Здравствуйте, уважаемые члены государственной экзаменационной комиссии.

Вашему вниманию представляется выпускная квалификационная работа на тему

«Применение генетического алгоритма для задачи понижения размерности пространства признаков»

Слайд 1

Одной из ключевых задач машинного обучения является задача снижения размерности пространства признаков.

В таких областях, как медицина, биоинформатика, промышленность оперирующих данными большой размерности с большим числом параметров, часто используются специальные алгоритмы интеллектуальной обработки данных для решения поставленных задач.

Для повышения эффективности работы этих алгоритмов, на практике часто применяются специальные методы, позволяющие выбирать наиболее релевантные комбинации параметров из рассматриваемых наборов данных.

В качестве таких методов предпочтительней всего использовать методы направленного поиска.

Слайд 2

Таким образом **целью данной работы** является исследование эффективности применения генетического алгоритма и некоторых его модификаций для задачи понижения размерности пространства признаков.

Для достижения поставленной цели были сформулированы **следующие задачи**:

- Реализовать исследуемый метод понижения размерности пространства признаков с применением ГА (в результате, проведения обзора существующих подходов к снижению размерности);

- Реализовать ГА с стандартными и модифицированными операторами;

- Выполнить оценку влияния модификации операторов ГА на эффективность понижения размерности;

- Выполнить сравнительный анализ исследуемого и часто используемых методов понижения размерности.

Теперь остановимся более подробно на способах понижения размерности пространства признаков.

Слайд 3

Существуют два подхода к снижению размерности:

Отбор признаков – это процесс выбора наиболее значимых признаков из исходного набора с сохранением их семантики для дальнейшего использования при построении модели обучения, визуализации и анализе данных и тд.

Выделение признаков – это, процесс снижения размерности пространства, в котором посредством специальных методов и алгоритмов из исходного набора признаков формируется новый набор меньшей размерности, обладающий иным семантическим наполнением.

Так как в данной работе будет проводиться исследование одного из методов отбора признаков давайте подробней остановимся и рассмотрим непосредственно их.

Слайд 4

В свою очередь отбор признаков можно разделить на следующие методы, а также их комбинацию:

Фильтры, оболочки (обертки) и встроенные методы. Принципиальные схемы работы этих методов, представленных на слайде.

В дальнейшем проводимые исследования по эффективности применению ГА для задачи понижения размерности будет проводиться с использованием метода обертка.

Слайд 5

Теперь рассмотрим более подробно задачу понижения размерности на основе метода обертки.

Детальная схема работы метода представлена на слайде под литерой а).

Так как в рамках данной работы исследуется применение одного из эвристических подходов для задачи понижения размерности, то в качестве алгоритма генерации подмножества признаков в данном методе будет выступать генетический алгоритм.

В качестве модели обучения будет фигурировать некоторый классификатор данных.

Слайд 6

Процесс выполнения отбора признаков с точки зрения самого ГА представлен на слайде пол литерой b).

Акцентируем внимание на том, что использование ГА в качестве генератора подмножества признаков приводит к тому, что модель обучения (некоторый классификатор) и оценка качества модели обучения будут являться неотъемлемой частью функцию приспособленности самого алгоритма, что представлено на слайде.

Слайд 7

Теперь подробней поговорим о количественные оценки эффективности отбора признаков и критерии завершения работы алгоритма.

Как отмечалось раньше в качестве функции приспособленности фигурирует классификатор данных, а также точность его работы.

На практики для количественной оценки работы классификаторов удобно использовать матрицу несоответствия. Структура матрицы и ее описание представлены на слайде.

Учитывая, что обучение и проверка классификатора будет проводиться на каждой итерации алгоритма для каждой особи из популяции, то в таком случае в качестве количественной меры лучше всего подходят простые оценки.

Свой выбор мы остановим на сбалансированной F1 мере. Формулы, для получения которой представлены на слайде.

Слайд 8

Теперь немного поговорим о реализации. Перечень основных операторов алгоритма и принятые стратегии представлены на слайде.

Дополнительно отметим, какие стратегии будут использоваться при реализации модифицированных операторов:

Оператор генерации нового поколения – выполняется в двух вариантах.

Стандартный использует операторы: выбор родителей, кроссинговер и мутацию.

Модифицированный, помимо перечисленных операторов, также используется принцип элитизма, где небольшое число самых приспособленных особей переходит в новое поколение без изменений.

