AUSTRIAN JOURNAL OF STATISTICS
Volume 35 (2006), Number 4, ???–???

On Speed of Stochastic CART Model Search

Márton Ispány and Ilona Krasznahorkay
University of Debrecen, Hungary

Abstract: Decision trees have proved to be commonly used nonlinear tools for supervised learning. This technique is a way to divide the space of the predictor variables into bricks in order to achieve as homogeneous partitions as possible. We improved the CART method proposed by Breiman et al. (1984) using a stochastic search, first suggested by Chipman et al. (1998) in the Bayesian framework. In this paper estimates are given for the rate of convergence and the mixing time of the MCMC method defined on decision trees.

Zusammenfassung: Entscheidungsbäume haben sich als häufig gebrauchtes, nichtlineares Werkzeug des beaufsichtigten Lernens erwiesen. Diese Technik ist ein Verfahren um den Raum der Prädiktoren zu zerteilen, so dass wir dadurch eine möglichst homogene Teilung erreichen. Wir haben die von Breiman et al. (1984) vorgestellte CART Methode mit einer stochastischen Suche verbessert, die erstmals bei Chipman et al. (1998) im Bayes'sche Rahmen empfohlen wurde. In dieser Arbeit werden Schätzer für Konvergenzrate und Mischzeit der MCMC Methode durch die Entscheidungsbäume gegeben.

Keywords: Decision Trees, MCMC Methods, Canonical Paths.

1 Introduction

Decision trees are used successfully in many diverse areas such as character recognition, medical diagnosis, expert systems, credit scoring, and fraud detection. They can be used for data exploration in such supervised data mining problems as description, classification and generalization. The decision tree is a multistage approach, a hierarchical, sequential structure that recursively partitions the data. Overviews of work on decision trees can be found in Murthy (1998) and Safavian and Landgrebe (1991).

In this paper we consider the classification problem of supervised learning. Namely let $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ be the vector of predictor variables (called feature vector in pattern recognition) in space $\mathcal{X} \subseteq \mathbb{R}^p$, and let Y be the categorical target variable in space $\mathcal{Y}$. For the sake of simplicity we suppose that Y is binary, i.e. $\mathcal{Y} = \{0, 1\}$. The goal of any classification scheme in general is to estimate Y based on the observations of $\boldsymbol{X}$. This theory requires a loss function for quantifying errors in prediction. The most common and convenient function is the squared error loss: $(Y - f(\boldsymbol{X})^2$. This leads us to the criterion of nonlinear least squares:

$$\mathbb{E}(Y - f(\boldsymbol{X}))^2 \to \min\,, \qquad f : \mathcal{X} \to \mathbb{R}\,.$$

The solution of this problem is the conditional expectation, which coincides with the conditional probability in this case

$$p(\boldsymbol{x}) := \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x})\,, \qquad \boldsymbol{x} \in \mathcal{X}\,.$$

The function p is known as the regression function. In general the values of p do not necessarily belong to the space $\mathcal{Y}$. Thus, in order to get usable decision rule, we approximate the regression function by a classifier. A classifier or decision rule d is a function that maps $\mathcal{X}$ into $\mathcal{Y}$. Many types of decision rules are known, e.g. nearest neighbor, neural network, support vector machine and decision tree, see the monograph of Hastie et al. (2001).

The emphasis of this paper is on decision trees focusing on the goodness of specific algorithms. We define and investigate stochastic algorithms as Markov chains on decision trees. We measure the "goodness" of these algorithms by their mixing time and apply canonical paths method developed by Jerrum and Sinclair (1996), Jerrum (1998) for estimating mixing time, see the recent monograph of Jerrum (2003). The main result of the paper is that the Metropolis-Hastings chains defined on the set of decision trees are rapid mixing, i.e. their mixing time can be bounded by a polynomial of the number of leaves of the tree.

The organization of the paper is as follows. In Section 2 we will present a short overview on decision trees. In Section 3 a brief summary is given on MCMC methods and two Metropolis-Hastings algorithms are defined on the set of decision trees. In Section 4 some basic results are introduced on mixing time of finite Markov chains. In Section 5 we state Theorem 2 and 3 on the mixing time of the defined Markov chains, applying canonical paths introduced in Section 4. The paper is finished with a short discussion.

2 Decision Trees

A binary tree $T = (V, E)$, where V is a finite non-empty set of vertices, and E is a set of edges, is an ordered tree such that each internal node has two sons, a left and a right one. Let $\ell(T)$ and $i(T)$ denote the set of leaves (terminal nodes) and internal nodes of the tree T, respectively. The decision tree is a hierarchical partition of the predictor space $\mathcal{X}$ along the predictor variables.

Definition 1. *By a decision tree D we mean a triple (T, f, s), where T is a binary tree, and $f : i(T) \to \{1, 2, \ldots, p\}$ and $s : i(T) \to \mathbb{R}$ are functions.*

If $v \in V$ is an internal node, then $f(v)$ and $s(v)$ indicate that we cut the node v in the $f(v)$th predictor variable at the value $s(v)$. Then the points in node v satisfying the relation $x_{f(v)} \leq s(v)$ belong to the left son of v, while the points fulfilling the relation $x_{f(v)} > s(v)$ belong to the right son of v. The index $f(v)$ is called the label of v, the real number $s(v)$ is the splitting value at v (see Safavian and Landgrebe, 1991 for details).

There are various heuristic methods for the construction of decision trees. They can be divided into four categories: bottom-up, top-down, hybrid and growing-pruning. In this paper we develop a stochastic growing-pruning algorithm by improving the CART (Classification and Regression Trees) algorithm introduced by Breiman et al. (1984).

