

**РЕФЕРАТ**

Курсовая работа содержит 24 страниц, 8 формул, 1 таблицу, 5 источников.

DATA MINING, [КЛАССИФИКАЦИЯ](https://www.intuit.ru/studies/courses/6/6/lecture/166?page=1#keyword1), ОБУЧЕНИЕ, С УЧИТЕЛЕМ, БЕЗ УЧИТЕЛЯ, ЧАСТИЧНОЕ ПРИВЛЕЧЕНИЕ УЧИТЕЛЯ, LEARNING, [SUPERVISED](https://www.intuit.ru/studies/courses/6/6/lecture/166?page=1#keyword18) , UNSUPERVISED , SEMI-SUPERVISED, [МНОГОМЕРНАЯ КЛАССИФИКАЦИЯ](https://www.intuit.ru/studies/courses/6/6/lecture/166?page=1#keyword25), [КЛАССИФИКАТОР](https://www.intuit.ru/studies/courses/6/6/lecture/166?page=2#keyword32), [TRAINING SET](https://www.intuit.ru/studies/courses/6/6/lecture/166?page=2#keyword35), ДИСКРИМИНАНТНЫЙ АНАЛИЗ, ПАРАМЕТРИЧЕСКИЙ ДИСКРИМИНАНТНЫЙ АНАЛИЗ

Объектом исследования являются задачи классификации, а также основные методы для их решения, в частности параметрический дискриминантный анализ.

Цель курсовой работы – применение данного метода для решения прикладной задачи из экономики.

СОДЕРЖАНИЕ

[Введение 3](#_Toc534720085)

[1 Классификация 6](#_Toc534720086)

[1.1 Общие сведения 6](#_Toc534720087)

[1.1.1 Что такое задача классификации 6](#_Toc534720088)

[1.1.2 Постановка классификационных задач 6](#_Toc534720089)

[1.1.3 Машинное обучение с учителем 7](#_Toc534720090)

[1.1.4 Машинное обучение без учителя 8](#_Toc534720091)

[1.1.5 Обучение с частичным привлечением учителя 8](#_Toc534720092)

[1.3 Дискриминантный анализ 9](#_Toc534720093)

[1.4 Параметрический дискриминантный анализ 9](#_Toc534720094)

[1.5 Алгоритм решения задач классификации с помощью параметрического дискриминантного анализа 10](#_Toc534720095)

[2 Актуальность применения метода в прикладных исследованиях 13](#_Toc534720096)

[2.1 Дискриминантный анализ в прогнозировании банкротства 13](#_Toc534720097)

[2.2 Дискриминантный анализ в определении целевой группы потребителей услуг 14](#_Toc534720098)

[2.3 Дискриминантный анализ в распознавании изображений 15](#_Toc534720099)

[2.4 Дискриминантный анализ в автоматическом определении происхождения изумруда 17](#_Toc534720100)

[3 Применение метода для решения прикладной задачи из экономики 19](#_Toc534720101)

[Заключение 23](#_Toc534720102)

[Список использованных источников 24](#_Toc534720103)

ВВЕДЕНИЕ

Иногда для объяснения какого-то понятия необходимо написать много текста, а иногда для объяснения сложнейших вопросов достаточно посмотреть на простую диаграмму. Для сокращения избыточности информации были придуманы математические формулы, чертежи, условные обозначения, программный код и т.д. Кроме того, важным является не только сама информация, но и ее представление. Понятно, что котировки акций более наглядно можно продемонстрировать с помощью графика, а математические формулы опишут законы Ньютона в более компактном виде.

В процессе развития информационных технологий, а также систем сбора и хранения данных – баз данных (databases), хранилищ данных (data warehousing), и с недавних пор, облачных репозиториев, возникла проблема анализа больших объемов данных, когда аналитик или управленец не в состоянии вручную обработать большие массивы данных и принять решение. Понятно, что аналитику необходимо каким-то образом представить исходную информацию в более компактном виде, с которой может справится человеческий мозг за приемлемое время.