Кроссинговер - выполняется в двух вариантах по многоточечной схеме.

Стандартный с равновероятностной передачей генов потомку от родителей.

Модифицированный передача генов потомку осуществляется с учетом значения функции приспособленности родительских особей;

Мутация – оператор выполняется в двух вариантах.

Статический – вероятность мутации, а также количество мутирующих генов остается неизменным на всем протяжении работы алгоритма.

Динамический – вероятность мутации отдельного гена и количество мутирующих генов завис от того на сколько близки с генетической точки зрения родительские особи.

Далее подробно поговорим об экспериментальных исследованиях.

Слайд 9

Для начала остановимся на оценке влияния модификации операторов ГА на эффективность понижения размерности пространства признаков.

Исследования будут проводиться на наборе данных предназначенным для диагностики неисправности ультразвуковых расходомеров жидкости.

Сперва мы рассмотрим работу метода с использованием стандартной реализации ГА. Результаты работы алгоритма представлены на слайде.

По имеющимся результатам работы можно сделать выводы.

На графике видно, что значение функции приспособленности наилучшей особи из популяции может снижаться от одного поколения к другому, это говорит о том, что наилучшая особь или ее гены не закрепляются в популяции на последующих итерациях расчета.

Отметим еще, что средние значения приспособленностей по популяции очень близко к максимальному, это говорит о преждевременной сходимости алгоритма и вырождении популяции в целом.

Слайд 10

Теперь мы рассмотрим работу ГА с использованием модифицированных операторов, результаты работы алгоритма представлены на слайде.

По имеющимся результатам работы можно сделать вывод, что применение стратегией элитизма, а также повышение шансов передачи генов от более приспособленной особи потомку позволяет сохранять наилучшие решения на всем протяжении работы алгоритма. Использование динамического оператора мутации позволяет уберечь алгоритм от преждевременной сходимости и вырождении популяции.

Слайд 11

Теперь поговорим о проведении сравнительного анализа.

Для выполнения сравнительного анализа будут использоваться 3 метода понижения размерности представленных на слайде.

Исследования будут проводиться с применением двух наборов данных разного размера, информация по наборам представлена на слайде;

Эффективность проделанной работы по понижению размерности будет осуществляться при помощи оценки качества итоговой классификации на наборах данных с редуцированным числом параметров.

Итоговая классификация будет выполняться при помощи нескольких классификаторов, перечень используемых классификаторов представлен на слайде.

Слайд 12

На слайде представлены результаты классификации наборов данных, полученных методами понижения рамерности, а также исходного набора данных.

Количество параметров, отобранных методами понижения размерности для выполнения классификации указано под наименованием метода. (про семантику в раздатке)

Дополнительно отметить, что сравнительный анализ проводился при равном числе отобранных и выделенных параметров.

В качестве количественной оценки при выполнении классификации использовалось значение площади под кривой ROC.

На основании полученных результатов можно сделать следующие выводы, для исследуемого алгоритма.

значительный прирост точности классификации по отношению к исходному набору данных наблюдается для того типа классификатора, который использовался при отборе признаков – kNN ~7 %;

для всех остальных классификаторов может наблюдаться снижение в точности их работы в пределах 5 % по отношению к исходному набору;

в сравнении с другими представленными методами снижения размерности, рассматриваемый алгоритм показывает соизмеримые результаты работы.

Слайд 13

Теперь для второго набора данных представим результаты классификации на исходном и редуцированном множестве признаков.

Отметим, что для методов отбора признаков проводилась предварительная обработка параметров при помощи фильтра, применялся так называемый гибридный метод.

И так, в сравнении с другими методами снижения размерности исследуемый алгоритм, также показывает соизмеримые результаты работы, отличия в точности классификации данных на редуцированных пространствах параметров находятся в пределах 10%.

Для данного набора данных исследуемый алгоритм показывает наилучший результат работы.

Слайд 14

В заключении отметим, что в ходе выполнения работы были достигнуты поставленные задачи.

- ✓ Выполнен обзор существующих подходов к снижению размерности пространства признаков, и выбран метод для исследования.

- ✓ Реализован ГА с модифицированными операторами для исследуемого метода.

- ✓ Проведена оценка влияния модификации операторов ГА на эффективность понижения размерности пространства признаков.

- ✓ Выполнен сравнительный анализ исследуемого и часто используемых методов понижения размерности пространства признаков.

- ✓ Все исследования проводились с использованием реальных наборов данных из открытых источников.