УДК 004.93

ПРОБЛЕМЫ СТАТИСТИЧЕСКИХ ОЦЕНОК ПРИ АНАЛИЗЕ НЕСБАЛАНСИРОВАННЫХ КЛАССОВ БОЛЬШИХ ДАННЫХ

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

Объединенный институт проблем информатики национальной академии наук Беларуси,
Республика Беларусь
E-mail:valerystar @ mail.ru

В.В. Старовойтов
Главный научный сотрудник ОИПИ НАН Беларуси, лауреат Государственной Премии Республики Беларусь (2002г.).

Аннотация. В статье утверждается что прикладная статистика в настоящее время не готова к анализу и обработке больших данных. Вычислять средние значения, дисперсию и прочие статистические характеристики для многочисленных и разнообразных классов объектов, относящихся к категории больших данных, бессмысленно и бесполезно. Одной из актуальных задач является классификация множества объектов на существенно различные по объему классы. К ним относятся реальные задачи разделения людей на заболевших некоторой болезнью и здоровых, сортировка электронной почты на спам и обычные сообщения и т.п. Разработано множество методов классификации данных. Результаты их работы описываются матрицами ошибок. По этим матрицам можно оценить качество классификации и выбрать лучший метод классификации определенных данных. До настоящего времени для оценки качества результатов классификации данных чаще всего используются функции Accuracy, Sensitivity, Specificity и F1. В результате экспериментальных исследований установлено, что указанные функции искажают истинные результаты классификации в случае существенного дисбаланса классов. Показано, что для оценки бинарной классификации из известных функций наиболее инвариантной к дисбалансу классов является функция AUC, которая вычисляет площадь под ROC-кривой. В случае бинарной классификации она равна среднему арифметическому значению функций Sensitivity и Specificity.

Ключевые слова: прикладная статистика, классификация несбалансированных данных, Accuracy, Sensitivity, Specificity, F1, AUC.

Введение. Шестьдесят лет назад цифровые компьютеры сделали информацию читаемой. Двадцать пять лет назад Интернет сделал ее доступной. С 2000-х поисковики создали единую (всемирную) базу данных. В 21-м веке самым ценным ресурсом на Земле становится информация, содержащаяся в цифровых данных различного вида. Таких данных накопилось столь много, что 10 лет назад стали широко применяться понятие «большие данные» или Big Data. Формального определения понятия «большие данные» нет. Неформально этим термином обозначают плохо структурированные данные огромных объемов, прирастающих с огромной скоростью.

Профессор Де Мауро проанализировал различные определения понятия “Big Data” и предложил следующее: «Большие данные – это информационный багаж, характеризующийся таким большим объемом, скоростью изменений и разнообразием, что для его преобразования

в нечто ценное требуются специальные технологии и аналитические методы» [1]. Понятие «большие данные» включает не только наборы данных гигантского объема до 10^{15}-10^{18} байт, но и методы сбора таких данных, их хранения и обработки. Для сравнения скажем, что масса Земли равна примерно $6*10^{23}$ граммов.

Как описывать, как использовать такие данные, анализировать их и извлекать содержащуюся в них информацию?

Прикладная статистика и большие данные. Математическая статистика – это область математики, в которой исследуются задачи анализа множеств количественных и качественных данных. В статистике хорошо проработаны задачи анализа небольших групп данных: несколько десятков или сотен единиц. В еще докомпьютерные времена был разработан ряд математически обоснованных критериев. Имеется закон больших чисел. Он гласит, что среднее значение конечной выборки из фиксированного распределения стремится к математическому ожиданию этого распределения. В статистике предложены методы описания и анализа множеств данных различной природы, проверки гипотез равенства средних, дисперсий, наличия линейной зависимости между данными двух множеств и т.п. Статистика продолжает развиваться в основном, используя вероятностный подход к анализу данных небольших объемов.

В начале 1980-х годов в СССР заговорили о появлении прикладной статистики [2-3]. Она нацелена на решение реальных задач. В неё входят ориентированные на прикладную деятельность статистические методы анализа данных, а также методологию организации статистического исследования и организацию компьютерной обработки данных, в том числе разработку и использование баз данных и статистического ПО.

