УДК 004.822

Классификация узлов ассоциативно-вербальной сети по когнитивным областям: этап оценки классификаторов и классификации узлов[1].

Ю.Н. Филиппович, А.В. Сиренко

Ю.Н. Филиппович – к.т.н., профессор каф. Медиасистем и технологий, Московского государственного университета печати им. Ивана Фёдорова

А.В. Сиренко – соискатель каф. Медиасистем и технологий, Московского государственного университета печати им. Ивана Фёдорова

y_philippovich@mail.ru, alexander.sirenko@gmail.com

Статья продолжает материал публикации, посвященной построению классификаторов узлов ассоциативно-вербальной сети по когнитивным областям. Приведены свойства рассчитанных классификаторов для ряда областей, оценка качества моделей выполняется посредством ROC-анализа. Рассмотрено использование построенных классификаторов для определения когнитивных областей узлов.

The paper continues publication, devoted to construction of classifiers for nodes of associative-verbal network. Here we provide properties of calculated classifiers and apply ROC-analysis of their quality. Constructed classifiers are considered to classify nodes.

Ключевые слова: ассоциативная сеть, классификация, когнитивная область, регрессионный анализ.

Key words: Associative network, classification, cognitive domain, regression analysis.

[1] В данной статье приводятся результаты исследований, выполненных при поддержке грантов РГНФ №12-04-12039в, №12-04-12059в и гранта Президента РФ №НШ-3661.2012.6

1. Введение.

В нашей статье, опубликованной в шестом номере журнала «Проблемы полиграфии и издательского дела» за 2012 год [1] описано построение классификаторов когнитивных областей для узлов ассоциативно-вербальной сети автоматизированной системы научных исследований психолингвистических экспериментов [2]. В качестве классификатора выбрана модель логистической регрессии:

$$r(Y) = \frac{1}{1+e^{-Y}}, \quad (1)$$

где Y — прогнозируемая линейная комбинация предикторных величин, которыми являются свойства контекста узлов ассоциативно-вербальной сети.

$$Y = c_{inter} + c_{in} * X_{in} + c_{out} * X_{out} + c_{com} * X_{com}, \quad (2)$$

где $c_{inter}, c_{in}, c_{out}, c_{com}$– коэффициенты регрессии;

X_{in}, X_{out}, X_{com}– предикторные переменные;

Одним из методов оценки качества модели линейной регрессии является ROC-анализ (Receiver Operator Characteristic) и вычисление показателя AUC – площади под графиком ROC [2,3].

При выполнении ROC-анализа требуется определить чувствительность (Se) и специфичность (Sp) модели:

$$Se = \frac{TP}{TP+FN},$$

$$Sp = \frac{TN}{TN+FP},$$

где TP – верно классифицированные положительные примеры;

FN – неверно классифицированные отрицательные примеры;

TN – верно классифицированные отрицательные примеры;

FP – неверно классифицированные положительные примеры;

Изменением порога отсечения классификационной модели получаем пары значений (Se, Sp), формирующие график ROC - зависимость Se от $(1 - Sp)$.

В таблице 1 представлены классификационные модели с наибольшим показателем AUC. Для ряда областей модель построить не удалось ввиду недостаточного количества положительных случаев классификации в обучающей выборке.

2. Оценка качества классификационной модели.

Между графиком ROC и традиционной характеристикой «точность-полнота» (precision-recall, сокращенно PR) есть взаимосвязь [4]: модель доминирует в пространстве ROC-кривой, тогда и только тогда, когда она доминирует в пространстве PR-кривой.

Дополнительно к ROC-анализу будем использовать вероятностные свойства коэффициентов предикторных переменных, а также показатели чувствительности и специфичности модели.

3. Классификация узлов сети.

В результате расчета параметров модели нам доступны величины:

$C_{in}, C_{out}, C_{com}, C_{inter}$ – математические ожидания коэффициентов входящих, исходящих и промежуточных связей, а также смещения регрессионной функции (2);

$\sigma_{in}, \sigma_{out}, \sigma_{com}, \sigma_{inter}$– среднеквадратичные отклонения соответствующих коэффициентов.

Программная среда R [5] при использовании функции классификации `predict` возвращает значение r формулы (1) и ее стандартное отклонение. Как было сказано ранее, мы не можем трактовать эту величину как вероятность принадлежности к области и использовать стандартное отклонение для расчета доверительных интервалов. Используем для классификации результаты ROC-анализа и параметры модели.

Ниже представлены параметры классификатора для области «дорога».

