IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-4, Issue-5 E-ISSN: 2347-2693

Survey on Attribute Oriented Induction Using Data Mining Techniques

S. Radha Priya[1*] and M. Devapriya[2]

[1] PG and Research Department of Computer Science, Government Arts College(Autonomous), Coimbatore, India.
[2] PG and Research Department of Computer Science, Government Arts College(Autonomous), Coimbatore, India.

Available online at: www.ijcseonline.org

Received: Apr/21/2016 Revised: May/04/2016 Accepted: May/18/2016 Published: May/31/2016

***Abstract*— Data and objects in databases often contain detailed information at primitive concept levels. It is useful to summarize a large set of data and present it at a high conceptual level. Attribute Oriented Induction(AOI) is a set-oriented data base mining method which generalizes the task-relevant subset of data attribute-by-attribute compresses it into a generalized relation and extracts from it the general features of data. The power of AOI is extraction from relational databases of different kinds of patterns including characteristic rules, discriminant rules, cluster description rules and multilevel association rules. The method is efficient, robust with wide applications and extensible to knowledge discovery in advanced database systems, including object-oriented, deductive and spatial database systems. This paper describes the broad classification of data mining techniques using AOI.**

Keywords- AOI, Clustering, Data mining, generalization.

I. INTRODUCTION

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can be databases, data warehouses, the web, other information repositories or data that are streamed into the system dynamically. There are a number of data mining functionalities. These include characterization and discrimination, the mining of frequent patterns, associations and correlations, classification and regression, clustering analysis and outlier analysis. Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks. Tasks can be classified into two categories: descriptive and predictive. Descriptive mining task characterize properties of the data in a target data set. Predictive mining tasks performs induction on the current data in order to make predictions.

II. BASIC METHODLOGY OF AOI

The general idea of AOI is to first collect the task-relevant data using a database query and then perform generalization based on the examination of the number of each attribute's distinct values in the relevant data set. The generalization is performed by either attribute removal or attribute generalization. Aggregation is performed by merging identical generalized tuples and accumulating their respective counts. This reduces the size of generalized data set. The resulting generalized relation can be mapped into different forms(e.g.,charts or rules) for presentation to the user. The following shows the basic methodology of AOI:

Data base → DM Query → Get Task Relevant Data →Prepare for Generalization →Prime Generalized Relation

I. III. FEASIBILITY OF AOI

A major strength of AOI over tuple oriented one is its efficiency in the induction process. Different possible combinations should be explored only when the relation has been generalized to a relatively small prime relation. AOI is robust and handles noise in exceptional cases elegantly and incorporates statistical information(using count) and generates disjunctive rules.

Count association facilitates incremental learning in large databases. When a new tuple is inserted into a database, its concepts are first generalized to the same concept level as those in the generalized relation and then merged naturally into the generalized relation. Association of count information, data sampling and parallelism can be explored in knowledge discovery.

As a generalization based method, AOI confines its power to the discovery of knowledge rules at general concept levels. The method is useful in most database-oriented applications which need to generalize some or all of the relevant attributes.

IV. AOI USING DATA MINING TECHNIQUES

AOI was first proposed, before the introduction of the data cube approach. Data cube can be viewed as a kind of multidimensional data generalization. Data generalization summarizes data by replacing relatively low-level values with higher-level concepts, or by reducing the number of dimensions to summarize data in concept space involving fewer dimensions. A concept hierarchy defines a sequence of mappings from a set of concepts to their higher level correspondences. It is usually partially ordered according to a general-to-specific ordering, with the most general concept usually defined by a reserved word "any" and the most specific concepts corresponding to the specific data in the database. Using a concept hierarchy, the discovered rules can be represented in terms of generalized concepts and stated in a simple and explicit form, which is desirable to most users.

