Distributed Computing. Vol. 63, 2003. – P. 551-563.
12. PACX-MPI: The Grid-computing library PACX-MPI extending MPI for computational Grid http://www.hlrs.de/organization/av/amt/research/pacx-mpi/ , свободный.
13. Глушков В.М., Цейтлин Г.Е., Ющенко Е.Л. Методы символьной мультиобработки. Киев: Наукова думка, 1980. – 252 с.
14. Ющенко Е.Л., Цейтлин Г.Е., Грицай В.П., Терзян Т.К. Многоуровневое структурное проектирование программ. Теоретические основы, инструментарий. М.: Финансы и статистика, 1989. – 208 с.
15. Капитонова Ю.В., Летичевский А.А. Математическая теория проектирования вычислительных систем. М.: Наука, 1988. – 296 с.
16. Зинкин С.А. Самомодифицируемые сценарные модели функционирования систем и сетей хранения и обработки данных (базовый формализм и темпоральные операции // Известия ВУЗов. Поволжский регион. Технические науки. №1, 2007. – С. 3-12.
17. Зинкин С.А. Самомодифицируемые сценарные модели функционирования систем и сетей хранения и обработки данных (реализация и свойства сценарных моделей) // Известия ВУЗов. Поволжский регион. Технические науки. №2, 2007. – С. 13-21.
18. Зинкин С.А. Сети абстрактных машин высших порядков в проектировании систем и сетей хранения и обработки данных (базовый формализм и его расширения) // Известия ВУЗов. Поволжский регион. Технические науки. №3, 2007. –С. 13-22.
19. Зинкин С.А. Сети абстрактных машин высших порядков в проектировании систем и сетей хранения и обработки данных (механизмы интерпретации и варианты использования) // Известия ВУЗов. Поволжский регион. Технические науки. №4, 2007. – С. 37-51.

TOWARDS THE MESSAGE PASSING INTERFACE IMPLEMENTATION AS A CLOUD SERVICE IN AGENT-ORIENTED METACOMPUTING SYSTEMS

Vashkevich N.P., Zinkin S.A., Karamysheva N.S.

In this paper we propose the models of the logic of managing global computer processes in computer networks with agent-oriented metacomputing cloud services based on the collective data transfers, allowing you to simplify the implementation of massively parallel applications in large-scale distributed systems. Offered for logical control of discrete processes to use the formalism of predicate asynchronous networks, allows abstract and structural synthesis of the functional architecture of agent-based metacomputer with the implementation of global and collective computing operations for a group of agents – the virtual nodes of metacomputer.

Keywords: *metacomputer, cloud computing, the logical process management, collective and computing opera-tions.*

Вашкевич Николай Петрович, Заслуженный деятель науки и техники РФ, д.т.н., профессор Кафедры вычислительной техники (ВТ) Пензенского государственного университета (ПГУ). Тел. (8-412) 36-82-27. E-mail: vt@alice.pnzgu.ru

Зинкин Сергей Александрович, д.т.н., профессор Кафедры ВТ ПГУ. Тел. (8-412) 36-82-27. E-mail: zsa49@yandex.ru

Карамышева Надежда Сергеевна, к.т.н., инженер-программист Кафедры ВТ ПГУ. Тел. (8-412) 36-82-27. E-mail: kagliari@yandex.ru

УДК 004.492.4

ПОСТРОЕНИЕ ОПТИМАЛЬНОГО СПАМ-ФИЛЬТРА НА ОСНОВЕ СОВМЕЩЕНИЯ СТАТИСТИЧЕСКИХ КЛАССИФИКАТОРОВ

Тарасов В.Н., Мезенцева Е.М.

В статье представлены критерии оптимальности для классификации сообщений на основе статистических методов с учетом ошибок первого и второго родов, доли ложных срабатываний и пропущенного спама. Приведены особенности тестирования и обучения спам-фильтра. Предложен подход к организации классификатора, который заключается в совместном использовании методов Байеса и Фишера. Для повышения качества фильтрации предложено использовать механизм анализа подмножеств пере-

сечения множеств, распознанных обоими методами по категориям (спам – не спам, ложные срабатывания и пропуск спама).

Ключевые слова: классификация сообщений, статистические методы, ошибки первого и второго рода, доля пропущенного спама, доля ложных срабатываний, подмножества пересечений, совмещенный фильтр.

Введение

В предыдущих работах авторов освещались вопросы реализации спам-фильтра в интерактивных частях сайтов в сети Internet на основе применения модифицированных методов Байеса и Фишера [1-2]. В данной статье излагаются подходы к тестированию разработанных классификаторов и оценке работы совмещенного фильтра.

