Sinteza
Međunarodna naučna konferencija Univerziteta Singidunum

Impact of Internet on Business Activities in Serbia and Worldwide
Uticaj Interneta na poslovanje u Srbiji i svetu

DOI: 10.15308/SINTEZA-2014-318-323

MACHINE LEARNING OF HYBRID CLASSIFICATION MODELS FOR DECISION SUPPORT

Vladislav Miškovic
Singidunum University, Belgrade

Abstract:
Machine learning methods used for decision support must achieve (a) high accuracy of decisions they recommend, and (b) deep understanding of decisions, so decision makers could trust them. Methods for learning implicit, non-symbolic knowledge provide better predictive accuracy. Methods for learning explicit, symbolic knowledge produce more comprehensible models. Hybrid machine learning models combine strengths of both knowledge representation model types. In this paper we compare predictive accuracy and comprehensibility of explicit, implicit, and hybrid machine learning models for several standard medical diagnostics, electronic commerce, e-marketing and financial decision making problems. Their applicability in different environments - desktop, mobile and cloud computing is briefly analyzed. Machine learning methods from Weka and R/Revolution environments are used.

Key words:
machine learning,
classification,
hybrid models,
decision support,
predictive accuracy,
comprehensibility.

INTRODUCTION

Machine Learning algorithms are used in data mining applications to retrieve hidden information that may be used in decision-making [1].

There are various basic learning methods like rule-based learning, case-based reasoning, artificial neural networks and decision trees learning. Every method has its own advantages and disadvantages. There are a lot of hybrid machine learning methods which attempt to combine several different learning methods to bring out the best from all of them [2], [3]. One approach for increasing the most important generalization property, prediction accuracy on unseen examples, is the method of combined classifiers or ensembles [2].

Numerous machine learning methods and appropriate knowledge representation models can be used to support decision making. For example, classification and regression methods can be used for learning decision trees, rules, Bayes networks, artificial neural networks and support vector machines [1], [2]. Their applicability and performances are problem-dependent, and according to the Generalization Conservation Law [4] or the No Free Lunch Theorem [5], the best machine learning method which is the best for every problem does not exist.

Hybrid machine learning systems combine or integrate different machine learning (and decision-making) models. Since each machine learning method works differently and exploits a different part of problem (input) space, usually by using a different set of features, their combination or integration usually gives better performance than using each individual machine learning or decision-making model alone. Hybrid models can reduce individual limitations of basic models and can exploit their different generalization mechanisms.

Machine learning is based on data from different sources and with different properties. There are appropriate methods to learn from sparse data, sequentially accessible data (*data streams*) and Big Data [6], which must be processed using distributed processing methods [6], [7].

MACHINE LEARNING METHODS

Machine learning

Machine learning is simply defined as the process of estimating unknown dependencies or structures in a system using a limited number of observations [1]. Typical machine learning tasks are classification, regression and clustering.

Machine learning methods are rote learning, learning by being told, learning by analogy, and inductive learning, which includes methods of learning by examples and learning by experimentation and discovery [1], [8].

Formal definition of inductive learning is that it is the process of estimating an unknown function or (input, output) dependency or structure of a system S using a limited number of observations *x* [1]. A set of functions which can be learnt and an estimation method for its best approximation are predefined by selection of a basic algorithm *A* and some background knowledge about the system *S*.

Induction is performed on a set of empirical data which is commonly called a training set (or a data set). Problem domain model creation is based on background knowledge about the problem under consideration and often ends by specifying of a set of attributes or variables x_i, i=1..n. Some of these attributes are irrelevant or redundant, and deteriorate the performances of a majority of learning algorithms. Irrelevant and redundant attributes removal is performed by attribute/feature selection methods [1], [9].

In the context of decision support, machine learning of *classifications* is of particular importance. A system learns to classify new cases to predefined discrete problem classes. Classification is a special kind of regression, its goal being to predict a numeric quantity instead of a discrete one.

