УДК 004.93

КАК ОЦЕНИВАТЬ РЕЗУЛЬТАТЫ КЛАССИФИКАЦИИ НЕСБАЛАНСИРОВАННЫХ БОЛЬШИХ ДАННЫХ?

В.В. Старовойтов
Главный научный сотрудник ОИПИ НАН Беларуси,доктор технических наук, профессор

Ю.И. Голуб
Старший научный сотрудник ОИПИ НАН Беларуси, кандидат технических наук, доцент

Объединенный институт проблем информатики Национальной академии наук Беларуси, Республика Беларусь.
E-mail:valerystar @ mail.ru.,6423506 @gmail.com.

В. В. Старовойтов
Главный научный сотрудник ОИПИ НАН Беларуси, доктор технических наук, профессор, лауреат Государственной Премии Республики Беларусь (2002г.).
Ю. И. Голуб
Старший научный сотрудник, кандидат технических наук, доцент, в 2019 г. получала стипендию Президента Республики Беларусь как талантливый молодой учёный.

Аннотация. Классификация больших данных неравномерно распределенных по классам является серьезной проблемой интеллектуального анализа данных. При массовом скрининге пациентов в соотношении больных и здоровых всегда имеет место дисбаланс классов. При определении, например, степени заболевания раком – аналогично. При существенном дисбалансе данных классическая функция точности (Accuracy) не учитывает особенности малых классов и может ошибочно посчитать лучшим вариант с множеством ошибок в малых классах. В статье приведены результаты сравнительного анализа 17 функций оценки качества классификации на примерах матриц ошибок для 7 классов реальных и искусственных данных. Показано, что 4 функции инвариантны к дисбалансу данных. Описаны их достоинства и недостатки.

Ключевые слова: классификация, несбалансированные данные, матрица ошибок, функции оценки точности.

Введение.

В научной литературе описано множество различных функций оценки результатов классификации данных [1-10]. Классификация выполняется на два или более заранее определенных классов. Выше оценка – точнее классификация. При этом известных функций десятки, значения их различны на одних и тех же результатах классификации. Возникает вопрос: какую функцию выбрать? А если взять несколько «самых лучших оценочных функций», можно ли с их помощью определить «самый лучший» классификатор?.

Многоклассовая классификация. Наиболее полно (но далеко не полностью!) исследованы вопросы бинарной классификации, когда данные разделяются на два заранее определенных класса [2, 8 – 9]. Например, разделить множество данных на два класса: мужчины и женщины; люди больные диабетом или нет и т. п.

Задачи, требующие автоматической классификации на несколько классов, менее изучены. Еще менее исследованы функции оценок результатов классификации данных с разным числом объектов в классах, т. е. не сбалансированных данных.

Оценочные функции, используемые в данной работе.

Будем использовать 17 функций, позволяющих оценить результаты многоклассовой

классификации данных, подробно описанных в статье [5]. В работе [2] показано, что только 4 из них обладают инвариантностью относительно баланса данных. Коэффициентом дисбаланса называется отношение числа объектов наибольшего класса к наименьшему и обозначается IR (от англ. imbalance ratio). В таблице 1 приведены формулы этих четырех функций.

Таблица 1. Основные функции оценок многоклассовой классификации

Обозначение	Математическое выражение
ACC – функция точности accuracy	$\mathrm{ACC} = \frac{1}{N}(c_{11} + c_{22} + \ldots + c_{NN})$
ACCBal – сбалансированная по классам функция точности [10]	$\mathrm{ACCBal} = \frac{1}{N}\left(\frac{c_{11}}{n_1} + \frac{c_{22}}{n_2} + \ \ldots\ + \frac{c_{NN}}{n_N}\right)$
SinACC – сбалансированная по классам функция точности на базе синусов [11]	$\mathrm{SinACC} = 1 - \frac{1}{N}\sum_{i=1}^{N}\frac{\sqrt{\sum_{j\neq i}^{N} c_{ij}^2}}{\sqrt{\sum_{j=1}^{N} c_{ij}^2}}$
AU1U – среднее всех сумм функций чувствительности и специфичности, вычисленных для всех вариантов бинарных матриц [c_{ii}, c_{ki}; c_{kk}, c_{ik}], построенных из матрицы ошибок C [5]	$\mathrm{AU1U} = \frac{1}{N(N-1)}\sum_{i=1}^{N-1}\sum_{k>i}^{N}\left(\frac{c_{ii}}{c_{ii}+c_{ki}} + \frac{c_{kk}}{c_{kk}+c_{ik}}\right)$

В данной статье также используются наименования других оценочных функций, формулы которых для экономии места опущены, их можно найти в работе [5]:

Kappa – каппа-функция Коэна;

Me. Precision – среднее арифметическое функций Precision, вычисленных для отдельных классов;

GMe. Precision – среднее геометрическое функций Precision, вычисленных для отдельных классов;

GMe. Sensitivity – среднее геометрическое функций Sensitivity, вычисленных для отдельных классов;

CosineCoef – косинусный коэффициент;

VM – среднее геометрическое функций Sensitivity и Precision, вычисленных для отдельных классов;

Fmicro – среднее арифметическое гармонических средних Sensitivity и Precision для отдельных классов;

Fmacro – гармоническое среднее Sensitivity и Precision для отдельных классов;

Jmacro – среднее арифметическое индексов Юдена для отдельных классов;

sInd – среднее арифметическое индексов сходства в ROC-пространстве на базе специфичности и чувствительности отдельных классов;

normMCC – нормализованный коэффициент Коэна с диапазоном значений [0, 1];

AUNU – средняя площадь под ROC-кривыми N бинарных матриц ошибок, построенных по принципу «один против всех» из матрицы C [5];

AUNP – то же с учетом весов, пропорциональных числу элементов каждого класса к сумме всех данных.

