

Lecture №6. Data analysis. Data management

Lecture purpose: to study data analysis bases.

Lecture content: methods of collection, classification and prediction. Decision trees. Processing large volumes of data. Methods and stages of Data mining. Tasks of Data mining. Visualization of data.

Data analysis

Data analysis is a process of data research, filtration, conversion and modeling to extract the useful information and decision-making. Data analysis has a set of aspects and approaches, covers different methods in different fields of science and activities.

To create the plan of data collection it is necessary to:

- 1) Define problems and formulate research objectives.
- 2) Realize a preliminary study of an interesting subject.
- 3) Develop concepts of research.
- 4) Make detail planning of a research.
- 5) Make a selection of information sources and collection of secondary data.
- 6) Estimate the obtained data and to make a decision how primary data are needed.
- 7) Define a method of collection of primary data – inquiry, observation, experiment.
- 8) Carry out directly the collection of primary information.
- 9) Provide results of the research (presentation).

- intuitive, which deal with opinions and estimates of experts;
- formalized which are already described in literature and based on which are already built forecasting models.

Decision trees are widely used in data slicing pane.

Decision trees – it the method of rules representation in hierarchical, sequential structure where a single node giving a decision corresponds to each object there.

All problems, which are solved by a tree method, can be integrated in the following three classes:

1) *Data description*: Decision trees allow storing information about data in the compact form, instead of them we can store a decision tree, which contains the exact description of objects.

2) *Classification*: Decision trees perfectly cope with tasks of classification, i.e. reference of objects to one of in advance known classes. The target variable shall have discrete values.

3) *Regression*: If a goal variable has continuous values, decision trees allow setting the dependence of the goal variable on independent (input) variables. For example, the tasks of numerical prediction (prediction of values of a goal variable) belong to this class.

There are many algorithms, which realize decision trees, among them CART, C4.5, NewId, ITrule, CHAID, CN2, etc. However, the following ones are most widespread:

CART (Classification and Regression Tree) – an algorithm of creation of a binary decision tree – dichotomizing classification model. Each node of such a tree in case of partition has only two descendants. The algorithm solves problems of classification and regression.

C4.5 is an algorithm of creation of a decision tree, in which the number of node's descendants is not restricted.

Data Mining Bases

Speaking about processing large volumes of data, you face the use of the term Data mining, which implies that there is a huge number of data.

Data Mining is a process of support of decision-making, which it is based on the search of hidden regularities (information templates) in data. It is such a technology intended for the search of regularities unobvious, objective and useful in practice in large volumes of data.

Tasks of Data Mining are sometimes called regularities or techniques. The main objectives of Data mining are: classification, clustering, prediction, association, visualization, analysis and detection of deviations, estimation, analysis of relations and summing up.

Methods and Data mining algorithms:

- artificial neural networks;
- decision trees;
- character rules;
- methods of the closest neighbor and the k-closest neighbor;
- method of reference vectors;
- Bayesian networks;
- linear regression;
- correlative regression analysis;
- hierarchical methods of cluster analysis;
- not hierarchical methods of cluster analysis, including algorithms of k-averages and k-median line;
- methods of search of the associative rules, including algorithm Apriori;
- method of limited search, evolutionary programming and genetic algorithms, various methods of data visualization and a lot of other methods.

Data Mining can consist of two or three stages.

Stage 1. Detection of regularities (free search).

Stage 2. Use of revealed regularities for prediction of unknown values (prognostic simulation).

Stage 3. Analysis of exceptions is the stage intended for detection and explanation of anomalies which are found in regularities.

Visualization of Data Mining tools.

Each of Data mining algorithms uses a certain approach to visualization. During the use of each of each of Data mining methods, to be exact, of its program implementation, we are viewers, by means of who it is possible to interpret the results, which are received because of operation of the appropriate methods and algorithms.

For decision trees such a viewer is a decision tree, list of rules and contingency table.

For neural networks, depending on the tool, it can be network topology, the diagram of change of error value showing training activity.

For Kekkonen's cards: cards of inputs, outputs, other specific cards.

For the linear regression, the regression line appears as a viewer.

For clustering: dendrograms scatter diagrams.

Diagrams and graphs of dispersion are often used for an assessment of quality of operation of this or that method.