AUC: 0,702	C_{in}: 1,818	C_{out}: 3,535	C_{com}: 2,559	C_{inter}: -5,739
	σ_{in}: 3,954	σ_{out}: 4,099	σ_{com}: 0,397	σ_{inter}: 0,267

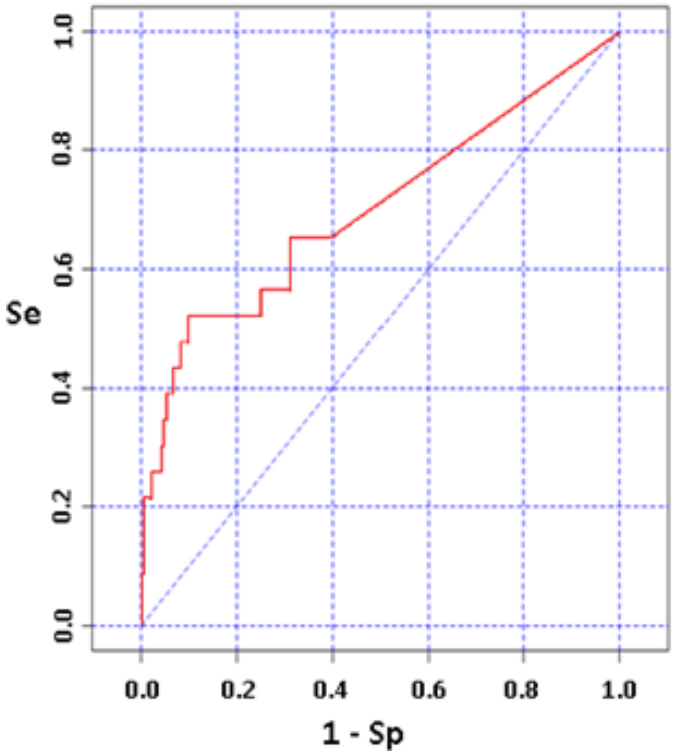

Рисунок 1. ROC-кривая классификатора области “дорога”.

Для определения принадлежности узла к области при некотором пороге отсечения, рассчитаем зависимость чувствительности от порога отсечения $Se = f(cutoff)$.

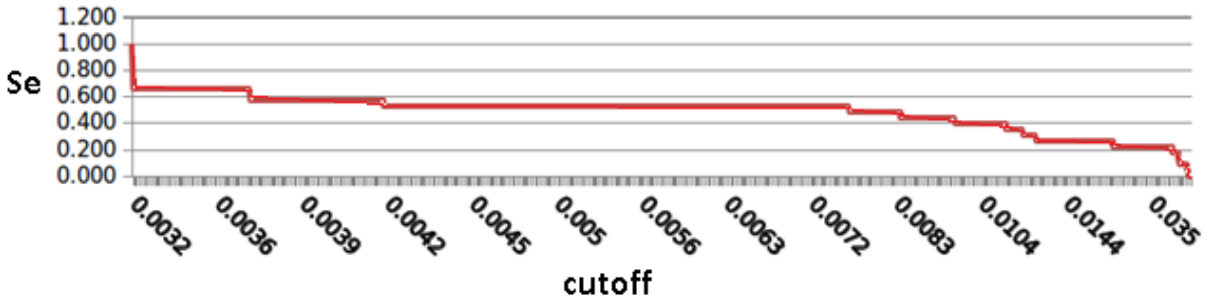

Рисунок 2. Зависимость чувствительности (Se) от порога отсечения (cutoff).

Чувствительность Se при некотором пороге отсечения $cutoff$ составляет вероятность принадлежности узла к области при значении регрессионной функции $r \geq cutoff$.

