Journal Of Turgut Ozal Medical Center
www.jtomc.org

Boosting Tree as a Stronger Approach in Classification: An Application of Carpal Tunnel Syndrome+

Handan Ankaralı[1], Gülhan Örekici Temel[2], Bahar Taşdelen[2], Aynur Özge[3]

[1]Düzce University Medical Faculty, Department of Biostatistics and Medical Informatics, Düzce, Turkey
[2]Mersin University Medical Faculty, Department of Biostatistics and Medical Informatics, Mersin, Turkey
[3]Mersin University Medical Faculty, Department of Neurology, Mersin, Turkey

Abstract

Aim: The Boosting Tree, one of the most successful combining methods. The principal aim of these combining algorithms is to obtain strong classifier with small estimation error from the combination of weak classifiers.

Material and Methods: We used boosting method to classify patients with Carpal Tunnel Syndrome. The individuals, who applied to Mersin University's Medical School's Neurology Main Scientific Branch's Electrophysiology Laboratory between the years of 2006 and 2010, with a pre-diagnosis of Carpal Tunnel Syndrome (CTS) were included in the study. Boosting Tree application was conducted in Statistica 7.0 software package.

Results: General success of the model in accurate classification according to the test data was found as 87.67%. Sensitivity and specificity of the latest model, when the test data were used, were calculated respectively as 85.65% and 92.36% .

Conclusion: The model can be used in CTS diagnosis as a successful method.

Key Words: Classification; Boosting Tree; Weak Classifiers.

This article presented at XIII. National Biostatistics Congress on 12-14 September 2011, Ankara, Kızılcahamam.

Sınıflamada Daha Güçlü Bir Yaklaşım Olan Boosting Ağacı: Karpal Tunel Sendromu Uygulaması

Özet

Amaç: Boosting ağaç yöntemi topluluk birleştirme yöntemlerinden en başarılı olanıdır. Birleştirme algoritmalarının temel amacı, zayıf sınıflayıcıların kombinasyonundan tahmin hatası düşük güçlü sınıflayıcılar oluşturmaktır.

Gereç ve Yöntemler: Bu çalışmada Karpal Tunel Sendromu vakaları boosting metodunu kullanılarak sınıflanmıştır. Mersin Üniversitesi Tıp Fakültesi Nöroloji Anabilim Dalının Elektrofizyoloji Laboratuvarına 2006-2010 tarihleri arasında Karpal Tünel Sendromu (KTS) ön tanısı ile başvuru yapan bireyler çalışmaya alınmıştır. Boosting Tree uygulaması Statistica 7.0 paket programında yapılmıştır.

Bulgular: Test verisi kullanıldığında ise modelin genel doğru sınıflama başarısı %87.67 olarak hesaplanmıştır. Test verisi kullanıldığında son modelin sensitivite ve spesifitesi ise sırasıyla %85.65 ve %92.36 olarak hesaplanmıştır.

Sonuç: Kullanılan modelin KTS tanısının konulmasında başarılı bir yöntem olarak kullanılabilir.

Anahtar Kelimeler: Sınıflama; Boosting Ağacı; Zayıf Sınıflayıcılar.

+12-14 Eylül 2011, Ankara-Kızılcahamam XIII. Ulusal Biyoistatistik Kongresi'nde sunulmuştur.

Orijinal Makale/Original Article

Received: 28.05.2012, **Accepted:** 01.06.2012

Corresponding Author:
Dr. Gülhan OREKİCİ TEMEL
Mersin Üniversitesi Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişimi Anabilim Dalı. Mersin, TÜRKİYE
Tel: +90 324 3610684-1028
e-mail:gulhan_orekici@hotmail.com

For citing/Atıf için:
Ankaralı H, Temel Orekici G, Taşdelen B, Özge A. Boosting tree as a stronger approach in classification: An application of carpal tunnel syndrome. J Turgut Ozal Med Cent 2012;19(4):228-33.
DOI:10.7247/jtomc.19.4.5

