ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 83 за 2022 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

Ю.Н. Орлов, В.О. Соловьев

Об оценке точности обработки больших потоков экспериментальных данных

Рекомендуемая форма библиографической ссылки: Орлов Ю.Н., Соловьев В.О. Об оценке точности обработки больших потоков экспериментальных данных // Препринты ИПМ им. М.В.Келдыша. 2022. № 83. 24 с. https://doi.org/10.20948/prepr-2022-83
https://library.keldysh.ru/preprint.asp?id=2022-83

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

Ю.Н. Орлов, В.О. Соловьев

Об оценке точности обработки больших потоков экспериментальных данных

Москва — 2022

Орлов Ю.Н., Соловьев В.О.

Об оценке точности обработки больших потоков экспериментальных данных

Исследуются вычислительные аспекты обработки большого объема экспериментальных данных, связанные с нестационарностью процесса, неточностью измерения, неточностью классифицирующих алгоритмов. Рассматриваются также ограничения применения байесовского подхода к задаче распознавания образов, когда максимум вероятности соответствия текущего состояния одному из базисных эталонов определяется путем разложения изучаемого фрагмента по известному базису.

Ключевые слова: нестационарный временной ряд, большие данные, базисные паттерны, классификация

Orlov Yu.N., Solovyev V.O.

To the accuracy estimation of the high intensive flows of experimental data

The computational aspects of processing a large volume of experimental data related to the unsteadiness of the process, measurement inaccuracy, and inaccuracy of classifying algorithms are investigated. The limitations of the Bayesian approach to the problem of pattern recognition are also considered, when the maximum probability of matching the current state to one of the basic standards is determined by decomposing the fragment under study according to a known basis.

Key words: non-stationary time series, big data, basis patterns, classification

Содержание

Введение

Данная работа носит методический характер. В ней мы описываем методики, которые следует применять к анализу большого количества данных, требующих автоматической классификации или иной обработки. Если по ряду причин данные оказываются нестационарными, т.е. взятыми из выборок, имеющих разные распределения, то было бы некорректно находить, допустим, средние значения или иные выборочные характеристики, так как они не будут отвечать статистическим критериям на предполагаемом уровне значимости.

В задачах статистической классификации данных часто возникает задача определения оптимального объема выборки, на основе которого можно сделать классификацию с заданной точностью в интервале с заданным уровнем доверия. Для разработки метода, позволяющего решать эту задачу, требуется критерий, связывающий между собой точность идентификации, уровень доверия и объем выборки.

Аналогичная задача возникает и при фильтрации шумовой составляющей при анализе потоков данных (сигналов), куда может вносить искажение и сам приемо-передающий или измерительный прибор.

В каждом таком случае следует формализовать различные причины ошибок, возникающих на стадии обработки и интерпретации данных. Мы предлагаем некоторую единую методическую процедуру, называемую построением согласованного уровня значимости, которая позволяет анализировать большие объемы как стационарных, так и нестационарных данных.

Во-первых, в работе обсуждается определение оптимального объема выборки для формирования базисных паттернов из нестационарного временного ряда данных.

Во-вторых, при нахождении доверительных интервалов позиционирования средних значений выборочных статистик вводится согласованный уровень значимости, поскольку наряду с уровнем доверия, который имеет статистический смысл, необходимо учитывать точность идентификации, имеющую смысл точности инструмента измерения.

В-третьих, при анализе нестационарных потоков данных следует оценивать априорный уровень близости между выборочными распределениями, который соответствует определенному уровню нестационарности. Для этого необходимо формализовать понятие уровня нестационарности и описать методику его применения.

В-четвертых, в работе анализируются погрешности вычислительного характера, которые связаны с классификацией многомерного вектора данных в базисе из эталонов классов. Обсуждаются методы наиболее вероятного эталона и эталона, ближайшего к данному фрагменту в определенной норме.

Эти аспекты представляются основными при анализе больших объемов экспериментальных данных.

1. Оптимальный объем выборки

Рассмотрим операцию усреднения данных по некоторому промежутку времени. Пусть $x(t)$ есть значение случайной величины ξ в момент времени t. Тогда текущие оценки статистических характеристик данного процесса (непрерывного или дискретного) используют усреднение определенных операторов $\Psi[x(t)]$ по некоторому промежутку $\Delta = [t_0, t_0 + T]$:

$$\langle \Psi[x(t)] \rangle_\Delta = \frac{1}{T} \int_{t_0}^{t_0+T} \Psi[x(t)] dt . \qquad (1)$$

В качестве оператора Ψ могут выступать выборочные моменты, операторы разностного дифференцирования, ядра интегральных преобразований типа разложения в ряд Фурье и др. Эти же средние характеристики могут быть получены посредством операций с плотностью выборочной функции распределения $f_\Delta(x,t)$: можно определить моменты этого распределения, характеристическую функцию и т.д.

Применение преобразований вида (1) для вычисления текущих оценок статистических характеристик нестационарного процесса $x(t)$ приводит к появлению принципиально неустранимых погрешностей: погрешности, возникающей за счет конечности времени усреднения, т.е. вследствие недостаточной репрезентативности объема выборки, а также погрешности, возникающей за счет изменения статистики на интервале усреднения. Будем называть погрешность оценки статистики (1) за счет конечности промежутка Δ погрешностью первого типа и обозначать Σ_1, а погрешность за счет нестационарности статистики – погрешностью второго типа и обозначать Σ_2.

Важно понимать, что по результатам наблюдений нельзя статистически отделить в конечной выборке фактор нестационарности от фактора нерепрезентативности. Поэтому обозначения Σ_1 и Σ_2 соответствуют не строго определяемым величинам, а просто указывают на то, что существуют разные факторы, приводящие к ошибкам. В то же время для минимизации ошибки прогноза желательно уметь анализировать влияние этих факторов по отдельности. С этой целью далее будут введены функционалы, которые определяются в основном одним из рассматриваемых факторов, хотя влияние другого и не может быть полностью исключено. Идея их введения состоит в следующем.

Если случайный процесс стационарный в узком смысле, то две его выборочных плотности функции распределения (далее ВПФР), построенные по любым двум различным выборкам, будут достаточно близки. Поэтому величину нестационарности можно оценить по близости между ВПФР. Разумеется, поскольку выборки конечны, то между соответствующими ВПФР будет некоторое различие независимо от того, является ряд стационарным или нет. Но с увеличением объема выборок это различие будет стремиться к нулю для стационарных распределений.

С другой стороны, величину статистического разброса естественно оценивать по дисперсии выборочного распределения. В ней также содержится влияние обоих факторов, но если окажется возможным зафиксировать допустимое расстояние между ВПФР и рассматривать только такие выборки, которые порождают достаточно близкие распределения, то дисперсия будет (приближенно) определяться в основном только квазистационарной статистикой ряда. Таким образом, расстояние между ВПФР и величину выборочной дисперсии при фиксированном значении этого расстояния можно рассматривать в качестве кандидатов для формального определения Σ_2 и Σ_1.

Формализуем поставленную задачу в терминах выборочных функций распределения (далее ВФР).

Итак, основной проблемой является минимизация функционала полной ошибки Σ в оценке статистики (1). Таким функционалом может быть, например, $\Sigma = \Sigma_1 + \Sigma_2$, или $\Sigma = \sqrt{\Sigma_1^2 + \Sigma_2^2}$ и т.п. Выбор наиболее адекватного функционала полной ошибки определяется условиями конкретной задачи и здесь не исследуется. Предполагается только, что процессы, приводящие к ошибкам первого и второго типов, независимы.