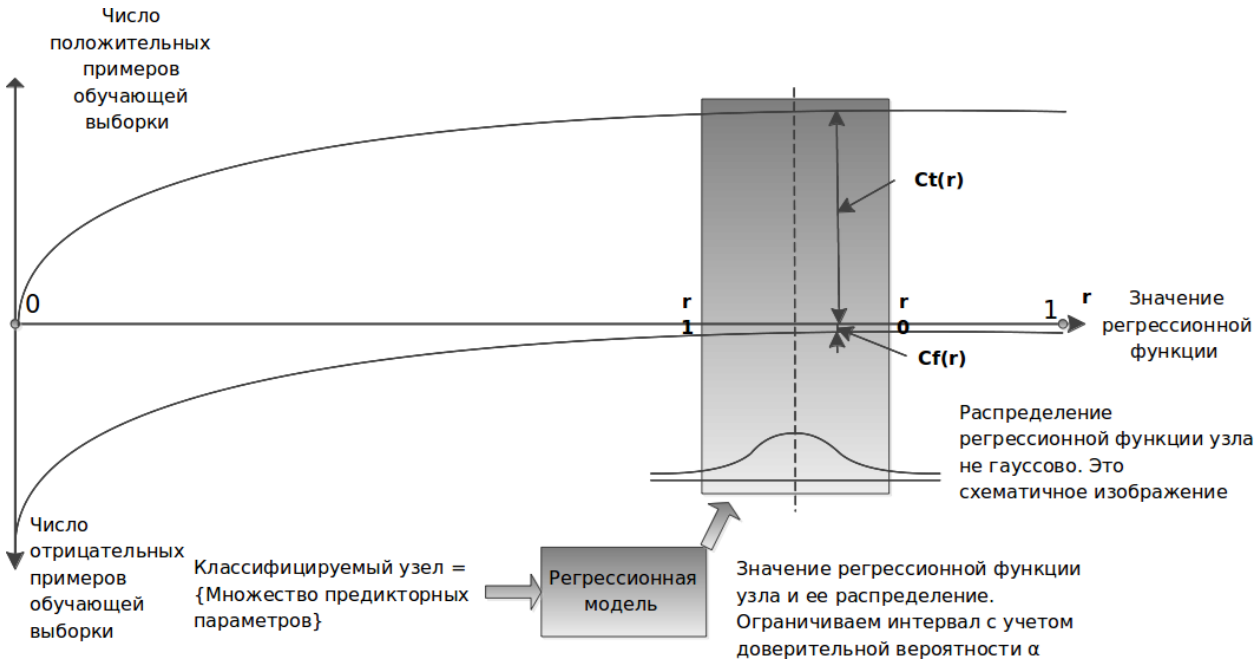

Рисунок 3. Расчет принадлежности узла области.

Пусть зависимость числа положительных примеров для некоторого r равна $Ct(r)$, отрицательных $Cf(r)$. Обозначим как функцию

$Part(r) = \frac{Ct(r)}{Ct(r)+Cf(r)}$, показывающую для каждого значения регрессионной функции статистическую вероятность принадлежности узла к области. $Part(r)$является результатом обработки узлов обучающей выборки. Тогда вероятность принадлежности узла к области будет определяться выражением:

$$P_{inarea} = \int_{r_0}^{r_1} Part(r) * f(r) * dr, \quad (2)$$

где r_0, r_1– пределы значений регрессионной функции для рассматриваемого доверительного интервала;

$f(r)$– функция распределения значения регрессионной функции.

Для заданной доверительной вероятности α и узла v_jграницы логистической функции r определяются ее минимальным и максимальным значениями:

$$r(v_j, \alpha) = [r_{min}, r_{max}].$$

В итоге получаем выражение для определения вероятности принадлежности узла области:

$P_{inarea} = \int_{r_{min}}^{r_{max}} Part(r) * f(r) * dr =$| Переходот r к y | =

$$\int_{y_{min}}^{y_{max}} Part\left(\frac{1}{1+e^{-1}}\right) * g(y) * r'_y * dy=$$

$$\int_{y_{min}}^{y_{max}} Part(\frac{1}{1+e^{-1}}) * q(y) * \frac{e^{-y}}{(1+e^{-y})^2} * dy=$$

$$\int_{y_{min}}^{y_{max}} \frac{e^{-y}}{(1+e^{-y})^2} * q(y) * Part\left(\frac{1}{1+e^{-y}}\right) * dy, \text{где}$$

$q(y)$– распределение величины y, определенной ранее через сумму случайных величин (2);

$$y_{min} = \log(\frac{r_{min}}{1+r_{min}});$$

$$y_{max} = \log\left(\frac{r_{max}}{1+r_{max}}\right).$$

В функции случайных величин yзначения x_{in}, x_{out}, x_{com}для рассматриваемого узла заданы. Параметры $C_{in}, C_{out}, C_{com}, C_{inter}$являются случайными с известными параметрами распределений.

4. Заключение.

Построенные регрессионные модели позволили оценить значимость выделенных прогностических признаков, а также выявить направления доработки классификатора. Показатель AUC может применяться для косвенной оценки моделей в совокупности с вероятностными характеристиками параметров моделей, но не в качестве единственного критерия ее качества.

