НЕКОТОРЫЕ ПОДХОДЫ К ПРОБЛЕМЕ ОШИБКИ МОДЕЛИ В СИСТЕМАХ DATA MINING

В.М. Резников

Общепризнанны принципиальное отставание и несоответствие возможностей теоретической науки потребностям практики. Это в полной мере относится к применению математических дисциплин. Анализ постановок задач, решаемых в математической статистике, показывает, что корректное применение статистических методов предполагает, что известны распределение данных и (или) независимость случайных величин, описывающих данные. Свойства независимости и закона распределения названы нами базовыми свойствами [1]. Сложность проверки соответствия данных предположениям теории, в том числе определения базовых свойств в данных, очень часто превышает трудности решения задачи с помощью статистических методов.

Предположим, что исследуемая задача решена. Необходимо оценить качество найденного решения. Одним из популярных критериев качества решений в математической статистике является свойство состоятельности. Статистическая оценка называется состоятельной, если она сходится по вероятности к искомому значению оцениваемого параметра. Статистическая теория не определяет минимальное количество данных, при котором применение математической статистики является обоснованным.

Эти и другие многочисленные несовершенства привели к тому, что стала широко критиковаться адекватность математической статистики практике научных исследований [2]. Отмечаются оторванность тематики математической статистики от запросов прикладной статистики, неадекватность постановок, неконструктивный характер теорем закона больших чисел, априористический характер математической статистики. В отношении последнего речь идет о том, что статистические закономерности существуют априори, а роль данных сводится к уточнению теоретических закономерностей.

Потребность в создании работающей статистической дисциплины реализовалась в появлении прикладного статистического анализа. Эта наука опирается на классические стохастические науки, если они могут быть применены, а также на ряд новых направлений статистической науки. Точками роста являются непараметрический статистический анализ, метрологическая концепция, компьютерное моделирование, бутстрэп, интервальный спастический анализ и некоторые другие направления [3]. По-видимому, для современного этапа развития науки характерны создание прикладных математических дисциплин и выделение их в самостоятельные науки.

В качестве примера другой научной дисциплины, частично опирающейся на строгую математику, а частично использующей эвристические методы и компьютерные технологии, приведем дисциплину, называемую «Data Mining». В отечественной научной литературе используется именно название «Data Mining», нам неизвестен русский эквивалент этого термина. Прикладной статистический анализ и Data Mining являются частично пересекающимися областями знания. Происхождение обеих дисциплин связано с математической статистикой. Кроме математической статистики Data Mining идейно связано с областью знания, относящейся к обучению компьютерных программ.

Системы Data Mining предназначены для получения нового знания как с помощью современных компьютерных методов, так и посредством традиционных формальных методов [4]. Во многих областях накоплены огромные массивы данных и предполагается практически неограниченное приращение данных, поэтому методология Data Mining предлагает эффективные принципы содержания, поиска и защиты информации. Сырые данные являются потенциальными носителями информации, для анализа данных в Data Mining используют модели обработки данных.

Особенность Data Mining – высокая технологическая обеспеченность, но при этом существенное отставание в развитии методологии исследований. Мы полагаем, что слабость методологии Data Mining связана с тем, что практически полностью игнорируется факт неуниверсальности формальных и компьютерных методов.

С одной стороны, общепризнано, что не существует универсальных моделей анализа данных. С другой стороны, границы применимости методов и особенности данных, для которых эти методы адекватны, как правило, не представлены. В практически единственной известной нам работе идея неуниверсальности методов в отношении разно-

образных приложений учтена следующим образом. Любой метод Data Mining явно или неявно предполагает: «1) тип данных; 2) использование определенного формального языка, позволяющего интерпретировать данные и выполнять формальные манипуляции; 3) класс гипотез, которые могут быть исследованы с помощью данного формального языка и специфических данных[5]».

