АКАДЕМИЯ НАУК РЕСПУБЛИКИ УЗБЕКИСТАН
ИНСТИТУТ МАТЕМАТИКИ И ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

На правах рукописи
УДК 519.95

Давронов Рифкат Рахимович

МИНИМАКСНЫЙ ПОДХОД К СТРУКТУРНОМУ АНАЛИЗУ ДАННЫХ МЕТОДОМ ЯДЕРНОЙ КЛАСТЕРИЗАЦИИ

05.13.01 - Системный анализ, управление и обработка информации

А В Т О Р Е Ф Е Р А Т

диссертации на соискание ученой степени
кандидата технических наук

Ташкент – 2012

Работа выполнена в Институте математики и информационных технологий Академии наук Республики Узбекистан

Научный руководитель	доктор технических наук, профессор **Адылова Фатима Туйчиевна**
Официальные оппоненты:	доктор технических наук, профессор **Нишанов Акрам Хасанович**
	кандидат технических наук **Умеров Хикмет Усниевич**
Ведущая организация	Национальный университет Узбекистана

Защита диссертации состоится «____» _____ 2012 г. в ___ часов на заседании специализированного совета Д. 015.17.02 при Институте математики и информационных технологий АН РУз по адресу: 100125, г. Ташкент, ул. Дурмон йули, 29.

С диссертацией можно ознакомиться в библиотеке Института математики и информационных технологий АН РУз.

Автореферат разослан «___» _____ 2012 г.

Ученый секретарь
специализированного совета Исмаилов М.А.

ОБЩАЯ ХАРАКТЕРИСТИКА ДИССЕРТАЦИИ

Актуальность работы. Усложнение объектов, изучаемых в последние годы наукой об управлении, приводит к появлению специфических задач, связанных с обработкой больших массивов информации. В связи с этим ощущается острая нужда в соответствующих методах. Сегодня подходы к решению задач обработки больших массивов данных уже сложились. Основное в этих подходах - стремление выявить неоднородность информации и разбить массивы на некоторые подмассивы, т.е. так или иначе сжать или упростить информацию. При этом используются сведения о связи между изучаемыми объектами (или параметрами объекта), различные матрицы связи (например, матрицы корреляций между параметрами объекта) и т.д. Другим подходом является попытка сокращения всякого рода переборов, характерных для методов целочисленного программирования.

Две главные проблемы обработки данных - точность и скорость при больших размерностях обрабатываемых массивов информации могут быть решены разработкой точных и быстрых алгоритмов, реализуемых на базе новых информационно-коммуникационных технологий. Настоящее исследование и посвящено решению этих двух проблем. Предлагается ядерный метод кластеризации в рамках структурного или интеллектуального анализа данных (Дата Майнинг) и оцениваются возможности его работы в распределенной среде.

Действительно, современные технологии структурного анализа данных, или Дата Майнинг, связаны с новым витком в развитии средств и методов обработки данных. Цель этого подхода состоит в выявлении скрытых правил и закономерностей в наборах данных, поскольку мозг человека не может выявить более двух-трех взаимосвязей даже в небольших выборках. Математическая статистика же зачастую оказывается эффективной лишь для проверки гипотез, уже полученных Дата Майнинг, причем формулировка гипотез выполняется автоматически в процессе вычислительного эксперимента. Сегодня эта технология широко применяется в бизнесе (банковское дело, страхование, торговля), телекоммуникациях. Зарубежный опыт показывает, что использование Дата Майнинг может дать экономический эффект, во много (10-70) раз превышающий первоначальные затраты.

Но особенно остро проблема Дата Майнинг стоит в медицине, молекулярной генетике и генной инженерии. Там конкретно ставится проблема обнаружения закономерностей в экспериментальных (вычислительная биология) и клинических (формализация правил диагностики и прогноза исхода заболевания) данных, в анализе сложных химических соединений, описание которых включает сотни и тысячи структурных элементов и их связей (вычислительная химия), в обработке изображений.

Таким образом, актуальность разработки методов структурного анализа данных заключается в их применении для анализа многомерных, неоднородных и нестационарных данных в системах, закономерности функционирования которых не могут быть описаны в виде статистических или аналитических моделей.

Сегодня в республике интенсивно развиваются те направления медицины, биологии и химии, где актуально использование технологии Дата Майнинг была показана выше. Исследования в этих центрах требуют развертывания так называемых виртуальных (*in silico*) лабораторий для обработки данных, где ключевую позицию занимают методы Дата Майнинг, являющиеся предметом данной диссертации. Виртуальные лаборатории - мировая тенденция в науке, если учесть возможности их реализации на ГРИД - технологиях, освоенных в Институте математики и информационных технологий АН РУз. Поэтому тема данной диссертации имеет большой потенциал для перспективы проведения НИР в республике на уровне современных требований науки и практики.

Степень изученности проблемы. Существует пять стандартных типов закономерностей, которые позволяют выявлять методы Дата Майнинг,- ассоциация, последовательность, классификация, кластеризация, прогнозирование. Настоящая диссертационная работа посвящена разработке методов и программных средств, реализующих новый подход к слоистой или ядерной кластеризации. Областью приложения результатов работы, определяющих ее актуальность, являются молекулярная биомедицина и обработка изображений.

Практические результаты применения традиционных алгоритмов классификации показывают, что, во-первых, получаемые классы неодинаковы между собой по средней плотности объектов, во-вторых, объекты внутри классов размещаются неоднородно - имеются относительно «густые» и более «разреженные» области, причем некоторые объекты весьма удалены от всех остальных и выпадают в обособленные классы при увеличении числа классов. В связи с этим возникла задача выделения из заданного множества взаимосвязанных элементов подмножества в определенном смысле «наиболее изолированных» элементов. Для этого было необходимо разработать метод, который, во-первых, обеспечивает глобальный экстремум соответствующего функционала, и, во-вторых, порождает классификацию с заранее не заданным числом классов. Эта постановка задачи является новой ступенью в развитии структурного анализа данных, основу которого заложили М.А.Айзерман, Э.М.Браверман, И.Б.Мучник, А.А.Дорофеюк, С.А.Айвазян, Е.Н.Кузнецов, А.Малишевский, Kempner Y., Pevsner J Dhillon I.,Modha D.и др., а также представители узбекской школы - М.М.Камилов, З.Т.Адылова, Ф.Т.Адылова, А.Х.Нишанов, Н.А.Игнатьев. Развиваемый в работе подход теоретически обоснован в

работах J.Mullat, B.Mirkin, I.Muchnik, T.Le, C.Kulikowski (1995-2011), реализованные приложения которого показали его перспективность.