Если классификация бинарная (N = 2), то AU1U = ACCBal по определению. Кроме того, AU1U = AUNU = AUNP, а для сбалансированных матриц ошибок все эти функции совпадают с ACC.

В настоящей работе значения всех оценочных функций нормализованы в диапазон [0, 0 +1].

Пример матрицы ошибок 20 несбалансированных классов.

Предположим, стоит задача разработки классификатора для разбиения некоторого множества данных на 20 заранее определенных классов. Это не очень большое число, при классификации символов классов может быть больше. В процессе настройки классификатора формируется матрица ошибок B размером 20x20 со средней ошибкой в каждом классе 35 %. В Таблицах 2 и 3 представлены для наглядности два фрагмента такой матрицы для случая, когда

число объектов в отдельных классах представляет собой геометрическую прогрессию с параметром 2 (как в старинной легенде про изобретателя шахмат).

$$B_j=B_1 q^{j-1} \quad (1)$$

где B_1 – число объектов меньшего класса, B_n – число объектов большего класса, q –параметр геометрической прогрессии, n – число членов прогрессии и число классов. Такие классы не сбалансированы поскольку IR = q^{n-1}, но результаты корректной классификации одинаково важны и для малых, и для больших классов. Общее число классифицируемых объектов в этом примере равно 104 857 502. Согласно сгенерированной случайно матрице ошибок к меньшему классу отнесены 99 объектов, к двум самым большим – 26 214 397 и 52 428 799. Средний процент ошибок классификации в каждом классе равен 35. В матрице B оценки экспертов расположены по вертикали, а предсказанные классификатором по горизонтали.

Таблица 2. Верхняя левая четверть матрицы ошибок B(1:10,1:10)

82	3	4	2	7	4	6	817	1046	1916
2	158	10	8	31	5	128	890	1206	689
1	1	263	2	37	3	242	353	381	451
0	3	14	623	37	2	223	480	1384	2048
0	1	1	14	1228	2	81	668	1386	706
1	2	4	15	16	3146	254	1137	401	1858
1	0	1	11	2	2	4053	401	2327	2479
1	2	9	15	39	1	17	3081	766	561
1	3	0	15	0	1	36	324	3317	1892
1	4	17	15	17	4	202	323	1415	25637

Таблица 3. Нижняя правая четверть матрицы ошибок B(11:20,11:20)

66682	516	1720	24524	22209	13437	169686	50604	436439	734446
2203	199487	12569	42610	29495	38833	315367	266766	459914	503280
2521	321	300191	57212	44910	19400	340328	336265	109488	1001666
2029	318	9320	162973	52994	8668	349239	25789	429397	288815
1359	341	2843	31023	1105153	1821	383749	14170	391036	1382632
1829	337	5027	28409	10078	2843948	91486	84078	330892	1351734
35	491	7188	18276	40216	14040	2468682	218180	376298	897692
4035	68	2809	38608	13651	19392	283595	9801775	418363	375445
349	345	5723	25031	24165	32480	95108	133753	20163847	1304298
4451	474	556	18559	45773	31593	322311	274475	151499	37265906

Как матрицу ошибок подобную B или больше ее проанализировать при выборе наилучшего варианта классификатора? Вручную это сделать невозможно. А автоматически? Классическая функция точности (Accuracy) для оценки результатов классификации несбалансированных данных не подходит [6]. Должна быть выбрана подходящая оценочная функция или алгоритм оценки результатов классификации. При настройке искусственных нейронных сетей такой анализ матриц ошибок повторяется многократно.

На рисунке 1 графически показаны значения 17 оценок матрицы ошибок B, описанной выше и имеющей 35 % ошибок в каждом классе. Значения отложены на единичных отрезках соединяющих центр и вершины правильного многоугольника.

Единица соответствует 100 % верной классификации, 0 – ни одного верно отнесенного объекта. Из рисунка видно, что значения функций нелинейно зависят от числа ошибок, т. е. при 50 % ошибок значение оценки может больше или меньше 0,5. Значение функции GMe. Precision практически равно нулю.

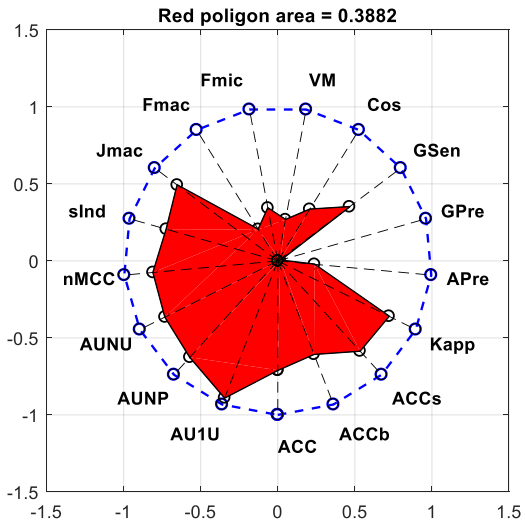

Рисунок 1. Визуальное представление 17 оценок точности, вычисленных по матрице B, представленной в таблицах 2 и 3, со случайно распределенными ошибками

Анализ матриц ошибок классификации кожных заболеваний.

Рассмотрим реальную медицинскую проблему – анализ дерматологических изображений кожи для выявления заболеваний (рисунок 2.).