Let $(v_0, v_1), (v_1, v_2), \ldots, (v_{k-1}, v_k)$, where $v_0, v_1, \ldots, v_k \in V$, be the path from the root v_0 to $v_k = v$. Let the subset $\mathcal{X}_v \subseteq \mathcal{X}$ consist of the points $\boldsymbol{x} \in \mathcal{X}$ which satisfy the constraint $x_{f(v_i)} \leq s(v_i)$ if v_{i+1} is the left son of v_i, and the constraint $x_{f(v_i)} > s(v_i)$ if v_{i+1} is the right son of v_i, respectively for $i = 0, 1, \ldots, k-1$. This means that the borders of the partition $\{\mathcal{X}_v, v \in \ell(T)\}$ are hyper-planes of the field $\mathcal{X}$. The decision tree classifier

d introduced by D is a map $\mathcal{X} \to \{0,1\}$ such that $d(\mathcal{X}_v) = 0$ or 1 for all $v \in \ell(T)$, and $d(\mathcal{X}_v) \neq d(\mathcal{X}_{v'})$ if v and v' have common father. In principle, d is not uniquely defined by D, but if we use for example majority voting, then the correspondence will be one-to-one.

The heart of the CART method is defining a risk function with which we can measure the goodness of a decision tree. Let $\varphi : [0,1] \to [0,1/2]$ be a non-negative, concave and symmetric loss function. Some well known examples of φ are

Misclassification: $\varphi(p) = -|p - 0.5| + 0.5\,,$

Gini-index: $\varphi(p) = 2p(1-p)\,,$

Entropy: $\varphi(p) = [-p\log p - (1-p)]\log(1-p)/(2\log 2)\,.$

First we define the theoretical risk function. Let $\lambda : \mathcal{X} \to \mathbb{R}_+$ be the generalized density function of the predictor variables $\boldsymbol{X}$ with respect to a σ-finite measure μ on $\mathbb{R}^p$. (In general μ is a mixed product of the counting measure and Lebesgue measure.) Introduce

$$P(\mathcal{X}_\nu) := \int_{\mathcal{X}_\nu} \lambda(\boldsymbol{x}) d\boldsymbol{x}\,, \qquad p_\nu := \frac{1}{P(\mathcal{X}_\nu)} \int_{\mathcal{X}_\nu} p(\boldsymbol{x})\lambda(\boldsymbol{x}) d\boldsymbol{x}\,,$$

for all $\nu \in \ell(T)$, where $\{\mathcal{X}_\nu, \nu \in \ell(T)\}$ is the partition induced by the decision tree (T, f, s).

Definition 2. *Let $D = (T, f, s)$ be a decision tree. Then the risk of D or its induced tree classifier d is defined as*

$$R(D) = R(T, f, s) := \sum_{\nu \in \ell(T)} \varphi(p_\nu) P(\mathcal{X}_\nu)\,. \tag{1}$$

The main advantage of risk function of this form is that we can compute it recursively. The optimal tree design may be posed as the following optimization problem:

$$\text{Minimize } R(T, f, s) \text{ with respect to } T, f, s.$$

One can easily see that in the continuous case (i.e. if λ has an absolutely continuous component) there is no optimal solution for this optimization problem because the size of the decision tree tends to infinite. Thus, we have to apply some constraint, e.g. we have to bound the number of leaves or the depth of the tree. For example, if we fix the number of leaves of the decision tree, then the above optimization problem can be solved in two steps:

$$\text{Step 1: for a given pair } (f^*, s^*)\,, \text{ find } T^* := \arg\min_T R(T, f^*, s^*)\,,$$

$$\text{Step 2: for a given } T^*\,, \text{ find } (f^*, s^*) := \arg\min_{f,s} R(T^*, f, s)\,.$$

In the first step we look for the optimal binary tree in the case when the feature and splitting value configuration is fixed. We suppose that the correspondence between the feature (predictor) variables with their splitting values and the leaves of the possible trees is uniquely defined, e.g. from the left to the right. In the second step we look for the optimal feature and splitting value configuration when the binary tree is fixed. Then these two steps are iterated until convergence. This approach is similar to the one in Kurzynski (1983) suggested in the Bayesian framework.

In practice the density function λ of the predictor variables and the conditional probability p are unknown. We estimate the quantities $P_\nu = P(\mathcal{X}_\nu)$ and p_ν, $\nu \in \ell(T)$, from a learning data set $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ in the following way:

$$\widehat{P}_\nu := \frac{1}{n}|\{i : \boldsymbol{x}_i \in \mathcal{X}_\nu\}|, \qquad \widehat{p}_\nu := \frac{|\{i : \boldsymbol{x}_i \in \mathcal{X}_\nu, y_i = 1\}|}{|\{i : \boldsymbol{x}_i \in \mathcal{X}_\nu\}|}$$

for all $\nu \in \ell(T)$, where $|A|$ denotes the cardinality of a set A.

The problem of designing a really optimal decision tree seems to be a very difficult problem. It is conjectured that the problem with the above or more general risk function to classify an unknown sample is NP-complete. Hyafil and Rivest (1976) proved similar result for decision tables, see also Murphy and McCraw (1991). Hoeffgen et al. (1995) showed that also the not so complex problem of finding a simple linear split is NP-hard. These facts supply motivation for finding efficient heuristics or stochastics for constructing near-optimal decision trees.