Будем определять функционал полной ошибки по формуле

$$\Sigma = \sqrt{\Sigma_1^2 + \Sigma_2^2}\,. \qquad (2)$$

Задача оптимизации объема T выборки в (1) в терминах функционала ошибки (2) состоит в определении такого объема T_{opt}, при котором $\Sigma(T_{opt}) = \min$, и построении соответствующего численного алгоритма нахождения T_{opt} для заданного нестационарного временного ряда. Оптимизационная задача возникает в силу того, что ошибки $\Sigma_1(T)$ и $\Sigma_2(T)$ как функции объема выборки имеют различное поведение. Именно: чтобы уменьшить ошибку $\Sigma_1(T)$ вследствие неполной статистической репрезентативности, следует увеличивать объем T, а для уменьшения ошибки $\Sigma_2(T)$ от влияния нестационарности на статистические характеристики временного ряда следует уменьшать объем выборки.

В настоящей работе способы дискретизации непрерывных случайных процессов не рассматриваются. Считается, что величина шага по времени, по истечении которого происходит измерение случайной величины, задана. Оптимизировать же объем выборки требуется для того, чтобы с заданной точностью спрогнозировать изменение выборочной функции распределения на некотором промежутке, называемом горизонтом прогноза. Требование соблюдения некоторой заданной точности прогнозирования приводит к постановке условия близости ВПФР, отнесенных к различным моментам времени, в подходящем функциональном пространстве. Формулировка такого условия близости объединяет две статистики данного временного ряда $x(t)$: статистику объема выборки $T_{opt}(t)$ и статистику горизонта планирования (или прогноза) $\tau(t)$. Именно: требуется определить $T_{opt}(t)$ так, чтобы ВПФР,

построенные по выборкам этих оптимальных объемов на промежутках $[t, t+\tau(t)]$, различались бы меньше, чем на заданное число ε, характеризующее уровень нестационарности.

Ошибкой δ прогноза временного ряда $x(t_i) \equiv x_i$ будем называть величину среднеквадратичного отклонения прогнозных значений $\tilde{x}_i$ от фактических x_i на промежутке горизонта планирования:

$$\delta = \sqrt{\frac{1}{\tau}\sum_{i=t+1}^{t+\tau}\delta_i^2}, \quad \delta_i = \tilde{x}_i - x_i. \tag{3}$$

Пусть значения ряда принадлежат некоторому конечному промежутку $[a;b]$, который без ограничения общности можно принять за отрезок $[0;1]$. Ошибкой прогноза ВПФР в момент времени t будем называть интегральное абсолютное отличие прогнозной ВПФР $\tilde{f}(x,t)$ от фактической $f(x,t)$, построенных по выборкам равных объемов. Обозначим соответствующую величину $\varepsilon(t)$:

$$\varepsilon(t) = \int_0^1 \left| \tilde{f}(x,t) - f(x,t) \right| dx. \tag{4}$$

Важно подчеркнуть, что ошибка прогноза ВПФР и ошибка прогноза временного ряда, по выборке из которого построена данная ВПФР, – это разные ошибки, определенные для разных математических объектов и в различных функциональных пространствах. Следовательно, чтобы использовать величины (3) и (4) совместно, надо придать корректный теоретико-вероятностный смысл получающемуся функционалу совокупной ошибки (2), поскольку сам по себе он лишь формализует некоторую идею, с помощью которой можно было бы оценить меру неточности прогноза нестационарных временных рядов. Такая формализация может быть проведена на основе нижеследующих оценок, получающихся из (3), если записать выборочную дисперсию в терминах ВПФР.

Предположим, что практическими соображениями диктуется некоторая максимально допустимая ошибка прогноза δ. Естественно, она не может быть меньше, чем оценка ошибки прогноза в смысле среднего квадратичного, т.е. чем корень из дисперсии, отвечающей прогнозной ВПФР $\tilde{f}$. Эту дисперсию обозначим через $\tilde{\sigma}^2$:

$$\tilde{\sigma}^2(t) = \int_0^1 \left(x - \tilde{\bar{x}}(t)\right)^2 \tilde{f}(x,t)dx, \quad \tilde{\bar{x}}(t) = \int_0^1 x\tilde{f}(x,t)dx. \tag{5}$$

Различие между прогнозными и фактическими средними величинами можно оценить из неравенства

$$\left|\tilde{\bar{x}} - \bar{x}\right| = \int_0^1 \left| x \cdot (\tilde{f} - f) \right| dx = \int_0^1 |x| \left| \tilde{f} - f \right| dx \le \varepsilon. \tag{6}$$

Рассмотрим ошибку (3) прогнозирования значения ряда в некоторый момент времени. Прогнозным значением ряда в момент времени t будем считать среднее значение $\tilde{\bar{x}}(t)$, определенное по прогнозному распределению

$\tilde{f}$. Как уже говорилось, формальной оценкой ошибки такого прогноза является $\tilde{\sigma}(t)$. Однако при сравнении с фактом следует учесть, что фактическая ВПФР изменилась по сравнению с прогнозной, так что оценка сверху квадрата фактической ошибки прогноза ряда в силу (6) составит величину

$$\delta^2 = \int_0^1 \left(x - \tilde{\bar{x}}\right)^2 f(x,t)dx = \int_0^1 \left(x - \tilde{\bar{x}} + \bar{x} - \bar{x}\right)^2 f(x,t)dx = \sigma^2 + (\bar{x} - \tilde{\bar{x}})^2 \le \sigma^2 + \varepsilon^2. \qquad (7)$$

Использованные при выводе оценки (7) величины ошибок имеют очевидный смысл: $\sigma = \Sigma_1$, $\varepsilon = \Sigma_2$. Таким образом, оценка (7) является обоснованием использования функционала (2) в качестве меры неточности прогноза нестационарного временного ряда.

Поскольку для стационарного ряда различие между двумя выборочными функциями распределения из одной и той же генеральной совокупности можно сделать сколь угодно малым, увеличивая объем выборки, то из формулы (7) следует, что ошибка прогноза будет тем меньше, чем точнее будет определена дисперсия ряда, т.е. чем большее количество членов этого ряда учитывается при построении ВПФР. Однако если ряд нестационарный, то увеличение объема выборки хотя и приводит к уменьшению ε, одновременно с этим в общем случае приводит к увеличению дисперсии σ^2. Отметим, что различие между дисперсиями двух ВПФР в соответствии с условием (4) оценивается сверху как

$$\left|\tilde{\sigma}^2 - \sigma^2\right| \le \int_0^1 x^2 \left|\tilde{f} - f\right| dx + \left|\tilde{\bar{x}}^2 - \bar{x}^2\right| \le \varepsilon + \left|\tilde{\bar{x}} + \bar{x}\right|\varepsilon \le 3\varepsilon. \qquad (8)$$

Таким образом, требуется определить такой объем выборки, при котором верхняя оценка величины в левой части (7) минимальна.

2. Согласованный уровень значимости

Теория стационарных случайных процессов позволяет по конечной выборке судить о том, с какой доверительной вероятностью ряд является стационарным в широком или узком смыслах. Соответствующая задача состоит в проверке статистических гипотез относительно того, насколько близко поведение выборочных средних значений к тому, что диктуется предельными теоремами. Более сложная задача оценки близости нестационарных распределений будет рассмотрена в следующих параграфах.

В стационарном случае выборочные средние с увеличением объема выборки должны стабилизироваться к соответствующим моментам генеральной совокупности, если последние существуют, а отклонения от этих постоянных величин оцениваются дисперсией выборочных моментов. Распределение отклонений в стационарном случае является нормальным.