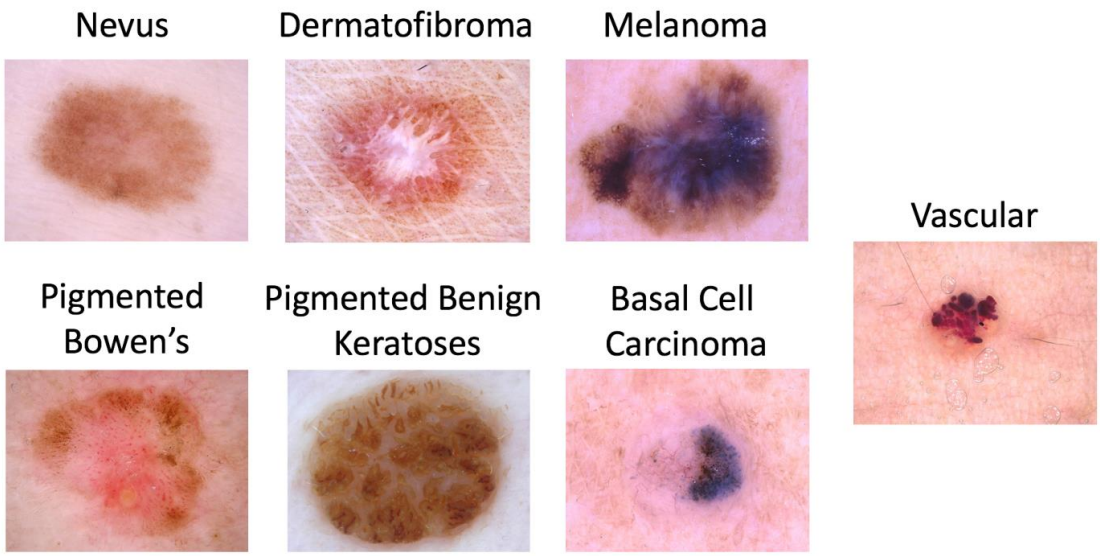

Рисунок 2. Примеры изображений семи вариантов поражений кожи

Рак кожи – это глобальная проблема здравоохранения, учитывая растущее распространение вредных ультрафиолетовых лучей в атмосфере Земли [12]. В настоящее время ежегодно во всем мире регистрируется порядка 123 000 меланом и 30 000 000 немеланом (IR=244) [13]. Самым эффективным вариантом снижения смертности от рака кожи является своевременная диагностика, поскольку выживаемость пациентов с меланомой за пятилетний период составляет 99 процентов при диагностике и скрининге на ранней стадии [14].

Есть много типов рака кожи, и меланома – самый смертельный из них. Дерматоскопия – важный метод визуализации для выявления меланомы и других кожных поражений. Однако классификация кожных повреждений на основе компьютерных методов диагностики является сложной задачей из-за нехватки маркированных данных и несбалансированного набора данных. Ранняя диагностика меланомы – важная проблема, которая может значительно улучшить

выживаемость пациентов [14]. Поражения кожи в первую очередь диагностируются визуально, а результат визуального осмотра кожных повреждений зависит от сходства и различий различных категорий изменений, поэтому диагностическая точность дерматологов составляет всего около 60 % [15]. Было проведено сравнительное исследование точности диагностики кожных заболеваний между системой на базе искусственного интеллекта и 112 дерматологами. Система показала точность выше большинства медиков [16].

Диагностическая точность кожных заболеваний, основанная на данных дерматоскопии, может быть увеличена до 75–84 % [17]. Несмотря на повышение диагностической точности с помощью дерматоскопии, скрининг кожных заболеваний с помощью дерматоскопических изображений требует много времени, а на диагноз влияют субъективные факторы врачей-дерматологов. Классификация кожных поражений с помощью автоматизированных систем компьютерной диагностики – сложная задача из-за вариабельности внешнего вида кожных поражений. Такие системы разрабатываются и показали большой потенциал [12, 15 – 19].

Набор общедоступная база дерматоскопических изображений ISIC2018 предоставлена Международной лабораторией визуализации кожи [19]. Она состоит из более 10 000 дерматоскопических изображений повреждений кожи, полученных в Австрии и Австралии. Это 24-битные изображения RGB размером 600×450 пикселей с разрешением 96 точек на дюйм. В базе выделены семь диагностических категорий поражений кожи (рис. 3.).

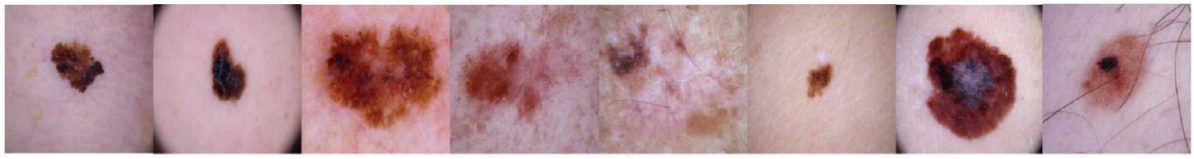

Рисунок 3. Примеры разнообразия вариантов фотографий меланомы из базы ISIC2018

Авторы статьи [12] разработали и сравнили 4 варианта классификаторов 7 категорий объектов. Основными показателями корректности классификации в статье были функции accuracy, sensitivity, specificity, average precision and balanced multiclass accuracy или (в русском переводе) точность, чувствительность, специфичность, средняя точность и сбалансированная мультиклассовая точность.

Матрица ошибок классификатора, названного в статье [12] Transfer-ResNet50, показана в оригинальной статье на Рисунке 11 (d), поэтому в данной статье она названа F11d. Матрицы, названные F11d2 и F11d22 – это производные от F11d, отличающиеся небольшими изменениями данных в 7-й строке (самый маленький класс).

