International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor :6.887
Volume 5 Issue XII December 2017- Available at www.ijraset.com

A Review on Association Rule Discovery over Two View Dataset

Ms. D.N. Jadhao[1], Prof. J. R. Mankar[2]
[1]M.E. Student , [2]Assistant Professor) Department of computer Engineering, K. K. Wagh Institute of Engineering Education & Research, Nashik Savitribai Phule Pune University, Maharashtra, India.

***Abstract*: An object having more than one distinct views are called as multi view data. These views provide distinct information about an object. Two views Dataset is a form of multi-view dataset. In association discovery the relationship among these two views is analysed. In pattern mining, intersecting patterns are extracted from each view. This work aims to find the relationship among these two techniques and study the existing strategies to mine association rules from two view dataset.**
***Keywords*: association rules discovery, two view data, multi-view data, Minimum description length, Redescription mining.**

I. INTRODUCTION

An object has more than one view called as multi-view Data. These views provide distinct information about an object. Consider an example of movie where genres and actors provide one view whereas movie tags and rating provide the other view. This movie example has two views that convey different information regarding the same objects are called as Two-view dataset. This work focuses on two view dataset. In two view dataset, dataset attributes are split in two sets to provide two different views about the same object. These two views are two disjoint set of attributes.

Two views of same object provide alternative information. Data analyst has a task to define association among these views. Analyst analyses the pattern that collectively describes the structure of data. These patterns should be non-redundant and compact.

Most of Existing work has been done in association discovery and pattern mining. But these techniques are not applied to multi-view dataset. These methods can be directly applied to two view dataset. If two views are merged together for analysis then it leads to loss of information about views.

Association rules [1] are mined in transaction item dataset. But this technique suffers from pattern explosion. Pattern explosion is nothing but a large dataset generated as a result of mining. This large rule dataset is impractical to analyse and interpret manually for analyst. Hence there is need to generate compact and non-redundant rules.

These rules are added in translation table. To limit the number of rules, Minimum Description Length (MDL) Principal rule score is evaluated. This technique is applicable only on Boolean dataset. Along with the rule mining, rule direction mining is also important task. The derived rules can be unidirectional or bidirectional depending upon the dependency present in dataset. If there is two way unidirectional rule dependency then it is called as bidirectional rule or symmetric rule.

In the following section, study related to various multi-view data mining technique is proposed.

II. LITERATURE WORK

In this section, the general idea about multi-view data mining techniques and comparative analysis of those techniques are studied. Multi-view K-means, EM clustering techniques are proposed in Multi-view Data Mining. Pattern mining in two view dataset is proposed in Exceptional Model Mining (EMM). To overcome the problem of pattern explosion in pattern mining, Minimum Description Length (MDL) Principle is used. In further sections pattern mining in two view dataset along with association rule discovery and data compression model is discussed.

A. Multi-view Data Mining

Lot of work has been done in existing systems on multi-view data. But very few techniques focuses on rule-based association discovery and pattern mining.

Multi-view data clustering solution [3] is proposed based on the conditional independence property of multiple views. In this technique multi-view K-means and EM clustering techniques are proposed.

The concept of parallel universe is proposed in [11]. In this technique multiple descriptor spaces are analysed simultaneously and a global model is generated based on local models. This technique provides new insights about the model with respect to universe

specific patterns and overlapping patterns. The process includes the analysis of structure of every individual view. This contradicts the basic concept of multi-view analysis as: multi-view analysis focuses on pattern and structure identification across different views.
Survey on Subspace clustering is proposed by Kriegel, Kröger, and Zimek [10]. Subspace clustering is the extension of classical clustering techniques in which clusters are generated in different subspaces within the given dataset. For such type of clustering high dimensional data is required.
This technique includes feature selection and clustering. This technique is advantageous than feature selection and clustering technique. This technique selects the relevant dimensions and allows the clusters generation process. It generates different views form data but these views contain overlapping attributes and it does not try to distinguish among generated views.
Survey is published on relationship between subspace clustering and pattern mining [12]. In this survey Jilles Vreeken and Arthur Zimek stated that there is strong relationship between these two approaches. These two approaches have same objective: extraction of interesting and non-trivial, unknown knowledge form data. Solutions for subspace clustering problem are also applicable for pattern mining problems and vice-versa.

B. Pattern mining for two-view data

Pattern mining in two view dataset is proposed in [5][9][13]. Exceptional Model Mining (EMM) [9] strategy is proposed for subgroup discovery. EMM finds the subset of data by focusing on model rather than targeting a single variable. It uses regression as a classification technique. EMM defines asymmetric rules based on the targeted model.
Redescription Mining technique is proposed in [5] [13]. Redescription mining is technique to find multiple descriptions of same entity.
Conceptual discussion related to importance of Redescription mining, feasibility of Redescription, problem faced by larger machine learning community is proposed in [5].
Along with Boolean data, categorical and real-valued data analysis [13] is done for finding Redescription of entity.
Unlike EMM technique RM finds the relationship in two ways i.e. bidirectional relationship identification. Hence it generates bidirectional rule called symmetric rule. Bidirectional rules extraction from the different views represents the complete association in data. Hence Redescription mining is good example of association rule discovery over two view dataset but this technique identifies individual high confidence rules without considering the relationship among discovered rules. This creates redundancy in mined rule data and bulk data is get generated.

