Hierarchical Decision Tree Induction in Distributed Genomic Databases

Amir Bar-Or, Daniel Keren, Assaf Schuster, and Ran Wolff

A. Bar-Or is with HP Cambridge research lab.

D. keren is with the Haifa University – Israel

A. Schuster is with the Technion – Israel Institute of Technology

R. Wolff is with the University of Maryland at Baltimore County

Abstract

Classification based on decision trees is one of the important problems in data mining and has applications in many fields. In recent years, database systems have become highly distributed, and distributed system paradigms such as federated and peer-to-peer databases are being adopted. In this paper, we consider the problem of inducing decision trees in a large distributed network of genomic databases. Our work is motivated by the existence of distributed databases in healthcare and in bioinformatics, and by the emergence of systems which automatically analyze these databases, and by expectancy that these databases will soon contain large amounts of highly dimensional genomic data. Current decision tree algorithms require high communication bandwidth when executed on such data, which large-scale distributed systems. We present an algorithm that sharply reduces the communication overhead by sending just a fraction of the statistical data. A fraction which is nevertheless sufficient to derive the exact same decision tree learned by a sequential learner on all the data in the network. Extensive experiments using standard synthetic SNP data show that the algorithm utilizes the high dependency among attributes, typical to genomic data, to reduce communication overhead by up to 99%. Scalability tests show that the algorithm scales well with both the size of the dataset, the dimensionality of the data, and the size of the distributed system.

Index Terms

data mining, distributed algorithms, decision trees, classification.

APPENDIX I

PROOFS OF BOUNDS

Lemma 1: For any $GiniIndex(P_1)$, $GiniIndex(P_2)$, n_1, n_2, an upper bound on $GiniIndex(P)$ is given by:

$$Upper\ bound = \frac{n_1 GiniIndex(P_1) + n_2 GiniIndex(P_2)}{n_1 + n_2} \tag{1}$$

Proof: Following the previous notations, $GiniIndex(P)$ is given by:

$$GiniIndex(P) = \frac{\frac{(a_{11}+b_{11})^2+(a_{12}+b_{12})^2}{a_{11}+a_{12}+b_{11}+b_{12}} + \frac{(a_{21}+b_{21})^2+(a_{22}+b_{22})^2}{a_{21}+a_{22}+b_{21}+b_{22}}}{a_{11}+a_{12}+a_{21}+a_{22}+b_{11}+b_{12}+b_{21}+b_{22}}$$

while the upper bound provided in this lemma is given by:

$$Upper\ bound =$$

$$\frac{\frac{(a_{11})^2+(a_{12})^2}{a_{11}+a_{12}} + \frac{(b_{11})^2+(b_{12})^2}{b_{11}+b_{12}} + \frac{(a_{21})^2+(a_{22})^2}{a_{21}+a_{22}} + \frac{(b_{21})^2+(b_{22})^2}{b_{21}+b_{22}}}{a_{11}+a_{12}+a_{21}+a_{22}+b_{11}+b_{12}+b_{21}+b_{22}}$$

Now, to prove that $Upper\ bound - GiniIndex(P) \geq 0$, it is enough for the following inequalities to hold:

$$1: \frac{(a_{11})^2+(a_{12})^2}{a_{11}+a_{12}} + \frac{(b_{11})^2+(b_{12})^2}{b_{11}+b_{12}} - \frac{(a_{11}+b_{11})^2+(a_{12}+b_{12})^2}{a_{11}+a_{12}+b_{11}+b_{12}} \geq 0$$

$$2: \frac{(a_{21})^2+(a_{22})^2}{a_{21}+a_{22}} + \frac{(b_{21})^2+(b_{22})^2}{b_{21}+b_{22}} - \frac{(a_{21}+b_{21})^2+(a_{22}+b_{22})^2}{a_{21}+a_{22}+b_{21}+b_{22}} \geq 0$$

It is straightforward to verify that inequality 1 is equivalent to:

$$\frac{(a_{11}b_{12} - a_{12}b_{11})^2}{(a_{11}+a_{12}+b_{11}+b_{12})(a_{11}+a_{12})(b_{11}+b_{12})} \geq 0$$

The above inequality is always satisfied, since both the numerator and denominator are always positive (all variables are non-negative). Inequality 2 is handled in the same manner. ■

Lemma 2: Let P_1, P_2, n_1, n_2 be given. Furthermore, let the candidate binary split decision divide P_1 into two subsets, P_1^{left} and P_1^{right}, with sizes of n_1^{left} and n_1^{right} ($n^1 = n_1^{left} + n_1^{right}$) respectively. Then a lower bound on $GiniIndex(P)$ is given by:

$$Lower\ Bound = \frac{GiniIndex(P_1)}{\left[1+\frac{n_2}{n_1}\right]\left[1+\max\left\{\frac{n_2}{n_1^{left}}, \frac{n_2}{n_1^{right}}\right\}\right]} \tag{2}$$