Возьмем частный пример из области прикладной статистики: математические методы диагностики [4]. Они делятся на параметрические и непараметрические. Первые основаны на предположении, что классы описываются распределениями из некоторых параметрических семейств. Обычно рассматривают многомерные нормальные распределения, при этом зачастую принимают гипотезу о том, что ковариационные матрицы для различных классов совпадают. На таких предположениях сформулирован классический дискриминантный анализ Фишера. Однако обычно нет оснований считать, что наблюдения извлечены из нормального распределения. Поэтому более корректными, чем параметрические, считаются непараметрические методы диагностики. Идея таких методов основана на лемме Неймана-Пирсона. Согласно этой лемме решение об отнесении вновь поступающего объекта к одному из двух классов принимается на основе отношения плотностей распределения двух классов. Если плотности распределения неизвестны, то применяют их непараметрические оценки плотностей, построенные по обучающим выборкам, а диагностическое решение принимают по их отношению. Таким образом, для решения задачи диагностики достаточно построить непараметрические оценки плотности для выборок объектов произвольной природы. Как это сделать в условиях больших данных никто знает.

В 2019 году в Беларуси провели республиканскую акцию, в которой участвовали стоматологи, онкологи и оториноларингологи [5]. Врачи оценивали статистику рака кожи в области лица, полости рта, на небных дужках, языке. Первичная выявляемость онкозаболеваний составила 6%, зато третью−четвертую стадии онкозаболеваний они диагностировали в 80% случаев. Эти результаты свидетельствуют о трудностях различия здорового и больного, т.е. о сложности задачи выделения информативных признаков и разбиения образцов исследуемых данных даже на два класса: здоровый и больной.

Современная статистика не готова анализировать большие данные. Методы классической статистики практически не применимы к анализу больших данным. Причинами этого (кроме их объема) являются децентрализованные способы хранения, плохая структуризация и взаимосвязь данных. Это привело к тому, что к настоящему моменту большие данные не могут быть статистически описаны.

В базе данных ImageNet (http://www.image-net.org) собрано более 14 миллионов аннотированных изображений, разбитых на 21 841 категорию. Например, породы собак представлены 120 классами и это лишь одна из категорий объектов. Как статистически обрабатывать такие множества данных пока не ясно. Вычислять средние значения, дисперсию и прочие статистические характеристики для подобных разнообразных данных бессмысленно и бесполезно. Аналогом бесполезности является величина средней зарплаты по стране.

Майский номер за 2018 год журнала Statistics & Probability Letters был целиком посвящен обсуждению проблемы «Роль статистики в эпоху больших данных» [6]. Было опубликовано 36 статей, написанных ведущими учеными-статистиками мира, однако кроме вывода о том, что что-то нужно делать никаких конкретных предложений они не дали. Таким образом статистика как наука существенно отстала от бурно увеличивающегося объема данных, которые требуется анализировать.

На сегодняшний день основным инструментов работы с большими данными являются глубокие нейронные сети. Их особенностью является принцип: чем больше данных использовано при обучении, тем лучше результат. Разработала, реализовала и положила начало практическому применению таких сетей компания Google. Она сначала обучила свои нейронные сети отличать кошек от собак, а в настоящее время уверенно распознавать их породы. С каждым годом расширяются области применения сетей этого типа. Глубокие нейронные сети уверенно распознают людей по их фотографиям, а также многие другие категории объектов, представленных большими объемами данных разных типов. При этом объемы обучающих выборок составляют миллионы объектов.

В интернете можно найти множество библиотек с открытыми кодами для применения глубоких нейронных сетей к решению самых разных задач. Очень интересным и полезным для проверки разных идей анализа и распознавания данных является платформа kaggle.com, принадлежащая компании Google LLC. На нее выкладываются различные задачи анализа данных, чаще всего не имеющие удовлетворительных решений, дается срок 2-3-4 месяца и объявляется конкурс на лучшее решение. Участники открыто обмениваются идеями и опытом, а компании, объявившие конкурс, получают быстрые прототипы решений научно-исследовательских задач.

Одним из важных инструментов статистического анализа данных является определение наличия корреляции между группами объектов. В книге Шонбергера и Кукера «Большие данные: революция, которая изменит то, как мы живем, работаем и думаем», 2013 года [7] сделано важное заявление: «*корреляция не равна причинно-следственной связи, она может быть ошибочной*». Например, маленьким детям говорят, что аисты приносят детей. Этому нашлось реальное подтверждение. Изучалась корреляция между числом аистов, свивших гнезда в южных районах Швеции, и рождаемостью в соседних поселениях. Корреляция между этими явлениями составила около 0.8. Оказалось, что синхронные изменения числа аистов и детей объяснялись изменением среднего уровня жизни жителей. При исключении этой искажающей переменной прежней корреляции уже не наблюдалось. Вывод: причинная зависимость не может быть выведена ни из какого наблюдаемого совместного изменения явлений.