Связь диссертационной работы с тематическими планами НИР. Диссертационная работа выполнялась в рамках темы НИР лаборатории медицинской информатики Института математики и информационных технологий АН РУз ФА-А17-Ф011: «Узбекистан- грид: инсталляция Грид-узла, адаптация и разработка программных и технических приложений в медицине, защите данных» 2009-2011 гг.

Цель исследования. Для систем поддержки принятия управленческих решений разработка модели обработки информации на основе структурного объединения в слоистые кластеры с использованием оценки связей элемента с подмножеством, которые вычисляются на структурных данных, таких как графы, или изображения. Модель основана на максимизации функции плотности, определенной как минимальное значение связей между множеством и его элементами, поэтому данный подход получил название минимаксного. Центральным для данной работы является понятие монотонной системы.

Задачи исследования.

- разработать параметрическую модель ядерной кластеризации как основу нового подхода структурного анализа информации для систем принятия управленческих решений;
- разработать новые алгоритмы модели ядерной кластеризации, включая параллельные версии, вариации используемой базовой монотонной системы, свободных параметров, метрик близости;
- решить алгоритмом ядерной кластеризации три различных по исходным данным задачи,- задачу диагонализации матриц связи, множественного выравнивания, сегментации изображения с тем, чтобы сравнить качество полученных результатов с известными решениями в указанных областях и эффективность работы алгоритма по скорости и требуемым ресурсам.

Объект и предмет исследования. Объект исследования – графы связи, отображающие результаты экспериментальных или иных исследований сложных систем из области биологии, медицины, обработки изображения. Предмет исследования – выявление скрытых закономерностей крупномасштабных массивов экспериментальных данных методом ядерной (слоистой) кластеризации.

Методы исследования. Используются методы: вариационного анализа данных, кластер-анализа, теории графов, объектно-ориентированного программирования.

Гипотеза исследования. В соответствии с гипотезой о компактности классов, кластеризация была ограничена поиском локального экстремума. Поэтому в данной диссертации исследуется процедура кластеризации, обеспечивающая глобальный экстремум в классе «жадных» алгоритмов.

Основные положения, выносимые на защиту:

- Разработка минимаксного метода структурного анализа данных;
- Разработка параметрической модели ядерной кластеризации;
- Разработка модели множественного выравнивания последовательностей на базе ядерной кластеризации;
- Разработка новых алгоритмов и программ, демонстрация эффективности их работы (множественное выравнивание, обработка изображений).

Научная новизна:

- Разработана параметрическая модель ядерной кластеризации и дана оценка ее эффективности;
- Разработаны параллельные алгоритмы диагонализации матриц связи и получены их сравнительные характеристики с алгоритмом ядерной кластеризации;
- Разработан алгоритм и программа множественного выравнивания белковых последовательностей методом ядерной кластеризации и получена ее оценка в сравнении с известными пакетами;
- Разработаны алгоритм и программа сегментации изображений методом ядерной кластеризации и получена оценка ее эффективности;
- Разработаны параллельные версии алгоритмов ядерной кластеризации.

Научная и практическая значимость результатов исследования. Разработанные программные средства с учетом проведенной оценки эффективности их применения могут быть использованы в научных исследованиях разного профиля, имеющих одну цель, - выявление скрытых закономерностей в больших массивах данных биомедицинской информации и обработка изображений, необходимых для систем поддержки принятия решений. Программные средства лицензированы и могут быть использованы в тех задачах бизнеса, систем связи, актуальность которых оговорена выше.

Реализация результатов. Результаты диссертационной работы в виде комплекса программ внедрены на кафедре информационных систем Муромского института Владимирского государственного университета и в Республиканском научном центре экстренной медицинской помощи Минздрава РУз.

Апробация работы. Основные результаты диссертационной работы доложены на: V International Conference “Distributed Computing and Grid Technologies in Science and Education” (Дубна, 2010); XXIII International Symposium on Nuclear Electronics and Computing (NEC’2011) (Болгария, Дубна, 2011); International Asian Workshop “ Problems of Complex Systems Optimization” (Ташкент, 2011). Основное содержание диссертации докладывалось и обсуждалось на научных семинарах лаборатории «Медицинская информатика» Института математики и информационных технологий АН РУз, на научном семинаре при специализированном совете Д.015.17.02 при ИМИТ АН РУз (2012).

Опубликованность результатов. По теме диссертационной работы опубликованы 9 научных статей и тезисов, среди них 6 журнальных статей. Получены 3 свидетельства на программное обеспечение, выданных Агенством по интеллектуальной собственности Республики Узбекистан (№ DGU 01998, 2010; № DGU 02126, 2011; № DGU 02239, 2011).

Структура и объем диссертации. Диссертация содержит 132 страницы, состоит из введения, трех глав, заключения, списка использованной литературы из 110 наименований и приложения.

ОСНОВНОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ

В первой главе работы обосновывается предлагаемый минимаксный подход к решению задач кластеризации в области биоинформатики, которая представляет собой междисциплинарные исследования по информационным технологиям, молекулярной биологии, химии, физике. С этой целью из сравнительного анализа методов кластеризации выводится основная теоретическая посылка диссертации: структурный анализ больших и неоднородных массивов данных эффективен в рамках минимаксного подхода (оптимизация критерия качества) методом ядерной кластеризации, обеспечивающего глобальный экстремум критерию, заданному в классе монотонных функций. Следует подчеркнуть, что базовым допущением работы является представление данных в виде графа.

Поскольку нам впервые приходится иметь дело с данными молекулярного уровня описания, то выявляются особенности обработки таких данных а именно, формулируется центральная проблема обработки данных в биоинформатике - поиск сходства. Конкретно, в данной работе с использованием предложенного методе решаются следующие задачи биоинформатики: множественное выравнивание последовательностей, анализ экспрессии гена и обработка изображений.

