundary of an SVM. Barakat and Diederich [6] improved the scalability of the method in [37] using a learning based approach that utilizes the support vectors and kernel parameters to obtain rules. Fung et al. [18] formulated rule extraction from hyperplane based linear classifiers as a constrained optimization problem. In [16], Diederich et al. presented an introduction of rule extraction from SVM and described several applications. The importance of bridging the gap between decisions of prediction algorithms and human interpretable rules as well as some limitations and opportunities for rule extraction from SVM were investigated.

B. Feature Selection

Some of the most successful feature selection algorithms combine a classification algorithm with some heuristics for post-processing. For reviews on feature selection in bioinformatics, we refer interested readers to [22], [43], and [33]. In general, feature selection can be categorized as filter methods, wrapper methods, or embedded approaches. Filter methods do not require a learning algorithm for a predictive model and are normally regarded as a pre-processing step of other algorithms. Wrapper methods use some search heuristics to find a subset of features that gives the best predictive model using a given learning algorithm. In embedded methods, feature selection is built into a learning algorithm.

The search algorithm is probably the most important component of a feature selection method. Several search strategies have been applied for feature selection [29]: branch-and-bound, divide and conquer, greedy method, evolutionary algorithm, annealing, etc. Among them, a greedy search strategy, such as forward selection (incremental search) or backward elimination, is one of the most popular techniques because of its computational efficiency, robustness to overfitting [22], and typically low variance.

C. Our Contribution

In many biological problems, building a good predictive model that explains the problem is the ultimate goal of modeling. It is hence necessary to simultaneously identify meaningful rules and important features to interpret the underlying biological principle.

High accuracy and compact representation (i.e., a small number of rules and a small number of features used) are two important requirements of rule extraction methods. Decision tree based methods usually generate a small set of rules. However, their accuracies are relatively low compared with SVM and RF. RF improves the accuracy, but generates a large number of rules, making it difficult to interpret the model. In this article, we take an iterative approach to sparse encoding of random forests to obtain refined rules without compromising the accuracy of a RF. With an ensemble of decision trees, a RF covers more candidate rules compared with a single decision tree, hence giving better performance. Sparse encoding keeps only a small number of rules that are the most discriminative. In a closely related work

by Friedman et al. [17], RF was used as an ensemble of rules for prediction in the context of rule fitting. By contrast, in our proposed method the context is sparse encoding. We can also handle multi-class data in addition to rule extraction and feature selection. For feature selection, we take an embedded approach with a greedy back elimination strategy.

The proposed combined rule extraction and feature elimination (CRF) method is a hybrid approach to rule extraction and feature selection in which they interact with each other. The result of rule extraction is used for feature selection. The selected features are then fed into RF and a sparse encoding step to extract important rules. The iterative alternating approach continues until the selected subset of features does not change.

II. Methods

In this section, we describe the proposed method. First, we present a binary encoding of the forest generated by random forests. It maps a sample point to a space defined by the entire set of leaf nodes (rules) of the forest. We then introduce a sparse encoding method to extract representative rules using the binary encoding of the forest. Changes in the distribution of features in the selected rules are used to select a subset of features, which are then used to construct a new RF to generate a new set of rules. This procedure is repeated until a stopping criteria is satisfied.

A. Binary Encoding of Random Forests

Random forests (RF) is an ensemble learning method which uses a large number of decision trees. A decision tree learning algorithm creates a tree model that predicts target values. It is performed by recursively splitting data into subsets using one of the variables. Gini impurity [8] and information gain [41] are two commonly used data splitting criteria.

As the path from a root node to a leaf node is interpreted as a decision rule, a RF is equivalently represented as a collection of decision rules. Because each sample traverses each decision tree from root node to one and only one leaf node, we define a binary feature vector to capture the leaf node structure of a RF. For sample x_i , the corresponding binary vector that encodes the leaf node assignment is defined as $X_i = [X_1, \dots, X_q]^T$ where q is the total number of leaf nodes in a forest,

$$X_j = \begin{cases} 1 & \text{if } x_i \text{ reaches the } j\text{-th leaf node,} \\ 0 & \text{otherwise} \end{cases} \quad j = 1, \dots, q.$$

Figure 1 illustrates the mapping. For illustration purpose, each sample traverses to a randomly selected leaf node on each tree. In real situations, the leaf node is selected based on the split rules on each node.

