РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ

А В Т О М Е Т Р И Я

№ 4 2000

УДК 681.5.015 : 681.518.25

А. И. Иванов, Л. Н. Сапегин, Е. А. Щигунова

(Пенза)

КОНТРОЛЬ КАЧЕСТВА УЧЕБНОГО МАТЕРИАЛА НЕЙРОСЕТЕЙ И СИСТЕМ БИОМЕТРИЧЕСКОЙ ИДЕНТИФИКАЦИИ ЛИЧНОСТИ

Приведена таблица, увязывающая качество будущего обучения технических систем с усредненными показателями качества учебного материала. Показано, что увеличение модулей значений коэффициентов корреляции между контролируемыми параметрами существенно ухудшает качество настройки нейросети биометрической системы идентификации личности.

Введение. В настоящее время актуальными являются вопросы предварительной оценки потенциально достижимого качества обучения различных технических систем. С данной задачей приходится сталкиваться при обучении систем биометрической аутентификации личности, а также при настройке искусственных нейросетей. Желательно еще до обучения некоторой технической системы знать потенциальные возможности предъявляемого системе учебного материала. Последнее дает возможность, с одной стороны, вовремя прекратить процесс обучения системы, достаточно приблизившись к пределу, а с другой стороны, еще до обучения можно скорректировать учебный материал в сторону его улучшения. Последнее оказывается крайне важным для систем биометрической аутентификации личности по голосу или по динамике воспроизведения подписи. Эксплуатация биометрических систем этого класса показала, что возможны эффекты их переобучения, когда попытка улучшить качество системы за счет увеличения объема обучающей выборки приводит к обратному эффекту. Кроме того, необходимо уметь предсказывать потенциальную стойкость парольной фразы (парольного слова) к попыткам подделки. Все эти задачи в конечном итоге сводятся к измерению качества учебного материала (далее – учебника) и последующему сравнению его с некоторым эталоном.

1. **Вычисление параметров учебника.** Все перечисленные выше задачи могут быть решены в первом приближении, если многообразие влияющих факторов свести к трем переменным и соответственно рассматривать функции качества учебника следующего вида: $\bar{P}_1(p; r; n)$, $P_2(p; r; n)$, где n – число учитываемых параметров; r – усредненное значение модулей коэффициентов корреляции между параметрами; p – усредненное значение вероятностей ошибок первого и второго рода при принятии решения по одному параметру; P_1 и $\bar{P}_2$ – рассматриваемые как функции от p, r, n вероятности ошибок перво-

Таблица 1

p_1/p_2	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
0,1	0,1	0,144	0,184	0,221	0,259	0,304	0,352	0,413	0,5
0,2	0,144	0,2	0,247	0,292	0,335	0,384	0,437	0,5	
0,3	0,184	0,247	0,3	0,349	0,395	0,447	0,5		
0,4	0,221	0,292	0,349	0,4	0,450	0,5			
0,5	0,259	0,335	0,395	0,450	0,5				
0,6	0,304	0,384	0,447	0,5					
0,7	0,352	0,437	0,5						
0,8	0,413	0,5							
0,9	0,5								

го и второго рода при принятии решения по совокупности n параметров. Исследования показали, что в виде функций качества обучения удобно принять равные вероятности ошибок первого и второго рода ($P_1 = P_2 = P(p; r; n)$) технической системы в целом (точка EER-статистики правильных и ошибочных решений). В этом случае конечное решение задачи становится проще, так как зависит от меньшего числа параметров.

С этой же целью упрощения в качестве первой переменной p имеет смысл рассматривать равные вероятности ошибок первого и второго рода ($p_1 = p_2 = p$), получающиеся при учете одного входного среднестатистического параметра. В первом приближении оценка p может быть получена путем усреднения вероятностей ошибок первого и второго рода по каждому входному параметру:

$$p \approx \frac{1}{2n} \sum_{k=1}^{n} (p_{k1} + p_{k2}), \tag{1}$$

где n – полное число входных параметров; k – номер параметра.

