ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

Survey on Multi Relational Graph Classification

Smital Deore, Prof. Soniya Mehata

Dept. of Computer Engineering, Alard college of Engineering and Management, Pune, India

ABSTRACT: An increasing number of data mining applications involve the analysis of complex and structured types of data and require the use of expressive pattern languages. Many of these applications cannot be solved using traditional data mining algorithms. This observation forms the main motivation for the multi-disciplinary field of Multi-Relational Data Mining (MRDM). Unfortunately, existing “upgrading” approaches, especially those using Logic Programming techniques, often suffer not only from poor scalability when dealing with complex database schemas but also from unsatisfactory predictive performance while handling noisy or numeric values in real-world applications. However, “flattening” strategies tend to require considerable time and effort for the data transformation, result in losing the compact representations of the normalized databases, and produce an extremely large table with huge number of additional attributes and numerous NULL values (missing values). As a result, these difficulties have prevented a wider application of multi relational mining, and post an urgent challenge to the data mining community. To address the above mentioned problems, this article introduces a multiple view approach—where neither “upgrading” nor “flattening” is required— to bridge the gap between propositional learning algorithms and relational databases and current research challenges in the field of Multi relational classification based on Multi View Learning.

I.INTRODUCTION

Most real-world data are stored in relational databases. So to classify objects in one relation, other relations provide crucial information. Traditional mechanism cannot convert relational data into a single table without expert knowledge or loosing essential information. Multi-relational classification automatically classifies objects using multiple relations. Vast amounts of real world data are routinely collected into and organized in relational databases. Most of today’s structured data is stored in relational databases. Thus, the task of learning from relational data has begun to receive significant attention in the literature. Unfortunately, most methods only utilize “flat” data representations. Thus, to apply these single-table data mining techniques, we are forced to incur a computational penalty by first converting the data into this
“flat” form. Patterns of activity that, in isolation, are of limited significance for classification but, when combined/related, will improve the performance of system. Multi relational classification aims at discovering useful patterns across multiple inter-connected tables (relations) in a relational database. Traditional machine learning approaches assume a random sample of homogeneous data from single relation but real world data sets are multi-relational and heterogeneous. Current solution does not scale well and cannot realistically be applied when considering database containing huge amount of data.

II.RESEARCH BACKGROUND

The Multi View Classification (MVC) approach employs the multi-view learning framework to operate directly on multi-relational databases with conventional data mining methods. The approach works as follows.
1: Information Propagation Stage: The Information Propagation Stage, first of all, constructs training data sets for use by a number of view learners, using a relational database as input. The Information Propagation Element propagates essential information from the target relation to the background relations, based on the foreign key links. In this way, each resulting relation contains efficient and various information, which then enables a propositional learner to efficiently learn the target concept.
2: Aggregation Stage: After the Information Propagation, the Aggregation Stage summarizes information embedded in multiple tuples and squeeze them into one row. This procedure is applied to each of the data sets constructed in the Information Propagation Stage. In this stage, aggregation functions are applied to each background relation (to which the

essential information from the target relation were propagated). By applying the basic aggregation functions in SQL, new features are created to summarize information stored in multiple tuples. Each newly constructed background relation is then used as training data for a particular view learner.

3: Multiple Views Construction Stage: In the third phase of the MRC algorithm, the Multiple Views Construction Stage constructs various hypotheses on the target concept, based on the multiple training data sets given by the Aggregation Stage. Conventional single-table data mining methods (view learners) are used in order to learn the target concept from each view of the database separately. In this stage, a number of view learners, which differ from one another, are trained.

4: View Validation Stage: All view learners constructed in the Multiple Views Construction Stage is then evaluated in the the View Validation Stage. The trained view learners need to be validated before being used by the meta learner. This processing is needed to ensure that they are sufficiently able to learn the target concept on their respective training sets. In addition, strongly uncorrelated view learners are preferred.

5: View Combination Stage: In the last step of the MRC strategy, the resulting multiple view learners (from the View Validation Stage) are incorporated into a meta learner to construct the final classification model. The meta learner is called upon to produce a function to control how the view learners work together, to achieve maximum classification accuracy. This function, along with the hypotheses constructed by each of the view learners, constitutes the final model.