Introduction

In medical applications, the choice of statistical methods for diagnosis of a given syndrome is an important topic. In parallel to developments on bioinformatics techniques, classification methods based on decision trees have been frequently used for more reliable diagnosis. In addition to their reliability, the accuracy of clinical diagnosis is especially important because after true diagnosis developing a successful treatment plan is also critical. Recently, according to empirical comparisons of these statistical methods in different datasets, the weakness of them have been discussed and combining algorithms such as bagging and boosting have been used to improve classification performance (1). The principal aim of these combining algorithms is to obtain strong classifier with small estimation error from the combination of weak classifiers (2).

Decision trees are weak (base) classifiers which have been used frequently in medical diagnosis. Bagging, bootstrap aggregating, combines a large number of classifiers with re-sampling. Boosting combines re-weighted weak classifiers linearly to find strong classifier. According to literature, boosting produces even better results than bagging (3).

In this paper, we used boosting method to classify patients with Carpal Tunnel Syndrome (CTS) and investigate literature, CTS is the most commonly seen nerve entrapment syndrome and its diagnostic accuracy measures must be carefully interpreted (4).

Boosting Method

The Boosting, one of the most successful combining methods, was proposed by Schapire (5). The most popular algorithm AdaBoost was introduced by Freund and Schapire in 1995 and also extended to multi-class problems (6-8). This algorithm was called as Real AdaBoost for two-class dependent variable and AdaBoost.MH for more than two classes (9). The weak classifiers used in AdaBoost algorithm are single-split classification tress (10). In Boosting, a sequence of trees is obtained reweighting data after each classification tree. In each stage of boosting the weight of wrongly classified patients are increased while the weight of correctly classified patients is decreased (11).

Figure 1 shows a schematic of the AdaBoost procedure (12). The process of averaging weighted classifiers not only reduces the fitting error rate but also protects against overfitting (13). The systematic process of AdaBoost algorithm starts with uniform distribution of weights over training samples of patients. Using a classifier f(x) and confidence index, each case is classified initially. Increasing the weights on misclassified patients, each case are re-classified. The process is repeated until convergence a sign function is used for final decision.

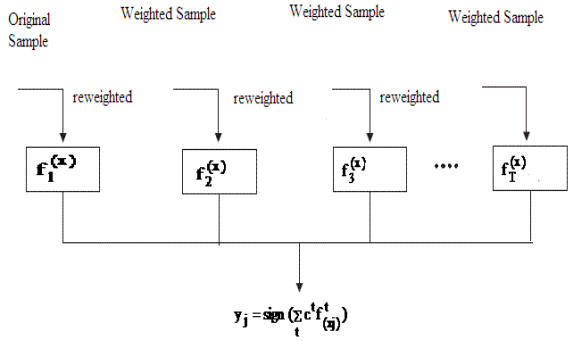

Figure 1. The schema of the AdaBoost procedure

Advantages of Method

Boosting method can be used for both categorical and continuous dependent variables. There is no limitation about the distribution of independent variables. They can be continuous, categorical or mixed type of distributions. In traditional classification models, the data is separated to training and test samples in certain proportions (%70: %30 or %60: %40) respectively.

First, the model is fitted to training sample in learning stage. Then, the validation of model is tested in remaining sample. Similarly, when Boosting tree is building, all data is separated to subgroups randomly. Because Boosting is a bootstrap-based method, all observations are being moved by chance and they are used in modeling. Trees built by this way resist over fitting, since the boosting both reduces the training classification

error and maximizes the classification margin separating the two classes (14,15).

Disadvantages of Method

According to simulation studies, the performance of boosting method is affected from small sample size and the number iteration. The number of iterations needed should be large as possible to resist over fitting (16).