Проверка статистической гипотезы состоит в следующем. По имеющейся выборке вычисляется некоторая статистика, распределение вероятностей которой известно, если считать, что осуществлена именно проверяемая

гипотеза. Эта статистика называется критерием. Затем вычисляется предполагаемая вероятность реализации критериальной статистики в данной выборке. Проверяемая гипотеза H_0 называется нулевой. Гипотеза H_1, состоящая в том, что гипотеза H_0 неверна, называется альтернативной.

Подчеркнем, что статистическая гипотеза, в отличие от иных гипотез, относится к функции распределения случайной величины, а не к значению самой случайной величины. Например, гипотеза о том, что среднее значение случайной величины равно 0,5, является статистической, а гипотеза о том, что определенное будущее значение временного ряда равно 0,5, таковой не является. Статистика не предсказывает значения случайных величин, которые могут быть наблюдены в эксперименте, а оценивает параметры их распределений.

Уровень значимости – это априори заданное значение вероятности реализации критерия, несовместимое с признанием случайности. Событие называется значимым, а не случайным, если вероятность его появления меньше, чем принятый уровень значимости. Соответствующее значение статистики критерия называется критическим. Дополнение уровня значимости до единицы называется уровнем достоверности. Если значение статистики критерия, вычисленное по имеющейся выборке, больше критического, то на выбранном уровне значимости проверяемая гипотеза отклоняется.

Поскольку принятие или отклонение гипотезы носит вероятностный характер, существует вероятность ошибочного решения. При этом различают ошибки двух типов. Ошибка первого рода состоит в отклонении верной гипотезы. Вероятность ошибки первого рода не выше уровня значимости. Ошибка второго рода состоит в принятии ложной гипотезы. Эффективность или мощность критерия, выбранного для проверки гипотезы H_0, оценивается вероятностью отклонения ложной гипотезы. Чем выше эта вероятность, т.е. чем меньше вероятность ошибки второго рода, тем более мощным является критерий.

Обычно гипотеза указывает область, называемую доверительной, где могут находиться истинные параметры, характеризующие задачу. Зададим некоторое число $\alpha > 0$ как вероятность ошибки в определении этой области. Вероятность $1-\alpha$ называется коэффициентом (или уровнем) доверия. Для гипотез одностороннего оценивания границы ε, такой, что $P(x > \varepsilon) = \alpha$, доверительная область определяется $1-\alpha$-квантилем распределения проверяемой статистики относительно x, плотность распределения которой есть $f(x)$. Двусторонняя оценка интервала (a, b), содержащего x, на уровне значимости α определяется из условия $\int_a^b f(x)dx = 1-\alpha$. Если распределение статистики близко к симметричному, то полагают $a = -b$ и требуют выполнения условия $\int_b^{+\infty} f(x)dx = \alpha/2$.

Для стационарных распределений отклонения средних значений сумм случайных независимых величин от их генеральных значений распределены асимптотически нормально, поэтому в статистике изучаются главным образом выборки из нормального распределения. Тогда отличие изучаемого распределения от стационарного можно обнаружить, анализируя его выборочные моменты и определяя, насколько точно можно удовлетворить теореме Леви-Линдеберга.

Если выборка не является нормально распределенной, то стандартные формулы для получения доверительных интервалов неприменимы. Следует анализировать собственно выборочное распределение в условиях, когда генеральное распределение не известно. В этом случае задача о принадлежности двух выборок одной генеральной совокупности решается статистикой Колмогорова-Смирнова $S_N = \sup_x \left| F_{1,N}(x) - F_{2,N}(x) \right|$, для которой справедлива асимптотика

$$\lim_{N\to\infty} P\left\{ 0 < \sqrt{\frac{N}{2}}\, S_N < z \right\} = K(z). \qquad (9)$$

Это означает, что если для двух ВФР было найдено значение S_N и вычислена величина $z = \sqrt{\frac{N}{2}} S_N$, то величина $1 - K(z)$ приближенно считается равной вероятности того, что $\sqrt{\frac{N}{2}} S_N \geq z$. Задав уровень значимости α, считаем, что если $1 - K(z) < \alpha$, то осуществилось маловероятное событие, несовместимое с понятием случайности, и эти выборки следует считать различными. Некоторая неопределенность вывода состоит в том, что надо априори задать желаемый уровень малости критерия $1 - K(z)$. Какую вероятность следует считать достаточной для того, чтобы признать выборки одинаковыми?

Следовательно, при анализе выборок определенной длины неправильно задавать априори желаемый уровень значимости, так как для заданной длины N выборки лишь при одном значении $\varepsilon = \varepsilon_0(N)$ вероятность превышения значения ε_0 равна значимости используемого для этой цели критерия, и это значение находится из уравнения

$$1 - \varepsilon_0 = K\left(\sqrt{\frac{N}{2}} \varepsilon_0 \right). \qquad (10)$$

Решение уравнения (10) единственно, поскольку правая часть как функция ε_0 монотонно возрастает от нуля до единицы, а левая монотонно убывает от единицы до нуля.

Какой уровень значимости α следует априори задать для получения практически разумных результатов? Подчеркнем, что точность оценки, например, среднего значения случайной величины дает не только уровень

значимости, но и содержащий это значение интервал. Возможная неприемлемость оценки возникает тогда, когда доверительный интервал оказывается слишком широк, либо уровень значимости слишком велик. Поэтому важно согласовать обе этих величины.

Уменьшая уровень значимости, мы расширяем интервал доверия, но тем самым увеличивается ошибка практического применения полученных оценок точно так же, как она увеличивается при высоком уровне значимости и узком доверительном интервале. Так, практически ненужным является результат, состоящий в том, что с доверительной вероятностью 1,0 среднее значение некой величины заключено в промежутке $[0;1]$, представляющем все множество допустимых значений этой величины. Точно так же не представляет интереса знание о том, что среднее значение заключено в узком промежутке с доверительной вероятностью, близкой к нулю.

С практической точки зрения выбранный уровень значимости α означает, что при проведении большого числа однотипных экспериментов по проверке статистической гипотезы в условиях, когда она заведомо верна, в доле α случаев формально будет допущена ошибка первого рода. Результаты экспериментов равновероятны, так как выборки берутся из одного и того же распределения. Поэтому практическая точность критерия принятия гипотезы равна наибольшему из двух чисел: априорному уровню значимости и доле меры доверительного интервала в мере множества допустимых исходов. Для равномерно ограниченной случайной величины удобно нормировать ее значения на промежуток $[0;1]$. В этом случае можно определить согласованный доверительный интервал условием равенства указанной доли и уровня значимости. Пусть $f_T(x)$ есть выборочная плотность функции распределения изучаемого признака. Тогда согласованным доверительным интервалом $(a;b)\subset[0;1]$ на уровне значимости ε называется интервал $(a;a+\varepsilon)$, где согласованный уровень значимости $\varepsilon(a)$ при заданном левом конце интервала находится из условия

$$\int_a^{a+\varepsilon} f_T(x)dx = 1-\varepsilon\,. \qquad (11)$$

Сам же левый конец интервала находится в общем случае неоднозначно. Имеет смысл выбрать такое значение a, при котором достигается наилучшая точность в оценке изучаемого показателя, т.е.

$$a_{opt} = \arg\min_a \varepsilon(a)\,. \qquad (12)$$

Описанный подход позволяет определить фактическую ошибку, совершаемую при практическом использовании того или иного статистического критерия. Возможная неприемлемость оценок возникает тогда, когда доверительный интервал оказывается слишком широк. Повышая уровень значимости, мы расширяем этот интервал, но тем самым увеличивается ошибка практического применения полученных оценок. Если же интервал сузить, то уровень значимости будет снижаться. Так, практически ненужным является

результат, состоящий в том, что с доверительной вероятностью 0,99 среднее значение некой величины, равное, по оценкам, 0,5, заключено в промежутке $[0;1]$. Точно так же не представляет интереса попадание среднего 0,5 в промежуток $[0{,}49;\ 0{,}51]$ с вероятностью, близкой к нулю. Для практической приемлемости статистических оценок требуется, чтобы точность позиционирования среднего значения совпадала бы с уровнем значимости.