Data Mining (извлечение, сбор данных, добыча данных) – это и есть процесс обнаружения в "сырых" данных ранее неизвестных нетривиальных практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Data Mining является одним из шагов Knowledge Discovery in Databases.

Информация, найденная в процессе применения методов Data Mining, должна быть нетривиальной и ранее неизвестной, например, средние продажи не являются таковыми. Знания должны описывать новые связи между свойствами, предсказывать значения одних признаков на основе других и т.д. Найденные знания должны быть применимы и на новых данных с некоторой степенью достоверности. Полезность заключается в том, что эти знания могут приносить определенную выгоду при их применении. Знания должны быть в понятном для пользователя не математика виде. Например, проще всего воспринимаются человеком логические конструкции "если … то …". Более того, такие правила могут быть использованы в различных СУБД в качестве SQL-запросов. В случае, когда извлеченные знания непрозрачны для пользователя, должны существовать методы постобработки, позволяющие привести их к интерпретируемому виду.

Алгоритмы, используемые в Data Mining, требуют большого количества вычислений. Раньше это являлось сдерживающим фактором широкого практического применения Data Mining, однако сегодняшний рост производительности современных процессоров снял остроту этой проблемы. Теперь за приемлемое время можно провести качественный анализ сотен тысяч и миллионов записей.

Классификация является самой простой и одновременно наиболее часто решаемой задачей Data Mining.

В данной курсовой работе будет подробно рассмотрен один из способов решения задачи классификации – параметрический дискриминантный анализ.

Целью курсовой работы является обзор параметрических методов классификации с обучением.

Для достижения поставленной цели необходимо решить следующие задачи:

– сделать обзор параметрических методов классификации и круга задач, решаемых данными методами;

– разобрать алгоритм дискриминантного анализа и реализовать его на примере малой размерности;

– изучить работы по применению дискриминантного анализа в Data Mining.

# 1 Классификация

## **1.1 Общие сведения**

### 1.1.1 Что такое задача классификации

В процессе решения широкого круга социально-экономических и исследовательских задач возникает необходимость анализа и формализации задач, связанных со сравнением и классификацией объектов, явлений, систем.

В самой общей формулировке под классификацией понимается разделение рассматриваемой совокупности объектов или явлений на однородные, в определенном смысле, группы либо отнесение каждого объекта из заданного множества классифицируемых объектов к одному из заранее известных классов, при этом классифицируемое «заданное множество» может состоять из единственного объекта. Термин «классификация» используется, в зависимости от контекста, для обозначения как самого процесса «разделения-отнесения», так и его результата.

### 1.1.2 Постановка классификационных задач

В учебнике [1] дается подробная постановка классификационных задач. Полагается, что каждый (*i*-й) из анализируемых объектов, (процессов, явлений) характеризуется значениями определенного набора признаков (свойств) *,* т.е. речь идет о классификации многомерных наблюдений , где .

Таким образом, на «входе» задачи классификации исследователь имеет:

* Априорные сведения о классах: о числе классов, об общем виде или некоторых свойствах законов распределения наблюдений X внутри каждого из классов, о диапазонах изменения анализируемых показателей. Основные источники априорных сведений о классах – результаты предыдущих аналогичных исследований, теоретические, предметно-профессиональные соображения, экспертные оценки;
* Обучающие выборки (training set)

|  |  |
| --- | --- |
|  |  |

где *k* – общее число (априори заданное) классов, а о наблюдениях известно, что все они характеризуют объекты, принадлежащие *j*-му классу. Такие выборки обычно получают с помощью специальным образом организованных статистических выборочных обследований анализируемой совокупности объектов.

В частном случае , т.е. в ситуации, когда исследователь не имеет в своем распоряжении обучающих выборок (unsupervised learning), говорят о задаче классификации без обучения и пользуются, соответственно методами кластер-анализа. Если же обучающие выборки имеются в распоряжении исследователя, то решается задача классификации при наличии обучения (supervised learning) и используется для ее решения методы дискриминантного анализа, в частности, параметрического.

