Методические указания к практическим занятиям

Программное обеспечение

WEKA – свободное программное обеспечение для анализа данных и машинного обучения, написанное на Java в Университете Уаикато (Новая Зеландия), распространяющееся по лицензии GNU GPL.

Weka представляет собой набор средств визуализации и алгоритмов для интеллектуального анализа данных и решения задач прогнозирования с графической пользовательской оболочкой для доступа к ним.

Weka позволяет выполнять такие задачи анализа данных, как подготовку данных, отбор важных признаков, классификацию, кластеризацию, анализ ассоциативных правил и визуализацию результатов.

Дистрибутив для установки: https://www.cs.waikato.ac.nz/ml/weka/downloading.html

Последняя версия для Windows x64: http://prdownloads.sourceforge.net/weka/weka-3-8-2-x64.exe

Версия 3.6.12 для машин на Windows 7: https://sourceforge.net/projects/weka/files/weka-3-6-windows/3.6.12/weka-3-6-12.exe/download

Версия 3.6.9 для 32 бит https://download.cnet.com/Weka-32-bit/3001-10254_4-75852610.html

Авторские учебные пособия

Целых А.Н. Современные методы прикладной информатики в задачах анализа данных : учебное пособие по курсу «Методы интеллектуального анализа данных» / Целых А.Н., Целых А.А., Котов Э.М.. — Ростов-на-Дону, Таганрог : Издательство Южного федерального университета, 2021. — 130 с. — ISBN 978-5-9275-3783-9. — Текст : электронный // IPR SMART : [сайт]. — URL: https://www.iprbookshop.ru/117165.html

Целых А.Н., Целых А.А., Котов Э.М., Краснощёков Е.Е. Методы интеллектуального поиска и анализа данных [Электронный ресурс] : монография / ЮФУ, ИТА, ИКТИБ, Каф. ИАСБ ; сост.: А. Н. Целых [и др.]. - Ростов н/Д : Изд-во ЮФУ, 2014 (Таганрог). – 232 с. – URL: http://ntblib.tgn.sfedu.ru/download/Resource/20344

Целых А.Н., Целых А.А., Котов Э.М. Методы интеллектуального анализа данных в системах принятия решений: Учебное пособие. – Таганрог: Изд-во ЮФУ, 2013. – 129 с. – URL: https://hub.sfedu.ru/repository/material/800820775/

Практические работы

1.1. Практическая работа №1

Классификация с использованием деревьев решений

Классификация – это упорядоченное по некоторому принципу множество объектов, которые имеют сходные классификационные признаки (одно или несколько свойств), выбранные для определения сходства или различия между этими объектами.

Целью классификации является построение оптимальной модели классификации, которая использует прогнозирующие атрибуты (вектор признаков) в качестве входных параметров для получения значения зависимого атрибута. Критерием оптимальности в нашем случае является показатель *точности*, рассчитываемый как отношение числа верно классифицированных примеров к их общему числу в выборке. При этом модель должна с высокой точностью классифицировать новые, ранее не рассмотренные примеры.

Для проведения классификации набор исходных данных разбивают на два множества: обучающее и тестовое. Оба множества содержат входные и выходные (целевые) значения атрибутов. В обучающем множестве (*training set*) выходные значения зависимых атрибутов предназначены для обучения (конструирования) модели. В тестовом множестве (*test set*) выходные значения используются для проверки работоспособности модели.

Подходы к оценке эффективности модели: на обучающем множестве (*Use training set*), на тестовом множестве (*Percentage split*), метод k-перекрестной проверки (*Cross-validation*).

В задачах обучения с учителем важно предупреждать *проблему переобучения* – нежелательное побочное явление, при котором ошибка алгоритма на обучающей выборке снижается, а на тестовой выборке растет.

Алгоритмы классификации. Для построения модели классификации в среде Weka используйте следующие алгоритмы на основе деревьев решений (в скобках приводятся параметры алгоритмов, требующие настройки):

- zeroR
- oneR
- Id3
- J4.8/C4.5 (unpruned, minNumObj, reducedErrorPruning, numFolds)
- JRip

Задание.

1. Используйте файлы данных в формате arff согласно вариантам заданий (табл. 1).

Таблица 1

Варианты заданий

Вариант 1 (Первая буква фамилии А-Ж)	Вариант 2 (Первая буква фамилии З-М)	Вариант 3 (Первая буква фамилии Н-С)	Вариант 4 (Первая буква фамилии Т-Я)
Contact-lenses.arff	Iris.arff	Labor.arff	Zoo.arff

2. Изучите файлы данных. Для каждого набора данных поочередно примените все доступные алгоритмы классификации с параметрами по умолчанию. Оцените точность классификаторов, используя обучающее множество, тестовое множество (2:1) и метод 10-перекрестной проверки. Объясните различие в результатах.
3. На основе результатов 10-перекрестной проверки сравните классификаторы между собой и отранжируйте по убыванию точности.
4. Настройте алгоритм J4.8 с предредукцией. Для этого установите значение параметра unpruned True. Далее произвольно изменяйте значения параметра minNumObj (минимальное число вершин в листьях) в диапазоне от 1 до 15. Оцените точность

классификаторов. Сравните с результатами алгоритма J4.8, полученными на предыдущем этапе.

