Consistency of Associative Classification based on Decision Lists

Anonymous Author(s)
Affiliation
Address
email

Abstract

We analyze properties of Associative Classification (AC) procedures based on decision lists [Rivest, 1987] which make them statistically consistent.

1 Introduction

Given an i.i.d. training sequence $D_n := \{(X_n, Y_n), (X_n, Y_n), ..., (X_n, Y_n)\}$ with $X_i \in \mathcal{X}$ and $Y_i \in \{0, 1\}$, associative classification procedures construct partition rules as follows:

Step 1. Generate base classifiers[1] of the form: $\pm\mathbf{1}[X \in A_j]$ satisfying certain empirical properties. The rules themselves are of the type: if $X \in A_j$, then $Y = 1$ or 0. The restriction on the right hand side to classification class attributes makes them a special case of general association rule mining.The candidate sets A_j belong to a collection $\mathcal{A}$. There are many ways to specify this collection including discretizing $\mathcal{X}$ into cells.

Step 2. Order these base classifiers as a decision list to minimize misclassification error over D_n. The decision list thus obtained can be represented as an equivalent partition rule. We will use this representation for analysis.

2 A generic AC procedure

Here is a description of a generic AC procedure. For each subsequent specific model, the points of departure will be the way base classifiers are generated in Step 1.

Step 1. **Association rule or base classifier generation:**

A collection of candidate sets $\mathcal{A}$ is first defined. Each element $A_j \in \mathcal{A}$ is then evaluated using D_n to see if obeys certain properties. These can be the popular minimum support or minimum confidence properties among others. A base classifier is then defined using the selected A_j as $\pm\mathbf{1}[X \in A_j]$. Let the total number of such selected base classifiers be L_n (including possibly a default base classifier).

For example, let the properties which need to be satisfied be exactly minimum support and minimum confidence. Let s% of data D_n have $(X_i, Y_i) \in A_j \times \{1\}$, and c% of D_n have $Y_i = 1$ conditioned on $X_i \in A_j$. If both (s, c) are above predefined threshold percentages, then A_j is selected. It represents the association rule: $X \in A_j \Rightarrow Y = 1$. A_j is called the body of the association rule, $Y = 1$ is the head, s is the support, and c is the confidence.

[1]These are called association rules among other names in the data mining literature. This step is typically called rule mining. We have used the name 'base classifiers' instead to distinguish them from 'decision rule' used in the statistics literature.

Step 2. **Decision list and partition rule representation:**
Order these overlapping sets $\{A_j\}$ into a decision list[2] by minimizing the empirical misclassification error over all permutations possible. Let the best ordered sequence be $\{A_j^*\}$.
To get a partition rule representation for the decision list, create a non-overlapping collection of *cells* $\{B_j\}$ as follows:

$$\begin{aligned} B_1 &= A_1^*, \\ B_2 &= A_2^* \setminus A_1^*, ... \\ B_{L_n} &= A_{L_n}^* \setminus \cup_{j=1}^{L_n-1} A_j^*. \end{aligned}$$

These form a (random or data dependent) partition[3] of $\mathcal{X}$ denoted as $\mathcal{P}_n = \pi_n(D_n)$ where π_n is a n-sample partitioning rule or function. π_n associates every $D_n \in (\mathcal{X} \times \{0,1\})^n$ with a measurable partition of $\mathcal{X}$. The classifier $g_n : \mathcal{X} \mapsto \{0,1\}$ which is a measurable function of X and D_n is then defined as follows:

$$g_n(x) = \begin{cases} 1 & \text{if } x \in B_j \text{ and } \sum_{i:X_i \in B_j} Y_i > \sum_{i:X_i \in B_j} (1 - Y_i) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The above classifier is also called a 'natural' decision rule Devroye et al. [1996] because of the use of majority voting in each cell of the partition. Without loss of generality, it defaults to class 0 in case of ties. We have suppressed the dependency of $\mathcal{P}_n$ and g_n on D_n for simplicity. Note that the ordering of sets $\{A_j\}$ is done using the misclassification error induced by sets $\{A_j\}$ whereas the decision rule g_n depends on $\{B_j\}$ and uses majority voting. We are interested in the consistency of the related decision list classifier which takes the majority vote on slightly different sets:

$$g_n^{AC}(x) = \begin{cases} 1 & \text{if } x \in B_k \subseteq A_j^* \text{ and } \sum_{i:X_i \in A_j^*} Y_i > \sum_{i:X_i \in A_j^*} (1 - Y_i) \\ 0 & \text{otherwise} \end{cases}$$

where A_j^* is the set containing B_k with $j = \inf\{k : x \in A_k^*\}$.