Критерии оптимальности при классификации сообщений на основе статистических методов

Предположим, что вся совокупность сообщений подразделяется на классы A и B, где A – класс спам-сообщений, B – класс не спам-сообщений. Задача соотнесения сообщения к одному из этих классов напрямую связана со статистической проверкой гипотез: простой гипотезы H_A: $X \in A$ против альтернативной H_B: $X \in B$, где X есть признак, характеризующий сообщения. Как известно из математической статистики, при этом если сообщение принадлежит классу A, а его соотнесли к B, то будет допущена ошибка первого рода, условная вероятность которой равна α – уровень значимости. Это будет ошибкой выбора альтернативной гипотезы H_B вместо правильной H_A. Если же справедлива гипотеза H_B, но при этом выбрана H_A, то будет допущена ошибка второго рода, условная вероятность которой равна β.

Ошибка первого рода, или ложноотрицательная ошибка, будет допущена, если спам-фильтр ошибочно пропускает нежелательное сообщение, определяя его как не спам (пропуск спама или недостаточная полнота метода). В то же время как спам-фильтр способен распознать большой процент нежелательных сообщений, может стать более важной задача минимизация числа ошибочных отсечений нужных сообщений, т.е. минимизация ошибки второго рода.

Ошибка второго рода, или ложноположительная ошибка, будет допущена, если спам-фильтр ошибочно классифицирует легитимное сообщение как спам (ложные срабатывания или точность метода). Спам-фильтр будет эффективным при низком уровне таких ошибок, то есть при минимальной ошибке второго рода. Однако в настоящее время все антиспамовые системы имеют корреляцию между ошибками первого и второго родов.

В классификаторах обычно допускают компромисс между приемлемым уровнем ошибок первого и второго рода и при принятии решения используют пороговые значения, которые могут варьироваться. От этого зависит, насколько классификатор более «строгий» или более «мягкий». В качестве порогового значения выбирают уровень значимости, который задают при проверке статистических гипотез. При этом повышение чувствительности фильтра приводит к увеличению ошибки первого рода (пропуск спама), а понижение чувствительности – к увеличению ошибки второго рода (ложные срабатывания).

Критерий оптимальности Байеса

Для оценки качества классификации, необходимо учитывать потери, связанные с ошибками первого и второго родов. Для этого пространство признака X разделим точкой x_0 на два полупространства X_A и X_B. Пусть c_1 – условная цена ошибки первого рода, c_2 – условная цена ошибки второго рода, $P(A)$ – априорная вероятность класса A, $P(B)$ – априорная вероятность класса B, $P(A) + P(B) = 1$. Величины c_1 и c_2 зависят от коэффициентов матрицы цен $C_{2x2} = \{c_{ij}\}$ и от ошибок первого и второго родов:

$$c_1 = c_{12}\,\alpha + c_{11}(1-\alpha); \quad (1)$$

$$c_2 = c_{21}\beta + c_{22}(1-\beta). \quad (2)$$

Эти величины также называют условными рисками при справедливости гипотез H_A и H_B соответственно.

В соответствии с теорией принятия решений введем решающее правило классификации, минимизирующее функцию потерь (риска) [3]:

$$R = c_1 P(\mathrm{A}) + c_2 P(B), \quad (3)$$

где c_1 и c_2 определяются выражениями (1) и (2).

Функция (3) представляет собой средний риск, причем зависящий от порогового значения x_0, так как величины c_1 и c_2 через ошибки первого и второго родов зависят от значения x_0, следовательно, ошибки первого и второго родов коррелированы между собой.

Минимальное значение R_{min} функции риска (3), достигаемое в точке x_0, называется байесовским риском.

$$\frac{f_1(X)}{f_2(X)} = \frac{c_{21} - c_{22}}{c_{12} - c_{11}} \cdot \frac{P(B)}{P(A)}, \quad (4)$$

где $f_1(X)$ и $f_2(X)$ плотности вероятностей распределения признака X по классам A и B соответственно.

