Выполнимость гипотезы простоты выборки для комбинированных признаковых описаний в задаче классификации временных рядов

Сергей Иванычев.

sergeyivanychev@gmail.com

Contents

1 Введение

В работе рассматривается задача классификации временных рядов в задаче классификации действий человека по временным рядам, порождаемым датчиками носимых устройств. Классификация временных рядов является частным случаем классификации объектов сложной структуры. Из-за того, что подобные задачи возникают во многих областях, например, в обработке сигналов, биологии, финансах, метеорологии, существует довольно много техник ее решения.

В нашей работе нас интересует решение задачи классификации временных рядов путем построения промежуточного признакового пространства. Этот метод применим не только к задаче классификации классификации данных с носимых устройств, так как к объектам сложной структуры можно свести соответствующие ряды из других задач. В общем случае подход с промежуточным признаковым пространством разделим на два этапа.

- На первом этапе для сегментов временных рядов, которые выступают в роли объектов (которые, вообще говоря, могут быть различной длины и даже частоты дискретизации) вычисляются некоторые статистики или добываются некоторые экспертные оценки. В результате на каждый объект мы имеем некоторый набор показателей одной природы и из одного пространства.
- Над вторичным пространством этих показателей (то есть преобразованными объектами) работает некоторый алгоритм классификации (например ...), который обучается на ”вторичной” выборке.

Эти этапы зависимы, так как классификатор, используемый во втором этапе может потребовать от обучающей выборки выполнимость некоторых гипотез и, в частности, гипотезы простоты выборки, что может быть обепечено только корректным первым этапом. Выполнимость гипотезы простоты выборки, находящейся в промежуточном пространстве необходима для корректной работы алгоритмов классификации.

В нашей работе мы рассматриваем при каких условиях отображение объектов сложной структуры порождает *простую* выборку, то есть случайную, однородную и независимую, а также предлагаем пути построения соответствующей выборки.

2 Обзор литературы

Work in progress.

3 Постановка задачи классификации

Рассмотрим некоторый временной ряд, то есть функцию определенную на множестве временных меток.

$$S : T \to \mathbb{R} \text{ где } T = \{t_0, t_0 + d, t_0 + 2d \ldots\}, |T| < \infty$$

Зададим некоторую ширину сегмента $k \in \mathbb{N}$, тогда объектом s_i мы назовем набор

$$s_i = (S(t), S(t-d), S(t-2d), \ldots, S(t-(k-1)d)) \in \mathfrak{S}$$

Необходимо восстановить зависимость $y = f(s), f : \mathfrak{S} \to \{+1, -1\}$. Для этого задана обучающая выборка

$$\mathfrak{D} = \{(s_i, y_i)\}_{i=1}^{l}, \quad y_i \in \{+1, -1\}$$

а также функция потерь

$$L(f(s), y)$$

Таким образом мы решаем задачу оптимизации

$$\hat{y} = \arg\min_{y \in Y} \sum_{i=1}^{l} L(f(s_i), y_i)$$

3.1 Комбинированное признаковое описание

Пусть $\mathfrak{G}$ — множество функций вида $g : \mathfrak{S} \to \mathbb{R}^m$, где $m = m(g)$, то есть это множество отображений пространства объектов сложной структуры в пространство действительных чисел некоторой размерности (для каждой функции размерность может быть своя). В G могут лежать, например

- Множество моделей локальной аппроксимации сигнала
- Множество статистик
- Множество экспертных оценок каждого из сложных объектов

Возьмем конечный поднабор этих функций $G = [g_1 \ldots g_k] \subset \mathfrak{G}$. Обозначим сумму размерностей образов функций из набора как

$$n_G \triangleq \dim(\mathrm{Im}(g_1)) + \dim(\mathrm{Im}(g_2)) + \ldots + \dim(\mathrm{Im}(g_k))$$

Тогда G индуцирует отображение $G : \mathfrak{S} \to \Theta \subset \mathbb{R}^{n_G}$, причем в векторах образа первые $\dim(\mathrm{Im}(g_1))$ компонент соответствуют образу g_1, следующие $\dim(\mathrm{Im}(g_2))$ соответствуют g_2 и так далее. Θ называется *признаковым пространством* объектов сложной структуры $\mathfrak{S}$. Тогда, мы можем искать f в семействе суперпозиций $h(g(\cdot), \gamma)$, где

- g — это признаковое отображение
- $h(\cdot, \gamma)$ — параметрическое отображение Θ в $\{+1, -1\}$, которое фактически соответствует некоторому алгоритму машинного обучения, параметризованного вектором гиперпараметров γ.

Для отображения из параметрического пространства задана функция ошибки $L(h(g(s_i), \gamma), y_i)$, тогда задача разбивается на два этапа:

- Поиск и вычисление отображения $X = \{g(\{s_i\}_{i=1}^{l}), y_i)\}$.
- Определение оптимальных парметров γ в задаче оптимизации

$$\hat{\gamma} = \arg\min_{\gamma} \sum_{i=1}^{l} L(h(x_i, \gamma), y_i)$$

Основное допущение, принимаемое в данном методе является допущение о том, что выборка в признаковом пространстве объектов является простой. В данной работе мы рассматриваем, для каких признаковых пространств это допущение справедливо, а также предлагаем способы построения таких выборок.