We call the space of X_i 's a leaf node space. In this space, each sample is mapped to a vertex of a hypercube. Each dimension of the rule space is defined by one decision rule. Therefore, for sample x_i , X_i essentially defines which rules are active (X_j of X_i is 1) and which are inactive ($X_j = 0$). The above binary mapping has been applied in [32] to combine RF with large margin learning.

B. Rule Extraction using Sparse Encoding

Using the above mapping, we obtain a new set of training samples, $\{(X_1, y_1), (X_2, y_2), \dots, (X_p, y_p)\}$, where X_i is a vector of binary attributes and $y_i \in \{1, 2, \dots, K\}$ is the corresponding class label. We consider classifiers of the form

$$y = \arg \max_{k \in \{1, \dots, K\}} (w_k^T X + b_k) \quad (1)$$

where the weight vector w_k and the scalar b_k define the linear discriminant function for the k -th class.

As each binary attribute represents a decision rule, the weights in (1) measure the class specific importance of rules: the magnitude of a weight indicates the importance of the rule. Clearly, in the above classifier, a rule can be removed safely if its weights for all K classes are 0. Rule extraction is therefore formulated as a problem of learning the weight vectors.

We consider the following learning problem using 1-norm regularization:

$$\min_{w_k, \xi_{ik}} \left(\lambda \sum_{k=1}^K \|w_k\|_1 + \sum_{i=1}^p \sum_{k=1, \dots, K, k \neq y_i} \xi_{ik} \right)$$

$$\text{s.t.} \quad (w_{y_i} - w_k)^T X_i + b_{y_i} - b_k + \xi_{ik} \geq 1$$

$$\xi_i \geq 0, \quad i = 1, \dots, p \quad (2)$$

The objective function consists of two terms. Because of the sparse favoring property of the 1-norm regularization, the first term, $\sum_{k=1}^K \|w_k\|_1$, controls the number of nonzero weights, and hence the number of rules extracted. The second term is the sum of slack variables ξ_{ik} . Because a nonzero slack variable indicates a misclassified sample, the second term is correlated with the empirical error. The tradeoff between the sparseness of a solution and the empirical error is determined by the regularization parameter, λ . 1-norm sparse encoding has been widely applied in statistics and machine learning, e.g., [48], [57], and [58]. The above optimization is a linear program (LP).

C. Combined Rule Extraction and Feature Elimination

The distribution of features in a RF is determined by the RF learning process. It is in general different from the distribution of features among the rules extracted from (2). We take advantage of this difference to select features based on the assumption that important features are retained in the decision rules extracted. Features that do not appear in the rules

extracted from (2) are removed because they have no effect on the classifier defined in (1). In this way, we can select rules and features together.

The regularization parameter λ is chosen by cross validation on the training set. It is possible to further select rules from a RF built on the selected features to get a more compact set of rules. This motivates an iterative approach. Features selected in the previous iteration are used for constructing a new RF. A new set of rules is then extracted from the new RF. This process continues until the selected features do not change.

D. An Algorithmic View

The following pseudo code illustrates the overall workflow of the algorithm:

Require:

Initial feature set $\mathbf{F}$

Training set $\mathcal{X}$ of p samples

Ensure:

Selected Features $\mathbf{F}_f$

Selected Rules $\mathbf{R}_f$

```

1:  $i \leftarrow 1, \mathbf{F}_0 \leftarrow \emptyset, \mathbf{F}_i \leftarrow \mathbf{F}$ 
2: while  $(\mathbf{F}_i - \mathbf{F}_{i-1}) \neq \emptyset$  do
3:   Run random forests on  $\mathcal{X}$  with feature set  $\mathbf{F}_i$ 
4:   Random forests generates a set of rules  $\mathbf{R}_i$ 
5:   Use  $\mathbf{R}_i$  to encode  $\mathcal{X}$ 
6:   Solve the linear program in (2) to get  $\mathbf{w}_k$ 's and cross
   validation training accuracy  $\mathbf{C}_i$ 
7:   index  $\leftarrow \bigcup_{k=1, \dots, K} \text{index of } (\mathbf{w}_k > \text{threshold, which is set by}$ 
   user and is a small positive number or 0)
8:    $\mathbf{R}_i \leftarrow \mathbf{R}_i(\text{index})$ 
9:    $\mathbf{F}_{i+1} = \{\text{Features used in } \mathbf{R}_i\}$ 
10:   $i \leftarrow i + 1$ 
11: end while
12:  $i^* = \arg \max_i \mathbf{C}_i$ 
13:  $\mathbf{F}_f \leftarrow \mathbf{F}_{i^*+1}$ 
14:  $\mathbf{R}_f \leftarrow \mathbf{R}_{i^*}$ 
15: return  $\mathbf{F}_f, \mathbf{R}_f$ 