Правая часть в выражении (4) называется отношением правдоподобия, которое является постоянным для данного выбора c_{ij}. При этом, если имеет место неравенство $\frac{f_1(X)}{f_2(X)} > \frac{c_{21} - c_{22}}{c_{12} - c_{11}} \cdot \frac{P(B)}{P(A)}$, то наблюдаемый вектор X относится к классу A, в противном случае вектор X относится к классу B. Если же выполняется равенство $\frac{f_1(X)}{f_2(X)} = \frac{c_{21} - c_{22}}{c_{12} - c_{11}} \cdot \frac{P(B)}{P(A)}$, то наблюдаемый вектор X относится к одному из классов A или B. Последнее выражение представляет собой уравнение границы классов A и B. Сформулированное решающее правило принятия решений относится к так называемым правилам Байеса [3-4].

Такую методику можно применять ко многим практическим задачам, формулируя их в терминах статистической теории принятия решений, но эта теория ограничена допущением, что плотности вероятностей $f_1(X)$ и $f_2(X)$ известны. Но чаще всего на практике функции $f_1(X)$ и $f_2(X)$ неизвестны, и необходимо определить их оценки $\tilde{f}_1(X), \tilde{f}_2(X)$ по обучающим выборкам аппроксимативными методами [3-4]. Последнее сильно будет тормозить работу классификатора. Учитывая данный факт, в работе выбран следующий подход, заключающийся в получении оценок $\tilde{f}_1(X), \tilde{f}_2(X)$ только на малых обучающих выборках объема 100-200 на стадии обучения фильтра, а сам критерий оптимальности с получением таких оценок в работу программы не включен.

Результаты х экспериментов на обучающих выборках, позволили определить оптимальные пороговые значения для принятия решений: $x_в$ = 0,95 для верхнего порога и $x_н$ = 0,4 для нижнего порога. Тем самым были установлены жесткие рамки по спаму и обычные для не спам-сообщений. Такие пороговые значения обеспечивают минимум пропуска в спам нужных сообщений, то есть минимум ложных срабатываний. Следует заметить, что каждый администратор сети может иметь свои предпочтения и легко может задать угодные ему значения пороговых значений. Для оценки алгоритмов фильтрации сообщений используют комбинацию двух критериев оценки качества:

Доля пропущенного спама

Этот параметр обозначает количество нормальных сообщений оцененных как спам к общему количеству сообщений со спамом в тестовом наборе. Этот критерий характеризует пропуск спама алгоритмом.

Таким образом, 15% пропусков (85% уровень фильтрации спама) означают, например, что всего пришло 1000 спамовых писем, из которых 150 не было распознано как спам.

Доля ложных срабатываний

Это отношение количества нормальных сообщений, ошибочно распознанных алгоритмом как спам, к общему количеству нормальных сообщений в тестовом наборе. Этот критерий характеризует, насколько редко алгоритм ошибается, помечая нормальные сообщения как спам.

Таким образом, 0,5% ложных срабатываний означают, например, что всего пришло 1000 легитимных сообщений, и из них 5 было ошибочно признано спамом. Критерий на основе доли «пропущенного спама» является наиболее понятным интуитивно – если один классификатор распознает 70%, а второй – 80% спама, то второй спам-фильтр можно считать лучшим. Однако, с другой стороны, необходимо знать, что повышение уровня распознавания влечет за собой рост количества ложных срабатываний. Это означает, что ошибки первого и второго родов при принятии решений в задаче классификации коррелированны между собой. Поэтому оба критерия нужно рассматривать совместно, причем оценка количества ложных срабатываний должна иметь приоритет при составлении суммарной оценки фильтра. Полностью избавиться от данных ошибок невозможно, но необходимо максимально уменьшить их число.

Особенности тестирования и обучения спам-фильтра

Так как мы ведем речь о байесовских и им подобных классификаторах, то есть об обучаемых пользователем системах, то при их тестировании необходимо использовать тот режим обучения, который будет применяться при реальной эксплуатации. Таким образом, необходимо проводить регулярное обучение системы по пропущенному спаму и ложным срабатываниям, сделав это тем самым штатным режимом эксплуатации спам-фильтра в конкретной организации.

Достоверные результаты тестирования можно получить при выполнении следующих необходимых условий:

- тестирование должно проводиться с установкой спам-фильтра на тот же поток сообщений на форуме, где его предполагается в дальнейшем использовать;
- продолжительность тестирования – несколько недель;
- объем тестируемого набора документов - от нескольких сотен до нескольких тысяч сообщений в день;
- анализ результатов с использованием корректного определения категорий спама/не спама, ложных срабатываний и пропуска спама.

Из сказанного следует, что спам-фильтр, обученный в одной организации, не обязательно будет хорошо работать в другой организации. Это является недостатком всех обучаемых фильтров. Его можно устранить быстрым обучением фильтра, организовав «коллективное обучение» совместно на веб-сервере верхнего уровня управления спам-фильтром и серверах уровня сайтов, на основе обмена данными по модификации базы данных веб-сервера.

