УДК 681.513

С.А. Жиляев

Динамический выбор размерности опорной функции в вероятностных алгоритмах МГУА на основе внешнего дополнения

Предложено определять размерность вероятностной опорной функции статистических алгоритмов МГУА с помощью внешнего дополнения. Приведен алгоритм реализации описанной методики. Указаны типы задач, для которых новый алгоритм наиболее эффективен. Получен сравнительный результат оценки точности алгоритма на примере решения актуальной практической задачи.

It is suggests to define the dimension of a probabilistic reference function in stochastic GDMN algorithms with the help of the external compliment. The algorithm of implementing the described technique is presented. The types of problems for which a new algorithm is the most efficient are indicated. A comparative accuracy estimation of the algorithm is obtained on the example of solving the actual practical problem.

Запропоновано визначати розмірність імовірнісної опорної функції статистичних алгоритмів МГУА за допомогою зовнішнього доповнення. Наведено алгоритм реалізації описаної методики. Вказано типи задач, для яких новий алгоритм є найбільш ефективним. Отримано порівняльний результат оцінки точності алгоритму на прикладі розв'язання актуальної практичної задачі.

Введение. Мировой экономический кризис вынудил бизнес быть рациональнее. Телевидение, радио, внешняя реклама, а также другие традиционные способы продвижения товаров и услуг на рынке уступают технологиям Интернет, которые зачастую более рентабельны. Если еще недавно потеря потенциального покупателя не была критичной, то сейчас идет жесткая борьба за каждого посетителя. Точная идентификация потребностей клиента – это ключевой фактор успешной продажи [1], что в настоящее время становится вопросом выживания для многих компаний.

Практическая задача прогнозирования предпочтений Интернет-клиентов на основе данных, характеризующих их персональные компьютеры, чрезвычайно актуальна. Огромные бюджеты лидеров электронной коммерции (*Google*, *Amazon*, *eBay* и т.д.), затрачиваемые на поиск оптимального решения, растут. Тем не менее, общепринятого подхода к проблеме не существует.

В публикации [2] рассмотрен алгоритм, решающий поставленную задачу. Несмотря на то что применение *GMDH MULTI* [3, 4] для оптимизации байесовского классификатора позволяет учитывать только «полезные» признаки из множества характеристик персональных компьютеров Интернет-клиентов при прогнозировании покупательских предпочтений, приведенный алгоритм использовал предположение об условной независимости признаков классификации, которое негативно отражалось на его точности. В статье предлагается новый подход динамического выбора размерности опорной функции статистических алгоритмов МГУА с помощью внешнего дополнения. Модифицированные соответствующим образом алгоритмы в зависимости от качества тренировочной выборки позволяют уменьшить негативное влияние гипотезы об условной независимости признаков либо вовсе его исключить.

Математическая постановка задачи

Пусть даны конечные множества K – продаваемые продукты и T – m-мерное множество векторов – признаков Интернет-клиента $(x_1, x_2, \ldots, x_m) \in T$, где $x_i \in \{0, 1\}$. Существует отображение $F: T \to K$, которое каждому клиенту–покупателю ставит в соответствие приобретенный им продукт. Стоимость товаров множества K одинакова. Множество признаков состоит из двух непересекающихся подмножеств $T = L \cup E$, $L \cap E \cup \varnothing$. Назовем L экспериментальным множеством, а E – экзаменационным. Требуется, зная купленный продукт каждого из клиентов экспериментального множества, определить приобретенные товары покупателей экзаменационного множества так, чтобы количество несовпадений с реальными покупками было минимально, т.е. известно $F_L = F\big|_L$ – сужение отображения F на множество L, нужно построить отображение $F_E: E \to K$ такое, что $\left|\{x \in E \,|\, F(x) \neq F_E(x)\}\right| \to \min$.

Пример. Пусть множество продуктов состоит из двух элементов $K = \{antivirus, systemdoctor\}$, размерность m пространства признаков T равна четырем.

Отображение $F: T \to K$ задается табл. 1.

Таблица 1

Подмножество T	Вектор признаков $x \in T = L \cup E$	Значения F на x
L	(0,0,0,1)	*systemdoctor*
	(0,0,1,0)	*systemdoctor*
	(1,0,1,0)	*antivirus*
	(1,1,0,0)	*antivirus*
	(0,1,1,1)	*systemdoctor*
	(1,1,1,1)	*antivirus*
	(1,0,1,0)	*antivirus*
E	(0,0,0,0)	*systemdoctor*
	(1,1,0,0)	*antivirus*
	(1,0,0,1)	*antivirus*

Известно, что $F_L = F\big|_L$, необходимо определить $F_E: E \to K$ – покупки клиентов с векторами признаков из экзаменационного множества E так, чтобы количество несовпадений с реальными покупками, определенными последними тремя строками табл. 1, было минимальным.