Proof: Recall that $GiniIndex(P)$ is given by:

$$GiniIndex(P) = \frac{\frac{(a_{11}+b_{11})^2+(a_{12}+b_{12})^2}{a_{11}+a_{12}+b_{11}+b_{12}} + \frac{(a_{21}+b_{21})^2+(a_{22}+b_{22})^2}{a_{21}+a_{22}+b_{21}+b_{22}}}{a_{11}+a_{12}+a_{21}+a_{22}+b_{11}+b_{12}+b_{21}+b_{22}}$$

since all entries are positive, the above expression is bounded from below by (recall that $n_1 = a_{11} + a_{12} + a_{21} + a_{22}, n_2 = b_{11} + b_{12} + b_{21} + b_{22}$):

$$\frac{1}{1+\frac{n_2}{n_1}} \frac{\frac{(a_{11})^2+(a_{12})^2}{(a_{11}+a_{12})\left[1+\frac{b_{11}+b_{12}}{a_{11}+a_{12}}\right]} + \frac{(a_{21})^2+(a_{22})^2}{(a_{21}+a_{22})\left[1+\frac{b_{21}+b_{22}}{a_{21}+a_{22}}\right]}}{n_1}$$

But this expression is clearly bounded from below by

$$\frac{1}{1+\frac{n_2}{n_1}} \frac{\frac{\frac{(a_{11})^2+(a_{12})^2}{a_{11}+a_{12}}+\frac{(a_{21})^2+(a_{22})^2}{a_{21}+a_{22}}}{n_1}}{1+\max\left\{\frac{n_2^{left}}{n_1^{left}}, \frac{n_2^{right}}{n_1^{right}}\right\}} =$$

$$\frac{1}{1+\frac{n_2}{n_1}} \frac{GiniIndex(P_1)}{1+\max\left\{\frac{n_2^{left}}{n_1^{left}}, \frac{n_2^{right}}{n_1^{right}}\right\}}$$

and the proof follows immediately by noting that $n_2 \geq n_2^{left}, n_2^{right}$. Note that n_2^{left}, n_2^{right} don't need to be known – only the total size n_2.

We note here that, in general, it is impossible to derive a lower bound on the Gini index of a union; this is possible only if one population is quite smaller than the other, in which case a bound can be derived from the relative sizes and the Gini index of the larger population. Nevertheless, as we show in our experiments the lower bound becomes especially useful when outliers occur in the data. In such cases small populations report attributes which are entirely different from those reported by the majority of the population. By using the lower bound the algorithm can overcome small amounts of missing data (about the attributes reported by the majority) and by that avoid the need for additional communication. ■

A. Information Gain Function

Lemma 3: For any $InfoGain(P_1)$, $InfoGain(P_2)$, n_1, n_2, an upper bound on $InfoGain(P)$ is given by

$$\frac{n_1 InfoGain(P_1) + n_2 InfoGain(P_2)}{n_1 + n_2} \tag{3}$$

Proof: Recall that $InfoGain(P)$ equals

$$InfoGain(P) =$$

$$\frac{(a_{11}+b_{11})\log((A+B)_{11}) + (a_{12}+b_{12})\log((A+B)_{12})}{n_1+n_2} +$$

$$\frac{(a_{21}+b_{21})\log((A+B)_{21}) + (a_{22}+b_{22})\log((A+B)_{22})}{n_1+n_2}$$

where the following definitions have been adopted for brevity:

$$A_{11} = \frac{a_{11}}{a_{11} + a_{12}}, A_{12} = \frac{a_{12}}{a_{11} + a_{12}},$$
$$A_{21} = \frac{a_{21}}{a_{21} + a_{22}}, A_{22} = \frac{a_{22}}{a_{21} + a_{22}}$$

Similar notations are used for B and $A + B$, e.g., $(A+B)_{11} = \frac{a_{11}+b_{11}}{a_{11}+b_{11}+a_{12}+b_{12}}$).

Similarly, $InfoGain(P_1)$ and $InfoGain(P_2)$ are given by

$$InfoGain(P_1) =$$
$$\frac{a_{11}\log(A_{11}) + a_{12}\log(A_{12}) + a_{21}\log(A_{21}) + a_{22}\log(A_{22})}{n_1}$$
$$InfoGain(P_2) =$$
$$\frac{b_{11}\log(B_{11}) + b_{12}\log(B_{12}) + b_{21}\log(B_{21}) + b_{22}\log(B_{22})}{n_2}$$

We now define the auxiliary variables

$$\lambda_1 = \frac{a_{11} + a_{12}}{a_{11} + a_{12} + b_{11} + b_{12}},\ \lambda_2 = \frac{a_{21} + a_{22}}{a_{21} + a_{22} + b_{21} + b_{22}}$$

