УДК 004.934.2

Е.Ю. Костюченко

Обработка естественной информации на основе аппарата нейронных сетей

Анализ естественной информации является актуальной проблемой и представляет интерес как с точки зрения извлечения содержания этой информации, так и с точки зрения идентификации субъекта, которому эта информация принадлежит. В статье рассмотрен нейросетевой подход к анализу естественной информации и описаны подход к определению информативности параметров естественной информации и критерий для определения информативности с учетом применения аппарата нейронных сетей.
Ключевые слова: нейронные сети, естественная информация, информативность, критерий оценки правильности работы.

Автоматизация обработки естественной информации является актуальной проблемой, стоящей перед современным человечеством. Под естественной информацией в данном случае понимается информация, получаемая при осуществлении человеком некоторых операций, например: написание теста «от руки» (рукописный почерк), набор текста на клавиатуре (клавиатурный почерк), простановка подписи, разговор и т.д. Можно заметить, что естественный сигнал содержит как прямую непосредственную информацию, например в случае набора текста на клавиатуре – некоторый набранный текст, так и информацию скрытую, например информацию о личности человека, набирающего этот текст. Можно видеть, что естественные сигналы соответствуют так называемым динамическим биометрическим характеристикам [1], и, как следствие, содержащаяся в них естественная информация может использоваться как для извлечения прямой информации, так и для идентификации личности.

Анализ естественной информации может быть представлен в виде последовательного выполнения следующих шагов:

1. Выбор параметров, по которым будут производиться анализ и извлечение этих параметров из естественного сигнала.
2. Выбор метода анализа параметров сигнала и применение этого метода к извлеченным параметрам.
3. Анализ результата и либо переход к следующему уровню анализа (результат принимается за новый сигнал и повторяется выполнение шагов 1–3), либо принятие решения.

Методы анализа параметров могут быть разделены на следующие группы:

1. Статистический анализ [2–3].
2. Спектральный анализ с применением преобразования Фурье, вейвлет-преобразований и др. [4].
3. Анализ на основе аппарата скрытых марковских моделей [5].
4. Применение генетических алгоритмов [6].
5. Применение аппарата искусственных нейронных сетей [7].

При этом возможно совместное использование этих методов как на одном, так и на различных уровнях анализа.

Описание количества входов нейронной сети и типа используемых параметров для различных задач анализа естественной информации, построенное на основе [8], представлено в таблице.

Видно, что при анализе естественно-речевой информации возникает проблема выбора и выделения параметров для анализа нейронной сетью, поскольку при анализе непосредственно параметров осциллограммы речевого сигнала обучение нейронной сети возможно, однако результат анализа работы на сигналах, не входивших в обучающую выборку, является неудовлетворительным.

Для оценки наличия в наборе данных информации может быть использовано понятие информативности. Существует множество подходов к оценке информативности параметров, описание их содержится в [9–10]. Эти подходы не учитывают специфику обработки параметра нейронной сетью, которая может на менее информативном параметре с точки зрения этих оценок показать лучшие результаты, чем на более информативном.

В данной работе предлагается использовать следующий подход:

1) выбирается параметр для оценки информативности;

2) выбираются параметры нейронной сети для оценки информативности;

3) из обучающей выборки выделяются выбранные параметры и подаются на входы нейронной сети для обучения;

4) после повторения процедуры обучения определенное количество раз в качестве оценки информативности выбирается функция (минимальная или средняя) ошибки на предъявленном обучающем наборе.

Применение нейронных сетей для анализа естественной информации

Вид естественной информации	Клавиатурный почерк, фиксированная фраза	Клавиатурный почерк, произвольная фраза	Подпись	Речь
Набор входных параметров	100–400	До 1500, но возможно ограничение 1000	Зависит от параметров, для гармоник – до 300	1000 и более, ограничено предельным размером входного слоя нейронной сети
Применение нейронных сетей	Двухслойный персептрон	Двухслойный персептрон	Двухслойный персептрон	Прямое решение простым применением двухслойного персептрона невозможно
Комментарии	Интервалы между нажатиями, время залипания, количество зависит от длины фразы	Зависит от отслеживаемых сочетаний (биграмм, триграмм и т.д.), статистические параметры	Параметры – результат анализа Фурье-зависимостей изменения координат и давления	Выбор параметров не определен. Зависит от окна анализа и частоты дискретизации

Определенная таким образом информативность учитывает особенности работы нейронной сети, поскольку формируется на основе результатов ее работы.

Существуют следующие критерии оценки правильности работы обученной нейронной сети на основе ошибок первого и второго рода, которые будут применяться на шаге 4 оценки информативности параметров: EER (Equal error rate) – уровень равной ошибки, CER (Classification error rate) – уровень ошибки классификации, MTER (Minimum total error rate) – уровень минимальной суммарной ошибки [11]. Главный недостаток этих характеристик – нет учета приоритета ошибок первого или второго рода. Критерий Байеса – критерий минимального среднего риска – учитывает приоритет ошибок, но не учитывает важность меньших ошибок по сравнению с большими. На практике возникает необходимость учета нелинейной зависимости критерия от ошибки.