5. Используйте алгоритм J4.8 с постредукцией. Для этого измените значение параметра unpruned на False, а значение параметра reducedErrorPruning – на True. Далее произвольно изменяйте значение параметра numFolds в диапазоне от 2 до числа примеров в базе данных. Тем самым, вы определяете объем данных, используемых для постредукции: одна часть используется для обрезки, остальные – для выращивания дерева решений. Оцените точность классификаторов. Сравните с результатами J4.8, полученными на предыдущих этапах.
6. На результатах п.4-5 проиллюстрируйте наглядным примером проблему переобучения.
7. По итогам работы определите оптимальные классификаторы для каждого набора данных.

Критерии оценки:

11-12 баллов выставляется студенту, если он своевременно выполнил все задачи, предусмотренные в практической работе, подготовил отчет в соответствии с требованиями преподавателя и в процессе защиты продемонстрировал наличие теоретических знаний в объеме содержания учебной дисциплины, относящейся к практической работе. Сумел ответить на дополнительные вопросы, связанные не только с процессом выполнения практической работы, но и с пониманием совершенных действий и решенных задач.

9-10 баллов выставляется студенту, если он выполнил от 70% задач, предусмотренных в практической работе, подготовил отчет в соответствии с требованиями преподавателя и в процессе защиты продемонстрировал наличие теоретических знаний в объеме содержания учебной дисциплины, относящейся к практической работе. Сумел ответить на вопросы, связанные с процессом выполнения практической работы.

7-8 баллов выставляется студенту, если он более чем на 50% выполнил поставленные в практической работе задачи, способен ответить на вопросы, касающиеся теоретической составляющей в объеме содержания учебной дисциплины, относящейся к практической работе.

0 баллов выставляется студенту, если он более чем на 50% не выполнил поставленные в практической работе задачи, не способен ответить на вопросы, касающиеся теоретической составляющей в объеме содержания учебной дисциплины, относящейся к практической работе.

1.2. Практическая работа №2

Классификация текстов

1. **Введение**

Интеллектуальный анализ текстов (*text mining*) связан с задачей выделения релевантной информации из текстов на естественном языке и поиском в ней закономерностей. Классификация текстов – это пример сложной проблемы интеллектуального анализа данных, в которой мы имеем дело с данными большой размерности и неоднородными шаблонами искомых данных.

2. **Классификация текстов в среде Weka**

Для представления текстовой информации в Weka используется строковый атрибут типа STRING. Такой атрибут может содержать огромное число значений и поэтому напрямую классификаторы с ним не работают. В исходных данных всего два атрибута: атрибут класса и строковый атрибут. Мы преобразуем значения STRING в множество атрибутов, представляющих частоту встречаемости слов в строке. Такое преобразование можно реализовать с помощью фильтра StringToWordVector (рис. 1). На первой вкладке нажмите кнопку Choose, найдите указанный фильтр и примените его с помощью кнопки Apply.

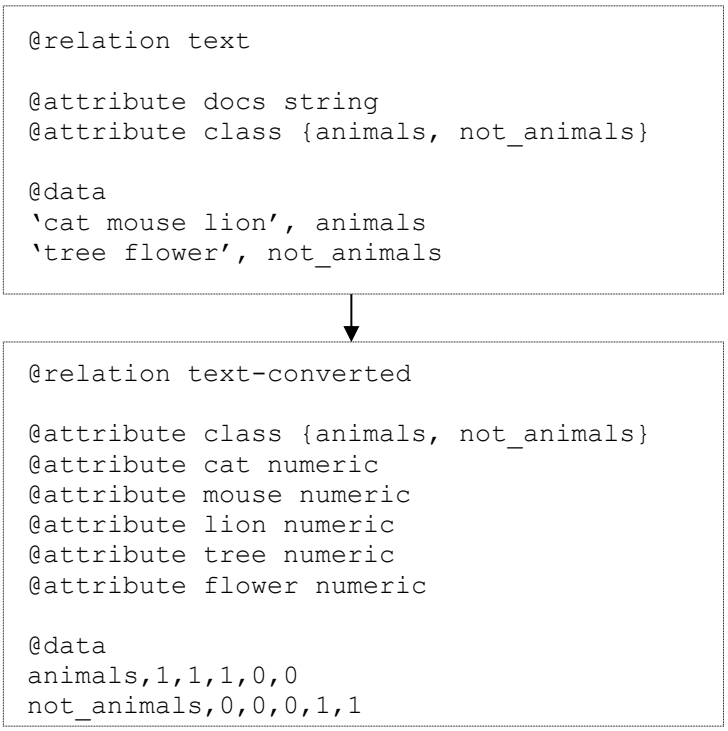

Рис. 1

Обратите внимание, что после преобразования атрибут класса обычно становится первым по счету. Однако в Weka по умолчанию атрибутом класса выбирается последний атрибут. Уточнить атрибут класса можно в выпадающем меню на вкладке Classify.