3 Preliminaries for Analysis

For any classification function g, let $L(g) := \mathbb{P}(\{g(X) \neq Y\})$ where (X, Y) is a new pair independent and identically distributed to each element of D_n. Let $L^* =: \inf_{g:\text{measurable}} L(g)$ and random variables $L_n := \mathbb{P}(\{g_n(X) \neq Y\}|D_n)$ and $L_n^{AC} := \mathbb{P}(\{g_n^{AC}(X) \neq Y\}|D_n)$. We want to show at least one of the following:

- Weak consistency for a given distribution: $\lim_{n\to\infty} \mathbb{E}[L_n^{AC}] = L^*$. That is, $\lim_{n\to\infty} \mathbb{P}(\{|L_n^{AC} - L^*| \leq \epsilon\}) = 1$ for all $\epsilon > 0$.
- Strong consistency for a given distribution: $\lim_{n\to\infty} L_n^{AC} = L^*$ with probability one. That is, $\mathbb{P}(\{\lim_{n\to\infty} L_n^{AC} = L^*\}) = 1$. This condition implies weak consistency.
- Universal consistency: For any distribution of (X, Y), strong consistency holds.

4 Model 1: Finite connected union

Here, the candidate sets in Step 1 are derived from the histogram rectangular grids. The modifications to the generic AC procedure are as follows:

Step 1. First discretize $\mathcal{X}$ into rectangular cells denoted by $\{A_j^0\}$. Let the volume of these cells be $K_{1,n}$. When cubic, the side length in each dimension is $\sqrt[d]{K_{1,n}}$. In this case, A_j^0 can be represented in terms of $K_{1,n}$ and a d-tuple of integers $\{k_i\}$ as the product set

[2]For an ordered sequence $\{A_j\}$, the corresponding decision list is: If $X \in A_1$, then b_1,else if $X \in A_2$, then b_2,....

[3]They obey two properties: (a) $B_i \cap B_j = \phi$ if $i \neq j$, and (b) $\cup_j B_j = \mathcal{X}$.

$\Pi_{i=1}^{d}[k_i \sqrt[d]{K_{1,n}}, (k_i+1)\sqrt[d]{K_{1,n}}), k_i \in \mathbb{Z}$. In the general case with d different side lengths $(h_1, ..., h_d)$ the expression is suitably modified.

Let $\mathcal{A}$ be defined as the collection of all A_j generated by finite connected unions[4] of these rectangular blocks which satisfy:

(a) Minimum support requirement with threshold s_n: If $\nu_{j,n}(A_j) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[Y_i = j, X_i \in A_j] \geq s_n$ for $j = 0$ or 1.

(b) Same majority in constituents: If A_j has majority of Y_i equal to 1, then $\nu_{1,n}(A_i^0) > \nu_{0,n}(A_i^0)$ for all $i \in I_j$. Same is the case when the majority is equal to 0.

Let the number of sets selected be $L_n - 1$. Add a default base classifier corresponding to the cell $A_{L_n} := \mathcal{X} \setminus \cup_{j=1}^{L_n-1} A_j$ (this need not be connected).

Step 2. Same as the generic AC procedure. Without loss of generality let B_{L_n} be the region of $\mathcal{X}$ corresponding to the default base classifier.

By controlling the way the three-tuple parameter vector $(K_{1,n}, K_{2,n}, s_n)$ is set, we can achieve consistency. Note that $K_{1,n}$ and $K_{2,n}$ are positive integers whereas minimum support threshold s_n is a real value in the set $[0, 1]$. First we consider the case when the parameter vector only depends on n and not on D_n. Even though these are only functions of n, the partition $\mathcal{P}_n$ is a function of D_n because of Step 2 involving minimization using D_n which necessitates the use of Theorem 5.2 instead of more direct techniques (e.g., [Stone, 1977]).

