Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 1, January 2014, pg.519 – 525

RESEARCH ARTICLE

Predictive Data Mining: A Generalized Approach

Meghana Deshmukh
Lecturer, Dept. of Computer Science and Engineering, Prof. Ram Meghe Institute of Technology &
Research Badnera,
Amravati, India
meghnadeshmukh9@gmail.com

Prof. S. P. Akarte
Asst. Prof. Dept. of Computer Science and Engineering, Prof. Ram Meghe Institute of Technology &
Research Badnera,
Amravati, India
s_akarte25@rediffmail.com

Abstract— In this paper, we included the ambitious task of formulating a general framework of data mining. We explained that the framework should fulfil. It should elegantly handle different types of data, different data mining tasks, and different types of patterns/models. We also discuss data mining languages and what they should support: this includes the design and implementation of data mining algorithms, as well as their composition into nontrivial multi step knowledge discovery scenarios relevant for practical application. We proceed by laying out some basic concepts, starting with (structured) data and generalizations (e.g., patterns and models) and continuing with data mining tasks and basic components of data mining algorithms (i.e., refinement operators, distances, features and kernels). We next discuss how to use these concepts to formulate constraint-based data mining tasks and design generic data mining algorithms. Finally this paper discussed about these components would fit in the overall framework and in particular into a language for data mining and knowledge discovery.

Keywords— data mining; data mining cycle; patterns; data mining methods; tasks

1. Introduction

When knowledge discovery in databases (KDD) and data mining have enjoyed great popularity and success in recent years, there is a distinct lack of a generally accepted framework for data mining. The present lack of such a framework is perceived as an obstacle to the further development of the field.

Much of the current research in data mining is about mining complex data, e.g., text mining, link mining, mining social network data, web mining, multi-media data mining.

As the complexity of the data analyzed grows, more expressive formalisms are needed to represent patterns found in the data. The use of such formalisms has been proposed within relational data mining and statistical relational learning; these are now used increasingly more often in link mining, web mining and mining of network data. Data preparation typically takes significant time and different data mining operations need to be applied and composed in practical applications. Arguably, there is insufficient support for humans carrying out the knowledge discovery process as a whole. Integration and compositionality of data mining operations/algorithms are called for.

2. The Inductive Databases and Queries

Inductive databases are an focusing area at the intersection of data mining and databases. In addition to normal data, inductive databases contain patterns. Besides patterns, models can also be considered. Inductive databases embody a database perspective on knowledge discovery, where knowledge discovery processes become query sessions. Ordinary queries can be used to access and manipulate data, while inductive queries (IQs) can be used to generate, manipulate, and apply patterns. KDD thus becomes an extended querying process in which both the data and the patterns that hold in the data are queried. the traditional KDD process model, where steps like pre-processing, data cleaning, and model construction follow each other in succession, by a simpler model in which all data pre-processing operations, data mining operations, as well as post-processing operations are queries to an inductive database and can be interleaved in many different ways. Given an inductive database that contains data and patterns, several different types of queries can be posed. Data retrieval queries use only the data and their results are also data: no pattern is involved in the query. In processing patterns, the patterns are queried without access to the data: this is what is usually done in the post-processing stages of data mining. Data mining queries use the data and their results are patterns: new patterns are generated from the data and this corresponds to the traditional data mining step.

When we talk about inductive queries, we most often mean data mining queries. A general statement of the problem of data mining involves the specification of a language of patterns and a set of constraints that a pattern has to satisfy. The latter can be divided in two parts: language constraints and evaluation constraints. The first part only concerns the pattern itself, while the second part concerns the validity of the pattern with respect to a given database. The IDB framework is appealing for data mining applications, as it supports the entire KDD process In inductive query languages, the results of one (inductive) query can be used as input for another: nontrivial multi-step KDD scenarios can be thus supported in IDBs, rather than just single data mining operations. More importantly, most of the existing approaches to constraint-based data mining and inductive querying work in isolation and are not integrated with databases or other data mining tools. Only few attempts at integration have been made, such as the approach of mining views. Answering complex inductive queries that involve different pattern domains and supporting complex KDD scenarios has also barely been studied.

3. The Basic Concepts of Data Mining

"Knowledge discovery in databases (KDD) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data",. According to this definition, data mining (DM) is the central step in the KDD process concerned with applying computational techniques to actually find patterns in the data. To arrive at a general theory/framework for data mining, we need to have general definitions for the above terms, including data, patterns and validity. The basic concepts of data mining include data, data mining tasks, and patterns/models. The validity of a pattern/model on a given set of data is related to the data mining task considered.