Для учета выявленных недостатков предлагается разработать собственный критерий. В качестве базового обобщенного критерия правильности работы нейронной сети предлагается выбрать следующий:

$$M(m_1, m_2) = ((a \times m_1)^s + (b \times m_2)^s)^{1/s}.$$

Здесь $m_1 \in [0,1]$ – значение ошибки первого рода; $m2 \in [0,1]$ – значение ошибки второго рода; $a \in [0,\infty]$ – коэффициент, учитывающий значимость ошибок первого рода; $b \in [0,\infty]$ – коэффициент, учитывающий значимость ошибок второго рода; $s \in [0,\infty]$ – коэффициент, учитывающий значимость больших ошибок по сравнению с меньшими.

С учетом границ реальных результатов работы нейронной сети (значение $M = 1$ заведомо достигается) значение критерия можно переписать как:

$$M(m_1, m_2) = \begin{cases} (m_1^s + (k \times m_2)^s)^{1/s}, M(m_1, m_2) < 1, \\ 1, M(m_1, m_2) \geq 1, \end{cases}$$

для случая, когда более важной является ошибка второго рода, или

$$M(m_1, m_2) = \begin{cases} ((k \times m_1)^s + m_2{}^s)^{1/s}, M(m_1, m_2) < 1, \\ 1, M(m_1, m_2) \geq 1, \end{cases}$$

для случая, когда более важной является ошибка первого рода.

Здесь $k \in [0, \infty]$ – коэффициент, учитывающий взаимную значимость ошибок первого рода. При таком определении значения критерия $M(m_1, m_2) \in [0, 1]$. Определенный таким образом критерий, с одной стороны, при определенных значениях коэффициентов k и s может быть преобразован к критериям Байеса, уровня минимальной суммарной ошибки, уровня ошибки классификатора, однако лишен их недостатков за счет введения этих коэффициентов.

Таким образом, в данной работе рассмотрены подходы к обработке естественной информации и выявлены проблемы, возникающие при анализе естественно-речевой информации с помощью аппарата искусственных нейронных сетей. Выявлена неприменимость традиционных подходов к оценке информативности параметров с точки зрения анализа с помощью нейронных сетей, введен подход к оценке информативности на основе ошибок первого и второго рода в результатах работы нейронной сети, введен новый критерий учета ошибок первого и второго рода с учетом их значимости и особенностей обработки речевого сигнала с применением нейронных сетей.

Литература

1. Иванов А.И. Биометрическая идентификация личности по динамике подсознательных движений. – Пенза: Изд-во Пенз. гос. ун-та, 2000. – 188 с.

2. Лапач С.Н. Статистика в науке и бизнесе / С.Н. Лапач, А.В. Чубенко, П.Н. Бабич. – Киев: Морион, 2002. – 640 с.

3. Айвазян С.А. Прикладная статистика. Т. 1. Теория вероятностей и прикладная статистика / С.А. Айвазян, В.С. Мхитарян. – М.: Юнити, 2001. – 656 с.

4. Дженкинс Г. Спектральный анализ и его приложения / Г. Дженкинс, Д. Ваттс. – М.: Мир, 1971. – 317 с.

5. Иконин С.Ю. Система автоматического распознавания речи SPIRIT ASR Egnine / С.Ю. Иконин, С.Ю. Сарана // Цифровая обработка сигналов. – 2003. – №3 [Электронный ресурс]. – Режим доступа: http://www.spirit.ru/articles/asr.html

6. Сапунов Г.В. Система автоматического распознавания речевых команд для параллельных архитектур: автореф. дис. ... канд. тех. наук. – М., 2006. – 28 с.

7. Хайкин С. Нейронные сети. – М.: Вильямс, 2006. – 1104 с.

8. Костюченко Е.Ю. Идентификация по биометрическим параметрам при использовании аппарата нейронных сетей / Е.Ю. Костюченко, Р.В. Мещеряков // Нейрокомпьютеры: разработка, применение. – М.: Радиотехника, 2007. – № 7. – С. 39–50

9. Загоруйко Н.Г. Методы распознавания и их применение. – М.: Сов. радио, 1972.

10. Костюченко Е.Ю. Методы определения информативности параметров при распознавании речи // Научная сессия ТУСУР – 2006: Матер. докл. Всерос. науч.-техн. конф. студентов, аспирантов и молодых ученых, Томск, 4–7 мая 2006 г. – Томск: В-Спектр, 2006. – Ч. 3. – С. 106–107.

11. Нгуен М.Т. Разработка алгоритмов построения оценок достоверности для систем распознавания речи: автореф. дис. ... канд. тех. наук. – М., 2008. – 22 с.

Костюченко Евгений Юрьевич
Ассистент каф. КИБЭВС, ТУСУР, т. 8-923-405-55-59
Эл. адрес: key@keva.tusur.ru

E.Y. Kostyuchenko

Processing of the natural information on the basis of neural networks

The summary: the analysis of the natural information is an actual problem and is interesting from the both point of view: of extraction of the maintenance of this information, and identification of the subject to which this information belongs. In article it is considered neural network approach to the analysis of the natural information and the approach to definition selfdescriptiveness of parameters of the natural information and criterion for definition selfdescriptiveness of taking into account application of the device of neural networks is described.

Keywords: the neural networks, the natural information, selfdescriptiveness, criterion of work correctness estimation.