Discrete AdaBoost Algorithm

Let training set with n observations (i=1,2,…,n) for p independent variables is given

$$D = (x_1, y_1),(x_2, y_2),...,(x_n, y_n)\} \quad (1)$$

Where $x_i = x_{1i},..,x_{pi})$ and $y \in + ,- \}$ (2)

AdaBoost algorithm for such sample includes following steps:

Step 1. Each subject in the training set is weighted equally.

$$(w_i^1 = \frac{1}{N}) \quad (i = ,2,..., N) \quad (3)$$

Step 2. For each iteration $(t = ,2,..., T)$

a.A data set with n subject is resampled with Bootstrap technique. Sampling probability of high weighted subject is more than the others.

b. A classifier f(x) is obtained using CART technique.

c. An indicator function is described to calculate classification error rate of f(x). The function for each iteration is given in following way. If a sample is misclassified 1, otherwise zero

$$err_i^t = \begin{cases} I(y(i)) = f(x_i) = 0 \\ I(y(i)) \neq f(x_i) = 1 \end{cases} \quad (4)$$

When a subject is wrong classified $err_i^t = $, otherwise $err_i^t =)$.

d. After the calculation of err_i^t for each iteration, weighted sum of all training set errors and confidence index (c^t) for f(x) classifier are calculated.

$$err^t = \sum_i (w_i^t err_i^t) \quad (5)$$

$$c^t = og(\frac{(1 - rr^t)}{err^t}) \quad (6)$$

The lower the weighted errors are the higher confidence index will be. e. All training sets are reweighted provided that

$$\sum_i v_i^{t+} = .$$

$$w_i^{t+} = v_i^t \exp(c^t err_i^t) \quad i = ,2,..., N \quad (7)$$

f. If $err_i^t \leq).5$ and t<T (t=t+1), steps (a)-(f) are repeated, otherwise iteration is stopped

Step 3. The performance of discrete AdaBoost algorithm is calculated using a test set. The final estimation for a sample in test set is combination of estimations from T classifiers.

$$y_j = sign(\sum_t c^t f_{(xj)}^t) \quad (8)$$

Where sign function is used to estimate dependent variable.

$$sign(*) = \quad (9)$$

Material and Methods

The individuals, who applied to Mersin University's Medical School's Neurology Main Scientific Branch's Electrophysiology Laboratory between the years of 2006 and 2010, with a pre-diagnosis of Carpal Tunnel Syndrome (CTS) were included in the study. Data set consists of 4076 incidences in total. 2517 (83.7%) of 3011 individuals with CTS taking place in the dataset were female patients and 491 (16.3%) of it were male patients. 868 (81.6%) of 1065 individuals in the control group were female and 196 (18.4%) of it were male. 3078 incidences of 4076 in total (2314 CTS + 764 healthy) were selected as learning data while the rest 998 incidences (697 CTS + 301 healthy) were selected as test data. The conducted electrophysiological measurements are independent variables. The condition of being a patient with CTS or not was considered as the dependent variable. Boosting Tree application was conducted in Statistica 7.0 software package.

Results

At the end of the conducted analysis, the model's success in accurate classification and significance rates of the variables were considered. Error rates of the subsequent trees, which were created according to the number of trees, are given on Figure 2

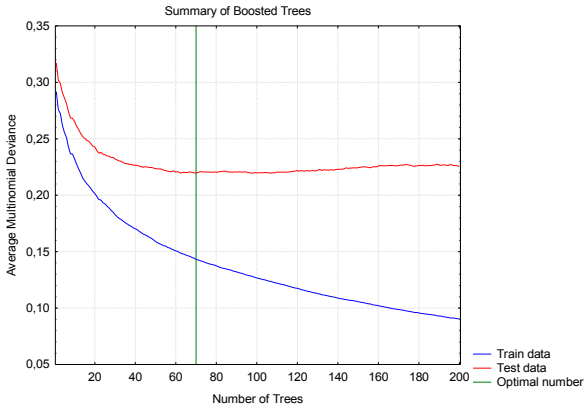

Figure 2. Variation in error rate according to number of trees

It is seen that as the number of trees increases, error rates in the learning data decrease. Considering learning and test data together, the optimum number of trees was found as 70. Classification tables according to the results from 70 trees created by using boosting algorithm based on the learning and test datasets are given on Tables 1 and 2, respectively.

