УДК 004.932

Г.А. Кравцов, канд. техн. наук
Ин-т проблем моделирования в энергетике
им. Г.Е. Пухова НАН Украины
(Украина, 03164, Киев, ул. Генерала Наумова, 15,
e-mail: hryhoriy.kravtsov@gmail.com)

Вычисления на классификациях. Оценка классификаторов

Существующие методы оценки классификаторов оперируют совокупностью классов, которые сопоставимы как по вероятности появления, так и по семантической взаимосвязи, т.е. семантически независимы. Разработанная теория вычислений на классификациях позволяет решать задачу оценки классификаторов на иерархических классификациях. Приведен пример расчета точности и полноты классов иерархической и плоской классификаций при одной и той же матрице неточностей.

Існуючі методи оцінки класифікаторів оперують сукупністю класів, які можуть бути співставлені як за ймовірністю появи, так і за семантичною взаємопов'язаністю, тобто семантично незалежними. Розроблена теорія обчислень на класифікаціях дозволяє розв'язати задачу оцінки класифікаторів на ієрархічних класифікаціях. Наведено приклад розрахунку точності та повноти для класів ієрархічної та плоскої класифікацій за умови тієї ж самої матриці неточностей.

К л ю ч е в ы е с л о в а: классификация, классификатор, семантика, точность, полнота, мера отличия.

Если относительно двух произвольных объектов выполнена задача определения классовой принадлежности [1] (т.е. определен класс каждого объекта), то модель вычислений на классификациях [2] позволяет определить меру отличия этих объектов в одной пространственной классификации. Если есть две произвольные классификации, то существует дуальная мера [1], отражающая семантические и структурные отличия. Полученные результаты теоретических исследований позволяют решить ряд прикладных задач, в том числе задачу оценки классификаторов, важность которой сложно переоценить в теории и практике машинного обучения [3, 4]. Однако до настоящего времени остаются неизвестными результаты исследований, посвященных оценке работы классификаторов на иерархических (многоуровневых) классификациях. Обычно задача определения

классовой принадлежности решается для множества классов, сопоставимых как по вероятности появления, так и семантически.

Рассмотрим задачу оценки классификатора при определении классовой принадлежности на иерархической классификации. Термины «классификация», «классифицирование», «классовая принадлежность» приведены [1]. Объект, в отношении которого выполнена задача определения классовой принадлежности, называется классифицированным объектом [2].

Согласно [5] классификатор: «1. Специалист по классификации; лицо, занимающееся классификацией. 2. Прибор для сортировки руды по крупности зерен (горн.)». В то же время, термин «классификатор» используется в значении систематизированного перечня именованных объектов, каждому из которых в соответствие дан уникальный код (например, Классификатор профессий Украины). В работе [6] термин «классификатор» используется в значении механизма (инструмента) определения классовой принадлежности.

Предлагается использовать термин «классификатор» в значении некоторой сущности, обладающей способностью определения классовой принадлежности (присваивание объекту некоторой метки класса). Здесь будем использовать слово «сущность», так как функции классификатора может выполнять человек, машина и человеко-машинный симбиоз. Данное определение хорошо соответствует формальному определению из работы [3]: «Классификатором называется отображение $\hat{c}: X \to C$, где $C = \{C_1, C_2, ..., C_k\}$ — конечное и обычно небольшое множество меток классов».

Очевидно, что при решении задачи определения классовой принадлежности классификатор может допускать ошибки. В теории и практике машинного обучения [3] разработан подход, позволяющий оценивать классификаторы с качественной стороны в количественных оценках: правильность, точность, полнота и F-мера. Указанные метрики определяют возможность классификатора выполнить задачу установления классовой принадлежности для некоторого тестового набора данных.

Согласно [3, 7] под правильностью (accuracy) классификатора понимают отношение числа правильно принятых классификатором решений к размеру тестовой выборки: $\text{Accuracy} = P / N$, где P — число документов, по которым классификатор принял правильное решение; N — размер обучающей выборки.

На примере классификации документов по числу классов для определения точности Д. Баженов [7] пишет: «...у этой метрики есть одна особенность, которую необходимо учитывать. Она присваивает всем документам одинаковый вес, что может быть не корректно в случае, если распределение документов в обучающей выборке сильно смещено в сторону

какого-то одного или нескольких классов. В этом случае у классификатора есть больше информации по этим классам и соответственно в рамках этих классов он будет принимать более адекватные решения. На практике это приводит к тому, что вы имеете accuracy, скажем, 80%, но при этом в рамках какого-то конкретного класса классификатор работает из рук вон плохо, не определяя правильно даже треть документов».

