PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1076 2005

TAKSONOMIA 12
Klasyfikacja i analiza danych – teoria i zastosowania

Dorota Rozmus
Akademia Ekonomiczna w Katowicach

DEKOMPOZYCJA BŁĘDU KLASYFIKACJI W ZAGREGOWANYCH MODELACH DYSKRYMINACYJNYCH

1. Wstęp

Pojęcie dekompozycji błędu predykcji wywodzi się z regresji i polega na rozbiciu spodziewanej wartości błędu na trzy składowe, tj.: szum (ang. *noise*), obciążenie (ang. *bias*) i wariancję (ang. *variance*). Podjęto także próby przeniesienia idei dekompozycji do zagadnienia klasyfikacji. Agregacja modeli [Breiman 1996] ma na celu obniżenie wartości błędu przez redukcję albo obciążenia, albo wariancji, albo obydwu tych wielkości jednocześnie. Jednak, aby porównać różne metody agregacji modeli dyskryminacyjnych ze względu na ich wpływ na wartości obciążenia i wariancji, należy być bardzo ostrożnym, bowiem różne sposoby dekompozycji błędu klasyfikacji zaproponowane w literaturze przedmiotu, dają różne wartości tych wielkości.

2. Zagregowany model dyskryminacyjny

W danym zbiorze uczącym: $\{(\mathbf{x}_1, y_1),(\mathbf{x}_2, y_1),...,(\mathbf{x}_N, y_N)\}$, gdzie $\mathbf{x}_i$ to wektor zmiennych objaśniających, a y to zmienna objaśniana, należy określić reguły pozwalające podać wartość zmiennej objaśnianej na podstawie znajomości wartości zmiennych objaśniających. Mamy tu do czynienia z zagadnieniem dyskryminacji (lub „uczenia z nauczycielem"), a problem określania reguł zależności jest rozwiązywany przez tzw. algorytm uczący. Buduje się zatem pewien model f, który w

kolejnym etapie zostanie wykorzystany do predykcji wartości zmiennej zależnej dla nowego obiektu **x**:

$$\hat{y} = f(\mathbf{x}). \tag{1}$$

Niech y oznacza rzeczywistą wartość zmiennej objaśnianej dla określonego obiektu **x**. Oceny jakości predykcji dokonujemy przez obliczenie wartości błędu:

$$Err = E_y[L(y, \hat{y})]. \tag{2}$$

Funkcja straty $L(y, \hat{y})$ mierzy koszt tego, że model błędnie przewidział wartość zmiennej objaśnianej. W regresji najczęściej stosowaną jest kwadratowa funkcja straty:

$$L(y, \hat{y}) = (y - \hat{y})^2, \tag{3}$$

a dla klasyfikacji – zero-jedynkowa funkcja straty:

$$L(y, \hat{y}) = \begin{cases} 0, & \text{gdy } y = \hat{y}, \\ 1, & \text{gdy } y \neq \hat{y}. \end{cases} \tag{4}$$

W dalszych rozważania rozpatrywany będzie błąd modelu zagregowanego. W związku z tym, że ten sam algorytm uczący będzie budował różne modele dla różnych prób uczących, wartość błędu będzie zależała od tego, jakie obiekty znalazły się w zbiorze uczącym. Zależność tę można wyeliminować przez zastosowanie wielu zbiorów uczących. Jeśli mamy zatem wyjściową próbę uczącą T, generujemy z niej metodą losowania obiektów kolejne próby uczące: $Z = \{T_1, T_2, ..., T_k\}$ i na podstawie każdej z tych prób budujemy osobny model. Próby te w ostatnim etapie będą połączone różnymi operatorami agregacji w jeden spójny model. Okazuje się, że błąd modelu zagregowanego, obliczony według formuły (5), jest mniejszy niż błąd poszczególnych modeli:

$$Err_{agr} = E_Z\{Err\} = E_Z\{E_y[L(y, \hat{y})]\}. \tag{5}$$

Błąd modelu zagregowanego podlega dekompozycji na trzy składowe. Pozwala to zbadać, w jaki sposób na jego wartość wpływają takie czynniki, jak np.: liczba modeli, liczebność poszczególnych prób uczących, sposób konstrukcji próby czy parametry samego algorytmu uczącego.

3. Dekompozycja błędu predykcji w regresji

Aby przejść do zagadnienia dekompozycji, na początku należy zdefiniować pojęcie modelu optymalnego i modelu zagregowanego.