Machine learning of classifications performs an estimation of an unknown dependence between input (*data*) and output of the considered system (*classifications*) based on available examples of correct classification. Estimated mapping is used to predict future output of the observed system for future input values.

Learning classifications includes learning mathematical or logical expressions, decision trees, rules, decision tables, graphs, networks, hypersurfaces and other useful knowledge representations.

Machine learning of redundant knowledge or ensemble methods is based on repetition of the machine learning process, each time with different elements: a different partition of a learning set and/or attributes, a different learning algorithm or some combination of these elements.

The goal is to learn a combined classifier which is better than any of its elements. This is possible if basic elements are sufficiently accurate and mutually different enough. Such diversity of ensemble elements can be increased by generating an appropriate partition of attributes for every classifier.

Machine learning methods for learning classifications

1) Methods for learning comprehensible knowledge
   Methods for learning comprehensible, human readable knowledge are especially appropriate in building knowledge based decision support systems/expert systems. Well known method are decision trees (DT) and rule learning (RL).
   An important new method is the *Hoeffding Tree* or the Very Fast Decision Tree (VFDT), introduced for incremental machine learning from *data streams* [10]. It stores a data stream only once and after that updates the tree.
   The name is derived from the Hoeffding bound, which states with probability 1 – δ that the true mean of a random variable of range *R* will not differ from estimated mean more than

$$\varepsilon = \sqrt{\frac{R^2 \cdot \ln(1/\delta)}{2n}}$$

   where *n* is a number of independent examples. This bound is not dependent of the probability distribution generating the examples, but more examples are needed to reach the same ε and δ as with distribution-dependent bounds.
2) Methods for learning implicit knowledge
   Implicit or distributed knowledge is subjective, empirical, hard to formalize, and not understandable for humans. It can be represented in form of Bayes or neural networks, support vectors or using the similarity function and learning examples by itself.
   The most used machine learning methods of this type are k-nearest-neighbours (kNN), Bayes networks, artificial neural networks (ANN), and support vector machines (SVM).
   Support Vector Machines method (SVM) is a very successful method of machine learning from examples [11] which is based on mapping of learning examples from input space to a new high dimensional, potentially infinite dimensional feature space in which examples are linearly separable. The method then finds an optimal hyperplane

$$\langle \mathbf{w}, \Phi(x) \rangle + b = 0$$

   where **w** is a matrix of coefficients, $\Phi(x)$ is a mapping function, and *b* is a constant. This hypersurface separates learning examples with a maximal margin or distance to the nearest learning example [11], [12]. Support vectors are a small set of critical border examples of each class, best separated by this hyperplane. Construction of an optimal hyperplane is performed using iterative algorithm which minimizes the error estimation function:

$$\frac{1}{2}\mathbf{w}^{\mathbf{T}}\mathbf{w} + C\sum_{i=1}^{n}\xi_i$$

   with the constraints

$$y_i(\mathbf{w}^{\mathbf{T}}\Phi(x_i) + b) \geq 1 - \xi_i,\ i = 1,..N, \xi_i \geq 0,\ i = 1,..n$$

   where **w** is a vector of coefficients, *b* is a constant, ξ is a slack variable (tolerance of overlapping linear non-separable classes of examples), *n* is a number of learning examples and *C* is a regularization parameter.
   SVM method uses linear functions to create discrimination borders in a high dimensional space. Non-linear discriminant function in an input space is obtained using inverse transformation (*kernel trick*).

3) Redundant knowledge machine learning methods
Methods of learning and combining redundant classifiers or ensembles are one approach for increasing prediction accuracy models on unseen examples, which is the most important generalization property.
Example of a method that successfully uses only symbolic classifiers in an ensemble is the *Random Forests* [13], which simultaneously uses two sources for diversity of its elements: (1) resampling of learning data and (2) resampling the attribute set as part of the induction process. The only basic machine learning method used is a decision learning algorithm called CART [14]. In addition, the *Random Forests* method can provide an estimation of attributes importance [13].