### 1.1.3 Машинное обучение с учителем

Большая часть практического машинного обучения использует обучение с учителем.

Обучение с учителем – это когда у нас есть входные переменные (*X*) и выходные переменные (*Y*), и мы используем алгоритм для изучения функции отображения *Y = f (X).*

Цель состоит в том, чтобы аппроксимировать функцию отображения так хорошо, чтобы при наличии новых входных данных (*X*) мы могли предсказать выходные переменные (*Y*) для этих новых данных.

Это называется контролируемым обучением, поскольку процесс обучения алгоритма из набора обучающих данных можно рассматривать как учителя, контролирующего учебный процесс. Мы знаем правильные ответы, алгоритм итеративно делает прогнозы на тренировочных данных и корректируется учителем. Обучение останавливается, когда алгоритм достигает приемлемого уровня эффективности.

### 1.1.4 Машинное обучение без учителя

Обучение без учителя – это когда у нас есть только входные данные (*X*) и нет соответствующих выходных переменных (*Y*).

Целью обучения без учителя является моделирование базовой структуры, чтобы узнать больше о данных. Это называется обучением без учителя, потому что, в отличие от вышеприведенного обучения, нет правильных ответов и нет учителя. Алгоритмы для обнаружения и моделирования таких структур данных могут быть выбраны на свое усмотрение. Проблемные задачи обучения без учителя могут быть далее сгруппированы в задачи кластеризации и ассоциации.

### 1.1.5 Обучение с частичным привлечением учителя

Это задача, когда у нас есть большой объем входных данных (*X*) и только некоторые из данных помечены (*Y*). Хорошим примером является фотоархив, в котором подписаны только некоторые изображения (например, собака, кошка, человек), а большинство не имеет маркировки.

Многие реальные задачи машинного обучения попадают в эту область. Это связано с тем, что маркировка данных может быть дорогой или длительной, поскольку для этого может потребоваться связь с экспертами в предметной области, в то время как немаркированные данные дешевые и их легко собирать и хранить. Мы можем использовать методы обучения без учителя, чтобы обнаружить и изучить структуру входных переменных. Также в этой задаче мы можем использовать методы обучения с учителем, чтобы делать наиболее точные прогнозы для немеченых данных, затем передавать эти данные обратно в алгоритм обучения с учителем в качестве обучающих данных и использовать модель для прогнозирования новых искомых данных.

## **1.3 Дискриминантный анализ**

Также в учебнике [1] говорится, что если исследователь (статистик, эконометрист) располагает наряду с классифицируемыми данными, так называемыми обучающими выборками, то для решения поставленной задачи классификации ему следует обратиться к методам дискриминантного анализа. Тогда, при этом каждый класс интерпретируется как одномодальная совокупность, закон распределения вероятностей которой оценивается по соответствующей обучающей выборке. Если априорные сведения позволяют сделать вывод об общем параметрическом виде закона распределения вероятностей каждого класса, то используют методы параметрического дискриминантного анализа. Если общий вид закона распределения внутри классов неизвестен, то обучающие выборки используются для получения непараметрических оценок внутриклассовых законов распределения вероятностей, а сами процедуры классификации называют методами непараметрического дискриминантного анализа.

## **1.4 Параметрический дискриминантный анализ**

Таким образом, параметрический дискриминантный анализ применяется при выполнении ряда предположений [1]:

* Предположения о том, что наблюдаемые величины – измеряемые  
  характеристики объекта имеют нормально распределение. Это предположение следует проверять. Умеренные отклонения от этого предположения допустимы;
* Предположения об однородности дисперсий и ковариаций наблюдаемых переменных в разных классах. Умеренные отклонения от этого предположения также допустимы.

Наиболее важным критерием правильности построенного классификатора является практика.