Матрицы, представленные на Рис. 4, взяты из трех разный публикаций. Исходные данные были взяты из доступной базы изображений [19], поэтому классы в них одинаковые, но в матрицах ошибок классы имеют разный порядок. T8 [18] и F11x имеют почти равное число объектов однотипных классов и почти одинаковый коэффициент дисбаланса IR≈58. У матрицы T3 [17] коэффициент равен почти 1 (см Таблицу 4). В Таблице 4 расставлены места согласно значению функции точности (Accuracy): T3, T8, затем F11d. Матрица T3 лидер по всм параметрам. Примеры точности классификации по отдельным классам приведены на Рисунке 5. Однако большинство функций (их значения выделены желтым) хоть и имеют близкие значения, но ставят матрицу F11d на второе место, а T8 на третье. Какой результат классификации объективно соотвествует второму месту?.

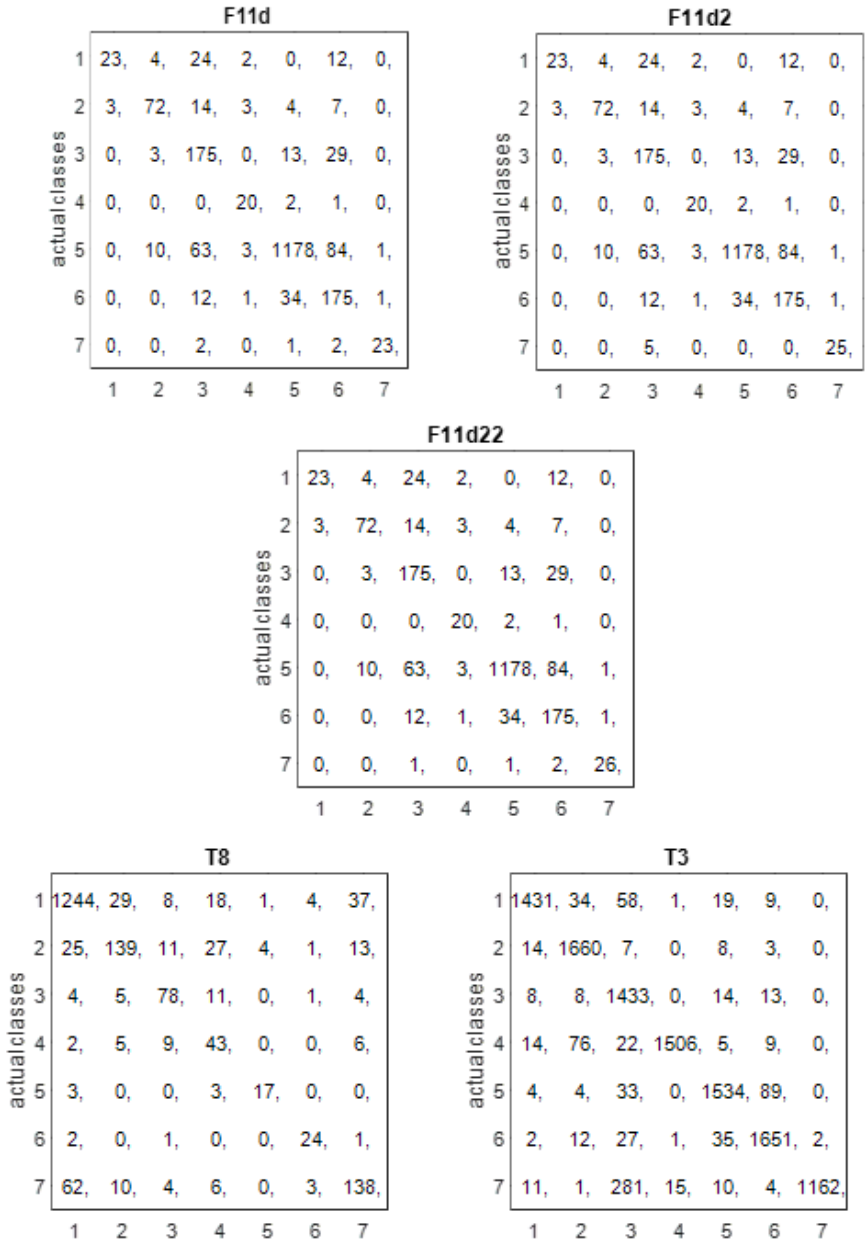

Рисунок 4. Матрицы ошибок с абсолютными значениями и разным дисбалансом данных

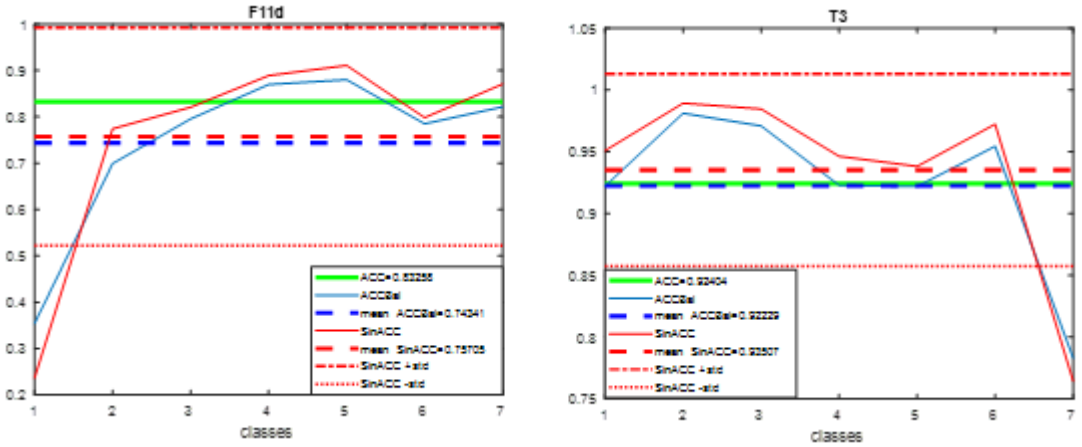

Рисунок 5. Результаты классификации по классам и средние значения трех вариантов точности: ACC, ACCBal, SinACC. Слева [15] и справа [17] классы имеют разный порядок и число объектов