Из сказанного выше понятно, что до настоящего времени уже были разработаны методы и программные средства решения рассматриваемых задач. Краткий обзор наиболее известных пакетов программ в биоинформатике и, в частности в кластеризации, приведены в главе для того, чтобы подтвердить правомочность и актуальность метода ядерной кластеризации по сравнению с динамическим программированием и скрытыми цепями Маркова, до сегодняшнего дня составляющими основу почти всего используемого программного обеспечения в биоинформатике. Тем не менее остаются нерешенными проблемы сложности используемых алгоритмов и потребности в мощных вычислительных ресурсах. Именно эти два вопроса «закрываются» в диссертации предложением метода ядерной кластеризации и распараллеливанием вычислений.

Тема диссертация касается направления Дата-Грид и проектов по распределенному структурному анализу, в частности решения задач кластеризации и обработки изображений на гриде. Поэтому в

заключительной части главы приведен краткий обзор архитектуры некоторых проектов, направленных на внедрение сервисов Дата Майнинг и извлечение знаний.

Причиной появления технологии распределенного Дата Майнинг (Distributed Data Mining) является большой объем распределенных по своей сути данных в современных системах управления. «Слив» их на один сайт для обработки ведет к перегрузке каналов и усложнению обрабатывающих алгоритмов. Построение единой БД для Дата Майнинг не в режиме распределения данных непрактично или просто невозможно. Распределенный Дата Майнинг есть среда, где каждый сайт имеет свой источник данных и алгоритмы структурного анализа, с помощью которых строятся локальные модели. Каждая такая модель представляет знания, полученные из локальных данных, но без возможностей обобщения. Поэтому сайты связаны с целью отслеживания глобальной информации, и здесь коммуникации являются самым узким местом. Выходить из этой ситуации предлагается алгоритмами кластеризации и классификации. Проблема эта актуальна в реальных условиях работы через созданный грид-портал в ИМИТ АН РУз, поэтому сформулированная постановка задачи распределенного Дата Майнинг в гриде является конкретным направлением дальнейших исследований.

Вторая глава работы полностью посвящена изложению теоретических основ минимаксного подхода к структурному анализу данных. Рассматривается метод структурного объединения в кластеры на основе оценки связей элемента с подмножеством, которые могут быть вычислены на структурных данных, таких как графы, или изображения. Метод основан на максимизации функции множества, определенной как минимальное значение связей между множеством и его элементами, и названной функцией плотности. Когда функция связи монотонная, слоистые кластеры могут быть легко найдены экономным алгоритмом. Метод, построенный на этой идее и названный методом ядерной кластеризации, является основой данной работы. Алгоритмы, построенные в идеологии ядерной (слоистой) кластеризации, относятся к классу так называемых «жадных алгоритмов» (greedy algorithms), которые предназначены для решения задач оптимизации, и обычно представляют собой последовательность шагов, на каждом из которых имеется возможность выбора. В жадном алгоритме всегда делается выбор, который является лучшим на этом шаге, и приводит к оптимальному решению глобальной задачи; соответствующая теорема доказана Томасом Х Кормен. Показано, что решение можно преобразовать так, чтобы в нем использовался жадный выбор,- такой подход принят в алгоритме ядерной кластеризации.

Проблема плотного ядра традиционно формализована в терминах функции $F(H)$, определенной на множестве подмножеств H конечного множества I. Функция выделяет подмножества по их «плотности» так, чтобы

подмножество H^*, на котором $F(H)$ достигает максимума, считается максимально плотным ядром в I. Как правило, решение проблемы максимизации может потребовать перебора экспоненциального числа подмножеств относительно количества элементов I (NP-проблема), но когда $F(H)$ относительно простая, это может быть сделано алгоритмом полиномиального типа. Вводится специальная структура для определения функции плотности как интегральной характеристики, выявляющей связи между подмножествами и их элементами. Для любого непустого подмножества $H \subseteq I$ функция плотности $F_\pi(H)$ является минимумом монотонной функции связи $\pi(i, H)$, $i \in H$.

Вводится и исследуется понятие слоистого кластера, который не только имеет максимально плотное ядро I, но также и вложенную цепь его оболочек, удельные веса которых монотонно уменьшаются с ростом оболочек. Определяется образ функции плотности как подмножество, которое строго отделено от остальных: его оценка, согласно функции множества, уменьшается, если какие-нибудь элементы добавлены, или если некоторые из его элементов удалены. Множество образов, которые являются подобными и вложенными, называется слоистой группой (кластером), если наименьший образ доставляет глобальный максимум функции плотности. В этом суть минимаксного подхода в структурном анализе данных. Существование слоистого кластера доказано для функций плотности в классе монотонных функций связи элемента и множества, на базе которых предложен экономный (жадный) алгоритм для последовательного его нахождения.

Метод ядерной кластеризации графа. Кластеризация графа означает его разделение на подграфы с сильно связанными узлами внутри каждого подграфа и слабыми связями между узлами разных подграфов. В главе предлагается новый метод объединения в кластеры, применимый к взвешенным или к невзвешенным графам, в которых каждый кластер состоит из очень плотной основной области, окруженной областью с более низкой плотностью. Метод очень эффективен и робастен в идентификации узлов, принадлежащих плотным ядрам кластеров. Множество узлов графа затем разделяется на группы, каждая из которых является представителем одного кластера. Эти группы, в конечном счете, расширяются до полных кластеров, покрывающих все узлы графа. Рассмотрим ненаправленный граф близости $G = (V, E, W)$, где V - узлы, E - ребра, W – матрица инциденций, w_{ij} – вес ребра между узлами i и j. В графах близости V представляет ряд объектов данных, $w_{ij} \geq 0$ - сходство объектов i и j. Более высокое значение w_{ij} отражает более высокую степень близости. Граф близости - естественное представление для набора данных, когда попарные отношения между объектами данных представлены в явном виде. Если данные представлены в пространстве признаков, граф близости может быть получен вычислением попарных коэффициентов сходства, использующих Евклидову или другие

метрики, как меру расстояний между точками в пространстве признаков. Это позволяет анализировать матрицу методами, основанными на подходах к анализу графов.

Метод ядерной кластеризации обнаруживает плотные подграфы в невзвешенном графе, или сильно «тяжелые» подграфы во взвешенном графе, с использованием утверждения, что сумма весов ребер в подграфах выше, чем сумма весов ребер, соединяющих различные подграфы. Таким образом, применяя метод кластеризации к графу близости, получаем ряд подграфов, где каждый подграф содержит группу подобных объектов, которые являются несходными с объектами из других подграфов. Известны различные подходы к кластеризации графа - максимальные клики, минимум порога, случайный поиск, спектральная кластеризация и т.д. В данной работе предлагается метод, который находит кластеры графа, сначала определяя узлы в самых плотных областях кластеров, затем - ядра кластеров и, наконец, приписывает все другие узлы к самому близкому кластеру.