Реальные связи и закономерности могут быть гораздо более глубокими и не очевидными. Например, Google предложила предсказывать возникновение эпидемий гриппа посредством анализа данных поисковых запросов об этой болезни [8]. Компания показала, что относительная частота определенных запросов сильно коррелирует с процентом визитов к врачу пациентов с симптомами, подобными гриппу. Поисковая система Google обработала порядка 450 миллионов математических моделей с целью уточнения условий поиска, сравнивая результаты с фактическими данными о случаях гриппа за 2007-2208 годы. Были найдены 45 условий поиска, использование которых давало коэффициент корреляции между

их прогнозом и официальными данными 97%. Это позволяет оценить уровень еженедельной активности гриппа в каждом регионе США с точностью до одного дня.

Поэтому в противовес классической статистике в марте 2008 года Питер Норвиг, директор исследований в Google, заявил на одной из конференций: «***Все модели ошибочны и все чаще вы можете добиться успеха без них***».

Классические оценки классификации несбалансированных данных работают некорректно. Актуальными направлениями анализа больших данных являются предсказание чего-либо в результате анализа данных и классификация данных. Очень важно корректно оценить результаты классификации, особенно в случае несбалансированных классов данных, и выбрать точнее работающий классификатор.

Анализ публикаций, посвященных классификации данных различной природы показывает, что в подавляющем большинстве из них используются стандартные статистические оценки, вычисляемые по матрицам ошибок, такие как Accuracy (правильность), Sensitivity (она же Recall или чувствительность), Specificity (специфичность), F1 (гармоническое среднее), area under ROC curve (AUC или площадь под ROC кривой). Причем первая оценка используется в более чем в 80% публикаций.

Приведем пример. [9] Эта статья посвящена анализу модели прогнозирования для диагностики рака молочной железы, как доброкачественной, так и злокачественной, на ранней стадии, поскольку это увеличивает шансы на успешное лечение. Сравниваются алгоритм классификации данных на базе SVM, Naive Bayes, k-NN, Decision Tree, а результаты оцениваются с помощью таких статистических мер, как точность, чувствительность, специфичность, значение положительного прогноза, значение отрицательного предсказания и площадь под кривой ROC. Исходные данные собраны из цифровых изображений и представлены в базе данных https://www.kaggle.com/uciml/breast-cancer-wisconsin-data. Данные представлены 32 признаками и размечены на два класса: злокачественная или доброкачественная опухоль. Всего дано 569 образцов: 357 доброкачественных, 212 злокачественных. Визуальное представление 10 первых признаков показано на Рис.1. Справа и слева от вертикальных отрезков показаны распределения признаков доброкачественных и злокачественных опухолей. Из рисунка видно существенное пересечение признаков обоих классов что означает невозможность 100% разделения двух классов. Основной оценочной функцией результатов классификации была функция Accuracy. Лидеры соревнования добились значений порядка 95-98%, а лучшими были признаны классификаторы, использовавшие PCA и SVM.

Приведем еще один пример. На сайте kaggle.com в конкурсе по определению мошенничества с банковскими транзакциями данные представляли собой 284807 корректных транзакций и 492 ложные (0,172% от всех операций). Дисбаланс классов составил 578:1. Применим этим данным простейший (необученный) классификатор, который относит все операции к классу корректных. Он будет имеет очень высокое значение функции Accuracy (99,827 %) и низкое значение Precision (практически 0,0 %), однако такой классификатор не выявит ни одной мошеннической транзакции. Какова польза от такого «правильного», т.е. почти безошибочного классификатора? Поэтому очень актуальным является вопрос: с помощью какой функции можно корректнее оценить результаты классификации несбалансированных данных? На данный момент прикладная статистка предложений не имеет.