All these methods of visual representation or data mapping can execute one of the functions:

- are an illustration of creation of a model (for example, representation of a structure (graph) of a neural network);
- help to interpret the received result;
- are means of an assessment of quality of a constructed model;
- combine the mentioned above functions (a decision tree, dendrograms).

Before using Data mining technology, it is necessary to analyze carefully its problems, restrictions and critical questions, connected with it, as well as to understand what this technology cannot do.

The analytics cannot replace Data Mining!

The technology cannot give answers to those questions, which were not asked. It cannot replace an analyst, and only gives him the powerful tool for facilitation and improving its operation.

The complexity of development and operation of Data mining application.

As this technology is a multidisciplinary area, for application development including Data Mining, it is necessary to involve experts from different areas, and to provide their high-quality interaction.

Control questions

- 1 What is data analysis?
- 2 What is data?
- 3 How can forecasting data methods be classified?
- 4 What is the regression in data analysis?
- 5 What is data visualization?
- 6 What is Data Mining?
- 7 What are data mining methods?
- 8 What is a decision tree?
- 9 What are the tasks of Data mining?
- 10 What are the algorithms of creation of a decision tree?

№6 дәріс. Деректерді талдау. Деректерді басқару

Дәріс мақсаты: деректерді талдау базасын зерттеу.

Дәріс мазмұны: жинау, жіктеу және болжау әдістері. Шешім ағашы. Деректердің үлкен көлемін өңдеу. Деректерді зияткерлік талдаудың әдістері мен кезеңдері. Деректерді зияткерлік талдаудың міндеттері. Деректерді визуализациялау.

Деректерді талдау

Деректерді талдау - бұл пайдалы ақпарат алу және шешім қабылдау үшін деректерді зерттеу, сүзу, түрлендіру және модельдеу процесі. Деректерді талдаудың көптеген аспектілері мен тәсілдері бар, ғылым мен қызметтің әртүрлі салаларындағы әртүрлі әдістерді қамтиды.

Деректерді жинау жоспарын құру үшін сізге қажет:

- 1) проблемаларды анықтау және зерттеу мақсаттарын тұжырымдау.
- 2) Қызықты пәнді алдын ала зерделеуді жүзеге асыру.
- 3) зерттеу тұжырымдамасын әзірлеу.
- 4) зерттеуді егжей-тегжейлі жоспарлау;
- 5) ақпарат көздерін іріктеуді және қайталама деректерді жинауды жүргізу.
- 6) алынған деректерді бағалау және деректердің қаншалықты бастапқы екендігі туралы шешім қабылдау
- 7) бастапқы деректерді жинау әдісін анықтау-анықтау, байқау, эксперимент.
- 8) бастапқы ақпаратты тікелей жинауды жүзеге асыру.
- 9) зерттеу (презентация) нәтижелерін ұсыну.

- сарапшылардың пікірлері мен бағалауларымен айналысатын интуитивті;

- формальды, олар әдебиетте сипатталған және олардың негізінде болжамды модельдер құрылған.

Шешім ағаштары деректерді кесу саласында кеңінен қолданылады.

Шешім ағаштары - бұл ережелерді иерархиялық, дәйекті құрылымда ұсыну әдісі, онда әр объект шешім беретін жеке түйінге сәйкес келеді.

Ағаш әдісімен шешілетін барлық тапсырмаларды келесі үш сыныпқа біріктіруге болады:

1) *деректер сипаттамасы*: шешім ағаштары деректер туралы ақпаратты ықшам түрде сақтауға мүмкіндік береді, олардың орнына объектілердің нақты сипаттамасы бар шешімдер ағашын сақтай аламыз.

2) *жіктеу*: шешім ағаштары жіктеу міндеттерін жақсы орындайды, яғни объектілерді алдын-ала белгілі сыныптардың біріне жатқызу. Мақсатты айнымалы дискретті мәндерге ие болуы керек.

3) *регрессия*: Егер мақсатты айнымалы үздіксіз мәнге ие болса, шешім ағаштары мақсатты айнымалының тәуелсіз (кіріс) айнымалыларға

тәуелділігін анықтауға мүмкіндік береді. Мысалы, бұл сыныпқа сандық болжау есептері кіреді (мақсатты айнымалы мәндерді болжау).