3.1 The Data Mining Task

The data mining tasks are of different types depending on the use of data mining result the data mining tasks are classified as:

3.1.1 Exploratory Data Analysis :

In the repositories vast amount of information's are available .This data mining task will serve the two purposes

(i).Without the knowledge for what the customer is searching, then

(ii) It analyzes the data These techniques are interactive and visual to the customer.

3.1.2 Descriptive Modeling:

It describe all the data, it includes models for overall probability distribution of the data, partitioning of the p-dimensional space into groups and models describing the relationships between the variables.

3.1..3 Predictive Modeling:

This model permits the value of one variable to be predicted from the known values of other variables.

3.1.4. Discovering Patterns and Rules :

This task is primarily used to find the hidden pattern as well as to discover the pattern in the cluster. In a cluster a number of patterns of different size and clusters are available .The aim of this task is "how best we will detect the patterns" .This can be accomplished by using rule induction and many more techniques in the data mining algorithm like. These are called the clustering algorithm.

3.1.5 Retrieval by Content:

The primary objective of this task is to find the data sets of frequently used in the for audio/video as well as images It is finding pattern similar to the pattern of interest in the data set

4. The Dual Nature: Patterns and Models

Patterns and models inherently have a dual nature. According to the definitions from the previous sections, they are functions that take as input data points and map them to probabilities, Booleans, class predictions or probabilities thereover, or cluster assignments. On the other hand, they can be treated as data structures and as such represented, stored and manipulated. For simple example: Suppose we have a frequent itemset consisting of the items bread and butter. We can view this as a set, namely {bread,butter}, and store it in a database. In this fashion, we can store the frequent itemsets derived from a set of transactions. On the other hand, from the functional viewpoint, it represents a mapping from transactions to Booleans. The transactions which contain the itemset, i.e., both bread and butter, are assigned the value true, i.e., the pattern holds true for such transactions. For example, the transaction {bread,butter,milk} subsumes our itemset and yields the value true, while {beer,peanuts,butter} does not and yields the value false.

4.1 Classes of Patterns and Models in data aspect

Many different kinds of predictive models have been considered in the data mining literature. Classification rules, decision trees and linear models are just a few examples. We will refer to these as model classes. In the case of patterns we will talk about pattern classes. A class of patterns CP on type T is a set of patterns P on type T, expressed in a language LP . Similarly, a class of models CM on types Td, Tc is a set of models M on types Td, Tc, expressed in a language LM. In the same fashion, we can define classes of probability distributions CD and clusterings CC. The languages LP/LM/LD/LC refer to the data part of the patterns and models. They essentially define data types for representing the patterns and models. For example, if we have data types Td = (Real, Real) and Te = Real, linear models would be represented by three real-valued coefficients and would need a data typeTl = (Real, Real, Real) to be represented. Suppose we have a dataset where data items correspond to descriptions of individuals, each individual being described by a tuple of the form (Gender, Age, HairColor), where Gender = Discrete({M,F}), Age = Real, HairColor = Discrete({Blond,Brown,Black,Red,Other}), and the target is of type Education = Discrete({None,Elementary,High,College,BSc,MSc,PhD}). The language of decision trees for this case would be the language of tree struc- tures with tests like HairColor=Blond in the internal nodes and predictions like Education=PhD in the leaves. The elements of this language (its alphabet) de- pend on the attributes and their values, and vary with the underlying data type.

4.2 Interpreters in function aspect

There is usually a unique mapping from the data part of a pattern/model to the function part. This takes the data part of a pattern/model as input, and returns the corresponding function as an output. The mapping we refer to is inherently second/higher order (Lloyd 2003) since it has a function as an output. This mapping can be realized through a so-called interpreter. An interpreter takes as input (the data part of) a pattern and an example, and returns the result of applying the (function part) of the pattern to the example. Given a data type d, an example E of type d, and a pattern P of type p :: d→ bool, an interpreter I returns the result of applying P to E, i.e., I(P,E)=P(E). The signature of the interpreter is i :: p→ d → bool. If we apply the interpreter to a pattern and an example, we obtain a Boolean value. In functional programming (Thompson 1999), we can evaluate the interpreter only partially, i.e., apply it only to the data part of a pattern, obtaining as a result the function part of the pattern. The partial evaluation iphas a signature d → bool. The interpreters map from the data part of a pattern/model to the function part. Suppose we are given a linear model with coefficients a, b, andc. The interpreter of linear models Il would, given a, b, andc, and a data tuple of the form (x,y), return the value of the linear combination ax + by + c. A partial evaluation/application of the interpreter to the tuple of constant coefficients of the linear model Il (a,b,c) would yield the linear function aX + bY + c: This linear function can then be applied to specific tuples (x,y) to yield predictions. The interpreter is crucial for the semantics of a class patterns/models: a class of patterns/models is only completely defined when the corresponding interpreter is defined (e.g., IP /IM for patterns/models are parts of the definition of the class CP /CM). To illustrate this, consider rule sets, which may be ordered or unordered. Both can actually be represented by the same list of rules: It is the interpreter that treats the rules as ordered or unordered. In the first case, the rules are considered in the order they appear in the list and the first rule that applies to a given example is taken to make a prediction. In the second case, all rules from the list that apply to a given example are taken, and their predictions combined to obtain a final prediction.