Since the MVC algorithm is based on the multi-view learning framework, it is able to use any conventional method to mine data from relational databases. Popular ensemble methods such as boosting, bagging and stacking, focus on constructing different hypotheses from subsets of learning instances. In contrast, an important characteristic of the multi-view learning is that this approach is more interested in learning independent models from disjoint features of the training data. That is, multi-view learning learns separate views from various disjoint-feature-based aspects of the data, leading to independent views on the target concept. node individually). Once the time series of the consecutive-graph similarities is obtained, Quality Control with Individual Moving Range [Montgomery, 1997] is used to spot the anomalous daily ENRON-graph instances. In contrast to the most of the previous works that detect anomalous graph instances, the following algorithms spot anomalous nodes in a graph sequence. The key idea in [Akoglu and Faloutsos, 2010] is the following: A node is anomalous at some time frame, if its “behavior” deviates from its past “normal behavior”. The authors build the “behavior” of the nodes by extracting various egonet node features (e.g., weighted and unweighted in- and out-degree, number of neighbors, number of triangles) from each snapshot of the graph sequence, and create a correlation matrix of node “behaviors” at each time window using Pearson’s correlation coeffi- cient. For each correlation matrix (one per time window), the principal eigenvector, which has one entry per node, is computed. By placing all the corresponding entries of the eigenvectors in a vector, the “eigen-behavior” vector of each node is obtained, and compared against its typical “eigen-behavior”, which is found by using averaging in the past time windows or SVD. The similarity between the “behaviors” is evaluated using the Euclidean dot-product. For low similarity between a node’s “behavior” and its past “behaviors”, the corresponding time window is reported as anomalous.

III.LITERATURE SURVEY

Along the same lines, the authors in [Papadimitriou et al., 2008] introduce five graph similarity functions for directed, time-evolving web graphs: vertex/edge overlap similarity, vertex ranking, vertex/edge vector similarity, sequence similarity, and signature similarity. Among these metrics, the one that performs best in terms of change detection in web graphs is the Signature Similarity (SS), which is based on the Sim Hash algorithm. This algorithm uses as features the nodes and edges of the input graphs, weighted appropriately by their Page Rank. [Berlingerio et al., 2012] use a graph similarity approach for discontinuity detection in daily instances of social networks. In a nutshell, NETSIMILE consists of three phases: (i) Feature Extraction. The focus is on local and ego net-based features (e.g., number of neighbors, clustering coefficient, average of neighbors’ degrees); (ii) Feature Aggregation. The node × features matrix of the first phase is converted to a single “signature” vector that consists of the median, mean, standard deviation, skewness and kurtosis of each extracted feature over all the nodes in the graph; (iii) Comparison. The signature vectors are compared using the Canberra Distance, and a single similarity score is produced for consecutive timestamps of the graph sequence. The days that have low similarity score with the surrounding days are characterized as anomalous. Another recent work, [Koutra et al., 2013b], proposes a complex graph-featurebased similarity approach, DELTACON, for discontinuity detection, which enjoys several desired properties. The intuition behind the method is to compare the pairwise node affinities of consecutive snapshots of the graph sequence. These node affinities are computed in this work by a fast variant of Belief Propagation [Koutra et al., 2011]. The matrices of pair wise node similarity matrices are then compared using the Matusita Distance (which is related to the Euclidean Distance), and the distance is finally transformed to similarity. A faster algorithm that avoids computing all the pair wise similarity scores is also proposed, and it is based on the idea of finding the similarity of all the graph nodes to non-overlapping groups of nodes (instead of each)

IV.ISSUES IN MULTI GRAPH LEARNING

The major challenges come from, the large high dimensional search spaces due to many attributes in multiple relations and the high computational cost in feature selection and classifier construction due to the high complexity in the structure of multiple relations [1].