```

III. Results and Discussion

In this section, we first describe the data sets used. We then present detailed results and discussion.

A. Data Sets

We applied the method to several data sets. The first data set is prospectr data [1], which is used to evaluate the rule extraction method described in Section II-B. The prospectr data set contains gene-based representations of human hereditary diseases. We combine the training set (OMIM training set) and test set (HGMD test set) described in [1]. There are 3586 samples each with 61 features. Most of them are numerical features. We removed features with missing values, resulting in 53 features. To get an idea of how well the method will perform on unseen data, we use five fold cross validation accuracy for this data set, while for the following data sets, we report test accuracies.

We then applied CRF to Cannabinoid (CB) receptor data sets and P-glycoprotein (P-gP) data set. CB compounds include published compounds reported as CB1 antagonists, CB2 agonists and CB2 antagonists, classified using cutoffs into active/inactive, selective/moderately selective/nonselective (for CB1 antagonists and CB2 agonists) and selective/nonselective (for CB2 antagonists), hence forming in total six data sets. For each data set, the compounds were divided into training and test set using a 70/30 split (see Table I). In order to balance the data set to have equal occupancy of each class, some randomly selected compounds from the Asinex database were put in the training and/or test sets as decoys and labeled as inactive or non-selective. For each data set two sets of features were calculated, a smaller set (~300) calculated using the MOE software [10] and a larger set (~3000) calculated using the DragonX software [36]. For the P-gP data set [39], compounds were classified as substrates or non-substrates of the P-glycoprotein transporter. The compounds were divided into training and test set making sure that each set had roughly the same structural diversity and activity distribution, based on Soergel index diversity analysis using calculated hashed binary fingerprints. The features used were calculated from physicochemical descriptors of the compounds. When using features from the DragonX software for CB1/CB2 or P-gP, constant features were removed before applying the algorithm. The statistics of the above data sets are listed in Table I.

Finally, we applied our method to four microarray gene expression data sets. One data set is Golub data [20] on human leukemias, Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML). There are 72 samples each containing expression values of 7129 probe sets (38 samples are used for training and 34 samples are used for testing). Another data set is Alon data [3], which contains 62 samples with 2000 gene expression values for each sample. The 62 tissue samples include 22 normal and 40 colon cancer tissues. The third data set, the Nutt data, is high-grade glioma brain cancer data [38] which contains 50 samples of 12625 expression values. The van't Veer data set [54] involves breast cancer data consisting of 78 samples each represented by 24418 expressions. Of these, 44 samples are in the "good" prognosis class whose patients remained free of disease after their initial diagnosis for an interval of at least five years. The remaining 34 samples are in the "poor" prognosis class whose patients had developed distant metastases within five years. Except

for the small Golub data set, 10 percent of the samples are used for test, while the rest are used for training.

B. Rule Extraction using Sparse Encoding

We applied the rule extraction method in Section II-B to the prospectr data to illustrate the effect of the regularization parameter λ . The experiment was done by separating all samples into five folds, and generating cross validation accuracies. The results shown are the average of accuracies on the validation sets. Figure 2 shows the number of rules extracted using different values of λ . When λ is zero, where there is no regularization, the number of rules extracted is in the upper middle range. As λ increases, the number of rules rises rapidly, then drops fast and then slowly approaches zero. The accuracy does not decrease significantly when λ is small as Figure 3 shows. Even with a large value of λ , say 40, where the number of rules extracted is very small, the accuracy is still considerably high.