Данный тип анализа заключается в статистической реализации правила:

|  |  |
| --- | --- |
|  |  |

где *k* – общее число классов, – удельный вес элементов класса *j*,  *–* функция, описывающая закон распределения вероятности анализируемого многомерного признака внутри класса *j*, когда известно, что , а – некоторый векторный параметр, значения компоненты которого внутри каждого из классов неизвестны. Тогда = , а неизвестное «значение» параметра , характеризующего закон распределения вероятностей в классе *j* заменяется в правиле (2) его статистической оценкой , построенной по обучающей выборке. Участвующие априорные вероятности либо считаются заданными, либо (если это возможно) также заменяются их статистическими оценками .

## **1.5 Алгоритм решения задач классификации с помощью параметрического дискриминантного анализа**

В результате, после всех теоретических выкладок, при различении нормальных (Гауссовских) классов устанавливается правило [1]:

|  |  |
| --- | --- |
|  |  |

где , а оценки параметров и параметров по наблюдениям обучающей выборки (1) имеют вид:

|  |  |
| --- | --- |
|  |  |
|  |  |

Соответственно, при различении двух нормальных классов (т.е. при  
k = 2) классифицирующая процедура (2) с учетом (3), (4) и (5) может быть описана в виде:

1. при: наблюдение следует отнести к классу 1, если

|  |  |
| --- | --- |
|  |  |

где оценка общей ковариационной матрицы вычисляется по формуле:

|  |  |
| --- | --- |
|  |  |

а оценки и вычислены в соответствии с (5). Функция , по значениям которой принимается решение об отнесении наблюдения **X** к 1-му или 2-му классу называется линейной дискриминантной функцией Фишера.

1. при: наблюдение следует отнести к классу 1, если

|  |  |
| --- | --- |
|  | (8) |
|  |  |

# 2 Актуальность применения метода в прикладных исследованиях

## **2.1 Дискриминантный анализ в прогнозировании банкротства**

В статье [2] рассматривается одна из самых актуальных проблем на сегодняшний день – проблема оценки финансового состояния с точки зрения экономических взаимоотношений, которая в большей степени сегодня является острой в сфере диагностики неплатежеспособности потенциального партнера. Способность клиента обслуживать свою задолженность вызывает собой большой интерес кредитных и страховых компаний. Здесь важно отметить разнообразие методов и моделей, которые применяются в целях оценки и прогнозирования несостоятельности, либо восстановления платежеспособности экономического субъекта. Одними из самых приемлемых и точных прогнозных моделей на сегодняшний день остаются модели, которые разработаны на основе многомерного дискриминантного анализа. При этом существует возможность оценить не только вероятность банкротства на данный момент, но и расчет такой вероятности на перспективу.

Во время процесса разработки модели прогноза банкротства с использованием дискриминантного анализа важно учитывать определенные обстоятельства. Так, в целях проводимого анализа уровень финансового состояния организации играет второстепенную роль, в данном случае результативным фактором является наличие процедуры банкротства. Поэтому существует объективная необходимость исследования двух соразмерных групп предприятий, одна из которых будет представлена организациями-банкротами, другая же – предприятиями, не подверженными данной процедуре. Дискриминантный анализ в таком случае позволяет установить наличие уровня зависимости между совокупностью выбранных переменных и состоянием организации (банкрот – не банкрот), а также выявить переменные, наиболее сильно характеризующие исследуемые позиции. Весь последующий прогноз осуществляется исходя из анализа данных за ряд лет, предшествующих процедуре банкротства.