- The idea of using heterogeneous learners will further increase understanding of the multiple views learning scheme.
- To study applying data preprocessing techniques such as feature selection in order to further improve the performance of the MVC algorithms.
- Also, prior work has shown that more complex aggregation functions can improve the generalization accuracy of relational learning. It would also be interesting to investigate this for MVC.
- It would also be interesting to examine the influence of different model combination techniques and view validation strategies
- Study different "goodness" heuristic measurements and their impact on these algorithms.
- Evaluating the method against learning tasks with more than two classes will be interesting to investigate.
- Study how the total tuples and imbalanced ratio in each resulting view impacts the result of the final combination model.
- Also, it would be very interesting to further investigate relational schemas with composite keys.
- In addition, another area for future work is the study which extends this approach to deal with relational data stored in the form of graph and social network.
- To investigate the behavior of the multiple view learning frameworks, while developing more sophisticated view construction techniques. In other words, the view construction procedure will search the entire feature space in order to determine how to better group the features into different views.
- Another area for future work is employing relational data mining algorithms as hypotheses construction methods, rather than generating relational features and then applying single-table learning strategies, while training a set of diverse individual view learners.
- Research has shown that popular ensemble methods such as Bagging, Boosting, and Stacking can significantly improve the predictive performance of an individual model in some cases. Through employing relational mining algorithms as view learners in the multiple view learning framework, will be able to explore the impact of popular ensemble techniques on the relational learning strategies.
- A novel approach is needed which can conduct both Feature and Relation Selection for efficient multi-relational classification.
- Join Graph can be further pruned to improve the classification time, by eliminating tables that may not really contribute much to the overall classification task.
- MVC can be extended to include selection of the right classifier at the table level.
- Consequently, no guidelines are available to select the best classifier for a particular type of data.
- In future, experimentation with different view combination techniques, such as majority voting and weighted voting can be future investigated.

V.CONCLUSION

In this paper we have demonstrated that Multi View Learning is inherently more powerful than other approaches of relational learning. There clearly is a large class of Data Mining problems that cannot be successfully approached using another relational learning without transformation. These problems, which can be characterized by the presence of relational structure within the database they deal with, can successfully be approached by the Multi View Learning. The presented overview of multi-faceted graph visualization aims to survey a vast field of visualization. To do so, despite the extent of the research in this field, we have devised a categorization based on different visual combination modalities and merely highlighted selected visualizations for each category. We have further presented our efforts to relate the existing surveys in this field to our categorization. While there exists no one-to-one mapping between those surveys and our categories, the overlap we could establish is considerable and the differences are mainly due to particularities of individual facets. We deem this “meta survey” to be an important step towards a better understanding of the space of possible visualization solutions for multi-faceted graphs altogether. This overview further points into directions for future research in the following three aspects:

REFERENCES

[1] Miao Zou, Tengjiao Wang,A General Multi-relational Classification Approach Using FeatureGeneration & Selection, Advanced Data Mining & Applications, Lecture Notes in Computer Science,2010,Vol 6441/2010,21-33.

[2] Jing-Feng Guo, An Efficient Relational Decision Tree Classification Algorithm, Third International Conference on Natural Computation (ICNC 2007).

[3] Yin, X., Han, J., Yang, J., and Yu, P.S., CrossMine: Efficient Classification across Multiple Database Relations, in Proceedings of the 2004 International conference on Data Engineering (ICDE'04), Boston, MA, 2004.

[4] Hongyan Liu, Xiaoxin Yin, and Jiawei Han, “d” , MRDM-2005, Chicago, 2005

[5] Arno Jan Knobbe, A Ph.D Theis on Multi relational Data Mining SIKS Dissertation Series No. 2004-15.

[6] Amir Netz, Integration of Data Mining and Relational Databases, Proceedings of the 26th International Conference on Very Large Databases, Cairo, Egypt, 2000

[7] Andreas Heß and Nick Kushmerick, Iterative Ensemble Classification for Relational Data: A Case Study of Semantic Web Services, 2007

[8] Anneleen Van Assche, Improving The Applicability Of Ensemble Methods In Data Mining, PhD Thesis, ISBN 978–90–5682–896–7, Katholieke University Leuven – B-3001 Heverlee (Belgium) 2008.

[9] Guo, H., Herna, L., Viktor.. Multirelational classification: a multiple view approach, Knowl. Inf. Systems, vol.17, pp.287–312, Springer-Verlag London. 2008

[10] PAN Cao, WANG Hong-yuan,,Multi-relational classification on the basis of the attribute reduction twice, Journal of Communication and Computer, ISSN 1548-7709, USA, Nov. 2009, Volume 6, No.11 (Serial No.60)

DOI: 10.15680/IJIRCCE.2015. 0312116