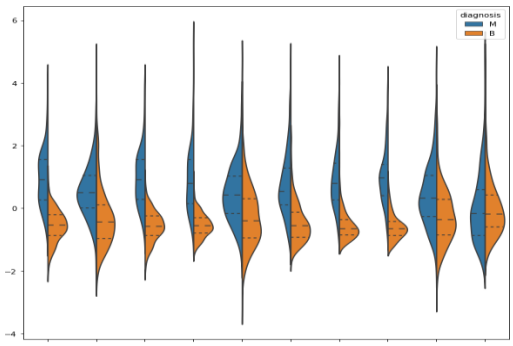


Рисунок *1*. – Первые 10 признаков из описания данных Breast Cancer Wisconsin Data Set. Все признаки двух классов существенно пересекаются

Приведем еще один пример. На сайте kaggle.com в конкурсе по определению мошенничества с банковскими транзакциями данные представляли собой 284807 корректных транзакций и 492 ложные (0,172% от всех операций). Дисбаланс классов составил 578:1. Применим этим данным простейший (необученный) классификатор, который относит все операции к классу корректных. Он будет имеет очень высокое значение функции Accuracy (99,827 %) и низкое значение Precision (практически 0,0 %), однако такой классификатор не выявит ни одной мошеннической транзакции. Какова польза от такого «правильного», т.е. почти безошибочного классификатора? Поэтому очень актуальным является вопрос: с помощью какой функции можно корректнее оценить результаты классификации несбалансированных данных? На данный момент прикладная статистка предложений не имеет.

В результате приведенного анализа и приведенных примеров можно сделать вывод о том, что современная статистика оказалась не готова обрабатывать реально большие данные. В настоящее время подобными задачами занимается только *data science* или наука о данных, относящаяся к искусственному интеллекту.

Оценки результатов классификации на базе набора KDDCUP99. Рассмотрим еще один пример множества несбалансированных данных. Это широко используемая и доступная база данных KDDCUP99 [10]. Она использовалась на 3-м международном конкурсе инструментариев для обнаружения знаний и интеллектуального анализа данных, который проводился Пятой международной конференцией по обнаружению знаний и интеллектуальному анализу данных в 1999 году. Задача конкурса заключалась в создании детектора несанкционированных вторжений в сеть на базе прогнозирующей модели, способной различать четыре типа вторжений и «нормальные» соединения. Все атаки разделены на четыре категории:

• DOS- сетевые атаки (Denial of Service attacks);

• R2L ((Remote-to-Local): несанкционированный доступ с удаленного компьютера:

• U2R (User-to-Root): несанкционированный доступ к данным сетевого администратора;

• Probing атаки – сканировании сетевых портов для получения конфиденциальной информации.

В таблице 1 приведены данные дисбаланса между классами этой базы. Самый большой дисбаланс между нормальными соединениями и атаками типа U2R. Данных первого класса в 1295,1 раз больше чем второго.

Таблица 1. – Дисбаланс между классами в базе KDD Train

класс	Normal	Dos	Probe	R2L	U2R	всего
размер	***67 343***	***45 927***	***11 656***	***995***	***52***	***125 973***
Normal		1,47	5,78	67,68	1 295,1	1,87
Dos			3,94	46,16	883,21	2,74
Probe				11,71	224,15	10,81
R2L					19,13	126,6
U2R						2 422,6

На рисунке 2 из статьи [11] показаны проекции образов исследуемых классов на две главные оси. Видно, что классы практически не разделимы без ошибок.

Оценим пять вариантов бинарной классификации на два класса U2R и Normal, по матрицам ошибок приведенным в Таблице 2. Первый вариант соответствует наилучшей классификации (всего 2 ошибки). Затем идут второй и третий. Четвертый и пятые самые плохие. В меньшем классе правильно определен лишь один объект. Два последних варианта соответствуют одному и тому же результату классификации, отличаются лишь перестановкой классов.

