УДК 519.7:004.855.5

**Комбинаторная оценка вероятности переобучения**

**на основе кластеризации и покрытий множества алгоритмов**

А. И. Фрей

Московский физико-технический институт (государственный университет)

Завышенность теоретических оценок обобщающей способности алгоритмов классификации остаётся открытой проблемой уже более сорока лет, начиная с работ В. Н. Вапника и А. Я. Червоненкиса [1]. На практике наиболее перспективным выглядит комбинаторный подход [2], в рамках которого уже удалось добиться улучшения качества логических закономерностей [3]. Данная работа направлена на дальнейшее повышение точности комбинаторных оценок вероятности переобучения за счет учета сходства между алгоритмами с близкими векторами ошибок.

Рассмотрим задачу классификации. Пусть  — генеральная выборка из  объектов,  — множество алгоритмов классификации, — бинарная функция потерь. Для произвольной подвыборки  определим *число* и *частоту* ошибок алгоритма  соответственно, как  и . *Методом обучения* называют отображение вида . Метод обучения ставит в соответствие произвольному множеству алгоритмов и обучающей выборке  некоторый алгоритм  В данной работе рассматривается метод *пессимистической минимизации эмпирического риска* (ПМЭР), действующий по правилу  где , 

Пусть  — множество всех разбиений генеральной выборки  на обучающую выборку  длины и контрольную выборку  длины  Следуя [2], определим *вероятность переобучения*  как долю разбиений  при которых переобученность  алгорима  превышает заданный порог :

 (1)

где  а квадратные скобки действуют по правилу ,

Введем на  отношение частичного порядка: означает, что  и . Если  и  такой, что  то будем говорить, что  *предшествует* , и записывать .

**Теорема 1.** *Пусть множество алгоритмов*  *представлено в виде разбиения на непересекающиеся подмножества* *, такие что внутри каждого* *алгоритмы допускают равное число ошибок на полной выборке. Пусть*  *— ПМЭР. Для каждого*  *рассмотрим порождающее и запрещающее множества*  *и* *:*



*Пусть, кроме этого, каждое подмножество вложено в объемлющее множество:* *Тогда для вероятности переобучения выполнена следующая оценка:*

 *(2)*

*где* *— верхняя оценка на вероятность*  *и введены следующие обозначения:*  *— число ошибок алгоритмов из* 

По результатам численого эксперимента оценка (2) оказывается точнее оценки, полученной методом порождающих и запрещающих множеств [3]. Кроме этого, новая оценка эффективно вычислима для семейств с существенно большим числом алгоритмов, т.к. сумма (2) содержит меньшее число слагаемых из-за кластеризации алгоритмов с близкими векторами ошибок.

**Литература**

1. *Vapnik V.N., Chervonenkis A.Y.* On the uniform convergence of relative frequencies of events to their probabilities. – Theory of Probability and Its Applications. – 1981. – N 16(2). – pp. 264-280.

2. *Воронцов К.В.* Точные оценки вероятности переобучения. – Доклады РАН. – 2009. – Т. 429. – №1. – С. 15-18.

3. *Vorontsov K.V., Ivahnenko A.A.* Tight combinatorial generalization bounds for threshold conjunction rules. – PReMI’11. – 2011. – pp. 66-73.