Generalnie rzecz ujmując, zmienna zależna y jest niedeterministyczną funkcją wartości zmiennych objaśniających, zatem niemal niemożliwe jest osiągnięcie zerowej wartości funkcji straty. Jednakże dla określonego problemu i określonej postaci funkcji straty można zdefiniować najlepszy możliwy model, zapewniający optymalną predykcję wartości zmiennej objaśnianej, oznaczoną przez y_*. Opty-

malną predykcją y_* dla określonego wektora wartości zmiennych objaśniających **x** jest predykcja, która minimalizuje wartość oczekiwaną błędu $E_y[L(y, y_*)]$. Wartość tę określa się w następujący sposób:

$$y_* = \arg\min_{y'} E_y[L(y, y')]. \tag{6}$$

Predykcja modelu zagregowanego, oznaczona przez y_m, dla dowolnej postaci funkcji straty L i zbioru prób uczących Z jest zdefiniowana jako:

$$y_m = \arg\min_{\hat{y}'} E_Z[L(\hat{y}, \hat{y}')]. \tag{7}$$

Model zagregowany zatem przewiduje taką wartość zmiennej objaśnianej $\hat{y}'$, której średni błąd w stosunku do wszystkich predykcji wartości zmiennej zależnej, otrzymanych na podstawie zbioru prób uczących $E_Z[L(\hat{y}, y_m)]$, jest minimalny.

Szum dla obiektu **x** definiuje się w sposób następujący:

$$N(\mathbf{x}) = E_y[L(y, y_*)]. \tag{8}$$

Jest to błąd wynikający z różnicy między predykcją optymalną a rzeczywistą wartością zmiennej objaśnianej. Szum jest nieuniknionym elementem składowym błędu, stanowi hipotetyczną dolną granicę błędu predykcji i jest niezależny od algorytmu uczącego.

Obciążeniem algorytmu uczącego dla obserwacji **x** jest systematyczny błąd spowodowany różnicą między predykcją optymalną a predykcją modelu zagregowanego:

$$B(\mathbf{x}) = L(y_*, y_m). \tag{9}$$

Wariancja obserwacji **x** to średni błąd spowodowany przez różnice między predykcją uzyskaną na podstawie poszczególnych zbiorów uczących a predykcją modelu zagregowanego:

$$V(\mathbf{x}) = E_Z[L(y_m, \hat{y})]. \tag{10}$$

Dekompozycja błędu predykcji rozbija spodziewaną wartość błędu na trzy składowe, tj.: szum, obciążenie i wariancję:

$$Err_{agr} = E_Z\{E_y[L(y, \hat{y})]\} = N(\mathbf{x}) + B(\mathbf{x}) + V(\mathbf{x}). \tag{11}$$

Rozważmy zagadnienie dekompozycji błędu predykcji w regresji z uwzględnieniem kwadratowej funkcji straty. Niech dany będzie obiekt **x**, dla którego prawdziwa wartość zmiennej objaśnianej wynosi y, i rozważmy algorytm uczący, który na podstawie zbioru prób uczących Z przewiduje wartość $\hat{y}$ dla tego obiektu. Wówczas spodziewana wartość błędu dana będzie wzorem:

$$Err_{agr} = E_Z\{E_y[(y - \hat{y})^2]\}. \tag{12}$$

Predykcja modelu optymalnego to wartość oczekiwana zmiennej objaśnianej:

$$y_* = E_y[y]. \tag{13}$$

Za predykcję modelu zagregowanego w zagadnieniu regresji przyjmuje się średnią arytmetyczną z wartości zmiennej zależnej przewidywanych przez modele zbudowane na zbiorach uczących ze zbioru Z:

$$y_m = E_Z[\hat{y}]. \tag{14}$$

Szum, obciążenie i wariancję dla obiektu $\mathbf{x}$ definiuje się za pomocą równości:

$$N(\mathbf{x}) = E_y\left[(y - y_*)^2\right], \tag{15}$$

$$B(\mathbf{x}) = (y_* - y_m)^2, \tag{16}$$

$$V(\mathbf{x}) = E_Z\left[(y_m - \hat{y})^2\right]. \tag{17}$$

Suma trzech tak zdefiniowanych elementów daje w efekcie spodziewaną wartość błędu:

$$Err_{agr} = E_Z\left\{E_y\left[(y - \hat{y})^2\right]\right\} = N(\mathbf{x}) + B(\mathbf{x}) + V(\mathbf{x}). \tag{18}$$