3 Metropolis-Hastings Algorithm on Decision Trees

The CART method suggested by Breiman et al. (1984) is a greedy, deterministic algorithm. This means that in each step all the possible splits are tested and the best one is chosen among them. There can be cases, when although we always choose the best split, at the end we will be far from the optimal tree. So the global optimum – as in many other cases – is not the consequence of the local optimum. That is why the stochastic search can help, and we apply the Metropolis-Hastings (MH) algorithm to find the optimal tree.

In the sequel of the paper we will apply the MH algorithm in several settings, thus we review the algorithm in general setup.

Let $q(x, y)$, $x, y \in \Omega$, be an arbitrary matrix of transition probabilities (transition kernel) on a finite state space Ω. (That is $q(x, y) \geq 0$ for all $x, y \in \Omega$, and $\sum_{y\in\Omega} q(x, y) = 1$ for all $x \in \Omega$.) Define the acceptance probability by $\alpha(x, y) = \min\{1, h(x, y)\}$, $x, y \in \Omega$, where

$$h(x, y) := \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}$$

is the Hastings quotient, and π is a probability measure on Ω. Note that $h(x, y) := +\infty$ if $\pi(x)q(x, y) = 0$. Introduce a Markov chain, called MH chain, $\{\xi_n, n \in \mathbb{N}\}$ in the following way. If $\xi_n = x$ then let

$$\xi_{n+1} = \begin{cases} y & \text{with probability } \alpha(x, y)\,, \\ x & \text{otherwise.} \end{cases}$$

Then $\{\xi_n, n \in \mathbb{N}\}$ is an ergodic Markov chain, and π is the unique stationary distribution of the chain (see Roberts, 1996).

The transition kernel of the above MH chain is

$$K(x, y) = q(x, y)\alpha(x, y) + r(x)\delta(x, y)\,,$$

where

$$r(x) = 1 - \sum_{y \in \Omega} q(x,y)\alpha(x,y)\,,$$

and $\delta(x,y) = 1$, if $x = y$, and 0 otherwise. In the applications we need such MH chains which have enough large loop probabilities, i.e. $K(x,x) \geq 1/2$, for all $x \in \Omega$. This can be done by relaxing the chain in the current state by introducing the relaxed transition kernel

$$\begin{aligned} K_r(x,y) &= \frac{1}{2}K(x,y) + \frac{1}{2}\delta(x,y) \\ &= \frac{1}{2}q(x,y)\alpha(x,y) + \left(1 - \frac{1}{2}\sum_{y \in \Omega} q(x,y)\alpha(x,y)\right)\delta(x,y)\,. \end{aligned}$$

Another way to achieve that the loop probabilities are greater than or equal to $1/2$ is to choose a proposal kernel q such that $q(x,x) = 1/2$, for all $x \in \Omega$. In fact, one can realize the relaxed MH chain by relaxing the proposal distribution. If q is an arbitrary proposal kernel, then let $q_r := q/2 + \delta/2$ be its relaxed proposal kernel. Then one can easily verify that $K_r(q) = K(q_r)$, where $K_r(q)$ denotes the relaxed MH kernel given by the proposal distribution q, and $K(q_r)$ denotes the MH kernel given by the relaxed proposal kernel q_r. Finally, we note that a Markov chain can be relaxed by tossing a fair coin at each step: in case of head we go further according to the original transition kernel, but in case of tail the chain remains in the current state.

In the first step of our optimization problem the feature and split configuration is fixed, hence the risk function depends on the binary tree, which is the skeleton of the decision tree. Thus we have to solve the following discrete optimization problem:

$$\text{Minimize } R(T) \text{ with respect to } T \in \mathcal{T}_\ell,$$

where $\mathcal{T}_\ell$ is the set of binary trees having ℓ internal nodes and $\ell + 1$ leaves. By the help of the risk function we define a probability distribution on the state space $\Omega := \mathcal{T}_\ell$. Consider the Gibbs distribution

$$\pi(T) := \kappa \exp\{-R(T)\}\,, \qquad T \in \mathcal{T}_\ell\,,$$

where $\kappa^{-1} := \sum_{T \in \mathcal{T}_\ell} \exp\{-R(T)\}$ is a normalizing constant.

In order to define our MH algorithm, we have to introduce an accessibility relation between the $\ell + 1$-leaved binary trees. By elementary tree we will mean a binary tree having one root and two terminal nodes. A subtree of a binary tree is such a subgraph which is a binary tree as well. By a prunable elementary tree of a binary tree we mean a subtree which is an elementary tree such that its terminal nodes are leaves in the original tree too. Let $T, T' \in \mathcal{T}_\ell$ such that $T \neq T'$. We say that T and T' are accessible from each other if they both have a common binary subtree which we get by removing a prunable elementary tree from each, respectively. (See, e.g., Figures 1 and 2.) This accessibility relation is denoted by $T \leftrightarrow T'$. Note that we can get T' from T by applying the PRUNE and GROW step of Chipman et al. (1998) successively. Denote by $p(T)$ the set of prunable elementary trees of T. One can easily check that the number of accessible trees from T

is $|p(T)|(\ell-1)$. We introduce the relaxed proposal distribution by choosing uniformly among the accessible trees:

$$q_r(T,T') := \begin{cases} \dfrac{1}{2p(T)|(\ell-1)}\,, & \text{if } T \leftrightarrow T'\,, \\ \dfrac{1}{2}\,, & \text{if } T = T'\,, \\ 0\,, & \text{otherwise.} \end{cases}$$