Machine learning methods for learning hybrid models can use combined models or hybrid ensembles or both.

HYBRID MACHINE LEARNING MODELS AND METHODS

The supervised learning problem is to find an approximation to an unknown function given a set of previously labelled examples. Different methods explore different hypothesis spaces, use different search strategies and are appropriate for different types of problems [15].

In case of decision trees the divide-and-conquer strategy is used. It has the ability to split the space of attributes into subspaces, which can then be fitted with different functions. This is the basic idea behind well-known tree based algorithms like CART [14] and C4.5 [16].

For classification problems, the methods that explore multiple representations are multivariate trees [14], [17]. Decision nodes of this class of algorithms can contain tests based on a combination of several attributes. For classification problems, multivariate decisions usually appear in internal nodes. For regression problems, they appear in leaf nodes.

Related work

According to [2], [18], there are a lot of hybrid machine learning methods developed in the past:

- Model Trees – multivariate trees with linear or some other functional models at the leaves [19], [20], [21].
- Perceptron Trees – combination of a decision tree and a linear threshold unit [22].
- Decision trees and Naive Bayes hybrid – a regular univariate decision tree where leaves contain a naive Bayes classifier built from the examples that fall at that node [23], [24].
- Functional trees – an extension of multivariate and model trees. They use functions at inner nodes or at leaves of decision trees [18].
- Model Class Selection – a hybrid algorithm that combines, in a single tree, nodes that are univariate tests, or multivariate tests generated by linear machines or instance-based learners [17].
- Meta decision trees – decision trees where leaves predict which classifier should be used to obtain a prediction [25].
- Stacked generalization – hybrid ensembles which are constructed from different base learning methods [26].
- Hybrid Hoeffding Trees – several hybrid variants of the basic method using Naive Bayes, functions and ensemble methods [7].

Typical examples of hybrid machine learning methods

Typical hybrid machine learning methods available in Weka environment [9] are:

- Model Trees: LMT (Logistic model trees) [21];
- Decision trees and Naive Bayes: NBTree [24];
- Functional trees: FT [18];
- Stacking generalization: StackingC [26];
- Hybrid Hoeffding Trees: HT and variants in meta library for massive online analysis (MOA) [7].

EXPERIMENTS

This work investigates applicability of selected basic and hybrid machine learning methods to solve typical unstructured decision making problems. For their unbiased comparison, all machine learning experiments are performed without using any external feature selection method.

Methods

Selected standard and hybrid machine learning methods are compared, together with ensemble and hybrid ensemble methods:

- Standard methods: *C45* (*J48*), C45Rules (*PART*), Linear Discriminant Analysis (*LDA)*, support vector machines (*LibSVM)*, k-nearest neighbours (IB*k*);
- Hybrid methods: *Functional Tree* (FT), *NBTree*, *Logistic Model Tree* (LMT), *Hybrid Hoeffding Tree*;
- Ensemble methods: *Random Forests*;
- Hybrid ensemble methods (meta): *Stacking*.

Datasets

As benchmark problems, we used standard decision making problems from finance (*German Credit*), medical diagnostics using gene expressions (*Breast Cancer*), e-commerce like recommendations (*Red-White Wine Quality*), e-mail filtering (*Spambase*) and direct marketing (*Direct Marketing*).

Brief descriptions of those decision making problems used as benchmark examples are the following:

1. *German Credit* – a well-known problem of inductive learning of credit approval policy for banking loans [30].

2. *Breast Cancer* – a problem of disease diagnostics on the basis of genetic expressions. Tissue samples are taken from healthy and ill patients, processed and deposited on a suitable DNA microarray chip with thousands oligonucleotide points whose intensity (the expression) corresponds to the activity of single genes in tissue samples [27], [28], [28].
3. *Spambase* – learning to decide whether an incoming e-mail is spam or not in order to automate e-mail classification [30].
4. *Quality* – a problem of learning to rank red and white wines slightly adapted to decide the class of wine quality (1..10) [30].
5. *Direct Marketing* – learn to decide/predict if a client in a direct marketing campaign of a banking institution will subscribe to the product (bank term deposit) or not, in order to minimize the number of phone calls needed [30].