После преобразования можно начать обучение и сравнение классификаторов.
В этой лабораторной работе мы будем использовать четыре классификатора:
- алгоритм ближайшего соседа Nearest Neighbor (*IBk*),
- байесовский Naïve Bayes (*NB*),
- алгоритм на основе деревьев решений *J48*,

- алгоритм на основе решающих правил *JRip*.

3. Вавилонская башня

3.1 Описание

Дано:

- Коллекция из 189 предложений об интеллектуальном анализе данных из Wikipedia (не перевод!) на английском, французском, испанском и немецком языках.

Найти:

- Классификаторы, которые распознают язык текста.

Данные содержатся в файле dataMining.arff

3.2 Задание

Перейдите от строковой переменной к множеству атрибутов как описано выше.

Обучите классификаторы и сравните их между собой методом 10-перекрестной проверки. *Какой алгоритм лучший?* Выпишите статистики точности, изучите матрицы несоответствий и деревья решений классификаторов *J48* и *JRip*. Используйте эту информацию для ответа на следующие вопросы:

Являются ли языки, представленные в тексте, родственными? Как вы можете это подтвердить?

Классификатор IBk показывает невысокую точность для маленьких значений параметра k. Когда вы изменяете значение (например, увеличиваете значение k до 59), точность существенно растет. Как вы это объясните?

Ввиду высокой размерности данных (1902 числовых атрибута), обучение некоторых классификаторов (*J48, JRip*) занимает много времени. Для уменьшения времени работы алгоритма вы можете сократить число атрибутов, оставив только наиболее релевантные. Это можно сделать с помощью одного из методов выбора атрибутов, представленных на вкладке Select Attributes. Перейдите на указанную вкладку; убедитесь, что в выпадающем меню выбран алгоритм класса; запустите процедуру с параметрами по умолчанию. После вычисления оценок значимости атрибутов, вы можете удалить ненужные атрибуты из набора данных, используя фильтр Remove. Для этого нажмите Choose; выберите указанный фильтр; в настройки фильтра вставьте номера отобранных атрибутов и допишите 1 через запятую; измените false на true для инверсии, чтобы оставить только значимые атрибуты и атрибут класса; нажмите Apply. Повторите эксперименты и выпишите значения статистик.

Как вы видите, кроме сокращения времени работы, классификатор IBk повысил свою точность при малых значениях k. *Как вы это объясните?*

4. Музыканты шутят

4.1 Описание

Дано:

- Коллекция из 409 шуток о музыке на английском языке с сайта http://www.mit.edu/~jcb/jokes/.

Шутки помечены как смешные и не смешные.

Найти:

- Классификаторы, которые распознают, смешная шутка или не смешная.

Данные представлены в файле musicJokes.arff

4.2 Задание

Обучите и сравните между собой классификаторы. Выпишите статистики и изучите матрицы несоответствий классификаторов.

Применяя фильтр StringToWordVector, вы сначала получите ошибку: название переменной class совпадает с названием одного из атрибутов (слов в тексте). Нажмите на слово StringToWordVector и в параметрах фильтра выберите настройку stopwordsHandler – Choose – WordsFromFile. Нажмите на слово WordsFromHandler, затем на stopwords. Создайте текстовый файл stopwords.txt, в котором содержится одно слово class. Укажите путь к этому файлу. Теперь вы можете применить фильтр.

Используя визуализацию деревьев решений J48 и JRip, можете ли вы предположить, что делает шутку смешной?

Поскольку ни один из классификаторов не имеет высокой точности, попробуйте улучшить работу классификаторов, выбрав наиболее релевантные атрибуты, используя один из методов на вкладке Select Attributes. *Какой алгоритм лучший?*

5. Американский и китайский английский

5.1 Описание

Дано:

- Коллекция из 132 пар предложений на английском, написанных американцами и китайцами, изучающими английский (http://www.englishdaily626.com/c-mistakes.php).

Найти:

- Классификаторы, которые распознают, написано предложение на английском американцем или китайцем.

Данные представлены в файле americanChineseStyles.arff

5.2 Задание

Обучите и сравните между собой классификаторы. Осуществите отбор важных атрибутов.

Точность классификаторов не превышает 50%. *Можно ли, используя классификаторы с невысокой точностью, получить приемлемые результаты классификации? Каким классификатором пользоваться?*

Критерии оценки:

11-12 баллов выставляется студенту, если он своевременно выполнил все задачи, предусмотренные в практической работе, подготовил отчет в соответствии с требованиями преподавателя и в процессе защиты продемонстрировал наличие теоретических знаний в объеме содержания учебной дисциплины, относящейся к практической работе. Сумел ответить на дополнительные вопросы, связанные не только с процессом выполнения практической работы, но и с пониманием совершенных действий и решенных задач.

9-10 баллов выставляется студенту, если он выполнил от 70% задач, предусмотренных в практической работе, подготовил отчет в соответствии с требованиями преподавателя и в процессе защиты продемонстрировал наличие теоретических знаний в объеме содержания учебной дисциплины, относящейся к практической работе. Сумел ответить на вопросы, связанные с процессом выполнения практической работы.