Theorem 4.1. *For the finite union AC procedure, if*

(a) $\sqrt[d]{K_{1,n}K_{2,n}} \to 0$,

(b) $nK_{1,n} \to \infty$ *and*

(c) $s_n \to 0$ *as* $n \to \infty$,

then

$$L_n^{AC} \to L^*$$

with probability one as $n \to \infty$.

Proof. Has two parts.

Part 1:

We show that the two conditions of Theorem 5.2 are satisfied:

Verifying condition (i): $\frac{1}{n}\log \Delta_n(\mathcal{F}_n^{(M)}) \to 0$.

Since, $\Delta_n(\mathcal{F}_n^{(M)}) = s(\mathcal{A}^{(M)}, n)$ by definition, we focus on $\mathcal{A}^{(M)}$. $\mathcal{A}^{(M)}$ is the collection of all sets which belong to the power set of some partition in $\mathcal{F}_n^{(M)}$. Consider an alternate set $\mathcal{A}_0^{(M)}$ which consists of all sets which belong to the power set of some rectangular grid partition parameterized by $K_{1,n}$. That is, this is the set we obtain when $s_n = 0$ and $K_{2,n} = 1$. For any other values of the parameters $(s_n, K_{2,n})$, $\mathcal{A}^{(M)} \subset \mathcal{A}_0^{(M)}$. Thus, the n-th shatter coefficients of these two collections are related as: $s(\mathcal{A}^{(M)}, n) \leq s(\mathcal{A}_0^{(M)}, n)$.

Computing the shatter coefficient of $\mathcal{A}_0^{(M)}$ can be done by computing two related quantities:

- The maximum number of cells m_n in a rectangular grid partition restricted to a ball with radius M, which is $\frac{c}{k_{1,n}}$ where c is a function of M and dimension d.
- The largest number of distinct partitions of any n point subset of $\mathbb{R}^d$ that can be induced by the rectangular grid partitions Δ^*. This is upper bounded by the same quantity for the collection of flexible grid partitions[5] since rectangular grid partitions are a subset of flexible

[4]Each selected A_j is of the form $\cup_{i \in I_j} A_i^0$ and for all $j, |I_j| \leq K_{2,n}$. Each constituent A_i^0 shares at least one $d-1$ dimensional face with some other constituent.

[5]The distance between two parallel hyperplanes can be flexible in contrast to rectangular grid partitions where this is fixed.

grid partitions for a fixed number of cells. The largest number of distinct partitions of any n points induced by flexible grid partitions with m_n cells is $\binom{n+m_n}{n}$. This follows from the equivalence to the number of ways of putting n points in $n + m_n$ boxes (also known popularly as the stars and bars problem in combinatorics).

We make use of the identity $s(\mathcal{A}_0^{(M)}, n) \leq 2^{m_n}\Delta^*$ (see Lemma 21.1 in [Devroye et al., 1996], Lemma 13.1 in [Györfi et al., 2002]) to get

$$\Delta_n(\mathcal{F}_n^{(M)}) = s(\mathcal{A}^{(M)}, n) \leq s(\mathcal{A}_0^{(M)}, n) \leq 2^{m_n}\binom{n+m_n}{n},$$

where $m_n = \frac{c}{k_{1,n}}$. Now,

$$\frac{1}{n}\log\Delta_n(\mathcal{F}_n^{(M)}) = \frac{m_n}{n}\log 2 + \frac{1}{n}\log\binom{n+m_n}{n}$$

