Подсекция 4: Интеллектуальный анализ данных (Big Data)

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ БОЛЬШИХ МАССИВОВ ДАННЫХ ДЛЯ ПОВЫШЕНИЯ КАЧЕСТВА МЕДИЦИНСКОЙ ДИАГНОСТИКИ

Н.Ю. Ильясова, А.В. Куприянов

Самарский государственный аэрокосмический университет им. академика С. П. Королёва (национально исследовательский университет),
Институт систем обработки изображений РАН

Предложен метод интеллектуального анализа данных больших массивов данных для решения крупномасштабных задач выявления причинно-следственных связей изменений диагностической информации на медицинских изображениях с различными видами заболеваний. В качестве интегральных показателей состояния сосудов глазного дна и коронарных сосудов сердца используется глобальный набор геометрических признаков, являющийся достаточно полной характеристикой диагностических изображений и позволяющий проводить эффективную диагностику сосудистой патологии. Для оценки информативности диагностических признаков сосудов по критерию эффективности классификации и формирования новых признаков для улучшения качества диагностики рассматривается метод дискриминантного анализа выборочных данных.

1 Введение

Ключевой проблемой современных информационных технологий является интеллектуальный анализ данных сверхбольшого объёма – «больших данных». В докладе академика И.А. Соколова о приоритетных направлениях исследований в информационных технологиях на первом месте указаны технологии сбора, хранения, обработки, поиска, анализа и визуализации сверхбольших данных. В соответствии с Прогнозом научно-технологического развития Российской Федерации на период до 2030 года, утверждённым Председателем Правительства Российской Федерации Д.А. Медведевым, к перспективным направлениям научных исследований относятся «Технологии обработки и анализа информации», включающие методы и технологии сбора, обработки, анализа и хранения сверхбольших объёмов информации. Целью работы является исследование методов и алгоритмов интеллектуального анализа больших массивов данных для решения крупномасштабных задач выявления причинно-следственных связей изменений диагностической информации на медицинских изображениях с различными видами заболеваний, а также разработка новых математических методов и алгоритмов распределённой обработки и распознавания биомедицинских изображений для систем удалённой диагностики. Предлагается единый подход к анализу различных классов изображений основанный на оценивании совокупности геометрических параметров выделяемых областей интереса, являющихся базовым набором признаков для дальнейшего диагностического анализа.

Для распознавания изображений на основе интеллектуального анализа больших массивов информации с применением методов дискриминантного анализа разработана технология формирования пространства эффективных признаков. В качестве интегральных показателей состояния сосудов глазного дна и коронарных сосудов предлагается использовать глобальный набор геометрических признаков, являющийся достаточно полной характеристикой диагностических изображений и позволяющий

проводить эффективную диагностику сосудистой патологии. На основе указанных методов создаются распределённые технологии и программное обеспечение для удалённой обработки, анализа и понимания изображений, предназначенные для реализации в автоматизированных телемедицинских системах не требующие знания априорных вероятностных моделей полезных сигналов, шумов и искажений. Разрабатываемые методы призваны повысить качество медицинской диагностики за счёт получения объективных численных оценок параметров биомедицинских изображениях с использованием больших объёмов массивов доступной информации.

2 Информационная технология интеллектуального анализа диагностических изображений.

Информационная технология интеллектуального анализа диагностических изображений включает метод формирования пространства эффективных признаков для классификации заданного набора изображений.

Методология выделения диагностически значимой информации на изображениях кровеносных сосудов основана на новой обобщённой математической модели кровеносных сосудов двух классов диагностических изображений: сосудов глазного дна и коронарных сосудов, характеризуемой набором геометрических параметров.

Геометрический подход к формированию диагностических признаков, которые в отличие от традиционных абстрактных спектрально-корреляционных признаков являются привычными и понятными для медиков, обладают наглядностью и учитывают специфику объекта, позволяет, в конечном счёте, повысить эффективность диагностики.

