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

УДК 004.383

Ширай
Станиславв Юрьевич

Алгоритмы кластеризации, основанные на теории интуиционистских нечетких множеств

АВТОРЕФЕРАТ
на соискание академической степени
магистра технических наук

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

Научный руководитель
Вятченин Д.А.
к.ф.н., доцент

Минск 2015

КРАТКОЕ ВВЕДЕНИЕ

Кластеризация – это процесс классификации объектов по группам без предварительного обучения модели. Кластеризация считается удобным инструментом обнаружения похожих объектов данных. Она разделяет похожие между собой объекты на различные не похожие друг на друга группы, называемые кластерами. Можно считать, что кластеризация – это процесс разделения объектов на группы таким образом, что схожесть объектов внутри одной группы значительно больше, чем схожесть объектов из различных кластеров.

Целью кластеризации является нахождение групп связанных между собой объектов и, таким образом, отыскание корреляций в больших наборах данных. Идея группировки данных является довольно простой и очень похожа на способ, каким думают люди сталкиваясь с большим объемом данных. Люди обычно стремятся свести большое число объектов к небольшому количеству классов или категорий, которые потом используются для дальнейшего изучения.

В настоящее время неуклонно возрастает интерес к нечетким моделям из-за возможностей их применения в условиях неполноты информации. Модели, как статических, так и динамических систем могут быть построены на основе положений теории нечетких множеств и нечеткой логики. Нечеткое моделирование не подменяет собой другие методологии моделирования сложных систем, в которых параметры системы могут получить численные оценки, но являются инструментом проектирования и анализа сложных систем в целом, а также различных из аспектов, в том случае, когда качественные элементы системы доминируют над количественными.

Нечеткая кластеризация особенно полезна тогда, когда границы между кластерами размыты. Нечеткая кластеризация является хорошо проработанной областью, а алгоритмы нечеткой кластеризации являются стандартным инструментом прикладной статистики и машинного обучения.

Развитием идей нечеткой кластеризации является подход, основанный на теории возможностей. Возможностные методы кластеризации накладывают менее строгие ограничения на искомый результат кластеризации, что делает их более общим и гибким методом обработки данных. Так же степени принадлежности объектов кластерам при таком подходе можно интерпретировать как степень типичности объекта в кластере.

Интуиционистская нечеткая кластеризация, основанная на теории интуиционистских нечетких множеств является современным подходом к обработке двусмысленных данных. С тех пор как Атанассов опубликовал свою первую статью на тему интуиционистских нечетких множеств было разработано множество приложений этой теории к практическим задачам анализа данных и принятия решений. Интуиционистские нечеткие множество позволяют учитывать при разработке алгоритмов не только значение принадлежности объекта кластеру, но и значение его непринадлежности этому кластеру. Значения непринадлежности особенно полезны в областях, таких как медицина, в которых проще отсеять неподходящие решения на основе непринадлежности, например, в медицине

проще отклонить диагноз при помощи анализа симптомов, чем подтвердить диагноз.

Совмещение интуиционистской нечеткой кластеризации с возможностным подходом так же полезно в задачах, данные для которых содержат шум. В таком случае объекты, представляющие из себя шум, будут иметь меньшую степень принадлежности к кластерам, интерпритируемую как значение типичности объекта по отношению к другим объектам кластера, и большее значение непринадлежности, т.е. нетипичности объекта по отношению к другим объектам кластера.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования

Целью диссертационной работы является разработка эвристических возможностных алгоритмов интуиционистской нечеткой кластеризации с задаваемым количеством интуиционистских нечетких кластеров, так же алгоритмы на основе операции транзитивного замыкания интуиционистского нечеткого отношения сходства.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Обобщить понятие частично разделенных нечетких кластеров на случай данных, представленных интуиционистскими нечеткими множествами и разработать на его основе алгоритм кластеризации с задаваемым количеством искомых кластеров.
2. Разработать алгоритм кластеризации, основанный на операции транзитивного замыкания интуиционистского отношения сходства, который отыскивает распределения объектов по заранее неизвестному количеству интуиционистских нечетких кластеров.
3. Провести экспериментальное исследование разработанных алгоритмов и подтвердить, что они адекватно относительно их особенностей кластеризуют наборы данных.

