

Оценка качества обучающих выборок в программном средстве формирования эталонов аэрокосмической системы лесоинвентаризации

В настоящее время ведётся активная работа по созданию системы мониторинга леса, которая позволит автоматизировать процессы лесоинвентаризации. Центральное место среди этих процессов занимает дешифрирование (классификация) аэрокосмических снимков. Под классификацией понимается процесс выделения на снимке объектов, относящихся к разным классам. Применительно к цифровым снимкам, классификация - это разделение всех пикселей снимка на группы, соответствующие разным объектам. Одним из этапов процесса классификации является формирование обучающей выборки (эталона). Обучающая выборка - это набор значений яркостей пикселей в разных спектральных каналах в пределах эталонного участка изображения. Эталонные участки формируются на предшествующих этапах. Полученные обучающие выборки для каждого эталонного участка используются в процессе непосредственно автоматизированного дешифрирования, таким образом результаты автоматизированной классификации снимков напрямую зависят от качества сформированной обучающей выборки, на основе которой производилось дешифрирование. Поэтому в рамках разработки программных средств формирования эталонов необходимо решить вопрос оценки качества формируемых обучающих выборок.

Качество обучающих выборок оценивают по различным критериям. Среди основных можно выделить следующие:

Однородность – отсутствие нехарактерных для класса значений яркости. Оценивается по величине стандартного отклонения значений яркости от среднего, а так же по результатам пробной классификации в

пределах специально отобранных контрольных участков.

Различимость – достаточное различие спектральных яркостей различных классов. Различимость можно оценить по степени перекрытия значений яркости классов и гистограмм яркости.

Характер распределения значений яркости - необходимо максимальное сходство распределения значений яркости с нормальным распределением и одномодальность гистограммы (особенно для статистических методов).

Таким образом, после формирования эталонов необходима автоматическая их проверка на соответствие данным критериям, после чего должно быть принято решение об их пригодности для классификации. В случае низкого качества обучающих выборок необходимо заново выбрать эталонные участки и провести эталонирование.

Для оценки соответствия обучающих выборок вышеприведённым критериям используются статистические показатели. В первую очередь, это стандартное отклонение (среднеквадратическое отклонение), используемое для определения однородности:

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}, \text{ где}$$

n — количество пикселей в выборке;

\bar{x} — среднее значение яркости пикселей выборки;

x_i — значение яркости текущего пиксела.

Поскольку стандартное отклонение показывает степень разброса значений выборки относительно среднего значения, то чем меньше значение отклонения, тем более компактно расположены значения яркостей пикселей в пространстве признаков и тем более высокое качество выборки.

Гистограммы распределения яркости, построенные на основе обучающих выборок, также могут использоваться для определения качества эталонов. Если гистограммы разных классов перекрываются (то есть, один или несколько интервалов значений одной гистограммы совпадает с интервалами другой), это значит, что при классификации пикселы, попавшие

в этот участок, не смогут быть точно отнесены к какому-либо классу и может возникнуть ошибка. Чем сильнее перекрываются гистограммы, тем больше пикселей находится в участке перекрытия и тем хуже качество обучающей выборки. В идеальном случае гистограммы всех участков не перекрываются, и результат будет точным. В случае наличия пересечений на точность дешифрирования будет влиять метод классификации.

Как правило, снимки содержат множество спектральных каналов, и в разных каналах может быть разная различимость. Поэтому с целью повышения качества формирования эталонов можно из множества каналов выбрать небольшое количество (2 — 4), в которых значения яркостей классов пересекаются менее всего.

Вид гистограммы также является важным критерием определения качества. Как было сказано выше, распределение значений выборки должно быть как можно более близко к нормальному. Это можно определить как визуально по гистограмме, так и аналитически с помощью выражения:

$$\frac{\sum_i |x_i - \bar{x}|}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$

Чем меньше значение этого выражения, тем ближе данное распределение к нормальному.

Кроме статистических показателей для оценки качества обучающих выборок применяют пробную классификацию. Эта классификация проводится на контрольных участках, либо на снимках с известным расположением объектов. Если в результате классификации достигнута необходимая точность (определённый процент пикселей классифицирован верно), то обучающая выборка пригодна для использования. При проведении пробной классификации можно также выбрать метод, наиболее пригодный для дешифрирования. Для этого необходимо поочерёдно проводить дешифрирование различными методами, и после этого сравнить точность, достигнутую каждым методом. Метод, с помощью которого была достигнута максимальная точность будет самым предпочтительным.

С целью повышения качества как обучающей выборки, так и классификации в целом, можно вести в файл эталонов информацию, не связанную со спектральными характеристиками снимка. В качестве такой информации могут служить данные о типе снимка, времени года, погодных условиях, природной зоне местности и т. п. С помощью данных о природной зоне местности можно убрать из рассмотрения объекты, которые никак не могут находиться в данной зоне. Данные о типе снимка, времени года и погодных условиях могут послужить для выбора данного файла эталонов для классификации снимков, сделанных в схожих условиях.

Приведённые методы оценки качества обучающей выборки позволяют с уверенностью судить о применимости выборки к конкретной задаче классификации, повысить качество как обучающей выборки, так и процесса классификации в целом. Комбинация нескольких методов ещё больше повышает качество.