C. Combined Rule Extraction and Feature Elimination

As our method is based on random forests, which is in turn based on decision trees, we compared our results with a decision tree-based rule extraction algorithm, RIPPER [12]. RIPPER uses a divide and conquer strategy and splits samples into a growing set and a pruning set. The growing set is used to generate decision trees, while the pruning set is used to prune the trees according to reduced error criteria. Once a rule is generated, all the samples covered by the rule are removed from the sets. This process repeats until all samples are covered.

1) Results on CB1 and CB2 data sets—Table II shows the number of rules generated by a RF and extracted by CRF on CB1 and CB2 data sets using the smaller feature sets (MOE software). On all six data sets, the number of rules selected by CRF is less than 1% of the total number of rules generated by the RF. Table III shows the test accuracy of the different methods. In each case, RF delivers the best performance. CRF is more competitive than RIPPER. On the larger feature sets, Table IV shows the number of rules generated by a RF and extracted by CRF. On all six data sets, the number of rules selected by CRF is less than 1% of the total number of rules generated by a RF. Table V lists the test accuracy. Again, RF produces the best prediction accuracy. CRF outperforms RIPPER. Note that although CRF's accuracy is slightly lower than RF, it only uses less than 1% of the rules of a RF, hence significantly improving interpretability.

The extracted rules are ranked with respect to the magnitude of weights. Table VI shows a sample of descriptors used in the top-ranked rule on one CB data set - CB1 antagonist activity with the larger DragonX feature set. Considering the top-ranked selected rule for each of the six CB dataset CRF models, we noted that the descriptors in the activity data sets are different from those in the corresponding selectivity data sets. This suggests that the activity and selectivity are determined by different physicochemical properties. For some data sets, the top-ranked rule contains descriptors of the same category, suggesting this category plays an important role in the determination of activity or selectivity. For some other data sets, the top-ranked rule contains descriptors from various categories, suggesting variable kinds of effects affect the determination of different classes.

2) Results on P-gP (P-glycoprotein) data set—As indicated in Table VII, classification is a harder problem for P-gP data than for the CB data. The accuracy of RF is low. CRF performs significantly better than RIPPER.

The descriptors that occurred in the top-ranked rule are listed in Table VIII. It is interesting to observe that in the top-ranked rule, except for 2 descriptors from the category ‘3D-MoRSE descriptors’, all the others are from different categories. These categories are highly independent and uncorrelated.

3) Results on Microarray data sets—Four microarray data sets were tested. The number of rules extracted are listed in Table IX. RIPPER consistently generates fewer rules on all but the Colon data set, for which CRF produces the least number of rules. The accuracies are presented in Table X. The accuracies of RF and CRF are comparable. CRF significantly outperforms RIPPER on all but Nutt data set, on which they tied. Overall, the rules extracted by CRF generate better prediction than RIPPER.

CRF generates three rules from the Golub data set. They are listed below. For ease of interpretation, the corresponding proteins of selected genes are presented in the rules.

1. **IF** “Zyxin” > 994.0 **THEN** AML.
2. **IF** “Catalase (EC 1.11.1.6) 5’ flank and exon 1 mapping to chromosome 11, band p13 (and joined CDS)” 2773.0 **AND** “HLA Class II Histocompatibility Antigen, DR Alpha Chain Precursor (HCHADACP)” 8774.5 **THEN** ALL T-Cell.
3. **IF** “Zyxin” 994.0 **AND** “HCHADACP” > 8953.0 **THEN** ALL B-Cell.

Zyxin is a zinc-binding adaptor phosphoprotein that is localized at focal adhesion plaque complex. Zyxin may regulate cytoskeletal dynamics and mediate adhesion-stimulated changes in gene expression in signal transduction [13], [25]. It has been shown by [56] computationally and biologically that Zyxin plays an important role in leukaemogenesis. Many other researchers have verified this, to list a few, [5], [47], and [55].

Catalase is an enzyme that catalyzes the conversion of hydrogen peroxide to water and oxygen [13], [25]. Kidson [28] found the association of catalase level with myeloid leukemia. Wang [2] also identified the catalase gene in 50 top-ranked genes.

HLA Class II Histocompatibility Antigen, DR Alpha Chain Precursor (HCHADACP) involves the presentation of foreign antigens to the immune system [13], [25]. Schoch et al. [45] found this protein helps to discriminate between distinct cytogenetic AML subtypes.