In this case the Hastings quotient is given by

$$h(T,T') = \frac{\pi(T')|p(T)|}{\pi(T)|p(T')|}\,.$$

The transition kernel of the MH chain induced by this proposal distribution is

$$K(T,T') = \begin{cases} \dfrac{\pi(T')}{\pi(T)}\dfrac{1}{2|p(T')|(\ell-1)}\,, & \text{if } T \leftrightarrow T' \text{ and } h(T,T') < 1\,, \\ \dfrac{1}{2|p(T)|(\ell-1)}\,, & \text{if } T \leftrightarrow T' \text{ and } h(T,T') \geq 1\,, \\ \dfrac{3}{2} - \displaystyle\sum_{\substack{T':T\leftrightarrow T' \\ h(T,T')<1}} \dfrac{\pi(T')}{\pi(T)}\dfrac{1}{2|p(T')|(\ell-1)} & \\ \quad - \displaystyle\sum_{\substack{T':T\leftrightarrow T' \\ h(T,T')\geq 1}} \dfrac{1}{2|p(T)|(\ell-1)}\,, & \text{if } T = T'\,, \\ 0\,, & \text{otherwise.} \end{cases}$$

In the second step of the optimization problem the skeleton T of the decision trees is fixed, thus the risk function depends only on the predictor variables and their splitting values. In this case the derived optimization problem is not necessarily discrete, for example if there is at least one continuous variable among the predictor variables. But we can discretise the problem by introducing a set of admissible splits. For an ordinal predictor variable the splitting values will be its possible values, while for a nominal variable more complicated splits, like subsets of its range, will be allowed. If a learning data set is given, then the splitting values will be the observed values in the sample. Hence, the number of the possible splits is not more than np where n is the sample size, and p is the number of predictors variables. In the sequel we suppose that the number of admissible splits is $k \in \mathbb{N}$. Finally, for the sake of simplicity, we suppose that each admissible split is allowed at each internal node of the binary tree T. This means that a split can occur at several internal nodes.

The optimization problem can be reformulated in the following way:

$$\text{Minimize } R(S) \text{ with respect to } S \in \mathcal{S}_T,$$

where $\mathcal{S}_T$ is the set of all split configurations on T. Our assumptions imply that $|\mathcal{S}_T| = k^\ell$. Define the Gibbs distribution in this case as

$$\pi(S) := \kappa \exp\{-R(S)\}\,, \qquad S \in \mathcal{S}_T\,,$$

where $\kappa^{-1} := \sum_{S \in \mathcal{S}_T} \exp\{-R(S)\}$ is a normalizing constant.

In order to define MH algorithm for the second step we introduce an accessibility relation between the splitting configuration. One can see that $S \in \mathcal{S}_T$ can be written as a function $S : v \mapsto (j, c)$, where $v \in i(T)$, and $(j, c), j \in \{1, \ldots, p\}$, $c \in \mathbb{R}$, is an admissible split for X_j, i.e. we cut the node v in the X_jth variable at value c. Let $S, S' \in \mathcal{S}_T$ such that $S \neq S'$. We say that S and S' are accessible from each other, denoted by $S \leftrightarrow S'$, if there exists $v \in i(T)$ such that $S(v) \neq S'(v)$ and $S(v') = S'(v')$ for all $v' \in i(T)$, $v' \neq v$. Define the relaxed proposal distribution as

$$q_r(S, S') = \begin{cases} \dfrac{1}{2\ell(k-1)}\,, & \text{if } S \leftrightarrow S'\,, \\ \dfrac{1}{2}\,, & \text{if } S = S'\,, \\ 0\,, & \text{otherwise,} \end{cases}$$

i.e., we choose uniformly among the accessible split configurations. The Hastings quotient is given by

$$h(S, S') = \frac{\pi(S')}{\pi(S)} = \exp\{-(R(S') - R(S))\}\,.$$

In this case the proposal distribution α coincides with the Metropolis sampler. The transition kernel of this MH chain is

$$K(S, S') = \begin{cases} \dfrac{\exp\{-(R(S') - R(S))\}}{2\ell(k-1)}\,, & \text{if } S \leftrightarrow S' \text{ and } R(S') > R(S)\,, \\ \dfrac{1}{2\ell(k-1)}\,, & \text{if } S \leftrightarrow S' \text{ and } R(S') \leq R(S)\,, \\ \dfrac{3}{2} - \displaystyle\sum_{\substack{S' : S \leftrightarrow S' \\ R(S') > R(S)}} \dfrac{\exp\{-(R(S') - R(S))\}}{2\ell(k-1)} - \\ - \displaystyle\sum_{\substack{S' : S \leftrightarrow S' \\ R(S') \leq R(S)}} \dfrac{1}{2\ell(k-1)}\,, & \text{if } S = S'\,, \\ 0\,, & \text{otherwise.} \end{cases}$$

Note that this MH chain is very similar to that one which is given in the graph-coloring problem, see Jerrum (1998, Section 3).

4 Mixing Time of Reversible Markov Chain

Let us define a discrete Markov chain on finite state space Ω with transition kernel $K : \Omega^2 \to [0, 1]$, i.e. let $K(x, y) \geq 0$ for all $x, y \in \Omega$, and $\sum_{y \in \Omega} K(x, y) = 1$ for all $x \in \Omega$. We suppose that the Markov chain is irreducible and aperiodic, thus ergodic with unique stationary distribution $\pi : \Omega \to [0, 1]$. The Markov chain is called time-reversible if it fulfills the detailed balance:

$$\pi(x)K(x, y) = \pi(y)K(y, x) \qquad \text{for all } x, y \in \Omega\,.$$

Note that an MH chain or relaxed MH chain is always reversible. Define the n-step transition kernel recursively by

$$K_n(x,y) = \sum_{z\in\Omega} K_{n-1}(x,z)K(z,y)\,, \qquad x,y\in\Omega\,.$$

Then $K_n(x,\cdot)$ is the distribution of the chain at time n given that x is the initial state.