## **2.2 Дискриминантный анализ в определении целевой группы потребителей услуг**

В статье, посвященной определению целевой группы потребителей услуг добровольного медицинского страхования с помощью дискриминантного анализа [3] рассматривается идея продвижения различных страховых продуктов, в частности, предложении услуг добровольного медицинского страхования (ДМС), где одной из самых главных задач для любой страховой компании является определение характеристик потенциальных потребителей страховых услуг. Целью такого исследования является определение основных характеристик потребителей услуг ДМС. В работе утверждается, что не существует единственно верного способа определения групп лиц, на которых должны быть направлены усилия по привлечению к страхованию. Источником информации для выявления детерминант спроса на услуги ДМС являются данные четырнадцатой волны наблюдения РМЭЗ за индивидами (2005г.). Для исследования вся совокупность индивидов была поделена на две группы. (имеющие - не имеющие полис ДМС). Для того, чтобы определить, какие из характеристик индивида и насколько существенно детерминируют его принадлежность к определенной группе, был проведен дискриминантный анализ, который позволил выявить наиболее информативные объясняющие переменные при классификации наблюдений на две группы. В качестве объясняющих переменных были выбраны социально-демографические и экономические характеристики индивида (пол, возраст, доход, наличие детей, возраст, населенный пункт, занятость трудовой деятельностью, принадлежность к профессиональной группе, образование, наличие пенсии), а также характеристики, которые описывают состояние здоровья (наличие проблем со здоровьем, недавняя госпитализация, наличие хронических заболеваний сердца, почек, желудочно-кишечного тракта и других хронических заболеваний, самооценка индивидом своего состояния здоровья, удовлетворенность жизнью, курение). Дискриминантный анализ в этом случае дал возможность принять решение о том, какие характеристики различают (дискриминируют) эти две группы, а также позволил на основе проведенной классификации предсказать, к какой группе будет относиться конкретное наблюдение, в данном случае – будет ли индивид с соответствующими характеристиками иметь полис медицинского страхования.

## **2.3 Дискриминантный анализ в распознавании изображений**

Рассмотрим один из важнейших вопросов измерения и обработки информации – проблему распознавания объектов. Итак, автоматизация процесса распознавания объектов ведёт к ускорению работы систем измерения и обработки данных и повышению их эффективности. Поэтому большой интерес к данной проблеме в различных областях наук и техники обусловлен многообразием прикладных задач, в которых используются либо сами изображения, либо результат их обработки. Необходимость в обработке и анализе изображений появляется не только при распознавании объектов, но и при изучении природных ресурсов Земли, управлении движущимися объектами, количественной оценке параметров объектов и др. При анализе и распознавании изображений отдельное изображение представляется в виде вектора признаков, каждый из которых отражает значение яркости одного пикселя. Таким образом, если изображение описывается матрицей пикселей   
a x b, размерность вектора равна произведению a на b. Так как вектор каждого изображения имеет достаточно большую размерность, задача обработки большого количества изображений является нетривиальной [4].

Большинство систем анализа изображений основывается на методах построения пространства признаков меньшей размерности. Задача снижения размерности важна еще и потому, что сложность большинства алгоритмов экспоненциально возрастает с увеличением размерности изображений, а практическая реализация таких алгоритмов требует мощных вычислительных средств. Одним из широко распространенных методов сокращения размерности изображений является метод главных компонент (МГК). В настоящее время для решения задачи поиска и распознавания лиц предлагаются множество алгоритмов, использующих МГК.

Любой набор изображений лиц может быть приблизительно восстановлен с помощью небольшого набора главных компонент и весовых коэффициентов для каждого изображения. Поскольку главные компоненты имеют ту же самую размерность, что и исходные изображения лиц, их часто называют «собственными лицами» или «собственными изображениями». Весовые коэффициенты, описывающие каждое изображение, называются главными факторами. Изображение хранится как набор главных факторов в базе данных в форме, которая может использоваться в качестве ключа поиска. В этом случае каждый фактор отображает не единичный пиксель, а группу пикселей, которые могут быть представлены в виде изображений (собственных лиц). Когда главные компоненты определяются на основе всех изображений, они создают полное пространство базисных векторов, то есть необходимая точность описания любого изображения достигается путем увеличения количества главных компонент. Если в качестве главных компонент используются все собственные вектора, то изображение восстанавливается точно.

Линейный дискриминантный анализ является очень мощной технологией для распознавания лиц, преобразующая исходное пространство изображений в низкоразмерное пространство признаков, в котором изображения классов группируются вокруг их центров, а центры классов удаляются друг от друга максимально, насколько это возможно.