The above rules have some interesting interpretations. For example, Rule 3 seems to suggest that if Zyxin is at a low level, and at the same time HCHADACP is at a high level, it is likely to be an ALL B-Cell case.

On the Colon data set, CRF extracts just one rule with an accuracy identical to that of RF:

1. **IF** “Collagen Alpha 2(XI) Chain” 84.878750 **AND** “Human vasoactive intestinal peptide (VIP) mRNA, complete cds.” > 50.168453 **AND** “Human mRNA for receptor of retinoic acid.” > 96.290622 **THEN** normal.

On the Nutt and van't Veer data set, 33 and 55 rules are generated, respectively. The accuracy of CRF is significantly better than RIPPER on the van't Veer data set. These results demonstrate the high potential of CRF in generating predictive models with good interpretability.

IV. Conclusions

We propose to use an ensemble of decision rules generated from random forests and one norm regularization to extract meaningful rules. The method selects a small number of rules (using a small number of features) while retaining performance comparable to RF. It also handles multi-class data which are common in practical applications. CRF balances the accuracy and interpretability of a model, as demonstrated on several data sets.

Acknowledgments

This material is based on work supported by the National Science Foundation under award numbers NSF EPS-0903787 and MCB-1027989. Cannabinoid (CB) Receptor Subtype data sets were from the project supported in part by USA National Institutes of Health (NIH) Grant Number 5P20RR021929 from the National Center for Research Resources (NCRR). This investigation was conducted in part in a facility constructed with support from research facilities improvement program C06 RR-14503-01 from the NIH NCRR. The content is solely the responsibility of the authors and does not necessarily represent the official views of NCRR or NIH.

References

1. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*. 6:55.2005; [PubMed: 15766383]
2. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS. SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics*. 22(6):773–774.Mar; 2006 [PubMed: 16423925]
3. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A*. 96(12):6745–6750.Jun; 1999 [PubMed: 10359783]
4. Andrews R, Diederich J, Tickle AB. Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems*. 8(6):373–389.1995;
5. Baker SG, Kramer BS. Identifying genes that contribute most to good classification in microarrays. *BMC Bioinformatics*. 7:407.2006; [PubMed: 16959042]
6. Barakat N, Diederich J. Eclectic Rule-Extraction from support vector machines. *International Journal of Computational Intelligence*. 2(1):59–62.2005;
7. Breiman L. Random Forests. *Mach Learn*. 45(1):5–32.2001;
8. Breiman, L, Friedman, J, Olshen, R, Stone, C. *Classification and Regression Trees* Monterey. CA: Wadsworth and Brooks; 1984.
9. Candes E, Tao T. Near optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*. 52:5406–5425.2006;
10. Molecular Operating Environment (MOE), version 2010.10. Chemical Computing Group; Montreal, Quebec, Canada: 2010.
11. Clark P, Niblett T. The CN2 induction algorithm. *Mach Learn*. 3:261–283.Mar.1989