Properties of machine learning problems used in this work are shown in Table I.

Software Used

All the used machine learning methods are publicly available, mostly in Weka environment [9]. Some of them are briefly compared with equivalent methods in R/Revolution environment, packages *kernlab*, *e1071* and *MASS* [31], [32].

Experimental methods

We use ten-fold cross validation as the only method of classification accuracy estimation for all the performed experiments.

Table 1. Descriptions of decision/learning problems (datasets) used

Problem/Dataset	# Attributes	# Examples	# Classis	% Majority
German Credit (*German*)	20	1,000	2	70.0
Breast Cancer (*Gene Expr*)	22,215	175	2	66.3
RWWineRatings (*Quality*)	12	6,497	10	43.7
Spambase (*Spam*)	57	4,601	2	60.6
Direct Marketing (*Direct*)	16	4,521	2	88.5

RESULTS

The main goal of this contribution is identifying an appropriate machine learning method for decision support that produces accurate and understandable results.

Table 2. Descriptions of decision/learning problems (datasets) used

Method	Problem				
	German	*Gene Expr*	*Quality*	*Spam*	*Direct*
C45	72,8	63,4	58,6	93,0	89,6
C45Rules	72,4	57,1	60,0	94,2	89,7
LDA[a]	75,7	-	53,3	88,8	89,9
LibSVM	76,3	**70,9**	65,2	92,5	89,3
kNN	74,5	66,9	65,0	90,8	89,0
Random Forests	76,3	66,3	**69,8**	**95,4**	89,8
Functional Trees	75,5	68,6	60,2	93,4	**90,2**
NBTree	75,3	-	56,7	93,2	89,6
LMT	75,9	-	60,6	93,7	**90,2**
Stacking	**76,4**	70,3	68,0	95,0	89,9
Hoeffding Trees	75,6	64,6	48,6	81,6	87,9

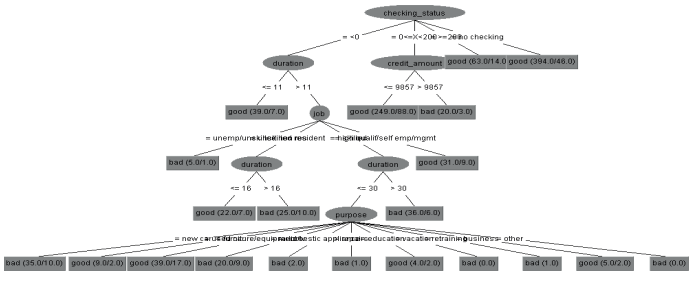
Classification using linear regression (no attribute selection)

In this work, we systematically estimate only the predictive accuracy of selected methods, Table II. The two ensemble methods considered, *Random Forests* and *Stacking*, are pointed by a different table cell colour.

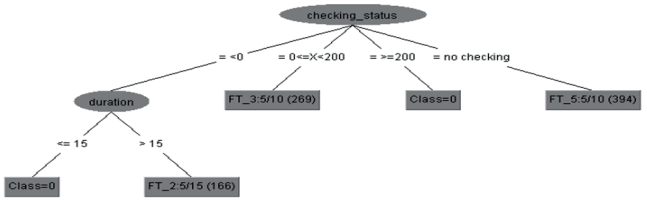
Understandability is estimated subjectively, by learning method type and size of resulting knowledge representation, Fig. 1.

Hybrid methods (FT, LMT) demonstrate small improvements in predictive accuracy only over standard comprehensible methods, as shown in Fig. 2.

Hybrid ensemble methods have predictive accuracy comparable to the standard ensemble method, Fig. 3.