Для оценки качества работы классификатора чаще используют показатели точности (precision) и полноты (recall) [3], полагая при этом, что классификация является неизменной.

В работе [7] следующим образом поясняется суть указанных метрик (под словом система подразумевается классификатор): «Точность системы в пределах класса — это доля документов, действительно принадлежащих данному классу относительно всех документов, которые система отнесла к этому классу. Полнота системы — это доля найденных классификатором документов, принадлежащих классу относительно всех документов этого класса в тестовой выборке».

Значения точности и полноты для каждого класса могут быть рассчитаны на основании следующей матрицы контингентности:

Категория (класс) A_i		Экспертная оценка	
		Положительная	Отрицательная
Оценка системы (классификатора)	Положительная	*TP*	*FP*
	Отрицательная	*FN*	*TN*

Здесь *TP* — истинно-положительное решение; *TN* — истинно-отрицательное решение; *FP* — ложно-положительное решение (ошибка первого рода [8]); *FN* — ложно-отрицательное решение (ошибка второго рода [8]). Тогда точность и полнота могут быть рассчитаны по формулам

$$\text{Precision} = \frac{TP}{TP+FP}, \quad \text{Recall} = \frac{TP}{TP+FN}.$$

Точность работы классификатора на выбранном классе определяется отношением числа объектов, которые отнесены классификатором к выбранному классу, к числу объектов, отнесенных к выбранному классу классификатором и экспертами.

Полнота определяется отношением числа объектов, корректно отнесенных классификатором к некоторому классу, к числу объектов, отнесенных к этому же классу экспертами.

Как указано в работе [7], на практике значения точности и полноты удобно рассчитывать с использованием матрицы неточностей (confusion

matrix). Если число классов относительно невелико (не более 100—150), этот подход позволяет наглядно представить результаты работы классификатора.

Матрица неточностей — это матрица размера $N \times N$, где N — число классов. Столбцы этой матрицы резервируются согласно экспертным решениям, а строки — согласно решениям классификатора. Когда выполняется задача определения классовой принадлежности документа из тестовой выборки, в матрице неточностей увеличивается на единицу число, стоящее на пересечении строки класса, определенного классификатором, и столбца класса, к которому действительно относится документ.

Пусть $P_{i,j}$ — число объектов класса A_i, классифицированных классификатором как относящиеся к классу A_j. Некоторая матрица неточностей имеет следующий вид:

	Полнота	0,96	0,94	0,72	Recall (A_j)	1,0
Точность	Классы	A_1	A_2	A_3	A_j	A_N
0,95	A_1	94	0	0	...	0
1,00	A_2	0	32	0	...	0
0,29	A_3	1	1	6	...	0
Precision (A_i)	A_i	...	...	...	$P_{i,j}$	...
0,98	A_N	0	0	1	...	78

С помощью такой матрицы рассчитать точность и полноту для каждого класса достаточно просто [7]. Точность равняется отношению соответствующего диагонального элемента матрицы и суммы всей строки класса, полнота — отношению диагонального элемента матрицы и суммы всего столбца класса. Так, для N классов справедливы формулы

$$\text{Precision}(A_i) = P_{i,i} \Big/ \sum_{j=1}^{N} P_{i,j}; \tag{1}$$

$$\text{Recall}(A_j) = P_{j,j} \Big/ \sum_{i=1}^{N} P_{i,j} \ . \tag{2}$$

На рис. 1 представлены два возможных варианта отношений между классами. Для одноуровневой классификации (рис. 1, *а*) справедливо неравенство $1 < i < N$, где N — число классов классификации; A_i — произвольный класс. Для многоуровневой классификации (рис. 1, *б*) верны