7-8 баллов выставляется студенту, если он более чем на 50% выполнил поставленные в практической работе задачи, способен ответить на вопросы, касающиеся теоретической составляющей в объеме содержания учебной дисциплины, относящейся к практической работе.

0 баллов выставляется студенту, если он более чем на 50% не выполнил поставленные в практической работе задачи, не способен ответить на вопросы, касающиеся теоретической составляющей в объеме содержания учебной дисциплины, относящейся к практической работе.

1.3. Практическая работа №3

Кластеризация

В настоящей лабораторной работе вы проведете эксперименты с тремя алгоритмами кластеризации.

1. **Алгоритм k-средних**

Алгоритм k-средних – это простой, прямой алгоритм кластеризации. Каждый кластер определяется центроидом, и экземпляры принадлежат к тому кластеру, евклидово расстояние до центроида которого минимально. Затем для каждого кластера находят новый центроид как среднее арифметическое наблюдений в кластере, что может привести к перестановкам экземпляров между кластерами. Итеративный процесс завершается, когда центроиды перестают меняться.

Загрузите файл “marble.arff”. Набор данных содержит описание 15 шаров, использованных в качестве примера на лекции. После загрузки файла удалите атрибут «цвет» («color»), для чего выберите его и нажмите кнопку «Удалить» («Remove»). Обратите внимание, что атрибут «класс» («class») содержит естественные названия групп, обсуждавшихся на лекции. Проверим теперь, сможет ли алгоритм k-means найти эти группы.

Перейдите на вкладку «Кластеризация» («Cluster»). Выберите алгоритм SimpleKMeans. Поскольку известны четыре естественных кластера, используйте 4 или 5 в качестве значения параметра k. Убедитесь, что в качестве режима кластеризации выбран «Classes to clusters evaluation».

2.1 Поскольку алгоритм k-средних находит локальный минимум, обычно рекомендуется запускать алгоритм несколько раз и использовать наилучший из результатов. Проведите эксперимент и изучите результаты. Насколько близко вы смогли приблизиться к «идеальному» разделению шаров на классы?

2. **Алгоритм Cobweb**

Данный алгоритм строит кластеры, пошагово добавляя элементы к дереву и включая их в существующий кластер, если это приводит к увеличению значения функции полезности, по сравнению с определением элемента в новый кластер. По необходимости существующий кластер может быть разбит на два новых кластера, если это положительно скажется на значении функции полезности. Результирующее множество кластеров называется «дендрограммой».

На вкладке «Preprocess» загрузите файл “marblespecific.arff” и удалите атрибут «цвет» («color»). Файл «marblespecific.arff» отличается от предыдущего набора данных в том плане, что атрибут «класс» («class») является уникальным для каждого шара, что позволяет распознавать шары в отчете Weka. Код для каждого шара состоит из четырех букв и цвета шара. В свою очередь, четыре буквы объясняют размер (Big / Small), расцветку (Monochrome / Polychrome), блеск (Shiny / Dull) и прозрачность (Transparent / Opaque).

Перейдите на вкладку «Кластеризация». Выберите алгоритм Cobweb. Установите флаг «Classes to clusters evaluation», чтобы алгоритм игнорировал значение атрибута «класс», но включил этот атрибут в финальный отчет для сравнения. Произведите кластеризацию.

3.1 Изучите результаты на основе визуализации дерева («Visualize tree»). Вы удовлетворены результатами кластеризации? Почему?

Увеличение значения параметра Cutoff алгоритма будет способствовать отнесению похожих шаров в один кластер.

3.2 Поэкспериментируйте со значением параметра Cutoff, пока вы не будете удовлетворены результатами кластеризации.

3. Алгоритм максимизации ожидания

Алгоритм EM – это вероятностный алгоритм кластеризации. Каждый кластер определяется вероятностями элементов обладать некоторыми конкретными значениями атрибутов и вероятностями принадлежности каждого элемента к кластеру. Для числовых значений это значение среднего и стандартное отклонение для каждого значения атрибута, для дискретных значений это вероятность для каждого значения атрибута.

Поскольку с дискретными значениями проще работать, а также для сравнения с результатами двух предыдущих алгоритмов, применим алгоритм EM к тому же набору данных. На вкладке «Preprocess» загрузите файл “marble.arff” и удалите атрибут «цвет» («color»).

Перейдите на вкладку «Кластеризация». Выберите алгоритм EM. Установите флаг «Classes to clusters evaluation». Произведите кластеризацию.

4.1 Изучите результаты. Сколько кластеров получилось? Почему? Можете ли вы получить другой результат с другим значением параметра seed?

Значение параметра numClusters по умолчанию -1, что позволяет алгоритму самостоятельно определять число кластеров. Если указать конкретное значение, алгоритм постарается обнаружить необходимое число кластеров.

4.2 Измените значение параметра на число кластеров, которое вы хотите получить (или немного больше) и повторите эксперимент. Проведите эксперименты с различными значениями seed. Как это влияет на результаты?

