2002年日本オペレーションズ・リサーチ学会
秋季研究発表会

データ分類におけるノイズ量の評価について

02502524 京都大学 *原口和也 HARAGUCHI Kazuya
01001374 京都大学 茨木俊秀 IBARAKI Toshihide

1 はじめに

データ解析, 知識発見の話題において, データの所属するクラスを判定する, いわゆる分類問題は重要な問題の一つである. 本研究で扱う個々のデータ $x=(x_1,x_2,\cdots,x_n)$ は n 個の属性に対して値を持つ n 次元ベクトルとして表現され, 正もしくは負いずれかのクラスに属する. データ x の属するクラスを $\Omega(x)$ と書き, 1 (正クラス) あるいは 0 (負クラス) の値をとる. 分類問題ではクラスが既知であるデータ (以下, 既知データ) の集合から知識を抽出し, それに基づいてクラスの未知なデータ (以下, 未知データ) のクラスを予測することが求められる. データのクラスを予測するシステムを分類器 (classifier) と呼ぶ. 分類器の性能は未知データのクラス予測における誤り率によって評価される. 分類器の例としてはニューラルネットワーク, 決定木などが知られている.

実在するデータ集合の多くは, データの要素値および所属クラスに誤りを含んでいる. このような誤ってデータの誤りをノイズと呼ぶ. 既知データの集合の信頼性はそれに含まれるノイズの量と反比例している. 本研究では既知データの集合に含まれるノイズの量を定量的に把握する手法を考察する.

2 用語の定義

データ x の各要素は定義域を持つので, それにしたがって定義されるすべてのデータの集合を S_∞ とする. 現実に手にいれることのできる既知データの集合 S_0 はこの一部分である. すなわち $S_\infty \backslash S_0$ は未知データの集合である.

データ集合 X を訓練集合とし, 分類器決定アルゴリズム C によって決定される分類器を C_X とする. 訓練集合には既知データのみ用いられる ($X \subseteq S_0$). C_X が予測したデータ x のクラスを $C_X(x)$ と書くと, 試験集合 X' のクラス予測における誤り率 $E(C_X,X')$ は

$$E(C_X,X') = \frac{|\{x \mid C_X(x) \neq \Omega(x), x \in X'\}|}{|X'|}$$

と計算される. 試験集合は誤り率の測定対象なので, 目的に応じて既知データ, 未知データのいずれを用いることも可能である ($X' \subseteq S_\infty$).

3 データ集合に含まれるノイズの量

既知データの集合 S_0 が与えられたとき, それに含まれるノイズの量を推定することを考える. この目的のために, 分類器によるクラス予測の揺らぎに着目する. 以下, 分類器決定アルゴリズム C は与えられているものとする. 予測の揺らぎとは, 試験集合 $X'(=S_\infty \backslash S_0)$ に含まれるデータ x' のクラスを分類器によって予測するとき, 訓練集合 X に依存して, 予測クラス $C_X(x')$ が異なる現象を言う. X としては S_0 のサイズ m の部分集合をランダムに選び, $X_1, X_2, ..., X_T$ とする. 予測の揺らぎが発生するとき, データ x' に対する添字集合 $I(x')=\{i \mid C_{X_i}(x') \neq \Omega(x'), i=1,2,...,T\}$ は空ではない. すなわち, $|I(x')|/T$ は x' の予測において上記のように X を選んで分類器を決定した場合の, 予測の誤り確率となる. $|I(x')|/T$ を $x' \in X'$ 全体について平均をとった値を E_m とする.

上記のように分類器を構成するとき, 予測の誤りの原因として以下の 2 つが挙げられる.

【原因 1】各 X_i が S_0 と離れた固有の傾向を持つこと.

【原因 2】各 X_i がノイズを含んでいること.

原因 1 の揺らぎは X のサイズ m を大きくすれば減少すると考えられることから, 以下のようにノイズの量 (原因 2) を推定することを考える.

仮定 E_m の値は, m が小さい場合には原因 1,2 の両方の影響を受け, 大きい場合には原因 1 の影響が小さくなる. 特にある m_0 以上の m に

おいて $E_m \approx E^*$ ならば, E^* は S_0 のノイズ量による誤り率を表している.

4 数値実験

E_m を明示的に求めることは実際にはできない. なぜなら (1) サイズ m の訓練集合の総数 T, および (2) 試験集合 X' のサイズは共に膨大であり, また (3) 未知データ $x' \in X'$ のクラスを知ることはできないからである. 本研究ではこの 3 つの問題点について以下のように考え, E_m の下界値 $\underline{E}_m$ を求めることを試みる.