Объектом исследования являются задачи классификации и алгоритмы их решения.

Предметом исследования является обработка данных, содержащих двусмысленности, т.е. содержащих информацию о наличии качественного различия значения признакового объекта.

Основной *гипотезой*, положенной в основу диссертационной работы, является предположение о возможности адекватного представления данных, содержащих двусмысленность с помощью аппарата теории интуиционистских нечетких множеств и возможности кластеризации представленных таким образом данных при помощи алгоритмов кластеризации, использующих тот же математический аппарат.

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

Работа выполнялась в соответствии научно-техническими заданиями и планами работ кафедры «Программное обеспечение информационных технологий», и хозяйственными договорами с предприятиями Республики Беларусь:

1. «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).

2. «Проведение исследований по созданию вибродиагностической системы определения качества изготовления и сборки узлов автомобилей БелАЗ» (х/д № 04-1079, № ГР 2005395, научный руководитель НИР – П. Ю. Бранцевич).

Личный вклад соискателя

Большинство результатов, приведенных в диссертации, получены соискателем в соавторстве с научным руководителем Д.А. Вятчениным.

Опубликованность результатов диссертации

По теме диссертации опубликовано 3 печатных работ, из них 2 статьи в рецензируемом издании, 1 работа в сборниках трудов и материалов международных конференцийю 1 работа принята на публикацию.

Структура и объем диссертации

Диссертация состоит из введения, общей характеристики работы, четырех глав, заключения, списка использованных источников, списка публикаций автора. В первой главе представлен анализ теории нечетких множеств и теории интуиционистских нечетких множеств, анализу расстояний и мер сходства, определенных для них, и основных алгоритмов нечеткой кластеризации. Вторая глава посвящена эвристическому возможностному подходу к нечеткой кластеризации. В третьей главе представлен анализ алгоритмов интуиционистской кластеризации. В четвертой главе предложены алгоритмы кластеризации, реализующие эвристический возможностный подход к интуиционистской нечеткой кластеризации, описан метод использования разработанных алгоритмов для решения задачи расстановки меток классов, полученных в результате кластеризации, а так же предложен способ представления численных данных в виде набора интуиционистских нечетких множеств.

Общий объем работы составляет 56 страниц, из которых основного текста – 49 страниц, 4 рисунка на 4 страницах, 4 таблицы на 5 страницах, список использованных источников из 30 наименований на 2 страницах.

ОСНОВНОЕ СОДЕРЖАНИЕ

Во **введении** определена область и указаны основные направления исследования, показана актуальность темы диссертационной работы, дана краткая характеристика исследуемых вопросов, обозначена практическая ценность работы.

В **первой главе** приведены сведения из теории нечетких множеств и интуиционистских нечетких множеств, основные определения и формулы, необходимы для дальнейшего описания как уже имеющихся, так и разработанных в рамках диссертационной работы алгоритмов.

Отдельный раздел первой главы отведен описанию и анализу расстояний между нечеткими множествами и расстояний и мер сходства между интуиционистскими нечеткими множествами, так как многие кластеризационные алгоритмы так или иначе используют их в своей работе и от правильного выбора мер сходства и расстояний в значительной степени зависят результаты кластеризации.

Еще один раздел отведен под описание и анализ классических алгоритмов нечеткой и возможностной кластеризации. Результаты, привденные в данной главе положили начало всей области нечеткой кластеризации и являются основной для большинства современных подходов к нечеткому кластерному анализу.

Результаты исследований, приведенных в первой главе, отражены в работах Заде, Атанассова, Шмидт и Кашпрчика, Дана и Бездека, Бустинце и Бурилло, Кришнапурама и Келлера, Пала и др.

Вторая глава просвещена описанию эвристического возможностного подхода к нечеткой кластеризации, который послужил основой для разработанных в рамках диссертации алгоритмов.

