Министерство образования Республики Беларусь
Учреждение образования
«Белорусский государственный университет
информатики и радиоэлектроники»

УДК 004.93´14

Петюкевич Наталья Станиславовна

Разведочный анализ данных на основе эвристической возможностной кластеризации

АВТОРЕФЕРАТ

на соискание степени магистра технических наук

по специальности 1- 40 80 05 - Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Научный руководитель
Вятченин Дмитрий Аркадьевич
к.ф.н., доцент

Минск 2016

КРАТКОЕ ВВЕДЕНИЕ

При необходимости обработки больших массивов данных на первой стадии возникает задача предварительного анализа имеющихся данных для:

- определения возможного числа классов, на которые «расслаивается» исследуемая совокупность;
- выделения в ней аномальных наблюдений;
- проецирования исследуемой совокупности на плоскость, к примеру, двух главных компонент или наиболее информативных признаков.

С этой целью может быть применен аппарат эвристической возможностной кластеризации.

Разведочный анализ данных традиционно используется, когда, с одной стороны, у аналитика имеется таблица многомерных данных, а с другой – информация о природе этих данных неполна или вовсе отсутствует. К задачам разведочного анализа данных традиционно относят:

- проецирование данных в двумерное пространство;
- обнаружение аномальных наблюдений и очистка данных;
- обнаружение возможного числа кластеров в исследуемой совокупности.

При этом проецирование исследуемой совокупности в двумерное признаковое пространство предполагает вначале нахождение соответствующих признаков, на плоскость которых будет проецироваться исследуемая совокупность.

В последние 20-30 лет огромную популярность получили методы автоматической классификации, основанные на концепции теории нечетких множеств, предложенной американским математиком Л.А. Заде, Это обусловлено высокой точностью и содержательной осмысленностью этих методов в сравнении с традиционными методами кластеризации. Необходимость на практике постоянно принимать решения в условиях неопределенности и нечеткой информации показывает, что теория нечетких множеств является стратегическим инструментом управления сложными системами. Технологии и алгоритмы, разработанные в рамках этой теории, являются универсальными по применимости. Сферы применения нечеткого кластерного анализа включают анализ данных, распознавание образов, сегментацию изображений

Подход, основанный на теории возможностей, является развитием идей нечеткой кластеризации. Возможностные методы кластеризации накладывают менее строгие ограничения на искомый результет, что и делает из более гибкими и общими методами обработки данных. Одним из таких подходов стал эвристический метод возможностной кластеризации, получивший дальнейшее развитие в задачах классификации и управления.

Направлением исследования данной работы является аппарат эвристической возможностной кластеризации, построение алгоритмов и оценка их эффективности в зависимости от выбора оператора агрегирования.

ОБЩАЯ ХАРАКТЕРИСТИКА

Цель и задачи исследования

Целью диссертационной работы является разработка методов разведочного анализа данных, основанных на аппарате эвристической возможностной кластеризации.

Для достижения поставленной цели необходимо изучить методы разведочного анализа данных, существующие походы к кластеризации, методы нечеткой и возможностной кластеризации и решить следующие задачи:

1. Разработка методов разведочного анализа данных.
2. Построение множества значений наиболее возможного числа классов в искомом распределении по нечетким α-кластерам.

Объектом исследования являются методы эвристической возможностной кластеризации.

Предметом исследования являются особенности аппарата эвристической возможностной кластеризации, позволяющие разработать методы наглядного представления многомерных данных и построения множества значений возможного числа нечетких кластеров в искомом распределении.

Связь работы с приоритетными направлениями научных исследований и запросами реального сектора экономики

Работа выполнялась в соответствии с научно-техническим заданием и планом работ кафедры «Программное обеспечение информационных технологий» по теме «Разработать модели, методы, алгоритмы для оценки параметров, повышения надежности и качества функционирования аппаратно-программных средств систем и сетей сложной конфигурации и внедрить в современные обучающие комплексы » (ГБ № 11-2004, № ГР 20111065, научный руководитель НИР – В. В. Бахтизин).

Личный вклад соискателя

Результаты, приведенные в диссертации получены соискателем в соавторстве с научным руководителем Д .А. Вятчениным.