Шешім ағаштарын жүзеге асыратын көптеген алгоритмдер бар, олардың ішінде CART, C4.5, NewId функциясы, ITrule, CHAID, CN2 пайдаланады және т. б. алайда, ең көп тарағандары:

CART (Classification and Regression Tree)-шешімдердің екілік ағашын құру алгоритмі-дихотомизациялаушы жіктеу моделі. Мұндай ағаштың әрбір түйіні бөлінген жағдайда тек екі ұрпағы болады. Алгоритм жіктеу және регрессия мәселелерін шешеді.

C4.5 - бұл шешім ағашын құру алгоритмі, онда түйіннің ұрпақтарының саны шектелмейді.

Деректерді зияткерлік талдау базалары

Деректердің үлкен көлемін өңдеу туралы айтатын болсақ, сіз көптеген деректердің болуын білдіретін Data mining терминін қолданасыз.

Деректерді зияткерлік талдау-бұл мәліметтерде жасырын заңдылықтарды (ақпараттық шаблондарды) іздеуге негізделген шешім қабылдауды қолдау процесі. Дәл осы технология үлкен көлемде нақты емес, объективті және пайдалы деректердің заңдылықтарын іздеуге арналған.

Деректерді интеллектуалды талдаудың міндеттері кейде заңдар немесе әдістер деп аталады. Деректерді зияткерлік талдаудың негізгі міндеттері: жіктеу, кластерлеу, болжау, біріктіру, визуализация, ауытқуларды талдау және анықтау, бағалау, байланыстарды талдау және қорытындылау.

Деректерді интеллектуалды талдау әдістері мен алгоритмдері:

- жасанды нейрондық желі;
- шешім ағашы;
- мінез ережелері;
- жақын көрші және K-жақын көршінің әдістері;
- қолдау векторлары әдісі;
- байесовский желілері ;
- сызықтық регрессия;
- корреляциялық регрессиялық талдау;
- кластерлік талдаудың иерархиялық әдістері;
- кластерлік талдаудың иерархиялық емес әдістері, оның ішінде k-орта және k-медианалық желі алгоритмдері;
- ассоциативті ережелерді іздеу әдістері, соның ішінде априори алгоритмі;
- шектеулі іздеу әдісі, эволюциялық бағдарламалау және генетикалық алгоритмдер, деректерді визуализациялаудың әртүрлі әдістері және басқа да көптеген әдістер.

Деректерді зияткерлік талдау екі-үш кезеңнен тұруы мүмкін.

Кезең 1. Үлгілерді анықтау (еркін іздеу).

Кезең 2. Белгісіз шамаларды болжау үшін анықталған заңдылықтарды қолдану (болжамдық модельдеу).

Кезең 3. Ерекше жағдайларды талдау-бұл заңдылықтарда кездесетін ауытқуларды анықтауға және түсіндіруге арналған кезең.

Деректерді болжалды талдау құралдарын визуализациялау

Деректерді интеллектуалды талдау алгоритмдерінің әрқайсысы визуализацияға белгілі бір тәсілді қолданады. Деректерді интеллектуалды талдау әдістерінің әрқайсысын, дәлірек айтсақ, оны бағдарламалық қамтамасыз етуді қолданған кезде біз көрермендер болып табыламыз, олардың көмегімен тиісті әдістер мен алгоритмдердің жұмысы нәтижесінде алынған нәтижелерді түсіндіруге болады.

Шешім ағаштары үшін мұндай қарау құралы шешім ағашы, ережелер тізімі және күтпеген жағдайлар кестесі болып табылады.

Нейрондық желілер үшін құралға байланысты бұл желілік топология, оқу белсенділігін көрсететін қате мәнін өзгерту диаграммасы болуы мүмкін.

Кекконен карталары үшін: кіру, шығу карталары, басқа да нақты карталар.

Сызықтық регрессия үшін регрессия сызығы қарау құралы ретінде көрсетіледі.

Кластерлеу үшін: дендрограммалардың шашырау диаграммалары.

Дисперсия диаграммалары мен графиктері көбінесе белгілі бір әдістің жұмыс сапасын бағалау үшін қолданылады.

Барлық осы визуалды бейнелеу немесе деректерді көрсету әдістері функциялардың бірін орындай алады:

- сипаттама моделін (мысалы, ұсыну құрылымы (баған) нейрон);
- алынған нәтижені түсіндіруге көмектесу
- салынған модельдің сапасын бағалау құралы болып табылады;
- жоғарыда аталған функцияларды біріктіріңіз (шешімдер ағашы, дендрограммалар).