Выбор размерности опорной функции условной вероятности

В классическом многорядном статистическом алгоритме МГУА используется парная вероятностная опорная функция $f(x_i, x_j) = \arg\max_{k^* \in K} p(x_i, x_j \,|\, k^*)$, а в алгоритме *MULTI* размерность опорной функции и вовсе равна единице. Рассмотрим мотивы использования столь низкой размерности опорных функций.

Первой причиной является необходимость наличия достаточного количества информации для вычисления совместных апостериорных вероятностей. Действительно, для вычисления вероятности $p(x_i, x_j, x_l \,|\, k^*)$ требуется приблизительно во столько раз больше экспериментальных данных, сколько значений может принимать компонента x_l, т.е. имеет место показательная зависимость требуемого объема экспериментальной выборки от размерности статистической опорной функции. Но, ограничивая алгоритм использованием лишь парных вероятностей, исследователь вносит субъективность в процесс построения модели, что может привести к потере желаемого решения или неточности классификации. Например, крупные компании, занимающиеся электронной коммерцией, проводят десятки, а иногда и сотни транзакций в секунду. Понятно, что доступ к экспериментальным данным не является для них ограничением при классификации.

В статье предлагается определять размерность вероятностной опорной функции с помощью внешнего дополнения согласно основному принципу МГУА. Таким образом, знания о структуре статистических связей между признаками клиента и его предпочтениями получены из выборки данных, а не заложены априорно в модель исследователем. Для нахождения оптимальной размерности ее следует постепенно увеличивать, каждый раз проводя поиск модели с помощью многорядного или многоэтапного алгоритма. При этом вначале точность классификации будет возрастать, так как с каждым шагом прогноз полнее учитывает взаимное влияние признаков. Но с некоторой итерации объема экспериментальной выборки может не хватать, чтобы корректно вычислить совместные апостериорные вероятности, поэтому точность моделирования начнет падать. Размерность опорной функции, при которой достигается минимум ошибки классификации, – оптимальна. Тот факт, что в поставленной задаче Интернет-покупатели характеризуются бинарными признаками, позволил увеличить размерность опорной функции даже при сравнительно небольшом количестве данных, имеющихся для прогнозирования.

Вторым аргументом в пользу использования парной вероятностной функции есть то, что количество моделей, генерируемых многорядным алгоритмом, является показательной зависимостью m^v от числа аргументов опорной функции – v с основанием, равным количеству возможных значений компонент вектора признаков – m. Снизить объем необходимых вычислений можно, проведя начальный поиск

многорядным алгоритмом при $v = 2$ и в дальнейшем рассматривая только признаки, вошедшие в лучшие найденные модели.

Обобщенный статистический алгоритм МГУА *MULTI*

Ниже приведен новый статистический алгоритм, построенный на базе алгоритма МГУА *MULTI*, который динамически определяет размерность опорной функции, уменьшая негативное влияние гипотезы об условной независимости характеристик клиентов, что в свою очередь повышает точность классификации.

На первом шаге рассматриваются модели вида

$$y = \arg\max_{k' \in K} p(x_i \mid k'), i \in \overline{1 \ldots m}.$$

Обозначим аргументы $x_{i_{h_1}}^{(1)}$, $i_{h_1} \in \overline{1 \ldots l}$ лучших l моделей с наименьшим значением внешнего критерия. На втором шаге вместо использования предположения об условной независимости признаков строим модели на основе совместной вероятности

$$y = \arg\max_{k' \in K} p(x_{i_{h_1}}^{(1)}, x_i \mid k'),\ i_{h_1} \in \overline{1 \ldots l},\ i \in \overline{1 \ldots m}.$$

На следующем шаге выбираем l пар аргументов $(x_{i_{h_1}}^{(2)}, x_{i_{h_2}}^{(2)})$, по которым построены модели с наименьшим внешним критерием, и строим зависимости вида

$$y = \arg\max_{k' \in K} p(x_{i_{h_1}}^{(2)}, x_{i_{h_2}}^{(2)}, x_i \mid k'),$$
$$i_{h_1} \in \overline{1 \ldots l},\ i_{h_2} \in \overline{1 \ldots l},\ i \in \overline{1 \ldots m}.$$

Продолжаем итерации до тех пор, пока внешний критерий не начнет падать вследствие недостатка экспериментальных данных для определения совместных вероятностей. Пусть такая ситуация сложилась на $q_1 + 1$-м шаге. Тогда на q_1-м шаге имеем l отобранных моделей

$$y = \arg\max_{k' \in K} p\left(x_{i_{h_1}}^{(q_1)},\ x_{i_{h_2}}^{(q_1)}, \ldots \right.$$
$$\left. \ldots, x_{i_{h_{q_1}}}^{(q_1)} \mid k'\right),\ h_p \in \overline{1 \ldots l},\ \ p \in \overline{1 \ldots q_1}$$