В первом разделе главы описываются основные понятия, необходимые для построения эвристических возможностной алгоритмов, такие, как четкие и нечеткие множества и отношения уровня α, определения понятий нечеткого кластера и его типичных точек. Приведены определения полностью и частично разделимых нечетких кластеров, указаны условия, которым они должны удовлетворять, описано понятие распределения элементов множества по нечетким кластерам, а так же описаны критерии выбора единственного распределения элементов по нечетким кластерам из множества возможных распределений.

Во втором разделе описана процедура работы эвристического возможностного нечеткого алгоритма кластеризации D-AFC(c), разбивающего множество элементов на заданное количество частично разделимых кластеров, приведена процедура декомпозиции матрицы нечеткого отношения сходства для построения последовательности уровней α, необходимых в рамках алгоритма для нахождения всех возможных множеств уровня α.

Третий раздел второй главы посвящен алгоритму D-AFC-TC, который разбивает множество элементов на заранее неизвестное количество полностью раз-

делимых кластеров. В нем приведена процедура построения замыкания нечеткого отношения сходства, полученного на основе описания кластеризуемых объектов в виде матрицы объект-признак, а так же приведен алгоритм нахождения порога нечеткого отношения сходства α при помощи так называемой эвристики скачка.

В **третьей главе** описаны существующие алгоритмы интуиционистской нечеткой и возможностной кластеризации. В ней рассматриваются, как обобщения классических оптимизационных алгоритмов, таких как Fuzzy c-means, так и новые реляционные алгоритмы, основанные на отношениях между интуиционистскими нечетким множествами.

В **четвертой главе** рассматривается эвристический возможностной подход к интуиционистской нечеткой кластеризации.

Первый раздел главы посвящен алгоритму D-PAIFC(c), являющемуся попыткой обобщить эвристический возможностной подход к нечеткой кластеризации, описанный во второй главе на случай интуиционистских нечетких множеств. В разделе даны основные определения, необходимые для такого обобщения – интуиционистские нечеткие кластеры и распределения элементов по ним, декомпозиция нечеткого интуиционистского отношения на пары порогов значений принадлежности и непринадлежности α и β, определены типичные точки интуиционистских нечетких кластеров, введено понятие полностью разделяемых интуиционистских нечетких кластеров, обобщен на случай интуиционистских нечетких множеств критерий выбора единственного распределения объектов по кластерам из множества возможных.

Второй раздел посвящен предложенному в рамках диссертационного исследования алгоритму эвристическому возможностному интуиционистскому нечеткому реляционному алгоритму кластеризации D-AIFC(c) [1], который разработан для разбиения множества объектов на заранее заданное число кластеров. С этой целью обобщается понятие частично разделяемых интуиционистских нечетких кластеров на случай данных, представленных интуиционистскими нечеткими множествами, а так же обобщаются условия, которым такие кластеры должны удовлетворять.

Третий раздел содержит разработанный алгоритм D-AIFC-TC [2], который кластеризует множество объектов на заранее неизвестное количество кластеров. С этой целью используется процедура построения транзитивного замыкания интуиционистского нечеткого отношения сходства, матрица которого используется для генерации полностью разделяемых интуиционистских нечетких кластеров. Так же вводится понятие прототипа интуиционистского нечеткого кластера, на основе сравнения которого с типичной точкой кластера производится выбор единственного решения задачи кластеризация из множества возможных решений [3].

Четвертый раздел четвертой главы содержит вариацию алгоритма D-AIFC-TC, использующую обобщение эвристики скачка для нахождения порогов принадлежности и непринадлежности интуиционисткого нечеткого отношения

сходства α и β для последующего использования их при построении распределений по заранее неизвестному количеству интуиционистских нечетких кластеров [4].

Пятый раздел содержит построенный алгоритм D-AIFC-PS(c), который строит распределение объектов по частично разделенным интуиционистским нечетким кластерам, используя частичное обучение. Его особенность заключается в том, что для каждого кластера определен помеченный объект из множества кластеризуемых объектов, для которого пользователем заданы степени принадлежности и непринадлежности и который должен принадлежать своему кластеру со степенью принадлежности равно либо большей, чем заданная пользователем и со степенью непринадлежности равной, либо меньшей, чем заданная пользователем.