Следует иметь в виду, что оценка (1) получается достаточно точной только тогда, когда вероятности ошибок первого рода p_{k1} и вероятности ошибок второго рода p_{k2} для всех k оказываются близкими. Если эти вероятности существенно отличаются, то их необходимо привести к одинаковым значениям. Процедура приведения к точке равновероятных ошибок может быть осуществлена путем аппроксимации данных с помощью табл. 1.

После приведения к одинаковым значениям вероятностей ошибок первого и второго рода данные усредняются по всем параметрам:

$$p \approx \frac{1}{n} \sum_{k=1}^{n} p_k. \tag{2}$$

Второй переменной функции качества учебника $P(p; r; n)$ является усредненное значение модуля коэффициентов корреляции между различными парами параметров:

$$r \approx \frac{1}{(n^2 - n)}\left(\left(\sum_{i=1}^{n}\sum_{j=1}^{n}|r_{ij}|\right) - n\right). \qquad (3)$$

В качестве третьего параметра (n) в функции качества используется число всех учитываемых системой параметров, когда речь идет об оценке качества учебника, предназначенного для обучения системы в целом. Число n можно рассматривать как число входов одного нейрона, когда оценивается качество имеющихся примеров для обучения одного конкретного нейрона.

2. **Идеальные учебники.** Следует иметь в виду, что перечисленные выше статистические характеристики учебного материала легко вычисляются и путем их сравнения можно сравнить качество совершенно разных учебных выборок. Будем считать, что из двух учебников, позволяющих получить по n параметров, лучшим является тот, у которого меньше p при одинаковых r. Очевидно также то, что при одинаковых p лучшим будет учебник с менее коррелированными параметрами. В пределе абсолютно идеальным учебником является учебник с нулевой корреляцией параметров и нулевыми вероятностями ошибок, однако этот случай тривиален ($P(0; 0; n) = 0$). Тривиален также идеальный учебник с нулевыми вероятностями ошибок ($P(0; r; n) = 0$). Для практики интересен только идеальный учебник ($P(p; 0; n) > 0$), с которым имеет смысл сравнивать худшие учебники с ненулевой корреляцией между параметрами.

3. **Синтез таблицы вероятностей.** Функция $P(p; r; n)$ имеет две непрерывные переменные $(p; r)$ и одну дискретную переменную (n). Если непрерывные переменные задать с некоторым шагом, то может быть получена таблица предсказания качества обучения по статистическим параметрам учебного материала. Табл. 2 получается средствами имитационного моделирования. В частности, моделирование идеального учебника ($P(p; 0; n) > 0$) весьма просто может быть осуществлено путем использования стандартной функции RND, позволяющей получить независимые (некоррелированные) случайные векторы из 1, 2, 3, ..., 9, 10 отсчетов. Для того чтобы смоделировать учебник с равными коэффициентами корреляции между параметрами, могут быть использованы рекуррентные процедуры следующего вида:

$$\begin{cases} \widetilde{x}_1 = a_1 x_1, \\ \widetilde{x}_2 = a_2 x_2 + b_2 \widetilde{x}_1, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \widetilde{x}_i = a_i x_i + b_i\left(\sum_{j=1}^{i-1}(-1)^{j+i}\widetilde{x}_j\right), \\ \ldots\ldots\ldots\ldots\ldots\ldots \end{cases} \qquad (4)$$

где $x_1, x_2, \ldots, x_n$ – вектор случайных независимых отсчетов с единичной дисперсией и нулевым математическим ожиданием; a_i и b_i – вычисленные зара-