Пусть статистика среднего значения $m_1(n)$ некоторой случайной величины ξ по выборке объема n такова, что $m_1(n) \neq 0$. Введем выборочную функцию распределения $f_n(x)$ отклонений фактических значений x_i от среднего значения. Тогда согласованным доверительным интервалом $(a;b)$ на уровне значимости α для среднего значения $m_1(n)$ будем называть интервал, удовлетворяющий условию

$$\frac{b-a}{2m_1(n)} = 1 - \int_a^b f_n(x)dx = \alpha . \qquad (13)$$

Смысл интервала (13) в том, что его полуширина составляет $100\alpha\%$ от среднего значения, причем α есть уровень значимости самого интервала.

3. Разбиение гистограммы

Рассмотрим равномерное разбиение гистограммы, когда ширина интервала равна $1/n$. Рассмотрим последовательность из N значений случайной величины ξ, которые попали в классовые интервалы $\Delta_i = [a_{i-1}; a_i)$, $a_i = (i-1)/n$. Элементы этой последовательности обозначим x_j, $j = 1, 2, ..., N$. Возникает вопрос: какова должна быть мелкость разбиения, чтобы с нужной точностью получить оценку ПФР, составив также и адекватное представление о виде распределения по виду ВПФР?

Очевидно, наилучшая точность в оценке вероятностей достигается при разбиении на один классовый интервал: тогда внутри него ВПФР совпадает с ВФР и равна единице. Однако практически этот результат бесполезен, ибо не дает представления о виде распределения. Если же взять число промежутков разбиения слишком большим, например, $n > N$, то оказывается невозможным трактовать частоты как оценки соответствующих вероятностей, поскольку тогда не может быть реализовано условие сходимости эмпирической вероятности к теоретической при $N \to \infty$. Следовательно, для каждого значения N существует оптимальное равномерное разбиение гистограммы, при котором и детализация распределения достаточна, и точность оценки ПФР оказывается приемлемой. Чтобы формализовать условие оптимальности, обратимся к классическому решению статистической задачи об оценивании доверительного интервала для статистики среднего значения на основе наблюдаемых данных.

Когда генеральная дисперсия σ^2 теоретического распределения не известна, а оценивается только по выборочной дисперсии $s_2(N)$, для получения

доверительного интервала генерального среднего значения μ следует рассматривать *t*-статистику Стьюдента, которая применительно к этой задаче имеет вид

$$t = \sqrt{N-1}\,\frac{|m_1(N)-\mu|}{\sqrt{s_2(N)}}. \qquad (14)$$

Здесь

$$m_1(N) = \frac{1}{N}\sum_{k=1}^{N} x_k \equiv \langle x \rangle_N, \quad s_2(N) = \left\langle \left(x - m_1(N)\right)^2 \right\rangle_N. \qquad (15)$$

Смысл статистики (14) в том, что на уровне значимости ε выражение $|m_1(N)-\mu|$ не превосходит величины

$$t_{1-\varepsilon/2}(N-1)\sqrt{s_2(N)}/\sqrt{N-1}, \qquad (16)$$

где $t_\alpha(N-1)$ есть α-квантиль распределения Стьюдента с $N-1$ степенью свободы. При больших N можно считать $N-1 \approx N$ и число степеней свободы в квантиле распределения Стьюдента взять для простоты бесконечным (тогда это распределение совпадает с нормальным).

Согласно центральной предельной теореме, отклонение выборочного среднего значения $m_1(N)$, определяемого в (15) по выборке длины N, от генерального среднего распределено асимптотически нормально с нулевым средним и стремящейся к нулю дисперсией σ^2/N, где σ^2 есть дисперсия этой величины по гипотетической генеральной совокупности. Рассмотрим в качестве такой случайной величины саму эмпирическую частоту $p_i(N)$ попадания в i-й классовый интервал. Формально эмпирическая частота есть среднее выборочное значение случайной величины, называемой индикатором I_{ik} принадлежности результата k-го наблюдения в i-му классовому интервалу. Индикатор I_{ik} определяется формулой

$$I_{ik} = \begin{cases} 1, & x_k \in \Delta_i; \\ 0, & x_k \notin \Delta_i. \end{cases} \qquad (17)$$

Как видно из формулы (17), эмпирическую частоту можно записать в виде

$$p_i(N) = \langle I \rangle = \frac{1}{N}\sum_{k=1}^{N} I_{ik}.$$

Выборочная дисперсия эмпирической частоты $p_i(N)$ равна

$$s_2(N;i) = \left\langle \left(I - \langle I \rangle\right)^2 \right\rangle = \left\langle I^2 \right\rangle - \langle I \rangle^2 = \langle I \rangle - \langle I \rangle^2 = p_i(N)\cdot\left(1 - p_i(N)\right). \qquad (18)$$

Таким образом, из (16) следует, что оценка генеральной частоты f_i^* заключена в интервале

$$\left| p_i(N) - f_i^* \right| \le t_{1-\varepsilon/2}(N-1)\sqrt{s_2(N;i)}/\sqrt{N-1}. \qquad (19)$$

Зададим уровень значимости ε равным уровню неопределенности в позиционировании доверительного интервала $\left|p_i(N) - f_i^*\right|$. Поскольку исходным требованием является близость оценки к генеральной частоте, естественно потребовать выполнения условия

$$\sum_{j=1}^{n}\left|p_i(N) - f_i^*\right| \le \varepsilon . \qquad (20)$$

Очевидно, условию (20) можно удовлетворить, потребовав выполнения более жесткого условия – близости каждой из частот в отдельности:

$$\left|p_i(N) - f_i^*\right| \le \varepsilon f_i^* . \qquad (21)$$

Таким образом, если для каждой эмпирической частоты выполнить условие $t_{1-\varepsilon/2}\dfrac{\sqrt{s_2(N;i)}}{\sqrt{N}} \le \varepsilon f_i^*$, то одновременно с выполнением условия (20) будет достигнут и требуемый уровень значимости для статистики Стьюдента в (19). Однако если некоторые вероятности в результате выбранного разбиения на классовые интервалы сами оказались малы, много меньше ε, то нет необходимости требовать, чтобы и они были оценены с той же точностью. Поэтому уместно для каждой вероятности выбрать свою точность аппроксимации ε_i и считать, что требуемый в целом уровень значимости определяется средневзвешенной по разбиению точностью, так что

$$t_{1-\varepsilon/2} = \frac{1}{\Sigma_N(n)}\sum_{i=1}^{n}\sqrt{s_2(N;i)}\,t_{1-\varepsilon_i/2} , \qquad (22)$$

где сумма, определяющая влияние мелкости разбиения гистограммы на точность оценки эмпирических вероятностей, равна

$$\Sigma_N(n) = \sum_{i=1}^{n}\sqrt{s_2(N;i)} = \sum_{i=1}^{n}\sqrt{p_i(N)\cdot\left(1-p_i(N)\right)} . \qquad (23)$$

Сумма (23) выражает качество приближения плотности гистограммой, поскольку чем меньше сумма, тем выше точность оценки ВПФР, т.е. тем меньше число ε. С увеличением числа интервалов сумма (23) возрастает, что означает снижение точности оценки ВПФР. В стационарном случае эта сумма представляет собой некий эффективный функционал учета особенностей графика функции плотности при аппроксимации плотности генерального распределения.

