Заключение

Анализ гистограмм распределения признаков пространственного спектра показал, что области распределения значений фактора Малиновской для различных групп кристаллограмм практически не пересекаются. Это свидетельствует об эффективности применения предложенной технологии анализа факторов формы пространственного спектра для классификации дендритных кристаллограмм.

Литература

1. Мартусевич, А. К. , Кристаллографический анализ: общая характеристика / А. К. Матрусевич // Вятский медицинский вестник. - 2002. - N. 3. С. 59-61.

2. Куприянов, А. В. Сегментация текстурных изображений на основе оценивания локальных статистических признаков / А. В. Куприянов // Вестник СГАУ. – 2008. – №2 (15). – С. 245-252.

И.А. Лёзин, А.О. Авдиенко

РЕШЕНИЕ ЗАДАЧИ КЛАССИФИКАЦИИ И МНОГОКРИТЕРИАЛЬНОГО ПОИСКА ПРИ ИМПОРТЕ БОЛЬШИХ МАССИВОВ ДАННЫХ

(Самарский государственный аэрокосмический университет имени академика С.П. Королева (национальный исследовательский университет))

По условию задачи, на вход автоматизированной системы поступает файл в некотором формате (например, CSV) с записями о сотрудниках организации, которые нужно импортировать в базу данных, предварительно провалидировав и разделив на классы. Записи могут быть невалидными, повторяющимися, уже существующими в базе, новыми, и т.д. Объем файла довольно велик, несколько десятков тысяч строк.

Каждая запись содержит M полей, то есть в файле M колонок и N строк. Некоторые колонки используются для поиска дубликатов записей с учетом их приоритета.

Подробнее о возможных сложностях и проблемах:

- в файле могут быть записи вообще без колонок, по которым проводится проверка;
- имя задается в 3 колонках, при этом никто не может гарантировать, что фамилия будет стоять в графе "Фамилия";
- ни одна из поисковых колонок не уникальна среди всех записей, уникальность поддерживается только по группе атрибутов;
- файл импортируется не весь целиком, а по частям, поэтому непосредственно в момент валидации мы не видим всего файла целиком.

По результатам импорта формируется отчет.

При работе с большими массивами данных особое внимание следует уделить оптимальности алгоритмов обработки с точки зрения быстродействия и затрачиваемой памяти. Некоторые допущения и возможные упрощения, сделанные для сравнительно небольших размерностей решаемой задачи, зачастую неприменимы к достаточно большим объемам входных данных. Особенно это касается задач поиска и алгоритмической классификации, при решении которых зачастую требуется неоднократный просмотр имеющихся входных массивов.

При решении поставленной задачи использовался следующий алгоритм – входной файл полностью разбирается на лексемы и загружается в оперативную память. Уже в процессе загрузки записи валидируются и разделяются на классы по набору эвристических критериев. Далее записи, уже прошедшие обработку, разделяются на группы, которые при необходимости импортируются в базу данных (batch update); если добавление не требуется, то они просто учитываются при формировании отчета.

Некоторые из использовавшихся принципов и эвристик:

– ёперебор сочетаний значений колонок, отвечающих за один атрибут, с дальнейшей перестановкой;
– отброс некоторых более приоритетных полей при наличии заполненной комбинации нескольких менее приоритетных;
– повторная проверка записей с целью перестройки конечного решения в случае появления новых дубликатов.

Система достаточно хорошо разделяет входные данные на классы (число ошибок близко к нулю), но её быстродействие довольно невелико. На данном этапе планируются дальнейшие исследования по увеличению производительности алгоритмов. В дальнейшем планируется применение нейронных сетей для решения задачи классификации, что позволит достигнуть большего быстродействия при той же точности классификации.

Литература

1. Муравьиный алгоритм [Электронный ресурс] – http://habrahabr.ru/post/221237/.

2. Метрические алгоритмы классификации [Электронный ресурс] – http://www.ccas.ru/voron/download/MetricAlgs.pdf.

3. Теория автоматической классификации [Электронный ресурс] – http://stu.sernam.ru/book_stat3.php?id=83.