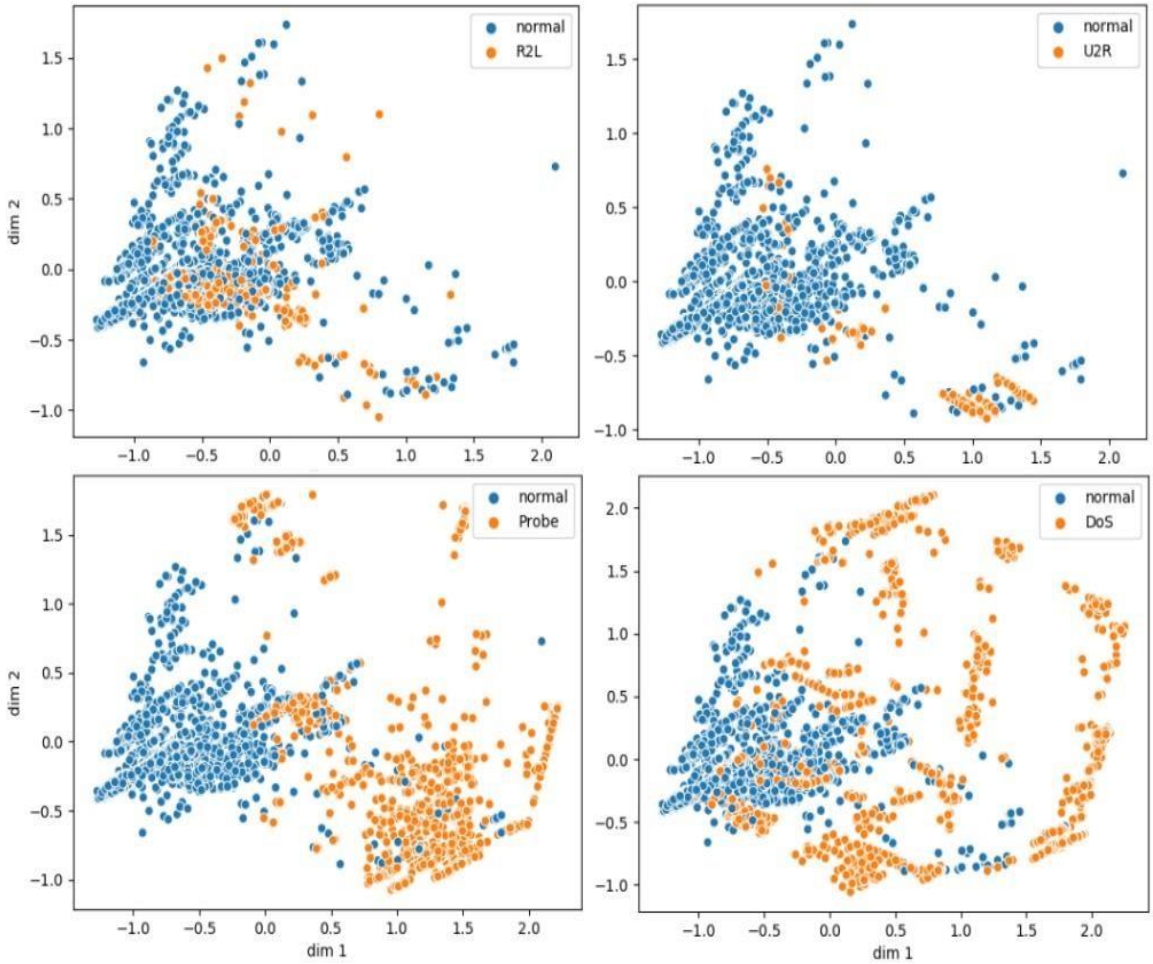


Рисунок *2.* – Проекция пар классов на 2 главные оси, выбранные методом главных компонент (PCA). Классы не разделимы

Таблица 2. –Матрицы пяти вариантов ошибок классификации объектов классов U2R и Normal

U2R (52 объекта)		предсказанные классы	
		Yes	No
реальные классы, вариант 1	Yes	51	1
	No	1	67342
вариант 2	Yes	41	10
	No	10	67332
вариант 3	Yes	32	1
	No	20	67342
вариант 4	Yes	1	1
	No	51	67342
вариант 5	Yes	67342	51
	No	1	1

Нами было исследовано около 20 известных функций оценки качества бинарной классификации со сбалансированными и несбалансированными классами данных [12]. Часть из них представлена в Таблице 3.

Вычислим функции из Таблицы 3 к матрицам ошибок и запишем в Таблицу 4.