или в более удобной записи

$$y = \arg\max_{k' \in K} p\Big(\bigwedge_{p \in \overline{1 \ldots q_1}} x_{i_{h_p}}^{(q_1)} \mid k'\Big), h_p \in \overline{1 \ldots l},\ p \in \overline{1 \ldots q_1}.$$

Таким образом, построена первая группа признаков с совместной вероятностью, характеризующая предпочтения клиента. Далее добавляем вторую группу, рассматривая модели вида

$$y = \arg\max_{k' \in K} \left(p\left(\bigwedge_{p \in \overline{1 \ldots q_1}} x_{i_{h_p}}^{(q_1)} \mid k'\right) \cdot p\left(x_i \mid k'\right)\right),$$
$$h_p \in \overline{1 \ldots l},\ i \in \overline{1 \ldots m}.$$

Снова выбираем l наилучших моделей и конструируем с помощью их аргументов новый ряд

$$y = \arg\max_{k' \in K} \left(p\left(\bigwedge_{p \in \overline{1 \ldots q_1}} x_{i_{h_p}}^{(q_1+1)} \mid k'\right) \cdot p\left(x_{h_1}^{(q_1+1)}, x_i \mid k'\right)\right),$$
$$h_p \in \overline{1 \ldots l},\ i \in \overline{1 \ldots m}.$$

Продолжаем данную процедуру до тех пор, пока внешний критерий не начнет падать вследствие невозможности корректно определить совместные вероятности второй группы характеристик. Пусть это произошло на $q_2 + 1$-м шаге. Тогда на q_2-м шаге получим l отобранных моделей

$$y = \arg\max_{k' \in K} \left(p\left(\bigwedge_{p \in \overline{1 \ldots q_2}} x_{i_{h_p}}^{(q_2)} \mid k'\right) \cdot p\left(\bigwedge_{g \in \overline{q_1+1 \ldots q_2}} x_{i_{h_g}}^{(q_2)} \mid k'\right)\right),$$

$$h_p \in \overline{1 \ldots l},\ h_g \in \overline{1 \ldots l},\ p \in \overline{1 \ldots q_1},\ g \in \overline{q_1 + 1 \ldots q_2}.$$

Аналогично алгоритм продолжает работу до того момента, пока добавление новой группы признаков с совместной вероятностью уже не приводит к улучшению внешнего критерия. Конечная модель будет иметь вид

$$y = \arg\max_{k' \in K} \prod_{j=0}^{n} p\left(\bigwedge_{p_j \in \overline{q_j+1 \ldots q_{j+1}}} x_{i_{h_{p_j}}}^{(q_n)} \mid k'\right),$$

где $q_0 = 0$, n – количество групп признаков с совместной вероятностью.

Отметим, что в случае использования малой выборки экспериментальных данных, не пригодной для корректного вычисления совместной вероятности даже для пар атрибутов классификации, решение будет совпадать с байесовским классификатором, оптимизированным с помощью алгоритма *MULTI* [2]

$$y = \arg\max_{k' \in K} \prod_{j=0}^{r} p\left(x_j \mid k'\right),\ \ j \in \overline{1 \ldots r}\ .$$

В случае же достаточно большой выборки экспериментальных данных, полученное ре-

шение может совершенно не использовать предположение об условной независимости и выглядеть так:

$$y = \arg\max_{k' \in K} p\left(\bigwedge_{j \in \overline{1 \ldots r}} x_j \mid k' \right),\ r \in \overline{1 \ldots m}.$$

Разработанный алгоритм показывает наилучшие результаты в случае сравнительно большой выборки экспериментальных данных с атрибутами, имеющими немного возможных значений. Это обуславливается тем, что если обозначить количество возможных значений признаков – n, то число векторов в экспериментальной выборке, необходимое для корректного вычисления совместной вероятности размерности m, равно

$$\text{const} \cdot n^m.$$

Результаты

При решении практической задачи, поставленной ранее, использовались следующие алгоритмы: байесовский классификатор [5], МГУА *MULTI* с линейной опорной функцией [3, 4], итерационный многорядный алгоритм МГУА с линейной, ковариационной [6–8] и байесовской опорной функцией [9], байесовский классификатор с оптимизированной структурой [2], а также впервые применялся обобщенный статистический алгоритм МГУА *MULTI*, динамически определяющий размерность опорной функции с помощью внешнего дополнения. Все перечисленные алгоритмы тестировались на трех выборках покупателей с различным количеством продаваемых продуктов. Выборка I содержала Интернет-клиентов, купивших один из двух товаров {*antivirus*, *systemdoctor*}, в выборке II присутствовали покупатели трех товаров {*antivirus*, *systemdoctor*, *drivecleaner*}, а в III – четырех {*antivirus*, *systemdoctor*, *drivecleaner*, *errorsafe*}. В экспериментальной части выборок для каждого продукта находилось по 400 купивших его клиентов, а в экзаменационной – по 100. В каждом из тестов потребители характеризовались 1281 признаком. В табл. 2 приведены сравнения точности распознавания предпочтений покупателей в каждом из экспериментов.