Деректерді интеллектуалды талдау технологиясын пайдаланбас бұрын, оны мұқият талдау қажет оның қиындықтарын, шектеулерін және сыни мәселелерін, сондай-ақ бұл технологияның не істей алмайтынын мұқият талдау қажет.

Аналитика деректердің зияткерлік талдауын алмастыра алмайды!

Технология сұралмаған сұрақтарға жауап бере алмайды. Ол аналитиканы алмастыра алмайды, тек оған жұмысын жеңілдету және жақсарту үшін қуатты құрал береді.

Деректерді зияткерлік талдау қосымшасын әзірлеу мен пайдаланудың күрделілігі. Бұл технология пәнаралық сала болғандықтан, қосымшаларды әзірлеу үшін, оның ішінде деректерді зияткерлік талдау үшін әртүрлі саладағы мамандарды тарту, сондай-ақ олардың сапалы өзара іс-қимылын қамтамасыз ету қажет.

Бақылау сұрақтары

1 деректерді талдау дегеніміз не?

2 деректер дегеніміз не?

- 3 деректерді болжау әдістерін қалай жіктеуге болады?
- 4 деректерді талдауда регрессия дегеніміз не?
- 5 деректерді визуализациялау дегеніміз не?
- 6 деректерді зияткерлік талдау дегеніміз не?
- 7 деректерді зияткерлік талдау әдістері дегеніміз не?
- 8 шешім ағашы дегеніміз не?
- 9 деректерді зияткерлік талдаудың міндеттері қандай?
- 11 шешім ағашын құру алгоритмдері қандай?

Лекция №6. Анализ данных. Управление данными

Цель лекции: изучение баз анализа данных.

Содержание лекции: методы сбора, классификации и прогнозирования. Дерево решений. Обработка больших объемов данных. Методы и этапы интеллектуального анализа данных. Задачи интеллектуального анализа данных. Визуализация данных.

Анализ данных

Анализ данных - это процесс исследования, фильтрации, преобразования и моделирования данных для извлечения полезной информации и принятия решений. Анализ данных имеет множество аспектов и подходов, охватывает различные методы в различных областях науки и деятельности.

Для создания плана сбора данных необходимо::

- 1) определить проблемы и сформулировать цели исследования.
- 2) осуществить предварительное изучение интересного предмета.
- 3) Разработка концепции исследования.
- 4) Сделайте детальное планирование исследования.
- 5) произвести отбор источников информации и сбор вторичных данных.
- 6) оценить полученные данные и принять решение о том, насколько необходимыми являются первичные данные.
- 7) определить метод сбора первичных данных-дознание, наблюдение, эксперимент.
- 8) осуществлять непосредственно сбор первичной информации.
- 9) представить результаты исследования (презентации).

-интуитивные, которые имеют дело с мнениями и оценками экспертов;
- формализованные, которые уже описаны в литературе и на основе которых уже построены прогнозные модели.

Деревья решений широко используются в области среза данных.

Деревья решений - это метод представления правил в иерархической, последовательной структуре, где каждому объекту соответствует отдельный узел, дающий решение.

Все задачи, решаемые методом дерева, могут быть объединены в следующие три класса:

1) описание данных: деревья решений позволяют хранить информацию о данных в компактном виде, вместо них мы можем хранить дерево решений, содержащее точное описание объектов.

2) Классификация: деревья решений прекрасно справляются с задачами классификации, т. е. отнесения объектов к одному из заранее известных классов. Целевая переменная должна иметь дискретные значения.

3) регрессия: если целевая переменная имеет непрерывные значения, деревья решений позволяют установить зависимость целевой переменной от независимых (входных) переменных. Например, к этому классу относятся задачи численного прогнозирования (предсказания значений целевой переменной).

Существует множество алгоритмов, реализующих деревья решений, среди которых CART, C4.5, функция newid, ITrule, CHAID, CN2 пользуются и т. д. Однако наиболее распространенными являются следующие:

CART (Classification and Regression Tree) – алгоритм построения бинарного дерева решений-дихотомизирующей классификационной модели. Каждый узел такого дерева в случае разбиения имеет только два потомка. Алгоритм решает задачи классификации и регрессии.