Таблица 3. – Основные функции оценки качества бинарной классификации

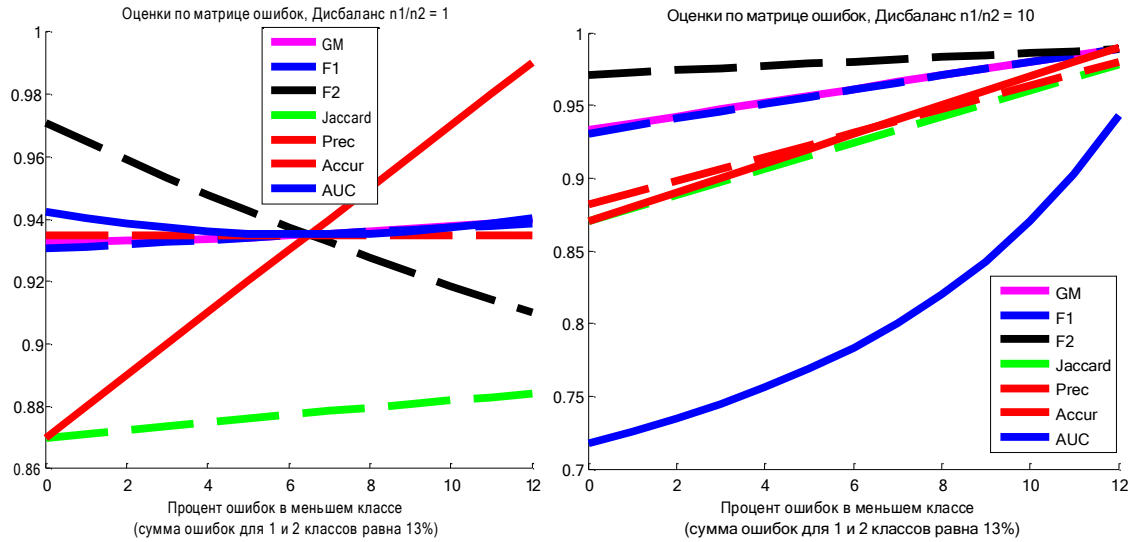
№	Обозначение функции	Формула
1	Sensitivity и Recall, чувствительность	$\frac{tp}{tp+fn}$
2	Specificity, специфичность	$\frac{tn}{tn+fp}$
3	Precision, точность	$\frac{tp}{tp+fp}$
4	Accuracy, правильность	$\frac{tp+tn}{n}$
5	Jaccard index, индекс Жаккара	$\frac{tp}{tp+fn+fp}$
6	*F*1, гармоническое среднее	$\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$, при $\beta = 1$
7	*F*2	$\frac{(1+\beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}$, при $\beta = 2$
8	геометрическое среднее, Geometric mean (GM)	$\sqrt{\text{Precision} \times \text{Recall}}$
9	area under ROC curve (AUC), площадь под ROC кривой	$\frac{\text{Sensitivity} + \text{Specificity}}{2}$
10	Cohen's kappa, каппа Коэна	$\frac{(tp+fp)(tp+fn)+(fn+tn)(fp+tn)}{n^2}$
11	MCC, коэффициент корреляции Мэтьюса	$\frac{(\text{tp} \times \text{tn} - \text{fp} \times \text{fn})}{\sqrt{(\text{tp}+\text{fp}) \times (\text{tp}+\text{fn}) \times (\text{tn}+\text{fp}) \times (\text{tn}+\text{fn}}}$

где $n = tp + fp + fn + tn$.

Таблица 4. – Сравнительные результаты оценок классификации

номер	функция	вар. 1	вар. 2	вар. 3	вар. 4	вар. 5
1	Sensitivity	0.9808	0.8039	0.9697	0.5000	0.9992
2	Specificity	1.0000	0.9999	0.9997	0.9992	0.5000
3	Precision	0.9808	0.8039	0.6154	0.0192	1.0000
4	**Accuracy**	***1.0000***	***0.9997***	***0.9997***	***0.9992***	***0.9992***
5	Jaccard	0.9623	0.6721	0.6038	*0.0189*	*0.9992*
6	F1	0.9808	0.8039	0.7529	0.0370	0.9996
7	F2	0.9808	0.8039	0.8696	0.0833	0.9994
8	GM	0.9808	0.8039	0.7725	0.0981	0.9996
9	**AUC**	**0.9904**	**0.9019**	**0.9847**	**0.7496**	**0.7496**
10	Cohen's kappa	0. 9904	0.9019	0.8764	0.5185	0.5185
11	MCC	0. 9904	0.9019	0.8862	0.5490	0.5490

Проанализируем результаты пяти вариантов классификации, собранные в Таблице 4. Первые четыре функции – это популярные меры качества, вычисляемые по матрице ошибок классификации. Корректные оценки должны быть инвариантны к перестановке классов в матрице ошибок, однако функции Sensitivity, Specificity, Precision, индекс Jaccard, F1 и F2 не инвариантны к такой перестановке, т.е. это неудовлетворительные меры. Accuracy инвариантна, но она не учитывает дисбаланс классов (см. варианты 4 и 5 в Таблице 4). При 100% необнаружении объектов меньшего класса (атак типа U2R) и одной ошибке классификации нормальных объектов Accuracy равна 0.9992, что означает почти 100% правильность классификации. Это называется парадоксом Accuracy. Данный пример показывает, что функция Accuracy не годится для оценки качества классификации несбалансированных данных.



