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

Диагностика предполагает разделение изображений на классы: с патологией и без.

Изображение можно описать набором признаков – отличительных особенностей для различения. Для такого объекта как медицинское изображение количество признаков может быть больше 20.

От этого входного набора признаков сильно зависит точность классификации. Т.е. при изменении множества признаков получается другая точность классификации.

Актуальность выбранной темы обосновывается тем, что своевременное распознавание патологических процессов в организме человека приведет к оказанию необходимой медицинской помощи. Проблема классификации патологий не может быть решена с достаточной точностью, если важные взаимодействия между оригинальными признаками, не будут приниматься во внимание.

Проблема исследования. 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 улучшения точности 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 построения 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 об области. В то же время индуктивное логическое программирование и аннотации удовлетворяют последним двум критериям, но не позволяют настраивать свои операторы. Генетическое обладает свойствами всех названных критериев, поэтому было выбрано в качестве метода для построения признаков медицинских изображений.

Генетическое программирование основывается на принципах биологической эволюции, таких как: 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, с помощью которого будем оценивать пригодность данного набора признаков.

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

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.