C4.5 - это алгоритм построения дерева решений, в котором число потомков узла не ограничено.

Базы Интеллектуального Анализа Данных

Говоря об обработке больших объемов данных, вы сталкиваетесь с использованием термина Data mining, который подразумевает наличие огромного количества данных.

Интеллектуальный анализ данных - это процесс поддержки принятия решений, в основе которого лежит поиск скрытых закономерностей (информационных шаблонов) в данных. Именно такая технология предназначена для поиска закономерностей неочевидных, объективных и полезных на практике в больших объемах данных.

Задачи интеллектуального анализа данных иногда называют закономерностями или методами. Основными задачами интеллектуального анализа данных являются: классификация, кластеризация, прогнозирование, объединение, визуализация, анализ и выявление отклонений, оценка, анализ связей и подведение итогов.

Методы и алгоритмы интеллектуального анализа данных:

- искусственная нейронная сеть;
- дерево решений;
- правила характера;
- методы ближайшего соседа и k-ближайшего соседа;
- метод опорных векторов;
- байесовские сети;
- линейная регрессия;
- корреляционный регрессионный анализ;
- иерархические методы кластерного анализа;

- неиерархические методы кластерного анализа, в том числе алгоритмы k-средних и k-медианной линии;
- методы поиска ассоциативных правил, в том числе алгоритм априори;
- метод ограниченного поиска, эволюционное программирование и генетические алгоритмы, различные методы визуализации данных и множество других методов.

Интеллектуальный анализ данных может состоять из двух или трех этапов.

Этап 1. Обнаружение закономерностей (свободный поиск).

Этап 2. Использование выявленных закономерностей для прогнозирования неизвестных величин (прогностическое моделирование).

Этап 3. Анализ исключений-это этап, предназначенный для выявления и объяснения аномалий, которые обнаруживаются в закономерностях.

Визуализация инструментов интеллектуального анализа данных

Каждый из алгоритмов интеллектуального анализа данных использует определенный подход к визуализации. При использовании каждого из методов интеллектуального анализа данных, а точнее, его программной реализации, мы являемся зрителями, с помощью которых можно интерпретировать результаты, полученные в результате работы соответствующих методов и алгоритмов.

Для деревьев решений таким средством просмотра является дерево решений, список правил и таблица непредвиденных обстоятельств.

Для нейронных сетей, в зависимости от инструмента, это может быть топология сети, диаграмма изменения значения ошибки, показывающая обучающую активность.

Для карт Кекконена: карты входов, выходов, другие специфические карты.

Для линейной регрессии линия регрессии отображается как средство просмотра.

Для кластеризации: диаграммы рассеяния дендрограмм.

Диаграммы и графики дисперсии часто используются для оценки качества работы того или иного метода.

Все эти методы визуального представления или отображения данных могут выполнять одну из функций:

- являются иллюстрацией построения модели (например, представления структуры (графа) нейронной сети);
- помогают интерпретировать полученный результат;
- являются средствами оценки качества построенной модели;
- комбинирует вышеупомянутые функции (дерево решений, дендрограммы).

Прежде чем использовать технологию интеллектуального анализа данных, необходимо тщательно проанализировать ее проблемы, ограничения и критические вопросы, связанные с этим, а также понять, что эта технология не может сделать.

Аналитика не может заменить интеллектуальный анализ данных!

Технология не может дать ответы на те вопросы, которые не были заданы. Она не может заменить аналитика, а лишь дает ему мощный инструмент для облегчения и улучшения своей работы.

Сложность разработки и эксплуатации приложения интеллектуального анализа данных. Поскольку эта технология является междисциплинарной областью, для разработки приложений в том числе и для интеллектуального анализа данных, необходимо привлекать специалистов из разных областей, а также обеспечивать их качественное взаимодействие.

Контрольный вопросы

- 1 Что такое анализ данных?
- 2 Что такое данные?
- 3 Как можно классифицировать методы прогнозирования данных?
- 4 Что такое регрессия в анализе данных?
- 5 Что такое визуализация данных?
- 6 Что такое интеллектуальный анализ данных?
- 7 Что такое методы интеллектуального анализа данных?
- 8 Что такое дерево решений?
- 9 Каковы задачи интеллектуального анализа данных?
- 10 Каковы алгоритмы построения дерева решений?