To measure the closeness to stationarity at time n we use variation distance:

$$\Delta_x(n) = \frac{1}{2}\sum_{y\in\Omega} |K_n(x,y)-\pi(y)| = \max_{A\subseteq\Omega} |K_n(x,A)-\pi(A)|\,.$$

Then the rate of convergence to stationarity from initial state x can be described by the mixing time:

$$\tau_x(\varepsilon) = \min\{n : \Delta_x(m)\le\varepsilon \text{ for all } m\ge n\}\,.$$

The Markov chain is called rapid mixing if $\tau(\varepsilon) := \max_{x\in\Omega}\tau_x(\varepsilon) \le P(d,\varepsilon^{-1})$, where P is a polynomial and d is the dimension of Ω.

There are a number of methods for estimating the mixing time. These techniques include e.g. geometric tools (canonical paths, conductance), analytic tools (estimating the spectral gap by Nash or Sobolev inequality) and comparison and coupling tools.

In this paper we apply the canonical paths technique for decision trees. The heart of this technique is an undirected graph $(\Omega,\mathcal{E})$ introduced by the Markov chain, where the vertex set is the state space Ω, and the edge set $\mathcal{E}$ is defined by

$$\mathcal{E} := \{(x,y)\in\Omega\times\Omega : \pi(x)P(x,y)>0\}\,.$$

A path γ in $(\Omega,\mathcal{E})$ is a sequence of vertices $\gamma=(x_0,\dots,x_k)$ such that $(x_{i-1},x_i)\in\mathcal{E}$, $i=1,\dots,k$. Equivalently, γ can be viewed as a sequence of edges $\gamma=(e_1,\dots,e_k)$ with $e_i=(x_{i-1},x_i)\in\mathcal{E}$, $i=1,\dots,k$. The length of such a path γ is $|\gamma|=k$. For each pair $(x,y)\in\Omega\times\Omega$ let $\Gamma(x,y)$ denote the set of every path $\gamma=(x_0,\dots,x_k)$, $k\in\mathbb{N}$, in $(\Omega,\mathcal{E})$ which starts from x (i.e. $x_0=x$) and ends at y (i.e. $x_k=y$), and have no repeated vertices (i.e. $x_i\neq x_j$, if $i\neq j$). For each $(x,y)\in\Omega\times\Omega$ choose exactly one path $\gamma(x,y)$ in $\Gamma(x,y)$ which is called the canonical path belonging to (x,y). Denote the set of canonical paths by $\Gamma := \{\gamma(x,y) : x,y\in\Omega\}$. Define the probability of the edge $e=(x,y)$ by $Q(e)=K(x,y)\pi(x)$. Finally, introduce the maximal edge loading of the set Γ of canonical paths as

$$\rho(\Gamma) = \max_{e\in\mathcal{E}} \frac{1}{Q(e)} \sum_{\substack{x,y\in\Omega:\\ e\in\gamma(x,y)}} \pi(x)\pi(y)|\gamma(x,y)|\,.$$

The basic result of Sinclair (see Sinclair, 1992, Jerrum, 1998, Theorem 4.1) gives an estimate for the mixing time by the help of maximal edge loading.

Theorem 1. *Let $K(x,y)$, $x,y\in\Omega$, be the transition kernel of a reversible, ergodic Markov chain on the finite state space Ω with stationary distribution π and loop probabilities $K(x,x)\ge 1/2$ for all $x\in\Omega$. Let Γ be a set of canonical paths with maximal edge loading $\rho(\Gamma)$. Then the mixing time of the Markov chain satisfies*

$$\tau_x(\varepsilon) \le \rho(\Gamma)(\log\pi(x)^{-1} + \log\varepsilon^{-1})\,, \qquad \textit{for all } x\in\Omega \textit{ and } \varepsilon>0.$$

It is well known, that there is a strong connection between the mixing time and the so-called spectral gap of the Markov chain. In fact, $1/\lambda \leq \rho(\Gamma)$, where λ is the spectral gap of the Markov chain. (See Theorem 3.2.1 in Saloff-Coste (1997).)

5 Speed of MH Algorithm on Decision Trees

In the effective MH algorithms the mixing time can be bounded by the dimension of the state space which is significantly less then the cardinality of the state space. For decision trees the number of internal nodes (or leaves) plays the role of the dimension.

In the first optimization step the accessibility relation defines a graph on the set of $\ell+1$-leaved trees with ℓ internal nodes, where $\ell \in \mathbb{N}$. The cardinality of the state space of trees with ℓ internal nodes is $\frac{1}{\ell+1}\binom{2\ell}{\ell} \approx c \cdot \ell^{-1/2}(\ell+1)^{-1} \cdot 4^{2\ell}$.

Definition 3. *Let $\ell \in \mathbb{N}$. The $(\ell+1)$-leaved tree-graph is the undirected graph $(\mathcal{T}_\ell, \mathcal{E})$, where the vertex set is the set of $(\ell+1)$-leaved trees $\mathcal{T}_\ell$, and the edge set $\mathcal{E}$ is the set of pairs (T, T'), $T, T' \in \mathcal{T}_\ell$, such that $T \leftrightarrow T'$.*

On Figures 1 and 2 one can see the 4-leaved and 5-leaved tree-graph, respectively. In this case the set of canonical paths can be given in such a way that every edge of the $(\ell+1)$-leaved tree-graph is part of not more than $|\mathcal{T}_\ell|$ number of canonical paths. The proof of this claim follows from the fact, that for every separation of this graph into two disjoint subgraphs the number of the disappearing edges of the original graph fulfills

$$\frac{|\mathcal{G}|(|\mathcal{T}_\ell| - |\mathcal{G}|)}{|\mathcal{T}_\ell|} \leq |e(\mathcal{G})|\,, \tag{2}$$

where $\mathcal{G}$ denotes one of the subgraphs after the separation, and $e(\mathcal{G})$ is the set of the disappearing edges.