Since $nK_{1,n} \to \infty$ from condition (b) of the theorem, $\frac{m_n}{n} = \frac{c}{nK_{1,n}} \to 0$. Also,

$$\begin{aligned}\frac{1}{n}\log\binom{n+m_n}{n} &\approx \frac{1}{n}\left[(n+m_n)\log(n+m_n) - m_n\log m_n - n\log n\right] \\ &= \frac{1}{n}\left[n\log(\frac{n+m_n}{n}) + m_n\log(\frac{n+m_n}{m_n})\right] \\ &= \log(1+\frac{m_n}{n}) + \frac{m_n}{n}\log(1+\frac{n}{m_n}),\end{aligned}$$

where, the first equation follows from using the Stirling approximation. Since $\frac{m_n}{n} \to 0$, $\log(1 + \frac{m_n}{n}) \to 0$. Further, $\lim_{n\to\infty}\frac{m_n}{n}\log(1+\frac{n}{m_n}) = \lim_{x\to 0} x\log(1+\frac{1}{x}) = 0$. Thus, condition (i) of Theorem 5.2 is satisfied.

Verifying condition (ii): We show that all cells of the random partition $\mathcal{P}_n$ shrink. In particular, the diameter of the largest cell $\sup_{B_j\in\mathcal{P}_n\setminus B_{L_n}}\sup_{x_1,x_2\in B_j}\|x_1 - x_2\| \leq \sqrt{d}\sqrt[d]{K_{1,n}K_{2,n}} \to 0$ as $n \to \infty$ by condition (a) of the theorem. Thus, condition (ii) of Theorem 5.2 is satisfied for $\{B_j\}_{j=1}^{L_n-1}$. This is because, for each closed ball S_M and scalar $\gamma > 0$, there exists a finite n_0 dependent on γ such that for all $n > n_0$, $\text{diam}(B_j \cap S_M) \leq \text{diam}(B_j) \leq \gamma$. This is achieved by picking n_0 such that $\sqrt{d}\sqrt[d]{K_{1,n_0}K_{2,n_0}} \leq \gamma$. Thus,

$$\lim_{n\to\infty}\mu(\{x : \text{diam}(B_j \cap S_M) > \gamma\}) = 0.$$

The diameter of the last cell B_{L_n} is controlled via the parameter s_n. In particular, condition (c) on $s_n \to 0$ ensures that $\text{diam}(B_{L_n}) \to 0$. That is, for all $\gamma > 0$, there exists a small enough s_n such that the measure of the default base classifier (which depends on s_n)

$$\mu\{x : \text{diam}(B_{L_n} \cap S_M) > \gamma\} \leq \mu\{x : \text{diam}(B_{L_n}) > \gamma\} = 0.$$

Satisfying the two conditions of Theorem 5.2 ensures that $L_n \to L^*$ with probability one.

Part 2:
We show that $L_n = L_n^{AC}$ for all D_n. That is, the two random variables are the same. This is true if $g_n^{AC}(x) = g_n(x)$ point-wise. Consider a cell B_j. Then $g_n^{AC}(x) = g_n(x)$ for all $x \in B_j$ if

$$\left[\sum_{i:X_i\in A_k^*} Y_i > \sum_{i:X_i\in A_k^*}(1-Y_i)\right] \Leftrightarrow \left[\sum_{i:X_i\in B_j} Y_i > \sum_{i:X_i\in B_j}(1-Y_i)\right],$$

where A_k^* is the set containing B_j with $k = \inf\{i : x \in A_i^*\}$ in the optimal ordering done in Step 2 of the AC procedure.

For the forward direction, let $\left[\sum_{i:X_i\in A_k^*} Y_i > \sum_{i:X_i\in A_k^*}(1-Y_i)\right]$. We then make use of the special property outlined in Step 1 of the AC procedure: If A_k^* has majority of Y_i equal to 1, then $\nu_{1,n}(A_i^0) > \nu_{0,n}(A_i^0)$ for all $i \in I_{k^*}$. Since $B_j \subseteq A_k^*$, it is also a union of some of these base cells

and $\nu_{1,n}(B_j) > \nu_{0,n}(B_j)$ which is the same as $\left[\sum_{i:X_i \in B_j} Y_i > \sum_{i:X_i \in B_j}(1-Y_i)\right]$. The same argument also holds in the reverse direction. Thus, the majority rule is not affected when the cells change from A_k^* to $B_j \subseteq A_k^*$. □

