Boosted Decision Trees for Diagnosis Type of Hypertension

Michal Wozniak

Chair of Systems and Computer Networks, Wroclaw University of Technology
Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland
michal.wozniak@pwr.wroc.pl

Abstract. The inductive learning algorithms are the very attractive methods generating hierarchical classifiers. They generate hypothesis of the target concept on the base on the set of labeled examples. This paper presents some of the decision tree induction methods, boosting concept and their usefulness for diagnosis of the type of hypertension (essential hypertension and five type of secondary one: fibroplastic renal artery stenosis, atheromatous renal artery stenosis, Conn's syndrome, renal cystic disease and pheochromocystoma). The decision on the type of hypertension is made only on base on blood pressure, general information and basis biochemical data.

1 Introduction

Machine learning [1] is the attractive approach for building decision support systems. For this type of software, the key-role plays the quality of the knowledge base. In many cases we can find following problem:

- the experts can not formulate the rules for decision problem, because they might not have the knowledge needed to develop effective algorithms (e.g. human face recognition from images),
- we want to discover the rules in the large databases (data mining) e.g. to analyze outcomes of medical treatments from patient databases; this situation is typical for designing telemedical decision support system, which knowledge base is generated on the base on the large number of hospital databases,
- program has to dynamically adapt to changing conditions.

Those situations are typical for the medical knowledge acquisition also. For many cases the physician can not formulate the rules, which are used to make decision or set of rules given by expert is incomplete.

In the paper we present two type of decision tree induction algorithm and we discus if boosting methods can improve the quality of decision tree for the real medical problem.

The content of the work is as follows. Section 2 introduces idea of the inductive decision tree algorithms. In Section 3 we describe mathematical model of the hypertension's type. Next section presents results of the experimental investigations of the algorithms. Section 5 concludes the paper.

J.L. Oliveira et al. (Eds.): ISBMDA 2005, LNBI 3745, pp. 223–230, 2005.

2 Algorithms

2.1 Decision Tree Induction

The most of algorithm as C4.5 given by R. J. Quinlan [3] or ADTree (Alternative Decision Tree)[13] are based on the idea of "Top Down Induction of Decision Tree". Therefore let us present the main idea of it

```
Create a Root node for tree
IF all examples are positive
      THEN return the single node tree Root with label
yes and return.
IF all examples are negative
      THEN return the single node tree Root with label no
and return.
IF set of attributes is empty
      THEN return the single node tree Root with label =
      most common value of label in the set of examples
      and return
Choose "the best" attribute A from the set of attributes.
  FOR EACH possible value vi of attribute
      1. Add new tree branch bellow Root, corresponding
         to the test A=vi.
      2. Let Evi be the subset of set of examples that
         has value vi for A.
      3. IF Evi is empty
            THEN bellow this new branches add a leaf node
            with label = most common value of label in
            the set of examples
            ELSE below this new branch add new subtree
            and do this function recursive.
  END
  RETURN Root
```

The central choice in the TDIDT algorithm is selecting "the best" attribute (which attribute to test at each node in the tree). Family of algorithm based on ID3 method [4] (e.g. C4.5) uses the information gain (or its' modification gain ratio) that measures how well the given attribute separates the training examples according to the target classification. This measure based on the Shanon's entropy of learning set *S:*

$$Entropy(S) = \sum_{i=1}^{M} - p_i \log_2 p_i \tag{1}$$

where p_i is the proportin of *S* belonging to klas i $\left(i \in \boldsymbol{M}, \boldsymbol{M} = \{1, 2, ..., M\}\right)$.

The information gain of an attribute *A* relative to the collection of examples *S*, is defined as

$$Gain(S, A) = Entropy(S) - \sum_{c \in values(A)} \frac{|S_v|}{|S|} Entropy(S_v), \tag{2}$$

where $values(A)$ is the set of all possible values for attribute A and S_v is the subset of S for which $A = v$. The future implementations of decision tree induction algorithm use measure based on defined in (2) information gain (e.g. information ratio [6]).