For example let $\mathcal{G}$ be the set of such $(\ell+1)$-leaved trees which contain the subtree showed on Figure 3. In this case inequality (2) has the following form

$$\frac{|\mathcal{T}_{\ell-1}|(|\mathcal{T}_\ell| - |\mathcal{T}_{\ell-1}|)}{|\mathcal{T}_\ell|} \leq |\mathcal{T}_{\ell-2}| \cdot \ell\,. \tag{3}$$

Using the relation between the number of ℓ- and $(\ell+1)$-leaved trees we get

$$\left(1 - \frac{\ell+1}{2(2\ell-1)}\right)\frac{2(2\ell-3)}{\ell} \leq \ell\,.$$

The left side of this inequality can be bounded by

$$\frac{2(2\ell-3)}{\ell} \leq \frac{2 \cdot 2\ell}{\ell} \leq 4\,,$$

so (3) holds for trees having $\ell \geq 4$ leaves. For $\ell = 2, 3$ one can easily prove the inequality.

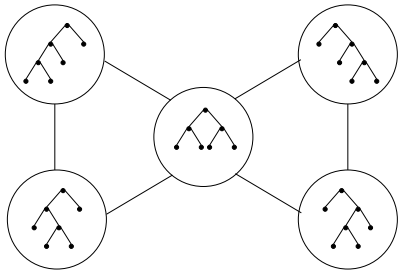

Figure 1: The 4-leaved tree-graph

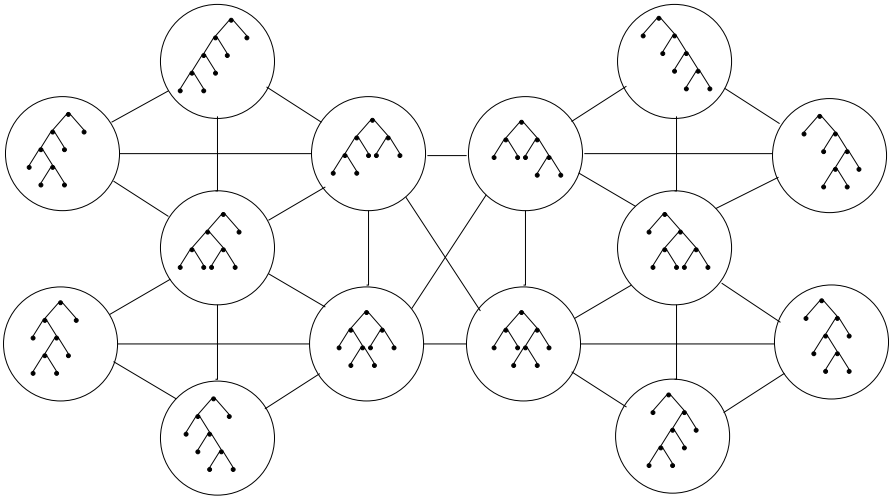

Figure 2: The 5-leaved tree-graph

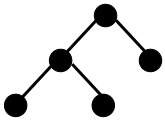

Figure 3: The subtree that every ℓ-leaved tree contains in $\mathcal{G}$

The canonical paths can be chosen such that their maximal length is $\ell - 1$, i.e.

$$\gamma_* = \max_{T,T'} |\gamma(T, T')| = \ell - 1 \,. \tag{4}$$

One possible canonical path in the 5-leaved tree-graph is showed on Figure 4 with bold lines.

Theorem 2. *The mixing time of the MH chain defined on the* $(\ell + 1)$*-leaved tree-graph can be bounded by*

$$\tau_T(\varepsilon) \leq \exp\{2C_R\}\ell^3(\ell \log 4 + \log \varepsilon^{-1} + C_R) \,, \tag{5}$$

for all starting tree $T \in \mathcal{T}_\ell$*, where* $C_R := R_{\max} - R_{\min}$*, and* $R_{\min} := \min\{R(T) : T \in \mathcal{T}_\ell\}$ *and* $R_{\max} := \max\{R(T) : T \in \mathcal{T}_\ell\}$.

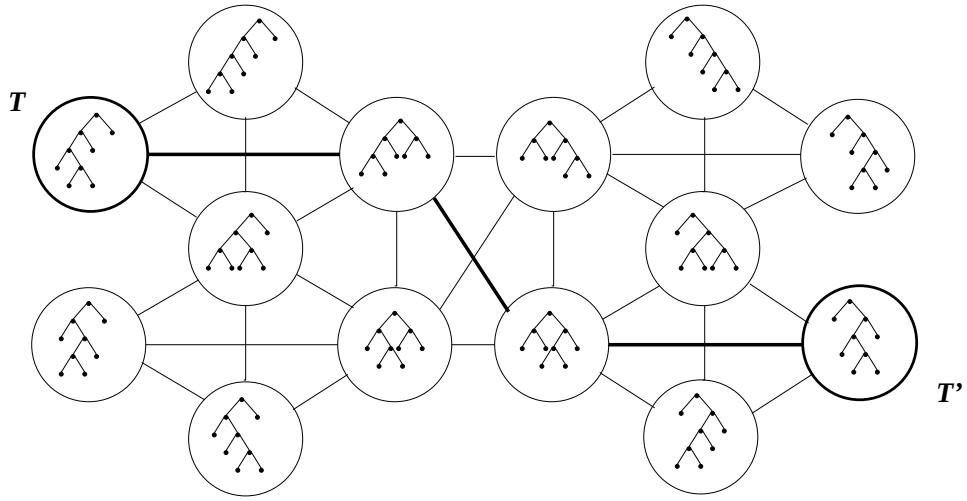

Figure 4: One canonical path in the 5-leaved tree-graph

Remark 1. *The quantity C_R is the range of the risk function. By the definition* (1) *of the risk function one can easily see that $C_R \le 1/2$. Thus the main part in* (5) *is* $3.76 \times \ell^4$.