4. Koncepcje dekompozycji błędu klasyfikacji

W klasyfikacji spodziewany błąd predykcji to prawdopodobieństwo błędnej klasyfikacji:

$$Err_{agr} = E_Z\{E_y[1(y \neq \hat{y} \mid \mathbf{x})]\} = P(y \neq \hat{y} \mid \mathbf{x}) = 1 - \sum_y P_y(y \mid \mathbf{x}) \cdot P_Z(\hat{y} \mid \mathbf{x}). \tag{19}$$

Optymalna predykcja w klasyfikacji przypisuje obiektowi $\mathbf{x}$ najbardziej prawdopodobną klasę, zgodnie z warunkowym rozkładem prawdopodobieństwa przynależności do klas:

$$y_* = \arg\max_y P_y(y \mid \mathbf{x}). \tag{20}$$

W zagadnieniu klasyfikacji odpowiednikiem predykcji modelu zagregowanego z regresji jest taka wartość zmiennej objaśnianej, która każdemu obiektowi $\mathbf{x}$ przypisuje klasę, najczęściej wskazywaną przez modele na podstawie różnych zbiorów uczących ze zbioru Z:

$$y_m = \arg\max_{\hat{y}} P_Z(\hat{y} \mid \mathbf{x}). \tag{21}$$

Szum w klasyfikacji definiuje się jako:

$$N(\mathbf{x}) = 1 - P_y(y_* \mid \mathbf{x}). \tag{22}$$

Obciążenie definiowane jest przez funkcję wskaźnikową:

$$B(\mathbf{x}) = 1(y_* \neq y_m \mid \mathbf{x}). \tag{23}$$

Zatem w klasyfikacji są obciążone wszystkie te obiekty, dla których predykcja modelu zagregowanego nie zgadza się z predykcją optymalnego.

Wariancja dla obiektu $\mathbf{x}$ w klasyfikacji jest obliczona ze wzoru:

$$V(\mathbf{x}) = 1 - P_Z(y_m \mid \mathbf{x}). \tag{24}$$

Okazuje się jednak, że w klasyfikacji przez analogię do regresji, sumując wartość szumu, obciążenia i wariancji nie otrzymamy wartości błędu klasyfikacji:

$$Err_{agr} = E_Z\{E_y[1(y \neq \hat{y} \mid \mathbf{x})]\} \neq N(\mathbf{x}) + B(\mathbf{x}) + V(\mathbf{x}). \tag{25}$$

Z tego też względu w literaturze pojawiły się liczne propozycje definiowania obciążenia i wariancji w zagadnieniu klasyfikacji. Ta różnorodność wynika m.in. z braku zgody co do własności, którymi powinny się charakteryzować te wielkości w zero-jedynkowej funkcji straty.

Jedną z pierwszych propozycji była dekompozycja Konga i Diettericha [1995]. Można ją było zastosować tylko w zbiorach danych pozbawionych szumu, ponieważ nie uwzględniała składnika będącego miarą wartości szumu.

Kong i Dietterich zaproponowali, by spodziewany błąd klasyfikacji definiować jako:

$$Err_{agr} = 1 - P_Z(y = \hat{y} \mid \mathbf{x}). \tag{26}$$

Obciążenie dla obserwacji $\mathbf{x}$ jest albo równe zeru, gdy model dokonał prawidłowej klasyfikacji, albo 1 w przeciwnym razie:

$$B(\mathbf{x}) = \begin{cases} 0, & \text{gdy } y_m = y, \\ 1, & \text{gdy } y_m \neq y. \end{cases} \tag{27}$$

Wariancję zaś – jako różnicę między spodziewanym błędem klasyfikacji a obciążeniem:

$$V(\mathbf{x}) = Err_{agr} - B(\mathbf{x}). \tag{28}$$

Tak zdefiniowana wariancja, w obserwacjach obciążonych, może przyjąć wartość ujemną.