Теперь из (23) получаем, что

$$\sum_{i=1}^{n} t_{1-\varepsilon_j/2}\frac{\sqrt{s_2(N;i)}}{\sqrt{N}} = t_{1-\varepsilon/2}\frac{\Sigma_N(n)}{\sqrt{N}} \le \varepsilon\sum_{i=1}^{n} f_i^* = \varepsilon ,$$

откуда на уровне значимости ε следует оценка

$$\frac{t_{1-\varepsilon/2}}{\varepsilon} \le \frac{\sqrt{N}}{\Sigma_N(n)} . \qquad (24)$$

При заданной точности ε и способе разбиения гистограммы формула (24) для знака равенства дает оценку на минимальную длину выборки, при которой

эта точность достигается в среднем. Поскольку функция $t_{1-\varepsilon}$ табулирована (см., напр., [5]), то функция $t_{1-\varepsilon}/\varepsilon$ известна. Она монотонно убывает с ростом ε, поэтому к ней существует обратная, значение которой и дает верхнюю оценку точности определения эмпирических вероятностей по заданному разбиению гистограммы. Обозначим для краткости

$$\varphi(\varepsilon)=\frac{t_{1-\varepsilon}}{\varepsilon},\quad \psi=\varphi^{-1},\quad z\equiv z(N,n)=\frac{\sqrt{N}}{\Sigma_N(n)}. \tag{25}$$

Тогда точность оценки ВПФР определяется формулой

$$\varepsilon=2\psi(2z). \tag{26}$$

4. Фильтрация шума

Во многих задачах статистического распознавания образов надо очистить сигнал (искомый неизвестный образ) от случайных возмущений, связанных с неточностью измерений и неточностью функционирования самой изучаемой системы. Задачи фильтрации в линейной постановке достаточно подробно исследованы, тогда как нелинейная фильтрация изучена менее детально. Мы будем использовать носитель эмпирической плотности совместного распределения случайной величины и ее приращений для распознавания нелинейно коррелированных временных рядов. Сложность анализа состоит в том, что сама мелкость разбиения гистограммы для кластеризации наблюдаемых значений случайной величины является определенным фильтром и требует оптимизации.

Кратко изложим основные подходы к линейной фильтрации шумов при анализе стационарных случайных процессов. Случайный процесс $x(t)$ называется стационарным в широком смысле, если его математическое ожидание $a=Mx(t)$ не зависит от t, а автокорреляционная функция (АКФ) зависит только от разности моментов времени:

$$A(t,s)=M\big(x(t)-a\big)\big(x(s)-a\big)\equiv B(t-s).$$

Спектральным представлением действительного скалярного случайного процесса с АКФ $B(t)$ называется представление

$$B(t)=\int_{-\infty}^{+\infty}e^{i\lambda t}dQ(\lambda)\ . \tag{27}$$

Преобразование $x(t)\to y(t)$ процесса $x(t)$ называется линейным фильтром, если процесс $y(t)$ представляется в виде

$$y(t)=\int_{-\infty}^{\infty}h(t-s)x(s)ds\ . \tag{28}$$

Функция $h(t)$ называется импульсной переходной функцией фильтра, а ее Фурье-образ $H(\omega)$ называется частотной характеристикой фильтра.

Известно, что всякий стационарный процесс $x(t)$ может быть представлен единственным образом в виде суммы некоррелированных между собой процессов: сингулярного (детерминированного) и регулярного (случайного).

Сингулярная составляющая стационарного случайного процесса определяется как некоторая динамическая система, а регулярная составляющая ищется на основе следующего утверждения (теорема Вольда).

Для того чтобы стационарный процесс $y(t)$ был регулярным, необходимо и достаточно, чтобы он представлял собой преобразование фильтрации (28) некоторого случайного процесса $x(t)$ с независимыми приращениями (т.е. белого шума).

В этой связи линейные преобразования фильтрации приобрели большую практическую значимость в анализе и моделировании случайных процессов. Однако линейная теория имеет естественные ограничения по эффективности своей применимости.

Рассмотрим в качестве типичного примера задачу фильтрации шума при измерении некоторого полезного сигнала $x(t)$. Пусть процесс измерения состоит в том, что прибор известным образом искажает сигнал, умножая его на некоторую функцию $C(t)$ (пусть она для простоты постоянна), а также добавляет в показания некоторый шум $V(t)$ в виде стационарного случайного процесса с нулевым средним и заданной дисперсией σ_V^2. Итак, вместо величины $x(t)$ наблюдается величина

$$z(t) = Cx(t) + V(t). \tag{29}$$

По ней требуется построить оптимальную в среднеквадратичном оценку измеряемой величины $x(t)$, если эта величина и шум не коррелированы.

Линейная по z оценка $\hat{x}$ при нулевых средних имеет вид

$$\hat{x}(t) = az(t). \tag{30}$$

Параметр a выбирается из условия

$$\frac{\partial}{\partial a}\sum_k \left(x_k - \hat{x}_k\right)^2 = -2\sum_k \left(x_k - \hat{x}_k\right)z_k = 0.$$

Подставляя в это уравнение связь $z_k = Cx_k + V_k$ и учитывая, что $\sum_k x_k V_k = 0$, получаем

$$\sum_k \left(Cx_k + V_k\right)\left(x_k - aCx_k - aV_k\right) = \left(C - aC^2\right)\sigma_x^2 - a\sigma_V^2 = 0,$$

откуда следует

$$a = \frac{C\sigma_x^2}{C^2\sigma_x^2 + \sigma_V^2}. \tag{31}$$

Если шума нет, то $\sigma_V^2 = 0$ и, естественно, $a = 1/C$. Но если шум есть, то наряду с гипотезой об отсутствии корреляции сигнала и шума надо еще знать

дисперсию σ_V^2 шума, что на практике, как правило, не реализуется. Следовательно, полученное решение (31) полезно главным образом теоретически как пример существования метода фильтрации.

Пусть далее наблюдаемый сигнал (29) дискретный, а $C=1$. Если известна не только дисперсия шума и фильтруемого сигнала, но и более общие характеристики – их автокорреляционные функции $B_x(n)$, $B_V(n)$, а также ковариационная функция $A_{xz}(k,k+n)=B_{xz}(n)$, то можно построить соответствующий фильтр (фильтр Винера). Пусть $h(n)$ — искомая импульсная характеристика фильтра. Тогда на выходе из него вместо z_n наблюдается сигнал

$$y_n=\sum_{k=0}^{n} z_k h(n-k). \tag{32}$$

Будем строить фильтр, исходя из минимизации среднего квадрата ошибки фильтрации $(y_n-x_n)^2$. При этом сама ошибка оказывается ортогональной в смысле среднего значения фильтруемому сигналу: $M\left(z_n(y_n-x_n)\right)=0$. Это условие, записанное через АКФ, называется уравнением Винера-Хопфа:

$$B_{zx}(j)=\sum_{k=0}^{n} B_z(j-k)h(k). \tag{33}$$

Если время наблюдения бесконечно, то (33) представляет собой свертку функций так, что их фурье-образы входят в виде произведения, что позволяет определить частотную характеристику фильтра:

$$H_{opt}(\omega)=\frac{S_{zx}(\omega)}{S_z(\omega)},\quad B(t)=\frac{1}{2\pi}\int_{-\infty}^{+\infty} S(\omega)e^{i\omega t}d\omega. \tag{34}$$

Если сигнал и шум не коррелированы, то $S_{zx}(\omega)=S_z(\omega)=S_x(\omega)+S_V(\omega)$ и частотная характеристика фильтра имеет вид

$$H_{opt}(\omega)=\frac{S_x(\omega)}{S_x(\omega)+S_V(\omega)}. \tag{35}$$

Однако спектральная функция шума $S_V(\omega)$ не определяется непосредственно по измеряемому сигналу на входе фильтра. Следовательно, для успешной фильтрации надо иметь «библиотеку» подходящих корреляционных функций, после чего методом проб и ошибок подбирать ту, которая действительно будет фильтровать сигнал.

