Цель исследования:

Целью работы является создание и теоретическое обоснование методов оценки объема многофакторных выборок, учитывающих вид модели классификации и более точных по сравнению с известными методами; создание методов классификации малых выборок

Предмет исследования:

Оценить минимальный объём выборки — количество производимых измерений некоторого параметра или набора параметров, необходимый для выполнения некоторых ранее сформулированных условий.

Исследуемая проблема:

Исследование направлено на решение проблемы выбора моделей при классификации выборок малой мощности. По заданной выборке, включающей многокритериальное описание объектов и метки класса объектов, требуется получить оценку структурных параметров, получить асимптотическую оценку необходимого объема выборки и указать предпочтительный подход к решению задачи классификации. Для классификации объекта требуется получить оценку параметров выбранной модели и выполнить анализ ошибок классификации.

Решаемая в данной работе задача:

В данной работе основное внимание уделяется байесовским методам оценки объёма выборки. Оценка объёма выборки в байесовской постановке включает оценку апостериорного распределения $p(D|w)$ параметров модели. При отсутствии наблюдаемых данных, апостериорное распределение $p(w|D) = p(D|w)p(w)/p(D)$ совпадает с априорным $p(w)$, как в классических методах оценки объёма выборки. Разница между вторым и третьим случаями заключается только в способе оценки распределения $p(w|D)$ — на основе сэмплированных, либо реально наблюдаемых данных.

Предлагаемое решение:

Статистические методы позволяют оценить объем выборки, исходя из предположений о распределении данных и информации о соответствии наблюдаемых величин предположениям нулевой гипотезы. В случае, если объем исследуемой выборки достаточен или избыточен, возможно применение методов, основанных на наблюдении за изменением некоторой характеристики процедуры построения модели при увеличении объема выборки. В частности, наблюдая за отношением качества прогнозирования на контрольной выборке и обучающей выборке, определим достаточный объем выборки как соответствующий началу переобучения. Таким же образом производится оценка объема выборки в рамках предлагаемого метода: предлагается считать объем выборки достаточным, если расстояние между распределениями, оцененными на подвыборках данного объема, достаточно мало. Такой подход не требует дополнительного обобщения на случай многих переменных. Кроме того, оценку можно производить как при наличии предположений о распределении данных, так и в их отсутствие.

Анализ сильных и слабых сторон предлагаемого решения:
Недостатком данного подхода является то что количественные оценки возможно получить лишь в случае, когда объем выборки избыточен. В противном случае метод позволяет лишь определить, является ли текущий объем выборки достаточным

Работа или работы описывающие наиболее близкие решения:
http://svn.code.sf.net/p/mlalgorithms/code/PhDThesis/Motrenko/doc/

http://svn.code.sf.net/p/mlalgorithms/code/Group874/Motrenko2014KL/

Цель эксперимента, на каких данных будет выполнен эксперимент:
Сравнение подходов (least confidence sampling), (max entropy sampling) со случайным выбором и информацией Lindley (Lindley). Эксперименты планируется проводить на выборках объёма M = 1000.