При изучении отчета вы увидите базовую вероятность кластера и значение вероятности для каждого атрибута, которую можно получить делением значения дискретной оценки (“discrete estimator”) на сумму оценок по каждому атрибуту («total»), например:

Атрибут: размер (size)
Discrete Estimator. Counts = 2.41 11.86 (Total = 14.26)

означает, что вероятность значения «большой» (big) для атрибута «размер» составит 2.41/14.26 = 0.169.

Порядок значений атрибутов следующий:

“size” {big, small}
“colouring” {monochrome, polychrome}
“shininess” {shiny, dull}
“transparency” {transparent, opaque}

Чтобы определить «лучший» кластер для шара, нужно перемножить базовую вероятность кластера и вероятности для каждого из значений атрибутов. Нормализованное число даст ожидаемую вероятность шара быть отнесенным к каждому из кластеров.

Для следующего эксперимента установите значение параметра seed равным 14, а число кластеров – 2. Произведите кластеризацию.

4.3 Пусть у вас имеется маленький, одноцветный, неяркий, непрозрачный шар. В каком кластере вы ожидаете его увидеть? (Для расчетов используйте стандартное приложение «Калькулятор».)

Заключение

Вы провели эксперименты с тремя алгоритмами на одном наборе данных.

5.1 Приведите (как минимум) одно преимущество и один недостаток каждого алгоритма.

5.2 Какой из трех алгоритмов кластеризации вы предпочитаете использовать для набора данных «Шары».

Критерии оценки:

11-12 баллов выставляется студенту, если он своевременно выполнил все задачи, предусмотренные в практической работе, подготовил отчет в соответствии с требованиями преподавателя и в процессе защиты продемонстрировал наличие теоретических знаний в объеме содержания учебной дисциплины, относящейся к практической работе. Сумел ответить на дополнительные вопросы, связанные не только с процессом выполнения практической работы, но и с пониманием совершенных действий и решенных задач.

9-10 баллов выставляется студенту, если он выполнил от 70% задач, предусмотренных в практической работе, подготовил отчет в соответствии с требованиями преподавателя и в процессе защиты продемонстрировал наличие теоретических знаний в объеме содержания учебной дисциплины, относящейся к практической работе. Сумел ответить на вопросы, связанные с процессом выполнения практической работы.

7-8 баллов выставляется студенту, если он более чем на 50% выполнил поставленные в практической работе задачи, способен ответить на вопросы, касающиеся теоретической составляющей в объеме содержания учебной дисциплины, относящейся к практической работе.

0 баллов выставляется студенту, если он более чем на 50% не выполнил поставленные в практической работе задачи, не способен ответить на вопросы, касающиеся теоретической составляющей в объеме содержания учебной дисциплины, относящейся к практической работе.

1.4. Практическая работа №4

Ассоциативные правила

Rules Wizard (разработчик BaseGroup Labs) – простое в использовании приложение, решающее задачи поиска ассоциативных связей (Association Rules). Программа позволяет выявлять ассоциативные правила, используя таблицу транзакций и представлять их разными способами - в виде форматированного текста, дерева, перекрестной и обычной таблицы.

Ассоциативные связи – это зависимости вида: если произошло событие A, то с определенной вероятностью произойдет событие B. Примером такого правила, служит утверждение, что покупатель, приобретающий Хлеб, приобретет и Молоко с вероятностью 75%.

Впервые задача поиска ассоциативных правил была предложена для нахождения типичных шаблонов покупок, совершаемых в супермаркетах, поэтому иногда ее еще называют анализом рыночной корзины (market basket analysis). Наибольшее использование ассоциативные правила находят в торговле, так как позволяют сказать, какой дополнительный товар может приобрести клиент, если он уже купил некоторые товары.

Задание: провести анализ рыночной корзины на основе набора данных о покупках в супермаркете (976 покупателей, 4852 покупки, 49 наименований товаров).

Объясните значение поддержки и достоверности для ассоциативного правила вида:

Если *Молоко* ***И*** *Печенье*
То *Чипсы*
Поддержка = 0,72%
Достоверность = 77,78%

1. Для БД транзакций из файла smarket.txt (в папке RulesWizard/Data) произведите анализ рыночной корзины со значениями поддержки s=0,3~100,0 и достоверности c=70,0~98,0. Изучите представления результатов в виде форматированного текста, дерева ассоциативных правил, перекрестной таблицы и таблицы правил. Зафиксируйте условия поиска, количество найденных правил и интервалы, на которых лежит значение поддержки и достоверности.

2. По дереву ассоциативных правил определите тройку товаров-бестселлеров. Перечислите все двухэлементные ассоциативные наборы-лидеры продаж (s>5%).

3. Для трех ассоциативных правил с низкой (<0,5%), средней (~1%) и высокой (>5%) степенью поддержки, соответственно, сделайте предположение, почему именно эти товары оказались в одной корзине. Составьте портрет-характеристику покупателя.

4. Используя представление в виде перекрестной таблицы, выясните, наличие каких товаров в корзине с большой вероятностью определит покупку ветчины, шоколада и кофе.