Если время наблюдения конечно, то вместо (34) рассматривается выборочная ковариационная матрица суммарного сигнала, построенная по выборке некоторой длины *N*:

$$B_z(n)=M\left(Z_n Z_n^T\right),\ Z_n^T=\left(z_n,z_{n-1},...,z_{n-N+1}\right). \tag{36}$$

В итоге решением уравнения Винера-Хопфа является оптимальный фильтр с передаточной функцией $H_n^T=\left(h_0,h_1,...,h_{N-1}\right)$ такой, что

$$H_{n,opt}=B_z^{-1}(n)r_{zx}(n),\ r_{zx}^T(n)=M\left(x_n Z_n^T\right). \tag{37}$$

Однако следует признать, что величина $r_{zx}(n)$ в правой части (37) неизвестна. Аналогичный вывод приходится сделать и относительно более сложных задач. Возникает вопрос: можно ли реализовать на практике некоторый способ распознавания случайного процесса, если его свойства априори не известны?

Предположим, что между элементами x и y двух временных рядов имеется прямая функциональная связь, т.е. $y=\varphi(x),\ x,y\in[0;1]$. Тогда, построив совместное распределение вероятностей $f_N(x,y)$ по выборке длины N, мы обнаружим, что его носитель находится в квадратах $y_i=\varphi(x_j)$, в соответствии с разбиением гистограммы на n равных кластеров. Точность, с которой мы можем говорить о такой функциональной связи, равна точности позиционирования случайных величин в классовых интервалах, т.е. $1/n$. Уровень достоверности полученной оценки равен интегралу от плотности распределения по выбранной доле носителя. Поскольку в данном случае вне отмеченных квадратов нет точек носителя совместного распределения, уровень достоверности равен единице.

Если функциональной связи нет, то при фиксированном номере j интервала по первому аргументу мы обнаружим отличные от нуля значения функции $f_N(j,i)$ для нескольких номеров i интервалов по второму аргументу. При этом с увеличением длины выборки носитель совместного распределения занимает все большую долю области разбиения гистограммы. Это означает, что путем потери точности можно получить достоверную оценку функциональной связи даже в отсутствие таковой, но будет ли это удовлетворять исследователя? Насколько точно нужно позиционировать искомое значение, чтобы не только вероятность его принадлежности определенному интервалу не была исчезающе малой, но и сам интервал существенно отличался бы от всего множества значений случайной величины? Для этого вместо априори задаваемого уровня значимости следует ввести согласованный критерий совместной оценки точности и уровня значимости. Введем величину

$$\delta=\int_{\Omega}dxdy \tag{38}$$

как меру множества $\Omega(x,y)$, принадлежащего носителю совместного распределения, на котором можно говорить о функциональной связи между x и y. Величина δ будет точностью, с которой установлена эта связь, а величина

$$\alpha=\int_{\Omega}f(x,y)dxdy \tag{39}$$

будет давать уровень достоверности найденной связи. Как величина δ, так и α зависят от множества $\Omega(x,y)$. Выберем такую часть этого множества, для которого $\alpha+\delta=1$. Величину $\delta=1-\alpha$ будем называть согласованным (с множеством Ω) уровнем значимости распознавания.

Тогда фильтрацию можно осуществить следующим образом. Отметим ту ячейку гистограммы совместного распределения $f_N(j,i)$, в которой для j-го

классового интервала значение функции $f_N(j,i)$ максимально. Таких ячеек при данной мелкости разбиения будет n, т.е. потенциальная точность распознавания будет равна $\delta = 1/n$. Если при этом сумма отмеченных вероятностей будет приблизительно равна величине $1-\delta$, то задачу распознавания считаем законченной. Если же уровень доверия оказался существенно меньше, то увеличиваем меру носителя, участвующего в распознавании, включая в него в каждом j-ом столбце одну из двух соседних ячеек дополнительно к отмеченной ранее, именно ту из них, в которой вероятность больше. Тем самым повысится величина интеграла в (39), но ухудшится точность (38), она станет равной $\delta = 2/n$. Постепенно ухудшая точность, добиваемся согласования значимости с уровнем доверия, так что $\alpha + \delta = 1$.

После того, как определена полоса носителя Ω, в которой непараметрически была распознана зависимость $y(x)$ в виде собственно самого этого множества Ω, переходим к фильтрации. Рассматриваем пару значений (x_1, y_1). Если эта пара принадлежит множеству Ω, то она выбрасывается из анализируемой совокупности, а если она лежит в ячейке, не входящей в это множество, то остается в нем. В итоге остаются только пары, не связанные функциональной зависимостью. Их статистику можно изучить отдельно от основной зависимости, отфильтрованной на предыдущем этапе.

5. Распознавание и классификация

Задача распознавания состоит в том, чтобы по измеряемым значениям параметра $x \in X \subset R^n$, относящегося к исследуемой системе, определить, в каком состоянии $s \in S \subset R^m$ находится (находилась или будет находиться) система в тот или иной момент времени по отношению к моменту измерения. Состояния s непосредственно не измеримы. Таким образом, распознавание – это отображение $V : X \to S$. Соответствующая функция $s = V(x)$ называется решающим правилом или управляющим функционалом [1].

Измеряемым параметром может быть, например, температура воздуха в определенном географическом месте в определенное время суток, биржевой курс ценной бумаги в определенный момент времени и т.п. Это – одномерный параметр, характеризующий систему (соответственно, погоду или биржу). В качестве измеряемого параметра может выступать упорядоченная последовательность из N наблюденных значений, называемая выборкой (или выборочной траекторией) длины N из временного ряда. Такая выборка – это вектор $x \in R^N$. Объектом анализа может быть также некоторый функционал от выборочной траектории: размах, среднее значение, дисперсия, медиана или иной квантиль распределения, собственно выборочная функция распределения или ее плотность, представленная в виде гистограммы. В частности, если наблюдаемая величина принимает дискретный набор из n значений, то гистограмма представляет собой набор эмпирических частот (или выборочных вероятностей) реализации указанных значений, т.е. это вектор $\mathbf{f}_N(x) \in R^n$, где

нижний индекс N показывает, что этот вектор построен по выборке длины N. Состоянием системы, о котором надо судить по выборке, является, например, тренд вверх или вниз биржевой цены, причем такой, который интересен конкретному торговцу, ожидающему определенную прибыль.

Если решающее правило $s = V(x)$ априори не очевидно, анализируется доступный массив данных для статистического оценивания условной вероятности $P(x|s)$ того, что при наблюдении значения x система находится в состоянии s. Предполагая, что оцененные вероятности правильно отражают скрытую от наблюдателя зависимость $s(x)$, решающее правило выбирается на основе байесовского принципа максимальной вероятности [1]. Именно, считается, что наблюдаемому значению x отвечает то значение состояния s, для которого $P(x|s) = \max$. Этот подход минимизирует ошибку распознавания, понимаемую как долю неверно распознанных состояний.