Мы полагаем, что многие проблемы неуниверсальности приложений снимаются решением двух взаимосвязанных проблем: проблемы объекта теории и проблемы ошибки модели. Проблема объекта заключается в учете характерных особенностей данных и выборе модели, наиболее адекватной для такого рода данных. Проблема ошибки модели заключается в формальном описании и учете влияния фоновых факторов с целью элиминации или уменьшения их влияния на отклонение реальных данных от гипотетических модельных данных.

Многие классические модели статистики, не удовлетворяют приведенным выше требованиям (1)–(3) и не являются робастными. Небольшое отклонение данных от исходных предположений влечет за собой получение грубых оценок. Отсутствие робастности ясно проявляется при обработке больших массивов данных. Поскольку характерная особенность Data Mining – это огромные полигоны данных, постольку видятся актуальными разработка и применение моделей, учитывающих указанные требования.

В доступной нам литературе проблема объекта представлена крайне скудно. Впервые на методологическом уровне эта проблема была поставлена Р. фон Мизесом [6]. Теория Мизеса предполагает применение моделей для данных, соответствующих идеальному объекту теории. Мизес сформулировал проблему объекта, но он не предложил конструктивного подхода к решению этой задачи.

Проблема объекта была сформулирована Мизесом для фундаментальных эмпирических вероятностных теорий. Объекты разработанной Мизесом теории – это бесконечные последовательности частот, закон образования которых неизвестен, и эти последовательности удовлетворяют двум требованиям, являющимся эмпирическими законами. Согласно первому требованию последовательность частот имеет предел. Согласно второму требованию, которое называют требованием невозможности системы игры, счетные подпоследовательности, выделяемые из исходной последовательности, должны иметь предел, совпадающий с пределом для всей совокупности данных.

В связи с требованиями Мизеса возникает следующий вопрос: относятся ли эти требования к абстрактным объектам или они предназначены для реальных данных? Существует три ответа на этот вопрос: 1) требования относятся к абстрактным данным; 2) требования относятся одновременно к абстрактным и реальным данным; 3) требования относятся к реальным данным.

В другой работе нами исследован подход Мизеса к решению практической задачи [7]. Решение Мизеса основано на сравнении теоретической и эмпирической дисперсий. Это свидетельствует в пользу первого варианта решения, потому что согласно требованиям Мизеса теоретический закон образования последовательности частот неизвестен, поэтому второе и третье предположения о характере требований Мизеса не являются адекватными. Таким образом, Мизес не предложил решения для сформулированной им проблемы.

Для того чтобы показать сложность или даже невозможность решения проблемы объекта, мы ввели понятие базового свойства математической теории.

Определение. Базовые свойства объектов теории удовлетворяют двум требованиям. Во-первых, наличие этих свойств логически невыводимо на основе информации о других (в том числе и базовых) свойствах объектов теории. Во-вторых, на основе базовых свойств доказаны фундаментальные теоремы. В теории вероятностей это теорема закона больших чисел.

Теорема. Пусть μ – число наступлений события A в n независимых испытаниях и p есть вероятность наступления A в каждом из испытаний. Тогда каково бы ни было число $\varepsilon > 0$, имеет место следующий результат:

$$\lim_{n\to\infty} P\{\mu / n - p(A)| < \varepsilon\} = 1. \qquad (1)$$

Аргументы в пользу эпистемологической значимости этой теоремы следующие.

1. В работе А.Н.Колмогорова и Б.В.Гнеденко [8] отмечается, что познавательная ценность теории вероятностей связана только с предельными теоремами. Основной аргумент в пользу этого утверждения связан с тем обстоятельством, что решение неасимптотических задач не может являться объектом достаточно общей теории.

2. В работе Б.В.Гнеденко [9] приведена такая аргументация: «В практической деятельности, да и в общетеоретических задачах, большое значение имеют события с вероятностями близкими к единице или нулю».

3. Утверждается, что сама по себе близость величин $p(A)$ и m/n не доказывает устойчивости, так как вероятность этой близости может быть мала.

Аргументы против эпистемологической значимости этой теоремы состоят в следующем.