Опубликованность результатов диссертации

По теме диссертации опубликована 1 работа в сборнике трудов и материалов научной конференции аспирантов, магистрантов и студентов, 1 работа принята на публикацию.

Структура и объем диссертации

Диссертация состоит из введения, четырех глав, заключения, списка использованных источников, списка публикаций и приложений.

Общий объем работы составляет 67 страниц, из которых основного текста –51 страница,7 рисунков на 3 страницах, 3 таблицы на 2 страницах, список использованных источников из 31 наименований на 3 страницах и 3 приложения на 11 страницах.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Диссертация состоит из введения, общей характеристики работы, четырех глав, заключения, списка использованных источников, списка публикаций и приложений.

Во введении указана область исследования, дана краткая характеристика исследуемых вопросов.

В первой главе диссертации рассматриваются основные понятия разведочного анализа данных, сущность и типологизация задач снижения размерности анализируемого признакового пространства, основные методы разведочного анализа данных, модели данных, форма задания исходной информации, типы оптимизируемого критерия информативности искомого набора признаков. Сформулированы цели его использования.

Во второй главе рассматривается нечеткий и возможностный подход к решению задач кластеризации, описываются основные понятия теории нечетких множеств, нечетких отношений и нечетких чисел, алгоритмы нечеткой и возможностной кластеризации, рассматривается математические модели, лежащие в основе построения методов снижения размерности, приведены схемы алгоритмов нечеткой и возможностной кластеризации.

В третьей главе диссертации даются основные понятия эвристической возможностной кластеризации, определение α-кластера, приведены общие меры кластерной валидности: коэффициент разбиения, энтропия разбиения , компактность и индекс разделения. Рассматриваются алгоритмы эвристической возможностной кластеризации двух типов: реляционные алгоритмы и алгоритмы, основанные на прототипах. В первом случае матрицей исходных данных служит матрица нечеткой толерантности T, являющаяся разновидностью

матрицы «объект-объект», а во втором случае – матрица вида «объект-признак».

В четвертой главе описывается метод построения треугольного нечеткого числа и гауссовского нечеткого числа, виды операторов агрегирования, методология построения множества возможного числа классов в искомом распределении по нечетким α-кластерам. Рассмотрено применение полученного алгоритма классификации для конкретного набора данных .

ЗАКЛЮЧЕНИЕ

В ходе выполнения работы был проведен обзор методов нечеткой и возможностной кластеризации, рассмотрены основные определения и обзор эвристических методов возможностной кластеризации.

В результате проделанной работы построен алгоритм для оценки нижней границы для числа кластеров c_{min} и верхней границы для числа кластеров c_{max} для множества $\{c_{min}, \ldots, c_{max}\}$ наиболее возможного числа нечетких кластеров в искомой структуре кластеризации. Основой предложенного метода явились эвристический D-AFC-TAGA алгоритм и гауссовские нечеткие числа . В качестве наглядного примера применения предложенного метода классификации рассматриваются данные по ирисам Андерсона. Предложенный подход может быть обобщен для случая множества реляционных данных с использованием эвристического D-PAFC- алгоритма возможностной кластеризации. Так как обнаружение возможного числа кластеров в исследуемой совокупности является одной из задач РАД, рассмотренный в данной работе метод применим для разведочного анализа данных.

Полученные результаты могут быть применены для решения задач обработки данных в различных областях, в том числе в медицине и экономике, в автоматизированных системах контроля и поддержки принятия решения в условиях неопределенности и нечеткой информации.

СПИСОК РАБОТ

1. Петюкевич, Н. С. Снижение размерности исследуемых данных на основе эвристической возможностной кластеризации / Н. С. Петюкевич // Компьютерные системы и сети : материалы 52-й научной конференции аспирантов, магистрантов и студентов. (Минск, 25 - 30 апреля 2016 года). – Минск : БГУИР, 2016. – С. 74 - 75.

2. Viattchenin, D. Estimation of Bounds of the Set of Possible Number of Fuzzy Clusters in a Sought Clustering Structure / D. Viattchenin, N. Petsiukevich, A. Damaratski // Communications on Applied Electronics (accepted).