Таблица 2

Значение коэффициентов a_i и b_i для r от 0,1 до 0,9

r	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
a_1	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
a_2	0,995	0,9798	0,9539	0,9165	0,8660	0,8000	0,7141	0,6000	0,4359
b_2	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
a_3	0,9909	0,9661	0,9282	0,8783	0,8165	0,7416	0,6508	0,5374	0,3839
b_3	0,0909	0,1667	0,2308	0,2857	0,3333	0,3750	0,4118	0,4444	0,4737
a_4	0,9874	0,9562	0,9117	0,8564	0,7906	0,7135	0,6225	0,5114	0,3635
b_4	0,0833	0,1429	0,1875	0,2222	0,2500	0,2727	0,2917	0,3077	0,3214
a_5	0,9845	0,9487	0,9003	0,8421	0,7746	0,6969	0,6064	0,4970	0,3526
b_5	0,0769	0,125	0,1579	0,1818	0,2000	0,2143	0,2258	0,2353	0,2432
a_6	0,9820	0,9428	0,8919	0,8321	0,7637	0,6860	0,5960	0,4879	0,3458
b_6	0,0714	0,1111	0,1364	0,1538	0,1667	0,1765	0,1842	0,1905	0,1957
a_7	0,9798	0,9381	0,8854	0,8246	0,7559	0,6782	0,5887	0,4817	0,3411
b_7	0,0667	0,1000	0,1200	0,1333	0,1429	0,1500	0,1555	0,1600	0,1636
a_8	0,9779	0,9342	0,8803	0,8189	0,7500	0,6724	0,5834	0,4771	0,3377
b_8	0,0625	0,0909	0,1071	0,1176	0,1250	0,1304	0,1346	0,1379	0,1406
a_9	0,9762	0,9309	0,8762	0,8144	0,7453	0,6679	0,5793	0,4735	0,3351
b_9	0,0588	0,0833	0,0967	0,1052	0,1111	0,1154	0,1158	0,1212	0,1233

нее весовые коэффициенты. Весовые коэффициенты a_i и b_i могут быть найдены различными способами, в том числе с помощью процедур, приведенных в [1]. Примеры значений весовых коэффициентов для равномерного закона распределения случайных отсчетов приведены в табл. 2.

Линейные преобразования (4) позволяют получить вектор случайных значений с корреляционной матрицей следующего вида:

$$\begin{bmatrix} 1 & -r & r & \dots & -r \\ -r & 1 & -r & \dots & r \\ r & -r & 1 & \dots & -r \\ \dots & \dots & \dots & \dots & \dots \\ -r & r & -r & \dots & 1 \end{bmatrix}.$$

Очевидно, что с помощью линейных преобразований (4) можно имитировать учебники различного качества. В частности, для биометрических систем идентификации личности могут быть смоделированы учебники «свой», «чужой». Далее возможно обучение на моделях учебников реальной технической системы и статистическая проверка ее качества. Формально реальную техническую систему биометрической идентификации личности можно рассматривать как некий вариант двухслойной нейросети. Имитационное

моделирование позволяет получить зависимость вероятностей ошибок биометрических систем идентификации личности от качества учебного материала.

По результатам имитационного моделирования построена табл. 3 предсказания вероятности ожидаемых выходных ошибок, которые приводятся ниже для разных значений p, r, n. Таблица получена на выборке из 500 векторов, параметры которых имеют закон распределения значений, близкий к нормальному. Число n в таблице можно интерпретировать как число входов нейрона нижнего слоя. Тогда данные таблицы можно рассматривать как ожидаемую вероятность ошибок первого и второго рода на выходе нейрона нижнего слоя.

4. **Вырождение учебников.** Верхняя часть табл. 3 соответствует наиболее выгодному идеальному учебнику без корреляции между параметрами вектора, нижняя часть – полностью вырожденному учебнику с единичной корреляцией между параметрами вектора (учебника). Из приведенной таблицы видно, что по мере увеличения корреляции между параметрами учебного материала качество решения нижнего нейрона падает. Случай $r = 1{,}0$ тривиален и не дает выигрыша при любом числе входов нейрона в сравнении с принятием решения на основе анализа всего одного параметра: $P(p; 1{,}0; n) = p$.