1. В теореме закона больших чисел теоретическая величина «вероятность успеха» задана априори, и тем самым теорема не может быть использована для определения этой уже известной вероятности [10].

2. Заключение теоремы говорит о вероятности события, представляющей собой разность вероятности события A и частоты события A. Под событием в прикладных науках понимается результат реального опыта. Поскольку вероятность события не может быть результатом реального испытания, постольку заключение теоремы не имеет прямой эмпирической интерпретации.

При эмпирической интерпретации вероятность события определена тогда и только тогда, когда соответствующая частота является устойчивой. При эмпирическом подходе внешняя вероятность в выражении (1) должна быть заменена частотой, и тогда заключение примет следующий вид:

$$\lim_{n\to\infty} \omega_n \{\mu / n(A) - p(A)| < \varepsilon\} = 1. \qquad (2)$$

Здесь символ ω обозначает частоту, а символ n – большое число. Пусть вероятность события A близка к нулю. Тогда устойчивость частоты события A определяется с помощью частоты события, являющегося частотой события A.

Продолжая аналогию, для того чтобы убедиться в устойчивости частоты от частоты события A, необходимо использовать более сложную конструкцию, а именно, требуется частота от частоты события B. При этом последнее событие, в свою очередь, является частотой события A. Этот процесс не имеет окончания.

3. Предположим, что производятся длительные статистические эксперименты. В многократно проведенных экспериментах была показана

близость величин $p(A)$ и m/n. Поэтому с прагматических позиций аргумент о случайной близости величин несостоятелен.

Таким образом, теорема не предоставляет ни концептуальные, ни операциональные средства для решения проблемы объекта эмпирической теории.

Специально проблема объекта в теориях Фишера и Неймана – Пирсона не выделена. Тем не менее некоторые разделы статистической теории, такие как критерии согласия, оценка параметров распределений, и некоторые принципы, например принцип рандомизации, в определенной степени обеспечивают решение проблемы объекта. Возникает вопрос: позволяют ли эти средства дать надежное удовлетворительное решение данной проблемы? Известно множество аргументов против того, что существующие методы и принципы являются адекватными решению проблемы объекта. К таким аргументам относятся, в частности, следующие.

1. В основе проверки гипотез лежит принцип Коорнота [11]. Согласно этому принципу маловероятные события невозможны. Данный принцип имеет только прагматическую, а не логическую значимость.

2. Проверка гипотез предполагает использование событий с малыми вероятностями. Эмпирическая проверка правильности событий порядка 10**(–N) является трудоемким процессом, она требует осуществления 10**(N+1) экспериментов.

3. Принцип рандомизации не обеспечивает решение проблемы объекта[12]

Эти и многие другие аргументы показывают, что теоретические методы статистики не обеспечивают решение проблемы объекта.

Очевидно, что проблемы объекта и ошибки модели должны решаться одновременно. В первом приближении предлагаются следующие подходы к классификации моделей ошибки.

Модель ошибки задана независимыми случайными величинами, имеющими нормальное распределение, с математическим ожиданием, равным нулю. Данная модель ошибки лежит в основе метода наименьших квадратов. Если модель ошибки задана правильно, то метод обеспечивает получение статистических оценок, являющихся несмещенными, состоятельными, эффективными и т.д.

Классические методы статистики обладают рядом недостатков. Критерий состоятельности предполагает, что чем больше данных, тем точнее статистические оценки. На самом деле в областях знания, где ценится точность результатов, например в метрологии, известно, что точность качества оценивания измерений может только уменьшиться при превышении определенного числа измерений. Кроме того, при превышении определенного числа данных минимальные отклонения ошибки модели от фактического положения дел ведут к большим погрешностям.