Concept hierarchies can also be generated automatically based on data semantics/ data distribution statistics. Different concepts based on different viewpoints or preferences. A concept hierarchy can also be in the shape of lattice or DAG(Directed Acyclic Graph). It is sometimes preferable to perform induction in parallel along with more than one hierarchy and select a promising one based on some intermediate generalization rules. David wailok chenung suggested that a rule based concept graph substantially increase the representation and induction power of AOI generalization[1]. Jagannatha reddy suggested the technique of extracting prediction rules from the predicted class label has the advantage that the number of units required can be reduced using attribute relevance analysis[2]. Yu-Ying Wu-yen- Liang Chen and ray-I- Chang applied AOI in two novel mining approaches to generate multiple-positive and multiple negative generalized knowledge[3].

Carter and Hamilton proposed more efficient methods of AOI[4]. Cheung proposed a rule based conditional concept hierarchy AOI and thereby allows different tuples to be generalized through different paths depending on other attributes of a tuple[1]. Hsu extended the basic AOI algorithm for generalization of numeric values[5]. Chen and shen proposed a dynamic programming algorithm, based on AOI to find generalized knowledge from an ordered list of data[6]. Savasere and Yuan combines positive associations with domain knowledge and only part of the rules are focused[7].

A hybrid rule induction strategy is proposed by Jiang and Abidi[8]. The star schema is examined by Warnars with AOI based on characteristic rule and non rule based concept[9]. AOI and Emerging Pattern[EP] is to use its typical strength of extracting high level emerging knowledge of data. An attribute-oriented concept tree ascension technique is applied in generalization, which reduces the computational complexity of database learning processes[10].

A hybrid rule induction strategy is proposed by Jiang and Abidi[8]. The star schema is examined by (Warnars)[9] with AOI based on characteristic rule and non rule based concept. AOI and Emerging Pattern[EP] is to use its typical strength of extracting high level emerging knowledge of data. An attribute-oriented concept tree ascension technique is applied in generalization, which reduces the computational complexity of database learning processes(Jiawei Han, Cai, Nick Cercone)[10].

Time complexity is derived using Path-Id by Hoi-Yee which helps to generalize the concept of the attributes in the database relations[11]. Shu-meng Huang designed a new classification algorithm to classify the multi-valued attributes[12]. Lukas Tanutama improved corporate users activities using slightly modified AOI approach to mine the network data[13]. Son Dao and Bad Perry proposed an application of data mining techniques to heterogeneous database schema integration[14]. The syntax attribute-oriented induction algorithm(SAOI) for stock prices and open financial statements are listed for textile and financial industries by Victor[15]. Isabella evaluates a range of feature selectors systematically with respect to their efficiency in improving the performance of the classifiers for sentiment analysis[16,17,18,19].

A. AOI for mining Characteristic rules:

Characteristic rule is an assertion which characterizes a concept satisfied by all or majority number of examples in the class undergoing learning(target class). For example, the symptoms of a specific disease can be summarized by a characteristic rule[12].

For mining characteristic rules, the AOI is performed in the following steps:

- Initial data collection
- Derivation of the generalization plan for each attributes
- Prime relation derivation

Each mining request is conditioned on a subset of attributes identified as "common" between multiple databases[1]. AOI influences AOI-HEP(High Level Emerging Pattern) by using AOI characteristics rule algorithm which was run twice with two input datasets, derived from the same dataset in order to create two rule sets which are then processed with HEP algorithm[16]. The novel star schema is more powerful than current AOI, since it can produce small number final generalization tuples and there is no ANY in the results[9].

B. Discovery of different kinds of rules:

AOI can be used to discover discriminant rules, cluster description rules and multi-level association rules.

i)Mining Discriminant rules:

A discriminant rule is an assertion which discriminates concepts of the class being examined (the target class) from other classes (called contrasting classes). A discriminant rule can be discovered by generalizing the data in both target class and contrasting classes synchronously in an attribute oriented fashion and excluding the properties that overlap in both classes in the generalized rule. The machine-learning techniques such as decision tree method like ID3 and C4.5 have been used to classify the objects and find discriminating behaviours.

ii) Mining Cluster Description rules:

In a large database, it is often desirable to cluster data according to data semantics called conceptual clustering and associate rules with such clusters.

iii) Mining Multiple-Level Association Rules:

An association rule represents an association relationship among a set of patterns(values) in a database. The method first finds large patterns at a high concept level and progressively deepens the search to find such patterns among their descendants at lower conceptual levels until it reaches the primitive level concepts in the database. Wu and Antonie[2009] presented a Apriori based framework for mining positive and negative associative rules.

