УДК 621.3.049.77–048.24:537.2

РАСПОЗНАВАНИЕ ЭМОЦИЙ С ИСПОЛЬЗОВАНИЕМ КЕПСТРАЛЬНОГО ПРЕДСТАВЛЕНИЯ РЕЧЕВОГО СИГНАЛА И МЕТОДА ОПОРНЫХ ВЕКТОРОВ

Краснопрошин Д.В., магистрант гр.255741

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Вашкевич М.И. – доктор. техн. наук

Аннотация. Экспериментально исследуется возможность применения метода опорных векторов (МОВ) для классификации эмоций в человеческой речи. Представлен вариант реализации классификатора (на основе МОВ) с использованием линейной ядерной функции. Показано, что полученная модель позволяет определять эмоции с точностью до 85%.

Ключевые слова. М*етод опорных векторов, МОВ, распознавание, цифровая обработка сигналов, машинное обучение*.

Введение

Одной из актуальных прикладных задач, связанных с созданием эффективного человеко-машинного взаимодействия является построения интерфейса, приближенного к естественным условиям. Для решения данной задачи требуется, чтобы компьютер был способен воспринимать текущую ситуацию и реагировать в соответствии с этим восприятием. Одним из условий для адекватного восприятия является понимание эмоционального состояния пользователя.

Среди основных способов выражения человеческих эмоций важная роль отводится его речи. За последние годы было проведено большое количество исследований по распознаванию (классификации) эмоций на основе речи [1-2].

Существуют различные варианты решения данной проблемы. В частности, можно отметить подходы, основанные на использовании нейронных сетей, байесовского классификатора в сочетании с методом максимального правдоподобия, скрытых марковские модели и т. д. [1]

В данной работе предлагается подход для классификации человеческих эмоций с использованием метода опорных векторов (МОВ).

Набор данных

При проведении исследования в качестве исходного набора данных использовался Toronto emotional speech set (TESS) [3].

Набор данных TESS представляет собой речевые аудиофайлы в формате .wav (16 бит, 48 кГц). Общее количество файлов: 2800. Озвучка была выполнена двумя профессиональными актрисами (в возрасте 26 и 64 лет), озвучивающих два лексически совпадающих высказывания с нейтральным североамериканским акцентом. Для обеих актрис английский язык является родным. Обе имеют университетское и музыкальное образование. Речевые эмоции включают выражения спокойствия, счастья, грусти, гнева, страха, удивления и отвращения. Каждое выражение производится на двух уровнях эмоциональной интенсивности (нормальный, сильный) с дополнительным нейтральным выражением.

Анализ речевого сигнала

Для построения системы по распознаванию эмоций в речи требуется провести предобработку исходных данных. Основной задачей предобработки является удаление шума, повышение высоких частот сигнала и получение плоского частотного спектра сигналов и частотных характеристик.

Еще одним важным шагом является выделение и выбор признаков. Обычно выделяется тональность и ее изменение, скорость произношения и другие спектральные характеристики.

В рамках данной работы для извлечения признаков использовалась техника на основе расчета Мел-частотных кепстральных коэффициентов. Данная техника подразумевает следующие шаги:

1) *Предыскажение***:** увеличивает величину энергии на более высокой частоте. В случаях, когда рассматривается частотная область звукового сигнала для звонких сегментов, таких как гласные, видно, что энергия на более высокой частоте намного меньше, чем энергия на более низких частотах. Повышение энергии на более высоких частотах повысить точность и производительность модели;

2) *Кратковременное преобразование Фурье***:** это особый вид преобразования Фурье, благодаря которому можно узнать, как частоты в сигнале меняются во времени. Он работает, разрезая ваш сигнал на множество небольших сегментов и выполняя преобразование Фурье каждого из них. В результате обычно получается каскадный график, показывающий зависимость частоты от времени;

3) *Расчет набора из M-фильтров***:** используется для моделирования свойств человеческого слуха на этапе выделения признаков, что позволяет улучшить производительность модели. Поэтому мы будем использовать мел-шкалу, чтобы сопоставить фактическую частоту с частотой, которую воспринимают люди. Формула отображения приведена ниже:

Отметим, что человеческий слух менее чувствителен к изменению энергии звукового сигнала при более высокой энергии по сравнению с более низкой энергией. Логарифмическая функция также имеет аналогичное свойство, при низком значении входного x градиент логарифмической функции будет выше, но при высоком значении входного градиента значение меньше. Поэтому мы применяем log к выходу Mel-фильтра, чтобы имитировать человеческий слух.

4) *Дискретное косинусное преобразование (ДКП)*: Проблема с полученной спектрограммой заключается в том, что коэффициенты банка фильтров сильно коррелированы. Поэтому нам нужно декоррелировать эти коэффициенты. Для этого применяется ДКП.

В результате мы получим набор чисел, являющихся мел-частотными кепстральными коэффициентами (МЧКК).

Классификация

Метод опорных векторов выполняет классификацию путем построения N-мерных гиперплоскостей, которые оптимально разделяют данные на отдельные категории. Классификация достигается путем построения в пространстве входных данных линейной (или нелинейной) разделяющей поверхности. Идея данного подхода заключается в преобразовании (с помощью функции ядра) исходного набора данных в многомерное пространство признаков. И уже в новом пространстве признаков добиться оптимальной в определенном смысле классификации.