Для формальной экспликации недостатков метода наименьших квадратов рассмотрим следующий пример [13]. Пусть осуществляется измерение искомой величины q, результаты n измерений $q_1, q_2, \ldots, q_n$. Ошибки измерений описываются случайными величинами ξ_i. Предполагается, что они являются независимыми, одинаково распределенными по нормальному закону с нулевым математическим ожиданием. Пусть наблюдения равноточные с дисперсией σ_I, $i = 1, \ldots, n$. Тогда дисперсия оценки

$$\sigma(q^{\wedge}) = \sigma / \sqrt{n} . \qquad (3)$$

Пусть связь произвольных результатов измерений описывается коэффициентом корреляции k_{ij}. Будем для простоты считать, что $k_{ij} = k$ для любых $i \neq j$. Тогда легко получить, что

$$\sigma(q^{\wedge}) = \sigma\sqrt{(1-k)/ n + k} . \qquad (4)$$

В случае $n = 1000$, $k = 0{,}01$ $\sigma(q^{\wedge})$ оказывается в 11 раз больше по сравнению со случаем отсутствия корреляции.

Во избежание получения сверхоптимистичных статистических оценок, недостижимых при нарушениях модельных предположений, был разработан так называемый гарантирующий подход [14]. Гарантирующий подход предполагает, что получение точных значений дисперсии ошибки и точных значений коэффициентов корреляций, особенно в случае малых значений этих коэффициентов, является нереалистичной задачей. Реалистичные предположения формулируются следующим образом: $| k_{ij} | \leq k_{\max}$, $\sigma_i \leq \sigma_{\max}$. Выполнение данных предположений соответствует наихудшему распределению ошибок. Решение, полученное в этом случае, называется гарантированным, так как гарантируется, что при произвольном распределении ошибок результат будет не хуже.

В работе П.Е.Эльясберга [15] обосновано, что формула (4) естественно возникает при реалистичных предположениях, например при учете коэффициента корреляции, при учете неодинаковой точности измерений и в ряде других случаев. Если коэффициент корреляции не равен нулю и единице, то при бесконечном увеличении объема данных дисперсия ошибки стремится к $\sigma(q^{\wedge}) = \sigma\sqrt{k}$. В этом случае существует предельный объем данных, превышение которого ухудшает качество оценивания.

При проведении исследований в области гуманитарных наук, например в социологии, данные являются, как правило, вариабельными. В то же время неизвестно распределение результатов и нет никакой информации об ошибке наблюдений. В этом случае не представляется возможным воспользоваться моделями ошибок ни в классическом варианте статистики, ни в гарантирующем варианте.

Метрологическая концепция статистики предназначена для определения статистических характеристик данных на основе синтеза идей метрологии и теории обучения, используемой в Data Mining. Эта концепция была создана Ю.И.Алимовым [16]. Концепция демонстрирует доминирование эмпирических величин над теоретическими. Реальный статус имеют данные исследований. Теоретические величины априори не существуют. Теоретическое среднее существует, если эмпирические средние величины являются устойчивыми.

Основная задача прикладной статистики – определение степени устойчивости статистических характеристик. Важность этой задачи связана с тем, что ее решение обеспечивает предпосылки для точного прогноза. Определение устойчивости статистических характеристик – важный случай проблемы объекта. Ю.И.Алимов критикует понятие доверительного интервала, так как оно не имеет частотной интерпретации. В качестве альтернативы вводится понятие невероятностного доверительного интервала, адекватное для частотных интерпретаций.

Методология и алгоритмизация невероятностного доверительного интервала для оценивания математического ожидания показаны на примере оценивания среднего. Предположим, что в результате исследования объекта наблюдения получены следующие результаты: $x_1, x_2, \ldots, x_n$. Существует два способа эмпирического определения статистических характеристик с одновременным измерением точности измерений.

Первый способ имеет название «метод многих серий». Данные разбиваются на k равных интервалов, в каждом из которых по m чисел. В каждом интервале определяется среднее арифметическое: $s_1, s_2, \ldots, s_k$. Далее

определяются минимальное и максимальное среднее: $s_{\min}$ и $s_{\max}$. Модуль разности для $s_{\min}$ и $s_{\max}$ определяет ошибку измерений. Если эта величина меньше или равна предполагаемой точности, то в качестве среднего может быть взято любое значение из интервала $s_{\min}$, $s_{\max}$.