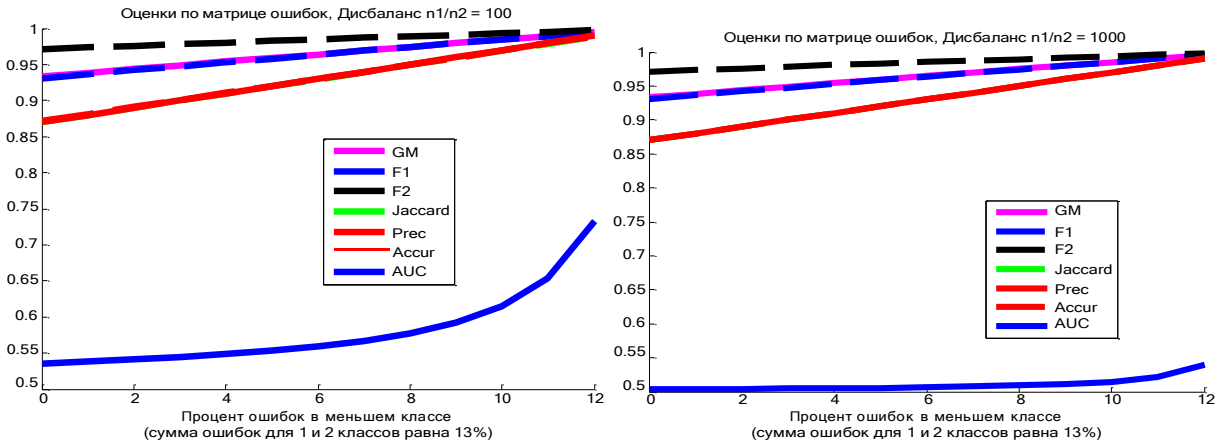


Рисунок *3.* – Графики семи оценочных функций при суммарных ошибках в двух классах равным 13%. Слева вверху без дисбаланса, справа вверху с дисбалансом 1:10, слева внизу с дисбалансом 1:100, справа внизу с дисбалансом 1:1000

На Рисунке 3 показано как изменяются значения семи функций из Таблицы 3 без дисбаланса и при дисбалансе классов 1:1000, процент ошибок в меньшем классе изменяется от 0 до 13%. При отсутствии дисбаланса функция Accuracy имеет постоянное значение, а GM, F1 и AUC имеют примерно такие же значения. При увеличении дисбаланса функция GM практически совпадает с F1 Значения функций Precision, Accuracy и индекса Jaccard также практически совпадают. Значения всех функций кроме AUC выше 0.87 и стремятся к 1 с возрастание числа ошибок в меньшем классе. Значения 1 всех функций соответствуют абсолютно точной классификации данных. График функции AUC демонстрирует наиболее адекватную оценку качества классификации. Значения уменьшаются при росте дисбаланса.

Функции MCC, Cohen's kappa, Youden_index, AUC инвариантны к перестановке классов в матрице ошибок. Отметим что две последних не инвариантны к транспонированию матрицы ошибок. Инвариантность к транспонированию позволяет произвольно записывать значения ошибок I и II рода.

Бинарная классификация часто применяется при разделении на классы больных и здоровых, нормальных и имеющих отклонения, и т.п. Выполним разделение данных базы KDD Train на два класса normal и anomaly. Матрица ошибок приведена в Таблице 5. Дисбаланс классов невелик и равен 1,33.

Таблица 5. –Матрица ошибок разбиения базы KDD Train на два класса: нормальные соединения и угрозы

	anomaly	normal
anomaly	9362	3471
normal	298	9413
число объектов в классе	9660	12884
	1: 1.33 дисбаланс	

Вычислим 12 оценочных функций, приведенных выше. Результаты приведены в Таблице 6. Поскольку дисбаланс невелик функция Accuracy, показавшая 83,28%, достаточно близка к реальной оценке результата классификации. Функции F1, GM, AUC, kappa, MCC имеют примерно такие же значения. Однако с учетом результатов предыдущего эксперимента рекомендуется применять AUC, либо kappa и MCC.