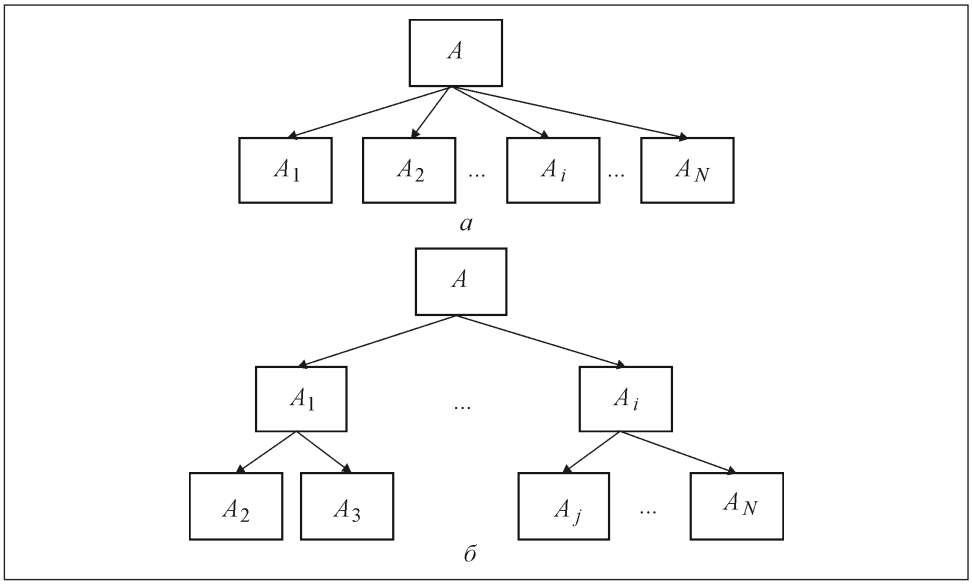

Рис.1. Одноуровневая (*а*) и многоуровневая (*б*) классификации

следующие ограничения: $1< i< N, 1< j< N, i \neq j$. Необходимо заметить, что представленная на рис. 1, *б*, двухуровневая классификация является частным случаем многоуровневой, в которой неделимые классы [1] (листья ориентированного дерева) могут иметь различные ранги [2] .

Напомним [2], что мера отличия на классификации определяется так:

$$\overline{O}(A_I, A_Y) = 1 - \frac{R(A, A_I \cdot A_Y) + 1}{R(A, A_I \cdot A_Y) + R(A_I, A_I \cdot A_Y) + R(A_Y, A_I \cdot A_Y) + 1}, \qquad (3)$$

где A_I и A_Y — классы в A: I, Y, — произвольные пути уточнения плоскости деления классификации A^i; R — относительное расстояние между классами A_I и A_Y иклассификации A, равное числу уникальных операций уточнения от ближайшего общего обобщающего класса. Если даны классы $A^i_{[a]}$, $A^i_{[b]}$, $A^i_{[a,\,b]}$ классификации A, ассоциативная бинарная операция обобщения классов классификации имеет вид [2]

$$A^i_{[a]} \cdot A^i_{[b]} = A,\ A^i_{[a]} \cdot A^i_{[a,\,b]} = A^i_{[a]},\ A^i_{[b]} \cdot A^i_{[a,\,b]} = A,\ A^i_{[a]} \cdot A = A.$$

Здесь A является идемпотентом, или нулем, операции обобщения классов относительно самой себя, $A \cdot A = A$, и всех уточняющих классов классификации $A^i_I \cdot A = A$.

Согласно (3) мера отличия для одноуровневой классификации между двумя любыми классами A_i и A_j имеет вид $\overline{Q}(A_i, A_j) = 0$, если $i = j$,

и $\overline{Q}(A_i, A_j) = 2/3$, если $i \neq j$. Для многоуровневой классификации $\overline{Q}(A_i, A_j) = 0$, если $i = j$. Если $i \neq j$, ничего определенного сказать нельзя.

Введем мультипликатор ошибки классифицирования $M_{I,Y}$, определяющий скорректированную меру отличия между классами классификации:

$$\begin{gathered} M_{I,Y} = \overline{O}(A_I, A_Y), \text{ если } I \neq Y, \\ M_{I,Y} = 1, \text{ если } I = Y. \end{gathered} \tag{4}$$

Представим (1) в виде уравнения

$$\text{Precision}(A_i) = \frac{M_{i,i}P_{i,i}}{\sum_{j=1}^{N} M_{i,j}P_{i,j}} = \frac{P_{i,i}}{\sum_{j=1}^{N} M_{i,j}P_{i,j}} \tag{5}$$

и проанализируем его свойства. Будем полагать, что (5) есть теоретическая точность классификатора на классификации. Очевидно, что для (5) выполняется условие

$$\frac{P_{i,i}}{\sum_{j=1}^{N} M_{i,j}P_{i,j}} \geq \frac{P_{i,i}}{\sum_{j=1}^{N} P_{i,j}},$$

т.е. теоретическая точность больше или равна используемой на практике. Если $P_{i,j} = 0$ при $i \neq j$, а $P_{i,i} \neq 0$, то теоретическая и практическая точности равны единице:

$$\frac{P_{i,i}}{\sum_{j=1}^{N} M_{i,j}P_{i,j}} = \frac{P_{i,i}}{\sum_{j=1}^{N} P_{i,j}} = 1.$$

По аналогии с (5) формулу полноты (2) представим в виде

$$\text{Recall}(A_j) = \frac{M_{j,j}P_{j,j}}{\sum_{i=1}^{N} M_{i,j}P_{i,j}} = \frac{P_{j,j}}{\sum_{i=1}^{N} M_{i,j}P_{i,j}}. \tag{6}$$

Будем полагать, что (6) есть теоретическая полнота классификатора на классификации. Очевидно, что для (6) выполняется условие

$$\frac{P_{j,j}}{\sum_{i=1}^{N} M_{i,j}P_{i,j}} \geq \frac{P_{j,j}}{\sum_{i=1}^{N} P_{i,j}},$$

т.е. теоретическая полнота больше или равна используемой на практике.