2.2 Boosting

Boosting is general method of producing an accurate classifier on base of weak and unstable one[9-10]. The boosting often does not suffer from overfitting. AdaBoost is the most popular algorithm introduced in 1995 by Freund and Shapire [1]. Pseudocode of AdaBoost.M1 (one of the version of AdaBoost algorithm) is presented below[2]:

Input:

1. sequence of m examples $\{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$ with labels $y_i \in Y = \{1, ..., k\}$
2. weak learning algorithm **WeakLearn**
3. integer T specifying number of iterations

Initialize $D_1(i) = 1/m$ for all I

do for $t = 1, 2, ..., T$:

1. Call **WeakLearn**, providing it with the distribution D_t
2. Get back a hypothesis $h_t : X \to Y$.
3. Calculate the error of h_t: $\varepsilon_t = \sum_{i: h_t(x_i) \neq y_i} D_t(i)$.
4. If $\varepsilon_t > 1/2$, then set $T = t - 1$ and abort loop.
5. Set $\beta_t = \varepsilon_t / (1 - \varepsilon_t)$.
6. Update distribution D_t: $D_{t+1} = \frac{D_t}{Z_t} \times \begin{cases} \beta_t & if\ h_t(x_i) = y_i \\ 1 & otherwise \end{cases}$

where Z_t is a normalization constant (chosen so that D_{t+1} will be a distribution).

Output the final hypothesis: $h_{fin}(x) = \arg\max_{y \in Y} \sum_{t: h_x(x) = y} \log \frac{1}{\beta_t}$.

3 Model of Type of Hypertension (HT) Diagnosis

During the hypertension's therapy is very important to recognize state of patient and the correct treatment. The physician is responsible for deciding if the hypertension is of an essential or a secondary type (so called the first level diagnosis). The senior physicians from the Broussais Hospital of Hypertension Clinic and Wroclaw Medical Academy suggest 30% as an acceptable error rate for the first level diagnosis.

The presented project was developed together with Service d'Informatique Médicale from the University Paris VI. All data was getting from the medical database *ARTEMIS*, which contains the data of the patients with hypertension, whose have been treated in Hôpital Broussais in Paris.

The mathematical model was simplified. Hover the experts from the Broussais Hôspital, Wroclaw Medical Academy, regarded that stated problem of diagnosis as very useful.

It leads to the following classification of type of hypertension:

1. essential hypertension (abbreviation: essential),
2. fibroplastic renal artery stenosis (abbreviation: fibro),
3. atheromatous renal artery stenosis (abbreviation: athero),
4. Conn's syndrome (abbreviation: conn),
5. renal cystic disease (abbreviation: poly),
6. pheochromocystoma (abbreviation: pheo).

Although the set of symptoms necessary to correctly assess the existing HT is pretty wide, in practice for the diagnosis, results of 18 examinations (which came from general information about patient, blood pressure measurements and basis biochemical data) are used, whose are presented in table 1.

Table 1. Clinical features considered

No	Feature
1	Sex
2	body weight
3	High
4	Cigarette smoker
5	limb ache
6	Alcohol
7	Systolic blood pressure
8	Diastolic blood pressure
9	Maximal systolic blood pressure
10	Effusion
11	Artery stenosis
12	Heart failure
13	Palpitation
14	carotid or lumbar murmur
15	Serum creatinine
16	Serum potassium
17	Serum sodium
18	Uric acid

4 Experimental Investigation

All learning examples were getting from medical database *ARTEMIS*, which contains the data of 1425 patients with hypertension (912 with essential hypertension and the rest of them with secondary ones), whose have been treated in Hôpital Broussais.

We used WEKA systems [11] and our own software for experiments e.g. [15]. Quality of correct classification was estimated using 10 folds cross-validation tests.

4.1 Experiment A

The main goal of experiment was to find quality of recognition the C4.5 algorithm and its' boosted form. The obtained decision tree is shown in Fig.1.