## **2.4 Дискриминантный анализ в автоматическом определении происхождения изумруда**

Задача определения того, откуда произошли драгоценные камни, является одной из центральных в геммологии. Среди способов решения этой задачи выделяются два направления: методы, основанные на изучении включений, и методы, основанные на изучении характеристик объекта в целом, в том числе элементный и разные виды спектрального анализа [5]. Среди преимуществ второго направления выделяют экспрессность при подходящем выборе аналитической процедуры и отсутствие непреодолимых препятствий для алгоритмизации. Рентгенофлюоресцентный анализ (РФА) монолитного образца имеет такие преимущества, как отсутствие пробоподготовки и разрушения образца и возможность проведения измерений в составе изделия. Однако минус данного метода в том, что существует невозможность использования какого-либо стандарта, кроме внутреннего. Для алгоритмизации определения того, как произошел образец необходим поиск сигнатур (fingerprinting), то есть признаков, вычисляемых математически или алгоритмически и в то же время позволяющих принять решение о его происхождении без участия человека. Данная проблема относится к взаимопересекающимся областям знания: искусственный интеллект, машинное обучение, математическая статистика и глубокий анализ данных.

Существует единственная уникальная работа по использованию методов машинного обучения для классификации изумрудов. В ней использовались данные элементного анализа большого количества образцов, полученные электронным микрозондом, которые затем подавались на вход нейронной сети с одним скрытым уровнем. В соответствии с принятыми в сфере машинного обучения методами вначале нейронная сеть обучалась на 2/3 образцов, затем качество ее обучения проверялось на остальных образцах (кросс-валидация). Классификация выполнялась не по странам происхождения, а в соответствии с одной из классификаций месторождений по 5 категориям. В итоге было получено успешное разделение изумрудов по типам месторождений с малой долей ошибок в 3%.

Рентгенофлюоресцентные измерения монолитных образцов и последующее применение линейного дискриминантного анализа показывает свою перспективность для автоматического определения происхождения изумрудов. Однако малое число измерений пока не позволяет оценить надежность этого метода в соответствии с общепринятыми в машинном обучении критериями (с помощью кросс-валидации). В итоге наиболее важным является дальнейшее накопление экспериментальных и расчетных данных.

# 3 Применение метода для решения прикладной задачи из экономики

Специальное проведенное исследование [1] показало, что склонность фирм к утаиванию части своих налогов (и, соответственно – к уклонению от уплаты части налогов) в существенной мере определяется двумя показателями:

– соотношением «быстрых активов» и текущих пассивов;

– соотношением прибыли и процентных ставок (оба показателя оцениваются по определенной методике в шкале от 300 до 900 баллов).

В таблице 1 приведены значения этих показателей (данные налоговой инспекции) по 10 фирмам, уличенным в тех или иных формах уклонения от уплаты налогов. Кроме того, имеющаяся статистика и специально проведенные обследования свидетельствуют о том, что доля фирм, в той или иной форме уклоняющихся от налогов, достигает 50%. Наконец, статистическая проверка гипотез о нормальном характере распределения двумерного признака внутри каждой из анализируемых совокупностей фирм, уклоняющихся от уплаты налогов (совокупности 1) и платящих налоги (совокупности 2), и о равенстве их ковариационных матриц   
) дала неотрицательный результат, то есть мы можем считать, что имеющиеся у нас обучающие выборки извлечены из нормальных генеральных совокупностей с одинаковыми ковариационными матрицами.