(a) A whole C45 decision tree (J48)



(b) Top level of equivalent Functional tree (FT)

Fig. 1. Concept "German Credit Approval" described by (a) comprehensible decision tree, and (b) hybrid functional tree.

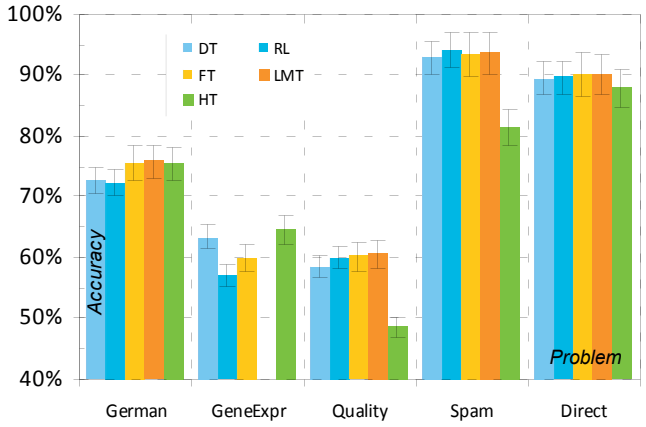


Fig. 2. Accuracy of basic comprehensible machine learning methods used compared to hybrid methods for five different problems.

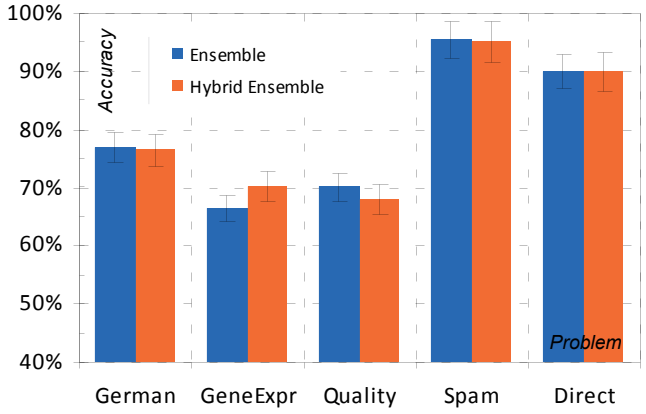


Fig. 3. Accuracy of ensemble and hybrid ensemble machine learning methods used for five different problems.

CONCLUSION

In this work, we have identified several promising machine learning methods suitable for learning knowledge useful for decision support. They produce both accurate and reasonably understandable results.

We systematically compared predictive accuracy of explicit, implicit and hybrid machine learning models for several standard medical diagnostics, electronic commerce, marketing, and financial decision making problems. Comprehensibility of new knowledge is subjectively evaluated.

Selected hybrid methods demonstrate improvement in predictive accuracy for five benchmark problems only with respect to comprehensible methods. The best method for every benchmark problem is different, but hybrid methods outperform standard comprehensible methods, and ensemble methods often outperform all other methods.

As expected, the Hoeffding trees and its variants, which are suitable for mobile computing, big data and/or data streams, demonstrate less accurate results for these batch problems which do not have huge numbers of learning examples.

REFERENCES

[1] Cherkassky V., Mulier F. M., Learning from Data: Concepts, Theory, and Methods, 2nd edition, John Wiley - IEEE Press, 2007.

[2] M. Wozniak, Hybrid Classifiers: Methods of Data, Knowledge, and Classifier Combination, Studies in Computational Intelligence, Vol. 519, Springer, 2014.

[3] P. Brazdil, C. Giraud-Carrier, C. Soares, R. Vilalta, Metalearning: Applications to Data Mining, Springer-Verlag, 2009.

[4] C. Schaffer, "A Conservation Law for Generalization Performance", in Proceedings of the Twelfth International Conference on Machine Learning, pp. 259-265, New Brunswick, NJ: Morgan Kaufmann, 1994.