C. AOI in advanced database systems:

The extension of AOI to knowledge discovery in advanced and/or special purpose databases, including object-oriented, deductive and spatial databases.

i)AOI in Object-Oriented Databases:

An object-oriented database organizes a large set of complex objects into classes which are in turn organized into class/subclass hierarchies with rich data semantics. Each object in a class is associated with

- An object-identifier
- A set of attributes which may contain sophisticated data structures, set-or-list-valued data, class composition hierarchies, multimedia data.
- A set of methods which specify computational routines or rule associated with object class.

ii)Integration of Knowledge discovery and deductive database techniques:

A deductive database is a database which consists of data, rules and integrity constraints and performs deductive reasoning on large set of data. Integration of deduction and induction mechanisms not only leads to discovery of new knowledge in deductive databases but also enhances the power of knowledge discovery mechanisms.

iii)AOI in spatial databases:

A spatial database system stores, manages and manipulates spatial (space-related) data, in most cases nonspatial data as well for geographic information systems and many other applications. The method used to represent the spatial and nonspatial data is spatial-dominant generalization and nonspatial dominant generalization.

Thus AOI approaches yield better results as combining different techniques by overcoming the drawback of each other and resulting in higher accuracy. Table 1 presents few AOI approaches using data mining:

TABLE I: Compilation of Attribute Oriented Induction in data mining

Author Name	Method Used	Methodology	Pros and Cons
Son Dao and Brad Perry(1995)	Schema Integration Process	Preclustering technique is used in heterogeneous databases and AOI is used to mine characteristic rules and classification rules	**Pros**: Generate discretized domain readily applicable to mining /integration techniques. High degree of coverage of data instances. **Cons**: Parallel mining techniques needed.
Qinshuang Jiang and Syed Sibte Raza Abidi	Decision Rule Acquisition workbench (DRAW)	Conceptual clustering and Rough set based rule induction algorithm and then AOI to yield high level decision rules and a minimal rule-set size.	**Pros:** Overall classification accuracy is high. **Cons:** It does not impose any static statistical parameters upon the data.
Spits-Warnars (2015)	Attribute Oriented High level Emerging Pattern (AOI-HEP)	AOI characteristic rule algorithm is used followed by high level emerging pattern algorithm	**Pros**:AOI-HEP pattern results in high level data.
Spits Warnars	Noval Star	Elimination of threshold number	**Pros**:Produce small number of final