Кроме того, из табл. 3 видно, как вырождается учебник при стремлении входной вероятности p к значению 0,5. В точке 0,5 учебник также вырождается независимо от значения коэффициента корреляции. Предельный случай $P(0{,}5; r; n) = 0{,}5$ бесполезен для практики и в принципе не позволяет улучшать ситуацию за счет роста числа входов у нейронов нижнего слоя.

5. **Использование таблицы для прогнозов.** Полученная табл. 3 может быть использована для разных целей. В частности, с помощью этой таблицы осуществляется прогноз качества работы биометрических систем идентификации личности по динамике подписи, по голосу и клавиатурному почерку. Проблема состоит в том, что параметры этих систем известны только некоторому среднестатистическому пользователю [2]. Обычно эти параметры получаются путем усреднения данных статистических испытаний по большой группе пользователей. При этом статистические параметры системы идентификации для конкретного пользователя могут быть существенно лучше или хуже в сравнении со среднестатистическими. Предложенная таблица позволяет получить коэффициент пересчета параметров системы применительно к данным конкретного пользователя. В первом приближении этот пересчет может быть осуществлен линейной интерполяцией. Тогда коэффициент пересчета находится как отношение соответствующих данных табл. 3:

$$K = \frac{P(p; r; n)}{P(\tilde{p}; \tilde{r}; \tilde{n})}, \qquad (5)$$

где p, r, n – заранее известные параметры среднестатистического пользователя; p, r, n – параметры конкретного пользователя, для которого строится прогноз.

Следует подчеркнуть, что при вычислении коэффициента пересчета (5) не может быть произвольным выбор значения числа контролируемых параметров n, $\tilde{n}$ в силу существенной нелинейности функции $P(p; r; n)$. Значение параметра $\tilde{n}$ должно быть заранее известно, а значение параметра n должно