Во всех прикладных задачах пространство состояний S дискретно в силу специфики вычислительных операций, хотя в теоретических моделях это пространство может быть и непрерывным. В настоящей работе мы будем предполагать, что система может пребывать в некотором конечном числе четко идентифицируемых состояний, так что s – это индекс, пробегающий натуральные значения от 1 до некоторого m. Если других состояний нет, а обучающая выборка для нахождения вероятностей $P(x|s)$ достаточно длинная, причем сами вероятности стационарны, то задача распознавания считается решенной. На практике, однако, часто возникают проблемы вычислительного свойства, препятствующие корректному распознаванию в смысле байесовского подхода, но некоторые следствия из этого подхода могут при всем при том оставаться весьма эффективными для распознавания, не будучи строго обоснованными в отсутствие байесовской корректности. В настоящей работе мы подробно рассмотрим пример такого распознавания, которое опирается на следствие из байесовской теории в виде близости плотностей распределений в норме суммируемых функций, но при этом ключевые условия (вероятностная интерпретация относительных частот или выпуклость конуса соответствующих состояний в пространстве S) не выполнены. Конечно, если вероятностное пространство заведомо неполно, то выбранный метод приведет к неверным результатам. Но если модель представляется адекватной, не следует относиться к таким ситуациям как к неким редким парадоксальным артефактам: авторы на практике сталкиваются с ними постоянно и на собственном опыте убедились, что отказ от теоретически хорошего, но по вычислительным причинам некорректного метода в пользу менее точного, но лучше обоснованного, в ряде случаев снижает точность распознавания до неприемлемой величины. Полезно исследовать причины такого феномена.

Будем рассматривать в качестве измеряемого параметра выборочную плотность функции распределения $f_N(k)$, построенную по выборке длины N, где индекс $k = 1, 2, ..., n$ нумерует классовый интервал, в который попадает наблюдаемое значение x. Такая трактовка измерения часто бывает более

содержательна, чем просто соответствие измеренной величины x номеру состояния s. Например, при анализе трендов в биржевых рядах никогда не бывает монотонного роста траектории на протяжении достаточно большого промежутка времени. На любой выборочной траектории локальные максимумы чередуются с локальными минимумами, но если приростов значений ряда одного знака оказывается заметно больше, чем другого, то такой фрагмент воспринимается как тренд. В то же время отдельный прирост $x(t+1)-x(t)$ не позволяет устойчиво идентифицировать текущую ситуацию на промежутке в N шагов как тренд. В работах [2-5] было исследовано предположение о том, что выборочные распределения, отвечающие явно выделяемым состояниям, таким как тренд вверх (вниз), кластеризуются. Действительно, на практике часто наблюдаются состояния, которые в смысле фрагмента траектории достаточно хорошо соответствуют определенным эталонам в терминах функций распределения или их плотностей. Эталон (или базисный паттерн) распределения временного ряда представляет собой средневзвешенное состояние распределений фрагментов траектории случайного процесса, входящих в выделенный кластер. В таком подходе эталоны в виде функций распределения характеризуют типовые состояния изучаемой системы.

Текущее состояние, распознаваемое по близости выборочного распределения к эталонному паттерну в определенной норме, может относиться как к локально установившемуся эталонному состоянию, так и к переходному состоянию. Это последнее, в свою очередь, может быть близким к новому эталону, а может представляться в виде линейной комбинации уже имеющихся эталонов. Тем самым возникает задача оптимального, т.е. с наименьшей ошибкой, разложения текущего состояния по базисным паттернам. Представляет интерес ситуация, когда любое выборочное состояние может быть с заданной точностью представлено как линейная комбинация базисных паттернов. В результате такого разложения будет решена как задача байесовского распознавания, так и указана вероятность, с которой состояние может относиться к какому-то другому типу.

Пусть векторы $\{\varphi_1,...,\varphi_m\}$, $\varphi_s \in R^n$ представляют собой гистограммы в количестве m штук, каждая из которых содержит n классовых интервалов, так что $\varphi_s(k)$ есть вероятность попадания наблюдаемого значения x в k-й классовый интервал при условии, что система находится в состоянии s. Тогда

$$\forall s \in \{1,...,m\} \ \sum_{k=1}^{n} \varphi_s(k) = 1. \qquad (40)$$

Пусть также вектор $\mathbf{f} \in R^n$, являющийся аналогичной гистограммой вероятностей, определяемой по текущей выборочной траектории длины N, принадлежит выпуклой оболочке векторов $\{\varphi_1,...,\varphi_m\}$, так что

$$\mathbf{f}(k) = \sum_{s=1}^{m} y_s \varphi_s(k), \quad 0 \le y_s \le 1. \qquad (41)$$

Тогда из (40) и (41) следует, что $\sum_{s=1}^{m} y_s = 1$.

Если разложение (41) получено, то, определив номер

$$s^* = \arg\max y_s, \tag{42}$$

строим решающее правило байесовского распознавания: состояние s^* отвечает наблюденному распределению $\mathbf{f}$. Заметим, что в этом случае в любой норме расстояние $\left\|\mathbf{f} - \varphi_{s^*}\right\|$ минимально среди всех расстояний $\rho_s = \left\|\mathbf{f} - \varphi_s\right\|$. Обратное, очевидно, неверно, поскольку только из наличия минимального из расстояний не следует существование выпуклой комбинации вида (41).

В работе [3] было показано, что хотя типовые паттерны состояний определяются с достаточной точностью, приближенное разложение вектора текущего состояния по паттернам, составляющим базис состояний временного ряда, во многих случаях не имеет вероятностной интерпретации: коэффициенты разложения либо отрицательны, либо больше единицы. Тем не менее, распознавание текущей ситуации по правилу

$$s^* = \arg\min\left\|\mathbf{f} - \varphi_s\right\| \tag{43}$$

продолжало оставаться весьма точным. Это происходит вследствие того, что вероятностное пространство состояний по факту не полно. Но также было выяснено, что увеличение количества базисных паттернов приводит к тому, что метод (42) становится неприменим в еще большем числе случаев. В [3] было высказано предположение, что наблюдаемый эффект в большей степени вычислительный, поскольку базисные паттерны близки между собой и матрица Грама базиса оказывалась плохо обусловленной. В результате проецирование из пространства большой размерности n (число классовых интервалов гистограммы) в пространство малой размерности m (число паттернов) оказывалось неустойчивым относительно малых шевелений элементов гистограмм. Отчасти это действительно так, но все же выбор паттернов в примерах биржевых рядов связан с процедурой экспертного отбора определенных ситуаций, т.е. не вполне объективен. Желательно было бы получить более веские доказательства того, что вычислительная процедура может воспрепятствовать применимости байесовского распознавания (42), но не ухудшать точности распознавания по методу (43).