Предполагаем, что каждый кластер входного графа имеет область высокой плотности (ядро), окруженную более разреженными областями (не-ядрами). Узлы в ядрах кластеров обозначены как «основные узлы», набор основных узлов - как «основное множество», а подграф, состоящий из основных узлов,- как «основной граф». Для графа, удовлетворяющего этому основному предположению, процедура, используемая в слоистом методе кластеризации, имеет два полезных свойства: 1) наименее непохожий узел удаляется при каждой итерации, поэтому узлы графа упорядочены таким образом, что узлы в более редких областях помещены перед узлами в более плотных областях. Другими словами, процедура идет от внешних к внутренним слоям графа; 2) снижение значений *плотности* значительно, когда процедура удаляет узлы, принадлежащие ядрам кластера графа. Мы использовали эти свойства, чтобы создать метод кластеризации графов, в котором у каждого кластера есть плотное ядро. Для каждого узла i из $H \subseteq V$ локальная плотность определяется как $d(i,H) = (\sum_{j \in H} w_{ij}) / |H|$. Узел с минимальной плотностью в H называется самым слабым узлом $H : \arg\min_{i \in H} d(i,H)$. Определяем минимальную плотность H как $D(H) = \min_{i \in H} d(i,H)$), чтобы измерить плотность самого слабого узла H.

Рассмотрим процедуру, которая итеративно вычисляет $D(G)$ и удаляет самые слабые узлы из G. Анализируя изменение значения минимальной плотности D, можем идентифицировать основные узлы, расположенные в плотных ядрах кластеров. Таким образом, если самый слабый узел будет в разреженной области, то значение D увеличится при удалении этого узла. Другими словами, следующий самый слабый узел, который будет удален, будет в области с более высокой плотностью. С другой стороны, если удаление самого слабого узла вызывает существенное понижение D, то этот узел сильно связан с рядом более сильных узлов в области высокой

плотности. Это - потенциально основной узел, потому что его удаление сильно уменьшает плотность узлов вокруг него.

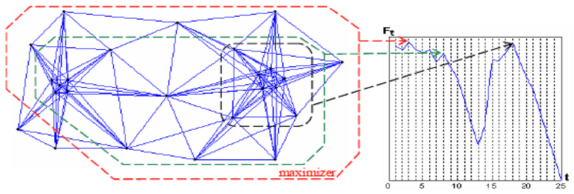

Рис.1. Граф G_1 с двумя кластерами, которые имеют два плотных ядра, и соответствующая кривая функции F_t

На рис.1 показан пример взвешенного графа G_1 из 25 узлов и 126 ребер с двумя плотными ядрами. Веса ребер обратно пропорциональны нормализованным длинам ребер и вычисляются как $w_{ij} = 1 - d_{ij} / \max_{x,y \in V} d_{xy}$, где d_{xy} - евклидово расстояние между x и y на плоскости. Непрерывно удаляя самые слабые узлы, мы постепенно идем в первое ядро группы слева, в то время как значения D имеют тенденцию увеличиваться. В некоторый момент при удалении узлов из ядра (t=8→13), D резко падает, потому что ядро кластера разрушается. После того как первый кластер полностью удален, значение D растет, но потом понижается снова, когда разрушается второе ядро кластера (t=21→25). Таким образом, уменьшение D_t выявляют основные узлы ядер кластера, что показано на рис.2.

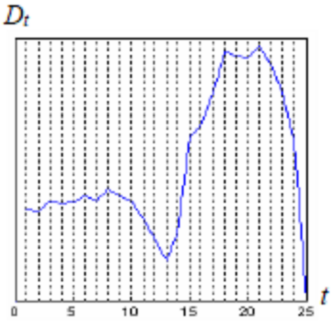

Рис.2. Изменение минимальной плотности D_t графа G_1.

Из сравнения рис. 1 и 2 видно, что удобнее использовать функцию минимальной плотности D_t. Алгоритм ядерной кластеризации состоит из четырех основных шагов.

Шаг 1. Вычисление последовательности изменения плотности. Это - итеративная процедура, которая вычисляет последовательность изменения плотности. При каждой итерации t вычисляем минимальную плотность D_t и набор самых слабых узлов M_t. Потом M_t удаляется из графа. Когда граф становится пустым, процедура возвращает последовательности D_ts и M_ts, которые содержат значения D и узлы, относящиеся к каждому значению D соответственно.

Шаг 2. Идентификация набора основных узлов. Основные узлы определяются, основываясь на последовательностях D_ts и M_ts, полученных на шаге 1 и двух параметрах,- $\alpha \in [0,1)$ и $\beta \in N$. Элементы M_t отбираются в качестве основных узлов, если соответствующие D_t удовлетворяют двум условиям:

(1) его степень уменьшения R, больше чем α, т.е. $R_t = (D_t - D_{t+1}) / D_t > \alpha$;

(2) $\exists k \in N$, что $D_t \in \{D_{k+1}, D_{k+2}..., D_{k+\beta}\}$, где $D_{k+1}, D_{k+2}..., D_{k+\beta}$ являются β последовательными значениями плотности, которые также удовлетворяют первому условию.

Шаг 3. Разделение основного набора на ядра кластеров. Набор основных узлов, идентифицированных на шаге 2, разделяется на группы, представляющие ядра кластеров. Число полученных групп автоматически определяет число кластеров. Поскольку на шаге 3 фактически решается проблема кластеризации, то любой метод кластеризации графа может использоваться для того, чтобы разделить основной набор. Отметим, что разделение ядер группы в основном наборе намного проще, чем кластеризация исходного графа.

Шаг 4. Расширение основных групп в полные кластеры. Этот заключительный шаг аналогичен решению проблемы классификации с учителем, где классифицируют на обучающей выборке, состоящей из известных классов данных. Здесь мы используем в своих интересах свойство шага 1 - движение от внешнего к внутренним слоям, т.е. по узлам в последовательности M_t, которые ранжированы от разреженных до плотных областей. При просмотре последовательности неосновные узлы приписываются к самым похожим кластерам, расширяя их ядра до полного кластера. Каждый неосновной узел n M_t идет к кластеру $C^* = \arg\max_C S(n,C)$, где $S(n,C)$ измеряет степень подобия узла n с C.

