ВЛИЯНИЕ СТАТИСТИЧЕСКИХ ХАРАКТЕРИСТИК ОБУЧАЮЩЕЙ ВЫБОРКИ НА ЕЁ РЕПРЕЗЕНТАТИВНОСТЬ

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

Искра В.В.

Татур М. М. – д-р. техн. наук, профессор

При практической оценке алгоритмов классификации важно отличать ошибки, вызванные несоответствиями выбранных алгоритмов, от ошибок, связанных с недостаточной репрезентативностью обучающей выборки. На данный момент не существует общепринятого математического определения понятия репрезентативности. Далее предлагается подход к определению понятия репрезентативности и исследованию влияния статистических характеристик обучающей выборки на её репрезентативность.

Исходя из материалов, приводящихся в литературе по социологии и статистике [1,2,3], можно принять следующее определение:

Репрезентативность – способность выборки представлять параметры генеральной совокупности, значимые с точки зрения задач исследования [2, 3].

Для выделения критериев, позволяющих оценить влияние статистических характеристик обучающей выборки на её репрезентативность, применим это определение к одному из известных подходов к формализации обучения классификаторов – принципу минимизации эмпирического риска [4,5].

В соответствии с принципом минимизации эмпирического риска, задача обучения классификатора с учителем представляется как минимизация функции, называемой функционалом риска:

$$R(w) = \int L\big(d, F(x,w)\big) dF_{X,D}(x,d)$$

где $x \in X$– входной сигнал, $w \in W$— настраиваемые параметры классификатора, $F(x,w)$– классификатор, $d = d(x)$– желаемый отклик,$L\big(d, F(x,w)\big)$– функция ошибки, $F_{X,D}(x,d)$– функция распределения примеров генеральной совокупности, $x, d \in X \times D$.

Так как $F_{X,D}(x,d)$ неизвестна, $R(w)$заменяется на:

$$R_{emp}(w) = \sum_{i=1}^{N} L\big(d_i, F(x_i, w)\big)\frac{1}{N}$$

где $\{(x_i, d_i)\}_{i=1}^{N} \in T$ – обучающая выборка.

При таком подходе понятие репрезентативности обучающей выборки можно переформулировать следующим образом:

Выборка T, обучающая машину $F(x,w)$ с использованием функции стоимости $L\big(d, F(x,w)\big)$, является репрезентативной в той мере, в которой минимум соответствующего функционала эмпирического риска $R_{emp}(w)$ близок к минимуму функционала риска $R(w)$.

Эти соображения позволяют ввести понятие функционала риска репрезентативности:

$$R_{repr}(w) = Q\left(R(w), R_{emp}(w)\right)$$

где $Q(.,.)$ – некоторый оператор, сравнивающий функции $R_{emp}(w)$ и $R(w)$.

В целях оценки состоятельности данного определения понятия репрезентативности обучающей выборки, а также в целях определения его взаимосвязи с традиционными методами, основанными на вычислении вероятности смещения оценки математического ожидания значений параметров выборки при предположении их нормального распределения, были экспериментально исследованы следующие величины:

- смещение среднего значения параметров обучающей выборки по сравнению с генеральной совокупностью;
- смещение дисперсии параметров обучающей выборки по сравнению с генеральной совокупностью;
- отклонение $R_{emp}(w)$ и $R(w)$для заданного классификатора;
- ошибка обобщения заданного классификатора.

Список использованных источников:

1. Большой толковый социологический словарь (Collins) // В 2 т. - т. 2 (П-Я), пер. с англ. - М.: Вече, АСТ, 1999. - с. 158.
2. Сотникова, Г.Н. Репрезентативность / Г. Н. Сотникова, Г.В. Осипов // Российская социологическая энциклопедия. - М.:НОРМА-ИНФА-М, 1998. - с. 445.
3. Ильясов, Ф. Н. Репрезентативность результатов опроса в маркетинговом исследовании // Социологические исследования. 2011. - № 3. - с. 112-116.
4. Хайкин, С. Нейронные сети: Полный курс / С. Хайкин. – М.: Издательский дом «Вильямс», 2006. - с. 140-146.
5. Vapnik, V. N. Principles of risk minimization for learning theory // Advances in Neural Information Processing Systems, 1992. - vol. 4. - p. 831-838.