В общем виде задача разложения вектора $\mathbf{f} \in R^n$ по заданному набору линейно независимых векторов $\{\varphi_1,...,\varphi_m\}$, $\varphi_s \in R^n$ сводится к нахождению вектор-строки $\mathbf{y}^T = (y_1,...,y_m)$, минимизирующей в смысле 2-нормы функционал $\left\|\mathbf{f} - \Phi\mathbf{y}\right\|$, где $\Phi_{n\times m}$ есть матрица, столбцы которой составляют векторы φ_s. Минимизация этого функционала осуществляется ортогональным проектированием вектора $\mathbf{f}$ на m-мерное подпространство, натянутое на векторы $\{\varphi_1,...,\varphi_m\}$. Это проектирование представляет собой так называемое QR-разложение [6] матрицы Φ в произведение специальной матрицы $Q_{n\times m}$,

такой, что $Q^T Q = I_{m\times m}$, и верхней треугольной матрицы $R_{m\times m}$. В результате такого разложения получается следующее представление вектора $\mathbf{f}$:

$$\mathbf{f} - \Phi\mathbf{y} = I\mathbf{f} - QR\mathbf{y} = \left(I - QQ^T + QQ^T\right)\mathbf{f} - QR\mathbf{y} = Q\left(Q^T\mathbf{f} - R\mathbf{y}\right) + \left(I - QQ^T\right)\mathbf{f}. \qquad (44)$$

Векторы, в виде суммы которых в последнем равенстве (5) представлено данное разложение, ортогональны:

$$\left(R\mathbf{y} - Q^T\mathbf{f}\right)^T Q^T\left(I - QQ^T\right)\mathbf{f} = \left(R\mathbf{y} - Q^T\mathbf{f}\right)^T\left(Q^T{}_{p\times n} I_{n\times n} - I_{p\times p} Q^T{}_{p\times n}\right)\mathbf{f} =$$

$$= \left(R\mathbf{y} - Q^T\mathbf{f}\right)^T\left(Q^T{}_{p\times n} - Q^T{}_{p\times n}\right)\mathbf{f} = \left(R\mathbf{y} - Q^T\mathbf{f}\right)^T O_{p\times n}\mathbf{f} = \mathbf{o}.$$

Второе слагаемое в (5) не зависит от коэффициентов разложения $\mathbf{y}$. Следовательно, с учетом ортогональности указанных слагаемых, минимальное по $\mathbf{y}$ значение нормы $\|\mathbf{f} - \Phi\mathbf{y}\|$ равно норме этого второго слагаемого и достигается тогда, когда первое слагаемое равно нулю: $R\mathbf{y} - Q^T\mathbf{f} = \mathbf{o}$.

Итак, оптимальное разложение определяется вектором

$$\mathbf{y}_{opt} = R^{-1}Q^T\mathbf{f}. \qquad (45)$$

Величина

$$\mathbf{r} = \mathbf{f} - \Phi\mathbf{y}_{opt} = \left(I - QQ^T\right)\mathbf{f} \qquad (46)$$

есть невязка разложения (44). Ошибкой разложения считается 2-норма невязки, т.е. величина $\delta = \|\mathbf{r}\| = \left\|\left(I - QQ^T\right)\mathbf{f}\right\|$. Относительная ошибка определяется как

$$\varepsilon = \frac{\delta}{\|\mathbf{f}\|} = \frac{\left\|\left(I - QQ^T\right)\mathbf{f}\right\|}{\|\mathbf{f}\|}. \qquad (47)$$

Пусть теперь, как это и бывает на практике, вектор текущего состояния $\mathbf{f}$ и матрица Φ базисных паттернов известны неточно. Неточность здесь имеет не измерительную, а статистическую природу, поскольку вместо генеральных совокупностей приходится иметь дело с выборочными распределениями. Возникает вопрос: как эта неточность повлияет на вычисление оптимального разложения, насколько эта процедура устойчива к малым возмущениям, какова в этом случае невязка? Положим

$$\xi = \max\left(\frac{\|\Delta\Phi\|}{\|\Phi\|}, \frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|}\right) \qquad (48)$$

и введем число обусловленности $\kappa(\Phi)$ матрицы Φ в смысле 2-нормы как отношение наибольшего и наименьшего ее сингулярных чисел. Поскольку матрица Φ по построению имеет полный столбцовый ранг, ее наименьшее сингулярное число строго больше нуля. Однако если базисные векторы оказываются близкими, то число обусловленности может быть очень большим. Согласно [7], 2-норма относительной вариации оптимального разложения оценивается сверху следующим образом:

$$\frac{\|\Delta\mathbf{y}\|}{\|\mathbf{y}\|} \le \xi\cdot\left(\frac{2\kappa(\Phi)}{\cos\theta} + \kappa^2(\Phi)tg\theta\right) + O\left(\xi^2\right), \qquad (49)$$

где $\sin\theta = \varepsilon$ есть синус угла между раскладываемым вектором $\mathbf{f}$ и вектором $\Phi\mathbf{y}_{opt}$.

Введем норму ρ_{ik} как расстояние между ПФР выборочных фрагментов i и k в норме суммируемых функций:

$$\rho_{ik} = \|f_i - f_k\| = \sum_{j=1}^{J(n)} |f_i(j) - f_k(j)|. \qquad (50)$$

Для каждого класса a построим плотность функции распределения $g_a^+(\rho)$ отклонений $\rho_{i_a,a}$ его фрагментов от эталона, а также распределение $g_a^-(\rho)$ отклонений $\rho_{k_b,a}$ чужих фрагментов от эталона.

Обозначим через $G^{\pm}(\rho) = \int_0^{\rho} g^{\pm}(r)dr$ соответствующие интегральные функции распределения расстояний между фрагментами и эталонами. Минимальное значение ρ, при котором $G^+(\rho) = 1$, обозначим ρ^+, а максимальное значение ρ, при котором $G^-(\rho) = 0$, обозначим ρ^-. Смысл введенных величин в том, что все ПФР фрагментов находятся на расстоянии не более ρ^+ от соответствующих авторских эталонов и на расстоянии не менее ρ^- от чужих эталонов. Величина $1 - G^+(\rho^-)$ есть вероятность ошибочно признать за принадлежащий классу «a» чужой фрагмент (ошибка второго рода, пропуск цели), а величина $G^-(\rho^+)$ есть вероятность ошибочно отвергнуть свой фрагмент, посчитав его за чужой (ошибка первого рода, ложная тревога). Назовем расстоянием разделения авторов такое значение ρ^*, для которого ошибка идентификации класса минимальна:

$$\rho^* = \arg\min\left(1 - G^+(\rho) + G^-(\rho)\right) = \arg\max\left(G^+(\rho) - G^-(\rho)\right). \qquad (51)$$

Построенная величина может служить верхним уровнем разделения для кластеризации фрагментов данных по классам.

Итак, в работе проведен анализ типовых ошибок, возникающих при обработке большого числа экспериментальных данных, и сформулированы методы коррекции ошибок.

Литература

1. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов. – М.: Наука, 1974. – 416 с.

2. Босов А.Д., Орлов Ю.Н., Федоров С.Л. О распределении рядов абсолютных приростов цен на финансовых рынках // Препринты ИПМ им. М.В. Келдыша, № 96, 2014. – С. 1-15.
URL: http://library.keldysh.ru/preprint.asp?id=2014-96

3. Кирина-Лилинская Е.П., Орлов Ю.Н., Федоров С.Л. Метод базисных паттернов в анализе нестационарных временных рядов // Препринты ИПМ им. М.В. Келдыша, № 7, 2016. – 20 с.
URL: http://library.keldysh.ru/preprint.asp?id=2016-7
4. Орлов Ю.Н., Осминин К.П. Методы статистического анализа литературных текстов. – М.: Эдиториал УРСС/Книжный дом «ЛИБРОКОМ», 2012. – 326 с.
5. Борисов Л.А., Орлов Ю.Н., Осминин К.П. Идентификация автора текста по распределению частот буквосочетаний // Препринты ИПМ им. М.В. Келдыша, № 27, 2013. – 27 с.
URL: http://library.keldysh.ru/preprint.asp?id=2013-27
6. С.К. Годунов. Современные аспекты линейной алгебры. – Новосибирск: Научная книга, 1997. – 388 с.
7. Деммель Дж. Вычислительная линейная алгебра. Теория и приложения (пер. с англ.). – М.: Мир, 2001. – 436 с.