Для взвешенных графов имеем:

(1) $S(n,C) = average_{i \in C, w_{ni} > 0} w_{ni}$ или

(2) $S(n,C) = \max_{i \in C,\ w_{ni} > 0} w_{ni}$

Для невзвешенных графов используем:

(1) $S(n,C) = average_{i \in C} w_{ni}$ или

(2) $S(n,C) = \sum_{i \in C} w_{ni}$.

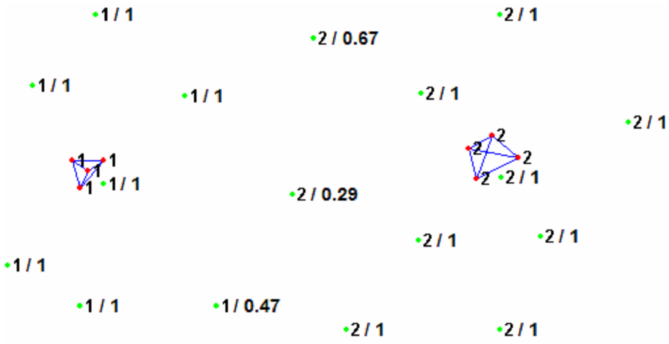

Рис.3. Результаты кластеризации графа G_1

На рис.3 приведены результаты кластеризации графа G_1, данного на рис.1 при α=0.4 и β=2. Черные жирные точки показывают основные узлы, образующие кластер.

Исследование параметров алгоритма ядерной кластеризации и времени его выполнения. Влияние параметров α и β на множество основных узлов имеет место, поскольку число узлов конечно, пространство параметров дискретно, и поэтому число возможных вариаций параметров конечно.

Если α равно константе, а β увеличивать с 1 до 3, то множество основных узлов уменьшается. С увеличением α и β одновременно сжимается множество ядер, и основные узлы приходится извлекать из глубины ядер кластера. Для большинства вариантов параметров получены похожие результаты кластеризации на 2 кластера.

Время выполнения. Если использовать матрицу смежности графа, то алгоритм имеет такую же сложность, как процедура поиска максимума. При $\pi(i,H)=\sum_{j\in H} w_{ij}$ и динамической памяти Фибоначчи шаг 1 реализуется за время $O(|E|+|V|\log|V|)$. На шаге 3 поиск в глубину и в ширину для нахождения связанных компонент занимает время $O(V_c+E_c)$, где V_c и E_c - множества узлов и ребер основного графа. Время выполнения шага 4 равно $O(|E|)$. Общее время выполнения алгоритма, следовательно, затрачивается на шаге 1 и варьирует от $O(|V|\log|V|$ для разреженного графа до $O(|V|^2)$ для графа насыщенного (плотного).

Сегментация изображения методом ядерной кластеризации. Изображение записывается графом близости, в котором узлы соответствуют пикселям изображения, а ребра отражают попарную близость между пикселями. Веса ребер вычислены по формуле функции близости, основанной на свойствах соответствующих пикселей: местоположение, яркость и цвет. В такой постановке сегментация изображения может быть решена методами кластеризации графа.

Поскольку мы хотим группировать соседние пиксели, у которых есть похожая интенсивность/цвет, вес ребра графа вычисляется как функция правдоподобия, основанная на местоположении и интенсивности/цвете

соседних пикселей. Для полутоновых изображений используем функцию веса вида:

$$w_{ij} = \begin{cases} e^{-\left(\frac{I(i)-I(j)}{s_I}\right)^2 - \left(\frac{dist(i,j)}{s_d}\right)^2}, & dist\ i\ j\ < r \\ 0 \quad - \text{ противном случае.} \end{cases} \qquad (1)$$

Здесь $I(i), I(j) \in [0,1]$ - интенсивность пикселей i, j соответственно, а $dist(i, j)$ является евклидовым расстоянием между ними. Для цветных изображений разница интенсивностей в (2) заменяется нормализованным евклидовым расстоянием между цветными компонентами пикселей:

$$w_{ij} = \begin{cases} e^{-\left(\frac{\|C(i)-C(j)\|_2}{\sqrt{3}s_I}\right)^2 - \left(\frac{dist(i,j)}{s_d}\right)^2}, & dist\ i\ j\ < r \\ 0 \quad - \text{ противном случае,} \end{cases} \qquad (2)$$

где $C(i)$ является вектором трех признаков, а именно, красным, зеленым и синим цветными компонентами пикселя i, поэтому $\|C(i) - C(j)\|_2 \in \left[0, \sqrt{3}\right]$ - ребро, связывающее два узла, только если расстояние между соответствующими пикселями будет меньше, чем r пикселей; следовательно, каждый узел соединен приблизительно с πr^2 другими узлами. Граф близости очень разрежен, поскольку $\pi r^2 << |V|$. При использовании функций веса (1) сильные ребра соединяют узлы, соответствующие пиксели которых близки друг к другу и имеют близкую интенсивность/цвет. Поэтому пиксели в сегменте с гомогенной интенсивностью/цветом имеют сильно связанные узлы в графе близости. С другой стороны, у пикселей на границах сегментов часто есть соседние пиксели с несходной интенсивностью/цветом, в этом случае соответствующие узлы обычно слабо соединены друг с другом.

Хороший результат метода сегментации на графе зависит от того, как сегменты изображения преобразуются в хорошо отделяемые кластеры графа близости. Поэтому здесь важны параметры настройки σ_I, σ_d и r, поскольку они определяют, как изображения будут представлены графом близости. Изучение и вычислительный эксперимент на различных изображениях показали, что имеет смысл устанавливать $\sigma_I = 0.07, \sigma_d = 8, r = 11$ для изображения размером меньше, чем 200×300.

Алгоритм сегментации, основанный на кластеризации его графа близости, реализуется пятью процедурами: граф близости G для входного изображения I; последовательность изменения плотности для G; множество основных пикселей; группы на множестве основных пикселей; сегментация изображения. Алгоритм прост и очень быстр, а поскольку используются

пиксели в ядрах сегментов, которые устойчивы и надежны, то метод устойчив к искажениям.