Для отбора наиболее эффективных признаков используется их корреляция с результатами экспертной оценки, дисперсионный анализ обучающей выборки или анализ ошибки диагностики с использованием отдельных характеристик. Производится оценка эффективности различных признаков для задачи автоматической диагностики и формируются рекомендации по использованию различных групп признаков в медицинской практике.

Информационная технология интеллектуального анализа диагностических изображений включает следующие новые методы и алгоритмы:

- метод и алгоритм повышения степени информативности признаков на основе дискриминантного анализа и формирование оптимальной выборки для обучения экспертной системы диагностики заболеваний;
- метод оценивания разделимости классов, не зависящий от распределений объектов в классах и от используемого классификатора;
- алгоритм уменьшения размерности пространства признаков и формирования новых информативных признаков, максимизирующих критерий разделимости на основе методов дискриминантного анализа, позволяющих повысить точность диагностирования степени патологии;
- технология формирования оптимальной выборки для обучения диагностической системы на основе исключения аномальных наблюдений, что также позволит повысить точность диагностирования заболеваний.

Разрабатываются проблемно-ориентированные распределённые программные комплексы анализа медико-диагностических изображений для выявления патологических изменений, включая инструментальные средства формирования количественных оценок степени патологии на основе экспертных заключений и предлагаемых методов классификации. Разрабатываемые программные комплексы призваны обеспечить пользователя возможностью управлять процессом проведения анализа и принятия решений [1]. Автоматизированные системы анализа количественных показателей позволяют стандартизировать постановку диагноза, значительно сократить время обследования и снизить его стоимость. Системы позволяют осуществлять анализ субклинических морфологических изменений патоморфологических элементов,

автоматизировать этапы диагностики и проводить количественный мониторинг патологических изменений диагностических образцов. Особенностью является использование элементов экспертных систем: база данных диагностических признаков, корреляционный, дискриминантный и кластерный анализ пространства признаков, прогноз степени патологии на основе экспертных оценок.

Система классификации и диагностических исследований [1] предоставляет средства проведения корреляционного и дискриминантного анализа для формирования пространства информативных признаков, средства формирования оптимальной выборки признаков по критерию эффективности разделения по группам патологии, средства кластерного анализа для фильтрации обучающей выборки с целью удаления недостоверных данных и получения нормативных значений признаков по группам патологии. Система интеллектуального анализа данных позволяет пользователю получать степень патологии, нормативные значения признаков для каждой степени патологии заболевания, прогноз вероятности развития заболевания, и обеспечит формирование диагностических решений.

3 Дискриминантный анализ для формирования пространства информативных признаков

Совместно с врачами Медико-Стоматологического Университета г. Москвы с кафедры Офтальмологии были проведены исследования на основе цифрового анализа изображений глазного дна. Была разработана методика диагностирования глазных заболеваний на основе оценки глобальных сосудистых характеристик (признаков). В работе рассматриваются геометрические признаки, предложенные в [2-4]. Такими признаками являются: средний диаметр, прямолинейность, чёткообразность, амплитуда колебаний толщины, частота колебаний толщины, извилистость толщины, амплитуда колебаний трассы, частота колебаний трассы, извилистость трассы, которые соответствуют диагностических признакам сосудов глазного дна.

При наличии двух или более классов (в нашем случае – 5 классов, включающих норму и 4 степени диабетической ретинопатии, сахарный диабет) задача выбора признаков состоит в отборе таких, которые являются наиболее эффективными с точки зрения разделимости классов [5, 6]. В дискриминантном анализе критерии разделимости классов формируются с использованием матриц рассеяния внутри классов и матриц рассеяния между классами [6, 7].