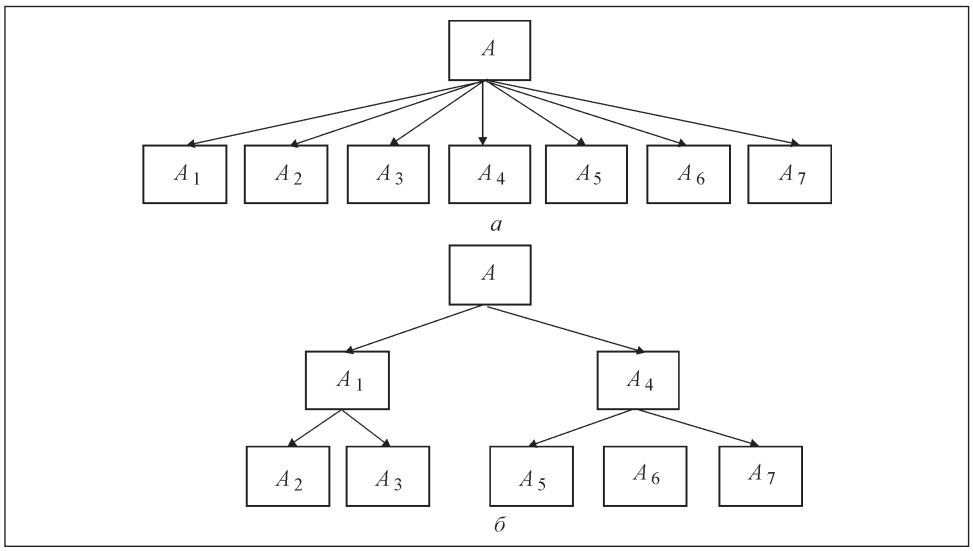

Рис. 2. Одноуровневая (*а*) и многоуровневая (*б*) классификации из семи классов (частный случай соответственно для рис. 1, *а* и *б*)

Следовательно, классификатор работает корректно, если выполняется условие $\mathrm{Recall}(A_i)=1$ и $\mathrm{Precision}(A_i)=1$, где $1<i<N$, т.е. для любого класса точность и полнота равны единице.

Рассмотрим изложенное на примере. Пусть даны две классификации, представленные на рис. 2, и следующая гипотетическая матрица неточностей (confusion matrix) для семи классов:

Классы	A_1	A_2	A_3	A_4	A_5	A_6	A_7
A_1	9	0	0	0	1	0	1
A_2	0	7	0	2	0	1	0
A_3	0	0	10	0	0	0	2
A_4	5	0	0	6	0	3	0
A_5	0	1	0	0	3	0	0
A_6	0	0	0	0	0	5	0
A_7	1	0	3	0	0	0	8

Для сравнения представим вычисления точности и полноты для каждого из классов классификаций, изображенных на рис. 2, *а* и *б*, с использованием матрицы неточности. Как видно из табл. 1 и 2, при одной и той

Таблица 1

Класс	Точность определения класса в классификации	
	одноуровневой (рис. 2, *а*)	многоуровневой (рис. 2, *б*)
A_1	0.870967741935484	0.8571428571428571
A_2	0.7777777777777778	0.7526881720430108
A_3	0.8823529411764706	0.8620689655172414
A_4	0.6428571428571428	0.7058823529411765
A_5	0.8181818181818181	0.8
A_6	0.6521739130434783	0.7352941176470589
A_7	0.8	0.7729468599033816

Таблица 2

Класс	Полнота определения класса в классификации	
	одноуровневой (рис. 2, *а*)	многоуровневой (рис. 2, *б*)
A_1	0.6923076923076923	0.6878980891719745
A_2	0.9130434782608695	0.8974358974358975
A_3	0.8333333333333334	0.8064516129032258
A_4	0.6428571428571428	0.6428571428571428
A_5	0.8181818181818181	0.7894736842105263
A_6	1.0	1.0
A_7	0.7499999999999999	0.7174887892376681

же матрице неточности для одного и того же числа классов, но с различной семантической структурой, показатели точности и полноты зависят от места класса в классификации и могут существенно различаться. Приведенный пример свидетельствует о том, что показатели полноты и точности изменяются при учете семантической взаимосвязи классов в классификации.