Множественное выравнивание белковых последовательностей. Известно только одно исключение из многих различных процедур множественного выравнивания - DIALIGN, который основан на абсолютно другом по сравнению с известными принципами множественного выравнивания. Это чисто комбинаторный принцип, в котором правильное множественного выравнивание определено как “непротиворечивая” система блоков. Целая система статистически оценена функцией, чье экстремальное значение связано с правильным выравниванием. Чтобы найти локальный максимум функции (точное глобальное решение является NP- трудной проблемой), выделяются основные комбинаторные структуры DIALIGN и предлагается новый способ их обработки в рамках метода ядерной кластеризации. Новизна здесь состоит в том, что результаты множественного выравнивания получают некоторую “внутреннюю структуру”, которая полезна в интерпретации результатов выравнивания.

Поясним концепцию множественного выравнивания, исходя из следующего простого наблюдения. Предлагается рассматривать вложенную структуру множества существенных хитов последовательностей, образующих гомологичные группы. Ядро группы последовательностей белка - множество сегментов его индивидуальных последовательностей с большой силой связи, которая уменьшается по мере нарастания числа слоев.

Третья глава работы содержит результаты вычислительных экспериментов, которые проводились таким образом, чтобы их результаты были сопоставимы с уже известными решениями. Это позволило дать оценку сильным и слабым сторонам минимаксного подхода к анализу данных и определить пути дальнейших исследований. Из традиционных методов кластерного анализа был выбран подход лингвистического анализа данных, поскольку в этой школе «вырос» и сегодня развивается метод ядерной кластеризации.

Были проведены эксперименты на двух известных матрицах связи, которые тестировались алгоритмами лингвистического анализа данных, - матрицы К.Холзингера и В.Небылицина.

Матрица корреляций К.Холзингера отражает результаты 24 психологических тестов, проверявшихся на 145 школьниках, и считается сложной. Для задания начальной группировки последовательно применялись два варианта известных алгоритмов корреляционных плеяд и «Объединение», а затем, на третьем шаге - вариационные алгоритмы, работающие по трем функционалам F_1 , F_2 , F_3.

В таблице 1 приведена конечная группировка на пять классов, полученная на исходной матрице К.Холзингера работы алгоритмов, описанных выше.

Таблица 1

Конечная группировка на пять классов

№	Функционалы		
	F_1	F_2	F_3
1	3,11	1,3,11,13	3,11
2	4,10	2,4,10,14	4,10
3	6,12	12,15,16	6,12
4	1,2,7,8,13-24,	5,6,7,8,9	5,7,8,9
5	20,23	17,18,19,20,21,22,23,24	1,2,13-24

Конечный результат группировки объектов сильно зависит от начального разбиения; группировки, полученные на функционалах F_1, F_2, F_3 после предварительной группировки методом корреляционных плеяд и алгоритмом «Объединение», далеки от идеального разбиения. Результаты по методу ядерной кластеризации (два варианта) не показали значительного улучшения классификации по сравнению с лингвистическим анализом, хотя были эффективнее алгоритма k-средних.

Вторая задача, на которой проводили испытание, является более простой по сравнению с матрицей К.Холзингера. Это матрица связи размерностью 11*11, которую необходимо разбить на четыре класса. Решения по всем трем функционалам F_1, F_2, F_3 совпали с идеальной группировкой; матрица быстро обрабатывается методом ядерной кластеризации, давая идеальное разбиение при $\pi(i, H) = \max_{j \in H} w_{ij}$.

Кроме того, проведен тест на задаче экспрессии гена, где проводилось сравнение наших решений при разных значениях параметров алгоритма и решений Le Thang и I.B.Muchnik из Университета Rutgers (США). Основной вывод состоит в том, что проблема задания параметров α и β алгоритма ядерной кластеризации является важной, поскольку они формально не определены, и их оптимизация выливается в человеко-машинную процедуру с предварительным заданием граничных значений интервалов изменений, что представляет собой сложную комбинаторную задачу.

В заключительной части главы приводятся результаты вычислительного эксперимента (ВЭ) с алгоритмом сегментации. Как показал ВЭ, для изображений, которые содержат смесь сегментов с различными размерами, увеличивая или уменьшая α, можно получить грубую или детальную сегментацию. Алгоритм исполнен также в параллельном варианте и сравнивается с нормализованным методом вырезки на графе, который является известным методом кластеризации графа. Обнаружение точной минимальной нормализованной вырезки - NP-трудная задача, поэтому получают приближенное решение. Сравнительный вычислительный

эксперимент на одном полигоне (графы близости) показал преимущество ядерной кластеризации, проявляющееся в скорости работы алгоритма. В табл. 2 даны времена выполнения сегментации на различных графах близости сравниваемых трех методов с использованием PC CPU of Core 2 Duo 3GHz и 2GBRAM .

Таблица 2

Времена выполнения нормализованных методов вырезки и ядерного метода кластеризации графа

Размер графа близости		Нормализованная вырезка (с)	Ядерный (с)		k-средних (с)
#узлы	#ребра		обычный	параллель	
$30*10^3$	$4.7*10^6$	7	0.3	0.2	7
$45*10^3$	$7.1*10^6$	12	0.4	0.3	12
$60*10^3$	$10.7*10^6$	22	0.7	0.5	25
$75*10^3$	$12.2*10^6$	50	0.9	0.65	40
$90*10^3$	$14.8*10^6$	91	1.0	0.8	80
$105*10^3$	$17.6*10^6$	116	1.2	0.93	100
$120*10^3$	$18.7*10^6$	-	1.5	1.1	-
$500*10^3$	$50*10^6$	-	4.2	3.2	-
$1000*10^3$	$70*10^6$	-	5.5	4.1	-

Для случаев, когда графы содержат больше, чем $105*10^3$ узлов и $18*10^6$ ребер, нормализованный метод вырезки не работает из-за ошибки «недостаточно памяти». Кроме того, преимущество ядерного метода состоит в том, что изменив β или α, можем легко получить новый результат перезапуском быстро исполняемых шагов 3, 4 и 5. Напротив, для нормализованного метода вырезки изменение k приведет к необходимости повторно вычислять вырезку с самого начала.

Еще раз подчеркнем, в дополнение к скорости и надежности, преимущество метода состоит в том, что параметры могут настраиваться на обработку определенного изображения.

Для того чтобы проверить метод ядерной кластеризации в множественном выравнивании, была обработана группа из пяти белковых последовательностей. Построение множественного выравнивания данным методом производилось в пять шагов: 1. выделение диагоналей (сервис DIALIGN); 2. построение матрицы весов, где объектами являются выделенные диагонали; 3. выделение ядер; 4. наращивание ядер. 5. построение множественного выравнивания с учетом полученной кластеризации.