The frequency of correct classification of this tree is 67,79% and the confusion matrix looks as follow

Table 2. Confusion matrix for decision tree

Real diagnosis						Recognized class
Athero	conn	essent	fibro	Pheo	Poly	
4	3	54	16	0	0	athero
0	44	92	11	0	0	Conn
2	25	878	7	0	0	Essent
4	5	64	40	0	0	Fibro
2	3	80	2	0	0	Pheo
0	2	81	6	0	0	Poly

We rejected the classifier because his quality did not satisfy expert. But we have to note that advantage of this tree is that the essential hypertension was recognized pretty good (96,26%).

We tried to improve quality of obtained classifier using boosting concept. Unfortunately new classifier had worse quality than original one (59,30%). The confusion matrix of the boosted C4.5 is presented in Tab.2

Table 3. Confusion matrix for boosted decision tree

Real diagnosis						Recognized class
Athero	conn	essent	fibro	pheo	Poly	
8	4	52	7	2	4	athero
2	22	106	10	4	3	Conn
19	47	790	34	12	10	Essent
3	6	86	9	8	1	Fibro
1	3	73	5	4	1	Pheo
1	4	70	0	2	12	Poly

4.2 Experiment B

Physician-experts did not accept classifiers obtained in Experiment A. After the discussion we simplified the problem once more. We were trying to construct classifiers which would point at essential type of hypertension or secondary one. We used two methods to obtain the classifiers:

1. Alternative Decision tree (ADTree),
2. C4.5 algorithm.

For each of classifier we check its' boosted form also. The results of tests is shown in Fig.2.

As we see the frequency of correct classification of ADTree algorithm is 79,16%. Unfortunately the quality of recognition the secondary hypertension is only 58,48%. We tried improve the quality of classifier by AdaBoost.M1 procedure and we obtained new classifier based on ADTree concept (we use 10 iterations), which frequency of correct classification grew to 83,30% and 72,90% of correct classified secondary type of hypertension. This results satisfied our experts.

Additionally we check the quality of C4.5 for the same dichotomy problem. We obtained the decision tree similar to tree in Fig.1 which quality is 74,74% and frequency of correct recognized secondary type of hypertension is 51,85%. The boosting procedure did not improve the average quality of C4.5 (72,42%) but strongly improved the recognition secondary type of hypertension (60,81%).