Table 1. Classification table for learning data.

Learning Data		**Predicted**		**Total**
		CTS	Control	
Observed	CTS	2100	214	2314
	Control	30	734	764
Total		2130	948	3078

The results on Table 2 were obtained by testing the results, which had been obtained from the learning set, on a test set relating to the classification success. Classification successes for the created model are given on Table 3.

Table 2. Classification table for test data.

Test Data		**Predicted**		**Total**
		CTS	Control	
Observed	CTS	597	100	697
	Control	23	278	301
Total		620	378	998

Table 3. Classification successes for the learning and test data.

	Sensitivity (%) [Confidence Interval]	**Specificity (%) [Confidence Interval]**	**Accuracy (%)**
Learning Data	90.75 [89.50-91.90]	96.07 [94.44-97.34]	92.07
Test Data	85.65 [82.83-88.17]	92.36 [98.75-95.09]	87.67

General success of the model in accurate classification according to the learning data was found as 92.07%. Sensitivity and specificity of the latest model, when the learning data were used, were calculated as 90.75% and 96.07% respectively.

General success of the model in accurate classification according to the test data was found as 87.67%. Sensitivity and specificity of the latest model, when the test data were used, were calculated as 85.65% and 92.36% respectively.

Graphics relating to classification tables based on the learning and test datasets are given on Figures 3 and 4.

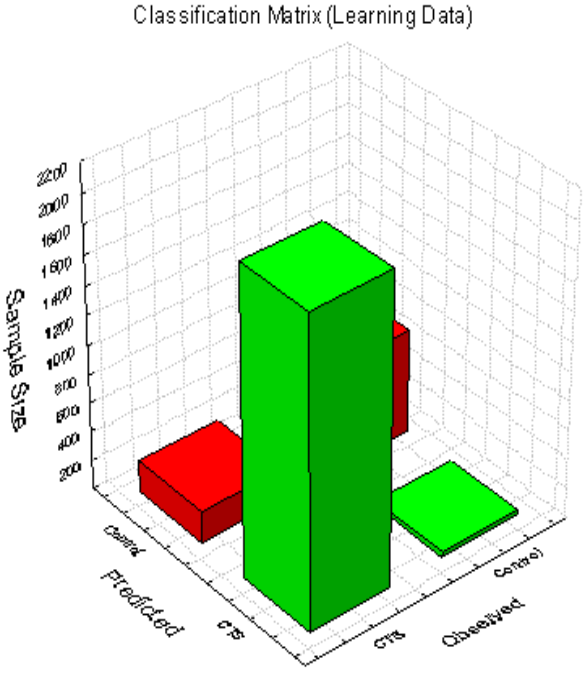

Figure 3. Classification matrix graphic belonging to the learning set

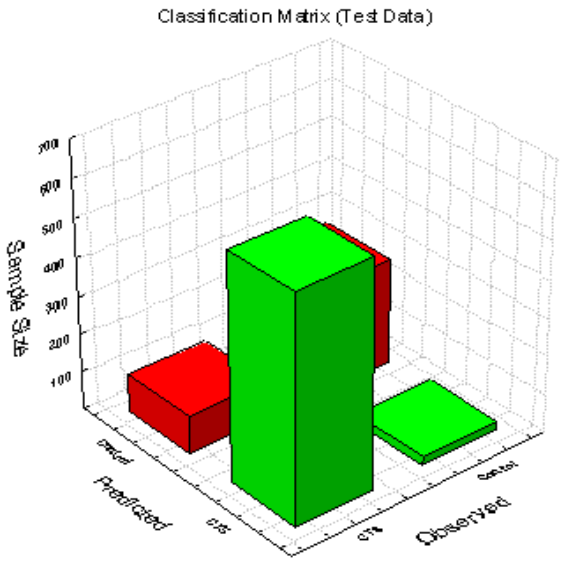

Figure 4. Classification matrix graphic belonging to the test set

Discussion

In this study, the boosting tree method, which is one of the population methods, was presented. The method is one of the population methods, which combine weak classifiers into a strong classifier. The method was applied to medical sciences through a dataset supplied by Neurology Main Scientific Branch. Carpal Tunnel Symptom is a diagnosis, whose gold standard is available.