C. Association Rule Discovery

Association rule discovery can be symmetric or asymmetric. In symmetric rule discovery, bidirectional rules are identified. In asymmetric rule discovery, unidirectional rules are discovered. There are lots of disadvantages of asymmetric rule discovery. On biggest issue is pattern explosion. Hence there is need to define some threshold or support value to filter these rules. But it is difficult to tune appropriate threshold or support value. To overcome this problem, the concept of closed frequent item sets [4] is proposed. This technique exponentially reduces the number of rules from the existing approach.
Constraint based pattern mining technique is proposed by Luc De Raedt and Albrecht Zimmermann [6].This is useful for classification purpose. It is single property of interest as a target and based on this target rules are discovered.
Statistical testing method [7] is applied for pattern discovery. To find patterns in a dataset it applies various statistical test using a Bonferroni correction. This rule selection strategy generates non-redundant strict rules form a given dataset.
For Association rule discovery whole dataset is scanned iteratively and hence it is time consuming. New algorithm is proposed for fast association rule discovery [2]. This technique discovers the rules in single dataset scan and improves the efficiency of discovery technique.

D. MDL : Model selection Technique

To overcome the problem of pattern explosion in pattern mining, Minimum Description Length (MDL) [1] Principle is used. Using this principle small non-redundant pattern based models are extracted. These models provide ease to the analyst to analyse the relationship among data manually. The generated compression models are also useful in clustering.
Non-redundant, small set of rules are extracted using association rule discovery over two view Boolean dataset [8]. The rules are unidirectional as well as bidirectional. Two views are two disjoint vocabularies. This uses translation table as a data-structure to save translation rules based on MDL principle. In this technique three translation rule generation algorithm are proposed to find

good translation table with moderate number of attributes. The comparative results show the good performance over existing technique but this technique is applicable on for Boolean data.

III. CONCLUSION

Existing work has been done in pattern mining and association discovery with different context. These techniques are not directly applied on two view dataset. Translation table proposes a solution for association rule mining over two view dataset. But this technique only focuses on Boolean dataset values. There is need to discover small non-redundant rules from categorical and real valued dataset. The above technique can be extended to multi-view association rule discovery model.

III. ACKNOWLEDGMENT

Authors would like to thanks Prof. Dr. K. N. Nandurkar, Principal and Prof. Dr. S. S. Sane, Head of Department of Computer Engineering, K.K.W.I.E.E.R., Nashik for their kind support and suggestions. We would also like to extend our sincere thanks to all the faculty members of the department of computer engineering and colleagues for their help.

REFERENCES

[1] RakeshAgrawal, Tomasz Imieli ´ nski, and Arun Swami.” Mining association rules between sets of items in large databases.” In Proc. of the 1993 ACM SIGMOD International Conference on Management of Data (SIGMOD’93), pages 207–216. ACM Press, 1993.
[2] Mohammed J Zaki, SrinivasanParthasarathy, MitsunoriOgihara, and Wei Li. “New algorithms for fast discovery of association rules.” In Proc. of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD’97), pages 283–286. ACM, 1997
[3] Steffen Bickel and Tobias Scheffer.” Multi-view clustering.” In Proc. of the 4th IEEE International Conference on Data Mining (ICDM’04), pages 19–26. IEEE Computer Society, 2004.
[4] Mohammed JaveedZaki. “Mining non-redundant association rules.” Data Mining and Knowledge Discovery, 9(3):223–248, 2004.
[4] LaxmiParida and NarenRamakrishnan. “Redescription mining: Structure theory and algorithms.” In Proc. of the 20th National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference (AAAI’05), pages 837–844. AAAI Press / The MIT Press, 2005.
[5] Luc De Raedt and Albrecht Zimmermann. “Constraint-based pattern set mining.” In Proc. of the 7th SIAM International Conference on Data Mining (SDM’07), pages 237–248. SIAM / Omnipress, 2007
[6] Geoffrey I. Webb. “Discovering significant patterns.” Machine Learning, 68(1):1–33, 2007.
[7] Peter D. Gr¨ unwald. “The Minimum Description Length Principle.” MIT Press, 2007.
[8] Dennis Leman, Ad Feelders, and Arno Knobbe. “Exceptional model mining.” In Proc. of the European Conference on Machine Learning and Knowledge Discovery in Databases, Part II, (ECML/PKDD’08), pages 1–16. Springer, 2008.
[9]] Hans-Peter Kriegel, Peer Kr¨oger, and Arthur Zimek. “Clustering high-dimensional data: A survey on subspace clustering, pattern based clustering, and correlation clustering.” ACM Transactions on Knowledge Discovery from Data, 3(1), 2009
[10] Bernd Wiswedel, Frank H¨oppner, and Michael R. Berthold. “Learning in parallel universes.” Data Mining and Knowledge Discovery, 21(1):130–152, 2010
[11] JillesVreeken and Arthur Zimek. “When pattern met subspace cluster.” In Emmanuel M¨ uller, Stephan G¨unnemann, Ira Assent, and Thomas Seidl, editors, MultiClust@ECML/PKDD, volume 772 of CEUR Workshop Proc., pages 7–18. CEU WS.org, 2011.
[12] Esther Galbrun and Pauli Miettinen. “From black and white to full color: extending redescription mining outside the boolean world.” Statistical Analysis and Data Mining, 5(4):284–303, 2012.
[13] Matthijs van Leeuwen and Esther Galbrun ,"Association Discovery in Two-View Data," IEEE Transactions on Knowledge and Data Engineering,Vol. 27, pp. 3190 - 3202, Issue. 12, Dec. 2015.