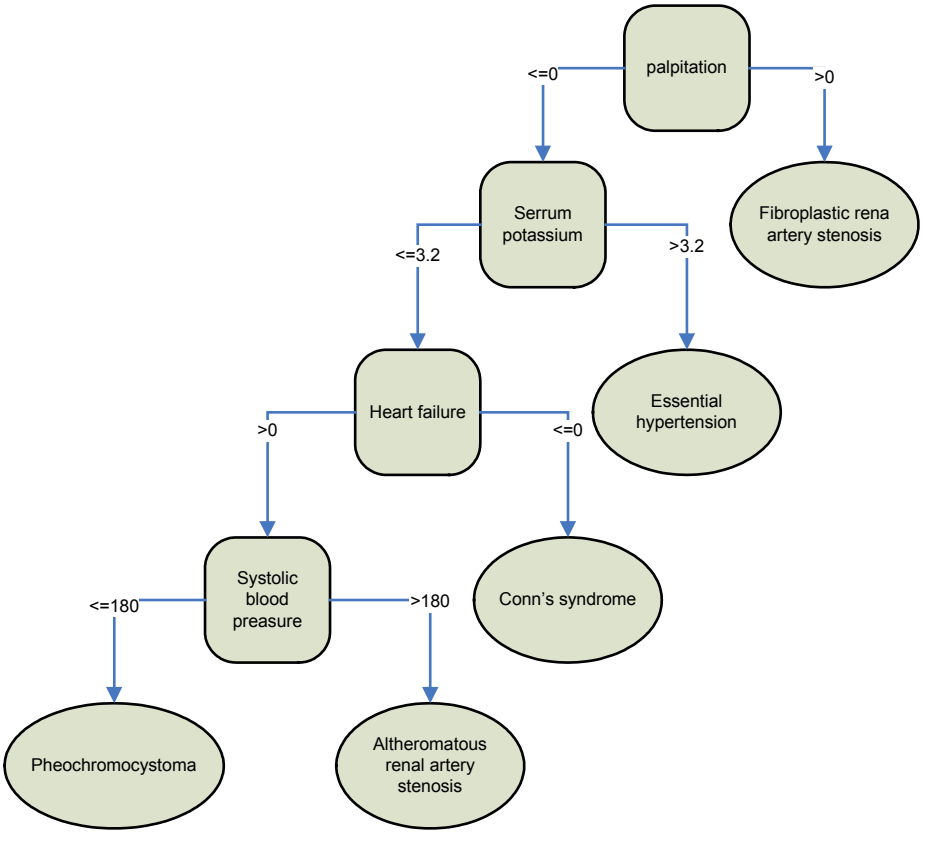

Fig. 1. Decision tree for hypertension diagnosis given by C4.5 algorithm

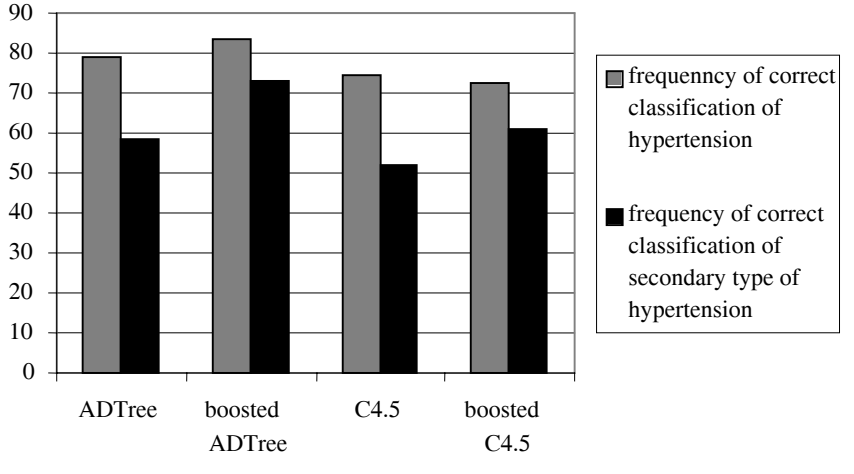

Fig. 2. Quality of recognition the essential and secondary type of hypertension

5 Discussion and Conclusion

The methods of inductive learning were presented. The classifiers generated by those algorithms were applied to the medical decision problem (recognition of the type of hypertension). The general conclusion is that *boosting* does not improve each classifier for each decision task. For the real decision problem we have to compare many classifiers. The similar observations were described by Quinlan in [7] where he did not observe quality improvements of boosted C4.5 for some of databases.

Most of obtained classifier (especially based on C4.5 method) did not satisfy experts. The best classifier (obtained for simplified decision problem) satisfied our expert. Now we want to construct classifier ensemble on the based on stacked classifier concept [5, 14] which idea is depicted in Fig.3. Obtained boosted ADTree classifier can be use for the first stage of recognition. Now we are working on the classifier of HT for patient with diagnosed secondary type of hypertension.