Таблица 3

Вход	Вероятность $P(p; 0; n)$ для некоррелированных $r = 0{,}0$ входных данных (идеальный учебник)								
$n = 1$	2	3	4	5	6	7	8	9	10
$p = 0{,}1$	0,039	0,012	0,004	0,0026	0,0001				
$p = 0{,}2$	0,101	0,068	0,049	0,029	0,024	0,013	0,008	0,006	0,004
$p = 0{,}3$	0,227	0,177	0,132	0,114	0,106	0,096	0,067	0,052	0,049
$p = 0{,}4$	0,352	0,341	0,315	0,294	0,268	0,256	0,242	0,224	0,217
$p = 0{,}5$	**0,5**	**0,5**	**0,5**	**0,5**	**0,5**	**0,5**	**0,5**	**0,5**	**0,5**
Вход	Вероятность $P(p; 0{,}2; n)$ для коррелированных $r = 0{,}2$ входных данных (идеальный учебник)								
$n = 1$	2	3	4	5	6	7	8	9	10
$p = 0{,}1$	0,056	0,038	0,021	0,020	0,014	0,011	0,007	0,006	0,005
$p = 0{,}2$	0,124	0,123	0,085	0,081	0,073	0,072	0,070	0,057	0,053
$p = 0{,}3$	0,268	0,232	0,206	0,192	0,186	0,185	0,181	0,159	0,157
$p = 0{,}4$	0,361	0,356	0,347	0,346	0,343	0,338	0,323	0,317	0,314
$p = 0{,}5$	**0,5**	**0,5**	**0,5**	**0,5**	**0,5**	**0,5**	**0,5**	**0,5**	**0,5**
Вход	Вероятность $P(p; 0{,}4; n)$ для коррелированных $r = 0{,}4$ входных данных (идеальный учебник)								
$n = 1$	2	3	4	5	6	7	8	9	10
$p = 0{,}1$	0,067	0,056	0,042	0,037	0,034	0,033	0,030	0,029	0,027
$p = 0{,}2$	0,155	0,146	0,138	0,127	0,126	0,123	0,120	0,104	0,098
$p = 0{,}3$	0,288	0,251	0,229	0,226	0,225	0,214	0,211	0,206	0,199
$p = 0{,}4$	0,382	0,375	0,371	0,370	0,359	0,352	0,352	0,346	0,340
$p = 0{,}5$	**0,5**	**0,5**	**0,5**	**0,5**	**0,5**	**0,5**	**0,5**	**0,5**	**0,5**
Вход	Вероятность $P(p; 0{,}6; n)$ для коррелированных $r = 0{,}6$ входных данных (идеальный учебник)								
$n = 1$	2	3	4	5	6	7	8	9	10
$p = 0{,}1$	0,080	0,076	0,065	0,063	0,056	0,052	0,051	0,050	0,050
$p = 0{,}2$	0,178	0,166	0,164	0,156	0,153	0,149	0,143	0,141	0,135
$p = 0{,}3$	0,282	0,278	0,276	0,271	0,269	0,262	0,260	0,255	0,248
$p = 0{,}4$	0,394	0,393	0,389	0,388	0,375	0,372	0,369	0,367	0,362
$p = 0{,}5$	**0,5**	**0,5**	**0,5**	**0,5**	**0,5**	**0,5**	**0,5**	**0,5**	**0,5**
Вход	Вероятность $P(p; 0{,}8; n)$ для коррелированных $r = 0{,}8$ входных данных (идеальный учебник)								
$n = 1$	2	3	4	5	6	7	8	9	10
$p = 0{,}1$	0,095	0,091	0,085	0,082	0,081	0,077	0,076	0,073	0,071
$p = 0{,}2$	0,195	0,189	0,182	0,180	0,179	0,173	0,169	0,167	0,162
$p = 0{,}3$	0,286	0,285	0,281	0,279	0,278	0,276	0,273	0,268	0,267
$p = 0{,}4$	0,396	0,395	0,394	0,391	0,389	0,385	0,384	0,374	0,371
$p = 0{,}5$	**0,5**	**0,5**	**0,5**	**0,5**	**0,5**	**0,5**	**0,5**	**0,5**	**0,5**
Вход	Вероятность $P(p; 1{,}0; n)$ для полностью коррелированных $r = 1{,}0$ входных данных (полностью вырожденный учебник)								
$n = 1$	**2**	**3**	**4**	**5**	**...**	**10**	**...**	**100**	**...**
$p = 0{,}1$	**0,1**	**0,1**	**0,1**	**0,1**	**...**	**0,1**	**...**	**0,1**	**...**
$p = ...$	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**	**...**
$p = 0{,}4$	**0,4**	**0,4**	**0,4**	**0,4**	**...**	**0,4**	**...**	**0,4**	**...**
$p = 0{,}5$	**0,5**	**0,5**	**0,5**	**0,5**	**...**	**0,5**	**...**	**0,5**	**...**

предоставляться соответствующим программным обеспечением. Только в этом случае появляется возможность предсказывать качество работы биометрической системы с данными конкретных людей на основе измерения усредненных значений p и r. Примеры использования табл. 3 для подобных расчетов даны в приложении.

6. **«Просеивание» и сортировка обучающей выборки.** При обучении нейросетей и биометрических систем может возникнуть эффект переобучения, когда добавление в обучающую выборку примеров не улучшает, а наоборот, ухудшает качество идентификации (распознавания образов). Подобпы новых примеров необходимо проверить качество добавляемой группы, вычислив p, r и обратившись к таблице. Если качество плохое, то добавлять новые примеры не следует.

Еще одной причиной переобучения может быть присутствие в обучающей выборке сильно коррелированных примеров. Если в обучающую выборку добавлять хорошие, но уже имеющиеся примеры, то значение r, вычисленное по формуле (3), будет увеличиваться. Легко показать, что добавление большой группы схожих между собой примеров даже с низкой внутренней корреляцией данных приводит к ухудшению среднего коэффициента корреляции (3), вычисленного по всей группе.