Author Name	Method Used	Methodology	Pros and Cons
(2010)	Schema	to control generalization	generalization tuples. No ANY in the result. More concept trees and more table concept tree must be created for join operation in Sql statements, automatically decrease query performance.
Jiawei Han, Yandong Cai and Mick Cercone m (1992)	Set-oriented database operation	Tree ascension technique is applied in generalization.	**Pros**: Reduces the computational complexity of database learning process. **Cons**:Experimentation has not been done on large size data.
Hoi-Yee Hwang and Wai-Chee Fu (1995)	Path-Id is used for efficient generalization	Top-down approach and bottom up approach.	**Pros**:The amount of disk I/O is O(log N) times less than the original method of AOI. **Cons**:For each generalization step we have to process all tuples of the path relation.
Yu Ying Wu,, Yen-Liang Chan and Ray-I-Chang (2009)	Generalized positive knowledge induction (GPKI) and Generalized negative knowledge induction (GNKI).	Generate multiple positive and negative generalized knowledge.	**Pros:**Simple and efficient way for knowledge generalization from a relational database. The method is efficient and scalable for relational databases.
Shu-meng Huang, Ping-Yu Hsu, Hwynh Nguyen and Nhat Lan (2013)	Modified AOI algorithm of mining multivalued attribute data.	MAOI algorithm converts the data to Boolean bit, using K-map and discovers the generalized knowledge.	**Pros:**Need not go through the process of data warehousing. No professional knowledge is required. It can deal with multivalued attributes. It can find out rules of relational data base.
David Wai-lok Chenung, Ada Wai-	Rule based AOI algorithm.	Deductive generalization rules, computational	**Pros:**Induction process is efficient. Experiment
Chee Fu and Jiawei Han		rules and hybrid rules are used and for information loss problem backtracking algorithm is used.	performed on medium sized data.
Qingshuang Jiang, Syed Sibte Raza Abidi	Decision rule Acquisition Work bench	Discovering conjunctive normal form decision rules from annotated data sets.	**Pros:**Concise rules and smaller rule sets are sufficient for classification accuracy.
Jen-Yin, Shen-Tsu Wang, Chein-Hsin Lin (2015)	Syntax attribute oriented induction algorithm (SAOI).	Original data is converted into binary values and after the process simplified induction ratio is calculated.	**Pros:** The financial performance indicators are screened through a standard process through SAOI algorithm. Calculation does not require graphic representation of modified karnaugh map for induction. **Cons:**The time sequence of data is not considered.It is not possible to read the rule at the highest induction degree.
Lukas Tanutama (2014)	Selected protocol Analyzer	Collected data from network transactions, through protocol analyzer wire-shark then undergoes data preparation, selection, integration and redundancy cleaning of collected data and then data mining process is used.	**Pros:**Concerned mapping of corporate user activity as reflected through the protocol attribute. **Cons**:Peer-to-peer protocol to download files that could compromise corporate network security. It could also expose the company vulnerable to legal litigation. The bandwidth capacity could become insufficient as peer-to-peer downloads are bandwidth consuming.

V ANALYSIS AND CONCLUSION

The given content have explored the power of an interesting induction method, the AOI, in data mining. As a set-oriented,

generalization-based data mining method, AOI, extracts effectively different kinds of knowledge rules, including characteristic rules, discriminant rules, cluster description rules, and multiple-level association rules in relational database systems. The method is efficient, robust, with wide applications and extensible to knowledge discovery in advanced database systems, including object-oriented, deductive and spatial databases. The method is generalization-based technique and requires the availability of some background knowledge. Knowledge discovered by AOI has interesting applications at query database knowledge, cooperative query answering . multiple layered database construction and semantic query optimization. Knowledge discovery represents an important and promising direction in the development of data and knowledge-base systems.

REFERENCES

[1] David Wai-lok Cheung, Ada Wai-Chee Fu and Jiawei Han "A Rule-Based Attribute-Oriented Approach", Knowledge Discovery in Databases.

[2] M.V. Jagannatha Reddy, B.Kavitha, "Extracting prediction rules for loan default using neural networks through Attribute Relevance Analysis",International Journal of Computer Theory and Engineering, Vokume 2,No.4,August **2010**,1793-8201.

[3] Yu-Ying Wu., Yen-Liand Chen., and Ray-I Chang, "Generalized Knowledge discovery from Relational Databases", International Journal of Computer Sciences and Network Security, Vol.9 No.6,June **2009.**

[4] C.L.Carter and H.J.Hamilton, "Efficient attribute-oriented generalization for knowledge discovery from large databases", IEEE Transactions on Knowledge and Data Engineering 10(2)(**1998**)193-208.

[5] C. Hsu," Extending attribute-oriented induction algorithm for major values and numeric values", Expert Systems with Applications 27(2)(**2004**)187-202.

[6] Y.L.Chen and C.C Shen, "Mining generalized knowledge from ordered data through attribute-oriented induction techniques", European Journal of Operational Research 166(1)(**2005**)221-45.

[7] A.Savasere, E.Omiecinski and S.Navathe," Mining for Strong Negative Associations in a Large Database of Customer Transactions", Proceedings of the Fourteenth International Conference on Data Engineering.(**1998**).