Таблица 2. Оценки для матриц ошибок, представленных на Рис.4

Оценочная функция	**F11d** 3 место	**F11d2**	**F11d22**	**T8** 2 место	**T3** 1 место
IR	58.2174	58.2174	58.2174	58.3043	1.1721
ACCuracy	0.8326	0.8328	0.8333	0.8402	0.9240
ACCBal	0.7434	0.7451	0.7499	0.7419	0.9223
SinACC	0.7570	0.7475	0.7621	0.7825	0.9351
Kappa	0.8483	0.8488	0.8491	0.8479	0.9557
Me. Precision	0.7753	0.7759	0.7769	0.7087	0.9313
GMe. Precision	0.7611	0.7614	0.7625	0.6904	0.9283
Gme.Sensitivity	0.7166	0.7181	0.7221	0.7343	0.9201
CosineCoef	0.7592	0.7604	0.7632	0.7251	0.9268
VM	0.7473	0.7484	0.7514	0.7220	0.9248
Fmicro	0.7590	0.7602	0.7631	0.7249	0.9267
Fmacro	0.7360	0.7371	0.7403	0.7187	0.9229
Jmacro	0.8547	0.8556	0.8580	0.8518	0.9548
sInd	0.8137	0.8150	0.8183	0.8077	0.9419
norm MCC	0.8507	0.8512	0.8515	0.8482	0.9563
AUNU	0.8547	0.8556	0.8580	0.8518	0.9548
AUNP	0.8809	0.8816	0.8814	0.8661	0.9558
AU1U	0.9428	0.9437	0.9442	0.9464	0.9864

Выполним балансировку матриц ошибок, разделив значения элементов на сумму строки, в которой они расположены После этой операции коэффициент дисбаланса равняется 1, а матрицы представлены на Рис 6 Оценки, вычисленные по ним, представлены в Таблице 5.

В Таблице 5 желтым цветов выделены изменившиеся фрагменты оценок по сравнению с несбалансированными представлениями матриц в Таблице 4 Все оценки, выделенные желтым, поставили точность классификации, представленную матрицей T8, на третье место, а матрицей F11d на второе Жирным курсивным шрифтом обозначены четыре оценки, которые не изменились Они выше у матрицы T8 по сравнению с матрицей F11d, кроме оценки ACCBal.

Очевидно, что результаты классификации, представленные матрицами ошибок F11d и T8, очень близки. Однако возникает вопрос: насколько инвариантны значения оценочных функций от относительно дисбаланса данных? При настройке классификаторов базирующихся на искусственных нейронных сетях приходится делать выбор лучшего классификатора десятки и сотни раз. Поэтому ответ на поставленный вопрос очень важен. Из Таблиц 4 и 5 видно, что только 4 из 17 анализируемых оценочных функций инвариантны к дисбалансу классифицируемых данных, а дисбаланс данных влияет на значение оценочной функции.

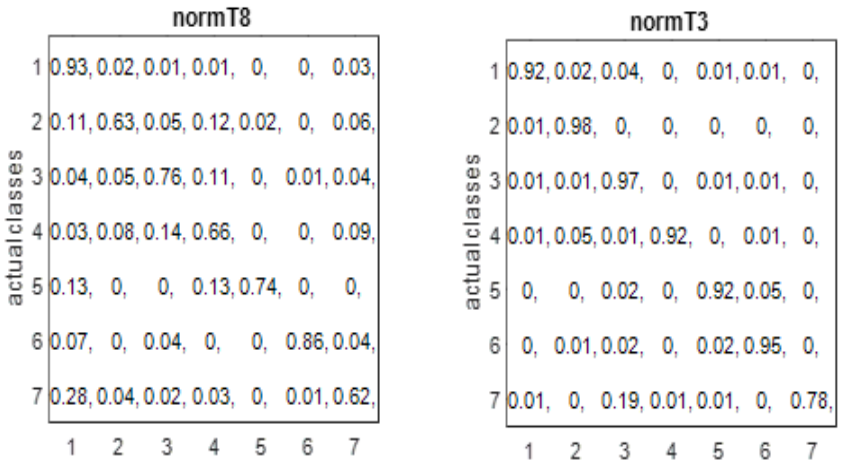

Рисунок 6. Примеры сбалансированных вариантов матриц ошибок T8 и T3

Таблица 3. Оценки сбалансированных матриц ошибок F11d, T8, T3

Оценочная функция	**F11d** 2 место	**T8** 3 место	**T3** 1 место
IR	1	1	1
ACCuracy	0. 7434	0. 7419	0.9223
ACCBal	***0.7434***	***0.7419***	0.9223
SinACC	***0.7570***	***0.7825***	0.9351
Kappa	0.8503	0. 8495	0.9547
Me. Precision	0.7951	0. 7680	0.9308
GMe. Precision	0.7754	0. 7552	0.9308
GMe.Sensitivity	***0.7166***	***0.7343***	0.9281
CosineCoef	0.7688	0.7549	0.9265
VM	0.7546	0.7500	0.9246
FMicro	0.7684	0.7547	0.9265
FMacro	0.7410	0.7451	0.9227
JMacro	0.8503	0.8495	0.9547
sInd	0.8120	0.8101	0.9417
norm MCC	0.8540	0.8514	0.9553
AUNU	0.8503	0.8495	0.9547
AUNP	0.8503	0.8495	0.9547
AU1U	***0.9428***	***0.9464***	0.9864

Анализ искусственных матриц ошибок классификации.