Для того чтобы проверить устойчивость оценивания, используют другие, ранее не использованные данные. Новое число интервалов $K >> k$, новое число данных в группе – $M >> m$. Если новые минимальные и максимальные средние попадают в старый интервал, то тем самым подтверждается гипотеза об устойчивости средних. В противном случае задаются новые концы интервалов.

Второй способ имеет название «метод удлиняющейся серии». Этот метод применяется, если данных недостаточно для объединения их в большое число групп. Минимальное число групп равно двум. Новые данные добавляются в ранее сформированные группы, а в остальных аспектах подходы совпадают. Методология свободна от недостатков концепции Фишера. Она адекватна для исследования устойчивости разнообразных статистических характеристик.

* * *

Современные системы Data Mining обеспечивают эффективные методы представления, хранения, поиска и защиты данных. Эффективный анализ данных предполагает решение ряда методологических проблем, в том числе проблемы объекта эмпирической теории и проблемы ошибки статистической модели. В современных системах Data Mining, рассчитанных на разнообразные приложения, должны быть представлены разные варианты задания ошибки модели. Это обеспечит корректное решение проблемы открытия нового знания и эффективное применение систем Data Mining.

Примечания

1. *Резников В.М.* Вероятностные концепции: анализ оснований и приложений. – Новосибирск: Новосибирск. гос. ун-т, 2005.

2. См.: *Алимов Ю.И.* Альтернатива методу математической статистики. – М.: Наука, 1980; *Эльясберг П.Е.* Измерительная информация: сколько ее нужно? Как ее обрабатывать? М., 1983; *Алимов Ю.И., Кравцов Ю.А.* Является вероятность нормальной физической величиной? // Успехи физических наук. – 1992. – Т. 162, № 7. – С. 149–182.

3. См.: *Орлов А.И.* Эконометрика. – М., 2003.

4. См.: *Kantardzic M*. Data Mining: concepts, models, methods, and algorithms. – 2002.

5. См.: *Vityaev E., Kovalerchuk B.* Relational methodology for Data Mining and knowledge discovery // Sixteenth International Workshop on Database and Expert Systems Applications, 1st International Workshop on Philosophies and Methodologies for Knowledge Discovery (22–26 August 2005, Copengagen, Denmark). – IEEE Computer Society, 2005. – P. 725–729.

6. См.: *Мизес Р.* Вероятность и статистика. – М.; Л., 1930.

7. См.: *Резников В.М.* Вероятностные концепции: анализ оснований и приложений…

8. См.: *Колмогоров А.Н., Гнеденко Б.В.* Предельные распределения для сумм независимых случайных величин. – М.; Л., 1949. – С. 7.

9. См.: *Гнеденко Б.В.* Курс теории вероятностей. – М.: Наука, 1988.

10. См.: *Ю.И.* Альтернатива методу математической статистики. – М.: Наука, 1980.

11. См.: *Howson C., Urbach P.* Scientific reasoning: The Bayesian approach. – Illinois, Open Court, 1996.

12. См.: *Worrall J.* What is evidence in evidence-based medicine? // Philosophy of science. – 2002.

13. См.: *Эльясберг П.Е.* Измерительная информация…

14. Там же.

15. Там же.

16. См.: *Howson C., Urbach P.* Scientific reasoning…

Институт философии и права
СО РАН, г. Новосибирск

Reznikov, V.M. **Some approaches to the problem of model error in Data Mining systems**

The paper proves that the object problem and the problem of model error are important for development of statistics methodology and correct use of formal methods. It shows that Mises was the first who has formulated the problem of empirical theory object, but he has not offer its decision. The paper also shows that the object problem has no decision in the probability theory. New heuristic non-strict approaches to decision of the model error problem are analyzed, these are the guaranteeing conception and the metrological empirical one.