Для улучшения прогностических качеств моделей требуется:

1. Изменить критерий отнесения узла к отрицательной выборке (например, подбором отрицательных примеров из областей, минимально пересекающихся с областью положительных примеров);

2. Увеличить число обучающих примеров по большинству областей,с использованием лексикографических объектов с определенной когнитивной областью (тематических словарей);

3. Учесть связи сети, представляющие собой модель гиперграфа, которые позволят увеличить длину учитываемых путей при расчете контекста узла (таблица 1).

На основе идеи когнайзера могут быть построены принципиально новые коммуникативные системы и технологии (программно-целевой деятельности, общения, обучения, поиска информации, синтеза и анализа мультимедиа-контента, автоматического перевода, извлечения и контроля знаний и др.), а также устройства и приспособления — интеллектуальные гаджеты, когнитивные акселераторы.

Область	AUC	C in	C in СКО	C out	C out СКО	C com	C com СКО	C inter	C inter СКО
электричество	0,944	-409,356	429170,362	-1656,263	545836,625	13,931	6,486	-7,970	0,826
эстрада	0,901	-545,860	968668,677	-3995,677	1032112,949	6,010	4,174	-8,022	0,883
мораль	0,864	27,307	8,261	3,972	3,369	3,917	1,521	-6,941	0,507
тайна	0,853	-678,332	1102024,537	-4520,675	1126775,038	3,331	4,823	-7,784	0,799
энтомология	0,839	16,055	16,406	-7,610	17,840	9,318	2,384	-7,688	0,688
рыбалка	0,834	1,964	3,300	2,691	6,269	4,199	0,974	-6,774	0,445
эмоции	0,805	-253,812	508241,543	-3165,856	980186,085	12,217	2,528	-8,570	1,000
поговорка	0,776	0,289	13,376	10,941	46,866	5,009	3,358	-7,158	0,558
стихия	0,776	23,611	29,062	136,855	187,919	-13,417	23,503	-7,577	0,712
праздник	0,767	2,065	3,043	5,372	3,541	2,618	0,451	-5,915	0,288
охота, рыбалка	0,762	3,936	3,203	36,758	11,392	-1,282	1,655	-5,196	0,221
геометрия	0,739	0,673	2,612	12,938	4,157	2,023	0,378	-5,472	0,232
рыбы	0,738	-691,124	113129,562	3806,224	750748,862	5,724	1,515	-7,157	0,555
математика	0,737	4,359	4,849	1,447	7,302	3,425	0,732	-6,389	0,377
семья	0,736	5,886	3,392	2,229	3,449	1,365	0,241	-5,445	0,231
природные явления	0,731	0,326	0,971	5,670	1,584	0,902	0,110	-4,842	0,176
интеллект	0,725	4,251	2,594	4,066	3,240	2,743	0,474	-6,035	0,310
лес	0,721	9,970	3,108	10,849	3,557	0,426	0,461	-5,892	0,296
астрономия	0,719	9,121	2,262	3,164	1,574	1,057	0,297	-5,070	0,207
секс	0,712	8,986	9,125	15,561	14,390	4,693	1,595	-6,422	0,384
деньги	0,709	6,144	2,586	1,688	3,244	0,689	0,058	-5,627	0,179
письменный текст	0,707	-5082,812	633999,428	-4300,358	857812,739	5,584	3,161	-6,626	0,451
чувства	0,705	4,804	1,352	7,954	1,668	0,739	0,134	-4,828	0,180
напитки	0,704	16,314	7,452	0,376	6,098	5,025	1,425	-7,087	0,525
дорога	0,702	1,818	3,954	3,535	4,099	2,559	0,397	-5,739	0,267

Таблица 1. Параметры классификационных моделей для областей с максимальной AUC.

Библиографический список

1. *Филиппович Ю.Н., Сиренко А.В.* Классификация узлов ассоциативно-вербальной сети по когнитивным областям: этап построения классификаторов. // Проблемы полиграфии и издательского дела. – 2012. – № 5

2. *Филиппович Ю.Н., Сиренко А.В.* Программный комплекс исследований психолингвистической модели вербального сознания на основе когнитивного и ассоциативного экспериментов // Вопросы психолингвистики. – 2011. – № 1. – С. 126-139.

3. Паклин Н. Логистическая регрессия и ROC-анализ - математический аппарат. Режим доступа: http://www.basegroup.ru/library/analysis/regression/logistic/ (дата обращения: 25.01.2011).

4. Davis J., Goadrich M. The Relationship Between Precision-Recall and ROC Curves. Pittsburgh, PA, 2006.

5. The R Project for Statistical Computing. Режим доступа: http://www.r-project.org/ (дата обращения: 25.01.2011).