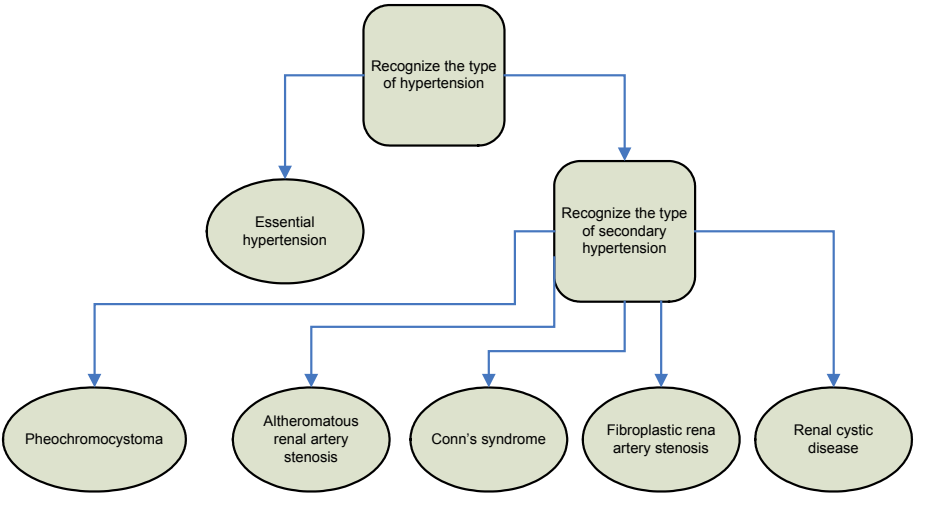

Fig. 3. Idea of stacked classifier of hypertension's type

The similar problem of computer-aided diagnosis of hypertension's type was described in [12] but authors used another mathematical model and implement Bayes decision rule. They obtained slightly better classifier than our, its' frequency of correct classification of secondary type of hypertension is about 85% (our 83,30%). Advantage of our proposition is simplified and cheaper model than presented in [12] (we use 18 features, authors of [12] 28 ones).

Advantages of the proposed methods make it attractive for a wide range of applications in medicine, which might significantly improve the quality of the care that the clinician can give to his patient.

This work is supported be The Polish State Committee for Scientific Research under the grant which is realizing in years 2005-2007.

References

1. Freund Y., Schapire R.E., A decision-theoretic generalization of on-line learning and application to boosting, *Journal of Computer and System Science,* 55(1),1997, pp. 119-139.
2. Freund Y., Schapire R.E., Experiments with a New Boosting Algorithm, *Proceedings of the International Conference on Machine Learning*, 1996, pp. 148-156.
3. Jain A.K., Duin P.W., Mao J., Statistical Pattern Recognition: A Review, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol 22., No. 1, January 2000, pp. 4-37.
4. Mitchell T., *Machine Learning*, McGraw Hill, 1997.
5. Opitz D., Maclin R., Popular Ensemble Methods: An Empirical Study, *Journal of Artificial Intelligence Research*, 11 (1999) 169-198
6. Quinlan J.R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993.
7. Quinlan J.R., Bagging, Boosting, and C4.5, *Proceedings of the 13th National Conference on Artificial Intelligence and Eighth Innovative Applications of Artificial Intelligence Conference, AAAI 96, IAAI 96*, Volume 1, Portland, Oregon, August 4-8, 1996, pp. 725-230.
8. Schapire R. E., The boosting approach to machine learning: An overview. *Proc. Of MSRI Workshop on Nonlinear Estimation and Classification*, Berkeley, CA, 2001.
9. Shapire R.E., The Strength of Weak Learnability, *Machine Learning*, No. 5, 1990, pp. 197-227.
10. Schapire R.E., A Brief Introduction to Boosting, *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, 1999.
11. Witten I.H., Frank E., Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Pub., 2000.
12. Blinowska A., Chatellier G., Bernier J., Lavril M., Bayesian Statistics as Applied to Hypertension Diagnosis, *IEEE Transaction on Biomedical Engineering*, vol. 38, n0. 7, July 1991, pp. 699-706.
13. Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer Verlag, New York 2001.
14. Puchala E., A Bayes Algorithm for the Multitask Pattern Recognition Problem – direct and decomposed approach, *Lecture Notes in Computer Science*, vol. 3046, 2004, pp. 39-45.
15. Koszalka L., Skworcow P., Experimentation system for efficient job performing in veterinary medicine area, *Lecture Notes in Computer Science*, vol. 3483, 2005, pp. 692-701.