(1) サイズ m の訓練集合の個数 T をある程度大きくすれば

$$E_m = \frac{1}{T}\sum_{i=1}^{T} E(C_{X_i}, X') \approx \frac{1}{T_0}\sum_{i=1}^{T_0} E(C_{X_i}, X')$$

と考えられる. ただし T_0 は S_0 のサイズ m の部分集合の総数である. 今回は, 予備実験の結果に基づいて $T = 500$ とする.

(2) 誤り率の計算に当たって X' に含まれるデータの全てを調べることは困難である. しかし実際には充分大きな $X'' \subseteq X'$ を用いると

$$E(C_{X_i}, X'') \simeq E(C_{X_i}, X')$$

の成立が期待できる. 今回は $|X''| = 5000$ とする.

未知データ x' は以下のように得る. まず, S_0 からランダムにデータ x を選び, パラメータ β を定める ($0 \le \beta \le 1$). x が持つ n 個の属性値のうち, ランダムに選んだ $\lfloor \beta n \rfloor$ 個の属性値を, 各属性の定義域における一様分布の下でランダムに変化させ, 得られたデータを x' とする.

(3) 上のように生成された x' のクラスを知ることはできないため, x' のクラスは C_{X_i} による全 T 回の予測のうち回数の多かった方のクラスと見なす. このように x' のクラスを定めると本来の誤り率の下界値が求められるので, この値を X'' の全体について平均をとり, E_m の下界値 $\underline{E}_m$ とする.

UCI Machine Learning Repository に存在する複数のベンチマーク用データ集合について, $\underline{E}_m$ を計算した. そのうちデータ集合 CAR($|S_0| = 1728$, $n = 6$) と HABER($|S_0| = 306$, $n = 3$) に関する結果をそれぞれ図 1,2 に掲げる. なお, 分類器決定アルゴリズムとしては C4.5 を使用している [1]. 図はそれぞれ横軸に $m/|S_0|$, 縦軸に $\underline{E}_m$ をとっている. 図中の各折れ線は試験集合生成のパラメータ β の値に対応している. m を大きくしたとき, $\underline{E}_m$ が一定値収束することなく減少していく CAR はノイズが少ないと考えられる. 一方 HABER では $\underline{E}_m$ は $m/|S_0|$ が 12.5% あたりで一定範囲値に収束しており, この誤り率の部分はデータに含まれるノイズを反映したものと考えられる.

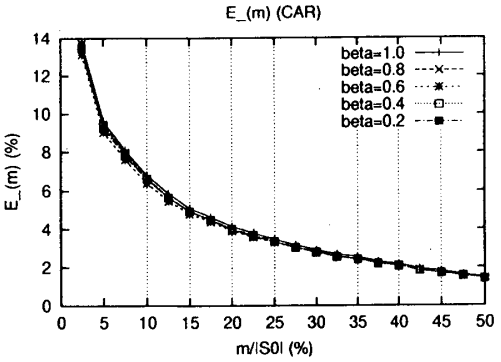

図 1: CAR に対する $\underline{E}_m$ の計算結果

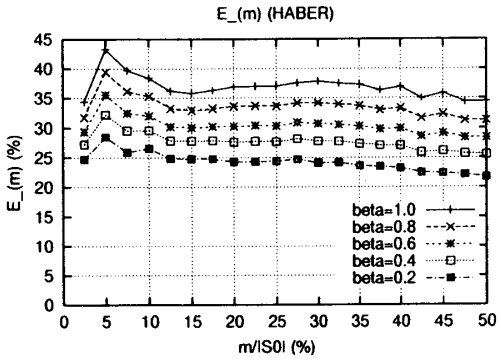

図 2: HABER に対する $\underline{E}_m$ の計算結果

次に, β が大きいほど, S_0 から「遠い」試験集合と考えられる. β の大小で $\underline{E}_m$ の変わらない CAR では S_0 と S_∞ の違いは少なく (実際, $S_0 = S_\infty$), HABER ではそうでないことがわかる.

5 おわりに

分類問題の立場からデータに含まれるノイズ量を知るための研究例は極めて少ない. したがって今後の発展が期待される一方, より洗練された理論の整備も課題となっている.

参考文献

[1] J.R.Quinlan, *C4.5: Programs for machine learning*, Morgan Kaufmann, 1993.

[2] 森下真一, 宮野悟, 発見科学とデータマイニング bit 5 月号別冊, 2000.