Tibshirani [1996] definiuje obciążenie w sposób następujący:

$$B(\mathbf{x}) = \left[1 - P_Z(y_m \mid \mathbf{x})\right] - \left[1 - P_y(y_* \mid \mathbf{x})\right]. \tag{29}$$

Wielkość ta niekoniecznie musi być nieujemna, dlatego, aby zapewnić dodatnią wartość dekompozycji błędu klasyfikacji, wyrażenie będące miarą wariancji definiuje jako różnicę między oczekiwanym błędem klasyfikacji a błędem modelu zagregowanego. Ze względu na to, że wielkość ta może również przyjmować wartość ujemną nazywa się ją efektem agregacji:

$$AE(\mathbf{x}) = Err - \left[1 - P_Z(y_m \mid \mathbf{x})\right]. \tag{30}$$

Suma tych dwóch elementów, tj. obciążenia i efektu, powiększona o wartość szumu daje wartość błędu klasyfikacji.

Kohavi i Wolpert [1995] zaproponowali zupełnie inną dekompozycję, w swojej formie zbliżoną do dekompozycji błędu kwadratowego w regresji. Nie łączy się ona z takimi problemami, jak: potencjalnie ujemna wariancja albo tylko dwie możliwe wartości obciążenia, tj. 0 lub 1.

Zamiast obciążenia, mającego postać (23), autorzy ci stosują kwadrat obciążenia, który przyjmuje wartości z przedziału między 0 a 1. Jest on definiowany w sposób następujący:

$$B^2(\mathbf{x}) = \frac{1}{2}\sum_y \left[P_y(y|\mathbf{x}) - P_Z(\hat{y}|\mathbf{x})\right]^2. \tag{31}$$

Wariancja jest nieujemną wartością liczoną według formuły:

$$V(\mathbf{x}) = \frac{1}{2}\left[1 - \sum_y P_Z(\hat{y}|\mathbf{x})\right]^2. \tag{32}$$

Autorzy tej dekompozycji proponują także swoją definicję szumu:

$$N^2(\mathbf{x}) = \frac{1}{2}\left[1 - \sum_y \left(P_y(y|\mathbf{x})\right)^2\right]. \tag{33}$$

Suma trzech tak zdefiniowanych elementów stanowi dekompozycję błędu klasyfikacji według Kohavi i Wolperta.

Breiman [1996] zaproponował dekompozycję dla zero-jedynkowej funkcji straty, obliczoną dla wszystkich obiektów w zbiorze rozpoznawanym. Nie zdefiniował jednak, czym jest obciążenie i wariancja dla pojedynczego obiektu **x**. Dekompozycja ta została oparta na przekonaniu, że tylko obciążone obiekty powinny brać udział w definiowaniu obciążenia i tylko nieobciążone obiekty powinny wpływać na wartość wariancji. Algorytm jest nieobciążony dla danego obiektu **x**, jeżeli klasyfikacja optymalna jest taka sama jak predykcja modelu zagregowanego.

Niech $\boldsymbol{B}$ będzie zbiorem wszystkich tych obserwacji, w których model jest obciążony, przez $\boldsymbol{U}$ oznaczmy zaś zbiór obserwacji, w których model jest nieobciążony. Obciążenie według Breimana definiuje się zgodnie z formułą:

$$B = P_{\mathbf{x}}(y_* = y, \mathbf{x} \in B) - E_Z\left[P_{\mathbf{x}}(\hat{y} = y, \mathbf{x} \in B)\right]. \tag{34}$$

Wariancję definiuje się ze wzoru:

$$V = P_{\mathbf{x}}(y_* = y, \mathbf{x} \in U) - E_Z\left[P_{\mathbf{x}}(\hat{y} = y, \mathbf{x} \in U)\right]. \tag{35}$$

Tak zdefiniowane obciążenie i wariancja są zawsze nieujemne.

W swojej drugiej dekompozycji Breiman [2000] na początku proponuje, by błąd klasyfikacji obliczyć jako sumę dwóch składników:

$$Err_{agr} = E_{\mathbf{x}}\left[N(\mathbf{x})\right] + E_{\mathbf{x}}\left[\sum_y \left(P_y(y_*|\mathbf{x}) - P_y(y|\mathbf{x})\right)P_Z(\hat{y}|\mathbf{x})\right], \tag{36}$$

tj. szumu i nieujemnego składnika, będącego miarą wzrostu wartości błędu ponad szum.