Proof. First we estimate the maximal edge loading of the previously defined set Γ of canonical paths. If $e = (T, T') \in \mathcal{E}$, than one can see that

$$Q(e) = \begin{cases} \dfrac{\pi(T')}{2|p(T')|(\ell-1)}\,, & \text{if } h(T,T') < 1\,, \\ \dfrac{\pi(T)}{2|p(T)|(\ell-1)}\,, & \text{if } h(T,T') \ge 1\,. \end{cases}$$

Since $|p(T)| \le \ell/2$, for all $T \in \mathcal{T}_\ell$, we have

$$Q^{-1}(e) \le \kappa^{-1}\ell(\ell-1)\exp\{R_{\max}\}\,. \tag{6}$$

On the other hand, by (4), we obtain

$$\sum_{(T,T')\in C(e)} \pi(T)\pi(T')|\gamma(T,T')| \le \kappa^2\gamma_*|C(e)|\exp\{-2R_{\min}\}\,, \tag{7}$$

where $C(e) := \{(T,T') \in \mathcal{T}_\ell \times \mathcal{T}_\ell : e \in \gamma(T,T')\}$. Finally,

$$\kappa^{-1} = \sum_{T\in\mathcal{T}_\ell} \exp\{-R(T)\} \begin{array}{l} \le \exp\{-R_{\min}\}|\mathcal{T}_\ell| \\ \ge \exp\{-R_{\max}\}|\mathcal{T}_\ell| \end{array},$$

which implies

$$\kappa \le \exp\{R_{\max}\}/|\mathcal{T}_\ell|\,. \tag{8}$$

By (6), (7) and (8) we obtain

$$\rho(\Gamma) \le \exp\{2C_R\}(\ell+1)(\ell-1)^2 \max_{e\in\mathcal{E}} |C(e)|/|\mathcal{T}_\ell|\,.$$

By the definition of canonical paths we get $|C(e)| \leq |\mathcal{T}_\ell|$ for all $e \in \mathcal{E}$, i.e. the number of canonical paths which are passing through the edge e is equal to or less than the number of binary trees having ℓ internal nodes.

Thus we proved that

$$\rho(\Gamma) \leq \exp\{2C_R\}(\ell+1)(\ell-1)^2 \leq \exp\{2C_R\}\ell^3\,.$$

It is well known that $|\mathcal{T}_\ell| = \frac{1}{\ell+1}\binom{2\ell}{\ell}$, which is the ℓth Catalan-number, see Cormen et al. (1990) Problem 13.4. Since

$$\log \pi^{-1}(T) = \log \kappa^{-1} + R(T) \leq \log |\mathcal{T}_\ell| + C_R\,,$$

the Stirling formula implies the assertion.

Example 1. In practice usually $\ell \leq 30$. Let for example $\ell = 30$ and $\varepsilon = 10^{-4}$. Then we have that $\tau_T(\varepsilon) \leq 3753561$, while one can see that the 30th Catalan-number is 6564120420.

In the second optimization step the accessibility relation defines a similar graph structure to the k-dimensional hyper-cube. We can think of a splitting configuration S as a vertex of the hyper-cube, and to an accessible pair (S, S') as an edge of the hyper-cube. Introduce a linear ordering $v_1, \ldots, v_\ell$ on the set of the internal nodes of the tree T, e.g. from up to down from the left to the right. Then a splitting configuration S can be written as $(S(v_1), \ldots, S(v_\ell))$.

Let $S = (S(v_1), \ldots, S(v_\ell))$ and $S' = (S'(v_1), \ldots, S'(v_\ell))$ be arbitrary splitting configurations. The canonical path $\gamma(S, S')$ from S to S' is composed of ℓ edges, 1 to ℓ, where edge i is simply

$$\begin{aligned} e = (&(S'(v_1), \ldots, S'(v_{i-1}), S(v_i), S(v_{i+1}), \ldots, S(v_\ell)),\\ &(S'(v_1), \ldots, S'(v_{i-1}), S'(v_i), S(v_{i+1}), \ldots, S(v_\ell)))\,, \end{aligned} \tag{9}$$

i.e. we change the split at the ith internal node. One can easily see that the length of every canonical path is exactly ℓ.

Theorem 3. *The mixing time of the MH chain, defined on the graph of splitting configurations with at most k admissible splits, can be bounded by*

$$\tau_S(\varepsilon) \leq 2\exp\{2C_R\}\ell^2(\ell \log k + \log \varepsilon^{-1} + C_R)$$

for all starting configuration S, where $C_R := R_{\max} - R_{\min}$, and $R_{\min} := \min\{R(S) : S \in \mathcal{S}_T\}$ and $R_{\max} := \max\{R(S) : S \in \mathcal{S}_T\}$.