Remark 4.2. When $s_n = 0$ and $K_{2,n} = 1$, one of the optimal solutions of the AC procedure above is the rectangular histogram rule. An upper bound on $K_{2,n}$ is: $K_{2,n} \leq \frac{c}{K_{1,n}}$. If $K_{2,n}$ is fixed arbitrarily, the theorem holds but in the most generality, we can allow $K_{2,n}$ to grow as $n \to \infty$ according to condition (a), but not too quickly. If $s_n = 0$ at $n = \infty$, the rectangular histogram rule is one of the optimal solutions for Step 2 of the AC procedure, and hence is consistent. Further, when $s_n = 0$ for any n, there is no default rule and all cells where the empirical measure is positive are selected for Step 2. In fact, there is no need for optimization in Step 2 and we can pick the rectangular histogram rule directly.

Remark 4.3. Condition (a) in the theorem is only letting all A_j and consequently B_j increase in diameter at a certain rate so that local changes can still be detected. Condition (c) is controlling the size of the default base classifier $A_{L_n} = B_{L_n}$. Condition (b) is ensuring that the log of the n-th shatter coefficient of the family of partitions is growing slower than n. Overall, this proof relies on the fact that the AC partitions retain properties of their histogram skeleton to attain consistency. *In particular, searching for the optimal permutation of sets $\{A_j\}$ has no effect on consistency. In practice however, they can vastly affect performance.*

5 Master Theorems

Let μ be the measure on $\mathcal{X}$ and μ_n be the related empirical measure. Let $\text{diam}(B) = \sup_{x,y \in B} \|x - y\|$. Let $B(x) = B_j$ such that $x \in B_j$ and $N(x) = n\mu_n(B(x)) = \sum_{i=1}^n \mathbf{1}[X_i \in B(x)]$. Sets $\{B_j\}$ depend on D_n though this dependence is suppressed in the notation. For $M > 0$, let $\mathcal{P}^{(M)}$ be a restriction of $\mathcal{P}$ to a closed ball of radius M centered at the origin. For each $M < \infty$, let $|\mathcal{P}^{(M)}| < \infty$. Let $\mathcal{F}^{(M)} = \{\mathcal{P}^{(M)} : \mathcal{P} \in \mathcal{F}\}$. Let $\mathcal{B}(\mathcal{P}^{(M)}) = \{A : A \in$ power set of $\mathcal{P}^{(M)}\}$ and $\mathcal{A}^{(M)} = \{A \in \mathcal{B}(\mathcal{P}^{(M)}) : \mathcal{P}^{(M)} \in \mathcal{F}^{(M)}\}$.

Definition 5.1. Shatter coefficient and Vapnik-Chervonenkis (VC) dimension for a family of partitions $\mathcal{F}$: The n-th shatter coefficient $s(\mathcal{A}, n)$ of a collection $\mathcal{A}$ of measurable sets is the maximal number of different subsets of n points that can be picked out by the class of sets $\mathcal{A}$. The largest integer k for which $s(\mathcal{A}, k) = 2^k$ is denoted $V_{\mathcal{A}}$ and is called the Vapnik-Chervonenkis dimension of class $\mathcal{A}$. For a family of partitions $\mathcal{F}^{(M)}$, define $\Delta_n(\mathcal{F}^{(M)}) = s(\mathcal{A}^{(M)}, n)$.

Theorem 5.2. *(Lugosi and Nobel [1996]) Let $\{\pi_1, \pi_2, ...\}$ be a fixed sequence of partitioning rules, and for each n let $\mathcal{F}_n$ be the collection of partitions associated with the n-sample partitioning rule π_n. If*

(i) For each $M < \infty$

$$\lim_{n\to\infty} \frac{\log(\Delta_n(\mathcal{F}_n^{(M)}))}{n} = 0$$

and

(ii) for all balls S_M and all $\gamma > 0$,

$$\lim_{n\to\infty} \mu(\{x : \textit{diam}(A(x) \cap S_M) > \gamma\}) = 0$$

with probability one,

then g_n corresponding to π_n satisfies $L(g_n) \to L^*$ with probability one.