Drugi składnik powyższej równości podlega rozbiciu na dwie składowe, tj. obciążenie:

$$B = P_y(y_*|\mathbf{x}) - P_y(y_m|\mathbf{x})P_Z(y_m|\mathbf{x}), \tag{37}$$

oraz rozproszenie, będące miarą wariancji (miara ta nie ma cech właściwych wariancji w zagadnieniu regresji, z czego wynika zmiana nazwy):

$$S = \sum_{y \neq y_m} \left(P_y(y_*|\mathbf{x}) - P_y(y|\mathbf{x})\right)P_Z(\hat{y}|\mathbf{x}). \tag{38}$$

Tak zdefiniowane wielkości przybierają wartości nieujemne.

Domingos [2000] proponuje, by zamiast ograniczania się do addytywnej dekompozycji błędu, a co za tym idzie – takiego definiowania obciążenia i wariancji, by addytywność ta była spełniona, należy najpierw zdefiniować ogólne pojęcie obciążenia i wariancji oraz ogólną postać dekompozycji, obowiązującą dla wszystkich postaci funkcji straty. Uogólniona postać dekompozycji jest dana wzorem:

$$\begin{aligned} Err_{agr} &= E_Z\left\{E_y\left[L(y,\hat{y})\right]\right\} = c_1 E_y\left[L(y,y_*)\right] + L(y_*,y_m) + c_2 E_Z\left[(y_m,\hat{y})\right] = \\ &= c_1 N(\mathbf{x}) + B(\mathbf{x}) + c_2 V(\mathbf{x}), \end{aligned} \tag{39}$$

gdzie c_1, c_2 – to czynniki przyjmujące różną postać, w zależności od postaci funkcji straty.

1. W przypadku kwadratowej funkcji straty: $c_1 = c_2 = 1$.
2. W przypadku zero-jedynkowej funkcji straty i obecności tylko dwóch klas:

$$c_1 = 2P_Z(\hat{y} = y_*) - 1, \tag{40}$$

$$c_2 = \begin{cases} 1, & \text{gdy } y_m = y_*, \\ -1, & \text{gdy } y_m \neq y_*. \end{cases} \tag{41}$$

Zatem w nieobciążonych obiektach wartość wariancji jest dodawana, natomiast odejmowana – w obiektach obciążonych.

3. W przypadku zero-jedynkowej funkcji straty i obecności wielu klas:

$$c_1 = P_Z(\hat{y} = y_*) - P_Z(\hat{y} \neq y_*) P_y(y = \hat{y} \mid y_* \neq y), \tag{42}$$

$$c_2 = \begin{cases} 1, & \text{gdy } y_m = y_*, \\ -P_Z(\hat{y} = y_* \mid \hat{y} \neq y_m), & \text{gdy } y_m \neq y_*. \end{cases} \tag{43}$$

5. Przykład empiryczny

Celem części empirycznej jest pokazanie, jak kształtują się wartości obciążenia i wariancji dla 0-1 funkcji straty, obliczone według różnych przedstawionych koncepcji. Do obliczeń wybrano benchmarkowy zbiór danych *Satimage* [Blake, Keogh, Merz 1998], do którego jest dołączony osobny zbiór testowy. Ze zbioru uczącego wylosowano 100 prób bootstapowych, każda taka próba posłużyła do budowy modelu w postaci drzewa klasyfikacyjnego. Następnie, z wykorzystaniem osobnego zbioru testowego, dokonano predykcji klasy na podstawie każdego ze 100 otrzymanych modeli. Całość obliczeń wykonano w programie *R* z zastosowaniem procedury *rpart*, uzyskane wyniki posłużyły zaś do wyznaczenia wartości obciążenia i wariancji.

Ze względu na to, że pomiar poziomu szumu jest problematycznym zagadnieniem, możliwym do przeprowadzenia *de facto* tylko w sztucznie generowanych zbiorach danych, przyjęto najczęściej stosowane rozwiązanie. Polega ono na przyj-

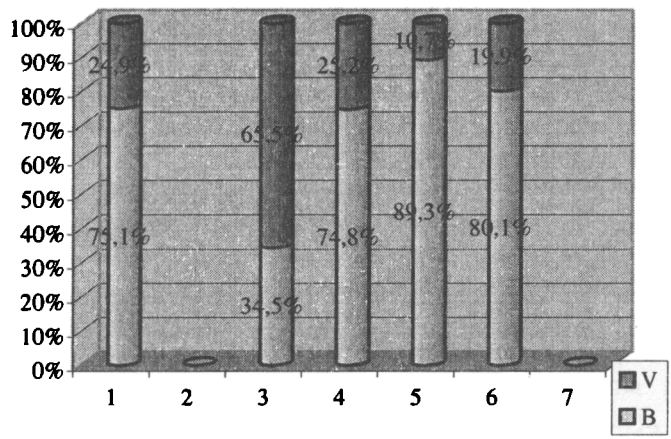
mowaniu założenia, że szum nie występuje i predykcja modelu optymalnego równa jest rzeczywistej wartości zmiennej objaśnianej.