5. Data Mining Life Cycle

The life cycle of a data mining project consists of six phases. The sequence of the phases is not rigid. Moving back and forth between different phases is always required. It depends on the outcome of each phase. The main phases are:

5.1. Business Understanding:

This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

5.2 Data Understanding:

It starts with an initial data collection, to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

5.3 Data Preparation :

In this stage , it collects all the different data sets and construct the varieties of the activities basing on the initial raw data.

5.4 Modeling:

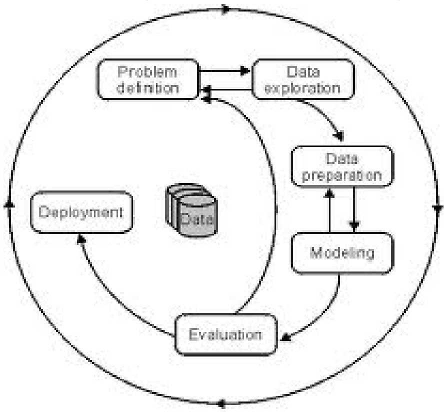
In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.

5.5 Evaluation:

In this stage the model is thoroughly evaluated and reviewed. The steps executed to construct the model to be certain it properly achieves the business objectives. At the end of this phase, a decision on the use of the data mining results should be reached.

5.6 Deployment :

The purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. The deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.



6. Data Mining Methods:

- Some of the popular data mining methods are as follows:
- Decision Trees and Rules
- Nonlinear Regression and Classification Methods
- Example-based Methods
- Probabilistic Graphical Dependency Models
- Relational Learning Models

6.1 The Scope of Data Mining

Data mining derives its name from the similarities between searching for valuable business information in a large database for example, finding linked products in gigabytes of store scanner data and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can prediction of trends and behaviors .

Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

• Artificial neural networks :

Non-linear predictive models that learn through training and resemble biological neural networks in structure.

• Decision trees :

Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) .

• Genetic algorithms :

Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.

• Nearest neighbor method :

A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k ³ 1). Sometimes called the k-nearest neighbor technique.

• Rule induction :

The extraction of useful if-then rules from data based on statistical significance. Many of these technologies have been in use for more than a decade in specialized analysis tools that work with relatively small volumes of data. These capabilities are now evolving to integrate directly with industry-standard data warehouse and OLAP platforms.

7. Related Work

When attempting to formulate a general framework for data mining, the potential set of related work items is dangerously large. Here we will give a biased sample of what we consider related work. Parts of it have been mentioned previously, while others have not been explicitly mentioned above even though they have made an intellectual influence during the writing of this article. Let us start with inductive databases and constraint-based data mining. Since the notion of inductive databases was introduced, a significant body of research has grown on these two topics: A survey can be found in the book edited by Boulicaut et al. (2005). An earlier collection of papers focussing on constraint- based data mining was edited by Bayardo (2002). Data mining query languages are also directly relevant: A survey article is presented by Boulicaut and Masson (2005). A more recent proposal for an SQL- based data mining query languages, which allows for the integration of various data mining operations at the data level, has been given by Kramer et al. (2006).

Finally, the IQL language proposed by Nijssen and De Raedt (2007/this volume), is very close in spirit to the discussion presented here: it recognizes the importance of functions and extends tuple relational calculus with a function and a typing system. However, it only allows for loose integration of data mining algorithms and does not support the creation of new algorithms. Another way to recognize the importance of functions is to use higher-order logic or functional programming to facilitate the implementation of data mining algorithms (for mining structured data). Lloyd (2003) uses higher-order logic to define structured data types and principled ways of constructing distances, features (which he calls predicates) and kernels. Allison (2004) uses functional programming to define data types and type classes for models (where models include probability distributions, mixture models and decision trees) that allow for models to be manipulated in a precise and flexible way.