Шестой раздел содержит алгоритм решения задачи присваивания меток классов кластеризованным объектам на основе разработанных в рамках диссертационного исследования алгоритмов.

Седьмой раздел содержит разработанный алгоритм конвертирования численных данных в интуиционистские нечеткие множества при помощи интуиционистского нечеткого комплемента. При помощи разработанных алгоритмов показано, что предложенный метод конвертирования позволяет корректно представить множество объектов в виде интуиционистских нечетких множеств и на примере размеченной тестовой выборки продемонстрировано, что предложенные алгоритмы способны корректно с учетом их специфики кластеризовать представление таким образом объекты.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. В рамках диссертационного исследования существенно расширен эвристический возможностный подход к интуиционистской нечеткой кластеризации, основывающийся на обобщения эвристической возможностной нечеткой кластеризации. Расширен формальный аппарат теории интуиционистских нечетких множеств, необходимый для построения алгоритмов интуиционистской нечеткой кластеризации.

2. Разработан метод кластеризации D-AIFC(c) данных, представленных интуиционистскими нечеткими множествами на заданное количество частично разделимых интуиционистских нечетких кластеров

3. Разработан метод кластеризации D-AIFC-TC данных, представленных интуиционистскими нечеткими множествами на основе обработки матрицы транзитивного замыкания интуиционистского нечеткого отношения сходства. Предложена вариация алгоритма, которая использует обобщения эвристики скачка для уменьшения количества необходимых вычислений.

4. Разработан метод кластеризации D-AIFC-PS(c) данных, представленных интуиционистскими нечеткими множествами, с использованием частичного обу-

чения для разделения объектов на задаваемое число частично разделяемых интуиционистских нечетких кластеров, позволяющий пользователю влиять на результат кластеризации путем задания размеченных объектов и присвоения им граничных значений принадлежности и непринадлежности.

5. Предложен метод расстановки меток классов кластеризованных объектов на основе разработанных алгоритмов кластеризации, который позволяет автоматически получать размеченные наборы данных в результат процесса кластеризации.

6. Разработан метод представления объектов, описанных численными признаками в виде матрицы типа объект-признак интуиционистскими нечеткими множествами. При помощи разработанных алгоритмов кластеризации продемонстрированно, что представленные в таком виде объекты можно корректно кластеризовать с учетом специфика работы конкретного алгоритма кластеризации.

Рекомендации по практическому использованию результатов

1. Полученные результаты формируют теоретическую и практическую базу для разработки ПО компьютерных систем для решения задач обработки данных в различных областях, в том числе они дают дополнительные возможности по интерпретации данных в медицине, экономике и при разработки автоматизированных военных систем.

2. Разработанные методы и алгоритмы могут применяться в автоматизированных системах контроля и поддержки принятия решения в условиях наличия неопределенности в исходных данных.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1. Viattchenin, D.A., Shyrai S. A relational heuristic algorithm of possibilistic clustering based on intuitionistic fuzzy sets / D.A. Viattchenin, S. Shyrai // International Hournal of Intelligent Systems and Applications (accepted)

2. Viattchenin, D.A., Shyrai S. Intuitionistic Heuristic Prototype-based Algorithm of Possibilistic Clustering. / D.A. Viattchenin, S. Shyrai // Communications on Applied Electronics = 2015 – Vol. 6 – P. 1 – 11.

3. Viattchenin, D.A., Shyrai S. Clustering Intuitionistic Fuzzy Data. Detection of an Unknown number of Intuitionistic Fuzzy Clusters in allotment / D.A. Viattchenin, S. Shyrai // Zhilina Internationl Conference (accepted).

4. Viattchenin, D.A., Shyrai S. Labeling procedure for results interpretation of heuristic possibilistic clustering of intuitionistic fuzzy data / D.A. Viattchenin, S. Shyrai // Proceeding of the 1st International Conference TACSIT 2015 – 2015 – Severodvenetsk – P. 17 – 21.