Труды научно-технической конференции кластера пензенских предприятий,
обеспечивающих **БЕЗОПАСНОСТЬ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ.**
Том 9, Пенза-2014, с. 29-31 (http://www.pniei.penza.ru/RV-conf/T9/C29)

ЛИНЕЙНОЕ СГЛАЖИВАНИЕ ГИСТОГРАММ БИОМЕТРИЧЕСКИХ ДАННЫХ, ИСКУССТВЕННО УВЕЛИЧИВАЮЩЕЕ ЧИСЛО СТЕПЕНЕЙ СВОБОДЫ ПРИ ОЦЕНИВАНИИ СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Серикова Н.И. (г. Пенза)

Априорная информация о законе распределения значений биометрических данных весьма и весьма значительна. Интеллектуальные средства биометрической аутентификации должны быть способны учитывать априорную информацию в виде знания закона распределения значений биометрических данных и знания об объеме тестовой выборки. Достоверность априорной информации должна проверяться по одному или нескольким критериям согласия. Например, для проверки может использоваться χ^2 критерий Пирсона совместно с критерием Джини [1].

При использовании критериев проверки статистических гипотез важно корректно выбирать число степеней свободы. В этом отношении выборки с малым числом данных оказываются трудно проверяемыми. Так при использовании χ^2 критерия Пирсона рекомендуется выбирать число степеней свободы (число столбиков, анализируемой гистограммы) таким, что бы среднее число опытов в каждом столбике было не менее 5. Если мы имеем выборку из 21 примера биометрических данных, то можем использовать гистограмму из 4 столбиков (21/5 = 4). Так, как мы по этим же данным вычисляем математическое ожидание и среднеквадратическое отклонение, число степеней свободы уменьшается до 2 (4-2=2). Этого явно недостаточно.

Известны способы увеличения тестовой (обучающей) выборки, например, пользуясь бутстрап-преобразованиями можно повторять данные в выборке или удалять данные из выборки. Одна из реализаций гистограммы из 9 столбиков (9 интервалов), построенной на 21 примере приведена на рисунке 1.

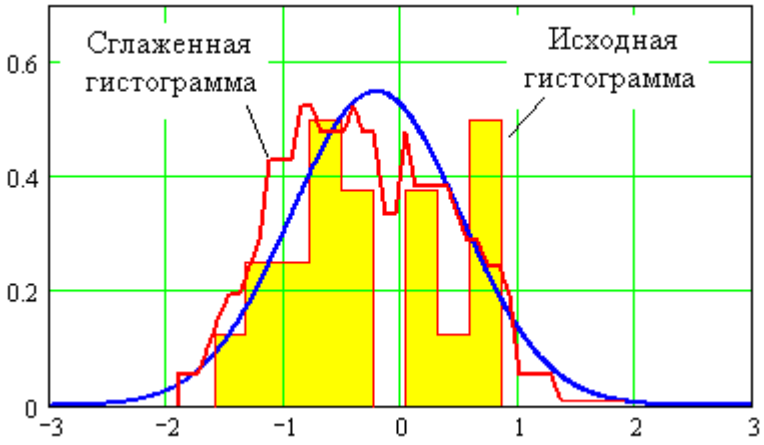

Рис. 1. Сглаживание реальной гистограммы цифровым линейным фильтром с нулевым фазовым сдвигом

Из рисунка 1 видно, что один пример, попавший в крайний правый столбик гистограммы может быть изъят из тестовой (обучающей) выборки так как заранее известно, что закон распределения нормален. Напротив во втором справа столбике гистограммы пример может быть повторен, так как этот столбик

мал. Очевидная ограниченность бутсрап-преобразований состоит в том, что они не способны заполнять пробелы внутри гистограммы. Так гистограмма рисунка 1 в центре имеет пробел, заполнить этот пробел бутсрап-преобразования не позволяют. Аналогичная ситуация отображена на рисунке 2, где так же в центре гистограммы имеется пробел исходных данных.

Рис. 2. Второй пример гистограммы, имеющей в центре пробел, который нельзя устранить бутсрап-преобразованями

Попытаемся устранить этот недостаток бутсрап-преобразований путем сглаживания гистограммы цифровым линейным фильтром с нулевым фазовым сдвигом. Для этой цели необходимо:

1. каждый из 9 столбиков гистограммы (рис. 1 и рис. 2) разбить на 10 интервалов, получив 90 интервалов;
2. присвоить каждому интервалу состояние «0», «1», «2», «3» исходя из условия попадания в эти интервалы 0, 1, 2, 3 биометрических данных;
3. задать окно сглаживающего линейного фильтра, которое должно быть больше ширины столбика исходной гистограммы (на рис. 1 и рис.2 окно выбрано равным 21 микро-интервалу или 2.1 ширины столбика исходной гистограммы);
4. сформировать кодовую последовательность, соответствующую входным данным, путем добавления в начало и конец состояний микро-интервалов по 21 нулевому состоянию (в итоге получим код «0000…01000200……..1000001000000…00000» длинной 90+21+21=132 дискетных значений;
5. осуществить линейное сглаживание кода путем подсчета суммы состояний, находящихся внутри скользящего окна сглаживающего фильтра с последующим делением результата на 21 и разместить результат фильтрации в центре окна сглаживания.

В конечном итоге мы получаем сглаженную гистограмму исходных данных, отображенную на рисунке 1 и рисунке 2. Из этих рисунков видно, что пробелы, существовавшие ранее в исходной гистограмме, исчезают. Теперь мы можем сглаженные данные использовать для формирования более гладкой гистограммы, имеющей не 4, а 30 столбиков (30 интервалов), объединив данные в 4 соседних микро-интервалах, сглаженной гистограммы. Это означает, что после проделанной цифровой обработки выборки из 21 примера исходных данных мы имеем возможность осуществлять проверку гипотез по критерию χ^2 с 30-3=27 степенями свободы. Появляется техническая возможность менять число степеней свободы, оптимизируя статистические вычисления.

Очевиден так же выигрыш от сглаживания гистограммы в случае проверки гипотез по критерию Джини. Этот критерий удобен тем, что вероятность подобия экспериментального распределения данных и гипотетического распределения близка к половине показателя критерия Джини:

$$P_0 \approx 1 - \frac{1}{2}\int_{-\infty}^{+\infty} |p(x) - \tilde{p}(x)| dx \qquad (1),$$

где P_0 – вероятность того, что теоретическое распределение $p(x)$ и экспериментальное распределение $\tilde{p}(x)$ совпадают.

Так из рисунка 1 видно, что ошибка классической гистограммы и нормального распределения составляет 0.55 (в центральном пробеле классической гистограммы). Однако после сглаживания данных предложенным цифровым фильтром в этом же месте ошибка составляет 0.2. Как следствие мощность критерия Джини увеличивается примерно в 0.55/0.2=2.75 раза. То есть в 2.75 раза должна подняться достоверность оценок, принимаемых с использованием критерия Джини.

ЛИТЕРАТУРА:

1. Кобзарь А.И. Прикладная математическая статистика для инженеров и научных работников. М. ФИЗМАТЛИТ, 2006 г., 816 с.
2. Болл Руд и др. Руководство по биометрии. / Болл Руд, Коннел Джонатан Х., Панканти Шарат, Ратха Налини К., Сеньор Эндрю У. // Москва: Техносфера, 2007. -368 с.

Материал поступил 27.04.2014, опубликовано по положительной рецензии доктора технических наук Малыгина А.Ю.