Матрица рассеяния внутри классов показывает разброс объектов относительно векторов математических ожиданий классов: $\mathbf{W} = \sum_{k=1}^{g} (\mathbf{X}_k - \bar{\mathbf{x}}_k)(\mathbf{X}_k - \bar{\mathbf{x}}_k)'$, где данным k – класса будут соответствовать вектора средних $\bar{\mathbf{x}}_k = [\bar{x}_{1k} \bar{x}_{21k} \ldots \bar{x}_{pk}]$, g – общее количество классов. Элементы *матрицы рассеяния между классами* B рассчитывается по формуле: $b_{ij} = \sum_{k=1}^{g} n_k \left(\bar{x}_{ik} - \bar{x}_i\right)\left(\bar{x}_{jk} - \bar{x}_j\right), i, j = \overline{1, p}$, $\bar{x}_i = \left(1/n\right)\sum_{k=1}^{g} n_k \bar{x}_{ik}$ – среднее значение признака i по всем классам, n_k – число объектов в k -м классе, $\bar{x}_{ik} = 1/n_k \sum_{m=1}^{n_k} \bar{x}_{ikm}$ – среднее значение признака в классе k, x_{ikm} – значение i-го признака для m-го объекта в k-м классе. Матрицы $\boldsymbol{W}$ и $\boldsymbol{B}$ содержат всю основную информацию о зависимости внутри классов и между классами. Для того чтобы получить критерий разделимости классов, нужно связать с этими матрицами некоторое число. Это число должно увеличиваться при увеличении рассеяния между классами или при уменьшении рассеяния внутри классов. Для этого наиболее часто используются критерии: $J_1 = tr(\mathbf{T^{-1}B})$, $J_2 = \ln |\mathbf{W^{-1}T}| = \ln\{|\mathbf{T}|/|\mathbf{W}|\}$, где $\mathbf{T} = \mathbf{B} + \mathbf{W}$.

Чем больше значение критерия – тем больше разделимость классов. Разработан следующий алгоритм формирования новых признаков, представленный на рисунке 1:

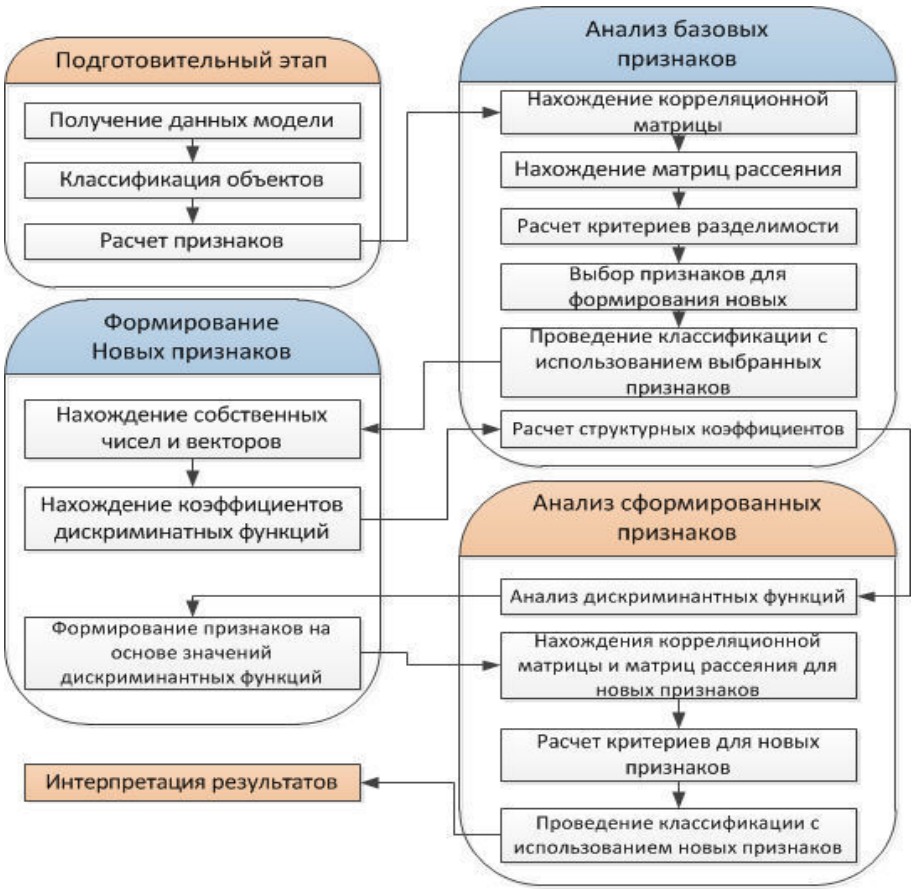

Рисунок 1 – Алгоритм проведения дискриминантного анализа признаков

4 Экспериментальные исследования

Был проведён ряд исследований на основе цифрового анализа изображений глазного дна, предназначенного для изучения особенностей формирования сосудистых нарушений при диабетической ретинопатии (ДР) 151 пациента с сахарным диабетом (СД). После обработки изображений выборка составила 8175 измерений, из них артериолы первого порядка – 1490, артериолы второго порядка – 2345, венулы первого порядка – 1960, венулы второго порядка – 2380. Врачи рассматривают венулы и артерии отдельно, так как в этих классах наблюдаются различные тенденции изменения сосудов при различных стадиях патологии.

При исследовании признаков можно сделать вывод о наличии двух сильно коррелированных групп признаков. В первую группу входят признаки, описывающие параметры трасс, в первую очередь прямолинейность и извилистость трассы, а вторую группу составляют признаки, характеризующие функцию толщины, такие как извилистость радиуса и чёткообразность. Для формирования новых признаков был произведён полный перебор исходных признаков для поиска комбинации новых признаков, которая максимизировала критерий разделимости. В результате был получен набор из четырёх признаков. Необходимо также отметить, что дальнейшее увеличение количества новых признаков не приводит к увеличению критерия разделимости.

При исследовании качества классификации было сформировано две выборки: обучающая и тестовая. На основе обучающей выборки был настроен классификатор, основанный на методе опорных векторов, с помощью которого классифицировалась тестовая выборка. Для синтеза классификатора используются только объекты обучающей выборки, которые не содержатся в тестовой выборке. Этот подход называют U-методом [6]. Объекты из истинного распределения могут быть заменены объектами, которые не были использованы для синтеза классификатора и независимы от объектов, по которым классификатор был синтезирован. Для реализации U-метода существует много

возможностей, при проведении исследования для оценивания вероятности ошибки классификации использовался метод исключения одного объекта.

Для уменьшения ошибки классификации осуществим фильтрацию исходной выборки методом кластеризации. Кластеризация проводилась алгоритмом *k-means* – итерационный алгоритм, который стремится минимизировать суммарное квадратичное отклонение точек кластера от центра этих кластеров. Каждая ГРУППА разбивалась на кластеры. Векторы признаков, которые не попали в нужный кластер (норма или 4 степени патологии), помечались как “выбросы” и отфильтровывались из выборки. Результаты, полученные в серии проведённых экспериментов, показали, что при формирования новых признаков всегда происходило улучшение общего критерия разделимости.

Таким образом, в результате дискриминантного анализа для каждой группы сосудов были определены лучшие признаки по критерию разделимости. Было показано, что в 4 группах эффективен свой набор глобальных геометрических признаков, что подтверждается клиническими исследованиями.

Группа		Исходная выборка			Отфильтрованная выборка		
		J_1	Повышение критерия	Ошибка	J_2	Повышение критерия	Ошибка
Артериолы 1 порядка	до	0,2593	19%	0,185	0,8896	39%	0,078
	после	0,3077		0,104	1,2386		0,024
Артериолы 2 порядка	до	0,3486	21%	0,144	0,9622	42%	0,083
	после	0,4219		0,090	1,3682		0,028
Венулы 1 порядка	до	0,3862	15%	0,128	1,1256	24%	0,105
	после	0,4434		0,096	1,4023		0,035
Венулы 2 порядка	до	0,3098	18%	0,162	1,1058	39%	0,072
	после	0,3656		0,113	0,8896		0,018