12. Cohen WW. Fast effective rule induction. In Proceedings of the Twelfth International Conference on Machine Learning Morgan Kaufmann. 1995:115–123.
13. T.U. Consortium. Ongoing and future developments at the universal protein resource. Nucleic Acids Research. 39(suppl 1):D214–D219.2011; [PubMed: 21051339]
14. Craven MW, Shavlik JW. Using sampling and queries to extract rules from trained neural networks. In Proceedings of the Eleventh International Conference on Machine Learning Morgan Kaufmann. 1994:37–45.
15. Craven, MW, Shavlik, JW. Extracting Tree-Structured representations of trained networks. In: Touretzky, DS, Mozer, MC, Hasselmo, ME, editors Advances in Neural Information Processing Systems. Vol. 8. The MIT Press; 1996. 24–30.
16. Diederich, J, editor Rule Extraction from Support Vector Machines, ser. Studies in Computational Intelligence. Vol. 80. Springer; 2008.
17. Friedman JH, Popescu BE. Predictive learning via rule ensembles. Annals Of Applied Statistics. 2(3):916–954.2008;
18. Fung G, Sathyakama S, Rao RB. Rule extraction from linear support vector machines. In KDD ACM Press. 2005:32–40.
19. Vahed, A, Omlin, C. Proceedings of the Rule Extraction from Trained Artificial Neural Networks Workshop, Society for the Study of Artificial Intelligence and Simulation of Behavior Workshop Series (AISB 96). University of Sussex; Brighton, UK: 1996. On the complexity of rule extraction from neural networks and network querying; 51–59.
20. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 286(5439):531–537.Oct; 1999 [PubMed: 10521349]
21. Gopalakrishnan V, Lustgarten JL, Visweswaran S, Cooper GF. Bayesian rule learning for biomedical data mining. Bioinformatics. 26(5):668–675.2010; [PubMed: 20080512]
22. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 3:1157–1182.Mar.2003
23. Huysmans J, Baesens B, Vanthienen J. Using rule extraction to improve the comprehensibility of predictive models. 2006
24. Jacobsson H. Rule extraction from recurrent neural networks: A taxonomy and review. Neural Comput. 17:1223–1263.Jun.2005
25. Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, Suzek B, Martin M, McGarvey P, Gasteiger E. Infrastructure for the life sciences: design and implementation of the uniprot website. BMC Bioinformatics. 10(1):136.2009; [PubMed: 19426475]
26. Jiang R, Yang H, Sun F, Chen T. Searching for interpretable rules for disease mutations: a simulated annealing bump hunting strategy. BMC Bioinformatics. 7(1):417.2006; [PubMed: 16984653]
27. Johansson U, Niklasson L. Rule extraction from trained neural networks using genetic programming. Joint 13th International Conference on Artificial Neural Networks and 10th International Conference on Neural Information Processing. 2003:13–16.
28. Kidson C. Two leukocyte enzymes with catalase activity in acute leukemia. Biochemical and Biophysical Research Communications. 9(1–2):138–141.1962; [PubMed: 14032492]
29. Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence. 97(1):273–324.1997;
30. Li J, Liu H, Downing JR, Yeoh AE-J, Wong L. Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (all) patients. Bioinformatics. 19(1):71–78.2003; [PubMed: 12499295]
31. Li J, Liu H, Ng S-K, Wong L. Discovery of significant rules for classifying cancer diagnosis data. Bioinformatics. 19(suppl 2):ii93–ii102.2003; [PubMed: 14534178]
32. Liu S, Chen Y, Wilkins D. Large margin classifiers and random forests for integrated biological prediction on mixed type data. Proc of the 7th Annual Biotechnology and Bioinformatics Symposium (BIOT). 2010:11–19.