Рассмотрим поведение этих же 17 функций на случайно сгенерированных матрицах, классы которых представляет собой геометрическую прогрессию с параметром 2, а число ошибок будет зафиксировано по классам. Пусть оно будет равна 1-ACCBal = 0,25, что соответствует 25 % ошибок в среднем в каждом классе как в матрицах ошибок F11d и T3 при коэффициенте IR=64. Сгенерируем по 100 матриц с ошибками, отнесенными к одному классу и к разным классам (рисунок 7) и оценим разброс ошибок на Рисунках 8 и 9.

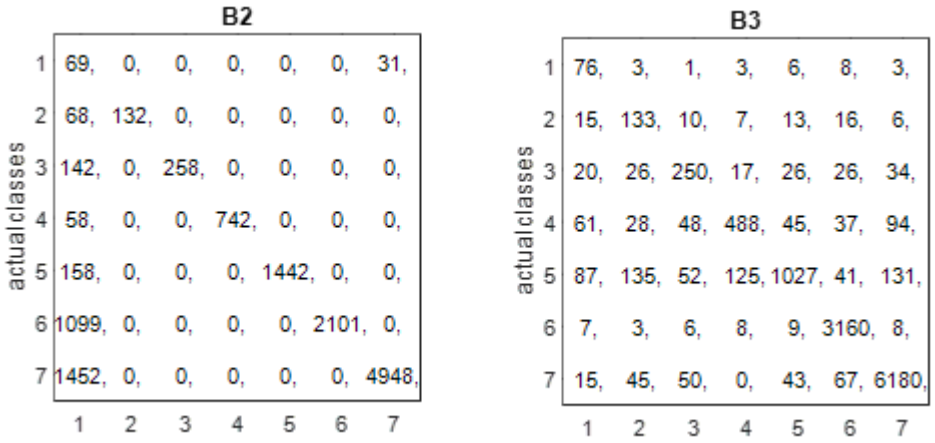

Рисунок 7. Примеры случайных матриц с 25 % ошибок, отнесенных к одному классу (a) и случайно распределенных по всем классам (b)

На Рисунке 10 представлены графики изменений четырех функций по ста матрицам. Из них ACC – самая нестабильная. Функция ACCBal выглядит самой стабильной, но это в среднем. Именно не ее базе определяется совокупность ошибок в матрице. На Рисунке 11 графически показаны оценки точности по классам матрицы B3, представленной на Рисунке 7. Функция SinACC имеет более узкий диапазон разброс значений по классам, чем функция ACCBal. С другой стороны, при равном числе ошибок в одном классе, но различном их распределении по классам, функция ACCBal будет всегда иметь одинаковые значения точности для этого класса, а функция ACCBal разные.

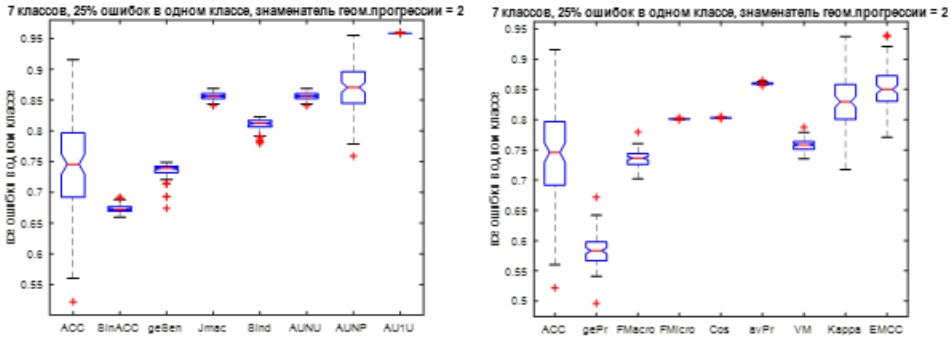

Рисунок 8. Сто случайно сгенерированных матриц с фиксированной среднеклассовой ошибкой 25 % отнесенных к одному классу и квартели разброса значений оценочных функций этих матриц

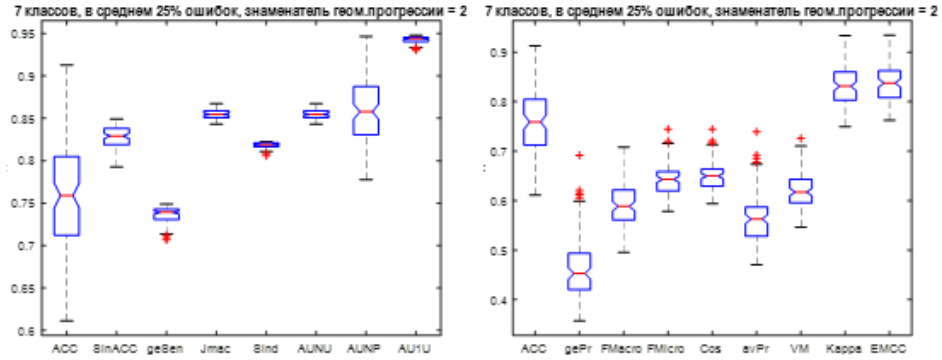

Рисунок 9. Сто случайно сгенерированных матриц с фиксированной среднеклассовой ошибкой 25 % случайно отнесенных к разным классам и квартили разброса значений оценочных функций этих матриц