Таб.1. Результаты дискриминантного анализа признакового пространства

Анализируя полученные результаты (таб.1), можно сделать вывод, что фильтрация позволяет значительно увеличить критерии разделимости признаков, а также уменьшить ошибку классификации в 2-4 раза. Исследования на четырёх группах сосудов показали, что для каждой группы важен свой набор диагностических признаков, что подтверждается клиническими исследованиями врачей. Например, для венул и артериол при патологических изменениях по-разному ведёт себя средний диаметр сосуда. Результаты исследований показали, что применение алгоритма формирования признаков привело к уменьшению ошибки классификации на классы патологий. В результате было получено увеличение критерия разделимости для ГРУППЫ 1 – на 39%, ГРУППЫ 2 – на 42%, ГРУППЫ 3 – на 24%, ГРУППЫ 4 – 39%. Также была получена дополнительная информация по использующимся признакам, такая как их информативность, выделены связи между некоторыми признаками.

5 Выводы

Для анализа информативности и формирования более эффективных диагностических признаков изображений кровеносных сосудов была применена процедура дискриминантного анализа, основанная на максимизации критерия разделимости. Был разработан алгоритм, основанный на отборе признаков, имеющих наибольшее значение критерия разделимости, а также на полном переборе с последующим формированием новых признаков, максимизирующих данный критерий. В результате дискриминантного анализа для каждой группы сосудов были определены лучшие признаки по критерию разделимости. Было показано, что в 4 группах эффективен свой набор глобальных геометрических признаков, что подтверждается клиническими исследованиями. Подсчитана ошибка классификации для каждой группы сосудов до и после работы алгоритма. Показано, что технология анализа признакового пространства по

группам, включающая алгоритм формирования пространства эффективных признаков, позволила повысить эффективность классификации сосуды по классам «норма» и различным степеням «патологии» (СД). При этом ошибка классификации была снижена до 1,8%-3,5% для различных групп патологий.

Литература

1. Ильясова, Н.Ю. Диагностический комплекс анализа изображений сосудов глазного дна // Биотехносфера. –2014. – №3. – С. 132-138.
2. Ильясова, Н.Ю. Информационные технологии анализа изображений в задачах медицинской диагностики / Н.Ю. Ильясова, А.В. Куприянов, А.Г. Храмов. – М.: Радио и связь, 2012. – 424 с.
3. Ильясова, Н.Ю. Оценивание геометрических признаков пространственной структуры кровеносных сосудов // Компьютерная оптика. – 2014. – Т. 38, № 3. – С. 529- 538
4. Ilyasova, N. Computer Systems for Geometrical Analysis of Blood Vessels Diagnostic Images // Optical Memory and Neural Networks (Information Optics). – 2014. – Vol.23, Issue 4. – P. 278-286.
5. Ильясова, Н.Ю. Измерение биомеханических характеристик сосудов для ранней диагностики сосудистой патологии глазного дна / Н.Ю. Ильясова, А.В. Куприянов, М.А. Ананьин, Н.А. Гаврилова // Компьютерная оптика. – 2005. – № 27. – С. 165-170.
6. Фукунага, К. Введение в статистическую теорию распознавания образов / К. Фукунага. – М.: Наука, 1979. – 270 с.
7. Факторный, дискриминантный и кластерный анализ / Дж.-О. Ким, Ч.У. Мьюллер, У.Р. Клекка [и др.]; под ред. И.С. Енюкова; пер. с англ. – М.: Финансы и статистика, 1989. – 215 с.
8. Ильясова, Н.Ю. Формирование признаков для повышения качества медицинской диагностики на основе методов дискриминантного анализа / Н.Ю. Ильясова, А.В. Куприянов, Р.А. Парингер // Компьютерная оптика. – 2014. – Т. 38, № 4. – С. 751-756. – ISSN 0134-2452.