5. Подберите значение поддержки, при котором формулируется ~100 ассоциативных правил. Ознакомьтесь со статьей «Зачем купил - не знаю» (газета «Новые известия»). Предложите не менее 3 «интересных» ассоциативных наборов товаров, которые рекомендуется размещать на одной полке или объединять в наборы со скидкой. Аргументируйте, почему эти ассоциативные правила можно считать «интересными».

Критерии оценки:

11-12 баллов выставляется студенту, если он своевременно выполнил все задачи, предусмотренные в практической работе, подготовил отчет в соответствии с требованиями преподавателя и в процессе защиты продемонстрировал наличие теоретических знаний в объеме

содержания учебной дисциплины, относящейся к практической работе. Сумел ответить на дополнительные вопросы, связанные не только с процессом выполнения практической работы, но и с пониманием совершенных действий и решенных задач.

9-10 баллов выставляется студенту, если он выполнил от 70% задач, предусмотренных в практической работе, подготовил отчет в соответствии с требованиями преподавателя и в процессе защиты продемонстрировал наличие теоретических знаний в объеме содержания учебной дисциплины, относящейся к практической работе. Сумел ответить на вопросы, связанные с процессом выполнения практической работы.

7-8 баллов выставляется студенту, если он более чем на 50% выполнил поставленные в практической работе задачи, способен ответить на вопросы, касающиеся теоретической составляющей в объеме содержания учебной дисциплины, относящейся к практической работе.

0 баллов выставляется студенту, если он более чем на 50% не выполнил поставленные в практической работе задачи, не способен ответить на вопросы, касающиеся теоретической составляющей в объеме содержания учебной дисциплины, относящейся к практической работе.

1.5. Практическая работа №5

Анализ связей в социальных сетях

Целью лабораторного занятия является изучение базовых понятий и техники анализа социальных сетей на основе программы UCINET 6 компании Analytic Technologies, являющейся стандартом для специалистов в области анализа социальных сетей.

Теоретическая часть

Очень часто распределенные системы, в частности сети сотовой связи, компьютерные сети и Всемирная Паутина обладают сложной топологией и имеют в своей основе социальные процессы. Основным подходом к изучению таких структур является анализ социальных сетей.

Анализ социальных сетей – это методология и методы исследования взаимодействий между социальными объектами (акторами) и выявление условий возникновения этих взаимодействий.

Основными методами анализа социальных сетей являются методы теории графов.

Данные о связях акторов представляются в виде *социоматрицы* - квадратной или прямоугольной таблицы, элементы которой соответствуют показателям силы связи, исходящей от актора в *i*-й строке к актору в *j*-м столбце.

Всякой социоматрице может быть взаимнооднозначно сопоставлен граф. Связи в графе могут быть ненаправленными (ребра) и направленными (дуги). Граф с заданными на нем дугами называется ориентированным, или орграфом. Вершины, соединенные ребром, являются смежными. *Степенью вершины* называется число ребер, соединенные с ней. *Исходящей степенью вершины* называется число дуг, исходящих из вершины, *входящей степенью* - число дуг, входящих в вершину.

Последовательность вершин, соединенных ребрами, составляет *цепь*. В цепи направление связей между вершинами не имеет значения. В *простой цепи* ни одна из вершин и ни одно из ребер не повторяются. Число ребер цепи называется ее *длиной*. Длина самой короткой цепи, связывающей две вершины, называется *расстоянием* между вершинами (без учета направления связей).

Последовательность вершин, соединенных дугами, называется *путем* (направление связей существенно). В *простом пути* ни одна из вершин и ни одна из дуг не повторяются. Число дуг, составляющих путь, называется его *длиной*. Длину самого короткого пути, связывающего две вершины, называют *расстоянием* между ними (с учетом направления связей). Орграф, в котором из каждой вершины существует путь к любой другой вершине, называется *сильно связным* (путешествовать можно лишь по направлению дуг). Орграф, в котором существует цепь из каждой вершины к любой другой, называется *слабо связным* (можно путешествовать против направления дуг).

Плотность графа вычисляется как отношение числа существующих связей к потенциально возможному.

Показатели заметности. Идея центральности вершин в графе, их "важности" начала разрабатываться одной из первых в анализе социальных сетей. Источник этой идеи можно усмотреть в понятии "звезды" - самого популярного человека в группе. Будем называть меру заметности актора в сети (неориентированном графе) *центральностью*, для входящих связей в орграфе - *престижем*, для исходящих связей - *экспансивностью*.

Простой и интуитивно понятный подход к измерению центральности индивидов основывается на идее *степени*. Центральность на основе степени тем выше, чем больше число связей вершины с другими вершинами в графе. Индексы центральности по степени являются локальными характеристиками положения вершины в графе - они учитывают непосредственных соседей, ближайшую окрестность вершины и в этом смысле поверхностны. В отношении типа «дружба» престиж можно интерпретировать как популярность, а экспансивность как форму коммуникабельности.