A new prediction was made by combining the predictions made by the trees, which are independent from each other, in creating the boosting tree. Model cannot be seen with a single tree as it is done in the methods of CART (17). Contributions of independent variables to formation of the model and the model's diagnostic accuracy sizes can be calculated at the end of the model. General success of the model in accurate classification according to the learning data was found as 92.07%. Sensitivity and specificity of the latest model, when the learning data were used, were calculated as 90.75% and 96.07% respectively. General success of the model in accurate classification according to the test data was found as 87.67%. Sensitivity and specificity of the latest model, when the test data were used, were calculated as 85.65% and 92.36% respectively. High sensitivity and specificity values for the model indicate that the results produced by the learning set are valid for the other data also; in other words, the model can be used in CTS diagnosis as a successful method.

Conclusion

Being able to use the entire dataset in creating model is important in this method. Furthermore, because it is possible to construct sample datasets, in a dataset in any size, in various amounts and sizes through the re-sampling by random sampling with replacement (bootstrap) from the original dataset, data, as much as possible, can be produced from the existing dataset.

References

1. Skurichina M, Duin R. Bagging, boosting and the random subspace method for linear classifiers. Pattern Anal Applic 2002;5:121-35.
2. Rodriguez JJ, Maudes J. Boosting recombined weak classifiers. Pattern Recogn Lett 2008;29: 1049-59.
3. Opitz D, Maclin R. Popular ensemble methods: an empirical study. J Artif Intell Res 1999;11:169-98.
4. D'Arcy CA, McGee S. The rational clinical examination. Does this patient have Carpal Tunnel Syndrome? JAMA 2000;283:3110-7.
5. Schapire RE. The strength of weak learnability. Mach Learn 1990;5:197-227.
6. Freund Y, Schapire RE. A decision-theoretic generalization of online learning and an application to boosting. computational learning theory: Second European Conference, EuroCOLT'95, 1995; p: 23-37, Springer-Verlag.
7. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 1997;55:119-39.
8. Schapire RE, Singer Y. Improved boosting algorithms using confidence-rated predictions. Mach Learn 1999;37:297-336.
9. Freund Y, Schapire RE. Experiments with A new boosting algorithm. Mach Learn 1996; Proceedings of the Thirteenth International Conference.
10. Zhang MH, Xu QS, Daeyaert F, Lewi PJ, Massart DL. Application of boosting to classification problems in chemometrics. Analytica Chimica Acta 2005;544: 167-76.
11. Death G. Boosted trees for ecological modelling and prediction. Ecology 2007;88:243-51.
12. He, P, Xu CJ, Liang YZ, Fang KT. Improving the classification accuracy in chemistry via boosting Technique**.** Chemometr Intell Lab 2004;70: 39-46.
13. Freund Y, Mansour Y, Schapire RE. Why averaging classifiers can protect against overfitting. Proceedings of the eighth international 2001; Workshop on Artificial Intelligence and Statistics.

14. Cherkassky V, Mulier FM. Learning from data: concepts, theory, and methods. 2007, 2nd Ed, Canada: John Wiley & Sons.
15. Schapire RE, Freund Y, Bartlett P, Lee WS. Boosting the margin: A new explanation for the effectivness of voting methods. Ann Statist 1998;26:1651-86.
16. Mease D, Wyner A. Evidence contrary to the statistical view of boosting. J Mach Learn Res 2008;9:131-56.
17. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. 1993; New York: Chapman&Hall.