Таблица 2

Выборка	Количество продуктов	Алгоритм	Точность, %
I	2	байесовский классификатор	72
		МГУА *MULTI* с линейной опорной функцией	86
		итерационный МГУА с линейной опорной функцией	83
		итерационный МГУА с ковариационной опорной функцией	83
		итерационный МГУА с байесовской опорной функцией	84
		байесовский классификатор, оптимизированный по МГУА *MULTI*	87
		обобщенный статистический алгоритм МГУА *MULTI*	91
II	3	байесовский классификатор	66
		МГУА *MULTI* с линейной опорной функцией	67
		итерационный МГУА с линейной опорной функцией	65
		итерационный МГУА с ковариационной опорной функцией	68
		итерационный МГУА с байесовской опорной функцией	74
		байесовский классификатор, оптимизированный по МГУА *MULTI*	76
		обобщенный статистический алгоритм МГУА *MULTI*	78
III	4	байесовский классификатор	45
		МГУА *MULTI* с линейной опорной функцией	44
		итерационный МГУА с линейной опорной функцией	45
		итерационный МГУА с ковариационной опорной функцией	45
		итерационный МГУА с байесовской опорной функцией	52
		байесовский классификатор, оптимизированный по МГУА *MULTI*	56
		обобщенный статистический алгоритм МГУА *MULTI*	60

Заключение. Итак, рассмотрена новая методика динамического определения размерности вероятностной опорной функции с помощью внешнего дополнения. Базируясь на основном принципе МГУА, она позволяет исследователю получать знания о структуре статистических связей между признаками объекта классификации и выходной величиной из выборки данных, а не закладывать их априорно, внося субъективность в процесс моделирования.

Основываясь на предложенной методике, разработан новый алгоритм, который является обобщением статистического алгоритма МГУА *MULTI*. В зависимости от качества выборки экспериментальных данных, алгоритм позволяет исключить или значительно снизить влияние гипотезы об условной независимости признаков классификации, что ведет к повышению точности моделирования. Разработанный алгоритм показывает наилучшие результаты в случае сравнительно большой выборки экспериментальных данных с компонентами векторов прогнозирования, имеющими немного возможных значений. В частности алгоритм эффективен для задач прогнозирования предпочтений Интернет-клиентов на основе данных, характеризующих их персональные компьютеры. Так как достаточное количество экспериментальных данных гарантируется высоким показателем числа электронных транзакций, проводимых работающими в данном сегменте компаниями.

Результаты практических экспериментов подтвердили преимущества разработанного алгоритма и целесообразность его использования, так как он позволил добиться увеличения точности прогнозирования от 15 до 19 процентов.

1. *Котлер Ф.* Основы маркетинга. – М.: Прогресс, 1990. – 736 с.
2. *Жиляев С.А.*, *Рябоконь Д.И.* Практическое применение байесовского правила принятия решений в качестве опорной функции алгоритма МГУА // УСиМ. – 2007. – № 6. – С. 73–79.
3. *Степашко В.С.* Конечная селекционная процедура сокращения полного перебора моделей // Автоматика. – 1983. – № 4. – С. 84–88.
4. *Степашко В.С.*, *Костенко Ю.В.* Исследование свойств комбинаторно-селекционного (многоэтапного) алгоритма МГУА // Моделирование и управление состоянием эколого-экономических систем региона. – К.: МНУЦИТ и С, 2000. – С. 84–100.
5. *Шлезингер М.И.*, *Главач В.* Десять лекций по статическому и структурному распознаванию. – К.: Наук. думка, 2004. – 535 с.
6. *Ивахненко А.Г.* Индуктивный метод самоорганизации моделей сложных систем. – К.: Наук. думка, 1982. – 296 с.
7. *Ивахненко А.Г.*, *Степашко В.С.* Помехоустойчивость моделирования. – К.: Наук. думка, 1985. – 216 с.
8. *Ивахненко А.Г. Мюллер Й.А.* Самоорганизация прогнозирующих моделей. – К.: Техніка, 1985. –223 с.
9. *Ивахненко А.Г.*, *Зайченко Ю.П.*, *Димитров В.Д.* Принятие решений на основе самоорганизации. – М.: Сов. радио, 1976. – 280 с.

Поступила 02.08.2010

Тел. для справок: (050) 941-2349 (Киев)
E-mail: *us5eme@gmail.com*

●