Учет семантической структуры классификации становится более актуальным в случае работы с обобщающими понятиями (umbrella terms) при классификации документов [9], что является доказательством востребованности разрабатываемой теории вычислений на классификациях в различных областях научных знаний. Представляет также интерес исследование влияния структуры классификации на показатели точности и полноты.

СПИСОК ЛИТЕРАТУРЫ

1. *Кравцов Г.* Мера отличия классификаций // Электрон. моделирование. — 2016. — **38**, № 4. — С. 81—97.
2. *Кравцов Г.* Модель вычислений на классификациях // Там же. — 2016. — **38**, № 1. — С. 73—87.
3. *Флах П.* Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных — М. : Изд-во «ДМК Пресс», 2015. — 400 с.
4. *Кохонен Т.* Самоорганизующиеся карты. — М. : Бином. Лаборатория знаний, 2008. — 655 с.
5. *Толковый* словарь русского языка / Под ред. Д.Н. Ушакова. Т. 1. — М. : Гос. ин-т «Сов. энцикл.»; ОГИЗ; Гос. изд-во иностр. и нац. слов, 1935.
6. *Хайкин С.* Нейронные сети. Полный курс — М. : Изд. дом «Вильямс», 2006. — 1104 с.
7. *Баженов Д.* Оценка классификатора (точность, полнота, F-мера). — [Электронный ресурс]. — Режим доступа: http://bazhenov.me/blog/2012/07/21/classification-performance-evaluation.html. — Дата доступа: май 2016.
8. *Харман Г.* Современный факторный анализ. — М. : Статистика, 1972. — 486 с.
9. *Struhl S.* Practical Text Analytics: Interpreting Text and Unstructured Data for Business Intelligence. Kogan Page; ed. — London, Philadelphia, New Delphi, 2015. — P. 272.

H.A. Kravtsov

THE CALCULUS OVER CLASSIFICATIONS.
ASSESSMENT OF CLASSIFIERS

The existing methods of classifier assessment use a set of classes which are comparable both by the probability of appearance and by semantical interrelation that is they are semantically independent. The developed theory of calculus over classification permits solving the issue of classifier assessment for hierarchical classifications. This papper contains the example of calculation of the precision and completeness of classes of plane-level and multi-level classification with the same confusing matrix.

K e y w o r d s: classification, classifier, semantic, precision, completeness, measure of difference.

REFERENCES

1. Kravtsov, H.A. (2016), “Measure of difference between classifications”, *Elektronnoe modelirovanie*, Vol. 38, no. 4, pp. 81-97.
2. Kravtsov, H.A. (2016), “Model of computations on classifications”, *Elektronnoe modelirovanie*, Vol. 38, no. 1, pp. 73-87.
3. Flakh, P. (2015), *Machinnoe obuchenie* [Machine learning: The art and science of algorithms that make sense of data], Izd-vo “DMK Press”, Moscow, Russia.
4. Kokhonen, T. (2008), *Samoorganizuyuschiesya karty* [Self-organizing maps], Binom, Moscow, Russia.
5. Ushakov, D.N. (1935), *Tolkovyi slovar russkogo yazyka* [Russian definition dictionary], Sovetskaya entsiklopediya, Moscow, Russia.
6. Khaykin, S. (2006), *Neyronnye seti. Polnyi kurs* [Neural networks: A comprehensive foundation], Izd. “Vilyams”, Moscow, Russia.
7. Bazhenov, D. (2012), *Otsenka klassifikatora. Tochnost, polnota, F- mera* [Classification performance evaluation. Precision, completeness, F-measure], available at: http://bazhenov.me/blog/2012/07/21/classification-performance-evaluation.html (accessed 2016).

8. Kharman, H.H. (1972), *Sovremennyi faktornyi analiz* [Modern Factor Analysis], Statistika, Moscow, Russia.
9. Struhl, S. (2015), Practical text analytics: Interpreting text and unstructured data for business intelligence, 1st edition, Kogan Page, London, Philadelphia, New Delphi.

Поступила 21.09.16;
после доработки 13.10.16

КРАВЦОВ Григорий Алексеевич, канд. техн. наук, докторант Ин-та проблем моделирования в энергетике им. Г.Е. Пухова НАН Украины. В 2000 г. окончил Севастопольский военно-морской ин-т им. П.С. Нахимова. Область научных исследований — кибербезопасность смарт-грид, криптография, программирование, разработка распределенных гетерогенных вычислительных систем.