Г. А. Ботвин, **П. П. Петтай** (Санкт-Петербург, СПбГУ)**. Интеллектуальные методы анализа данных.**

Решение задач анализа динамики и прогнозирования развития социально-экономических явлений и процессов в классической постановке ориентировано на использовании накопленных статистических данных. Наличие такой информации, тем не менее, не гарантирует возможность учета всех возможных факторов, неявно присутствующих и не всегда выявляемых при первичном анализе. Поэтому существенное развитие получают методы интеллектуального анализа данных. Одним из подходов к интеллектуальному анализу является многомерный анализ данных. В процессе принятия решений экономист-аналитик генерирует некоторую совокупность гипотез. Проверка гипотез осуществляется на основании информации о предметной области исследования. На практике наиболее удобным способом представления такой информации является выявление зависимости между некоторыми параметрами. Число факторных параметров, оказывающих существенное влияние на результативный признак (признаки), может варьироваться в достаточно широких пределах. Поэтому возникает необходимость в построении зависимостей результативных признаков от факторных показателей, оказывающих существенное влияние на конечные результаты и оценки силы связи каждого из них на конечные результаты. Традиционные средства и методы анализа, оперирующие факторами в парадигме реляционной модели данных, не могут в полной мере удовлетворять современным требованиям. Данные, количественно или качественно характеризующие тот или иной фактор, представляют собой точки в этом многомерном пространстве. В простейшем случае многомерную модель данных можно представить в виде гиперкуба. Над таким гиперкубом могут выполняться операции среза (формирование подмножества многомерного массива данных, соответствующих единственному значению одного или нескольких факторов), операции вращения (изменение расположения факторов относительно системы координат), консолидации и детализации (переход от детального представления данных к агрегированному или наоборот).

Таким образом, можно предположить, что парадигма принятия решений экономистами-аналитиками все в большей степени концентрируется на использовании методов интеллектуального анализа. Точкой бифуркации развития методов анализа данных можно считать переход к технологии Data Mining как процессу исследования и обнаружения современными средствами искусственного интеллекта и алгоритмическими методами в огромных информационных потоках скрытых знаний, которые ранее не были известны, нетривиальны, обладают практической полезностью и доступны для интерпретации аналитиками.

Практика аналитической деятельности и анализа динамики социально-экономических процессов и явлений приводит к выводу о том, что наиболее трудоемкими являются проблемы постановочной и алгоритмической формализации задач кластеризации.

На основе теории нечетких множеств разработан Fuzzy Classifier Means — FCM-алгоритм кластеризации данных.

Реализация алгоритма приводит к необходимости вычисления расстояний между векторами. Если признаки имеют различную размерность и/или измерены в различных шкалах, то расстояние между векторами станет абсолютно неинформативным.

Будем предполагать, что искомые нечеткие кластеры представляют собой *нечеткие множества*, образующие *нечеткое покрытие* исходного множества объектов кластеризации. Содержательный смысл предположения заключается в том, что объект принадлежит хоть какому-нибудь одному кластеру.

Минимизируемый функционал нелинейный и даже невыпуклый в силу специфики вычисления показателей. Если применить для решения данной оптимизационной задачи нелинейного программирования метод множителей Лагранжа, то в итоге получим достаточно сложную систему нелинейных уравнений, решить аналитически

которую не представляется возможным.

Авторами предложено обобщение итерационного FCM-алгоритма. В обосновании сходимости метода в качестве метрики использовался квадрат евклидова расстояния. Данная метрика удобна в том смысле, что она проще всего дифференцируется.

Для нахождения решения оптимизационной задачи нелинейного программирования в обобщенном методе FCM (в смысле использования *произвольных* метрик) целесообразно использовать генетические алгоритмы. Принципиальными особенностями генетических алгоритмов, позволяющими их применить в FCM-алгоритме, является отсутствие требований к дифференцируемости целевой функции и возможность отыскания *глобальных* экстремумов.

Практическая реализация алгоритма приводит к выводу о возможности использования предложенного подхода как одного из методов интеллектуального анализа данных.