Результат нашего вычислительного эксперимента сравнивался с результатами, полученными при помощи уже известных сервисов на основе трех принятых критериев CR1, CR2, CR3 (табл.3).

Таблица 3

Сравнение выравниваний, полученных различными сервисами

Сервис	CR1		CR2		CR3
	2	>2	2	>2	
ClustalW	18	14	10	36	39
DIALIGN	22	16	8	37	53
Ядерная кластеризация	17	20	15	34	71

Из табл. 3 следует, что ядерная кластеризация дает лучшие результаты, чем DIALIGN по двум критериям, но проигрывает по третьему критерию. ClustalW не обеспечивает хорошее сходство, а DIALIGN выигрывает за счет итеративности процедуры.

Таким образом, корректное множественное выравнивание должно быть аналогично процессу кристаллизации, основанному на поиске «слоистого кластера» структуры выровненного множества последовательностей.

В приложении приведены документы, подтверждающие практическое использование результатов диссертационной работы, а также копии трех свидетельств об официальной регистрации программы в Агентстве по интеллектуальной собственности Республики Узбекистан.

ЗАКЛЮЧЕНИЕ

В результате анализа многомерных, неоднородных и нестационарных данных в системах управления в различных областях, закономерности функционирования которых не могут быть описаны в виде статистических или аналитических моделей, предложены и реализованы новые алгоритмы кластеризации, обеспечивающие достижение глобального максимума.

В процессе работы получены следующие научные и практические результаты:

1. Развит минимаксный подход к структурному анализу данных, реализованный в виде метода ядерной кластеризации. Показана роль кластеризации как ключевой технологии структурного анализа данных в системах поддержки принятия управленческих решений, проведен системный анализ основных подходов. Поскольку большой объем данных требует ресурсов, показано современное решение проблемы структурного анализа на грид - инфраструктуре.

2. Вводится и исследуется понятие ядерного (слоистого) кластера, который не только имеет максимально плотное ядро, но также и вложенную цепь его оболочек, удельные веса которых монотонно уменьшаются с ростом оболочек. Метод эффективен и робастен в идентификации узлов, принадлежащих плотным ядрам кластеров. Теоретически обосновано главное преимущество минимаксного подхода, а именно возможность получать глобально оптимальные решения. Исследовано влияние параметров настройки алгоритма ядерной кластеризации на качество и время его выполнения.

3. Протестированы и сравнены на одном полигоне все предложенные в работе модификации метода ядерной кластеризации с известными методами кластеризации. Ядерный алгоритм дает скорость и точность выше, чем у алгоритмов лингвистического анализа.

4. Качество сегментации изображения, выполненной алгоритмом ядерной кластеризации (последовательная и параллельная версия), сравнивалось с широко используемым методом нормализованной вырезки графа и методом k - средних. Преимущество ядерной кластеризации, в дополнение к выигрышу по скорости, надежности и робастности состоит в том, что параметры могут изменяться в процессе настройки на обработку определенного изображения.

5. Алгоритмом ядерной кластеризации решена задача множественного выравнивания пяти белковых последовательностей, результаты сравнивались по трем критериям с двумя другими алгоритмами (DIALIGN, ClustalW), известными и используемыми в биоинформатике. Ядерная кластеризация дает лучшие результаты, чем DIALIGN по двум критериям, но проигрывает по третьему критерию.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1. Давронов Р.Р., Кузиев Б.Н., Минбоев У.Х. Математическая модель структуризации данных на базе теории монотонных систем // Узб. журнал «Проблемы информатики и энергетики.» – Ташкент, 2008. – № 1. – С.72-76.
2. Адылова Ф.Т., Давронов Р., Кузиев Б.Н., Алгоритмические проблемы биоинформатики и подходы к их решению // Доклады АН РУз. – Ташкент, 2010. – № 3. – С.21–25.
3. Давронов Р. Лингвистический анализ булевых матриц на базе монотонных систем / Вопросы вычислительной и прикладной математики: Сб. науч. тр. – Ташкент, ИМИТ АН РУз , 2010. – вып. 124. – С.128–134.
4. Adilova F.T., Davronov R., Dushanov E. An Effective Algorithm of Coring Clusterization for biodata analysis: comparative studies and potential for Grid Applications // Distributed computing and grid-technologies in scienceand education.–June 28 – 3, 2010. –Dubna, 2010. – P.19.
5. Адылова Ф.Т., Давронов Р.Р., Томскова А.А. Методы искусственного интеллекта (дата майнинг) в анализе биологических данных // Узб. журнал «Проблемы информатики и энергетики.» – Ташкент, 2010. –№ 4. – С.42-44.
6. Давронов Р.Р. Программа выявления структуры экспериментальных данных большого объема методом ядерной кластеризации // Агентстве по интеллектуальной собственности Республики Узбекистан. Свидетельство № DGU 01998. 26.07.2010 г.
7. Давронов Р.Р. Оптимизация принятия решений на основе диагонализации матриц связи методами лингвистического анализа и ядерной кластеризации // Проблемы оптимизации сложных систем. 17–27 октябрь. – Ташкент, 2011. – С.257-263.
8. Tomskova A., Davronov R. GRID-MAS conception: the applications in bioinformatics and telemedicine // XXIII International Symposium on Nuclear Electronics & Computing. 12-19 September. –Bulgaria, 2011. –P.253-257.
9. Давронов Р.Р. Множественное выравнивание белковых последовательностей, основанное на слоистой кластеризации // Кимёвий технология назорат ва бошқарув. – Ташкент, 2011. – №3. –С.85-87.
10. Давронов Р.Р., Подход к сегментации изображения на базе графа близости // Доклады АН РУз. – Ташкент, 2011. – № 6. – С.20–24.
11. Давронов Р.Р. Программа оптимизации выявления структуры данных методом ядерной кластеризации // Агентстве по интеллектуальной собственности Республики Узбекистан. Свидетельство № DGU 02126. 10.01.2011 г.
12. Давронов Р.Р., Томскова А.А. Программа множественного выравнивания последовательностей белка или ДНК при помощи кластеризации методом монотонных систем // Агентстве по интеллектуальной собственности Республики Узбекистан. Свидетельство № DGU 02239. 07.07.2011 г.