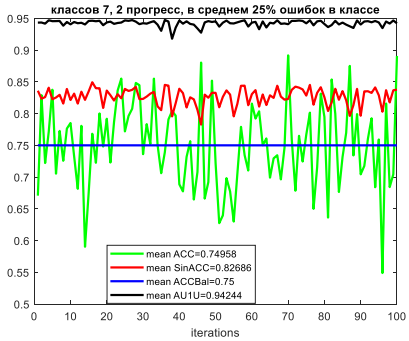

Рисунок 10. Графики средних оценок функций ACC (зеленый цвет), SinACC (красный), AU1U (черный), ACCBal=75 % (синий) по каждой из ста случайных матриц ошибок

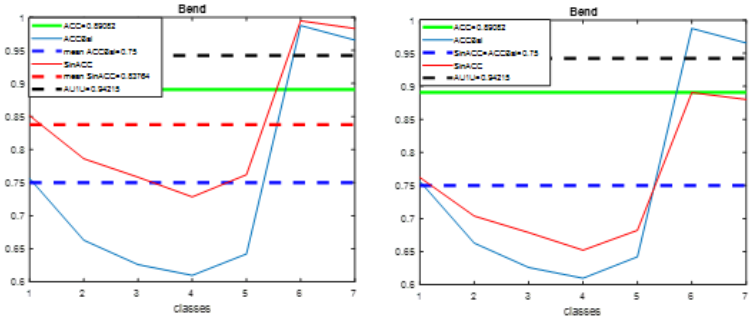

Рисунок 11. Оценки точности по классам матрицы B3. Справа оценки SinACC нормализованы так, чтобы средняя совпадала со средней ACCBal

На каждом графике Рисунке12 представлены квартели и диапазоны значений 17 оценочных функций, вычисленных для тысячи случайных матриц ошибок с несбалансированными классами и фиксированной средней ошибкой в классе, равной 1 %. Анализ графиков представленных на Рисунках 8, 9 и 12 показывает, что самый узкий диапазон значений имеет функция AU1U, но и разница ее значений при 25 и 1 % ошибок не велика, а сами значения очень близки к 1, что при больших процентах ошибок может быть неверно оценено. Следует отметить функцию GMeSensitivity. Она имеет узкий диапазон значений, близкий к реальному проценту ошибок, однако если среди анализируемых классов будет такой, у которого ни один объект не будет корректно распознан, значение этой функции будет равно 0. Функция SinACC также имеет узкий диапазон значений для матриц ошибок заданного типа. Ее особенностью является снижение общей оценки классификации, когда большая часть ошибок сосредоточена в одном классе, и повышение, когда ошибки распределены между классами боле равномерно. Это свойство может помочь автоматически анализировать распределение ошибок классификации по классам в процессе настройки классификатора. Функции GMePrecision, FMacro, FMicro, CosineCoef, MePrecision и VM имеют широкий разброс значений при одном проценте ошибок и не инвариантны к дисбалансу данных. Функции Kappa и normMCC не инвариантны к дисбалансу данных и имеют более широкий диапазон значений, чем функция ACCBal.

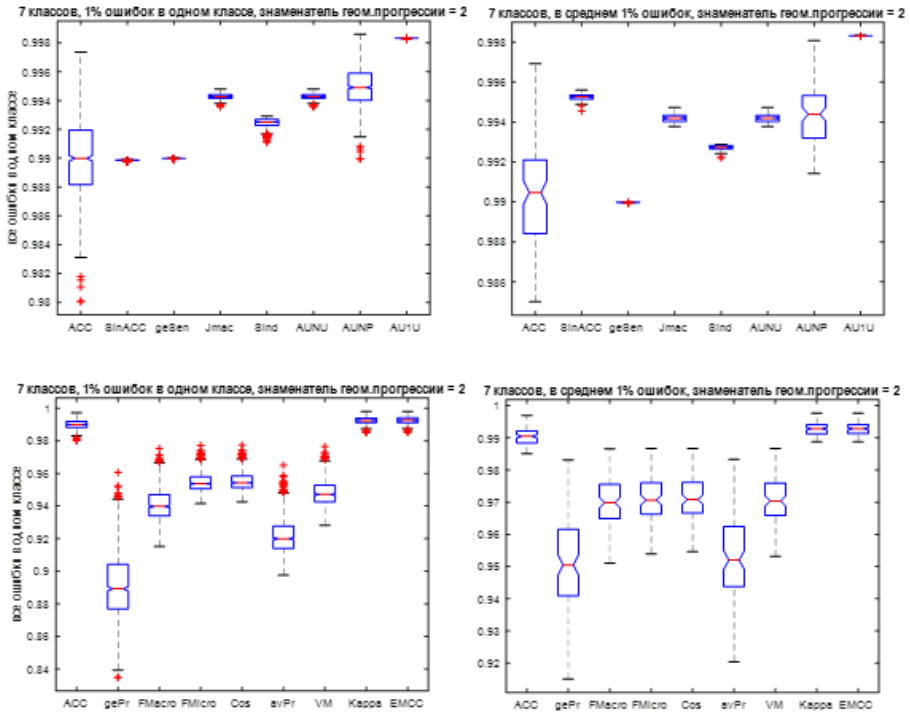

Рисунок 12. Статистика матриц ошибок построенных аналогично показанным на Рисунках 8 и 9 с 1 % ошибок в каждом классе.

Заключение.

В работе на примерах реальных и искусственных данных продемонстрированы особенности 17 функций, оценивающих результаты классификации несбалансированных данных. Из них только четыре функции инвариантны к дисбалансу данных и имеют узкий диапазон значений при анализе случайно сгенерированных матриц ошибок. Это – ACCBal, SinACC, AU1U и GMeSensitivity. Однако GMeSensitivity является средней геометрической величиной от функции Sencitivity, вычисленных для каждого класса, и если для некого класса ее значение будет равно 0 или очень мало, значение функции GMeSensitivity также будет равно 0 или очень мало даже при полном отсутствии ошибок классификации объектов других классов. Значения функции AU1U имеют очень узкий диапазон и близки к 1 независимо от процента ошибок классификации, что

может привести к выбору не самого лучшего классификатора в процессе его настройки.