Таким образом, при формировании обучающей выборки имеет смысл отбрасывать примеры с сильной внутренней и внешней корреляцией с уже имеющимися примерами. Принятие компромиссного решения о целесообразности расширения обучающей выборки должно приниматься с учетом данных приведенной таблицы. Подобная тактика в ряде случаев позволяет многократно улучшить качество обучающего материала. В итоге таблица дает возможность не только прогнозировать результат будущего обучения для конкретного пользователя, но и путем «просеивания» и сортировки примеров существенно улучшать качество учебного материала, тем самым упростив и улучшив процедуру обучения.

Заключение. Табл. 3 построена для алгоритма обучения конкретной системы биометрической аутентификации личности по динамике рукописного почерка «Рубеж». Естественно, что для иного алгоритма обучения потребуется синтез другой, похожей таблицы, на что потребуется всего несколько часов машинного времени. Преимущества же подобных таблиц очевидны: они дают возможность предсказывать результат обучения конкретной технической системы еще до проведения самого обучения. Как показала практика использования табл. 3, для прогнозов с приемлемой погрешностью вполне достаточно знания всего трех параметров обучающей выборки: p, r, n.

ПРИЛОЖЕНИЕ

Задача 1. Рукописный почерк человека-1 при контроле одного параметра дает равные вероятности ошибки первого и второго рода $p = 0{,}4$, если его сравнивать со среднестатистическим почерком. Для 100 разных почерков их сравнение со среднестатистическим почерком дает $p = 0{,}3$. Корреляция меж-

ду контролируемыми параметрами слова «Пенза», воспроизведенного человеком-1, составляет $r = 0{,}4$. Среднее значение корреляции между параметрами, характерное для среднестатистического человека, имеет значение $r = 0{,}6$. Требуется предсказать качество работы биометрической системы для человека-1 по отношению к среднему качеству ее работы, заявленному в рекламе.

Р е ш е н и е. Предположим, что нейросеть биометрической системы имеет нейроны с 10 входами и ее структура не меняется. Тогда по табл. 3 для среднего человека получим $P(0{,}3;\ 0{,}6;\ 1{,}0) = 0{,}248$. Для человека-1 найдем $P(0{,}4;\ 0{,}4;\ 1{,}0) = 0{,}34$. Для человека-1 биометрическая система будет давать вероятности ошибок примерно в 0,34/0,248 = 1,37 раза больше, чем вероятности, заявленные в рекламе.

Задача 2. Биометрическая система обеспечивает заявленную вероятность ошибок при контроле 50 параметров для среднестатистического пользователя. При рукописном воспроизведении слова «Пенза» человеком-1 система контролирует 80 параметров. Требуется оценить, как должен сказаться рост числа контролируемых параметров на качестве работы системы при условии совпадения p и r с условиями задачи 1.

Р е ш е н и е. Для среднего человека нейросеть будет иметь 5 нейронов нижнего слоя, использующих различные входные данные, и один нейрон с 5 входами во втором слое. По табл. 3 найдем $P(0{,}248;\ 0{,}6;\ 5) = (0{,}156 + 0{,}271)/2 = 0{,}21$. Аналогично для человека-1 второй слой будет состоять из 1 нейрона с 8 входами, для которого $P(0{,}34;\ 0{,}4;\ 8) = (0{,}211 + 0{,}352)/2 = 0{,}281$. За счет увеличения числа учитываемых параметров произошло некоторое улучшение качества решения, но оно все же будет хуже в 2,81/2,1 = 1,33 раза, чем заявлено в рекламе.

СПИСОК ЛИТЕРАТУРЫ

1. **Шалыгин А. С., Палагин Ю. И.** Прикладные методы статистического моделирования. Л.: Машиностроение, 1986.
2. **Иванов А. И.** Оценка систем биометрической аутентификации // Защита информации. Конфидент, 1998. № 2. С. 77.

Пензенский научно-исследовательский электротехнический институт,
E-mail: crystall@tl.ru

Поступила в редакцию
3 июня 1998 г.