[5] D. H. Wolpert, "The lack of a prior distinctions between learning algorithms and the existence of a priori distinctions between learning algorithms", Neural Computation, 8, 1341–1390,1391–1421, 1996.

[6] A. Rajaraman, J. Leskovec, J. D. Ullman, Mining of Massive Datasets, Cambridge University Press, 2011.

[7] A. Bifet, G. Holmes, R. Kirkby, B. Pfahringer, Data Stream Mining: A Practical Approach, Technical report, University of Waikato, May 2011

[8] R. Michalski, J. Carbonell, T. Mitchell (Eds.), Machine learning: An artifcial intelligence approach (Vol. I), San Francisco, CA: Morgan Kaufmann, 1983.

[9] I.H. Witten, E. Frank,. M.A. Hall, Data Mining: Practical machine Learning Tools and Techniques, 3rdEd, Elsevier Inc, 2011.

[10] G. Hulten, P. Domingos, "Mining High-Speed Data Streams", pp. 71-80, ACM Press, 2000.

[11] V.Vapnik, Statistical Learning Theory, John Wiley&Sons, 1998.

[12] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, 2004.

[13] L., Breiman "Random Forests", Machine Learning, 45, pp. 5-32, 2001

[14] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regresssion Trees, Wadsworth, Belmont, 1984.

[15] C. Perlich, F. Provost, J. Simonoff, "Tree induction vs. logistic regression: A learning-curve analysis", Journal of Machine Learning Research, 4, 211–255, 2003.

[16] R. Quinlan, C4.5: Programs for machine learning, Morgan Kaufmann Publishers, Inc., 1993.

[17] C.E. Brodley, P.E Utgoff, "Multivariate decision trees", Machine Learning, 19(1), 45–77, 1995.

[18] J. Gama, "Functional Trees", Machine Learning, 55, 219–250, Kluwer Academic Publishers, 2004.

[19] R. Quinlan, "Learning with continuous classes", In Adams,Sterling (Eds.), 5th Australian joint conference on artificial intelligence, pp. 343-348, World Scientific, 1992.

[20] I. Witten, E. Frank, Data mining: Practical machine learning tools and techniques with Java implementations, Morgan Kaufmann Publishers, 2000.

[21] N. Landwehr, M. Hall, E. Frank, "Logistic model trees", Machine Learning, 59(1/2), pp.161–205, 2005.

[22] P. E. Utgoff, "Perceptron trees: A case study in hybrid concept representations", In Proc. AAAI, pp. 601-606, 1988.

[23] I. Kononenko, B. Cestnik, I. Bratko, Assistant professional user's guide, Technical report, Jozef Stefan Institute, 1988.

[24] R. Kohavi, Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid, In: Second International Conference on Knoledge Discovery and Data Mining, 202-207, 1996

[25] Lj. Todorovski, S. Džeroski, "Combining Classifiers with Meta Decision Trees", Machine Learning, 50, 223–249, 2003.

[26] D. Wolpert, "Stacked generalization", Neural Networks, 5(2), 241-260,1992.

[27] Milosavljević M., Buturović LJ., "Analysis of One Class of Methods for Discriminative Selection of Gene Expressions" (in Serbian), Proc. 51. ETRAN Conference, Herceg Novi – Igalo, June 4-8, 2007.

[28] V. Miškovic, M. M. Milosavljević, "Application of Hybrid Symbolic Ensembles to Gene Expression Analysis", in Proceedings of 9th Symposium on Neural Network Applications in Electrical Engineering, p.95-98, Belgrade, September 2008.

[29] V. Miškovic, M. M. Milosavljević, "Application of Symbolic Inductive Learning Methods to Gene Expression Analyses", in Proceedings of 9th Symposium on Neural Network Applications in Electrical Engineering, p.99-102, Belgrade, September 2008.

[30] A. Frank, A. Asuncion, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science, 2010.

[31] The Comprehensive R Archive Network, http://cran.r-project.org/

[32] Revolution Analytics, http://www.revolutionanalytics.com