Таблица 6. – Оценки классификации, вычисленные по матрице ошибок из Таблицы 5

функция	значение	функция	значение
Sensitivity (recall)	0.9692	F2	0.9094
Specificity	0.7306	GM	0.8408
Precision	0.7295	AUC	0.8499
Accuracy	0.8328	Kappa	0.8361
Jaccard	0.7130	MCC	0.8496
F1	0.8324		

Функция AUC не идеальна, но она самая простая с точки зрения вычислений и согласно исследованиям, описанным в работе [12] наиболее устойчива к дисбалансу классов анализируемых данных. Авторы работы [13] пришли к аналогичному заключению. Они исследовали четыре варианта представления данных для классификации и пришли к выводу что, если в обучающей выборке присутствует шум или другие искажения, то меры, основанные на вероятности, не очень хороши для анализа результатов классификации, при этом функция AUC является лучшим вариантом оценки качества классификации.

Заключение. В работе затронута важная проблема: прикладная статистика в настоящее время не имеет методов обработки больших данных. Одной из актуальных прикладных задач анализа больших данных является разделение множеств объектов на существенно различные по объему классы, т.е. имеющие дисбаланс. Продемонстрировано, что при оценке результатов классификации данных с дисбалансом не стоит использовать классические функции, такие как Accuracy, Sensitivity, Specificity и F1. Для оценки качества классификации из известных наиболее инвариантной к дисбалансу классов является функция AUC, вычисляющая площадь под ROC кривой. В случае классификации на два класса она равна среднему арифметическому значению функций Sensitivity и Specificity.

Список литературы

[1.]De Mauro A, Greco M, Grimaldi M. A formal definition of Big Data based on its essential features // Library Review.- 2016.- Vol. 65.- No. 3.- C. 122-135.

[2.]Орлов А. И. О развитии прикладной статистики. — В сб.: Современные проблемы кибернетики (прикладная статистика). — М.: Знание, 1981, с.3-14.

[3.]Математическая и прикладная статистика: учеб. пособие / Ю.С. Харин, Е.Е. Жук.-Мн.: БГУ, 2005.- 279с.

[4.]Орлов А.И. Прикладная статистика, М.: Экзамен, 2004.-656с.

[5.]Рак полости рта у белорусов чаще всего находят на последних стадиях. Почему?

[6.]The role of Statistics in the era of big data // Statistics and Probability Letters.- Vol.136.- 2018.- 170p.

[7.]Mayer-Schönberger V., Cukier K. Big data: A revolution that will transform how we live, work, and think. – Houghton Mifflin Harcourt, 2014.

[8.]Ginsberg, J., et.al. Detecting influenza epidemics using search engine query data // Nature.- 2009.- Vol.457 (7232) .- C.1012-1014.

[9.]Sinha A, Sahoo B, Rautaray SS, Pandey M. Analysis of Breast Cancer Dataset Using Big Data Algorithms for Accuracy of Diseases Prediction // Int. Conf. on Computer Networks and Inventive Communication Technologies, 2019.- C. 271-277.

[10.] KDD Cup 1999 Data. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[11.] Odiathevar M., Seah W. K. G., Frean M. A Hybrid Online Offline System for Network Anomaly Detection // 28th Int. Conf. on Computer Communication and Networks, IEEE, 2019. – C.1-9.

[12.] Старовойтов В.В., Голуб Ю.И. Сравнительный анализ оценок качества бинарной классификации // Информатика.- 2020.- Т.17.- №1.

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

PROBLEMS OF STATISTICAL EVALUATIONS IN ANALYSIS OF IMBALANCED CLASSES OF BIG DATA

V.V. Starovoitov,
Doctor of Engineering Sciences
Chief Researcher, Professor

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

Abstract. The article stated that applied statistics is currently not ready for analysis and processing of big data. It is senseless and useless to calculate the average values, variance, and other statistical characteristics for numerous and diverse classes of objects that belong to the category of big data.One of the actual tasks is classification of object sets into classes that are significantly different in volume. There are real problems to divide people into those who are sick and healthy, sorting emails into spam and regular messages, etc. Many data classification methods have been developed. Their results are described by confusion matrices. Using these matrices, one can evaluate quality of classification and choose the best method. To date, the basic functions of classification quality assessment are Accuracy, Sensitivity, Specificity, and F1. As a result of experimental studies, it was found that these functions distort the true assessment of the classification quality in the case of a significant class imbalance. We have shown that to evaluate the binary classification the AUC function is the best among well-known functions. It calculates the area under the ROC curve. In the case of binary classification, it is equal to the mean value of the Sensitivity and Specificity.

Keywords: applied statistics, classification of imbalanced data, Accuracy, Sensitivity, Specificity, F1, AUC.