Для приукрашивания реальной картины при рекламировании спам-фильтров довольно часто используют результаты тестирования при условиях, когда один фильтр используется на наборе спам-сообщений, полученном при работе другого фильтра. При этом если у второго фильтра случится пропуск спама, обнаруженного первым фильтром, то второй считается хуже, чем первый. Однако на самом деле данное утверждение несправедливо, так как такое может случиться и с первым фильтром при тех же условиях тестирования. Поэтому необходимо проводить анализ множества результатов работы отдельных фильтров и подмножества их пересечений

Нами предлагается подход к организации классификатора, который заключается в совместном использовании методов Байеса и Фишера для повышения качества фильтрации на основе анализа подмножеств пересечения множеств, распознанных обоими методами (спам – не спам, ложные срабатывания и пропуск спама).

Пусть $\mathbf{S} = \{s_i\}$ ($i = 1 \ldots M$) – множество документов (сообщений), включающее как легитимные, так и спам-сообщения; $\mathbf{S}_B \subset \mathbf{S}$ и $\mathbf{S}_F \subset \mathbf{S}$ – множества документов, распознаваемые соответственно классификаторами Байеса и Фишера. Тогда подмножество – пересечение $S_B \cap S_F$ по всем вышеуказанным категориям может быть использовано для оценки качества работы совмещенного фильтра (см. рис. 1).

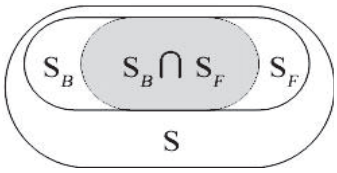

Рис. 1. Иллюстрация меры близости двух множеств S_B и S_F

Полнота такого пересечения $S_B \cap S_F$ также будет давать оценки для подмножеств $S_B \setminus S_F$ и $S_F \setminus S_B$. В качестве меры близости двух множеств S_B и S_F предложено использовать абсолютную меру $N(S_B \cap S_F)$ – число общих документов в этих множествах.

Таким образом, в работе в качестве оптимального критерия для оценки качества обучения спам-фильтра принимается максимальное значение меры по категориям l (спам – не спам, ложные срабатывания, пропуск спама):

$$N_l(\mathbf{S}_B^l \cap \mathbf{S}_F^l) \to \max.$$

После достижения наилучших показателей меры близости множеств S_B и S_F по всем категориям, администратор может сделать выбор, каким фильтром в дальнейшем ему пользоваться (см. рис. 2).

Заключение

Предложенная общая схема оценки качества обученности совмещенного фильтра включает следующие этапы (см. рис. 3):

- построение множеств спам-сообщений: легитимных, ложных срабатываний и пропуска спама на основе фильтра Байеса;
- построение множеств спам-сообщений: легитимных, ложных срабатываний и пропуска спама на основе фильтра Фишера;
- выполнение операции пересечения по спам-сообщениям, общим для обоих фильтров;
- выполнение операции пересечения по легитимным сообщениям, общим для обоих фильтров;
- выполнение операции пересечения по ложным срабатываниям, общим для обоих фильтров;
- выполнение операции пересечения по пропускам спама, общим для обоих фильтров.

В этом случае действительно будет возможно оценить все компоненты общей картины:

- спам-сообщения, которые улавливаются обоими фильтрами;
- спам-сообщения, которые улавливаются только фильтром Байеса или только фильтром Фишера;
- одновременные ложные срабатывания обоих фильтров;

- ложные срабатывания каждого фильтра по отдельности;
- одновременный пропуск спама обоими фильтрами;
- пропуск спама каждым фильтром по отдельности.

