Learning journal Data Mining - COMP40370

Student: Felipe Guth

Student Id: 14210231

15/09/15 – Data Mining

Data mining may sound complicate sometimes, expensive or not necessary for naïve users. The main positive point is the profitable real world uses. Data explosion nowadays generated an opportunity to extract valuable, non-trivial information. “We are drowning in data, but starving for knowledge”.

Definition of DM: Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases. Alternatives names are KDD, business intelligence, etc.

The range of potential applications for data mining is wide; one of the most common is market analysis and management. By taking information from different data sources it is possible to determine patterns of purchase over time, predict future requirements and adapt merchandising actions accordingly with the cluster of clients profile. Other examples of data mining applications are resource planning, planning and asset evaluation in corporations; Fraud detection; Astronomy and Sports.

The knowledge discover in databases (KDD) process, integrates a range of different sub-tasks such as data cleaning, data integration, data warehousing, selection of task-relevant data, data mining (which is the core), pattern evaluation and use of discovered knowledge.

The business intelligence process has the aim of making decisions. Previous steps include data sources, data warehouses, data exploration, data mining and data presentation. Different users participate of this process since DBA to data and business analysts, and end users.

Architecture of a Data Mining system consists of databases and data warehouse, database or data warehouse server, data mining engine, pattern evaluation and graphical user interface as a knowledge-base storage.

A set of different tools commercial and free are available such as RapidMiner, Weka, Tanagra, IBM SPSS modeller, Oracle Data Mining, etc.

17/09/15 – Data Warehousing

Definition of data warehouse by Inmon: "A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process".

Subject-oriented: focusing on subjects as customer, sales, product. Directed to useful issues on decision making, not on daily operations or transactions (databases).

Integrated: Using various data source (databases, flat files, online records). Operations for cleaning and transform the data are needed.

Time Variant: Ever structure on a data warehouse is connected to a time element. The time horizon of the data warehouse is longer than that of operational system.

Non-volatile: Data warehouse are physically separated from data bases in a different storage. Operational update of data does not occur in the data warehouse.

OLTP (On-Line Transaction Processing): Major task of traditional relational DB, day-to-day operations.

OLAP (On-line Analytical Processing): Focused on data warehouse system, data analysis and decision making.

It is important to separate data warehouse from the operational database to improve performance on both systems. Also, DBMS are tuned for OLTP whereas warehouses are tuned for OLAP.

22/09/15 – Data Warehouse – Multi-dimensional Data models

Multidimensional data model which views data in the form of a data are the base of a data warehouse. A data cube allows the display of data in multiple dimensions. Important definitions: An n-Dimensional base cube is called a base cuboid. The top most 0-D cuboid, which holds the highest-level of summarisation, is called the apex cuboid. The lattice of cuboids forms a data cube. In the conceptual modelling of data warehouses different schemas applied: Star Schema: Consists of a fact table in the middle connected to other dimension tables. Snowflake Schema: Refinement of star schema with some dimensional hierarchy is normalised into a set of smaller dimension tables. Fact Constellations: Multiple fact tables sharing dimension tables.

DMQL

Definition sample (star schema):

define cube sales_star [time, item, branch, location]:

Euros_sold = sum(sales_in_Euros),

avg_sales = avg(sales_in_Euros),

units_sold = count(*)

define dimension time as

(time_key, day, day_of_week, month, quarter, year)

define dimension item as

(item_key, item_name, brand, type, supplier_type)

define dimension branch as

(branch_key, branch_name, branch_type)

define dimension location as

(location_key, street, city, county, province, country)

24/09/15 - Data Warehouse – Multi-dimensional Data models

Measures:

Distributive: when the result of applying the function to n aggregate values is the same as apply to all data without partitioning. count(), sum(), etc. Algebraic: computed by algebraic function with M arguments, each obtained with a distributive aggregate function. Avg(), standard_deviation(), etc. Holistic: if there is no constant bound on the storage size needed to describe a sub-aggregate. Median(), mode(), rank(). Typical OLAP operations are Roll up (drill-up), Drill down (roll down), Slice and dice (project and select), Pivot (rotate). Views of design a data warehouse: Top-down view: allows of the relevant information necessary for data warehouse; Data source view: exposes information gathered from operational systems; Data warehouse view: fact tables dimension tables; Business query view: sees the data of warehouse from end-user perspective; Data warehouse Design process: Top-down: overall design and planning. Bottom-up: Experiment prototypes. From SW engineering point of view: Steps : planning, data collection, DW design, test and evaluation, DW deployment. Waterfall. Spiral. The tree data warehouse models are: Enterprise warehouse: collects information from entire organisation. Data Mart: a subset of corporate-wide data that is of value to a specific group of users. Virtual warehouse: a set of views over operational databases. OLAP server architectures: Relational OLAP (ROLAP): Use DBMS to store and manage warehouse data and OLAP middleware to support missing pices. Multidimensional OLAP (MOLAP): Array-based multidimensional storage engine. Hybrid OLAP (HOLAP): User flexibility, low level: relational, high-level: array. Specialised SQL servers: specialised support for SQL queries over star/snowflakes schemas.

29/09/15 – Data Pre-processing – Descriptive Data Summarisation

Data pre-processing is necessary to correct faulty data of the real world. A quality mining is dependent of good quality data. Data Quality: Perfect data: data is valid, complete. Not Perfect data: Data with no-serious flaws, but need some pre-processing. Verbal/Inspection data: Data with serious gaps. Soft data: Data relied on the memories of experienced personnel. Major tasks in Data Pre Processing are data cleaning which consists of fill in missing values, correct noisy data, remove outlier and resolve inconsistencies; Data integration that is integration of different data sources such

as databases and files; Data transformation which is defined by normalisation and aggregation operations; Data reduction that is reduce the representation in volume but generate same or similar analytical results.

Data Summarisation is the identification of properties of the data in descriptive properties or statistics. Important statistics are mean, median, standard deviation in conjunction with data dispersion as quartiles, inter-quartile range, and variance. Forms of data displaying and analysis are boxplot, histograms, quantile plot and scatter plot.

01/10/15 – Data Pre-Processing

In this lecturer we learned about data cleaning, data integration, data transformation and data reduction and discretisation. Data cleaning compress the activities of filling missing values, identify and treat noisy and inconsistent data. Workarounds involve the binning method, clustering and regression techniques to smooth data and remove outliers. Binning consists in dividing the data in intervals and then assign to each item a new value based on this new partition ranges.

Data integration is the process responsible to combine data from different sources. This process requires a careful analysis for redundant data and conflict representations. Redundant data can be detected using correlation-based analysis.

In the data transformation, Normalization scales the data to a specified range; the most used techniques are min-max normalisation and z-score.

Data reduction techniques seek to obtain a reduced representation of the dataset using a small quantity of the original data but yet producing similar results in the analytical solutions. Common techniques for data reduction are data cube aggregation, dimensionality reduction (feature selection), numerosity reduction, discretisation and concept hierarchy generation.

06/10/15 – Association Rules

Association rules are methods that seek to discover relations between items in a dataset. The discovery of frequent items relations provide information that can be used for decision making process.

{eggs,bread}=>{milk} , found in the sales transaction that eggs and bread together are likely to indicate also milk buying.

Support is the value that defines the frequency of an item or item-set defined as the proportion of transactions in a database which contains this item or item-set. If an item X appears 1 out of 5 times in a database its support is 1/5=0.2.

Confidence is a value that defines the proportion of transactions that contains an item or item-set X should also contain Y. conf(X=>Y)=sup(X U Y)/sup(X).

13/10/15 – Classification

The classification problem separates data accordingly to a given model and a training set in determined class labels creating a classifier that can be used to classify new entries. Prediction is the use of models to predict future results based on previous experiences that are learned by a classifier. In the supervised learning (classification) the training data is identified with labels indicating the respective class of each entry. New data is classified based on the training set learning. In the unsupervised learning (clustering) the data labels are unknown and the system tries to partition the data in different classes, named clusters, by measuring the similarity between items of the dataset.

In order to reduce the dimensionality of the problem and use the most relevant variables, attribute selection methods are applied, such as information gain and gini index. The information gain algorithm selects the attributes with highest information gain. By using the entropy measure, it is calculate the importance of a determined attribute to classify the dataset. Low entropy results in a good data separation. The Gini index measures the impurity of data and split the data in n classes.

20/10/15 - Classification

Classification in Large Datasets. While machine learning algorithms are interested in the end goal of learning, data mining are focus on discovering actual patterns, the algorithms must be capable of handling very large datasets with reasonable speed.