Функция ACCBal дает объективную среднюю оценку ошибок классификации несбалансированных данных и проста в вычислении. Функция SinACC отражает как общий уровень ошибок классификации, так и их распределение в отдельных классах.

Анализируя графики типа представленных на Рис.11, можно дополнительно контролировать результаты классификации при большом числе классов.

Список литературы

[1] Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks // Information processing & management. – 2009. – Vol. 45. – No.4. – pp.427-437.

[2] Старовойтов В.В., Голуб Ю.И. Сравнительный анализ оценок качества бинарной классификации // Информатика. – 2020. – 17(1). – С.87-101. doi.org/10.37661/1816-0301-2020-17-1-87-101.

[3] Hossin M, Sulaiman M.N. A review on evaluation metrics for data classification evaluations // International Journal of Data Mining & Knowledge Management Process. – 2015. – Vol. 5. – No.2. – pp.1-11. doi: 10.5121/ijdkp.2015.5201.

[4] Luque A, et al. The impact of class imbalance in classification performance metrics based on the binary confusion matrix // Pattern Recognition. – 2019. – Vol. 91. – pp.216-231.

[5] Ballabio D., Grisoni F., Todeschini R. Multivariate comparison of classification performance measures // Chemometrics and Intelligent Laboratory Systems. – 2018. – Vol. 174. – pp.33-44. doi.org/10.1016/j.chemolab.2017.12.004.

[6] Старовойтов В.В., Голуб Ю.И. Об оценке результатов классификации несбалансированных данных по матрице ошибок // Информатика. – 2021. – 18(1). – С.61-71. doi.org/10.37661/1816-0301-2021-18-1-61-71.

[7] Ferri C., Hernendez-Orallo J., Modroiu R., An experimental comparison of performance measures for classification // Pattern Recognition Letters. – 2009. – Vol. 30. – No.1. – pp.27-38.

[8] Freeman E.A., Moisen G.G. A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa // Ecological modelling. – 2008. – Vol. 217.– No.1-2. – pp.48-58.

[9] Chicco D., et al. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation // BioData Mining. – 2021, doi: 10.1186/s13040-021-00244-z.

[10] Garsia V., et al. Index of balanced accuracy: a performance measure for skewed class distributions // In Iberian Conference on Pattern Recognition and Image Analysis. – 2009 Jun 10. – pp.441-448. Springer, Berlin, Heidelberg.

[11] Starovoitov V., Golub Yu. New function for estimating imbalanced data classification results // Pattern Recognition and Image Analysis, 2020, Vol. 30, No. 3, pp. 295–302. doi:10.1134/S105466182003027X.

[12] Chaturvedi S.S., Gupta K., Prasad P.S. Skin lesion analyser: an efficient seven-way multi-class skin cancer classification using MobileNet // Advanced Machine Learning Technologies and Applications. – Springer, Singapore. – 2020. – p.165-176. doi: 10.1007/978-981-15-3383-9_15.

[13] World cancer report 2014 / in C. P. Wild, & B. W. Stewart (Eds.). Geneva, Switzerland: World Health Organization, 2014, pp. 482-494.

[14] Siegel R.L., Miller K.D., Jemal A. Cancer statistics // A Cancer Journal for Clinicians, 2021, Vol.71, No.1, pp. 7– 33. doi: 10.3322/caac.21654.

[15] Qin X.Z. et al. A GAN-based image synthesis method for skin lesion classification // Computer Methods and Programs in Biomedicine. – 2020. – Vol. 195. – p.105568. doi.org/10.1016/j.cmpb.2020.105568.

[16] Maron R.C., et al., Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks // European Journal of Cancer. – 2019. – Vol. 119. – pp.57-65.

[17] Kassani, S.H., Kassani, P.H. A comparative study of deep learning architectures on melanoma detection // Tissue and Cell. – 2019. – Vol. 58. – pp.76-83. doi: 10.1016/j.tice.2019.04.009.

[18] Al-Masni M.A., Kim D.H., Kim T.S. Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification // Computer Methods and Programs in Biomedicine. – 2020. – Vol. 190. – p.105351. doi.org/10.1016/j.cmpb.2020.105351.

[19] ISIC Skin image analysis Workshop and Challenge @ MICCAI 2018 [Электронный ресурс]. – Режим доступа: https://workshop2018.isic-archive.com.

HOW TO EVALUATE RESULTS OF IMBALANCED BIG DATA CLASSIFICATION

V.V. STAROVOITOV
Doctor of Engineering Sciences, Professor, Chief Researcher UIIP NAS of the Republic of Belarus

Y.I. Golub
Ph.D., Associate Professor, Senior Researcher UIIP NAS of the Republic of Belarus

United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Republic of Belarus
E-mail: valerystar @ mail.ru
United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Republic of Belarus
E-mail: 6423506@gmail.com

Abstract. Classification of imbalanced big data is an important data mining problem. In mass screening, there is always a class imbalance in the ratio of sick and healthy. When determining, for example, the degree of cancer, it is the same. If there is a significant imbalance in the data, the classic Accuracy function does not take into account the peculiarities of small classes and may erroneously consider the best option with many errors in small classes. The article presents the results of a comparative analysis of 17 functions for quality assessment of classification on examples of confusion matrices for 7 classes of real and artificial data. It is shown that 4 functions are invariant to data imbalance. Their advantages and disadvantages are described.

Keywords: classification, imbalanced data, confusion matrix, accuracy estimation functions.