Na podstawie otrzymanych wyników widać, że rzeczywiście suma wartości obciążenia i wariancji, obliczonych według ogólnej definicji, nie daje w rezultacie wartości błędu klasyfikacji.

Tablica 1. Wartości obciążenia i wariancji dla różnych dekompozycji

Koncepcja	Obciążenie	Wariancja	Błąd
Ogólnie	0,214	0,071	0,206
Kong i Dietterich	0,214	-0,008	0,206
Tibshirani	0,071	0,135	0,206
Kohavi i Wolpert	0,154	0,052	0,206
Breiman I	0,184	0,022	0,206
Breiman II	0,165	0,041	0,206
Domingos	0,214	-0,008	0,206

Źródło: obliczenia własne.



Rys. 1. Udział obciążenia i wariancji w wartości błędu klasyfikacji
Źródło: opracowanie własne.

Na podstawie pozostałych wyników widać dość spore różnice w wartościach interesujących nas wielkości, w zależności od sposobu liczenia. W dwóch przypadkach, tj. w dekompozycji Konga i Diettericha oraz Domingosa, które dają takie same wartości, otrzymano ujemną wartość wariancji, co sprawia, że dekompozycje są mało użyteczne ze względu na trudności interpretacyjne. W pozostałych przypadkach dodatnie wartości obydwu składników pozwalają zbadać, jaki jest wpływ błędu systematycznego i błędu średniego na wartość błędu klasyfikacji. Warto także zwrócić uwagę na wyniki otrzymane z dekompozycji według Tibshiraniego, które całkowicie odwracają proporcje między wartością obciążenia i wariancji. Najbardziej zbliżone do wartości obciążenia i wariancji, obliczonych według ogólnych reguł, są te otrzymane z dekompozycji Konga i Wolperta.

Decydując się na wybór jednej z dekompozycji błędu klasyfikacji, należy pamiętać, że każda z nich ma określone właściwości, ma także swoje wady i zalety. To, że pojawiło się tak wiele koncepcji, dowodzi, że pojęcie obciążenia i wariancji w zagadnieniu klasyfikacji jest zagadnieniem bardzo złożonym.

Literatura

Blake C., Keogh E., Merz C.J., *UCI Repository of Machine Learning Databases*, Department of Information and Computer Science, University of California, Irvine 1998.

Breiman L., *Arcing Classifiers*, Technical Report, Departament of Statistics, University of California, California 1996.

Breiman L., *Randomizing Outputs to Increase Prediction Accuracy*, „Machine Learning" 2000, 40 (3), wrzesień.

Domingos P., *A Unified Bias-Variance Decomposition for Zero-One and Squared Loss*, „Proceedings of the Seventeenth National Conference on Arificial Intelligence", AAAI Press, Austin 2000.

Kohavi R., Wolpert D.H., *Bias Plus Variance Decomposition for Zero-One Loss Functions*, „Proceedings of the Twelfth International Conference on Machine Learnig", Morgan Kaufmann, Tahoe City 1995.

Kong E.B., Dietterich T.G., *Error-Correcting Output Coding Corrects Bias and Variance*, „Proceedings of the Thirteenth International Conference on Machine Learnig", Morgan Kaufmann 1995.

Tibshirani R., *Bias, Variance and Prediction Error for Classification Rules*, Technical Report, Department of Statistics, University of Toronto, Toronto 1996.

ERROR DECOMPOSITION IN AGGREGATED DISCRIMINANT MODELS

Summary

The idea of error decomposition comes from regression where the square loss function is applied. Prediction error is decomposed into three components: noise, bias and variance.

There are also trials to apply the idea of error decomposition in classification. But the sum of those three components is different to the value of classification error. This is why several authors have proposed their own definitions of decomposition components for classification problem. In this paper we present and compare known decompositions for 0-1 loss.