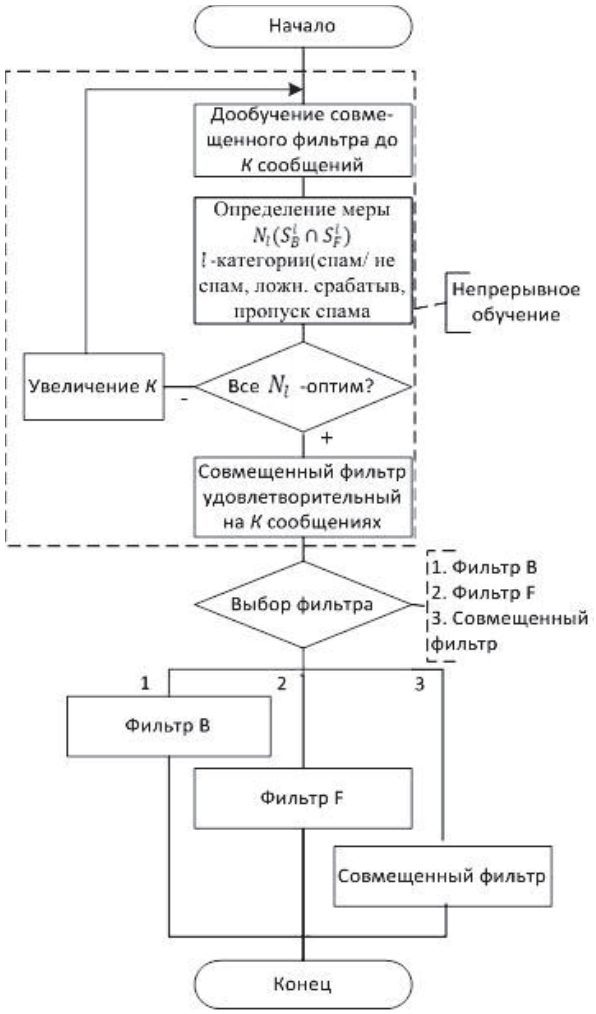

Рис. 2. Алгоритм оценки качества работы фильтров

Только имея такую полную картину, можно обоснованно сравнивать качество обученности совмещенного фильтра.

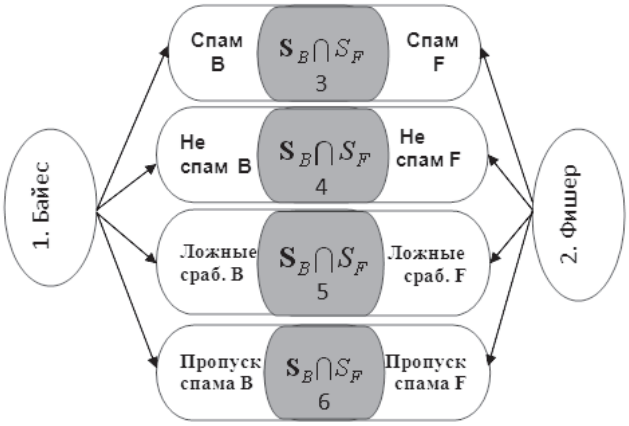

Рис. 3. Схема оценки качества работы совмещенного фильтра

Литература

1. Мезенцева Е.М., Тарасов В.Н. Защита компьютерных сетей. Веб программирование многомодульного спам-фильтра // Программная инженерия. № 4, 2012. – С. 27-32.
2. Мезенцева Е.М., Тарасов В.Н. Организация защиты компьютерных сетей. Метод многомодульной фильтрации спама на web-сайтах // Информационные технологии. №6, 2012. – С. 18-22.
3. Себестиан Г.С. Процессы принятия решений при распознавании образов. Киев: Техника, 1965. - 149 с.
4. Андерсон Т. Введение в многомерный статистический анализ. М.: Физматиз, 1963. – 500 с.

CREATION OF OPTIMUM SPAM FILTER ON THE BASIS OF COMBINATION OF STATISTICAL QUALIFIERS

Tarasov V.N., Mezentseva E.M.

Criteria of an optimality are presented in article for classification of messages on the basis of statistical methods taking into account errors of the first and second childbirth, a share of false operations and passed spam. Features of testing and training spam filter are given. Approach to the organization of the qualifier which consists in sharing of methods of Bayes and Fischer is offered. For improvement of quality of a filtration it is offered to use the me-chanism of the analysis of subsets of crossing of the sets distinguished by both methods on cat-egories (spam \not spam, false operations and spam admission).

Keywords: *message classification, statistical methods, errors of the first and second sort, share of passed spam, share of false operations, subsets of the crossings, the combined filter.*

Тарасов Вениамин Николаевич, д.т.н., профессор, заведующий Кафедрой программного обеспечения и управления в технических системах (ПОУТС) Поволжского государственного университета телекоммуникаций и информатики (ПГУТИ). Тел. (8-846) 228-00-13, 8-960-827-22-33. E-mail: vt@ist.psati.ru.

Мезенцева Екатерина Михайловна, старший преподаватель Кафедры ПОУТС ПГУТИ. Тел. (8-846) 228-00-13, 8-929-707-9-111. E-mail: katya-mem@psuti.ru