Decision three induction in data mining is relative fast is simply to understand and convertible in classification rules. The accuracy is comparable with other algorithms. Subclasses are SLQ, SPRINT, PUBLIC and RainForest.

Bayesian classifiers calculate probabilities for certain hypothesis of learning problems. In an incremental way prior knowledge is combined with observed data and the hypothesis are updated in each iteration. The cost of this algorithm is $O(n^2)$. $P(y/x) = P(x,y)*p(y)/p(x)$

Naïve Bayes classifier assumes that attributes are independent which reduces the computational cost. It works well when there is no correlation between the attributes.

22/10/15 - Clustering

Cluster is a collection of data objects that are similar to each other. Clustering analysis seeks to separate the dataset in distinct clusters that have different characteristics. This is an unsupervised learning where there are no pre-defined classes. Application are pattern recognition, image processing, spatial data analysis, economic science, marketing, insurance.

A good clustering method is the one that group very similar objects in a cluster. Requirements are scalability, ability to deal with noisy data and different types of attributes, etc.

27/10/15 - Clustering

There are different methods to measure the similarity/dissimilarity in clusters depending on the type of variables taken into account. The definitions of distance functions are usually very different for interval-scaled, Boolean, categorical, ordinal and ratio variables. The distance on interval-valued variables is calculated by absolute deviation (zscore). The distance between normal objects in usually calculated with Euclidean distance. In binary variables methods to measure the distance are contingency table, simple matching coefficient and Jaccard coefficient (if the binary variable is asymmetric). Nominal variables distance is measured by simple matching or by the creation of binary variables accordingly to the data and then measuring the distances. The distance in ordinal variables is measured by using the rank in the order, in an interval-scaled approach. For ration-scaled variables, one method is to treat them like interval-scaled variables; another option is apply logarithmic transformation or treat them as continuous ordinal data and treat their rank as interval-scaled. Major clustering techniques are partitioning algorithms (like k-means), hierarchy algorithms, density-based, grid-based and model based. I never heard about the three last techniques.

27/10 (Evening) Clustering

Today we learned how to write a learning journal. Also, K-means algorithm has a complexity of t*n*t, where n = dataset size, k = numbers of clusters, t=number of iterations. It is implemented in 4 steps. First of all, the user selects how many subsets k will be calculated (i.e. the number of clusters). Secondly, the centroids of the data set are calculated. Thirdly, each object is assigned to the nearest cluster. Lastly, the second and third are repeated until the clusters do not change. It often terminates at a local optimum. Weaknesses are the guessing the number of cluster k. It is unable to handle noisy data and outliers and it is not suitable to discover clusters with non-convex shapes, it is appropriate to find circle-like clusters. For categorical variables the PAM algorithm is used.

Hierarchical clustering. The Agglomerative algorithm in the beginning is started with n clusters. In the forward steps the clusters are merged accordingly with the distance matrix between them. The Devise algorithm is the reverse of the agglomerative algorithm. Each step specialize, unmerge the clusters accordingly with dissimilarity. A dendogram shows how the clusters are merged and organized. Hierarchical clustering methods have the complexity of n^2 and can never undo what was done previously.

K-means find clusters in a radial way. It is not a good solution to find uneven clusters and non-convex shapes.

29/10/15 - Clustering

K-means is a distance-based method that has the drawback of not being able to detect non-convex shapes. An option for this problem is the CURE algorithm. It uses multiple representative points to evaluate the distance between clusters. Given the number of points and partition the algorithm split the data to p partitions with size s/p. Then, partially cluster partitions into s/pq clusters. Eliminate outliers. Cluster partial clusters and then the data is label in disk. For clustering categorical data it is

used the ROCK algorithm that uses similarity function to detect the difference between elements. Density based clustering algorithms (Good solution but costly) have as features the discovering of clusters of arbitrary shapes, handle noise, one scan of the data and needing of density parameters and termination condition. The parameter Eps determines the maximum radius of the neighbourhood and MinPts is the minimum number of points in Eps-neighbourhood of that point. A point is Density-reachable from a point q with regard to Eps if there is a chain of density-reachable points between them. Density connected means that there is a continuity of the density through regions of points in the clusters.

DB-Scan algorithm is density-based. A cluster is defined as the max set of density-connected points.