В качестве ядра используется любая симметричная, положительно полуопределенная матрица K, которая составлена из скалярных произведений пар векторов x_i и x_j, где $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$, характеризующих меру их близости. А ϕ является функцией, формирующее ядро. В частности, примерами таких функций являются:

- **линейное ядро:**

$$K(x_i, x_j) = x_i^T x_j,$$

что соответствует классификатору на опорных векторах в исходном пространстве

- **полиномиальное ядро со степенью p:**

$$K(x_i, x_j) = (1 + x_i^T x_j,)^p$$

- **гауссово ядро (радиальная базисная функция):**

$$K(x_i, x_j) = \exp\left(\gamma \|x_i - x_j\|^2\right)$$

В качестве ядра для модели на основе МОВ была выбрана линейная функция. Значение параметра C (cost) (допустимый штраф за нарушение границы зазора) было равно единице.

Построение классификатора на опорных векторах с использованием перечисленных выше ядер можно, в частности, осуществить с помощью библиотеки sklearn, написанной на языке Python.

Описание эксперимента

Исходный набор данных был разбит на тренировочную (70%) и тестовую (30%) выборки.

Для оценки качества работы модели было вычислено среднее арифметическое (невзвешенное) полноты рассчитанной для каждого распознанного класса.

Полнота представляет собой отношение $\text{ИП}/(\text{ИП} + \text{ЛО})$, где ИП — количество истинных положительных результатов, а ЛО — количество ложноотрицательных результатов. Также под полнотой понимается интуитивно способность классификатора находить все положительные образцы.

Значение полноты лежит в диапазоне от 0 до 1.

В результате построения и обучения модели был получен классификатор, точность предсказаний которого при использовании тестового набора данных и вышеуказанной метрики качества достигала 85%.

Далее будет представлена мультиклассовая матрица спутывания (англ. Multiclass Confusion Matrix) представляющая собой таблицу или диаграмму, показывающая точность прогнозирования классификатора в отношении двух и более классов. Ячейки таблицы заполняются количеством прогнозов классификатора. Правильные прогнозы идут по главной диагонали от верхнего левого угла в нижний правый.

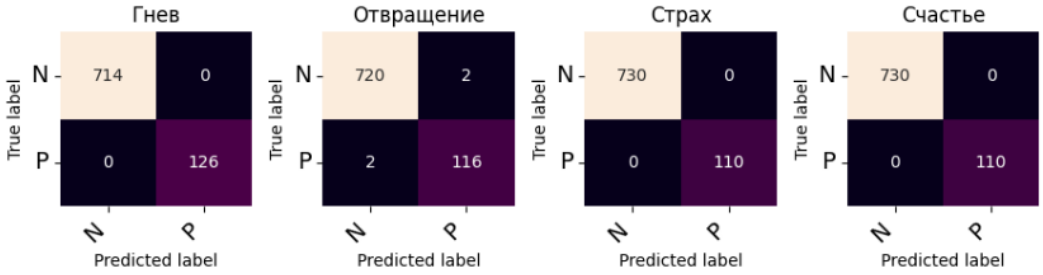

Рисунок 1 – Мультиклассовая матрица спутывания (часть 1)

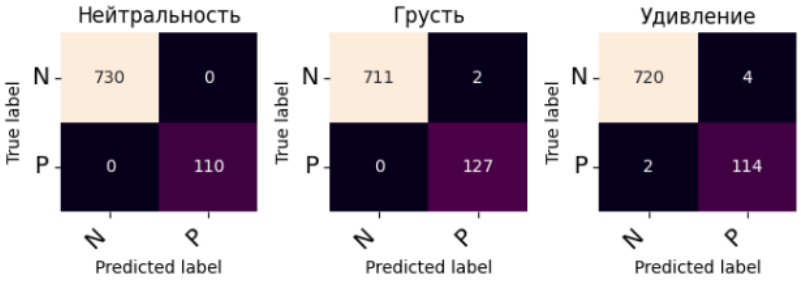

Рисунок 2 – Мультиклассовая матрица спутывания (часть 2)

Выводы

Анализ полученных результатов показал, что метод опорных векторов достаточно успешно справляется с задачей распознавания речевых эмоций. Тем не менее, для более комплексных входных данных (большее количество актеров разного пола и возрастов) этого метода может оказаться недостаточно. В связи с этим для решения обозначенной задачи, возможно, следует попробовать более сложные модели. Таковыми, например, являются скрытые Марковские модели, сверточные нейронные сети и долговременная память (особая разновидность архитектуры рекуррентных нейронных сетей, способная к обучению долговременным зависимостям), поскольку они лучше отражают временную динамику, включенную в речь человека.

Список использованных источников:

1. L. Chen, X. Mao, Y. Xue, and L. Cheng, "Speech emotion recognition: Features and classification models," Digital Signal Processing. Vol. 22, No. 6, pp. 1154-1160, 2012.

2. D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," Speech Communication. Vol. 48, No. 9, pp. 1162-1181, 2006.

3. Toronto emotional speech set (TESS) / Kaggle – Режим доступа: https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess. – Дата доступа: 03.03.2023.

UDC 621.3.049.77–048.24:537.

RECOGNITION OF EMOTIONS USING THE CEPSTRAL REPRESENTATION OF A SPEECH SIGNAL AND THE SUPPORT VECTOR MACHINE

Krasnoproshin D.V.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Vashkevich M.I. – PhD

Annotation. The possibility of using the support vector machine (SVM) for the classification of emotions in human speech is experimentally studied. A variant of the implementation of the classifier (based on SVM) using a linear kernel function is presented. It is shown that the resulting model allows you to classify emotions with an accuracy of up to 85%.

Keywords. support vector machine, SVM, *recognition, digital signal processing, machine learning*.