Вторая группа показателей центральности основана на идее *близости* данной вершины ко всем остальным вершинам графа. Центральным является тот индивид, который быстро взаимодействует с другими либо непосредственно, либо через небольшое число посредников. Близость определяется как величина, обратная сумме длин самых коротких путей от данного индивида ко всем остальным. Доступная интерпретация близости - ожидаемое время движения ресурса (например, информации) от любого участника сети к данному индивиду. Центральность по близости является глобальной мерой сети. Недостаток показателя в том, что он не определен для изолированных вершин, поскольку при отсутствии связи между вершинами расстояние между ними бесконечно.

Взаимодействие двух несмежных индивидов может находиться под контролем возможных посредников. При поисках работы, например, важно не то, сколько знакомых у претендента, а сколько знакомых у этих знакомых. Метод оценки центральности по *посредничеству* для вершины заключается в нахождении доли самых коротких путей, соединяющих все пары вершин, которые проходят через данную вершину. Это сумма вероятностей того, что другие акторы в своих взаимодействиях будут прибегать к посредничеству данного актора.

Центральность на основе *собственных векторов* – более сложный показатель, учитывающий связи вершины с вершинами, имеющими большой вес. Примером такой меры центральности является параметр Google PageRank.

Показатели центральности, основанные на степени, информационно бедны. Центральность по посредничеству и центральность на основе собственных векторов предпочтительны в силу того, что они имеют большую изменчивость значений и более интересную интерпретацию.

Задание

1. На листке бумаге нарисуйте произвольный граф на 6 вершин и 8 неориентированных ребер (без изолированных вершин). Назовите вершины латинскими буквами или словами.
2. Представьте нарисованный граф в виде матрицы 6x6, в которой «1» соответствует ребру между вершинами, «0» - ребро отсутствует.
3. В Блокноте *на основе подготовленной матрицы* создайте исходный текстовый файл net1.txt в следующем формате (где n – число вершин):
   dl n=**4** format=fullmatrix
   data:
   0 1 1 0
   1 0 1 1
   1 1 0 0
   0 1 0 0
4. Запустите программу UCINET (Пуск > Программы > Analytic Technologies > Ucinet 6 for Windows).
5. Загрузите файл (Data > Import > DL). В первом поле (Input text file in DL format) укажите путь к исходному текстовому файлу и нажмите ОК. Итоговый рабочий файл в формате UCINET будет иметь расширение *.##h. Откройте файл для просмотра (Data > Browse) и убедитесь, что импорт состоялся.
6. Запустите NetDraw (последняя иконка на панели инструментов UCINET) и откройте рабочий файл net1.##h (File > Open > Ucinet dataset > Network).
7. Перетаскивая вершины, приведите рисунок в соответствие с графом на бумаге. Сохраните рисунок в формате JPG (File > Save Diagram as > Jpeg) под именем net1.jpg.
8. Повторите шаги 3-7, но используйте теперь расширенный формат исходного файла net2.txt с названиями вершин:
   dl n=**4**
   labels:
   A,B,C,D
   data:

0 1 1 0
1 0 1 1
1 1 0 0
0 1 0 0

Файлы net1.txt, net1.##h, net1.jpg, net2.txt, net2.##h, net2.jpg служат отчетом по первой части работы.

Далее работа будет вестись с набором данных PADGETT (\Program Files\Analytic Technologies\Ucinet 6\DataFiles\PADGETT.##h) о социальных связях между 16 семействами Флоренции эпохи Возрождения. В файле содержатся две сети – супружеские и деловые связи. Мы будем работать с первой сетью. Для этого необходимо выделить сеть из набора, используя команду Data > Extract. В первом поле (Input dataset) укажите путь к исходному файлу, в поле Which matrices введите «1» и нажмите OK. В дальнейшем работайте с новым файлом PADGETT-Ext.##h

Отчет по второй части работы оформляется в виде файла net3.doc в формате Microsoft Word.

1. Изучите матрицу сети (Data > Display) и включите ее в отчет. (Используйте фиксированный шрифт, например Courier.)
2. Запустите NetDraw, загрузите сеть, сохраните рисунок в формате JPEG и включите его в отчет.
3. Вычислите плотность графа (NETWORK > COHESION > DENSITY). Проверьте правильность, рассчитав плотность вручную на основе матрицы из шага 1. Включите значение плотности и ход ручного расчета в отчет.
4. Вычислите расстояния между семействами (NETWORK > COHESION > DISTANCE). Проверьте правильность, рассчитав вручную расстояния между несколькими семействами на основе рисунка из шага 2. Включите матрицу Geodesic Distances и ход ручного расчета в отчет.
5. Вычислите и включите в отчет (только основные результаты – без статистик) следующие показатели центральности семейств:
    a. Центральность на основе степени (NETWORK > CENTRALITY > DEGREE).
    b. Центральность по посредничеству (NETWORK > CENTRALITY > FREEMAN BETWEENNESS > NODE BETWEENNESS).
    c. Центральность на основе собственных векторов (NETWORK > CENTRALITY > EIGENVECTOR).
6. Подготовьте для отчета таблицу, в которой 16 семейств будут отсортированы по убыванию показателя центральности на основе степени. Таблица – на 16 строк и 4 колонки (семейство, центральность на основе степени, центральность по посредничеству, центральность на основе собственных векторов). (Используйте Excel или инструмент Таблица в Word.)
7. Выберите произвольную пару семейств с одинаковой центральностью на основе степени, равной «4», но с совершенно разными оценками центральности по посредничеству. Используя социограмму из шага 2, предложите объяснение столько кардинальному различию в центральности по посредничеству. Сделайте то же самое для пары семейств с одинаковой центральностью на основе степени, равной «3». Отразите свой выбор и объяснение в отчете.
8. Выберите произвольную пару семейств с одинаковой центральностью на основе степени, равной «4», но с совершенно разными оценками центральности на основе собственных векторов. Используя социограмму из шага 2, предложите объяснение столь кардинальному различию в центральности на основе собственных векторов. Сделайте то же самое для пары семейств с одинаковой центральностью на основе степени, равной «3». Отразите свой выбор и объяснение в отчете.
9. Самоорганизация социальных сетей проявляется в виде сообществ – групп вершин с высокой плотностью ребер внутри группы и не высокой плотностью ребер между