Proof. We estimate the maximal edge loading $\rho(\Gamma)$. If $e = (S, S')$ is an edge, then one can see that

$$Q(e) = \begin{cases} \dfrac{\pi(S')}{2\ell(k-1)}\,, & \text{if } h(S,S') < 1\,,\\[2ex] \dfrac{\pi(S)}{2\ell(k-1)}\,, & \text{if } h(S,S') \geq 1\,. \end{cases}$$

Thus

$$Q(e) \geq \frac{\kappa}{2\ell(k-1)} \exp\{-R_{\max}\} .$$

Moreover, we obtain

$$\sum_{e \in \gamma(S,S')} \pi(S)\pi(S')|\gamma(S,S')| \leq \kappa^2 \ell \exp\{-2R_{\min}\}|\{(S,S') : e \in \gamma(S,S')\}| .$$

Suppose, that the edge e is given by (9). Consider the number of canonical paths $\gamma(S,S')$ that include e. The number of possible choices for S is $k^{\ell-i}$, as the first i position are fixed. By a similar argument the number of possible choices for S' is k^{i-1}. Thus the total number of canonical paths containing a particular edge e is $k^{\ell-1}$. Using this result, we get

$$\rho(\Gamma) \leq 2\ell^2(k-1)k^{\ell-1}\kappa \exp\{R_{\max} - 2R_{\min}\} \leq 2\exp\{2C_R\}\ell^2 ,$$

since $\kappa \leq k^{-\ell} \exp\{R_{\max}\}$. A similar argument to the end of Theorem 2 completes the proof.

Corollary. *For a learning data set with sample size n the mixing time can be bounded by*

$$\tau_S(\varepsilon) \leq 2\exp\{2C_R\}\ell^2(\ell(\log n + \log p) + \log \varepsilon^{-1} + C_R) .$$

Example 2. If $n = 10^6$, then we need just 2×10^6 steps to achieve the stationary distribution and the best decision tree instead of the 30th Catalan-number.

6 Conclusions

Decision trees are these days commonly used tools for solving the problem of supervised learning. For the construction of decision trees the CART method suggested by Breiman et al. (1984) is not always effective. This algorithm was improved by Chipman et al. (1998) with a stochastic search in Bayesian framework. This idea can be applied not only in that case when a prior distribution is given on the decision trees but also in the case considered in this paper when a discrete optimization problem is given.

We defined two kinds of MH chains on decision trees. We applied the method of canonical paths to estimate the rate of convergence and the mixing time. One can see, that using our method the optimal $\ell+1$-leaved decision tree can be reached in polynomial time. We think that our method can be extended for more general splits, see e.g. Loh and Vanichsetakal (1988).

A remaining task for us is to see how this method works in practice. How fast we could find the best tree when we have a learning data set and have to use it to estimate the distributions needed for the algorithm.

Acknowledgement

This work was partially supported by the Hungarian Scientific Research Fund Grant No. OTKA–T047067/2004.

References

Breiman, L., Friedman, J. H., Olsen, A. O., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth International Group.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian CART model search. *Journal of the American Statistical Association*, *93*, 935-960.

Cormen, T. H., Leierson, C. E., and Rives, L. R. (1990). *Introduction to Algorithms*. The Massachusetts Institute of Technology.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning. Data Mining, Inference and Prediction*. New York: Springer.

Hoeffgen, K. U., Simon, H. U., and Van Horn, K. S. (1995). Robust trainability of single neurons. *Journal of Computer System Sciences*, *50*, 114-125.

Hyafil, L., and Rivest, R. L. (1976). Constructing optimal binary trees is NP-complete. *Information Processing Letters*, *5*, 15-17.

Jerrum, M. (1998). Mathematical foundations of the markov chain monte carlo method. In *Probabilistic methods for algorithmic discrete mathematics.*

Jerrum, M. (2003). *Counting, Sampling and Integrating: Algorithms and Complexity*. Basel: Birkhäuser.

Jerrum, M., and Sinclair, A. (1996). The Markov chain Monte Carlo method: An approach to approximate counting and integration. In D. S. Hochbaum (Ed.), *Approximation Algorithm for NP-hard Problems* (p. 482-520). Boston.

Kurzynski, M. W. (1983). The optimal strategy of a tree classifier. *Pattern Recognition*, *16*, 81-87.

Loh, W. Y., and Vanichsetakal, N. (1988). Tree-structured classification via generalized discriminant analysis. *Journal of the American Statistical Association*, *83*, 715-725.

Murphy, P. M., and McCraw, R. L. (1991). Designing storage efficient decision trees. *IEEE Transactions on Computers*, *40*, 315-320.

Murthy, S. K. (1998). Automatic construction of decision trees from data: A multi-disciplinary survey. In *Data Mining and Knowledge Discovery, 2*. Boston: Kluwer Academic Publishers.

Roberts, G. O. (1996). Markov chain concepts related to sampling algorithms. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice* (p. 45-57). London: Chapman & Hall/CRC.

Safavian, S. R., and Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Transaction on Systems, Man and Cybernetics*, *21*, 660-674.

Saloff-Coste, L. (1997). Lectures on finite Markov chains. In P. Bernard (Ed.), *Lectures on Probability Theory and Statistics* (p. 301-413). Berlin: Springer.

Sinclair, A. (1992). Improved bounds for mixing rates of Markov chains and multicommodity flow. *Combinatorics, Probability and Computing*, *1*, 351-370.

Authors' address:
Márton Ispány and Ilona Krasznahorkay
Department of Applied Mathematics and Probability Theory
University of Debrecen
P.O. Box 12
H-4010 Debrecen, Hungary
E-mail: ispany@inf.unideb.hu and krasznil@inf.unideb.hu