Clearly

$$(A+B)_{11} = \lambda_1 A_{11} + (1-\lambda_1)B_{11},$$
$$(A+B)_{12} = \lambda_1 A_{12} + (1-\lambda_1)B_{12}$$
$$(A+B)_{21} = \lambda_2 A_{21} + (1-\lambda_2)B_{21},$$
$$(A+B)_{22} = \lambda_2 A_{22} + (1-\lambda_2)B_{22}$$

In order to make the notations less cumbersome, let us bound one summand of $InfoGain(P)$ (the other summands are handled similarly):

$$(a_{11} + b_{11})\log((A+B)_{11}) =$$
$$\frac{a_{11} + b_{11}}{(A+B)_{11}}(A+B)_{11}\log((A+B)_{11}) =$$
$$\frac{a_{11} + b_{11}}{(A+B)_{11}}(\lambda_1 A_{11} + (1-\lambda_1)B_{11})\log(\lambda_1 A_{11} + (1-\lambda_1)B_{11})$$

Since the function $x\log(x)$ is convex, the last expression is bounded from above by

$$\frac{a_{11} + b_{11}}{(A+B)_{11}}(\lambda_1 A_{11}\log(A_{11}) + (1-\lambda_1)B_{11}\log(B_{11})) =$$
$$a_{11}\log(A_{11}) + b_{11}\log(B_{11})$$

Hence $InfoGain(P)$ is bounded from above by

$$(a_{11}\log(A_{11})+a_{12}\log(A_{12})+a_{21}\log(A_{21})+a_{22}\log(A_{22})+$$
$$b_{11}\log(B_{11})+b_{12}\log(B_{12})+b_{21}\log(B_{21})+b_{22}\log(B_{22}))/(n_1+n_2)=$$
$$\frac{n_1 InfoGain(P_1)+n_2 InfoGain(P_2)}{n_1+n_2}.$$

■

Lemma 4: Let P_1, n_1, n_2 be given. Furthermore, let the candidate split decision divide P_1 into two subsets, P_1^{left} and P_1^{right}, with size n_1^{left} and n_1^{right} respectively. Then a lower bound on $InfoGain(P)$ is given by:

$$\left[\frac{1}{1+\frac{n_2}{\min\{n_1^{left},n_1^{right}\}}}\cdot\frac{1}{1+\frac{n_2}{n_1}}\right] InfoGain(P_1) \tag{4}$$

Proof: Following the proof of Lemma 3, let us now bound $(a_{11}+b_{11})\log((A+B)_{11}) = (a_{11}+b_{11})\log(\lambda_1 A_{11}+(1-\lambda_1)B_{11})$ from *below*. Since $\log(x)$ is concave, $(a_{11}+b_{11})\log(\lambda_1 A_{11}+(1-\lambda_1)B_{11})$ is bounded from below by $(a_{11}+b_{11})(\lambda_1\log(A_{11})+(1-\lambda_1)\log(B_{11}))$, which (since $\log(A_{11}), \log(B_{11})$ are negative and b_{11} is positive) is bounded from below by $\lambda_1 a_{11}\log(a_{11})$. Continuing in much the same way as in the previous proof, we obtain:

$$InfoGain(P) \geq \left[\frac{1}{1+\frac{n_2}{\min\{a_{11}+a_{12},a_{21}+a_{22}\}}}\cdot\frac{1}{1+\frac{n_2}{n_1}}\right] InfoGain(P_1)$$

Recall that $a_{11}+a_{12}=n_1^{left}$ and $a_{21}+a_{22}=n_1^{right}$, thus proving the lemma. ■

We summarize our results with the following theorems:

Theorem 1: Let P be a population of size n, and $\{P_1, P_2, ..., P_k\}$ a partition of P into k subpopulations of sizes $n_1, n_2, ..., n_k$ respectively. Let $G()$ denote the gain function (information gain or Gini index). Then an upper bound on $G(P)$ is given by:

$$G(P) \leq \frac{\sum_{i=1}^{k} n_i G(P_i)}{\sum_{i=1}^{k} n_i}.$$

Note that above theorem is a trivial generalization of lemmas 1 and 3.

Theorem 2: Let P be a population of size n, and $\{P_1, P_2\}$ a partition of P into two subpopulations of sizes n_1, n_2 respectively. Assume that the candidate split divides P_1 into two subsets, P_1^{left} and P_1^{right}, with sizes n_1^{left} and n_1^{right} respectively. Let $G()$ denote the gain function (information gain or Gini index). Then, lower bounds on $G(P)$ is given by:

$$G(P) \geq \frac{G(P_1)}{\left[1+\frac{n_2}{n_1}\right]\left[1+\frac{n_2}{\min\{n_1^{left},n_1^{right}\}}\right]}$$