Медицинское изображение – это структурно-функциональный образ органов человека. Врачи используют его для диагностики заболеваний и изучения анатомо-физиологической картины организма.

Актуальность выбранной темы обосновывается тем, что своевременное распознавание патологических процессов в организме человека приведет к оказанию необходимой медицинской помощи. Проблема классификации патологических процессов по данным медицинских изображений не может быть правильно решена, если важные взаимодействия между оригинальными признаками, не принимаются во внимание.

Проблема. Эффективность работы классификатора сильно зависит от входного множества признаков. Как выбрать оптимальное множество признаков для классификатора?

Объектом исследования выступает классификация изображений. Предметом исследования является алгоритм построения признаков для классификации.

Цель – нахождение и отработка методики построения признаков для решения задачи классификации.

Для достижения поставленной цели необходимо выполнить следующие задачи:

* изучить соответствующую литературу;
* разработать алгоритм построения признаков;
* реализовать программу по данному алгоритму;
* оценить эффективность работы алгоритма и сравнить с результатами классификации без построения признаков;
* обобщить полученные результаты и сделать соответствующие выводы.

Классификация заключается в прогнозировании значения категориального атрибута (класса) на основе значений признаков объекта при известном множестве обучающих примеров.

Концептуально любой метод построения признака можно рассматривать как выполнение следующих действий:

1) Выбор начального пространства признаков F0.

2) Преобразование F0 для построения нового пространства признаков FN.

3) Выбор подмножества признаков Fi из FN на основании определения полезности Fi для задачи классификации.

4) Если некоторые критерии завершения не достигнуты, то возвращаемся к шагу 3.

5) Иначе множество FT = Fi, где FT – это сконструированное пространство признаков.

Для автоматического построение признаков можно использовать деревья решений, индуктивное логическое программирование, аннотации и генетическое программирование. Деревья решений обладают гибкой настройкой операторов, но не учитывают взаимодействия и отношения признаков, а также не имеют возможности добавления дополнительных знаний. В то же время ИЛП и аннотации удовлетворяют последним двум критериям, но не позволяют настраивать свои операторы. Поэтому было выбрано ГП.

Данный метод основывается на принципах биологической эволюции: естественный отбор, скрещивание и мутация.

В качестве популяции будут выступать леса – наборы признаков. Каждое дерево леса представляет собой один сконструированный признак, состоящий из функций, исходных признаков и констант.

Будем параллельно развивать несколько наборов лесов. Каждому такому набору назначим свой уникальный классификатор.

Оценивать пригодность особей будем через ошибку классификации. Подаем классификатору набор сконструированных признаков и вычисляем отношение неправильно предсказанных значений класса к числу всех объектов.

Результаты. Была разработана программа, которая позволяет получить набор построенных признаков данных посредством генетического программирования. Для проверки ее работы был проведен эксперимент.

Заданы обучающая и тестовая выборки медицинских изображений. Они были получены из доступного в интернете ресурса и представляют собой КТ снимки легких с отмеченными областями поражения.

Определен набор классификаторов: классификатор ближайшего соседа, метод опорных векторов, алгоритм C4.5 деревьев решений, наивный байесовский классификатор, искусственная нейронная сеть. И задано множество исходных признаков. Максимальная ошибка классификации для прекращения работы программы была выбрана равной 0.1.

Результаты работы программы можно продемонстрировать в виде таблицы. Каждый классификатор сначала предсказывал значение класса по оригинальному набору признаков. Видно, что точность предсказания равна приблизительно 80%. Затем классификаторы использовали свой лучший построенный набор признаков для прогнозирования класса. Можно увидеть, что средняя точность предсказания стала более 90%. Следовательно, цель работы – увеличение точности классификации – была достигнута.