33. Ma S, Huang J. Penalized feature selection and classification in bioinformatics. *Briefings in Bioinformatics*. 9(5):392–403.2008; [PubMed: 18562478]
34. Markowska-kaczmar U, Trelak W. Extraction of fuzzy rules from trained neural network using evolutionary algorithm. In *European Symposium on Artificial Neural Networks (ESANN)*. 2003:149–154.
35. Martens D, Baesens B, Van Gestel T. Decompositional rule extraction from support vector machines by active learning. *Knowledge and Data Engineering, IEEE Transactions*. 21(2):178–191.2009;
36. Mauri A, Consonni V, Pavan M, Todeschini R. Dragon software: An easy approach to molecular descriptor calculations. *Match Communications In Mathematical And In Computer Chemistry*. 56(2):237–248.2006;
37. Núñez H, Angulo C, Català A. Rule extraction from support vector machines. In *Proceedings of European Symposium on Artificial Neural Networks*. 2002:107–112.
38. Nutt CL, Mani DR, Betensky RA, Tamayo P, Cairncross JG, Ladd C, Pohl U, Hartmann C, McLaughlin ME, Batchelor TT, Black PM, von Deimling A, Pomeroy SL, Golub TR, Louis DN. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res*. 63(7):1602–1607.Apr; 2003 [PubMed: 12670911]
39. The MDR1 ‘substrate’ and ‘non-substrate’ P-gP data was generously provided by the National Institute of Mental Health’s Psychoactive Drug Screening Program, Contract #HHSN-271-2008-00025-C (NIMH PDSP).
40. Quinlan JR. Induction of decision trees. *Machine Learning*. 1(1):81–106.Mar; 1986
41. Quinlan, R. C4.5: programs for machine learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc; 1993.
42. Rabuñal JR, Dorado J, Pazos A, Pereira J, Rivero D. A new approach to the extraction of ANN rules and to their generalization capacity through gp. *Neural Comput*. 16:1483–1523.Jul.2004 [PubMed: 15165398]
43. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*. 23(19):2507–2517.2007; [PubMed: 17720704]
44. Schmitz GPJ, Aldrich C, Gouws FS. ANN-DT: an algorithm for extraction of decision trees from artificial neural networks. *IEEE Transactions on Neural Networks*. 10:1392–1401.1999; [PubMed: 18252640]
45. Schoch C, Kohlmann A, Schnittger S, Brors B, Dugas M, Mergenthaler S, Kern W, Hiddemann W, Eils R, Haferlach T. Acute myeloid leukemias with reciprocal rearrangements can be distinguished by specific gene expression profiles. *Proceedings of the National Academy of Sciences*. 99(15): 10008–10013.2002;
46. Setiono R, Leow WK, Zurada JM. Extraction of rules from artificial neural networks for nonlinear regression. *IEEE Transactions on Neural Networks*. 13:564–577.2002; [PubMed: 18244457]
47. Su Y, Murali TM, Pavlovic V, Schaffer M, Kasif S. Rankgene: identification of diagnostic genes based on expression data. *Bioinformatics*. 19:1578–1579.2003; [PubMed: 12912841]
48. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*. 58:267–288.1996;
49. Tickle AB, Andrews R, Golea M, Diederich J. The truth will come to light: directions and challenges in extracting the knowledge embedded within trained artificial neural networks. *Neural Networks, IEEE Transactions on*. 9(6):1057–1068.1998;
50. Towell GG, Shavlik JW. The extraction of refined rules from knowledge-based neural networks. *Machine Learning*. 1993:71–101.
51. Vahed A, Omlin C. Rule extraction from recurrent neural networks using a symbolic machine learning algorithm. In *6th International Conference on Neural Information Processing*. 1999:712–717.
52. Zhou Z-H, Jiang Y, Chen S-F. Extracting symbolic rules from trained neural network ensembles. *AI Communications*. 16:3–15.2003;
53. Vapnik, VN. *Statistical Learning Theory*. Wiley-Interscience; New York: Sep, 1998
54. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R,

- Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 415(6871):530–536.Jan; 2002 [PubMed: 11823860]
55. Wang Y. Cancer classification using loss of heterozygosity data derived from single-nucleotide polymorphism genotyping arrays. :5864–5867.2006
56. Wang Y, Tetko IV, Hall MA, Frank E, Facius A, Mayer KFX, Mewes HW. Gene selection from microarray data for cancer classification-a machine learning approach. *Comput Biol Chem*. 29:37–46.Feb.2005 [PubMed: 15680584]
57. Zhu, J, Rosset, S, Hastie, T, Tibshirani, R. *Neural Information Processing Systems*. MIT Press; 2003. 1-norm support vector machines; 16
58. Zou H. An improved 1-norm svm for simultaneous classification and variable selection. *Journal of Machine Learning Research - Proceedings Track*. 2:675–681.2007;

Fig. 1.
Binary encoding of a random forest.

Fig. 2.
Number of rules extracted on prospectr data.

Fig. 3.
Cross validation accuracy on prospectr data.

TABLE I

Some statistics of data sets for Cannabinoid (CB) receptor subtypes CB1 and CB2, and for P-glycoprotein (P-gP).

Data	training set size	test set size	number of classes	number of features
CB1 antagonist activity	255	107	2	306 or 1972
CB1 antagonist selectivity	149	65	3	306 or 1880
CB2 agonist activity	288	128	2	306 or 1914
CB2 agonist selectivity	135	60	3	306 or 1808
CB2 antagonist activity	102	40	2	298 or 1863
CB2 antagonist selectivity	108	42	2	298 or 1837
P-gP	465	117	2	2054

TABLE II

Number of rules extracted from different methods on Cannabinoid (CB) Receptor Subtypes CB1 And CB2 data sets with the smaller MOE feature sets.

Data	RF	CRF
CB1 antagonist activity	7199	28
CB1 antagonist selectivity	9511	43
CB2 agonist activity	11495	67
CB2 agonist selectivity	9432	88
CB2 antagonist activity	10836	24
CB2 antagonist selectivity	6709	25

TABLE III

Accuracy of the different methods on Cannabinoid (CB) Receptor Subtypes CB1 and CB2 data sets with the smaller MOE feature sets.