Техника фанлари номзоди илмий даражасига талабгор Давронов Рифқат Рахимовичнинг 05.13.01 – Тизимли таҳлил, бошқарув ва ахборотни қайта ишлаш ихтисослиги бўйича «Ядроли синфлаштириш усули билан маълумотларни тузилмали таҳлил қилишда минимаксли ёндашув» мавзусидаги диссертациясининг

Р Е З Ю М Е С И

Таянч сўзлар: маълумотларни структурали таҳлили, яқинлик монотон функцияси, ядроли синфлаш, тасвирларни сегментлаш, оқсил кетма-кетликларини тўпламли текислаш.

Тадқиқод объектлари:яқинлик матрицаси, тасвирлар, графлар.

Ишнинг мақсади:яқинлик монотон функциялари асосида қатламли синфлаш моделини ишлаб чиқиш ва тадбиқ қилиш.

Тадқиқод методлари:дискрет математика, маълумотларни лингвистик таҳлил қилиш, класификациялаш, кластер – таҳлил, тасвирни қайта ишлаш методлари. Яратилган алгоритмлар C++, MATLAB, C# тилларида дастурлаштирилиб қўлланилди.

Олинган натижалар ва уларнинг янгилиги:

– яқинлик монотон функция зичлиги тушунчаси асосида ядроли синфлашнинг параметрли модели ишлаб чиқилган;
– ядроли синфлаш методи билан тасвирни сегментлаш масаласи ечилган, унинг эффективлиги k-ўртача ва нормаллаштириб қирқиб олиш методи билан солиштириб баҳоланган.
– ядроли синфлаш методи билан оқсил кетма-кетликлари тўпламли текислаш процедураси ишлаб чиқилган,унинг эффективлиги CLUSTAL ва DIALIGN пакетлари билан қиёслаб баҳоланган;

Амалий аҳамияти:тадқиқотлар натижасида олинган алгоритмлар ва дастурлар маълумотларни катта массивларини қайта ишлашда, тўпламли текислаш процедурасида ва тасвирларни сегментлашда қўлланилиши мумкин.

Татбиқ этиш даражаси ва иқтисодий самарадорлиги:диссертация иши натижалари Ўзбекистон Республикаси Соғлиқни сақлаш вазирлиги Республика тез тиббий ёрдам кўрсатиш илмий марказида холецистitларнинг турли кўринишларини компьютерли диагностика қилишда, шунингдек Владимир давлат университетининг Муром институти ахборот тизимлари кафедрасида тасвирларни қайта ишлаш масалаларини ечишда ва ўқув жараёнида дастурлар мажмуаси кўринишида фойдаланилди. Тадбиқ этиш самарадорлиги – ижтимоий.

Қўлланиш соҳаси: тадқиқот натижалари маълумотларни интелектуал таҳлил қилишда, биология ва тиббиётдаги бошқарув ечимларини қабул қилишни қўллаб қувватлаш тизимларида фойдаланилиши мумкин.

Р Е З Ю М Е

диссертации Давронова Рифката Рахимовича на тему: “Минимаксный подход к структурному анализу данных методом ядерной кластеризации» на соискание ученой степени кандидата технических наук по специальности 05.13.01- Системный анализ, управление и обработка информации

Ключевые слова: анализ структуры данных, монотонная функция близости, ядерная кластеризация, сегментация изображений, множественное выравнивание белковых последовательностей.

Объекты исследования: матрицы близости, изображения, графы.

Цель работы: разработка и исследование минимаксной модели слоистой кластеризации на основе понятия монотонной функции близости.

Методы исследования: методы дискретной математики, лингвистического анализа данных, классификации, кластер - анализа, обработки изображения. Программная реализация разработанных алгоритмов осуществлена на языках C++, MATLAB, C#.

Полученные результаты и их новизна:

- разработана параметрическая модель ядерной кластеризации, основанная на понятии плотности монотонной функции близости;
- решена задача сегментации изображений методом ядерной кластеризации, оценена ее эффективность по сравнению с методом нормализованной вырезки и k-средних;
- разработана процедура множественного выравнивания белковых последовательностей методом ядерной кластеризации, оценена ее эффективность с пакетами CLUSTAL и DIALIGN.

Практическая значимость: полученные в исследовании алгоритмы и программы могут быть использованы в обработке больших массивов данных, процедурах множественного выравнивания и сегментации изображений.

Степень внедрения и экономическая эффективность: результаты диссертационной работы в виде комплекса программ внедрены в Республиканском научном центре экстренной медицинской помощи Минздрава РУз для компьютерной диагностики различных форм холецистита, а также на кафедре информационных систем Муромского института Владимирского государственного университета для решения задач обработки изображений и использования в учебном процессе. Эффект от внедрения - социальный.

Область применения: результаты исследований могут быть использованы в интеллектуальном анализе данных, и системах поддержки принятия управленческих решений в медицине и биологии.

RESUME

Thesis of Davronov Rifkat Rakhimovich on the scientific degree competition of the doctor of philosophy in technical sciences on specialty 05.13.01 - System analysis, control and information processing, subject: “Minimax approach to data structural analysis based on core clusterization method”

Key words: data mining, monotonous proximity function, core clusterization, image processing, multiple alignment of protein sequences

Subjects of research: matrixes of proximity, images, directed and undirected graphs

Purpose of work: the development and study of minimax model of layered clusterization based on conception of monotonous proximity function

Methods of research: the methods of discrete mathematics, linguistic data analysis, classification, cluster analysis, image processing. Software was realized on C++, MATLAB, C#.

The results obtained and their novelty:

– development of core clusterization parametric model. Based on conception of monotonous proximity function;
– the method of image segmentation based on core clusterization, and its’ efficiency in comparing with method of normalized cut and k-means;
– the procedure of protein sequences multiple alignment based on core clusterization was evaluated in comparing with known software such as CLUSTAL and DIALIGN.

Practical value: developed software can be used in different applications in area of data and image processing, particularly in bioinformatics problems solution.

Degree of embed and economic effectivity: the results of dissertation were used as software package for image processing and teaching needs in information system department of Murom Institute of Vladimir state University University and as software for computing diagnostics of cholecystitis different forms in Republican Scientific Center of Emergency Medicine, Uzbekistan Ministry of Healthcare. Efficiency is social impact on research, treatment of patients, and education.

Field of application: Data Mining and decision-making systems in biology and healthcare.