Таблица 1 – Исходные данные

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Обучающая выборка (1) (фирмы, уклоняющиеся от налогов) | | Обучающая выборка (2)  (фирмы, не уклоняющиеся от налогов) | |
|  |  |  |  |
| 1 | 740 | 680 | 750 | 590 |

Продолжение таблицы 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Обучающая выборка (1) (фирмы, уклоняющиеся от налогов) | | Обучающая выборка (2)  (фирмы, не уклоняющиеся от налогов) | |
|  |  |  |  |
| 2 | 670 | 600 | 360 | 600 |
| 3 | 560 | 550 | 720 | 750 |
| 4 | 540 | 520 | 540 | 710 |
| 5 | 590 | 540 | 570 | 700 |
| 6 | 590 | 700 | 520 | 670 |
| 7 | 470 | 600 | 690 | 790 |
| 8 | 560 | 540 | 670 | 700 |
| 9 | 540 | 630 | 620 | 730 |
| 10 | 500 | 600 | 690 | 840 |
| 11 |  |  | 610 | 680 |
| 12 |  |  | 550 | 730 |
| 13 |  |  | 590 | 750 |

На фирме, не прошедшей проверку налоговой инспекции, зарегистрированы значения переменных .

Требуется предложить и реализовать метод, с помощью которого можно определить, к какой совокупности (1 или 2) следует отнести фирму с показателями .

Из условия задачи следует, что число анализируемых классов *k*=2, мы располагаем выборками объемов и , классы нормальны и различаются только средними значениями, а их удельные веса и одинаковы (). Следовательно, мы находимся в условиях применимости параметрического дискриминантного анализа и, в частности, можем воспользоваться линейной дискриминантной функцией Фишера вида (6).

Воспользовавшись формулами (5) и (7), мы на основании исходных данных, и получаем:

; ;

; .

Находим оценку общей ковариационной матрицы по формуле (7):

*,*

.

Для определения значения левой части формулы (8) вычисляем:

Наконец:

Следовательно, наблюдение должно быть отнесено к совокупности один, а это значит, что есть основания к тому, чтобы диагностировать анализируемую фирму как фирму, в той или иной форме уклоняющуюся от налогов.

ЗАКЛЮЧЕНИЕ

В курсовой работе были рассмотрены теоретические и практические основы машинного обучения (Machine Learning), в частности методы сбора и прогнозирования информации на основе эмпирических данных с помощью методов задач классификации, в частности параметрического дискриминантного анализа.

В процессе написания курсовой работы был сделан обзор параметрических методов классификации и круга задач, решаемых данными методами, был подробно разобран алгоритм дискриминантного анализа и далее реализован на примере малой размерности, а также мы изучили несколько реальных примеров работ применения дискриминантного анализа в Data Mining.

В результате, рассмотренный метод параметрического дискриминантного анализа был успешно применен для решения одной из прикладных задач экономики.

В итоге, можно сделать вывод, что дискриминантный анализ все еще является достаточно актуальным и одним из самых эффективных способов решения задач классификации на сегодняшний день ввиду его простоты, а также, в частных случаях, ввиду его вычислительных преимуществ перед остальными методами решения подобных задач классификации.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Прикладная статистика в задачах и упражнениях: Учебник для вузов / С. А. Айвазян, В. С. Мхитарян. – М.:ЮНИТИ-ДАНА, 2001. – 270 с.
2. Прогнозирование банкротства сельскохозяйственных производителей с использованием методов дискриминантного анализа / С. А. Кучеренко; Кубанский государственный аграрный университет. – Краснодар // Конкурсное производство и прогнозирование банкротства. – 2008. – №12 (117). – С. 73.
3. Определение целевой группы потребителей услуг добровольного медицинского страхования с помощью дискриминантного анализа / А. А. Трегубова // Учет и статистика. – № 4 (16). – 2009. – С. 77-82.
4. Об эффективности анализа и распознавания изображений методом главных компонент и линейным дискриминантным анализом / В.В. Мокеев, С. В. Томилов // Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника» – 2013. – Вып. 13 – №3 – С. 61–70.
5. Применение линейного дискриминантного анализа для автоматического определения происхождения изумруда по данным рентгенофлюоресцентного анализа / А. В. Поротников, М. П. Попов, Н. П. Горбунова // ЕЖЕГОДНИК-2012, Тр. ИГГ УрО РАН. – 2013. – №160. – C. 353–355.