группами. Выделите сообщества в сети, используя следующий алгоритм [Girman, Newman, 2002]:

a. Вычислите показатель посредничества для ребер в сети (NETWORK > CENTRALITY > FREEMAN BETWEENNESS > EDGE (LINE) BETWEENNESS).
b. Удалите ребро с максимальным значением посредничества (DATA > BROWSE, обнуляем удаляемое ребро, FILE > SAVE)
c. Изучите новый рисунок сети в NetDraw.

Повторяйте шаги a)-c), пока граф не начнет распадаться на несвязные компоненты, постепенно обнаруживая сообщества. Включите в отчет 2-3 рисунка, наглядно иллюстрирующие процесс выделения сообществ.

Критерии оценки:

11-12 баллов выставляется студенту, если он своевременно выполнил все задачи, предусмотренные в практической работе, подготовил отчет в соответствии с требованиями преподавателя и в процессе защиты продемонстрировал наличие теоретических знаний в объеме содержания учебной дисциплины, относящейся к практической работе. Сумел ответить на дополнительные вопросы, связанные не только с процессом выполнения практической работы, но и с пониманием совершенных действий и решенных задач.

9-10 баллов выставляется студенту, если он выполнил от 70% задач, предусмотренных в практической работе, подготовил отчет в соответствии с требованиями преподавателя и в процессе защиты продемонстрировал наличие теоретических знаний в объеме содержания учебной дисциплины, относящейся к практической работе. Сумел ответить на вопросы, связанные с процессом выполнения практической работы.

7-8 баллов выставляется студенту, если он более чем на 50% выполнил поставленные в практической работе задачи, способен ответить на вопросы, касающиеся теоретической составляющей в объеме содержания учебной дисциплины, относящейся к практической работе.

0 баллов выставляется студенту, если он более чем на 50% не выполнил поставленные в практической работе задачи, не способен ответить на вопросы, касающиеся теоретической составляющей в объеме содержания учебной дисциплины, относящейся к практической работе.

Вопросы

1. Понятие Data Mining.
2. Данные, информация и знания.
3. Единая методология обнаружения знаний.
4. Задача регрессии.
5. Задача классификации.
6. Задача кластеризации.
7. Задача анализа ассоциаций и последовательностей.
8. Наборы и типы данных.
9. Форматы хранения данных.
10. Качество данных. Очистка данных.
11. Методы отбора значимых признаков.
12. Фильтры. Оболочки.
13. Индукция деревьев решений.
14. «Обрезка» деревьев: предредукция и постредукция.
15. Решающие правила.
16. Алгоритмы ID3, CART, C4.5.
17. Алгоритмы ограниченного перебора.
18. Метод опорных векторов. Линейная и нелинейная разделимость.
19. Байесовская классификация.
20. Наивная байесовская классификация.
21. Типологический и таксономический анализ.
22. Статистические методы кластеризации. EM-алгоритм.
23. Метод k-средних. Меры расстояний.
24. Иерархические методы кластеризации.
25. Визуализация кластеров. Дендрограммы.
26. Диаграммы рассеивания.
27. Самоорганизующиеся карты Кохонена.
28. Концептуальная кластеризация.
29. Алгоритм Cobweb.
30. Графовые методы кластеризации.
31. Выделение связных компонент.
32. Нечеткая кластеризация.
33. FCM-алгоритм.
34. Меры интересности: поддержка, достоверность, лифт, уверенность.
35. Алгоритм Apriori. Задача анализа рыночных корзин.
36. Матрица несоответствий.
37. Метрики качества: правильность, полнота, точность, F-мера, чувствительность, AUC.
38. Проблема переобучения.
39. Стратификация данных.
40. Диаграмма выигрыша.
41. Ансамбли (комитеты) моделей. Бэггинг. Бэггинг с рандомизацией.
42. Бустинг (усиление) ансамбля классификаторов.
43. Метаклассификаторы. Стэкинг.
44. Понятия социальной сети и социального графа.
45. Позиционный и ролевой анализ в социальной сети.