In the above theorem, condition (1) deals with richness of the class of partitions of $\mathcal{X}$ and condition (2) says that the rule should eventually make local decisions. The proof involves splitting the difference between the error of g_n and g^* into estimation and approximation error: estimation error is bounded using VC theory and approximation error is bounded using cell diameter shrinkage arguments. Since g_n and g^* are in plugin form, they show L_1-convergence of the approximating function (which depends on D_n) to the a posteriori probability $\eta(x)$. $\mathcal{P}_n$ can depend arbitrarily on D_n. It extends earlier data independent results for a variety of partition based classifiers.

6 Related work

We know that certain tree classifiers, neural networks, generalized linear classifiers, structural risk minimization procedures, K-nearest neighbor, kernel classifiers and variants of histogram rules are statistically consistent to various degrees. Motivated by this, we look at associative classification procedures which are popular in computer science and data mining and address statistical consistency questions for the same. These procedures are broadly of two types. The first involves generating a weighted combination of class association rules while the second involves forming decision lists. The former lose interpretability and will not be dealt with here[6]. Among the latter decision list procedures, we assess a set of heuristic techniques which are used in practice and provide consistency results for the learned classifiers.

A decision list is structurally a decision tree with only one long path. Decision lists (and histogram rules) can be though of as a bottom up approach to constructing partition classifiers compared to decision tree classifiers which are mostly top down and recursive in nature. While both decision list based classifiers and tree classifiers are interpretable in general, there are key differences. For instance, the latter afford computational gains because of the recursive nesting nature whereas the former allow for enhanced interpretability since they have a single path in the resulting tree. Both associative classification procedures and decision tree classifiers belong to the class of partitioning classifiers and can handle heterogeneous inputs easily compared to other model representations like half-spaces, neural networks, graphical models [Domingos, 2012] where suitable preprocessing is needed. Associative classification procedures can be compared to the more closely related greedy rule induction learners like IREP and RIPPER.

In addition to much desired interpretability, studying the consistency of AC procedures is further motivated by the fact that these techniques reuse the rich rule mining technology already used by data mining practitioners. Table 1 summarizes a few popular AC procedures[7]. The first two are of the weighted majority type whereas the last method is of the decision list type.

Procedure	Rule generation	Ordering
CBA [Ma, 1998]	Apriori	Database coverage heuristic, interestingness measures
CMAR [Li et al., 2001]	FPgrowth	Database coverage heuristic, interestingness measures
ORC [Bertsimas et al., 2012]	Mixed-integer optimization	Minimize empirical error over orderings

Table 1: Some AC procedures. Most use Apriori [Agrawal et al., 1994] or FPgrowth [Han et al., 2000] for generating rules and either take a weighted majority or order them using support and confidence values (this can be done greedily or by solving an optimization formulation).

7 Conclusion

We presented the first results proving the consistency of associative classification procedures which are based on decision lists.

References

Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, volume 1215, pages 487–499, 1994.

Dimitris Bertsimas, Allison Chang, and Cynthia Rudin. ORC: Ordered rules for classification a discrete optimization approach to associative classification. 2012.

[6]Their consistency results can be established using ensemble method consistency results **?**

[7]CBA: Classification Based on Associations, CMAR: Classification based on Multiple Rules, ORC: Ordered Rules for Classification. Apriori has been cited more than 14000 times in the data mining and related literature.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. New York: Springer, 1996.

Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer, 2002.

Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In *ACM SIGMOD Record*, volume 29, pages 1–12. ACM, 2000.

Wenmin Li, Jiawei Han, and Jian Pei. CMAR: Accurate and efficient classification based on multiple class-association rules. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 369–376. IEEE, 2001.

Gábor Lugosi and Andrew Nobel. Consistency of data-driven histogram methods for density estimation and classification. *The Annals of Statistics*, 24(2):687–706, 1996.

Bing Liu Wynne Hsu Yiming Ma. Integrating classification and association rule mining. In *Proceedings of the 4th*, 1998.

Ronald L Rivest. Learning decision lists. *Machine learning*, 2(3):229–246, 1987.

Charles J Stone. Consistent nonparametric regression. *The annals of statistics*, 5(4):595–620, 1977.