Formulating an algebra for data mining that would be the equivalent of Codd's relational algebra for databases is probably the most ambitious goal in the con- text of the discussion presented here. The 3W-model (Johnson et al. 2000) was among the first to take an algebraic view on data mining: Finally, the compositionality of data mining operators, as discussed by Ramakrishnan et al. (2005), can be expected to play a crucial role in the general framework.

Conclusion

In this paper we briefly reviewed the various data mining applications. This review would be helpful to researchers to focus on the various issues of data mining. In future course, we will review the various classification algorithms and significance of evolutionary computing (genetic programming) approach in designing of efficient classification algorithms for data mining. This previous studies on data mining applications in various fields use the variety of data types range from text to images and stores in variety of databases and data structures. The different methods of data mining are used to extract the patterns and thus the knowledge from this variety databases. Selection of data and methods for data mining is an important task in this process and needs the knowledge of the domain. Several attempts have been made to design and develop the generic data mining system but no system found completely generic. Hence, for every domain the domain expert's assistant is compulsary. The domain experts shall be guided by the system to effectively apply their knowledge for the use of data mining systems to generate required knowledge. The domain experts are required to determine the variety of data that should be collected in the specific problem domain, selection of specific data for data mining, cleaning and transformation of data, extracting patterns for knowledge generation and finally interpretation of the patterns and knowledge generation.

References

[1]Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.

[2] Larose, D. T., "Discovering Knowledge in Data: An Introduction to Data Mining", ISBN 0-471-66657-2, ohn Wiley & Sons, Inc, 2005.

[3] Dunham, M. H., Sridhar S., "Data Mining: Introductory and Advanced Topics", Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1 st Edition, 2006

[4] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R... "CRISP-DM 1.0 : Step-by-step data mining guide, NCR Systems Engineering Copenhagen (USA and Denmark), DaimlerChrysler AG (Germany), SPSS Inc. (USA) and OHRA Verzekeringenen Bank Group B.V (The Netherlands), 2000".

[5] Fayyad, U., Piatetsky-Shapiro, G., and Smyth P., "From Data Mining to Knowledge Discovery in Databases," AI Magazine, American Association for Artificial Intelligence, 1996.

[6] Tan Pang-Ning, Steinbach, M., Vipin Kumar. "Introduction to Data Mining" , Pearson Education, New Delhi, ISBN: 978-81-317-1472-0, 3 rd Edition, 2009. Bernstein, A. and Provost, F., "An Intelligent Assistant for the Knowledge Discovery Process", Working Paper of the Center for Digital Economy Research, New York University and also presented at the IJCAI 2001 Workshop on Wrappers for Performance Enhancement in Knowledge Discovery in Databases.

[7] Baazaoui, Z., H., Faiz, S., and Ben Ghezala, H., "A Framework for Data Mining Based Multi-Agent: An Application to Spatial Data, volume 5, ISSN 1307-6884," Proceedings of World Academy of Science, Engineering and Technology, April 2005.

[8] Rantzau, R. and Schwarz, H., "A Multi-Tier Architecture for High-Performance Data Mining,A Technical Project Report of ESPRIT project, The consortium of CRITIKAL project, Attar Software Ltd. (UK), Gehe AG (Denmark); Lloyds TSB Group (UK), Parallel Applications Centre, University of Southampton (UK), BWI, University of Stuttgart (Denmark), IPVR, University of Stuttgart (Denmark)".

[9] Botia, J. A., Garijo, M. y Velasco, J. R., Skarmeta, A. F., “A Generic Data mining System basic design and implementation guidelines”, A Technical Project Report of CYCYTprojectofSpanish Government.1998.WebSite: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.53.1935

[10] Campos, M. M., Stengard, P. J., Boriana, L. M., “Data-Centric Automated Data Mining”, , Web Site.: www.oracle.com/technology/products/bi/odm/pdf/automated_data_mining_paper_1205.pdf

[11] Sirgo, J., Lopez, A., Janez, R., Blanco, R., Abajo, N., Tarrio, M., Perez, R., “A Data Mining Engine based on Internet, Emerging Technologies and Factory Automation,” Proceedings ETFA '03, IEEEz Conference,16-19Sept.2003.WebSite:www.citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.11.8955

[12] The Survey of Data Mining Applications.pdf

[13]Towards a General Framework for Data Mining.pdf