Data	RF	RIPPER	CRF
CB1 antagonist activity	95.3	91.6	94.4
CB1 antagonist selectivity	83.1	63.1	80.0
CB2 agonist activity	80.5	77.3	80.5
CB2 agonist selectivity	73.3	63.3	70.0
CB2 antagonist activity	90.0	75.0	75.0
CB2 antagonist selectivity	85.7	85.7	78.6

TABLE IV

Number of Rules Extracted from Different Methods on Cannabinoid (CB) Receptor Subtypes CB1 and CB2 Data Sets with larger DragonX feature sets

Data	RF	CRF
CB1 antagonist activity	6190	35
CB1 antagonist selectivity	7430	65
CB2 agonist activity	11403	58
CB2 agonist selectivity	16356	73
CB2 antagonist activity	14220	25
CB2 antagonist selectivity	12498	35

TABLE V

Accuracy of the different methods on Cannabinoid (CB) Receptor Subtypes CB1 and CB2 data sets with the larger DragonX feature sets.

Data	RF	RIPPER	CRF
CB1 antagonist activity	95.3	88.8	95.3
CB1 antagonist selectivity	80.0	64.6	72.3
CB2 agonist activity	78.1	74.2	79.7
CB2 agonist selectivity	71.6	65.0	68.3
CB2 antagonist activity	87.5	85.0	85.0
CB2 antagonist selectivity	85.7	83.3	85.7

TABLE VI

Sample descriptors in the top-ranked selected rule from Cannabinoid (CB) Receptor Subtypes CB1 and CB2 data sets.

Descriptors	Category	Description
T(O..O)	Topological descriptors	sum of topological distances between O..O
Mor02e	3D-MoRSE descriptors	3D-MoRSE - signal 02/weighted by atomic Sanderson electronegativities
BEHv1	Burden eigenvalue descriptors	highest eigenvalue n. 1 of Burden matrix/weighted by atomic van der Waals volumes
ESpm08d	Edge adjacency indices	Spectral moment 08 from edge adj. matrix weighted by dipole moments
RDF055u	RDF descriptors	Radial Distribution Function - 5.5/unweighted
Mor22u	3D-MoRSE descriptors	3D-MoRSE - signal 22/unweighted
G2u	WHIM descriptors	2nd component symmetry directional WHIM index/unweighted
GGI3	Topological charge indices	topological charge index of order 3
C-005	Atom-centred fragments	CH3X

TABLE VII

Results of different methods on P-glycoprotein (P-gP) data set.

Data	RF	RIPPER	CRF
P-gP	66.7	61.1	66.7

TABLE VIII

Summary of descriptors in top-ranked selected rule from P-glycoprotein (P-gP) data set.

Descriptors	Category	Description
SP11	Randic Molecular Profiles	shape profile no.11
EEig05x	Edge Adjacency Indices	Eigenvalue 05from edge adj. matrix weighted by edge degrees
QXXv	Geometrical descriptors	Qxx COMMA2 value/weighted by atomic van der Waals volumes
Mor13u	3D-MoRSE descriptors	3D-MoRSE - signal 13/unweighted
nHDon	Functional Group Counts	number of donor atoms for H-bonds (N and O)
AEige	Eigenvalue-based indices	Absolute eigenvalue sum from electronegativity weighted distance matrix
R3m+	GETAWAY descriptors	R maximal autocorrelation of lag 3/weighted by atomic masses
Mor10m	3D-MoRSE descriptors	3D-MoRSE - signal 10/weighted by atomic masses
L2m	WHIM descriptors	2nd component size directional WHIM index/weighted by atomic masses

TABLE IX

Number of rules extracted from different methods on microarray data sets.

Data	RF	RIPPER	CRF
Golub	4336	2	3
Colon	4522	2	1
Nutt	9314	4	33
Veer	7624	3	55

TABLE X

Accuracy of different methods on microarray data sets.

Data	RF	RIPPER	CRF
Golub	91.18	73.53	88.24
Colon	66.67	50.00	66.67
Nutt	80.00	80.00	80.00
Veer	71.43	37.50	71.43