[8] Qingshyang Jiang, Syed Sibte Raza Abidi," A Hybrid of Conceptual Clusters, Rough Sets and Attribute Oriented Induction For Inducing Symbolic Rules".

[9] Spits Warnars H.L.H., "Attribute Oriented Induction with Star Schema", International Journal of Database Management Systems(IJDMS)", Vol.2,No.2,May **2010.**

[10] Jiawei Han., Ywandong Cai and Nick Cercone., "Knowledge Discovery in Databases: An Attribute-oriented Approach"; Proceedings of the 18th VLDB Conference;**1992.**

[11] Hoi-Yee Hwang and Wai-Chee Fu.," Efficient Algorithms for Attribute-Oriented Induction"; KDD Proceedings;**1995**.

[12] Shu-Meng Huang., Ping-Yu Hsu., Hwynh Ngynh NguYen Nhat Lam, An Attribute-Oreinted Approach for knowledge Discovery from Relational Databases", Advances in information Sciences and Service Sciences (AISS) Volume5,Number3, Feb **2013.**

[13] Lukas Tanutama.," Frequency Count Attribute Oriented Induction of Corporate Network Data for Mapping Business Activity" ;EPJ Web of Conferences;**2014**.

[14] Son Dao., Brad Perry.," Applying a Data Miner to Heterogeneous Schema Integration", KDD Proceedings,**1995,**63-68.

[15] Jen-Yin Yeh., Shen-Tsu Wang., Chien-Hsin Lin., "Explore Financial Data Characteristics of Different Types of Enterprises During Rise in Stock Prices, Using a Semantic Attribute-Oriented Induction Algorithm"; 2012 International workshop on Information and Electronics Engineering (IWIEE).

[16] Spits Warnars.," Mining Frequent and Similar Patterns with Attribute Oriented Induction High Level Emerging Pattern (AOI-HEP) Data Mining Technique", International Journal of Emerging Technologies in computational and Applied Sciences(IJETCAS), 266-276,**2015**.

[17] Victor C.Cheng., C.H.C.Leung, Jiming Liu, Alfredo Milani ; "Probabilistic Aspect Mining Model for Drug Reviews"; IEEE Transactions on Knowledge and Data Engineering, Vol,26,No.8,Aug **2014**.

[18] Rafal A.Angryl, Jacck Czerniak, "Heuristic algorithm for interpretation of multi-valued attributes in similarity-based fuzzy relational databases", International Journal of Approximate Reasoning,**2010.**

[19] J.Isabella and Dr.R.M.Suresh,"Analysis and evaluation of feature selectors in opinion mining", International Journal of Computer Science and Engineering",Vol.3,No.6,Dec 2012-jan 2013.

Authors Profile

Mrs.S.Radha priya received her bachelors degree from Avinashilingam deemed University, Coimbatore in 1992 and Masters degree, MCA from Avinashilingam Deemed University in 1995 and M.Phil in computer science from Bharathiar University ,coimbatore .She is currently pursuing Ph.D.,and working as Assistant Professor in PG and Research department of Computer Science, Government Arts College(Autonomous), Coimbatore. She has published 5 research papers. Her main research work focuses on Data Mining, Big data Analytics, Opinion mining and sentiment analysis. She has 18 years of teaching experience and 4 years of research experience.

Dr.M. Devapriya received her M.Sc computer science from PSG College of Arts and Science, Coimbatore in the year 1991. She obtained her M.Phil and Ph.D., degree from Mother Teresa Women's University, Kodaikanal in the year 2001 and 2011 respectively. She is having around 20 years of teaching experience in various colleges. At present she is working as Assistant Professor in PG and Research Department of Computer Science, Government Arts College(Autonomous), Coimbatore. She has published more than 20 papers in National, International journals and conferences. Her area of interest includes Object Oriented Programming, Image processing, Multimedia, Network Security and Computer Networks. She is a member of various professional bodies.